11-20-2018

Crash Severity Analysis of Rear-End Crashes in California Using Statistical and Machine Learning Classification Methods

Alidad Ahmadi

Arash Jahangiri

Vincent Berardi

Sahar Ghanipoor Machiani

Crash Severity Analysis of Rear-End Crashes in California Using Statistical and Machine Learning Classification Methods

CRASH SEVERITY ANALYSIS OF REAR-END CRASHES IN CALIFORNIA USING STATISTICAL AND MACHINE LEARNING CLASSIFICATION METHODS

Alidad Ahmadi
Graduate Assistant
Department of Civil, Construction, and Environmental Engineering
San Diego State University
San Diego, CA 92182, United States
Tel: 760 798-5453; Email: aahmadi@sdsu.edu

Arash Jahangiri
Assistant Professor
Department of Civil, Construction, and Environmental Engineering
San Diego State University
San Diego, CA 92182, United States
Tel: 540 200-7561; Email: ajahangiri@mail.sdsu.edu

Vincent Berardi
Assistant Professor
Department of Psychology,
Chapman University
Orange, CA 92866, United States
Tel: 908 591-2948; Email: berardi@chapman.edu

Sahar Ghanipoor Machiani (corresponding author)
Assistant Professor
Department of Civil, Construction, and Environmental Engineering
San Diego State University
San Diego, CA 92182, United States
Tel: 619 594-1937; Email: sghanipoor@mail.sdsu.edu

ABSTRACT

Investigating drivers' injury level and detecting contributing factors that aggravate the damage level imposed on drivers and vehicles is a critical subject in the field of crash analysis. In this study, a comprehensive vehicle-by-vehicle crash dataset is developed by integrating five years of data from California crash, vehicles involved, and road databases. The dataset is used to model the severity of rear-end crashes for comparing three analytic techniques: multinomial logit (MNL), mixed multinomial logit (MMNL), and support vector machine (SVM). The results of the crash severity models and the role of contributing factors to the severity outcome of rear-end crashes are extensively discussed. In terms of prediction performance, all three models yielded comparable results; although, the SVM performed slightly better than the other two methods. The results from this study will inform aspects of our driver safety education and design, either vehicle or roadway design, required to be improved to alleviate the probability of severe injuries.

Keywords: traffic safety, crash severity classification, machine learning, mixed multinomial logit, support vector machine

1. BACKGROUND

According to crash statistics report presented annually by the Fatality Analysis Reporting System (FARS), in 2015, nearly 32 thousand people were killed in vehicle crashes throughout the United States ("Fatality Analysis Reporting System (FARS) Encyclopedia, National Highway Traffic Safety Administration (NHTSA)," n.d.). According to this report, since 1995, California, Texas, and Florida are among the states with the largest number of fatalities. The number of fatalities has decreased about 30 percent from the year the report was first initiated, from about 47500 fatalities in 1994 to about 32000 fatalities in 2015; however, the current number is still extremely high, which reflects the need to find remedies to decrease the rate more quickly. In this regard, the crash analysis in the context of traffic safety has become one of the main areas of focus among the traffic engineers.

Crash databases are usually built by using police reports and are comprised of information such as the status of the crash, driver's information, road segment detail, environmental factors, and traffic condition. Understanding crash models and identifying contributing factors and their significance are crucial as the outcomes can be used in higher-organizational and management-level actions to define countermeasures that could prevent future crashes, improve the standards for the roadway and network design, improve public health policies, provide better emergency services, alleviate driver's injury severity, and nurture safer driving experience.

Crashes are naturally randomly-occurring incidents. Statistical models aid in better understanding the variability of these random events by examining the factors associated with them. There is a large body of research in the context of crash analysis. A review paper written by Lord and Mannering discusses the most common methodologies that have been used in studying crash frequency as well as the issues associated with them (Lord & Mannering, 2010). Savolainen

et al. (Savolainen, Mannering, Lord, & Quddus, 2011) published a similar paper discussing the most common methodologies for studying crash-severity analysis. A comprehensive study was carried out by Mannering and Bhat (Mannering & Bhat, 2014) in which an updated list of the most recent methodologies since the two previous studies [(Lord & Mannering, 2010) and (Savolainen et al., 2011)] was presented. Their analysis also focused on demonstrating how the crash analysis approach has evolved over time and how issues identified in previous studies have been addressed in more recently-introduced, advanced models (Mannering & Bhat, 2014). Most recently, Mannering et al. conducted a detailed discussion of how different statistical techniques can address the unobserved heterogeneity focusing on both injury-severity analysis and analysis of accident likelihood (Mannering, Shankar, & Bhat, 2016). In the remainder of this section, the most recent and relative studies concerning the modeling of injury-severity in crash databases are discussed.

Several studies that have developed injury-severity analysis models have focused on how the model's classification performance would change in absence or presence of various factors. For example, Li et al. (Z. Li, Wang, Liu, Bigham, & Ragland, 2013) have concentrated on demographic attributes which are believed to vary from one city to other cities. They showed the impact of implementing spatial attributes in development of crash prediction models in the framework of a Geographically Weighted Poisson Model (GWPM). Investigating the crash data and socio-demographic attributes of 58 counties in California, they compared the GWPM model's performance to the popular Generalized Linear Models (GLM) in predicting fatal crashes (Z. Li et al., 2013). In another study, Wu et al. (Q. Wu et al., 2014) used mixed logit models to analyze the injury-severity in single-vehicle crashes and crashes that involve two or more than two vehicles (Q. Wu et al., 2014). In a similar approach, focusing on truck-involved crashes, Chen et al. (C. Chen, Zhang, Tian, Bogus, & Yang, 2015) employed a modified Hierarchical Bayesian Random

Intercept model to predict the truck driver's injury severities based on a two year database of truck-involved crashes on rural roadways from New Mexico (C. Chen et al., 2015). Ye and Lord (Ye & Lord, 2014) investigated how the sample size can influence the performance of the Multinomial Logit, Ordered Probit, and Mixed Logit. Their results indicated that the sample size has a significant influence on the model performance. For example, a Mixed Logit model requires a much larger sample size compared to an Ordered Probit model, which can provide reasonable results even with small sample sizes (Ye & Lord, 2014).

The most basic crash severity modeling approaches, such as binary Logit and Probit models, have evolved into more advanced parametric and non-parametric models. These advanced models can address more of the unobserved characteristics of the data, which were not examined in earlier models. One such class of analytic approaches are mixed models (Milton, Shankar, & Mannering, 2008; Q. Wu et al., 2014; Yasmin & Eluru, 2013; Ye & Lord, 2014). For instance, Milton et al. (Milton et al., 2008) used the mixed logit model that is able to differentially account for the effect of independent variables on the level of severity over different road segments (Milton et al., 2008). Previous models had considered this effect to be the same on all road segments which results in biased results. The mixed feature of the model allows the coefficient of each variable affecting the injury-severity (explanatory variables) to vary across all individuals in the crash database in order to consider the heterogeneous effect and correlations of unobserved factors (Savolainen et al., 2011; K. E. Train, 2009). In another study, when studying single-vehicle crashes in the state of California from 2003 to 2004, Kim et al. (Kim, Ulfarsson, Kim, & Shankar, 2013) employed mixed logit models in injury-severity analysis. Their results indicated the importance of considering population heterogeneity according to the notable differences in results for different age groups, especially when comparing the older age group to younger drivers (Kim et al., 2013).

Developing separate models for two different age groups of older and younger drivers and for two gender groups on three different road surfaces, Morgan and Mannering (Morgan & Mannering, 2011) utilized the mixed logit analysis to assess the effect of those factors on the severity of crashes in single-vehicle crashes. In a more recent study, Behnood et al. utilized mixed logit modeling and compared it to latent-class models using the eight-year pedestrian-injury database in Chicago (Behnood & Mannering, 2016a). A similar modeling approach has shown promising application in other studies as well (for examples see (Anastasopoulos & Mannering, 2011; Aziz, Ukkusuri, & Hasan, 2013; Behnood & Mannering, 2015, 2016b; Cerwick, Gkritza, Shaheed, & Hans, 2014; Malyshkina, Mannering, & Tarko, 2009; Manner & Wünsch-Ziegler, 2013; Moore, Schneider IV, Savolainen, & Farzaneh, 2011; Ye & Lord, 2011)).

More recently, non-parametric modeling methods such as SVM, Artificial Neural Networks, and Decision Table/Naïve Bayes (C. Chen, Zhang, Yang, Milton, & Alcántara, 2016) have become popular in the crash analysis studies. Non-parametric models are not built on the assumptions made based on the distribution properties of the data, which is usually the base for parametric models. An advantage of non-parametric models over parametric models is that they do not require the pre-defined relationship between the dependent and explanatory variables (Z. Li, Liu, Wang, & Xu, 2012).

SVM, a supervised machine learning technique, is one of the methods of classification/regression used in many different transportation and traffic safety-related areas (for examples see (Balali & Golparvar-Fard, 2016; Jahangiri & Rakha, 2015; Jahangiri, Rakha, & Dingus, 2016; X. Li, Lord, Zhang, & Xie, 2008; C.-H. Wu, Ho, & Lee, 2004)). Chen et al. (S. Chen, Wang, & van Zuylen, 2009) demonstrated the application of SVM as an incident detection tool to identify traffic incidents that reduce the capacity of the road (S. Chen et al., 2009). In the

context of active traffic management, Yu and Abdel-Aty (Yu & Abdel-Aty, 2013) showed the application of SVM in real-time risk analysis to predict crash occurrences (Yu & Abdel-Aty, 2013). Studying 326 freeway sections around the state of Florida, Li et al. (Z. Li et al., 2012) employed SVM and Ordered Probit (OP) to predict the injury severity of individual crashes. Using the Radial Basic Function (RBF), they indicated that the SVM model performed better than the OP in terms of the percent of correct predictions. The result was achieved by comparing the two models with multi-class response (five injury-severity levels). It has also been demonstrated that classification results for a two-level SVM also resulted in a significant improvement in the prediction accuracy of SVM model (Z. Li et al., 2012).

Yu et al. (Yu & Abdel-Aty, 2013) also developed fixed parameter logit, SVM (with RBF), and random parameter logit for four-year data collected from a mountainous freeway section in Colorado. Comparing three models, they indicated that SVM and random parameter models provided a better fit than the fixed parameter logit models. In a more recent study, based on a two-year crash data from New Mexico, Chen et al. (C. Chen, Zhang, Qian, Tarefder, & Tian, 2016) studied the application of SVM in mapping the injury severity in rollover crashes. Their result indicated that the SVM model provided reasonable performance in terms of predicting the injury severity (C. Chen, Zhang, Qian, et al., 2016). Unlike the study conducted by Li et al. (Z. Li et al., 2012), where every single crash were taken as a single research unit, Chen et al. has taken each individual driver/vehicle as the research unit and taken into analysis a variable with the number of vehicles involved in the crash as an independent variable to the model (C. Chen, Zhang, Qian, et al., 2016).

In this paper, crash databases consisting of information on the crash, environment, vehicles, and occupants for five consecutive years (2007-2011) in the State of California were integrated.

To the best of our knowledge, this vehicle by vehicle integrated database has not been used to develop crash models. Three approaches for modeling the severity of rear-end crashes, -support vector machine (SVM), multinomial logit (MNL), and mixed multinomial logit (MMNL)-were applied to this database and compared. The remainder of this paper is structured as follows. The next section describes the methodology including data description and methods used to analyze the data. Then, results of the classification methods are presented and discussed for all three methods. The last section provides conclusions and future directions.

2. METHODOLOGY

2.1. DATA DESCRIPTION

The data used in this study was obtained from the Highway Safety Information System (HSIS) in the state of California for five consecutive years from 2007 to 2011. The database consists of three tables including the crash database, vehicle database, and road database. The rows in the crash database are built based on each case of a crash and includes information such as weather condition and lighting that is common among all vehicles involved in the crash. The vehicle database includes information specific to each vehicle such as driver's age, sex, and vehicle type. Finally, the road database is built based on information of each road segment such as a number of lanes and terrain level. In order to perform a vehicle by vehicle analysis for the purpose of this study, it was required to attach the information of each crash to every single vehicle involved in that specific crash, and then add the road information to each vehicle. This task was performed using a Matlab script.

The observations for the rear-end crashes were extracted from the database. The severity-injury levels in the dataset include 5 levels of (1) Property damage only, (2) Complaint of pain, (3)

Other visible injury, (4) Severe injury, and (5) Fatal. To select a subset of the independent variables for developing the models, all variables were individually examined to determine how well they classified severity levels using the Area Under the Curve (AUC) measure to rank the variables. Top variables that provided the best results in predicting the severity level were used in the full modeling procedure. The selected variables were age, sex, terrain level, weather condition, lighting, vehicle type, number of lanes, and crash cause. The models were developed on 9468 individual vehicles with 70 percent of the data randomly selected to build the training model and the remaining 30 percent left as the test data to compare the performance of three models in terms of their prediction performance.

2.2. MULTINOMIAL LOGIT (MNL) AND MIXED MULTINOMIAL LOGIT (MMNL) REGRESSION

Mixed Multinomial Logit regression, also known as random parameters logit model, is a generalized form of the Multinomial Logistic regression in which the coefficients of any of the variables are allowed to vary across the individuals and not be limited to a fixed value. Consequently, it allows the model to take into account the heterogeneity of the population. For the standard logistic regression (multinomial logit), the probability of individual *i* experiencing the severity level of *l* from the set of severity outcomes *J* is (Croissant, 2012):

$$P_{il} = \frac{e^{\beta' x_{il}}}{\sum_{j=1}^{J} e^{\beta' x_{ij}}}$$

Here, x is the factor and β' is the fixed coefficient for all individuals. In the mixed multinomial logit, each individual has their own coefficient β_i', and probabilities are described as probability of the individual *i*, conditional on the vector of individual-specific coefficient β_i, experiencing severity level *l* as:

$$P_{il} \mid \beta_i = \frac{e^{\beta_i' x_{il}}}{\sum_{j=1}^{J} e^{\beta_i' x_{ij}}}$$

Since there are many observations, and finding an individual coefficient for each observation may not be of interest, the coefficients are considered to be random variables and the probabilities of each individual *i* is found conditional on the vector of random coefficients β. Later, the average of the probabilities for all values of β is found to obtain the unconditional probability. Given that β_i has the density of $f(\beta, \theta)$ (θ as distribution parameters of β), for one individual coefficient, the unconditional probability for individual *i* experiencing injury level of *l* is:

$$P_{il} = \int_{\beta} (P_{il}|\beta_i)\, f(\beta|\theta) d\beta$$

Here, the function f indicates the density function for β with θ defining the parameters of the density function. Solving this integral becomes more complicated when there is more than one parameter, which requires defining a separate β for each of the random variables. This necessitates simulation techniques. In this study, 'mlogit' package was utilized through the R software to perform the mixed logit analysis. Using 200 draws was found to be sufficient for results to converge. For more detail on how the simulation process is performed by the package and how to define simulation parameters, we refer the readers to (Croissant, 2012; K. Train & Croissant, 2012; K. E. Train, 2009).

2.3. BINARY LOGIT

Binary Logit is the simple form of the multinomial logit (detailed above) with two instead of multiple outcomes. In this study, we build a binary model with the set of same variables for each of the severity levels. In each model, the first level is whether a specific severity occurs and the second level is whether that specific severity does not. For example, regarding the property damage

only level, the first outcome is when the severity was property damage only, and the second outcome is when the severity is not property damage only. In other words, the second outcome is the combination of the 4 remaining levels in the dataset (e.g. all possible outcomes except property damage only). The procedure is repeated for each level which makes a total of 5 binary models. It is expected that the binary logit does not yield results as efficient as multinomial logit; however, comparing results for all 5 models, we can investigate how disparate each of severity levels is from the rest of population.

2.4. SUPPORT VECTOR MACHINES (SVM)

In addition to multinomial logit and mixed multinomial logit, SVM, a supervised machine learning algorithm, was employed to classify severity levels. SVM is known as a powerful method of classification (and also regression) problems as it tries to find the best possible decision boundaries between different classes. In model development, SVM applies the function $\phi(.)$ to transform the data from *X* space into some *Z* space. This transformation rearranges the data in such a way that the classification becomes an easier task. SVM was first introduced in (Boser, Guyon, & Vapnik, 1992) for separable data and was further expanded in (Cortes & Vapnik, 1995) for non-separable data. The SVM objective function maximizes the margin between different classes and at the same time accepts some errors that can be regulated through a penalty parameter. The formulation of SVM is as follows.

$$\min_{w,b,\xi}\left(\frac{1}{2}w^T w + C\sum_{n=1}^{N}\xi_n\right) \qquad (1)$$

Subject to:

$$y_n(w^T\phi(x_n) + b) \geq 1 - \xi_n\,, n = 1, \ldots, N \quad (2)$$

$$\xi_n \geq 0\,, n = 1, \ldots, N \quad (3)$$

Where,

w	Parameters to define decision boundary between classes
C	Regularization (or penalty) parameter
ξ_n	Error parameter to denote margin violation
b	Intercept associated with decision boundaries
$\phi(x_n)$	Function to transform data from X space into some Z space
y_n	Target value for the n^{th} observation

To develop a multi-class classification model using SVM, one versus one approach based on a voting strategy has been employed; models were developed using only two classes (e.g. fatal vs severe injury, fatal vs other visible injury, and so forth). To predict a new observation, all these models are used to produce votes for different classes. The class with the highest vote is identified as the predicted class. Since there are five classes for crash severity, combinations of two classes from five classes results in ten different combinations, which led to the development of ten models.

3. RESULTS

3.1. MULTINOMIAL LOGIT AND MIXED MULTINOMIAL LOGIT REGRESSION

Table 1 and Table 2, respectively, represent the estimation results for the coefficient of each variable in the MNL and MMNL models as well as overall performance. Property damage only severity level is selected as the base of comparison in the analysis and all the other severity

levels are compared to it. Levels of each independent variable are shown below the name of each variable with the first level representing the base level.

Selecting Random Parameters

To find the variables in which the coefficients vary across the individuals (random parameters in MMNL), all independent variables were checked to determine how strong the hypothesis of coefficients varying across individuals is. In the first run, all variables were put in the model assuming that the distribution of their coefficients was normal. The existence of variability among the coefficient (having the random effect) was signaled by a statistically significant variable that also had a standard deviation that is statistically different from zero (p-value lower than 0.05 in this study). For instance, according to Table 2, the coefficient associated to "4 lanes" level of the variable "Number of lanes" was found to be a significant factor for severe injury crashes and have a high standard deviation that is statistically different from zero (p-value = 0.00). Examining the normal distribution of coefficient associated with this variable, it was found that almost 74 percent of the sample has a negative coefficient and the rest are positive. This, in fact, indicates that this factor decreases the probability of crash resulting in a severe injury for 74 percent of the population and increases the probability for the remaining 26 percent of the population. In a similar manner, significant variables with random effect were selected to remain in the model as variables with random effects and the rest of the variables that did not show the mixed effect w considered to have fixed coefficients.

Variables Significance

Only a few factors including, driving along a dark road with street light, driving on 4 lanes and 6 lanes roads, and speeding show the heterogeneity effect. Results from MNL and MMNL are very similar with many variables showing significance in both MNL and MMNL. However, in order to

discuss the random effects as well, we used results found from MMNL (Table 2) as the basis for variable significance analysis.

Several variables were found to have a significant impact on different levels of severity. The results for different age groups show that aging has an increasing effect on the probability of getting killed in a crash. This, in fact, is consistent with the results from (Yasmin, Eluru, Pinjari, & Tay, 2014) in which they maintain the increase in the risk of fatal crashes for drivers of age 65 or above (compared to other age groups) for different types of crashes. Similar results were also achieved by (Q. Wu et al., 2014) in a study on the contribution of a variety of factors on different levels of injury in multi-vehicle crashes, where the probability of fatality was increased for drivers of 65 years or older. Some other studies by (Kim et al., 2013; Xie, Zhang, & Liang, 2009; Yasmin & Eluru, 2013) have also confirmed the same finding.

Regarding the drivers' gender, it was found that compared to female drivers, male drivers are less likely to complain about pain after a crash. An explanation might be that men are physically stronger. There is also evidence that men are more likely to be killed compared to female drivers which raise the idea that male drivers might be more aggressive drivers than female drivers. This is in line with the finding by (Shankar, Mannering, & Barfield, 1996) where they maintain greater probability of fatal and disabling injury for crashes involved male drivers. Different results, however, have been concluded from the gender analysis. For instance, (Q. Wu et al., 2014) found that in multi-vehicle crashes, the coefficient varies across genders as for some female drivers it increases the probability of a fatal crash and for some, it reduces the probabilities. On the other hand, in some studies such as (Abdel-Aty, 2003; Xie et al., 2009), it was found that female drivers are more likely to get killed in a crash with the same circumstances.

Investigating driving on different terrain levels, the outcome of a crash is more likely to result in a severe injury or fatality when driving on mountainous and rolling terrains compared to flat terrains. This was expected since driving on mountainous or rolling terrains is usually more complex than driving on the flat terrains. This also confirms the results from (Shankar et al., 1996), where they indicated that high proportion of horizontal curves was found to increase the likelihood of a possible injury crash. As for driving in different weather condition, it was found that driving in foggy condition increases the probability of crashes leading to fatality or severe injuries compared to clear weather condition.

Results from the lighting condition of the street indicated that driving in a low light condition such as dusk or dark without street lights increases the probability of a fatal crash compared to driving in daylight. This is in line with the results from (Kim et al., 2013; Q. Wu et al., 2014; Xie et al., 2009), where they also found that undesirable lighting increases the propensity of a crash to be more severe or fatal. In this study; however, driving at night when there is street light found to have mixed effect. Evaluation of the distribution of the coefficient shows the existence of a statistically high standard deviation of 1.83 (p-value = 0.00) and a mean of 0.55 for the coefficients of the "Dark - street light" level of variable "Light". This might be associated to the fact that some drivers tend to drive more carefully when their range of view is limited due to the lack of proper lighting while some drivers may not feel the need to compensate in this situation.

Comparing different vehicle types, it was found that motorcycle riders and truck drivers are more prone to higher levels of severity and fatality and less likely to experience property damage and surface injuries compared to passenger cars.

Increase in the number of lanes for most of the levels found to have a negative effect on the log odds of the crash to become fatal or severe. This, in fact, raises the idea that crashes are

more likely to have higher levels of severity on two-lane roads than roads and highways with a higher number of lanes.

Regarding crash causes, behavior such as following too closely and speeding are less likely to lead to fatality or severe injury compared to the situation where the driver is under influence of the alcohol. A similar result was achieved in other studies such as (Xie et al., 2009; Yasmin & Eluru, 2013).

3.2. BINARY LOGIT

Table 3 demonstrates the summary of the results for the binary logit model. Results from both Table 1 and Table 3 indicate the same effects from selected factors in different ways. For example, looking at the motorcycle data, from Table 1 (multinomial approach), we see that motorcycle drivers are more prone to being killed in accidents compared to passenger car drivers. The same result could be ascertained by comparing the estimated coefficients/significances retrieved from each of the five binary logits. Starting from the first binary model (property damage only vs. all levels) there is a low negative coefficient that increases and eventually becomes positive as you move toward the fifth binary model (fatal vs. all levels). The same concept applies to other variables as well.

3.3. SUPPORT VECTOR MACHINE (SVM)

The same train and test datasets that were used for the MNL and MMNL models were applied to develop and validate the SVM model using Gaussian kernels. All attributes (variables) were scaled as SVM usually performs poorly without feature scaling. This is a simple preprocessing step that could significantly impact the SVM model performance. In order to find

the optimal SVM model, two parameters, the regularization parameter, and the Gamma parameter, must be tuned. This parameter tuning was conducted for all ten models, one of which is shown in Fig. 1. Focusing on the fatal vs property damage only model, Fig. 1 illustrates how different parameter settings impact the area under curve (AUC) value. This figure shows a grid search of the Gamma and Cost parameters to achieve the best possible performance of one of the SVM models. The best AUC was achieved with cost and Gamam parameters being equal to 0.32 and 0.1, respectively. The darker regions in the figure represent higher values of AUC, which correspond to better performance. The optimal performance of the other SVM models was achieved in the same fashion.

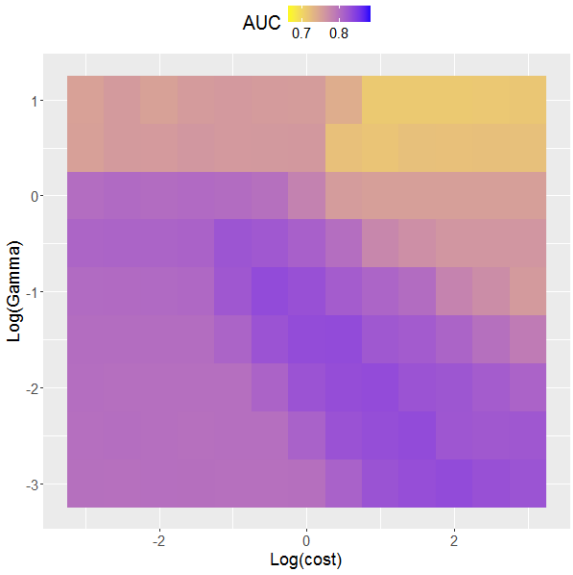

Figure 1 SVM parameter tuning for the fatal vs property damage only.

4. DISCUSSION OF PREDICTION RESULTS

In this section, the SVM, MNL (both binary and multi-class), and MMNL are compared in terms of their predicting accuracy on the test data. Table 4 demonstrates the number of true and false predictions in each model as well as for each class.

While both binary and multinomial logit models indicate similar information regarding the effect of factors on the outcome of the accident, they do not maintain the same accuracy when used for prediction. Based on Table 4, the binary models built for the first three levels (property damage only, complaint of pain, and other visible injury) did not maintain good prediction results based on the test data, where every single prediction for the desired level (e.g. property damage only in the first model) is incorrect. Although binary models built for last two levels (severity and fatal) maintained better prediction results, the accuracy was still not very good. While the prediction results from binary models display estimation inefficiency, it also demonstrates how discrete the last two levels (severity and fatal) are from the other levels.

Regression and support vector approaches with all five levels in the model led to more accurate predictions. The prediction results are also included in Table 4 for comparison. Both MNL and MMNL models maintained a comparable accuracy, which is expected since most parameters were not found to have random effects. Yet it is important that random parameters are included to better report the effect of these variables. It should be noted that most machine learning algorithms such as SVM are considered as black-box methods that are difficult to interpret. In general, they could perform better than classical statistical methods such as logit models but have limited explanatory power. In the present study, the SVM performed slightly better in specific cases as explained below.

According to Table 4, the SVM has maintained greater accuracy for “other visible injury” and “fatal” class of the database. On the other hand, MNL and MMNL had greater accuracy for “property damage only”, “complaint of pain”, and “severe injury”. Although, in total, the number of true predictions by the SVM is slightly higher than the total for MNL and MMNL.

More detailed results regarding the prediction of each model in the form of confusion matrix are shown in Table 5. According to Table 5, the largest number of true predictions belongs to "Complaint of pain" level. On the other hand, the lowest number of true predictions belong to "Other visible injury" level. This is a broad, and somewhat vague, a category which is overlapping with the "Complaint of pain" category, so it is expected that it would be hard to predict. The remaining "Property damage only", "Severe injury", and "Fatal" levels have similar accuracy levels. The last row under each confusion matrix indicates the percent of true prediction for each individual level of severity.

5. CONCLUSION

Considering the significant losses resulting from crashes, predicting the likelihood of crash severity has been a vital line of research. Studying crashes and their contributing factors allow traffic engineers and practitioners to determine and prioritize dangerous crash prone situations, which in turn aids efforts to implement suitable countermeasures and enhance the level of safety. In this study, crash severity models were developed using three statistical approaches of MNL (binary and multi-class), MMNL, and SVM. The database used for the model training is an integrated database of the California crash database from 2007 to 2011 including information on driver, vehicle, and environment.

Studying effects of different factors on severity of accidents, results from both MNL and MMNL models showed that older and male drivers are more likely to have a fatal crash while complain of pain is observed more among female drivers compared to males. As expected, driving in foggy and mountainous or rolling terrains is associated with more fatal and severe injury crashes.

Similarly, low light conditions are associated to more fatal crashes. Regarding the vehicle type, motorcycle riders and truck drivers are more likely to experience a higher level of crash severity. Also, a fewer number of lanes was found to be associated with higher level of severity. Regarding the cause of crashes, improper turning and driving under influence were found to cause severe injuries and fatal crashes. In terms of random parameters, factors including driving on dark roads with a street light, driving on 4 and 6 lane roads, and speeding could have various impact on the severity outcome of accidents as for some people they have been a contributing factor and for the other, they have acted as the opposite.

Studying the performance of the models based on their prediction accuracy, the MNL and MMNL maintained comparable results. This result was expected since only a few of the variables were found to have a mixed effect, and those that found to have mixed effect were not dominant enough to make a significant change in the overall outcome of the model. The SVM approach maintained a slightly better prediction accuracy using the test data.

In the future, the crash analysis will be conducted for other crash types than rear-end crashes. Also, other techniques such as neural network and discriminant analysis method will be considered on the same dataset and results will be compared in terms of model performance.

ACKNOWLEDGEMENT

The authors would like to thank the Highway Safety Information System (HSIS) for their support and provision of the crash dataset. The authors wish to also thank the University Grants Program (UGP) at San Diego State University for partially sponsoring this study.

REFERENCES

Abdel-Aty, M. (2003). Analysis of driver injury severity levels at multiple locations using ordered probit models. *Journal of Safety Research*, *34*(5), 597–603. https://doi.org/10.1016/j.jsr.2003.05.009

Anastasopoulos, P. C., & Mannering, F. L. (2011). An empirical assessment of fixed and random parameter logit models using crash- and non-crash-specific injury data. *Accident Analysis & Prevention*, *43*(3), 1140–1147. https://doi.org/10.1016/j.aap.2010.12.024

Aziz, H. M. A., Ukkusuri, S. V., & Hasan, S. (2013). Exploring the determinants of pedestrian–vehicle crash severity in New York City. *Accident Analysis & Prevention*, *50*, 1298–1309. https://doi.org/10.1016/j.aap.2012.09.034

Balali, V., & Golparvar-Fard, M. (2016). Evaluation of Multiclass Traffic Sign Detection and Classification Methods for U.S. Roadway Asset Inventory Management. *Journal of Computing in Civil Engineering*, *30*(2), 04015022. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000491

Behnood, A., & Mannering, F. L. (2015). The temporal stability of factors affecting driver-injury severities in single-vehicle crashes: Some empirical evidence. *Analytic Methods in Accident Research*, *8*, 7–32. https://doi.org/10.1016/j.amar.2015.08.001

Behnood, A., & Mannering, F. L. (2016a). An empirical assessment of the effects of economic recessions on pedestrian-injury crashes using mixed and latent-class models. *Analytic Methods in Accident Research*, *12*, 1–17. https://doi.org/10.1016/j.amar.2016.07.002

Behnood, A., & Mannering, F. L. (2016b). The effects of drug and alcohol consumption on driver injury severities in single-vehicle crashes. *Traffic Injury Prevention*, *18*(5), 456–462. https://doi.org/10.1080/15389588.2016.1262540

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (pp. 144–152). New York, NY, USA: ACM. https://doi.org/10.1145/130385.130401

Cerwick, D. M., Gkritza, K., Shaheed, M. S., & Hans, Z. (2014). A comparison of the mixed logit and latent class methods for crash severity analysis. *Analytic Methods in Accident Research*, *3–4*, 11–27. https://doi.org/10.1016/j.amar.2014.09.002

Chen, C., Zhang, G., Qian, Z., Tarefder, R. A., & Tian, Z. (2016). Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention*, *90*, 128–139. https://doi.org/10.1016/j.aap.2016.02.011

Chen, C., Zhang, G., Tian, Z., Bogus, S. M., & Yang, Y. (2015). Hierarchical Bayesian random intercept model-based cross-level interaction decomposition for truck driver injury severity investigations. *Accident Analysis & Prevention*, *85*, 186–198. https://doi.org/10.1016/j.aap.2015.09.005

Chen, C., Zhang, G., Yang, J., Milton, J. C., & Alcántara, A. "Dely." (2016). An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier. *Accident Analysis & Prevention*, *90*, 95–107. https://doi.org/10.1016/j.aap.2016.02.002

Chen, S., Wang, W., & van Zuylen, H. (2009). Construct support vector machine ensemble to detect traffic incident. *Expert Systems with Applications*, *36*(8), 10976–10986. https://doi.org/10.1016/j.eswa.2009.02.039

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1023/A:1022627411411

Croissant, Y. (2012). Estimation of multinomial logit models in R: The mlogit Packages. *R Package Version 0.2-2. URL: Http://Cran. R-Project. Org/Web/Packages/Mlogit/Vignettes/Mlogit. Pdf*. Retrieved from ftp://193.1.193.75/disk1/cran.r-project.org/web/packages/mlogit/vignettes/mlogit.pdf

Fatality Analysis Reporting System (FARS) Encyclopedia, National Highway Traffic Safety Administration (NHTSA). (n.d.). Retrieved May 1, 2017, from http://www-fars.nhtsa.dot.gov/Main/index.aspx

Jahangiri, A., & Rakha, H. A. (2015). Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. *IEEE Transactions on Intelligent Transportation Systems*, *16*(5), 2406–2417.

Jahangiri, A., Rakha, H., & Dingus, T. A. (2016). Red-light running violation prediction using observational and simulator data. *Accident Analysis & Prevention*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0001457516302056

Kim, J.-K., Ulfarsson, G. F., Kim, S., & Shankar, V. N. (2013). Driver-injury severity in single-vehicle crashes in California: A mixed logit analysis of heterogeneity due to age and gender. *Accident Analysis & Prevention*, *50*, 1073–1081. https://doi.org/10.1016/j.aap.2012.08.011

Li, X., Lord, D., Zhang, Y., & Xie, Y. (2008). Predicting motor vehicle crashes using Support Vector Machine models. *Accident Analysis & Prevention*, *40*(4), 1611–1618. https://doi.org/10.1016/j.aap.2008.04.010

Li, Z., Liu, P., Wang, W., & Xu, C. (2012). Using support vector machine models for crash injury severity analysis. *Accident Analysis & Prevention*, *45*, 478–486. https://doi.org/10.1016/j.aap.2011.08.016

Li, Z., Wang, W., Liu, P., Bigham, J. M., & Ragland, D. R. (2013). Using Geographically Weighted Poisson Regression for county-level crash modeling in California. *Safety Science*, *58*, 89–97. https://doi.org/10.1016/j.ssci.2013.04.005

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, *44*(5), 291–305. https://doi.org/10.1016/j.tra.2010.02.001

Malyshkina, N. V., Mannering, F. L., & Tarko, A. P. (2009). Markov switching negative binomial models: An application to vehicle accident frequencies. *Accident Analysis & Prevention*, *41*(2), 217–226. https://doi.org/10.1016/j.aap.2008.11.001

Manner, H., & Wünsch-Ziegler, L. (2013). Analyzing the severity of accidents on the German Autobahn. *Accident Analysis & Prevention*, *57*, 40–48. https://doi.org/10.1016/j.aap.2013.03.022

Mannering, F. L., & Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, *1*, 1–22. https://doi.org/10.1016/j.amar.2013.09.001

Mannering, F. L., Shankar, V., & Bhat, C. R. (2016). Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research*, *11*, 1–16. https://doi.org/10.1016/j.amar.2016.04.001

Milton, J. C., Shankar, V. N., & Mannering, F. L. (2008). Highway accident severities and the mixed logit model: An exploratory empirical analysis. *Accident Analysis & Prevention*, *40*(1), 260–266. https://doi.org/10.1016/j.aap.2007.06.006

Moore, D. N., Schneider IV, W. H., Savolainen, P. T., & Farzaneh, M. (2011). Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection

and non-intersection locations. *Accident Analysis & Prevention*, *43*(3), 621–630. https://doi.org/10.1016/j.aap.2010.09.015

Morgan, A., & Mannering, F. L. (2011). The effects of road-surface conditions, age, and gender on driver-injury severities. *Accident Analysis & Prevention*, *43*(5), 1852–1863. https://doi.org/10.1016/j.aap.2011.04.024

Savolainen, P. T., Mannering, F. L., Lord, D., & Quddus, M. A. (2011). The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis & Prevention*, *43*(5), 1666–1676. https://doi.org/10.1016/j.aap.2011.03.025

Shankar, V., Mannering, F., & Barfield, W. (1996). Statistical analysis of accident severity on rural freeways. *Accident Analysis & Prevention*, *28*(3), 391–401. https://doi.org/10.1016/0001-4575(96)00009-7

Train, K., & Croissant, Y. (2012). Kenneth Train's exercises using the mlogit package for R. *R*, *25*, 0–2.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press.

Wu, C.-H., Ho, J.-M., & Lee, D. T. (2004). Travel-time prediction with support vector regression. *IEEE Transactions on Intelligent Transportation Systems*, *5*(4), 276–281. https://doi.org/10.1109/TITS.2004.837813

Wu, Q., Chen, F., Zhang, G., Liu, X. C., Wang, H., & Bogus, S. M. (2014). Mixed logit model-based driver injury severity investigations in single- and multi-vehicle crashes on rural two-lane highways. *Accident Analysis & Prevention*, *72*, 105–115. https://doi.org/10.1016/j.aap.2014.06.014

Xie, Y., Zhang, Y., & Liang, F. (2009). Crash Injury Severity Analysis Using Bayesian Ordered Probit Models. *Journal of Transportation Engineering*, *135*(1), 18–25. https://doi.org/10.1061/(ASCE)0733-947X(2009)135:1(18)

Yasmin, S., & Eluru, N. (2013). Evaluating alternate discrete outcome frameworks for modeling crash injury severity. *Accident Analysis & Prevention*, *59*, 506–521. https://doi.org/10.1016/j.aap.2013.06.040

Yasmin, S., Eluru, N., Pinjari, A. R., & Tay, R. (2014). Examining driver injury severity in two vehicle crashes – A copula based approach. *Accident Analysis & Prevention*, *66*, 120–135. https://doi.org/10.1016/j.aap.2014.01.018

Ye, F., & Lord, D. (2011). Investigation of Effects of Underreporting Crash Data on Three Commonly Used Traffic Crash Severity Models. *Transportation Research Record: Journal of the Transportation Research Board*, *2241*, 51–58. https://doi.org/10.3141/2241-06

Ye, F., & Lord, D. (2014). Comparing three commonly used crash severity models on sample size requirements: multinomial logit, ordered probit and mixed logit models. *Analytic Methods in Accident Research*, *1*, 72–85.

Yu, R., & Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis & Prevention*, *51*, 252–259. https://doi.org/10.1016/j.aap.2012.11.027

1 Table 1 - Summary of results for multinomial logit model

	Complaint of pain				Other visible injuries				Severe injury				Fatal			
Variable	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)
:(intercept)	0.26	0.31	0.84	0.40	1.39	0.27	5.11	0.00***	1.94	0.26	7.40	0.00***	1.22	0.28	4.44	0.00***
Age																
:Young adult																
:Adult	0.06	0.10	0.58	0.56	0.04	0.10	0.45	0.65	0.02	0.10	0.16	0.88	0.13	0.12	1.09	0.27
:Middle aged	0.03	0.11	0.31	0.76	0.11	0.11	1.03	0.30	0.09	0.12	0.78	0.44	0.27	0.13	2.10	0.04*
:Old	0.03	0.18	0.18	0.86	0.16	0.18	0.86	0.39	0.20	0.19	1.03	0.30	0.63	0.20	3.09	0.00**
Sex																
:Female																
:Male	-0.30	0.08	-3.70	0.00***	-0.06	0.08	-0.66	0.51	0.10	0.09	1.14	0.26	0.25	0.10	2.47	0.01*
Terrain																
:Flat																
:Mountainous	0.06	0.21	0.26	0.79	0.35	0.20	1.76	0.08 ‘	0.42	0.20	2.08	0.04*	0.29	0.22	1.31	0.19
:Rolling	0.11	0.09	1.14	0.26	0.07	0.10	0.76	0.45	0.16	0.10	1.64	0.10 ‘	0.34	0.11	3.24	0.00**
Weather																
:Clear																
:Cloudy	0.11	0.11	1.08	0.28	-0.19	0.11	-1.67	0.09 ‘	-0.21	0.12	-1.76	0.08 ‘	-0.15	0.13	-1.16	0.25
:Raining	-0.02	0.24	-0.10	0.92	0.12	0.23	0.53	0.60	-0.27	0.26	-1.04	0.30	-0.36	0.28	-1.28	0.20
:Fog	0.89	0.84	1.06	0.29	0.62	0.87	0.71	0.48	2.33	0.75	3.10	0.00**	2.22	0.77	2.87	0.00**
Light																
:Daylight																
:Dusk - Dawn	-0.43	0.22	-1.93	0.05 ‘	-0.31	0.22	-1.38	0.17	0.14	0.22	0.66	0.51	0.87	0.22	3.93	0.00***
:Dark - Street Lights	-0.01	0.12	-0.08	0.93	-0.08	0.12	-0.65	0.52	0.57	0.12	4.81	0.00***	1.06	0.13	8.29	0.00***
:Dark - No Street Lights	-0.13	0.14	-0.92	0.36	0.29	0.13	2.17	0.03*	1.00	0.13	7.78	0.00***	1.78	0.13	13.56	0.00***
Vehicle type																
:Passenger Car																
:Truck	-0.17	0.24	-0.69	0.49	0.36	0.21	1.72	0.09 ‘	1.13	0.19	5.88	0.00***	1.62	0.19	8.40	0.00***
:Motorcycle	1.72	0.78	2.21	0.03*	3.47	0.72	4.80	0.00***	4.73	0.72	6.62	0.00***	4.74	0.72	6.60	0.00***
:Pickup Truck	-0.01	0.11	-0.05	0.96	0.02	0.11	0.17	0.87	0.08	0.12	0.71	0.48	-0.04	0.13	-0.30	0.76
:Bus	1.04	1.16	0.90	0.37	1.11	1.16	0.96	0.34	2.39	1.07	2.23	0.03*	2.06	1.14	1.80	0.07 ‘
:Emergency Vehicle	0.22	0.55	0.40	0.69	-0.96	0.83	-1.17	0.24	-0.44	0.73	-0.60	0.55	-0.51	0.81	-0.63	0.53
Number of lanes																
:2																
:3	0.46	0.71	0.64	0.52	0.38	0.72	0.53	0.59	-1.00	0.85	-1.18	0.24	0.60	0.70	0.86	0.39

	Complaint of pain				Other visible injuries				Severe injury				Fatal			
Variable	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)
:4	-0.09	0.20	-0.46	0.65	-0.05	0.20	-0.25	0.80	-0.51	0.20	-2.57	0.01*	-0.79	0.21	-3.82	0.00***
:5	0.33	0.35	0.92	0.36	0.28	0.37	0.76	0.45	-0.18	0.36	-0.49	0.63	-0.52	0.38	-1.39	0.16
:6	-0.55	0.20	-2.76	0.01**	-0.36	0.20	-1.78	0.07 ‘	-1.15	0.20	-5.81	0.00***	-1.34	0.21	-6.44	0.00***
:7	-0.76	0.30	-2.53	0.01*	-0.32	0.29	-1.11	0.27	-0.84	0.29	-2.95	0.00**	-1.31	0.32	-4.13	0.00***
:8	-0.26	0.19	-1.39	0.16	-0.34	0.20	-1.73	0.08 ‘	-0.86	0.19	-4.56	0.00***	-1.22	0.20	-6.14	0.00***
:> 8	-0.46	0.19	-2.48	0.01*	-0.27	0.19	-1.44	0.15	-1.01	0.18	-5.46	0.00***	-1.33	0.19	-6.85	0.00***
Cause																
:Under influence																
:Following too closely	0.35	0.31	1.15	0.25	-1.40	0.28	-5.05	0.00***	-2.51	0.32	-7.94	0.00***	-3.07	0.43	-7.14	0.00***
:Improper Turn	0.34	0.40	0.86	0.39	-0.81	0.35	-2.31	0.02*	-0.45	0.32	-1.42	0.16	0.16	0.32	0.51	0.61
:Speeding	0.26	0.24	1.04	0.30	-1.31	0.19	-6.91	0.00***	-1.90	0.18	-10.45	0.00***	-1.80	0.19	-9.67	0.00***
:Other Violations	-0.01	0.29	-0.05	0.96	-1.30	0.24	-5.50	0.00***	-1.61	0.23	-7.00	0.00***	-2.00	0.26	-7.84	0.00***
:Other Than Driving	0.90	0.90	1.00	0.32	-0.75	0.93	-0.80	0.43	-0.32	0.84	-0.38	0.70	1.69	0.77	2.20	0.03*
:Unknown	-0.77	0.64	-1.19	0.23	-2.46	0.63	-3.87	0.00***	-3.35	0.72	-4.64	0.00***	-1.84	0.55	-3.33	0.00***
Model overall																
Log-Likelihood: -9693.1 McFadden R^2: 0.09 Likelihood ratio test : chisq = 1894.7 (p.value = < 2.22e-16)																

1

1 Table 2 - Summary of results for mixed multinomial logit model

	Complaint of pain				Other visible injuries				Severe injury				Fatal			
Variable	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)
:(intercept)	0.25	0.31	0.82	0.41	1.37	0.32	4.33	0.00***	1.97	0.27	7.34	0.00***	1.29	0.29	4.48	0.00***
Age																
:Young adult																
:Adult	0.06	0.10	0.61	0.54	0.07	0.13	0.55	0.58	0.02	0.11	0.18	0.86	0.17	0.13	1.30	0.19
:Middle aged	0.04	0.11	0.33	0.74	0.16	0.14	1.11	0.27	0.09	0.12	0.73	0.46	0.33	0.14	2.31	0.02*
:Old	0.03	0.19	0.17	0.86	0.10	0.24	0.40	0.69	0.18	0.21	0.89	0.37	0.75	0.22	3.34	0.00**
Sex																
:Female																
:Male	-0.30	0.08	-3.68	0.00***	-0.08	0.11	-0.75	0.46	0.13	0.10	1.35	0.18	0.28	0.11	2.48	0.01*
Terrain																
:Flat																
:Mountainous	0.06	0.22	0.26	0.80	0.46	0.25	1.81	0.07’	0.47	0.22	2.15	0.03*	0.29	0.23	1.27	0.20
:Rolling	0.11	0.10	1.15	0.25	0.02	0.12	0.19	0.85	0.13	0.11	1.24	0.22	0.40	0.12	3.43	0.00***
Weather																
:Clear																
:Cloudy	0.11	0.11	1.04	0.30	-0.26	0.15	-1.74	0.08’	-0.19	0.13	-1.49	0.14	-0.17	0.14	-1.17	0.24
:Raining	-0.04	0.24	-0.15	0.88	0.26	0.30	0.87	0.38	-0.26	0.28	-0.95	0.34	-0.32	0.30	-1.05	0.30
:Fog	0.91	0.85	1.07	0.28	0.30	1.00	0.30	0.77	2.59	0.80	3.23	0.00**	2.17	0.85	2.55	0.01*
Light																
:Daylight																
:Dusk - Dawn	-0.43	0.23	-1.89	0.06’	-0.39	0.29	-1.33	0.18	0.20	0.23	0.89	0.37	1.00	0.23	4.29	0.00***
:Dark - Street Lights	-0.01	0.12	-0.11	0.91	-0.22	0.17	-1.28	0.20	0.61	0.13	4.75	0.00***	0.55	0.36	1.51	0.13
:Dark - No Street Lights	-0.13	0.15	-0.87	0.38	0.20	0.17	1.16	0.25	1.07	0.14	7.68	0.00***	1.91	0.15	13.04	0.00***
Vehicle type																
:Passenger Car																
:Truck	-0.16	0.24	-0.64	0.52	0.32	0.26	1.24	0.21	1.23	0.22	5.71	0.00***	1.81	0.22	8.34	0.00***
:Motorcycle	1.73	0.79	2.19	0.03*	3.43	0.74	4.61	0.00***	4.94	0.73	6.78	0.00***	4.88	0.73	6.67	0.00***
:Pickup Truck	0.00	0.11	-0.02	0.98	-0.01	0.15	-0.05	0.96	0.09	0.13	0.68	0.50	-0.08	0.15	-0.53	0.60
:Bus	1.00	1.19	0.84	0.40	1.00	1.35	0.73	0.46	2.58	1.20	2.15	0.03*	2.18	1.27	1.72	0.09’
:Emergency Vehicle	0.23	0.55	0.42	0.67	-1.40	1.24	-1.13	0.26	-0.54	0.86	-0.63	0.53	-0.81	1.20	-0.67	0.50
Number of lanes																
:2																
:3	0.45	0.72	0.63	0.53	0.63	0.81	0.77	0.44	-1.01	0.88	-1.15	0.25	0.66	0.68	0.97	0.33

	Complaint of pain				Other visible injuries				Severe injury				Fatal			
Variable	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)
:4	-0.09	0.20	-0.45	0.66	0.03	0.27	0.11	0.91	-1.05	0.36	-2.92	0.00**	-1.28	0.30	-4.21	0.00***
:5	0.33	0.36	0.92	0.36	0.35	0.45	0.77	0.44	-0.19	0.37	-0.52	0.61	-0.72	0.42	-1.72	0.09’
:6	-0.54	0.20	-2.71	0.01**	-0.25	0.27	-0.92	0.36	-1.82	0.39	-4.62	0.00***	-1.82	0.32	-5.65	0.00***
:7	-0.75	0.30	-2.51	0.01*	-0.14	0.38	-0.37	0.71	-0.87	0.30	-2.88	0.00**	-1.37	0.34	-4.07	0.00***
:8	-0.26	0.19	-1.36	0.17	-0.27	0.25	-1.06	0.29	-0.89	0.19	-4.70	0.00***	-1.30	0.20	-6.47	0.00***
:> 8	-0.46	0.19	-2.44	0.01*	-0.12	0.26	-0.47	0.64	-1.04	0.18	-5.64	0.00***	-1.47	0.20	-7.47	0.00***
Cause																
:Under influence																
:Following too closely	0.35	0.31	1.14	0.26	-1.44	0.29	-5.05	0.00***	-2.67	0.35	-7.64	0.00***	-3.28	0.44	-7.49	0.00***
:Improper Turn	0.35	0.40	0.86	0.39	-0.85	0.36	-2.39	0.02*	-0.41	0.34	-1.22	0.22	0.19	0.34	0.55	0.58
:Speeding	0.26	0.25	1.04	0.30	-2.21	0.63	-3.53	0.00***	-1.97	0.19	-10.14	0.00***	-1.95	0.20	-9.60	0.00***
:Other Violations	-0.01	0.29	-0.04	0.97	-1.35	0.24	-5.55	0.00***	-1.63	0.24	-6.73	0.00***	-2.20	0.28	-7.85	0.00***
:Other Than Driving	0.88	0.95	0.92	0.36	-0.81	0.94	-0.86	0.39	-0.57	0.91	-0.63	0.53	1.83	0.79	2.33	0.02*
:Unknown	-0.77	0.64	-1.21	0.23	-2.52	0.66	-3.84	0.00***	-3.53	0.78	-4.56	0.00***	-2.04	0.67	-3.04	0.00**
Random parameters																
Light :Dark - Street Lights													1.83	0.61	2.98	0.00**
Number of lanes :4									1.66	0.59	2.82	0.00**	1.36	0.40	3.41	0.00***
Number of lanes :6									1.66	0.53	3.10	0.00**	1.23	0.44	2.78	0.01**
Cause :Speeding					-2.14	0.92	-2.34	0.02**								
Model overall																
Log-Likelihood: -9684.8 McFadden R^2: 0.09 Likelihood ratio test : chisq = 1911.3 (p.value = < 2.22e-16)																

1

2

1 Table 3 - Summary of results for binary logit model (each column belongs to one of the five binary logits)

	Property damage only Versus All levels				Complaint of pain Versus All levels				Other visible injuries Versus All levels				Severe injury Versus All levels				Fatal Versus All levels			
Variable	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)
:(intercept)	-2.90	0.23	-12.38	0.00***	-2.51	0.23	-10.71	0.00***	-1.35	0.18	-7.55	0.00***	-0.62	0.16	-3.88	0.00***	-1.55	0.18	-8.62	0.00***
Age																				
:Young adult																				
:Adult	-0.06	0.08	-0.73	0.46	0.02	0.08	0.22	0.82	0.00	0.08	-0.05	0.96	-0.05	0.08	-0.65	0.52	0.09	0.10	0.95	0.34
:Middle aged	-0.11	0.09	-1.20	0.23	-0.06	0.09	-0.71	0.48	0.03	0.09	0.34	0.74	-0.04	0.09	-0.41	0.68	0.19	0.10	1.84	0.07 ’
:Old	-0.20	0.15	-1.35	0.18	-0.16	0.15	-1.09	0.27	-0.01	0.15	-0.10	0.92	-0.02	0.15	-0.16	0.88	0.51	0.16	3.10	0.00**
Sex																				
:Female																				
:Male	0.07	0.07	1.03	0.30	-0.34	0.07	-5.16	0.00***	-0.02	0.07	-0.27	0.79	0.15	0.07	2.09	0.04*	0.31	0.09	3.61	0.00***
Terrain																				
:Flat																				
:Mountainous	-0.26	0.17	-1.53	0.13	-0.20	0.16	-1.21	0.23	0.16	0.15	1.13	0.26	0.25	0.14	1.76	0.08 ‘	0.02	0.16	0.12	0.90
:Rolling	-0.15	0.08	-1.88	0.06 ‘	0.00	0.08	-0.06	0.95	-0.06	0.07	-0.83	0.41	0.02	0.07	0.29	0.77	0.24	0.08	2.96	0.00**
Weather																				
:Clear																				
:Cloudy	0.08	0.09	0.92	0.36	0.24	0.09	2.77	0.01**	-0.15	0.09	-1.59	0.11	-0.15	0.09	-1.59	0.11	-0.06	0.10	-0.59	0.56
:Raining	0.09	0.20	0.48	0.63	0.06	0.20	0.29	0.77	0.28	0.18	1.53	0.13	-0.17	0.21	-0.82	0.41	-0.28	0.23	-1.23	0.22
:Fog	-1.63	0.73	-2.24	0.03*	-0.64	0.49	-1.32	0.19	-1.01	0.53	-1.92	0.05 ‘	1.07	0.31	3.51	0.00***	0.74	0.35	2.12	0.03*
Light																				
:Daylight																				
:Dusk - Dawn	0.05	0.17	0.27	0.78	-0.53	0.19	-2.76	0.01**	-0.39	0.19	-2.12	0.03*	0.08	0.18	0.46	0.64	0.97	0.18	5.51	0.00***
:Dark - Street Lights	-0.27	0.10	-2.86	0.00**	-0.29	0.10	-3.00	0.00**	-0.39	0.10	-4.13	0.00***	0.35	0.09	4.00	0.00***	0.91	0.10	9.16	0.00***
:Dark - No Street Lights	-0.66	0.11	-6.05	0.00***	-0.83	0.11	-7.31	0.00***	-0.40	0.10	-4.12	0.00***	0.37	0.09	4.17	0.00***	1.37	0.09	14.94	0.00***
Vehicle type																				
:Passenger Car																				
:Truck	-0.81	0.17	-4.75	0.00***	-0.99	0.19	-5.29	0.00***	-0.45	0.15	-3.04	0.00**	0.40	0.12	3.27	0.00**	1.09	0.12	8.99	0.00***

	Property damage only Versus All levels				Complaint of pain Versus All levels				Other visible injuries Versus All levels				Severe injury Versus All levels				Fatal Versus All levels			
Variable	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)	Parameter Estimate	Standard Error	t-value	Pr(>|t|)
:Motorcycle	-3.95	0.71	-5.55	0.00***	-2.18	0.33	-6.70	0.00***	-0.32	0.15	-2.09	0.04*	1.32	0.12	10.86	0.00***	1.10	0.14	8.12	0.00***
:Pickup Truck	-0.02	0.09	-0.18	0.86	-0.02	0.09	-0.25	0.80	0.01	0.09	0.11	0.91	0.10	0.09	1.06	0.29	-0.06	0.11	-0.58	0.56
:Bus	-1.64	1.03	-1.59	0.11	-0.50	0.64	-0.78	0.43	-0.45	0.63	-0.72	0.47	1.23	0.46	2.65	0.01**	0.54	0.58	0.93	0.35
:Emergency Vehicle	0.26	0.50	0.53	0.60	0.62	0.46	1.34	0.18	-0.89	0.75	-1.19	0.24	-0.25	0.63	-0.40	0.69	-0.35	0.72	-0.49	0.63
Number of lanes																				
:2																				
:3	-0.24	0.63	-0.38	0.70	0.33	0.46	0.71	0.48	0.26	0.45	0.57	0.57	-1.44	0.62	-2.31	0.02*	0.73	0.40	1.84	0.07 ‘
:4	0.31	0.17	1.81	0.07 ‘	0.21	0.15	1.44	0.15	0.30	0.15	2.00	0.05*	-0.23	0.13	-1.70	0.09 ‘	-0.57	0.15	-3.89	0.00***
:5	-0.05	0.31	-0.17	0.86	0.38	0.25	1.52	0.13	0.31	0.25	1.25	0.21	-0.21	0.23	-0.88	0.38	-0.62	0.26	-2.40	0.02*
:6	0.77	0.16	4.69	0.00***	0.07	0.15	0.50	0.62	0.38	0.15	2.57	0.01*	-0.51	0.14	-3.72	0.00***	-0.71	0.15	-4.71	0.00***
:7	0.75	0.23	3.19	0.00**	-0.21	0.24	-0.88	0.38	0.39	0.22	1.75	0.08 ‘	-0.18	0.21	-0.85	0.40	-0.77	0.25	-3.06	0.00**
:8	0.59	0.16	3.68	0.00***	0.27	0.14	1.94	0.05 ‘	0.23	0.14	1.62	0.10	-0.33	0.13	-2.58	0.01**	-0.76	0.14	-5.32	0.00***
:> 8	0.69	0.16	4.40	0.00***	0.10	0.14	0.74	0.46	0.41	0.14	2.93	0.00**	-0.42	0.13	-3.34	0.00***	-0.79	0.14	-5.70	0.00***
Cause																				
:Under influence																				
:Following too closely	1.58	0.24	6.68	0.00***	2.01	0.24	8.30	0.00***	0.00	0.20	0.00	1.00	-1.30	0.24	-5.36	0.00***	-1.76	0.38	-4.64	0.00***
:Improper Turn	0.35	0.30	1.18	0.24	0.72	0.30	2.44	0.01*	-0.58	0.22	-2.59	0.01**	-0.27	0.18	-1.52	0.13	0.65	0.18	3.67	0.00***
:Speeding	1.38	0.17	8.04	0.00***	1.65	0.19	8.78	0.00***	-0.11	0.11	-0.99	0.32	-0.85	0.09	-9.10	0.00***	-0.62	0.10	-6.20	0.00***
:Other Violations	1.41	0.21	6.84	0.00***	1.34	0.23	5.89	0.00***	-0.07	0.16	-0.43	0.67	-0.42	0.14	-3.01	0.00**	-0.86	0.18	-4.84	0.00
:Other Than Driving	-0.44	0.75	-0.59	0.56	0.54	0.56	0.95	0.34	-1.35	0.61	-2.21	0.03*	-1.19	0.46	-2.60	0.01**	2.05	0.35	5.79	0.00
:Unknown	2.27	0.45	5.03	0.00***	1.14	0.58	1.96	0.05 ‘	-0.68	0.55	-1.24	0.21	-1.74	0.63	-2.78	0.01**	0.28	0.43	0.64	0.52
Model Overall																				
	Log-Likelihood: -3127.9 McFadden R^2: 0.072 Likelihood ratio test : chisq = 482.72 (p.value = < 2.22e-16)				Log-Likelihood: -3171.6 McFadden R^2: 0.075 Likelihood ratio test : chisq = 514.85 (p.value = < 2.22e-16)				Log-Likelihood: -3171.6 McFadden R^2: 0.075 Likelihood ratio test : chisq = 514.85 (p.value = < 2.22e-16)				Log-Likelihood: -3207.9 McFadden R^2: 0.057 Likelihood ratio test : chisq = 386.92 (p.value = < 2.22e-16)				Log-Likelihood: -2570.5 McFadden R^2: 0.134 Likelihood ratio test : chisq = 797.17 (p.value = < 2.22e-16)			

1

Table 4 - summary of number of true and false predictions for each model

		Binary outcome Model										Multi-level outcome Model					
		Property damage only vs. all other levels		Complaint of pain vs. all other levels		Other visible injury vs. all other levels		Severe injury vs. all other levels		Fatal vs. all other levels		Multinomial model with all five levels		Mixed multinomial model with all five levels		Support vector machine with all five levels	
		True	False	True	False	True	False	True	False	True	False	True	False	True	False	True	False
Prediction result for individual levels	Property damage only	0	620									216	404	215	405	208	412
	Complaint of pain			0	579							275	304	280	299	270	310
	Other visible injuries					0	609					17	592	17	592	30	579
	Severe injury							29	563			225	367	222	370	213	379
	Fatal									48	392	133	307	132	308	158	282
Total prediction	Total	2219	621	2258	582	2231	609	2244	596	2391	449	**866**	1974	**866**	1974	**879**	1962
	AUC	0.65		0.66		0.54		0.67		0.74							

* Shaded areas indicate the combination of all remaining levels for each model.

Table 5 - Confusion matrix for multinomial logit (1), mixed multinomial logit (2), and support vector machine (3)

		Reference				
		Property damage only	Complaint of pain	Other visible injury	Severe injury	Fatal
Prediction	Property damage only	216	197	168	115	65
	Complaint of pain	275	275	236	138	68
	Other visible injury	24	12	17	19	10
	Severe injury	67	53	122	225	164
	Fatal	38	42	66	95	133
	True prediction in %	**35 %**	**47 %**	**3 %**	**38 %**	**30 %**

(1) Multinomial Logit

		Reference				
		Property damage only	Complaint of pain	Other visible injury	Severe injury	fatal
Prediction	Property damage only	215	192	166	115	64
	Complaint of pain	278	280	238	136	68
	Other visible injury	24	13	17	25	14
	Severe injury	67	50	117	222	162
	Fatal	36	44	71	94	132
	True predictions in %	**35 %**	**48 %**	**3 %**	**38 %**	**30 %**

(2) Mixed Multinomial Logit

		Reference				
		Property damage only	Complaint of pain	Other visible injury	Severe injury	fatal
Prediction	Property damage only	208	178	147	108	62
	Complaint of pain	266	270	229	121	66
	Other visible injury	20	21	30	19	15
	Severe injury	64	53	119	213	139
	Fatal	62	58	84	131	158
	True predictions in %	**34 %**	**47 %**	**5 %**	**36 %**	**36 %**

(3) Support Vector Machine