Road Accident Prediction Model Using Data Mining Techniques

Dattatray G. Takale[1*], Shubhangi D. Gunjal[2], Vajid N Khan[3], Atul Raj[4], Satish N. Gujar[5]

Abstract

Road Accident is an all-inclusive disaster with consistently raising pattern. In India according to Indian road safety campaign every minute there is a road accident and almost 17 people die per hour in road accidents. There are different categories of vehicle accidents like rear end, head on and rollover accidents. The state recorded police reports or FIR's are the documents which contains the information about the accidents. The incident may be self-reported by the people or recorded by the state police. In this paper the frequent patterns of road accidents is been predicted using Apriori and Naïve Bayesian techniques. This pattern will help the government or NGOs to improve the safety and take preventive measures in the roads that have major accident zones. Data mining (DM) techniques (artificial neural networks (ANNs) and support vector machines (SVM)) were used to model accident and incident data compiled from the historical data. Based on the R-Tools, results were compared with those from some classical statistical techniques (logistic regression (LR), revealing the superiority of ANNs and SVM in predicting and identifying the factors underlying accidents in toll road.

KeyWords: Apriori, Naïve Bayes, Pattern Prediction, Road Accident

DOI Number: 10.48047/NQ.2022.20.16.NQ880299 **NeuroQuantology2022; 20(16):2904-2911**

Introduction

Traffic accidents have been identified as a major contributor to morbidity and death rates in India. About 25,859 people were killed and 100,000 were injured in traffic-related incidents in India in 2017. The costs to society of traffic accidents may be enormous, both in terms of inconvenience to drivers and repair bills for public infrastructure. As a matter of fact, it happens on a global scale, with 1.17 million lives lost annually due to traffic accidents and another 10 million crippled or injured. In recent years, more and more incidents involving motor vehicles have been reported throughout India, according to data from the Indian National Police's Traffic Corp. That figure has been steadily climbing since 2014. The number of reported incidents increased from 2014's 95,906 to 2015's 98,970 and 2016's 105,374. The number of reported cases in 2012 peaked at 117,949 and declined to 100,106 in 2013, thus these increases stand in sharp contrast to those trends. The increasing rate of accidents from one year to the next has been linked to a variety of possible causes. The human aspect, the vehicle component, and the environment all play a role. When these four factors are considered together, it is widelyaccepted that human error is the primary cause of road accidents. However, the overall number of car accidents remains relatively constant year over year. There is undeniable evidence that methods that can accurately identify potential accident risk factors are in high demand.

When compared to other kinds of highways, the accident rate on toll roads is higher than it is on other kinds of roads. In connection to the elements that were discussed before, about 75% ofaccidents that occurred on toll roads were the result of human causes [2].

Corresponding author: Dattatray G. Takale
Address: [1]Assistant Professor, Department of Computer Engineering, Vishwakarma Institute of information Technology, SPPU, Pune, [2]Associate Professor, Department of Mechanical Engineering, Jaihind College of Engineering SPPU, Pune, [3]Assistant Professor, Department of Computer Engineering, Dhole Patil College of Engineering, Wagholi, Pune, [4]Lectural, Department of Computer engineering, Government Polytechnic, Sikandra, Kanpur, Dehat, [5]Professor, Department of Computer Engineering, JSPM College of Engineering, SPPU, Pune
E-mail: dattatray.takale@viit.ac.in

On toll roads, individual vehicle operators have a wide range of physical and mental abilities, different perceptions of risk, different reactions to external stimuli, and their operating abilities may be further complicated by varying degrees of self-inflicted impaired driving. Moreover, there is a high potential for accidents on toll roads because of the large number of drivers who are impaired. Because the majority of these countries either do not keep adequate records or do not keep any records at all, very little is known about the primary factors that contribute to highway traffic accidents in developing countries [3]. This is the primary reason why there is so little information available.

The creation of a TRA prediction model is the essential step that must be taken in order to enhance the current level of safety on India's toll roads. This model's purpose is to offer estimates or predictions of the number of traffic accidents that are devoid of impact bias caused by the phenomena of regression to the mean. To establish the impact of regulating one standard on the overall safety level of the road, it is vital for the toll operator business to have a thorough grasp of the link between the standards and the actual number of accidents that occur. In addition to this, it may act as a guide for a programme that improves toll roads, which is particularly useful for developing nations like India who are having trouble constructing roads that are in accordance with their toll road standard [4]. If the authority can determine which aspects of toll road standards contribute to the actual number of accidents, they will be able to properly direct highway expenditures towards the rehabilitation of roads that are insufficient. While the accident database system at the operator level is being upgraded, the creation of a TRA prediction model may aid other operators who have poor accident databases. To that end, the creation of a TRA prediction model has as its ultimate objective the enhancement of the current level of toll road traffic safety.

The TRA prediction model that was created for JORR is presented in this document. It is anticipated that the use of the model that was built would identify and prioritise the factors that contributed to the accident. Traditional statistical models relied on accident records to make predictions about the present state of road safety conditions; one of the limitations of these models is that the whole circumstances surrounding an accident cannot be known with certainty [5]. The TRA prediction model has been constructed by using a non-linear regression model with ANNs and SVM technique. This has been done on the basis of the requirement that the accident rate is discrete, and it cannot be negative and unique. The created model makes use of substantial amounts of data that are the property of toll road operators by processing DM data using r-miner. It is vital to build a nice and continuous pattern [6] since the recording of enormous volumes of data makes this need imperative. Therefore, the collection of data that has been obtained via the measurement of accidents may be employed in an organised and scalable scheme to assist the interpretation and prediction of accurate data [7]. Because it is vital to adopt new methods and make use of the most recent technology, The best model performance among the DM findings is what is utilised to construct the TRA prediction model for India's toll roads.

Literature Survey

It is a tremendous task to use proper system ways to strengthen safety measures, especially when the total number of existing data regularly rises. You may choose some previously conducted research that focuses mostly on the facts pertaining to local traffic accidents based on the current situation. One of the countries that has a significant number of people killed and injured in car accidents is Iraq. During the previous three years, road accidents in Iraq have resulted in an average of 24,000 fatalities every day (three persons per hour), and around 240,000 people are wounded each year as a direct consequence of those incidents. Because of this fact, a group of authors decided to categorise the factors that have the most significant impact on the degree to which drivers are hurt in automobile collisions. They utilised accident information taken from the records kept by the Ministry of Information and Technology of the Iraq Traffic Police Department during the years of 2006 and 2008. The desired outcome is a severity level of three: There were no injuries, illnesses, or fatalities recorded; the data covers more than 169,000 drivers. The decision factor is determined by the variable importance percentage (VIM) of each of the outputs generated by the CART (Classification and Regression Tree) algorithm. According to the findings, the use of a seat belt is the single most important factor in determining the severity of injuries sustained in a car accident; conversely, failure to use a seat belt substantially raises the risk of suffering serious injuries or even passing away.

Similar techniques were utilised by Chang and

Wang [3] to develop CART models that analyse the relationship between injury severity and driver/vehicle characteristics, path/environmental variables, and collision factors. This was done so that a connection could be made between the two variables. From their examination of collision data from 2001 in Taiwan (Taipei), they concluded that vehicle type was the most important element in establishing incident severity. Moreover, the author Yau-Ren Shiau and his coworkers utilised information gathered in central Taipei in their study. [4] We will use fuzzy robust principal module analysis, back propagation neural network, and Logistic regression mining to identify the most important contributing elements to the over 2,400 traffic road incidents that occurred in 2011. Combining the first two methods outlined above yields the maximum accuracy, at 85.89%.

Nayak et al. [5] we looked at over 42,000 records from the Queensland Department of Transportation and Australia's roads and traffic accident database between 2004 and 2007 and drew conclusions on the most dangerous routes to take and the causes of these events. The author employs a taxonomy that divides it into several phases of accident propensities (including certain streets, designs, or situations that have higher collision proportions than others), and this inspires a collaborative effort by another author [6]. The authors used a variety of statistical methods in their research, including a chi-squared test tree, regression trees based on f-tests, neural networks, logistic regression, and Bayesian models. If you compare the results of various test simulations to those of decision trees, you'll find that the former fare significantly better.

An additional research has been conducted on the topic of road traffic safety in the UAE. Road traffic accidents are the second largest cause of mortality in the UAE [7], with over 600 persons losing their lives annually as a result of automobile-related catastrophes. Between the years 2008 and 2010, the author made use of more than 1,800,000 data, each of which had 19 characteristics that covered various aspects of accidents, drivers, and the route or mode of transportation circumstances. The target attribute has the values "catastrophic," "severe," "moderate," and "slight," in addition to "none." The end analytical model was produced using the WEKA support tool and the subsequent method, which included a multilayer perceptron, Bayesian network, and J48 decision tree. A multi-layer perceptron was able to produce the model with the highest degree of accuracy (more than 99%), while a Bayesian network was able to produce the model with the highest rate of speed (0.17 seconds).

Collectively, the writers from the Transportation Department of the Hong Kong Government utilised comparable methods [8]. This data set includes 34,000 more records; genetic algorithms are used for feature selection; classification tests are also carried out using the WEKA mining kit; and J48 offers a more accurate classification model to predict the degree of injuries sustained in automobile collisions.

[9] outlines the fundamental parameters that should be considered while using data mining strategies for the purpose of improving traffic safety in the Andalusia area of Spain. The whole of the process is predicated on a component of the road that is referred to as the susceptibility improvement element. This component is described as a stretch of roadway that demonstrates road conditions that deviate from the best road safety criteria (eg, layout, signals, crossings, tunnels, etc.). In order to merge the datasets of the three distinct original datasets—element, roadways, and collapses—a variety of data mining techniques, such as decision trees, neural networks, and association rules will be put into practise.

Last but not least, the study of Flach et al. [10] was defined because the authors employed the same data set that was needed by the Hampshire National Council in the UK. Furthermore, the writers wished to understand the profile of road safety progress over the course of the previous 20 years. Seven methodological teams worked together to carry out the research and use a variety of data mining approaches, such as time series clustering, text mining, multi-relation data mining, subgroup finding, and association rule learning. Some of the results that were acquired may serve as examples, including the following: Imagine that there is an accident at a time when the speed limit is 60 miles per hour, it is after 8 o'clock at night, and the car only has two wheels. The collision will likely result in fatalities. This accounts for 70,000 of the total 100,000 records. In addition, the findings of an examination into the connection between motor vehicle collisions and the time of day or the day of the week were as follows: The majority of the mishaps happened between the hours of 8 a.m. and 4 p.m. and 5 p.m.

All of the studies that have been proposed have a few things in common, including the following: they

use small data samples for specific time periods or local areas, which, when taking into account the specifics of those areas, can be seen as a positive; however, on the other hand, some significant problems that are universal may be overlooked; Nearly all studies produce highly accurate models, but this value is highly dependent on the feature range method that was used. In many cases, it appears that traditional data mining algorithms are used as decision trees because they are easy to understand and interpret. Experiments involving vast volumes of data, such as those detailed in this article, call for the development of new approaches and settings that are conducive to their implementation. As a direct consequence of this, the in-memory technique [11], which was stated in the platform H2O and R languages, was selected as a suitable grouping for a specific objective.

Proposed Work

As part of our research, we have developed an application that, using the information that is currently accessible on road accidents, is able to provide predictions regarding the likelihood of the occurrence of accidents. In order to acquire a dataset from this road accident data, pre-processing of the data is performed. During the step known as "data pre-processing," steps such as "cleaning," "where null and garbage values are removed," and "normalisation" of the data are carried out. This is followed by "feature selection," where only the features that are pertinent to the original dataset are chosen for inclusion in the final dataset. Following this, the dataset is analysed using a variety of data mining methods. This dataset goes through the process of clustering. Following that, the clusters are put through further algorithms such as Support Vector Machines (SVM) and Apriori. The data that is being used for the study has an unknown distribution, and we need to sort out the frequent and infrequent items that are included in the dataset. Therefore, the former (SVM) is used to predict the probable risk of accidents, and the latter (Apriori) is applied to perform rule mining, which is to generate a frequent item set based on given support and confidence values. Both of these methods are being used in conjunction with each other. Rules have been developed after taking into account the many diverse combinations of elements that have led to accidents of varied sorts and degrees of severity on a variety of roadways and under a range of climatic conditions. The selected support and confidence values for the often occuring item sets suggest that there is a larger possibility of a certain combination of qualities resulting to an accident. This is the case since these values reflect the frequency with which these item sets occur. For instance, based on the rule mining that was done, the probability of an accident happening even when the weather is nice in a junction due to excessive speeding is high, and according to the training dataset, it has the potential to end in a fatality. This conclusion was reached after analysing the data. A SVM classification has been employed so that each accident occurrence may be characterised as belonging to either a high risk category or a low risk category. In order to get the findings that have been interpreted, a number of data mining methods and exploratory visualisation approaches have been used to the accident dataset. Figure 1 depicts the overall architecture of the model that was used for the research presented in this article. The estimate of the many different components that contribute to road accidents helps to identify how much of a role each factor has in creating accidents. These analysis and studies aid in giving solutions with the goal of lowering the accident rate and the number of fatalities that occur as a result of such accidents.

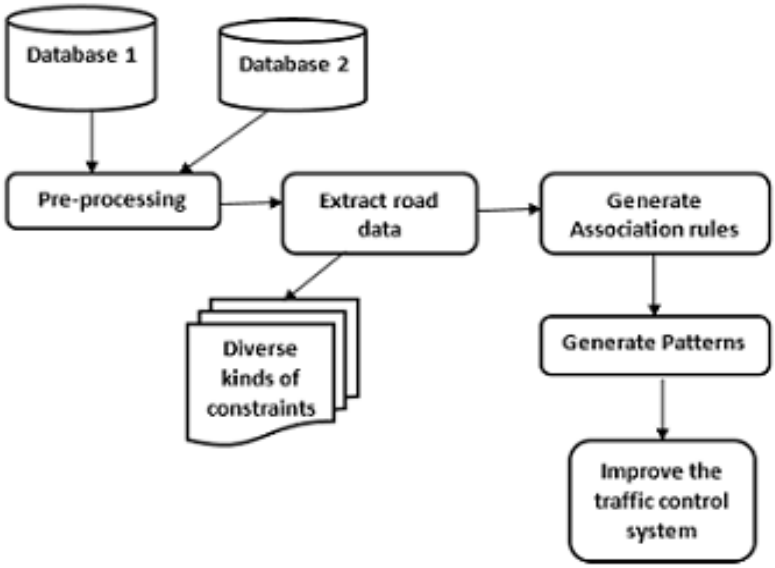

Fig.1: Architecture of the implemented model

A. Database

The dataset that was used in this investigation was retrieved from the Open Government Data (OGD) Platform, which is located in India. The model was developed with the use of datasets including information on incidents that occurred in the Bangalore area during the years of 2014 and 2017. This dataset includes details such as the date, time, and location of accidents, the nature of the accident, such as whether it was a head-on collision or caused due to over-speeding, skidding, or other causes, the type of the road, such as whether it was a straight road or a curved road, the number of lanes that were present, whether it was a junction of multiple roads, the number of fatalities, and other similar details. For the purpose of the current research, a model may be developed based on the interaction of all of these components. However, this cannot be modelled by using a straightforward deterministic model; rather, in order to get the desired outcomes, a stochastic modelling approach is required. The addition of supporting machine learning algorithms to data mining methods is thus required as a result of this need.

The unprocessed raw data on road accidents that was collected is pre-processed so that it can be used to create the dataset that will be fed into the model. The model is then trained further using the training data, and it is made to make predictions regarding the possible risk of accidents for an area that a user will input. The user is also presented with a graphical representation of the data depending on the statistics that were collected. The operation of this model can be broken down into four modules: rule mining, risk prediction, graph plotting, and new data entry. Each module performs a different function. The Apriori Algorithm is utilised in the process of rule mining. This allows us to produce a frequent item set based on the dataset that is used as input. The SVM (Support Vector Machine) Algorithm, which is most commonly employed for classification, is utilised in the process of risk prediction. The categorisation work is done once the sample data set has been taken as an input. This module is able to forecast the likelihood of accidents occuring in a certain location. The bar charts are generated by the plot graph based on the weather, previous accidents, and the factors that caused the accidents. The New Data Entry module is what's used to report any new accidents that have taken place.

C. Software and Languages Used

Python is the programming language that was used in the development of the application, and the Anaconda Spyder software was used for the actual implementation.

D. Simulation

R tools are utilised in order to carry out the simulation. In order to obtain interpreted results from the accident dataset, a number of data mining techniques and exploratory visualisation techniques are applied. A user interface that is interactive can be developed with the help of the R tools. Therefore, by plotting a variety of graphs, charts, and other statistical and graphical representations, we will be able to conduct an analysis of the various factors that contributed to the accidents.

Results And Discussion

Urban Versus Rural

Analysis of urban and rural road traffic accidents shows that rural areas are more prone to traffic accidents. Compared with the number of accidents in rural areas, the total number of urban road accidents is small. The table shows that rural roads require substantial investment and improvement to reduce accidents in rural areas.

Table 1: Number of people killed/injured in urban and rural areas

Category	Accident	Killed	Injured
Rural	403598	121126	2589
Urban	15461	4091	890

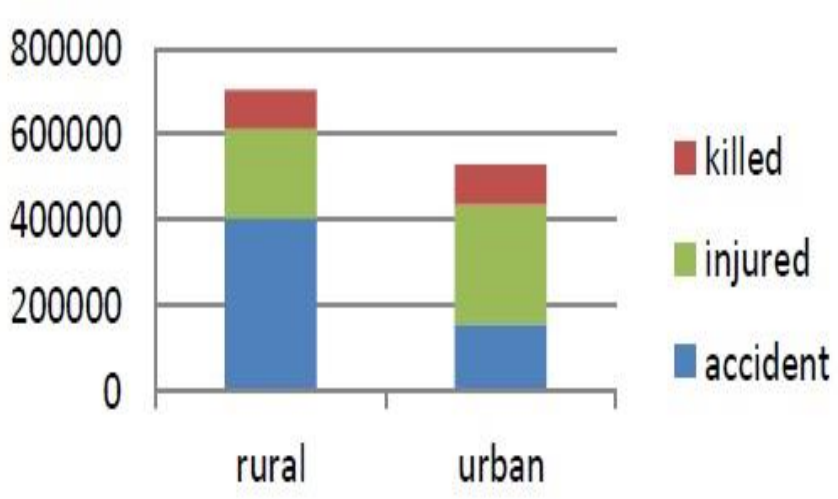

Figure 2: x- axis: accidents Y –axis-person killed on road type

2. Age of Persons Killed (Gender wise) in Road Accidents

In the deaths of road traffic accidents, the male and female genders clearly showed that the total number of men and women killed during the year was 1,27,435 and 23,332 respectively.

Table 2: Accident on person's age groups

Age Group	Male	Female
Less than 18	8347	2275
18-25	27417	4358
25-35	32609	5467
35-45	28564	4994
45-60	18592	3582
60 & above	6964	1850
Age not known	4960	806
total	127453	23332

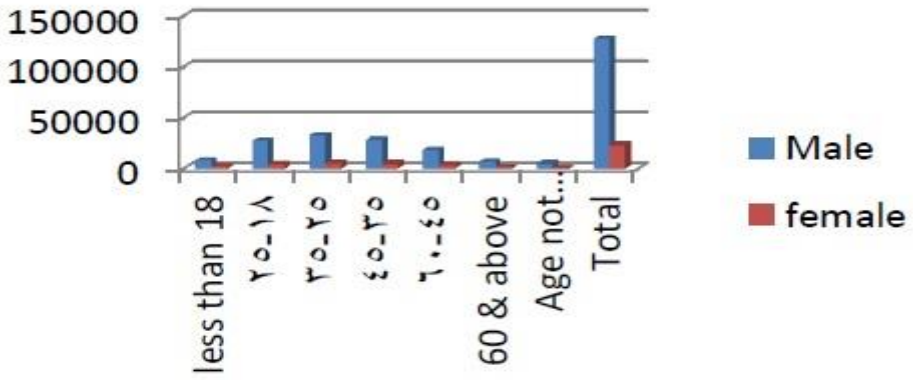

Figure 3: Chart for the accident on person's age groups

The trend in accidents revel that young people in the age of 20 to 35 males disobey the traffic rules and hence prone to huge number of causalities in the road accidents.

Conclusion

In this paper, a road accident prediction model has been developed and implemented, taking into consideration different possible causative factors. The range of factors chosen for the study are limited to mainly the condition of the road, weather influences and the nature of accident cause. The emotional state of mind and experiential influence of the driver have not been considered as in past literature. Figures 5, 6, 7 and 8 are indicative of the various parameters that have been used in the study and creation of the prediction model. Figure 5 shows a comparative analysis of the number of accidents reported for each type of accident, such as head-on collision, over-speed, skidding and so on. Figure 6 shows the weather type observed for the reported accidents whereas figure 7 indicates the activity initiated in response after each reported accident. Figure 8 shows a graphical representation of the increasing number of accident cases reported due to heavy duty vehicles. All these data have been included in the dataset used for this study. This model has been used in the creation of an application that can be used to predict the probability of risk of accident over an area inputted by the user. The user interface of the model based application outputs a graphical visualization of the factors that have been responsible for causing accidents relative to a specified area in the past. Based on this, a categorical prediction as high or low risk relative to accident occurrences is made for an area chosen by the user. The overall model has helped to give an understanding of the combinations of factors that have proven fatal in accident scenarios. A provision to further improve the dataset for future use has also been made in the form of an option to enter details of new accident cases.

From the statistical results, it can be seen that the rural mortality rate is higher, while the city is lower. Statistical analysis also includes other limiting factors such as the age of the vehicle, the type of vehicle, the age group of the person, and the category of road users. The predicted data results are displayed in a graphical representation. Graphical representations help the public understand accident metrics that help reduce mortality.

References

Srivastava AN, Zane-Ulman B. (2005). Discovering recurring anomalies in text reports regarding complex space systems. In Aerospace Conference, IEEE. IEEE 3853-3862.

Ghazizadeh M, McDonald AD, Lee JD. (2014). Text mining to decipher free-response consumer complaints: Insights from the nhtsa vehicle owner's complaint database. Human Factors 56(6): 1189- 1203. http://dx.doi.org/10.1504/IJFCM.2017.089439.

Chen ZY, Chen CC. (2015). Identifying the stances of topic persons using a model-based expectationmaximization method. J. Inf. Sci. Eng 31(2): 573-595. http://dx.doi.org/10.1504/IJASM.2015.068609.

Williams T, Betak J, Findley B. (2016). Text mining analysis of railroad accident investigation reports. In 2016 Joint Rail Conference. American Society of Mechanical Engineers V001T06A009- V001T06A009. http://dx.doi.org/10.14299/ijser.2013.01.

Suganya, E. and S. Vijayarani. "Analysis of road accidents in India using data mining classification algorithms." 2017 International Conference on Inventive Computing and Informatics (ICICI) (2017): 1122-1126.

Sarkar S, Pateshwari V, Maiti J. (2017). Predictive model for incident occurrences in steel plant in India. In ICCCNT 2017, IEEE, pp. 1-5. http://dx.doi.org/10.14299/ijser.2013.01.

Stewart M, Liu W, Cardell-Oliver R, Griffin M. (2017). An interactive web-based toolset for knowledge discovery

from short text log data. In International Conference on Advanced Data Mining and Applications. Springer, pp. 853-858. http://dx.doi.org/10.1007/978-3- 319-69179-4_61.

Zheng CT, Liu C, Wong HS. (2018). Corpus based topic diffusion for short text clustering. Neurocomputing 275: 2444-2458. http://dx.doi.org/10.1504/IJIT.2018.090859.

ArunPrasath, N and Muthusamy Punithavalli. "A review on road accident detection using data mining techniques." International Journal of Advanced Research in Computer Science 9 (2018): 881-885

George Yannis, Anastasios Dragomanovits, Alexandra Laiou, Thomas Richter, Stephan Ruhl, Francesca La Torre, Lorenzo Domenichini, Daniel Graham, NioviKarathodorou, Haojie Li (2016). "Use of accident prediction models in road safety management – an international inquiry". Transportation Research Procedia 14, pp. 4257 – 4266.

Anand, J. V. "A Methodology of Atmospheric Deterioration Forecasting and Evaluation through Data Mining and Business Intelligence." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 2, no. 02 (2020): 79-87.

Prayag Tiwari, Sachin Kumar, Denis Kalitin (2017). "Road-User Specific Analysis of Traffic Accident Using Data Mining Techniques". International Conference on Computational Intelligence, Communications, and Business Analytics. 10.1007/978-981-10-6430-2_31.

Kaur, G. and Er. Harpreet Kaur. "Prediction of the cause of accident and accident prone location on roads using data mining techniques." 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (2017): 1-7.

Irina Makarova, Ksenia Shubenkova, Eduard Mukhametdinov, and Anton Pashkevich, "Modeling as a Method to Improve Road Safety During Mass Events", Transportation Research Procedia 20 (2017) 43.