Analysis of crash injury severity on two Trans-European Transport Network corridors in Spain using discrete-choice models and random forests

Bahar Dadashova[1,a], Blanca Arenas-Ramires[b], Jose Mira-McWillaims[b], Karen Dixon[a] and Dominique Lord[a, c]

[a]Texas A&M Transportation Insitute, Texas A&M University System, 2935 Research Pkwy, College Station 77843-3135, TX, USA
[b]University Institute of Automobile Research (INSIA), Technical University of Madrid (UPM), Jose Gutierrez Abascal 2, 28006 Madrid, Spain
[c]Zachary Department of Civil and Environmental Engineering, Texas A&M University System, 2935 Research Pkwy, College Station 77843-3135, TX, USA

Abstract

Objective

The objective of this paper is to identify the list of crash severity contributing factors and evaluate their impact on multiple-vehicle crashes on two high use Trans-European interurban, freight corridors in Spain (southern Europe): Madrid - Irùn and Barcelona – Almerìa.

Methods

We have used both logistic regression and random forests to identify crash severity predictors and estimate their impacts on crash outcomes. Although both statistical methods can provide useful information to help explain the safety implications of highway crashes, using both methods may further enable a more comprehensive understanding of this phenomenon. For this effort, we disaggregated the crash data into different crash types (i.e., head-on, angle, sideswipe and rear-end) and analyzed this data using roadway design elements, driver characteristics, and environmental factors. To identify the most important predictors of crash severity, we used the random forests data mining approach. We then used ordered logit models to estimate the effect of external factors on the severity of each crash type. Finally, we assessed the accuracy of the model estimates using bootstrap sampling.

Results

The results of data mining analyses indicated that roadway design factors such as horizontal and vertical curvature, superelevation, and lane and shoulder width are among the most important factors associated with crash severity. The results of logistic regression show that the impact of the selected roadway element on the crash outcome is conditional on the crash type and the direction of the effects is not always consistent.

Conclusions

The contribution of this paper to the existing literature is two-fold: the first important contribution of the paper is related to the safety analysis of two of the most important freight corridors in Spain and southern Europe. The second contribution of this paper is to address the existing gap in the literature relating to the comparison and compatibility of data mining and the logistic regression model.

Keywords*: Crash injury severity; crash type; roadway geometry design; logistic regression; random forests; bootstrap.*

[1] Corresponding author: Bahar Dadashova, PhD; Tel.: (979)317-2137; Email: b-dadashova@tti.tamu.edu.

Introduction

In this paper, we have analyzed crash severity data collected from two interurban freight corridors in Spain: the corridor connecting Madrid (central region) with Irùn (north) and the corridor connecting Almerìa (south-east) with Barcelona (north-east). These corridors constitute a part of the trajectory of two International Rail Freight Corridors across Spain, as part of the TEN-T (Trans-European Transport Network) Program, and are the subjects of major transport policy changes (MFOM). Crash data were collected during three years, 2010-2012.

Crash injury severity depends on various factors, such as roadway and roadside geometrics, driver characteristics and environmental factors. Studies exploring the impacts roadway design elements on crash severity found that design speed, horizontal and vertical alignment, superelevation, number of lanes, and lane and shoulder widths are the most significant factors affecting crash severity. The greater rate of vertical alignment (grade) and superelevation have been mainly found to have a negative impact on both crash frequency and severity (Bauer and Harwood, 2013; Savolainen and Mannering, 2007; Kopelias et al., 2007). Wider lane and shoulder widths are mostly found to have a positive effect on decreasing the crash frequency; however, their effect on the crash severity might be negative in some cases (Stamatiadis et al., 2009; Karlaftis and Golias, 2002, Zhu et al., 2010). This counterbalancing effect has been explained by the fact that wider lanes may result in increased operating speeds which consequently might increase the crash severity.

In roadway safety, driver behavior is a significant contributing factor to the crash occurrence and a substantial amount of research has been dedicated to this area (Lee, 2008; Dewar et al., 2007). Drivers perceive the surrounding environment and act upon through attention and information processing, vision, perception-reaction time and speed choice. All these functions can be affected by the driver's age and gender, fatigue (sleepiness and dizziness), driver impairment (alcohol and drug) and distraction (texting). The presence of other passengers may this factor driving behavior although it's effect has not been thoroughly analyzed.

Although not always homogenous, however, crash severity can be categorized as an ordinal categorical process. This categorization is normally based on KABCO (where K is fatal, A is suspected serious injury, B is non-incapacitating injury, C is possible injury and O is no injury or Property Damage Only -PDO) and Abbreviated Injury Scale (AIS) scales. KABCO stands for Although these two systems use different scales and are assigned by police (KABCO) and hospital (MAIS) however the crash injury level can be translated from one system to another (Burch et al., 2014). Because of this categorization, existing literature has documented the use of discrete-choice ordered models, such as ordered logit and probit models, mixed logit, nested models, as well as data mining tools, such as classification and regression trees and random forests, for analyzing severity data (see Savolainen et al., 2011, Harb et al., 2009; Khan et al., 2015; Shi and Abdel-Aty, 2015). Data mining tools are useful methods for detecting and interpreting higher-order interactions observed among the crash-contributing factors which might be a very difficult task when using logistic regression. However, the data mining tools are a

black box in nature and cannot be used for developing the crash severity prediction models, nor can they be used to quantify the impact of the contributing factors on injury severity probabilities. In this paper, we are using random forests and logistic regression models to identify the list of the most influential crash-contributing factors and estimate their impacts on the severity of crashes on two Trans-European Network corridors in Spain: Madrid - Irùn and Barcelona – Almerìa.

Methods

Data Overview

The data used in this paper originated from two different sources: 1) police crash reports and 2) roadway geometric data. The crash reports were obtained from the General Directorate of Traffic, while the roadway geometry design data were obtained from the Road Inventory Database from the General Directorate of Highways of the Ministry of Transportation. The first database contained crash reports, crash locations, and various crash-contributing factors such as driver, roadway, vehicle and weather conditions and was collected from 2010 to 2012. A total of 3,722 fatal and injury crashes took place during this period (Appendix Table A1).

The second database contains road geometry design information that includes route type, number of through lanes, lane width, shoulder width, median type and width, horizontal and vertical curvature and superelevation. The two corridors were divided into 185 roadway segments stretching across 1,262.55 kilometers of highway. The corridors have three facility types: *autovia* (divided, multi-lane freeway), *autopista* (divided, multi-lane toll freeway), and *carretera* (two-lane, two-way rural highway). Table 1 shows the number and length of roadway segments per facility type and corridor.

Table 1. Road Types and Corresponding Length

The two databases were integrated using the distance from the origin (DFO) and the unique route ID. The descriptive statistics of all the variables are shown in the Appendix (Table A2).

Random Forests

The random forests method was proposed by Breiman (2001) and is considered to be one of the most efficient classification methods. RF method has garnered mostly favorable reviews when compared to logistic regression, quadratic discriminant analysis, support vector machines, classification and regression trees, and others (Verikas et al., 2011). The random forests method is based on the bagging principle and the random subspace method (Breiman, 2001) that relies on constructing a collection of decision trees with random predictors. The general architecture of the random forests using decision trees is described below (Verikas et al., 2011):

i. Generate a bootstrap sample of size N_c from the overall data, N to grow a $tree_B$ by randomly selecting the predictors $X = \{x_{i,i=1,\dots,I}\}$ (we will call this bootstrap sample, a *cluster*).

ii. Use the predictor x_i at the node n of the $tree_B$ to vote for the class label k_B in this node. At each node, only one predictor providing the best split is selected.

iii. Run the out-of-bag data $(N - N_c)$ down the $tree_B$ to obtain the misclassification rate, i.e. out-of-bag error rate, $OOBER_B$.

a. Repeat i.-iii. for a large number of trees until the minimum out-of-bag error rate, $OOBER_B$, is obtained.

b. Assign each observation to final class k by a majority vote by averaging over the set of trees.

In random forests, a large number of trees are used to decrease the out-of-bag (OOB) error rate. To avoid the overfitting of random forests, it is suggested to tune the hyperparameters, such as limiting the maximum depth and hence the number of samples in the leaf. OOB data are also used to estimate the importance of the variables. OOB error rate and variable importance are the two most important byproducts of random forests.

Variable importance ranking is measured by the classification accuracy and Gini impurity coefficient. The classification accuracy measure computes the mean decrease in the classification accuracy of the OOB data $(N - N_c)$. This importance measure shows how much the mean squared error (MSE) or the impurity increase when the specified variable is randomly permuted. If prediction error does not change by permuting the variable then the importance measures will not be altered significantly which in turn will change the MSE of the variable only slightly (low values). This implies that the specified variable is not important. On the contrary, if the MSE significantly decreases during the permutation of the variable then the variable is deemed as important. The classification accuracy measure of the variable is averaged over the number of trees, *B*, used to construct the random forests:

$$MDA(x_i) = \frac{\sum_{tree=1}^{B} MDA^{tree}(x_i)}{B} \qquad (2.1)$$

where $MDA(x_i)$ is the average importance rate of the variable x_i and $MDA(x_i)$ is the importance rate of the same variable in $tree = \{tree_{b,b=1,\dots,B}\}$.

The mean decrease in Gini impurity coefficient computes the contribution of the variable to the homogeneity of the nodes and leaves in the resulting RF. The Gini impurity coefficient is a measure of homogeneity from 0 (homogeneous) to 1 (heterogeneous):

$$MDG^n(x_i) = 1 - \sum_{k=1}^{K} p(k|n) \qquad (2.2)$$

where $MDG^n(x_i)$ is the Gini impurity coefficient of the variable x_i at the node n; $p(k|n)$ is the probability of class k in node n (weights) and K is the number of classes. Each time a specified

variable is used to split a node, the Gini impurity coefficient for the child nodes is calculated and compared to that of the parent node. Usually, after the split of a node, the impurity in the child node becomes smaller than the parent node. The changes in Gini impurity coefficient are summed for each variable and normalized at the end of the calculation. Summing up the Gini impurity coefficient measures for each variable all over the trees gives the importance rate which is often consistent with the permutation importance measure (Breiman, 2001), thus the variable with the higher impurity is deemed as more important.

Logistic Regression Models

Logistic regression is frequently used in crash severity analysis. In logistic regression, the outcome, or dependent variable is a binary or dichotomous factor. For example, the outcome of a crash could be a fatality, injury or PDO. In this paper we are only considering two outcomes, hence we will use the logit model. Logit model estimates the probability ($p(y_i)$) of a crash (y_i) resulting in either fatal injury (KA) or non-fatal injury (BC) crash given the values of the explanatory variables ($\boldsymbol{X}$):

$$logit(p(y_i)) = \log\left(\frac{p(y_i)}{1 - p(y_i)}\right) \tag{2.3}$$

where:

$$p(y_i) = \frac{\exp(\boldsymbol{\beta} \cdot \boldsymbol{X})}{1 + \exp(\boldsymbol{\beta} \cdot \boldsymbol{X})} \tag{2.4}$$

$$\boldsymbol{\beta} \cdot \boldsymbol{X} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{2.5}$$

where β_k is the estimated impact of variable x_k on the response variable y_i. Increasing β_k indicates increasing severity, since the higher order of y_i, refers to fatality or severe injury as shown in equation (2.4).

Finally, the results of the logistic regression were replicated using *bootstrap* sampling in order to test the consistency of estimates. Bootstrapping is a computer-based method to assess the accuracy measure of model estimates (Efron and Tibshirani, 1986). Low bootstrap bias and standard error indicate that the coefficient estimate is reliable.

Results

Data Mining Using Random Forests

We first grouped the data into four clusters where each cluster represents one of the following crash types: head-on, angle/turning movement, sideswipe and rear-end. We then built random forests by setting the optimal number of trees based on the out-of-bag (OOB) sample error rate. Figure 1 shows that after tree #40 the OOB error rate of all clusters becomes highly steady thus indicating that using 100 trees for building the random forests would be a rational choice.

Figure 1. Out-of-bag error rate ($tree_{B=100}$)

After determining the number of optimal trees, we ran a random forests analysis to determine the list of the most important variables associated with crash severity as shown in Figure .

Figure 2. Variable importance based on random forests

According to both the Gini impurity and mean decrease coefficients the most important variables affecting crash severity are grade, superelevation, radius, roadbed width, lane width, and shoulder width.

We then used the list of most important variables to build a decision tree where each node was assigned a class label, k, that shows the most likely outcome of the crash once it takes place: $k_1 = Fatal\ Injury$ or $k_2 = Non - fatal\ Injury$. The decision trees are constructed using the most important variables and the tree with the minimum OOB error rate since the tree with a lower error rate is assumed to be a stronger classifier (Figure). The number of nodes was limited to 8-12 while the depth of the tree was set to be lower than 50 in order to avoid overfitting. Note that the decision tree built using random forests is not a common practice and does not represent the crash as a system. Firstly, because the variable importance and error rate, which are the most important statistics in RF analysis, are averaged over 100 trees and selecting only one tree, even the one with the minimal error rate, may not show the true importance ranking of the variable since it could be masked by another correlated input. Secondly, by limiting the nodes to 8 or 12 (opposed to the original depth of a tree > 50) might cause estimation bias. However, we assume that extracting decision trees might help us to explain the outcome of RF better and understand the direction of effects for most important variables.

Figure 3. Decision tree for crash severity

The first split of the tree represents grade (vertical alignment) variable (grade >-6.05%) showing that it is the most important variable explaining crash injury severity. Note that a negative grade indicates a sag curve while the positive grade indicates a crest. Crashes occurring on the curves with a grade value >-6.05% have a higher probability of resulting in a severe injury or fatality. The second split of the tree occurs by roadbed width (<10.3 m) thus showing that this variable is the second most important variable explaining crash injury severity. The third and fourth splits occurred by shoulder width (>2.7 m) and grade (-3.5%) variables. The terminal node created at the end of these splits (1, 2, 3 and 4) shows that crashes occurring at a road segment with a roadbed width smaller than 10.3 m and shoulder width wider than 2.7 m are more likely to result in a fatality if the grade is between -3.5% and -6.5%, (-6.5% < grade <-3.5%). Otherwise, a minor injury is likely to be the outcome of the crash. In other words, for the given conditions, a steeper grade would affect the crash outcome negatively as far as the crash severity is concerned.

The fifth split of the tree is the radius (i.e., horizontal curve). A radius value larger than 84.2 m can be expected to be associated with crashes that are less severe, at a location with a wider lane (< 10.3 m). A smaller radius (i.e., sharp horizontal curve) may have a significantly negative

impact on the crash outcome if the grade (6th split) is between -6.5% and 4.5%. The radius will have a less severe impact on crash severity if the grade is higher than 4.5 m.

The lane width variable again creates the seventh split of the tree. The results in this terminal node show that if the grade is larger than 6.5% and the shoulder is narrower than 2.7 m, then a crash occurring at the segment with a relatively wider roadbed (7.05 m < lane width < 10.3 m) is most likely to result in a minor injury. If the lane is narrower than 7.05 m, crashes are more likely to result in a fatality given that the grade is larger than 6.5%.

Logistic Regression and Bootstrapping

We first conducted a multicollinearity test to remove the variables that may impact model performance. As can be observed in Figure 4, there is high collinearity between route type, the number of through lanes and median and roadbed widths, as well as left shoulder width and median and roadbed widths. Also, there is a high-collinearity between pavement conditions and weather conditions. Therefore, the number of through lanes, left shoulder width, and weather conditions were removed, and the logistic model was run with the remaining variables for each crash type: head-on, angle, sideswipe, and rear-end crashes. The final model for each crash type was based on the Akaike information criterion (AIC) statistic, variable significance, and the consistency of the variable effects. Because the coefficient estimates of radius and daylight were not interpretable, these variables were also removed from the final models.

Figure 4. Multicollinearity test

Because the logistic regression was estimated based on a small number of crashes (e.g., 212 head-on crashes) the goodness of fit statistics of the variable coefficient estimates may be biased. Therefore, to assess the accuracy of the variable coefficient estimates, we used bootstrapping techniques. To estimate the accuracy of the logit model estimates, we developed a bootstrap sampling with 1,000 replications for each of the crash types. Appendix Table A3 shows the results of logistic regression and bootstrap sampling in terms of the odds ratio, standard error of coefficient estimates, p-value and bootstrap bias. The results of bootstrapping show that the bias is quite small for all of the variable estimates.

Driver-related variables affecting the severity of all crash types are found to be the age group and gender, presence of alcohol, sleep, and the number of passengers. According to the results of logistic regression, crashes involving middle and older adults have higher probability of resulting in a fatality; the odds ratio is particularly high for head-on crashes; the probability of head-on, angle, sideswipe and rear-end crash resulting in a fatality, and severe injury is 2.7, 1.4, 1.07, and 1.2 times higher for middle-aged driver, and 1.4, 1, 4, 0.7 and 1.6 times higher for older drivers compared to younger drivers. This result is supported by the existing literature and is explained to occur due to younger adults being more resilient and having a higher recovery rate (Yasmin et al., 2014). The results of the logistic regression show that angle, sideswipe and rear-end crashes

involving male drivers have a lower probability of resulting in a fatality and severe injury, which may be due to their stronger constitution compared to females (Yasmin et al., 2014).

Alcohol is found to be associated with less severe injuries for head-on crashes. Although counterintuitive, however, this result does not seem to be very far fetched. For example, a study by Friedman (2012), found that intoxication may have a substantial protective effect, and may result in less severe injuries once a crash takes place.

Sleep and drowsiness, on the other hand, is found to increase the crash severity significantly; odds of a head-on, angle, sideswipe and rear-end crash resulting in a fatality is found to be 3.8, 2.5, 1.2 and 2.6 times higher if the driver is sleepy. Finally, the number of passengers is found to reduce the severity of head-on and rear-end crashes, and increase the severity of angle and sideswipe crashes.

Roadway design and environmental factors affecting crash severity are grade, superelevation, lane and shoulder width, and time of day. According to the estimation results, the increasing grade is found to increase the severity of head-on, angle and rear-end crashes by 10% (odds ratio = 1.1). It does not seem to affect the severity of the sideswipe crashes. Super-elevation is also found to increase the severity of crashes by approximately 10%; except for sideswipe crashes, where the odds ratio is equal to 1.7.

The impact of the lane and shoulder widths on crash severity will depend on the collision type. Lane width was found to have a positive effect in decreasing the severity of head-on (by 0.9), sideswipe (by 0.9), and rear-end (by 0.6) crashes, while it was found to increase the severity of angle crashes (by 1.1). Wider left side shoulder (towards the median) is associated with more severe angle and sideswipe crashes, and less severe rear-end crashes. Wider right shoulder width, on the other hand, is associated with more severe angle and rear-end crashes, and less severe sideswipe crashes. Although wider lane and shoulder widths are mostly found to have a positive effect on decreasing the crash frequency, their effect on the crash severity might be negative in some cases (Zhu et al., 2010; Stamatiadis et al., 2009). This counterbalancing effect has been explained by the fact that wider lanes may result in increased operating speeds which consequently might increase the crash severity (Stamatiadis et al., 2009).

Discussions

In this paper, we have used the roadway geometry design, driver and environmental factors to explain the severity outcomes of head-on, angle, sideswipe and rear-end crashes on two TENT freight corridors in Spain using modern and innovative data-driven tools. We conducted data analyses through random forests and logistic regression. Both methods have been previously used in roadway safety applications; however, their compatibility has not been thoroughly addressed.

The results of data analysis indicate that the estimation results of random forests and logistic regression match for the most part. Both models indicate that roadway design characteristics grade, superelevation, lane and shoulder widths, and driver characteristics such as age group,

sleep, and the number of passengers are important variables for explaining the crash severity. However, the two approaches do not always produce similar results; for example, although radius and roadbed width were found to be important in explaining crash severity, these variables were not included in the logistic regression estimation either because of the collinearity or because their inclusion in the model did not improve model performance. On the other hand, the gender of the driver and impairment (alcohol) were included in the logistic regression model, although these two variables were found to be the least important according to random forests. Aside from this finding, estimation results are mostly confirmatory and concur with the existing literature.

The contribution of this paper to the existing literature is two-fold: the first important contribution of the paper is related to the safety analysis of two of the most important freight corridors in Spain and southern Europe. The second contribution of this paper is to address the existing gap in the literature relating to the comparison and compatibility of data mining and the logistic regression model. We do however acknowledge that several gaps need to be addressed in the future. One of the most important gaps is related to the limitation of data; although a considerable effort was made for collecting and integrating data, some of the important crash-contributing factors were not available (e.g., speed limit) hence were not included in the study.

Acknowledgments

This work has been carried out in the framework of the MODALTRAM - TRA2011-28647-C02-01 Research Project "*Development of an integrated methodology for the assessment of effects on Safety and Environment, for the road and rail modal shift*", of the Spanish National Research Plan 2011-2016, Ministry of Economy and Competitiveness (MINECO). The authors would like to thank the two anonymous reviewers and the editors for their feedback and comments that have helped to significantly improve the quality of the paper.

References

Bauer K, Harwood D. Safety effects of the horizontal curve and grade combinations on rural two-lane highways. *Transportation Research Record: Journal of the Transportation Research Board*, 2013;2398:37-49.

Breiman, L. Random forests. *Machine Learning*, 2001;45(1):5-32.

Burch C, Cook L and Dischinger P. A comparison of KABCO and AIS injury severity metrics using CODES linked data. *Traffic Injury Prevention*, 2005;5(6):627-630.

Chang LY, Chen WC. Data mining of tree-based models to analyze freeway crash frequency. *Journal of Safety Science*, 2005;36:365-375.

Das A, Abdel-Aty M, Pande A. Using conditional inference forests to identify the factors affecting crash severity on arterial corridors. *Journal of Safety Research*, 2009;40(4):317-327.

Dewar RE, Olson PL, Gerson JA. Human factors in traffic safety, 2007.

Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1986;54-75.

Friedman LS. Dose–response relationship between in-hospital mortality and alcohol following acute injury. *Alcohol*, 2012:46(8):769-775.

Harb R, Yan X, Radwan E, Su X. Exploring precrash maneuvers using classification trees and random forests. *Accident Analysis & Prevention*, 2009;41(1):98-107.

Karlaftis MG and Golias I. Effects of road geometry and traffic volumes on rural roadway crash rates. *Accident Analysis & Prevention,* 2002;34(3):357-365.

Khan G, Bill AR, Noyce DA. Exploring the feasibility of classification trees versus ordinal discrete choice models for analyzing crash severity. *Transportation Research Part C: Emerging Technologies*, 2015;50:86-96.

Kopelias P, Papadimitriou F, Papandreou K, Prevedouros P. Urban Freeway Crash Analysis: Geometric, Operational, and Weather Effects on Crash Number and Severity. *Transportation Research Record: Journal of the Transportation Research Board*, 2007;2015:123-131.

Lee JD. Fifty years of driving safety research. *Human Factors*, 2008;50(3):521-528.

Savolainen PT, Mannering FL, Lord D, Quddus MA. The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives. *Accident Analysis & Prevention,* 2011;43(5):1666-1676.

Savolainen PT, Mannering FL. Probabilistic models of motorcyclists' injury severities in single-and multi-vehicle crashes. *Accident Analysis & Prevention*, 2007;39(5):955-963.

Shi Q, Abdel-Aty M. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. *Transportation Research Part C: Emerging Technologies*, 2015;58:380-394.

Stamatiadis N, Pigman J, Sacksteder J, Ruff W, Lord, D. *Impact of shoulder width and median width on safety.* Washington DC: National Cooperative Highway Research Program Report 633; 2009.

Verikas A, Gelzinis A, BacauskieneM. Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 2011; 44(2):330-349.

Wijnen W, Stipdonk H. Social costs of road crashes: An international analysis." *Accident Analysis & Prevention*, 2016;94:97-106.

Yasmin S, Eluru N, Bhat CR, TayR. A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. *Analytic methods in accident research*, 2014;1: 23-38.

Zhu H, Dixon KK, Washington S, Jared D. Predicting single-vehicle fatal crashes for two-lane rural highways in Southeastern United States. *Transportation Research Record: Journal of the Transportation Research Board*, 2010;2147:88-91.

Appendices

Table A1. Number of Crashes per Crash Severity and Crash Type.

Corridor	Crash Severity	Crash Type	2010	2011	2012	Total
Barcelona-Almeria	Non-fatal Injury (BC)	Head-On	31	26	18	75
		Angle	181	145	92	418
		Sideswipe	229	181	94	504
		Rear-End	490	472	236	1,198
		Total	**931**	**824**	**440**	**2,195**
	Fatal Injury (KA)	Head-On	48	28	23	99
		Angle	48	35	27	110
		Sideswipe	50	21	8	79
		Rear-End	62	53	19	134
		Total	**208**	**137**	**77**	**422**
Madrid-Irun	Non-fatal Injury (BC)	Head-On	2	6	4	12
		Angle	38	31	25	94
		Sideswipe	57	24	32	113
		Rear-End	246	269	224	739
		Total	**343**	**330**	**285**	**958**
	Fatal Injury (KA)	Head-On	2	13	11	26
		Angle	25	23	12	60
		Sideswipe	6	2	2	10
		Rear-End	29	14	8	51
		Total	**62**	**52**	**33**	**147**

Table A2. Description of Variables.

Quantitative Variable	Min	Max	Mean	S.D.
Grade (%)	-7%	7.4%	-0.2%	1.5
Lane Width (meters)	2.5	8.3	3.7	0.56
Median Width (meters)*	0	40	7.6	6.8
Number of Through Lanes (both directions)	2	13	4.2	1.8
Number of Passengers	0	>4	1.63	1.12
Radius (meters)*	0 (tangent)	9,000	5,292.1	3,917.2
Roadbed Width (meters)	3.4	25	9.1	2.7
Shoulder Width, Right (meters)*	0	5.1	1.9	0.7
Shoulder Width, Left (meters)*	0	5.1	1.2	0.5
Super Elevation (%)	-8.9%	8.5%	-0.9%	2.4

Qualitative Variable	Description
Age group	1=Young (18-35); 2=Middle (36-55); 3 = Old (56-80).
Alcohol	0 = Not Present; 1 = Present
Crash Type	1 = Head-on Crash; 2 = Angle Crash; 3 = Sideswipe Crash; 4 = Rear-end Crash
Crash Severity	0 = Injury; 1 = Fatal
Corridor	1= Barcelona-Almeria; 2=Madrid-Irun;
Daytime and Lighting Conditions	1=Daylight; 2=Twilight; 3=Illuminated night time; 4=Poorly illuminated night time; 5=Nighttime
Gender	0=Male ; 1=Female
Pavement Conditions	1=Dry and clean; 2=Humid; 3=Wet; 4=Frozen; 5=Snow covered; 6=Dirt; 7=Grease.
Province	Alava; Alicante; Almeria; Barcelona; Burgos; Castellon; Madrid; Murcia; Segovia; Tarragona; Valencia.
Route Type	1= Divided multi-lane freeway; 2= Divided multi-lane toll road; 3= Two-way two-lane highway.
Sleep & Drowsiness	0 = Not Present; 1 = Present
Traffic Density	1=Fluid; 2= Dense; 3=Congested.
Weather Conditions	1=Fine; 2=Dense Fog; 3=Light fog; 4= Drizzle; 5= Rain; 6= Hail; 7= Snow; 8= Strong Wind.
Year	2010; 2011; 2012

*0 indicates that the corresponding element does not exist on that road segment (where every segment is 10 meters).

Table A3. Logistic Regression and Bootstrap Results

Exogenous Variables	Head-on Crashes				Angle Crashes				Sideswipe Crashes				Rear-end Crashes			
	Odds Ratio	Std. Err.	p-value	Bootstrap Bias	Odds Ratio	Std. Err.	p-value	Bootstrap Bias	Odds Ratio	Std. Err.	p-value	Bootstrap Bias	Odds Ratio	Std. Err	p-value	Bootstrap Bias
Intercept	2	1.4	0.6	0	0.1	0.7	<0.01	0.2	0	1	<0.01	0.1	0.4	2	1.4	0.6
Age Group (1 if middle-aged, 0 young)	2.7	0.4	0	-0.1	1.1	0.2	0.6	0.1	1.1	0.3	0.8	0.1	1.2	2.7	0.4	0
Age Group (1 if old, 0 young)	1.4	0.5	0.5	0	1.4	0.3	0.2	0.1	0.8	0.4	0.5	0.1	1.7	1.4	0.5	0.5
Alcohol (1 if present, 0 otherwise)	0.2	0.7	0	0.8	0.7	0.4	0.5	0.1	0.9	0.5	0.8	0.2	0.9	0.2	0.7	0
Gender (1 if male, 0 female)	0.8	0.4	0.6	0	1	0.2	0.9	0	0.7	0.3	0.3	0	0.5	0.8	0.4	0.6
Grade (%)	1.1	0.1	0.5	0.2	1.2	0.1	<0.01	0.1	1	0.1	1	2.1	1.2	1.1	0.1	0.5
Lane Width (meters)	0.9	0.3	0.7	2.5	0.9	0.1	0.9	0	1	0.2	0.5	0	0.6	0.9	0.3	0.7
Number of Passengers	0.9	0.2	0.4	0.4	1	0.1	0.8	0	1.1	0.1	0.6	0	1	0.9	0.2	0.4
Daytime (1 if twighlight, 0 otherwise)					1.6	0.5	0.3	0	0.5	0.8	0.4	0				
Daytime (1 if night, illum., 0 otherwise)	0.2	0.9	0	0	0.9	0.4	0.8	0	3.9	0.5	<0.00	0	0.4	0.2	0.9	0
Daytime (1 if night, poor illum., 0 otherwise)	0.1	0.7	0	0	0.7	0.4	0.5	0	4.1	0.4	<0.01	0	1	0.1	0.7	0
Daytime (1 if night, no ilum. 0 otherwise)	0.9	0.4	0.8	0	1.9	0.3	<0.01	0	2.3	0.3	<0.01	0	2.4	0.9	0.4	0.8
Shoulder Width, Left (meters)	1	0.3	0.9	0	1.5	0.1	<0.01	0	1.7	0.2	<0.01	0	0.9	1	0.3	0.9
Shoulder Width, Right (meters)	1	0.3	0.9	0	1.1	0.1	0.7	0	0.8	0.2	0.3	0	1.2	1	0.3	0.9
Sleep (1 if drowsy, 0 otherwise)	3.8	0.5	0	-0.1	2.5	0.4	<0.01	0	1.3	0.5	0.6	0.1	2.6	3.8	0.5	0
Super Elevation (%)	1.1	0.1	0.3	0	1.1	0	0.2	0	1.7	0	0.2	0	1	1.1	0.1	0.3
Goodness of Fit (AIC)	259				757				531				1,199			

Tables

Table 1. Road Types and Corresponding Length

Road Type	Madrid-Irún		Barcelona-Almería		Total	
	Segments	Length (km)	Segments	Length (km)	Segments	Length (km)
Total	90	548.1	95	714.3	185	1262.5
Divided multi-lane freeway	49	254.7	54	300.6	103	555.4
Two-way two-lane highway	33	207.1	--	--	33	207.1
Divided multi-lane toll road	8	86.2	41	413.7	49	499.9

Figures

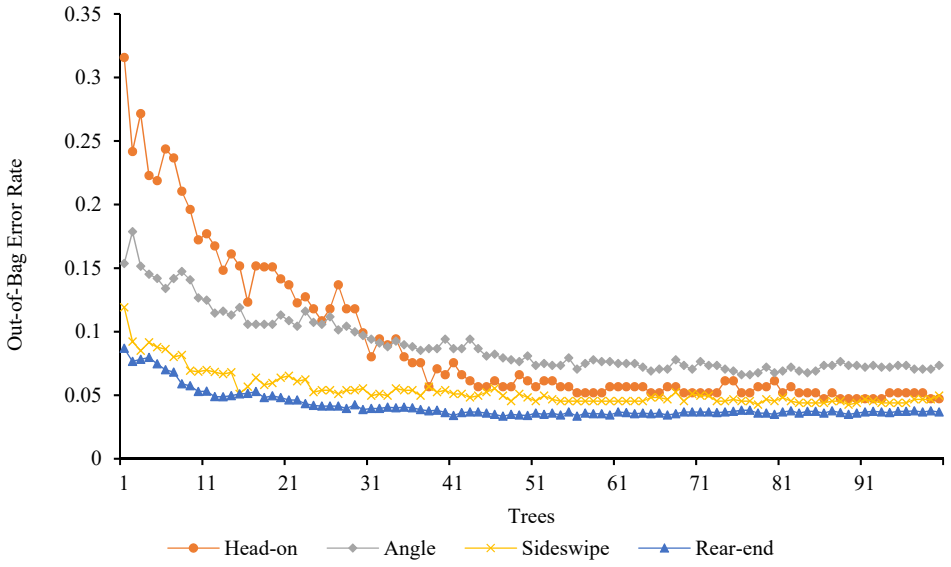

Figure 1. Out-of-bag error rate ($tree_{B=100}$)

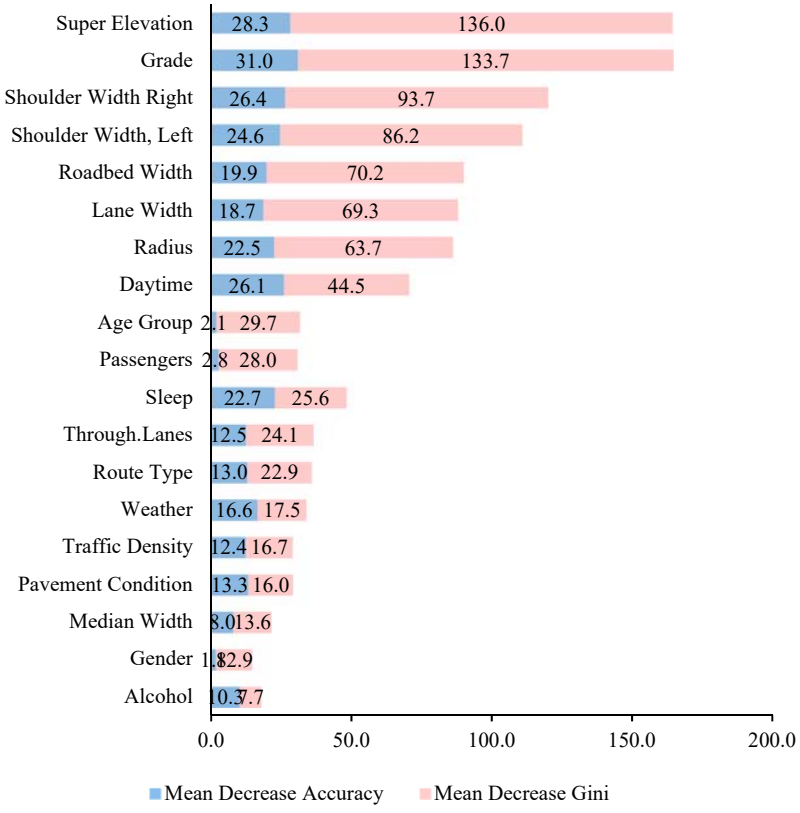

Figure 2. Variable importance based on random forests

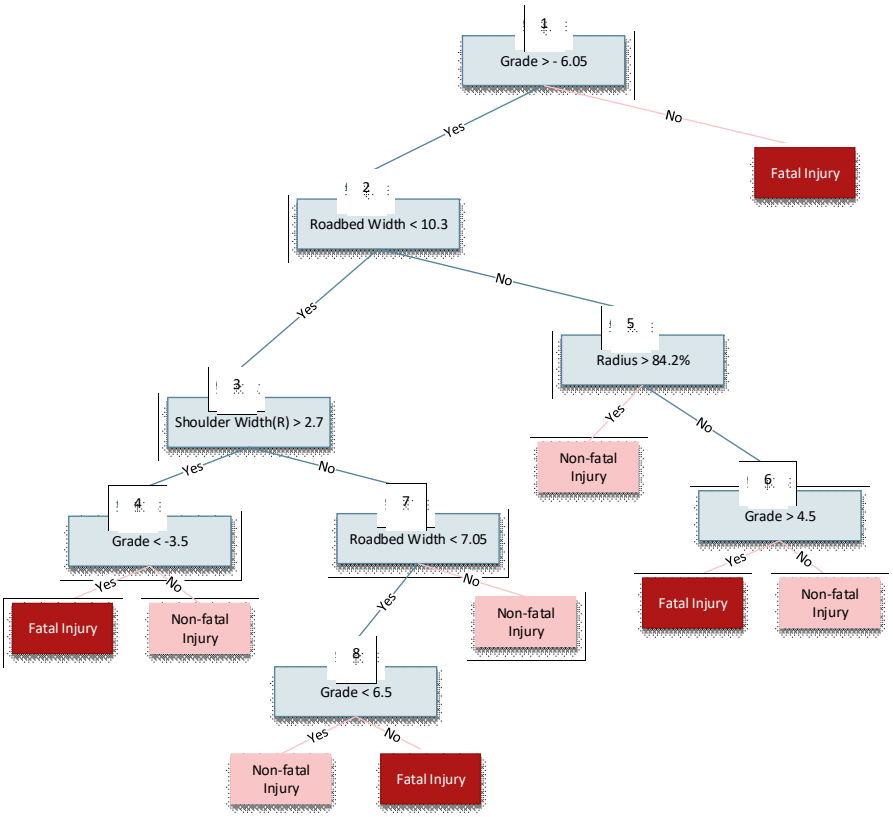

Figure 3. Decision tree for crash severity

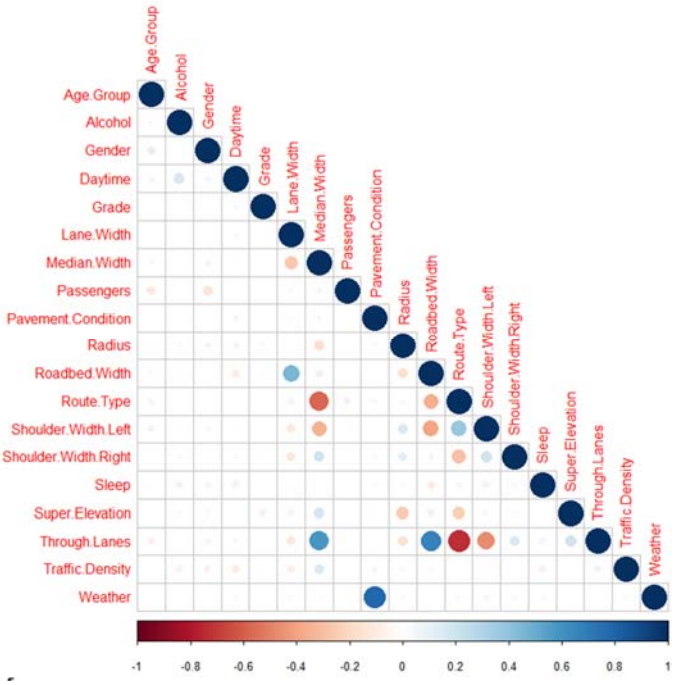

Figure 4. Multicollinearity test