Accident Analysis and Prevention 134 (2020) 105251

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

A data mining approach to investigate patterns of powered two-wheeler crashes in Spain

Alfonso Montella[a], Rocìo de Oña[b], Filomena Mauriello[a], Maria Rella Riccardi[a], Giuseppe Silvestro[a]

[a] University of Naples Federico II, Department of Civil, Architectural and Environmental Engineering, Via Claudio 21, 80125 Naples, Italy
[b] University of Granada, TRYSE Research Group, Department of Civil Engineering, Spain

ARTICLE INFO

Keywords:
Powered two-wheelers
Data mining
Classification trees
Rules discovery
Injury severity
Run-off-the-road crashes
Head-on crashes

ABSTRACT

Powered two-wheelers (PTWs) are growing globally each year as they are considered an attractive alternative to cars (flexible, small, affordable, fast and easy to park), especially on congested traffic situations. However, PTWs represent an important challenge for road safety. In fact, in 2016, Spain ranked fifth in terms of PTW fatalities among EU 28. For this reason, this paper aims to investigate which are the patterns among crash characteristics contributing to PTW crashes in Spain. Data from 78,611 crashes involving PTWs occurred in Spain in the period 2011–2013 were analyzed. The analysis was performed by using classification trees and rules discovery which are suitable models aimed at extracting knowledge and identifying valid and understandable patterns from large amounts of data previously unknown and indistinguishable. The response variables assessed in this study were severity and crash type. As a result, several combinations of road, environmental and drivers' characteristics associated with severity and typology of PTW crashes in Spain were identified. Based on the analysis results, several countermeasures to solve or mitigate the safety issues identified in the study were proposed.

From the methodological point of view, study results show that both the classification trees and the a priori algorithm were effective in providing non-trivial and unsuspected relations in the data. Classification trees structure allowed a simpler understanding of the phenomenon under study while association discovery provided new information which was previously hidden in the data. Given that the results of the two different techniques were never contradictory, we recommend using classification trees and association discovery as complementary approaches since their combination is effective in exploring data providing meaningful insights about PTW crash characteristics and their interdependencies.

1. Introduction

Powered two-wheelers (PTWs) are increasing rapidly in number in the last decade (ITF, 2015; IRTAD, 2018) and their use will continue to grow and evolve (Ivers et al., 2016). What they offer is an attractive form of transport for daily commuting, work or simply pleasure. Their use produces a double positive effect. The first is the opportunity to make better use of the existing road system. Therefore, it follows that switching from car to PTW is bound to increase network capacity. The second is the chance to reduce the environmental damage because the CO2 contribution of PTWs to overall transport is marginal (ACEM, 2010).

Despite these positive characteristics, PTWs represent an important challenge for road safety. This mode of transport accounts for more than 286,000 deaths each year globally – about 23% of all road traffic deaths (WHO, 2017). This alarming number of potentially avoidable deaths highlights the need for increased attention around PTWs and their use in road safety policy. In the EU28, 25,651 people lost their lives on EU roads in 2016 (European Commission, 2018a) and 4358 motorcyclists lost their life in traffic crashes, accounting for 17% of the road fatalities even if they accounted only for 2% of the passenger-kilometers in the EU roads (126 billion passenger-kilometers out of 5508 billion passenger-kilometers). Since the use of PTWs continues to grow globally each year, the result is that, in spite of a remarkable improvement in traffic safety for all road users, motorcyclists have seen their exposure to road risk increasing while the fatality of other road users declined significantly from 2001 to 2016 (global road fatalities reduction equal to 53% vs PTW fatalities reduction equal to 37%).

In Spain, the number of PTWs has increased in time and it is more than doubled in fifteen years - from 1,517,208 in 2002 to 3,211,474 in 2016 (DGT, 2018). PTW represents the main category of involved vehicle in a fatal crash after car, accounting for 22% of the Spanish total

E-mail address: alfonso.montella@unina.it (A. Montella).

https://doi.org/10.1016/j.aap.2019.07.027
Received 14 May 2019; Received in revised form 27 June 2019; Accepted 26 July 2019
Available online 09 August 2019
0001-4575/

traffic fatalities in the period 2000–2016. In 2016, Spain ranked fifth in terms of PTW fatalities among EU 28 (European Commission, 2018a).

In 2010, in agreement with the UN General Assembly, with the declared "Decade of Action for Road Safety" and with the goal to "stabilize and then reduce" the predicted increase in road traffic fatalities, the European Commission proposed to carry on with the target of halving the overall number of road fatalities by 2020. Since the effective target seems being too hard to be reached by 2020, a new road safety target was announced for the period 2020–2030 on 17 May 2018. The Members reconfirm 'Vision Zero' as long-term objective and base this framework on the 'Safe System' approach addressed in preventing death and serious injury (European Commission, 2018b). In the framework of these policy orientations, one of the priority actions considered by the Commission is the improvement of the safety of vulnerable road users, in particular motorcyclists for whom crash statistics are particularly worrying (European Commission, 2010). To increase safety of PTWs, the EU Member States have already put in place good practice. In the Netherlands, several steel barriers with motorcycle protection systems were installed (ERF, 2018). Buckinghamshire County Council positioned hazard marker posts on the outer edge of the bend to focus the rider's eyes on the vanishing point based on the principle that "where you look is where you go" and works on the basis that if you can hold the drivers' eye around a bend then they are likely to successfully negotiate it (IHE, 2018).

As PTWs are a major social concern, the Association of European Motorcycle Manufacturers with the support of the European Commission and other partners conducted the MAIDS study, an extensive in-depth investigation of motorcycle and moped crashes with the aim of identifying the causes and consequences of PTW crashes (ACEM, 2004). The major findings of this study are as follows: in 37% of cases the primary contributing factor was a human error of the PTW rider while in 50% of cases the primary contributing factor was a human error of the other vehicle (OV) driver; among the primary contributing factors, over 70% of the OV driver errors were due to the failure to perceive the PTW; roadside barriers presented an infrequent but substantial danger to PTW riders, causing serious lower extremity and spinal injuries as well as serious head injuries; road surfaces had defects in 30% of cases. Recently, the Federal Highway Administration developed the Motorcycle Crash Causation Study (MCCS) focusing on the unique circumstances that caused PTW crashes and providing a possible comprehensive data collection in order to better understand the nature of collision (Nazemetz et al., 2019). The major findings of this study are as follows: single-vehicle crashes and fatalities were overrepresented at night; curves were overrepresented for both single-vehicle and fatal crashes; motorcycle's inappropriate speed was a contributory factor in 29 percent of crashes and was overrepresented in fatal crashes. Overall, what these studies highlighted is the need of reliable data source in the development of proper countermeasures that will reduce the frequency and severity of PTW crashes. Furthermore, even if numerous studies investigated the factors influencing the risks of fatal crashes in general, the factors influencing the risks of PTW fatal crashes are still not clear. Moreover, the dynamics of motorcycles and vulnerability of their riders make the countermeasures implemented to reduce crash severity at global level not always effective in reducing the risks of serious crashes involving PTWs. Thus, an updated assessment of motorcycle crash risk factors is needed to understand why motorcycle crashes continue to occur and to determine how to improve motorcycle crash prevention.

Identifying factors that affect crash injury severity and understanding how these factors affect injury severity is critical in planning and implementing highway safety improvement programs. The number of studies dealing with crash severity prediction is increasing over time (Chang et al., 2016; Cunto and Ferreira, 2016; Mannering et al., 2016; Milton et al., 2008; Ye and Lord, 2014), highlighting how this issue is becoming important for the scientific community. Great emphasis is given to serious injuries crashes also at political level. The EU set the target of halving the number of serious injuries in the EU by 2030 from the 2020 baseline using a common definition based on the MAIS 3+ trauma scale.

Most studies used discrete outcome models treating injury severity as either a nominal or ordered variable (Savolainen et al., 2011). The most commonly used are the logit and probit models (Ye and Lord, 2014) which address the issue of unobserved heterogeneity (Mannering et al., 2016) by using random parameters including variations in the effect of variables across the sample population that are unavailable to the analyst (Chang et al., 2016; Cunto and Ferreira, 2016; Milton et al., 2008). Recently, some studies used data mining techniques, such as Kashani et al. (2014) and Montella et al. (2012), to detect interdependence as well as dissimilarities among crash characteristics and provide insights for the development of safety improvement strategies focused on PTWs. Although some studies have already been developed in Spain (Hidalgo-Fuentes and Sospedra-Baeza, 2018; Perez-Fuster et al., 2013), they did not carry out an in-depth research exploring the main underlying relationships between the contributing factors of PTW fatal or serious crashes. In fact, Hidalgo-Fuentes and Sospedra-Baeza (2018) analyzed the differences on the characteristics of motorcycle crashes between males and females, and Perez-Fuster et al. (2013) characterized offenders and non-offenders motorcyclists based on the motorcyclist characteristics and environmental factors. For this reason, literature review highlights the need to carry out more studies on PTW crash contributory factors and circumstances. To fill this research gap, this study investigated a large database of PTW crashes in Spain by data mining techniques. Specifically, the principal purpose of this study is to discover non-trivial and unsuspected patterns existing among the crash characteristics contributing to the severity of PTW crashes and the patterns contributing to the PTW crash type. Another aim of this paper is to investigate the importance of these contributing factors on predicting both dependent variables (PTW crash severity and PTW crash type). The reason for using data mining technique was related to the nature of the crash phenomenon. A crash can be defined as a rare, random, multi-factor event always preceded by a situation in which one or more road users fail to cope with the road environment (Montella, 2011). Moreover, data mining technique permits to find out non-trivial and unsuspected relations in the data, which is not possible with conventional statistical models. Thus, while collisions will continue to occur, deaths and serious injuries are largely preventable and the identification of PTW crash contributory factors and their interdependences by means of data mining techniques can provide useful insights for the development of effective countermeasures.

The remainder of the paper is organized as follows: Section 2 provides details of the statistical methodologies used in the analysis; Section 3 describes the crash data; Section 4 explains the results of the classification trees and of the association analysis methodologies; and finally, a discussion of the results and recommendations aimed at reducing motorcycle crashes and improving safety were provided.

2. Methodology

2.1. General aspects

Data mining tools are suitable models aimed at extracting knowledge and identifying valid and understandable patterns from large amounts of data previously unknown and indistinguishable. Data mining is focused on the search and finding of patterns in data rather than the confirmation of hypotheses (Das et al., 2019). The powerful aspect is that data mining does not need any assumptions and a priori probabilistic knowledge about the phenomena under studying, differently from conventional statistical and artificial intelligent models that have pre-defined underlying relationships between dependent and independent variable. In the study, two data mining techniques were used: (1) classification trees and (2) rules discovery. Classification tree has been largely employed in transportation (De Oña et al., 2015) and

road safety analysis (López et al., 2014; López and de Oña, 2017; Montella et al., 2011, 2012; Moral-García et al., 2019). In addition, the association discovery was implemented since its use in data mining has successfully identified sets of crash contributory factors highlighting obscured patterns or rules (Das et al., 2019; Pande and Abdel-Aty, 2009).

Due to the large number of patterns considered, classification trees and association rules suffer from an extreme risk of type I error, that is of finding patterns that appear due to chance alone (Webb, 2007). To overcome this problem, the study data were randomly split in two data sets: (1) a training sample (70% of the total dataset, 55,166 crashes) and (2) a test sample (30% of the total dataset, 23,445 crashes). The training sample was used to generate the classification tree structures (see Section 2.2) and all the potential association rules (see Section 2.3). The tree structure of the training tree was applied to the test sample and nodes with the same class in both the learned and the test trees were validated, removing the nodes with higher risk of type I error and obtaining the validated trees (see Section 2.2). Similarly, the test sample was used to test the statistical significance of the rules discovered with the training data by hypothesis testing and reduce the error rate (see Section 2.3). Finally, Receiver Operating Characteristics (ROC) analysis was performed to evaluate the diagnostic performance of the validated trees (see Section 2.4).

2.2. Classification trees

Tree-based methods are non-linear and non-parametric data mining tools for supervised classification and regression problems. A tree is an oriented graph (Fig. 1) where a root node (which contains all data) is divided on the basis of an independent variable (splitter) into a finite number of leaf nodes in order to create groups such that the samples within the same group are as homogenous (pure) as possible. To achieve this, a set of candidate split rules is created, which consists of all possible splits for all variables included in the analysis. These splits are then evaluated and ranked based on the Gini reduction criterion (de Oña et al., 2012, 2013). The terminal nodes present a low degree of impurity compared to the root node. Any segmentation methodology is characterized by the definition of the following steps: (1) the partitioning criterion to define the optimality function when choosing the best partition of the objects into homogeneous subgroups; (2) the stopping rule to arrest the growing procedure to build up the tree; and (3) the assignment rule to identify either a class or a value as label of each terminal node.

We focused on the framework of the CART algorithm introduced by Breiman et al. (1984). In our study, the heterogeneity at any node t is evaluated in terms of an impurity measure $i_Y(t)$ expressed by the Gini index, which is calculated as follows:

$$i_Y(t) = 1 - \sum_j p(j/t)^2 \tag{1}$$

where $P(j|t)$ is the proportion of observations in the node t that belong to the class j. If a node is 'pure', all the observations in the node belong to one class and the impurity (node) will be equal to zero.

The total impurity of any tree T is defined as follows:

$$i_Y(T) = \sum_{t \in \tilde{T}} i_Y(t) p(t) \tag{2}$$

where $i_Y(t)$ is the impurity of the node t, $p(t) = N(t)/N$ is the weight of the node t, N(t) is the number of observations falling in node t, N is the total number of observations, and $\tilde{T}$ is the set of terminal nodes of the tree T. The total impurity of the tree is reduced by finding at each node of the tree the best partition s* of the observations into disjoint classes such that it induces the highest decrease in the impurity of the response variable Y when passing from the node t to the children nodes t_l and t_r:

$$\max_{s \in S} \Delta i_Y(t, s) = \max_s \{ i_Y(t) - p_l i_Y(t_l) - p_r i_Y(t_r) \} \tag{3}$$

where s ∈ S includes the set of splits generated by all predictors, t_l and t_r are the children nodes of the node t, respectively the left node and the right node, and p_l and p_r are the proportions of observations in node t falling into the left and right node.

Tree growing stopped basing on two criteria: (1) the reduction in the Gini measures is less than a prespecified minimum fixed equal to 0.0001; and (2) the maximum size of the tree, choosing the maximum number of levels of the tree equal to 4.

In our study, variables of interests are unbalanced, especially as far as the variable severity where the most important class, i.e. serious and fatal injuries, represents a small proportion of the observations. Thus, it was introduced a posterior classification ratio (PCR) which compares the classification of the terminal nodes of the tree with the classification of the root node (López et al., 2014; Montella et al., 2012). The PCR was calculated as follows:

$$\mathrm{PCR}(j|t) = \frac{p(j|t)}{p(j|t_{root})} \tag{4}$$

where $p(j|t)$ is the proportion of observations in node t that belong to the class j and t_{root} is the root node of the tree.

The assignment of the class to each node was performed selecting the class j* with the greatest value of PCR:

$$j^* | t : \max_j PCR(j|t) \tag{5}$$

Once the PCR was assessed for the training tree, it was calculated again in the testing tree with the aim of validating the nodes with the same class.

The tree growing process was applied to the training sample focusing on severity and crash type. The result was two different tree structures, one for each response variable. Subsequently, the tree structure was applied to the test sample to obtain a test tree, one for each variable, used for validation and for evaluating the accuracy of the classifier. At each final node of the test tree, the class assignment was compared with the assignment performed in the learned tree. As a result, only nodes with the same class in both the learned and the test trees were validated.

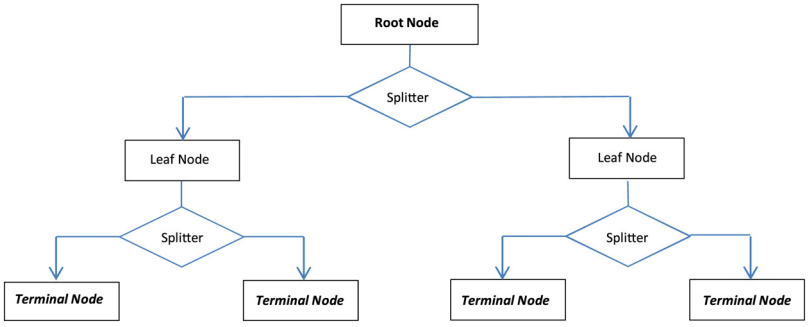


Fig. 1. An example of classification tree.

For the validated trees, the Variable Importance Measure (VIM) has been calculated. VIM is a predictor ranking based on the contribution of the predictors in building up the tree. The importance of a variable xj is defined by the following equation (Kashani et al., 2014):

VIM (xj) = $\sum_{t=1}^{T} \frac{n_t}{N}$ ΔGini (xj, t) (6) where ΔGini (xj, t) is the Gini reduction at a node t that is achieved by splitting by the variable xj; n_t /N is the proportion of the observations in the dataset that belong to node t; T is the total number of nodes and N is the total number of observations. Dividing by the highest value obtained for a variable, it is possible to provide the normalized importance of each variable. Data were analyzed by the SPSS software.

2.3. Rules discovery

2.3.1. Association rules

Association rules discovery is a methodology for identifying potential interaction among a large number of crash-related factors that can potentially discover relationships between crash attributes and the variable in assessment that are not well known from current research works (Li et al., 2018). The association rules analysis may be considered as a process of looking through all possible multidimensional contingency tables and extracting the most interesting conclusions (Pande and Abdel-Aty, 2009). Basing on the relative frequency of the number of times the sets of items occur alone and in combination in a dataset, the association rules were extracted with the form "A → B", where A and B are disjoint item-sets: A is the antecedent and B is the consequent.

Associations discovery was performed using the a priori algorithm according to the methodology introduced by Agrawal et al. (1993). The a priori algorithm uses simple and repetitive steps examining candidate item-sets to find frequent item-sets. Then, it uses the new candidate item-sets produced using frequent item-sets to find new frequent item-sets until no newer item-sets can be produced. The strength of the association rule can be measured in terms of the values of support, confidence, and lift (Das et al., 2019), where support is the percentage of the entire data set covered by the rule, confidence measures the reliability of the inference of a generated rule, and lift is a measure of the statistical interdependence of the rule.

Supports are calculated as follows:

$$\text{Support (A} \rightarrow \text{B)} = \frac{\#(A \cap B)}{N}\text{; Support (A)} = \frac{\#(A)}{N}\text{; Support (B)} = \frac{\#(B)}{N} \quad (7)$$

where support(A → B) is the support of the rule, support(A) is the support of the antecedent, support(B) is the support of the consequent, #(A ∩ B) is the number of crashes where both the condition A (antecedent) and the condition B (consequent) occur, #(A) is the number of crashes with A antecedent, #(B) is the number of crashes with B consequent, and N is the total number of crashes in the dataset.

Confidence is calculated as follows:

$$\text{Confidence} = \frac{\text{Support (A} \rightarrow \text{B)}}{\text{Support (A)}} \quad (8)$$

Lift is calculated as follows:

$$\text{Lift} = \frac{\text{Support (A} \rightarrow \text{B)}}{\text{Support (A)} \times \text{Support (B)}} \quad (9)$$

The lift of the rule relates the frequency of co-occurrence of the antecedent and the consequent to the expected frequency of co-occurrence under the assumption of conditional independence. A lift value lower than 1 indicates negative interdependence between the antecedent and the consequent. A lift value equal to 1 designates independence, and a value greater than 1 indicates positive interdependence (i.e., the number of times the sets of items occur together is greater than they would if they were independent of each other). The higher the lift, the greater the strength and the interest of the association rule. It is desirable for the rules to have a high level of support, a large confidence, and a lift value considerably greater than one. Thus, minimum values for support, confidence and lift are needed. Then, each rule with n+1 items is validated by verifying that each variable produced a lift increase (LIC). The LIC ensures that each additional item in the rules lead to an increase in term of lift (López et al., 2014). The rules with only one item in the antecedent are used as a starting point, rules with more items are selected over simpler rules if the LIC condition satisfied the minimum threshold of 1.05 (López et al., 2014; Montella et al., 2011, 2012). LIC is calculated as follows:

$$\text{LIC} = \frac{\text{Lift}_{A_n}}{\text{Lift}_{A_{n-1}}} \quad (10)$$

where A_n is the antecedent of rule with n items, and A_{n+1} is the antecedent of rule with n + 1 items.

Consistently with previous studies previous studies (Li et al., 2018; López et al., 2014), the threshold values for support (S), confidence (C), and lift (L) were set as follows: S ≥ 0.1%, C ≥ 1.0%, and L ≥ 1.20. This study used open source software R and R package 'arules' to conduct the analysis (Hahsler et al., 2018).

The first set of rules was generated by the training sample. The test sample was used to test the statistical significance of the rules discovered with the training data. We used the binomial test to verify the statistical significance of deviations of the rule support measure from the theoretically expected value when antecedent and consequent items are independent, using the rules obtained with the training data and applying the binomial test to the test data. We set the significance level α equal to 0.05, that means there is a 5% chance that a spurious rule passes the significance test.

2.3.2. Classification trees

To integrate the results of the classification trees and the association discovery, the results of the classification trees were converted into rules. All the splits of the parent nodes are the antecedent of the rule whereas the class of the terminal node is the consequent. Support, confidence and lift (equal to the posterior classification ratio) were calculated for each terminal node t. The threshold values for support (S), confidence (C), lift (L), and lift increase (LIC) were set equal to the association rules thresholds: S ≥ 0.1%, C ≥ 1.0%, L ≥ 1.20, and LIC ≥ 1.05.

2.4. Statistical evaluation

Receiver Operating Characteristics (ROC) analysis was performed to evaluate the diagnostic performance of the validated trees on the basis of its classification accuracy in the test sample. It is a widely used graphical plot that illustrates the ability of a classifier created by plotting the true positive rate (TPR) on the vertical axis against the false positive rate (FPR) on the horizontal axis at various threshold settings (Fig. 2).

The true positive rate measures the fraction of positive that are correctly identified and is also known as sensitivity while the false positive rate measures the fraction of negative cases that are mistakenly categorized as positives and is also evaluated as 1 - specificity. Since there is a trade-off between the sensitivity and the specificity, if the sensitivity (i.e., the correct positives ratio) increases, the specificity (i.e., the correct negatives ratio) will decrease, and vice versa.

Sensitivity and Specificity were assessed as follows:

$$\text{TPR} = \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (11)$$

$$\text{FPR} = 1 - \text{Specificity} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (12)$$

where the True Positives (TP) refer to the number of Predicted Positive cases that were correctly classified, False Positives (FP) refer to the

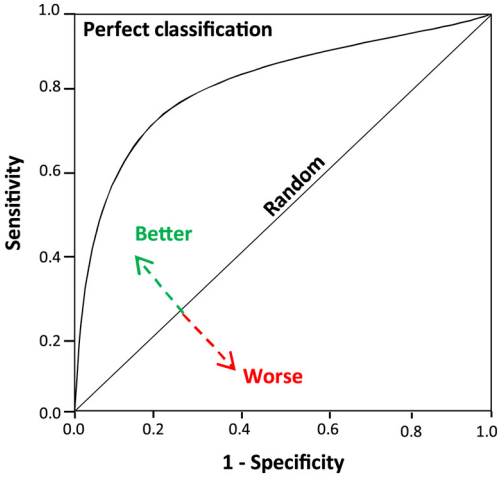


Fig. 2. An example of ROC curve.

number of Predicted Positive cases that were erroneously classified (the outcome is incorrectly predicted as positive when it is actually negative), True Negatives (TN) refer to the number of Predicted Negative cases that were correctly classified and False Negatives (FN) refer to the number of Predicted Negative cases that were erroneously classified (the outcome is incorrectly predicted as negative when it is actually positive). TP + FN refers to the total number of positives equals to the number of correct (true) positives plus the number of false negatives whereas FP + TN refers to the total number of negatives equals to the number of correct negatives plus the number of false positives.

When ROC curve is created, the Area Under the Curve (AUC) could finally show the classification accuracy of the model. An AUC value varies between 0 and 1. It can be inferred that if the AUC is ≥ 0.90, the model is considered to have outstanding discrimination, if the AUC is ≥ 0.80 and < 0.90, the model is considered to have excellent discrimination and finally, if the AUC is ≥ 0.60 and < 0.80, then the model would be considered to have acceptable discrimination (Kashani et al., 2014). An AUC equal to 0.50 represents a random prediction and an AUC value < 0.50 is considered as poor performance.

3. Data

Crash data were provided by the Spanish General Traffic Crash Directorate (DGT) for the Spanish road network, relative to the three-years period 2011–2013. The data were collected into three different subsets including: (1) a dataset containing 35 different fields describing general and specific characteristics of the crash focusing the information on road safety elements, on crash contributory factors, and on general characteristics such as weather, crash time, and crash type; (2) a dataset containing 9 variables describing the involved vehicles focusing the information on vehicle type and its age and condition; and (3) a dataset containing 19 variables describing people (driver, passenger or pedestrian) involved in the crash. A detailed description of the DGT data can be found in https://sedeapl.dgt.gob.es. Finally, the three different data sets were combined in one data set.

The original data set consisted of 255,661 crashes. Only 78,611 crashes where at least one s was involved (31% of the total crashes) were extracted from the data set and were used for the subsequent analyses.

As shown in Table 1, the data set was rearranged and 19 categorical variables were selected: (1) area, (2) road type, (3) lighting, (4) weather, (5) pavement, (6) driver PTW gender, (7) driver PTW age, (8) driver PTW outcome, (9) Vehicle B driver gender, (10) Vehicle B driver age, (11) Vehicle B driver outcome, (12) pedestrian gender, (13) pedestrian age, (14) pedestrian outcome, (15) alignment, (16) involved vehicles, (17) PTW type, (18) crash type, and (19) severity.

The crash severity classification was based on the most severe injury to any person involved in the crash. According to the agreed international definition, a traffic crash fatality was every single person that dies in the crash or within the 30 days following it. As far as non-fatal injuries, injuries were classified as severe or slight. Severe injury represents any injured person whose condition required hospitalization for more than 24 h; and slight injury was any injured person who did not meet the severe injury definition. Thus, the crash severities were as follows: fatal (n = 4,459; 5.7% of the total crashes), serious (n = 7,318; 9.3% of the total crashes), and slight (n = 66,834; 85.0% of the total crashes). Because the different categories of the variable severity were not balanced and this issue affected the performance of the model, the variable severity was transformed in two categories: SI = crashes producing slight injuries (n = 66,834; 85.0% of the total crashes) and killed or seriously injured (KSI) = crashes producing fatalities or serious injuries (n = 11,777; 15.0% of the total crashes).

Descriptive statistics (Table 1) show some categories with higher crash severities, such as rural area, nighttime, foggy weather, male PTW drivers, young and old PTW drivers, old pedestrians hit by PTWs, curve geometry, run-off-the-road and head-on crashes.

4. Results

The response variables assessed in this study were severity and crash type. All dependent variables have been treated as nominal. Severity is defined as the level of injury sustained by the most severely injured person involved in the crash, crash type describes the type of collision and involved vehicles identifies the vehicles involved in the crash. The classification trees were developed in sequential order, removing the variable severity in the second tree. In the association rules analysis, the rules were first ordered by the consequent and then ordered by the decreasing value of the lift. For each two-items rule, the three items rules having the same antecedent of the parent rule were ordered again by the decreasing value of the lift, and so on.

Classification trees produced 9 validated rules. The application of the a priori algorithm identified 207 significant rules. For each response variable, the value of the area under the curve was calculated (Table 2) to assess the accuracy of the tree model. Most responses are classified with AUC greater than 0.60, showing that the model has acceptable classification accuracy in most cases.

4.1. Severity

Severity was classified in two categories: KSI which includes fatal and serious injury crashes (15.0%) and SI which includes slight injury crashes (85.0%). The classification tree (Fig. 3) produced 14 validated terminal nodes but only 4 nodes (rules T1_11, T1_14, T1_17, T1_19 in Table 3) satisfied the LIC criterion (Eq. (10)), identifying as predictors the variables road type, area, alignment, crash type, involved vehicles, pavement, weather, and lighting. The tree method produced an AUC equal to 0.66 for KSI and 0.67 for SI. Fig. 4 shows the normalized importance of the variables in the tree model. Seven variables were identified as influencing the classification accuracy (Fig. 4). Road type, area and alignment had the greatest normalized importance and were identified by the a priori algorithm as antecedents in 40 out of the 44 significant rules (Table 3).

The primary split of the classification tree was the road type. The two nodes identified by this split have a substantial difference in the proportion of KSI crashes: 30.9% in node 2 vs. 11.4% in node 1. Road types in node 2, i.e. road types with the highest crash severity, are motorways, rural autonomous, rural national, rural provincial, urban autonomous, urban national, and urban provincial. The second split was the alignment, with crashes in the curves showing the greatest proportion of crash severity (37.7% in node 5 vs. 28.2% in node 6). The third split was the crash type and identified the node with the greatest crash severity. This is the node 11, with 55.3% of KSI crashes (posterior

Table 1
Descriptive statistics of crash data.

Variable	Code	Count	Percent	KSI (%)	Variable	Code	Count	Percent	KSI (%)
Area					Vehicle B driver gender				
Urban	U	54,836	69.8	11.8	Male	Ma	34,339	60.7	13.9
Rural	R	23,775	30.2	22.3	Female	Fe	10,127	17.9	11.9
Total		78,611	100.0	15.0	Missing	Missing	12,144	21.5	13.5
Road Type					Vehicle B driver outcome				
Urban Municipal	Um	50,652	64.4	10.6	Uninjured	Unj	27,635	48.8	16.1
Urban Provincial	Up	968	1.2	31.3	Slight Injured	Sli	3,923	6.9	12.6
Urban Autonomous	Uac	1,145	1.5	34.8	Serious Injured	Sei	148	0.3	100.0
Urban National	Un	478	0.6	31.2	Dead	Dd	97	0.2	100.0
Rural Municipal	Rm	8,061	10.3	12.2	Missing	Missing	24,807	43.8	9.9
Rural Provincial	Rp	4,123	5.2	29.5	Pedestrian gender				
Rural Autonomous	Rac	5,096	6.5	29.1	Male	Ma	240	4.7	10.4
Rural National	Rn	2,153	2.7	32.0	Female	Fe	201	3.9	11.4
Motorway	Mw	5,424	6.9	17.3	Missing	Missing	4,667	91.4	20.3
Other	Ot	511	0.7	41.1	Pedestrian age				
Lighting					0–18	0–18	75	1.5	9.3
Day	Dy	60,624	77.1	14.7	19–25	19–25	43	0.8	7.0
Night	Nt	17,987	22.9	16.0	26– 45	26– 45	101	2.0	5.0
Weather					46– 65	46– 65	76	1.5	13.2
Clear	Cl	71,341	90.8	15.2	> 65	> 65	100	2.0	20.0
Rainy	Rn	5,616	7.1	12.1	Missing	Missing	4,713	92.3	20.1
Foggy	Fg	231	0.3	18.2	Pedestrian outcome				
Snow	Sw	94	0.1	13.8	Uninjured	Unj	37	0.7	0.0
Other	Ot	1,329	1.7	17.6	Slight Injured	Sli	367	7.2	1.9
Pavement					Serious Injured	Sei	5	0.1	100.0
Dry	D	68,714	87.4	15.2	Dead	Dd	37	0.7	100.0
Wet	W	7,467	9.5	13.1	Missing	Missing	4,662	91.3	20.3
Slippery	Sl	870	1.1	15.6	Alignment				
Frozen	Fr	107	0.1	15.9	Tangent	Tan	34,842	44.3	13.4
Snowy	Sw	30	0.0	10.0	Curve	Cu	8,944	11.4	27.3
Other	Ot	1,423	1.8	12.6	Intersection	Int	33,545	42.7	13.4
PTW driver gender					Other	Ot	1,280	1.6	13.0
Male	Ma	52,893	67.3	16.6	Involved Vehicles				
Female	Fe	10,451	13.3	8.6	PTW-Car	Car	45,936	58.4	13.1
Missing	Missing	15,267	19.4	13.8	PTW Single Vehicle	SV	19,960	25.4	18.7
PTW driver age					PTW - Pedestrian	Ped	2,041	2.6	19.9
0–18	0–18	3,629	4.6	17.0	PTW-PTW	PTW	2,754	3.5	12.2
19–25	19–25	9,948	12.7	12.1	PTW-Bike	Bike	871	1.1	11.1
26–45	26–45	33,409	42.5	14.8	PTW-Truck	Truck	6,646	8.5	17.0
46–65	46–65	13,515	17.2	18.0	PTW-Other	Ot	403	0.5	1.1
> 65	> 65	1,364	1.7	24.0	PTW type				
Missing	Missing	16,746	21.3	13.4	Motorcycle	Motorcycle	55,835	71.0	15.6
PTW driver outcome					Moped	Moped	22,776	29.0	13.5
Uninjured	Unj	2,703	3.4	10.6	Crash Type				
Slight Injured	Sli	51,238	65.2	2.8	Angle	An	28,992	36.9	13.5
Serious Injured	Sei	4,863	6.2	100.0	Falling from the Vehicle	FfV	5,557	7.1	15.6
Dead	Dd	3,006	3.8	100.0	Head-On	HO	2,114	2.7	25.6
Missing	Missing	16,801	21.4	13.1	Hit obstacle	Hobs	1,736	2.2	15.0
Vehicle B driver age					Hit parked Vehicle	HpV	443	0.6	20.3
0–18	0–18	899	1.6	15.4	Hit Pedestrian	Ped	5,108	6.5	19.5
19–25	19–25	5,797	10.2	12.6	Rear-End	RE	16,663	21.2	10.6
26–45	26–45	22,100	39.0	12.9	Run-Off-the-Road	ROR	10,898	13.9	21.1
46–65	46–65	12,077	21.3	14.7	Other	Ot	7,100	9.0	14.7
> 65	> 65	2,470	4.4	16.4	Severity				
Missing	Missing	13,267	23.4	13.1	Killed or Seriously Injured	KSI	11,777	15.0	100.0
					Slight Injured	SI	66,834	85.0	0.0

classification ratio equal to 3.68). This node is characterized by head-on and hit-pedestrian crashes occurring in curve on motorways, rural autonomous, rural national, rural provincial, urban autonomous, urban national, and urban provincial roads. Similarly, head-on, hit-pedestrian and hit-parked vehicles crashes in tangents and intersections occurring on the same road types (node 14) have very high crash severity, i.e. 43.5% of KSI crashes. Road types in node 1, i.e. road types with the lowest crash severity, are rural municipal and urban municipal roads. In these road types, two significant nodes were identified: node 19 (KSI = 32.3%) and node 17 (KSI = 19.0%). Node 19 includes head-on, hit obstacle, hit parked vehicle, hit pedestrian and run-off-the-road (ROR) crashes occurring on curves and involving PTW-single vehicle and PTW-bike. Node 17 includes angle, falling from the vehicle, and rear-end crashes occurring on curves and involving PTW-single vehicle and PTW-truck.

The a priori algorithm identified several rules which added new knowledge (Table 3). In detail, the algorithm identified ten 2-item rules, twenty-one 3-item rules and thirteen 4-item rules. Two-item rules were related to road type (urban provincial, rural national, urban national, rural provincial, and rural autonomous), alignment (curve), crash type (head-on and ROR), PTW driver age (old drivers, > 65), and area (rural). Interestingly, three-item and four-item rules identified several combination of crash characteristics associated with strong increase in the proportion of KSI crashes. As expected, hit-pedestrian crashes in rural area were associated with high proportion of KSI crashes (rule 33, confidence = 43%, lift = 2.88). Rural autonomous roads, head-on crashes, and male PTW drivers were associated with the greatest crash severity (rule 15, confidence = 60%, lift = 4.02). Both

Table 2
AUC values for each response variable.

Variable	AUC
Severity	
KSI	0.66
SI	0.67
Crash type	
Run-off-the-road	0.77
Hit Pedestrian	0.74
Angle	0.68
Falling from the vehicle	0.68
Head-on	0.66
Hit obstacle in carriageway	0.64
Rear-end	0.62
Hit parked vehicle	0.58

ROR and head-on crashes in the curves were associated with high crash severity (rule 19, confidence = 37%, lift = 2.49; rule 28, confidence = 31%, lift = 2.05). Male PTW drivers further increased severity of ROR crashes in the curves (rule 29, confidence = 36%, lift = 2.38; LIC = 1.16). Overall, seventeen rules included crashes occurring in the curves, with nine rules including both curve crashes and rural road types or rural area. PTW driver age in the range 46–65, i.e. old middle-aged drivers, increased the severity of curve crashes (rule 20, confidence = 35%, lift = 2.33; LIC = 1.32).

4.2. Crash type

Crash type was classified in nine categories, with the largest being angle (36.9%), rear-end (21.3%) and run-off-the-road (13.8%). The classification tree (Fig. 5) produced 5 terminal nodes (rules T2_7, T2_9, T2_11, T2_14, T2_17) which satisfied the LIC criterion (eq. 10). Seven variables were identified as influencing the classification accuracy (Fig. 6); alignment, road type, area, pavement, and, with a small importance value, weather, PTW type, and lighting. The a priori algorithm identified further 173 rules (Table 4, Table 5, Table 6, and Table 7).

The first split was in relation to the area. In rural area (node 1), the crash type with the greatest posterior classification ratio (Eq. (4)) was the run-off-the-road (31.8% of the total crashes). ROR proportion further increased in single vehicle crashes, with pavement condition different from slippery, occurring on curves (node 17, ROR = 55.8%). In single vehicle crashes on slippery pavement, the crash type with the greatest posterior classification ratio was the falling from the vehicle (node 7, FfV = 48.3%). In multi vehicle crashes on rural autonomous, rural municipal, rural national, and rural provincial roads, the crash type with the greatest posterior classification ratio was head-on (node 9, HO = 6.4%). In urban area, the two significant terminal nodes (node 11 and node 14) exhibited as class value hit pedestrian (node 11, Ped = 14.8%, PCR = 2.33) and angle crash (node 14, An = 55.8%, PCR = 1.51). Hit pedestrian crashes were associated with single vehicle crashes on dry pavement whereas head-on crashes were associated with multi vehicle crashes at intersections.

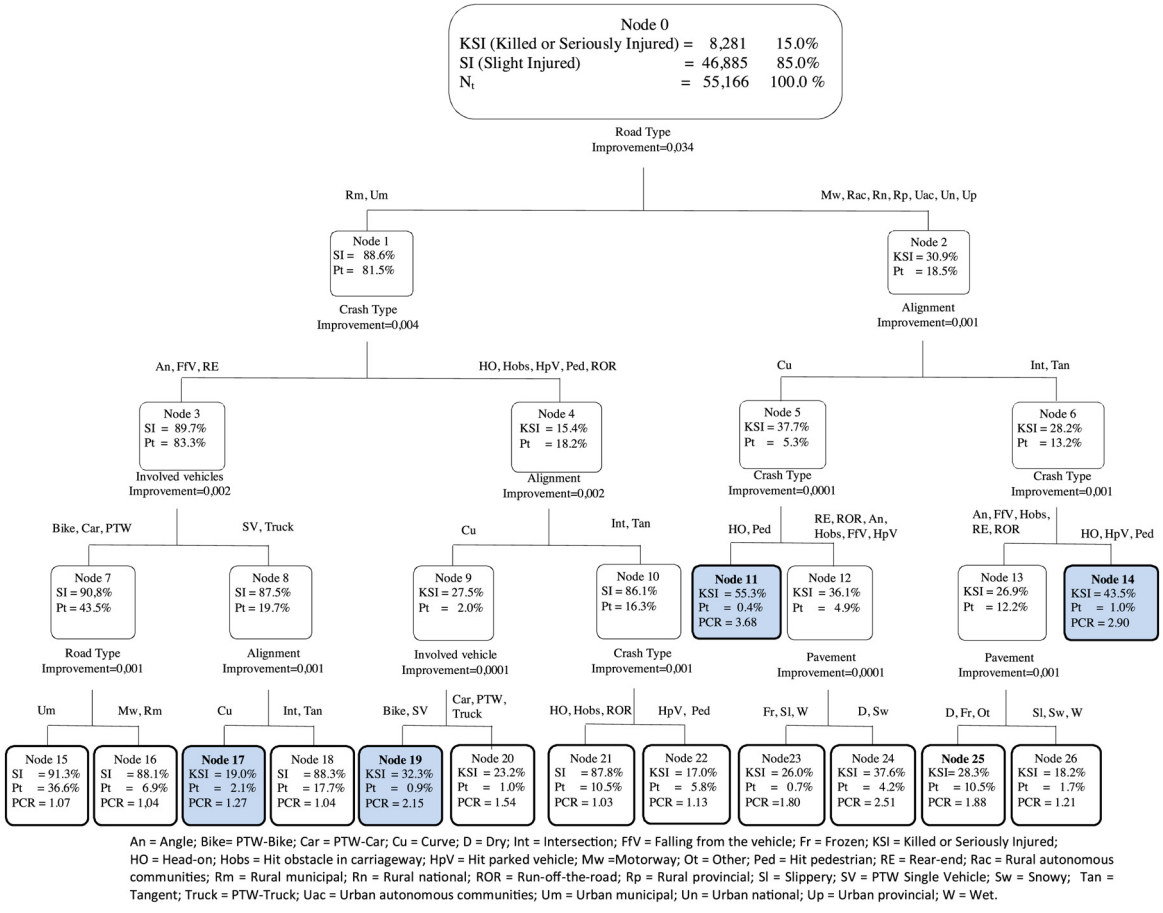


Fig. 3. Classification tree with KSI severity as consequent.
An = Angle; Bike = PTW-Bike; Car = PTW-Car; Cu = Curve; D = Dry; Int = Intersection; FfV = Falling from the vehicle; Fr = Frozen; KSI = Killed or Seriously Injured; HO = Head-on; Hobs = Hit obstacle in carriageway; HpV = Hit parked vehicle; Mw = Motorway; Ot = Other; Ped = Hit pedestrian; RE = Rear-end; Rac = Rural autonomous communities; Rm = Rural municipal; Rn = Rural national; ROR = Run-off-the-road; Rp = Rural provincial; Sl = Slippery; SV = PTW Single Vehicle; Sw = Snowy; Tan = Tangent; Truck = PTW-Truck; Uac = Urban autonomous communities; Um = Urban municipal; Un = Urban national; Up = Urban provincial; W = Wet.

Table 3
Rules with KSI severity as consequent.

Rule ID	Association rule Antecedent	Consequent	S %	C %	Lift	LIC
1	Road Type = Up	KSI	0.4	32.1	2.15	n.a.
2	Road Type = Up & Alignment = Cu	KSI	0.2	39.3	2.63	1.22
3	Road Type = Rn	KSI	0.8	31.3	2.10	n.a.
4	Road Type = Rn & Alignment = Cu	KSI	0.2	37.7	2.53	1.20
5	Road Type = Rn & Alignment = Cu & PTW driver age = 26-45	KSI	0.1	47.9	3.21	1.27
6	Road Type = Rn & Alignment = Cu & PTW driver gender = Ma	KSI	0.2	45.1	3.02	1.19
7	Road Type = Un	KSI	0.2	29.8	2.00	n.a.
8	Road Type = Rp	KSI	1.5	29.1	1.95	n.a.
9	Road Type = Rp & PTW driver gender = Ma	KSI	1.2	33.4	2.24	1.15
10	Road Type = Rp & PTW driver gender = Ma & Involved vehicles = SV	KSI	0.5	35.8	2.40	1.07
11	Road Type = Rp & Alignment = Cu	KSI	0.5	33.4	2.24	1.15
12	Road Type = Rp & Crash Type = HO	KSI	0.1	37.8	2.54	1.30
13	Road Type = Rac	KSI	1.9	28.9	1.94	n.a.
14	Road Type = Rac & Crash Type = HO	KSI	0.1	51.5	3.45	1.78
15	Road Type = Rac & Crash Type = HO & PTW driver gender = Ma	KSI	0.1	60.0	4.02	1.17
16	Road Type = Rac & PTW driver gender = Ma	KSI	1.3	36.1	2.42	1.25
17	Road Type = Rac & Alignment = Cu	KSI	0.6	31.4	2.11	1.09
18	Alignment = Cu	KSI	3.0	26.4	1.77	n.a.
19	Alignment = Cu & Crash Type = HO	KSI	0.2	37.1	2.49	1.41
20	Alignment = Cu & PTW driver age = 46-65	KSI	0.6	34.7	2.33	1.32
21	Alignment = Cu & PTW driver age = 46-65 & Involved vehicles = SV	KSI	0.4	36.9	2.47	1.06
22	Alignment = Cu & PTW driver gender = Ma	KSI	2.3	32.4	2.17	1.23
23	Alignment = Cu & Area = R	KSI	1.8	31.4	2.11	1.19
24	Alignment = Cu & Area = R & PTW driver age = 46-65	KSI	0.3	37.9	2.54	1.20
25	Alignment = Cu & Area = R & PTW driver gender = Ma	KSI	1.4	36.3	2.43	1.15
26	Alignment = Cu & Area = R & PTW driver age = 26-45	KSI	0.8	36.1	2.42	1.15
27	Alignment = Cu & PTW driver age = 26-45	KSI	1.3	31.1	2.08	1.18
28	Alignment = Cu & Crash Type = ROR	KSI	0.9	30.6	2.05	1.16
29	Alignment = Cu & Crash Type = ROR & PTW driver gender = Ma	KSI	0.8	35.5	2.38	1.16
30	Crash Type = HO	KSI	0.6	25.0	1.68	n.a.
31	PTW driver age = > 65	KSI	0.4	22.7	1.52	n.a.
32	Area = R	KSI	6.6	22.1	1.48	n.a.
33	Area = R & Crash Type = Ped	KSI	0.2	43.0	2.88	1.95
34	Area = R & Crash Type = HO	KSI	0.4	38.1	2.56	1.73
35	Area = R & PTW driver age = 46-65	KSI	1.3	25.9	1.73	1.17
36	Area = R & PTW driver age = 46-65 & Involved vehicles = Truck	KSI	0.2	34.7	2.33	1.35
37	Area = R & Involved vehicles = Truck	KSI	0.7	25.2	1.69	1.14
38	Area = R & Involved vehicles = Truck & PTW driver gender = Ma	KSI	0.5	30.5	2.05	1.21
39	Area = R & PTW driver gender = Ma	KSI	4.7	24.6	1.65	1.11
40	Area = R & PTW driver gender = Ma & Involved vehicles = SV	KSI	1.9	26.3	1.77	1.07
41	Area = R & Involved vehicles = SV	KSI	2.5	23.7	1.59	1.07
42	Area = R & Involved vehicles = SV & PTW driver age = 26-45	KSI	1.1	26.3	1.76	1.11
43	Crash Type = ROR	KSI	2.9	20.7	1.39	n.a.
44	Crash Type = ROR & Involved vehicles = SV	KSI	1.5	25.0	1.68	1.21
T1_11	Road Type = Mw/Rac/Rn/Rp/Uac/Un/Up & Alignment = Cu & Crash Type = HO/Ped	KSI	0.2	45.6	3.06	1.47
T1_14	Road Type = Mw/Rac/Rn/Rp/Uac/Un/Up & Alignment = Tan/ Int & Crash Type = HO/Ped/HpV	KSI	0.4	39.9	2.68	1.54
T1_17	Road Type = Um / Rm & Crash Type = RE Angle / FfV/Ot & Involved vehicles = SV / Truck / Ot & Alignment = Cu	KSI	0.4	18.2	1.22	1.52
T1_19	Road Type = Um/Rm & Crash Type = ROR/HO/Hobs/Ped/HpV & Alignment = Cu & Involved vehicles = SV/Bike	KSI	0.3	27.9	1.87	1.17

α_{crit} = 0.05; Cu = Curve; D = Dry; HO = Head-on; Ma = Male; Nt = Night; R = Rural; Rac = Rural Autonomous Communities; Rn = Rural National; ROR = Run off the road; Rp = Rural provincial; SV = PTW single vehicle; Tan = Tangent; Truck = PTW-truck; Ped = PTW-Pedestrian; Un = Urban National; Up = Urban Provincial.

Moreover, the a priori algorithm identified several rules showing a significant effect on the crash type of different crash characteristics and their combination. The most critical variables were alignment, area, involved vehicle, pavement, and road type.

As far as falling from the vehicle crashes (Table 4), consistently with the classification tree results (node 7) most rules involved slippery pavement, rural area and involvement of a single vehicle. However, the a priori algorithm identified other important features such as wet pavement, rainy weather, old PTW driver (> 65), very young PTW driver (0–18), and moped PTW type. The lift of the 2-items rule with slippery pavement as antecedent (rule 45) is 3.79, thus showing that the association between slippery pavement and falling from the vehicle is very strong. Slippery pavement in rural area exhibited a lift value equal to 4.11 (rule 47). A smaller, albeit significant, association was with the wet pavement (rule 70, lift = 1.60). Gender was involved in 14 rules, with 9 rules involving males and 5 rules involving females. The rule with the highest lift involving males was the rule 62 (lift = 3.54). In this rule the antecedent consisted of male driver, single vehicle, rain, and daytime. The rule with the highest lift involving females was the rule 67 (lift = 3.42). In this rule the antecedent consisted of female driver, single vehicle, and moped PTW type. Driver age was involved in 12 rules, with 2 rules involving very young drivers (0–18), 2 rules involving young middle-aged drivers (26–45), 7 rules involving old middle-aged drivers (46–65), and 1 rule involving older drivers (> 65). Rules with very young drivers involve single vehicle and daytime (lift = 2.69). Rules with young middle-aged drivers involve rural area and single vehicle (lift = 2.66). Rules with old middle-aged drivers involve rural area/rural municipal roads and single vehicle (lift = 3.04/3.87) or wet pavement (lift = 2.53). The rule with older drivers is a 2-item rule (only the age as antecedent) with lift equal to 1.56.

As far as run-off-the-road crashes (Table 5), consistently with the classification tree results (node 17) most rules involved rural area (e.g., rule 110, lift = 2.34) or road types in rural area (e.g., rule 89,

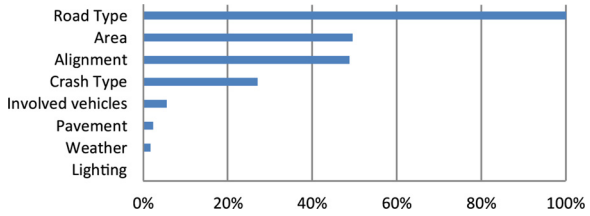


Fig. 4. Importance of the variables for severity.
An = Angle; Bike = PTW-bicycle; Car = PTW-car; Cu = Curve; D = Dry; FfV = Falling from the vehicle; Fr = Frozen; HO = Head-on; Hobs = Hit obstacle in carriageway; HpV = Hit parked vehicle; Int = Intersection; Mw = Motorway; Ot = Other; Ped = Hit pedestrian; PTW = PTW-PTW; R = Rural; Rac = Rural autonomous communities; RE = Rear-end; Rm = Rural municipal; Rn = Rural national; ROR = Run-off-the-road; Rp = Rural provincial; Sl = Slippery; SV = PTW Single Vehicle; Sw = Snowy; Tan = Tangent; Truck = PTW-truck; U = Urban; Uac = Urban autonomous communities; Um = Urban municipal; Un = Urban national; Up = Urban provincial; W = Wet.

lift = 2.89), curve alignment (e.g., rule 119, lift = 1.94), and poor pavement conditions, such as slippery pavement (e.g., rule 128, lift = 1.83) and wet pavement (e.g., rule 134, lift = 1.66). Young middle-aged drivers, combined with curve alignment and rural area/rural national roads/rural provincial roads (rules 113/108/102, lift = 3.67/4.04/3.45) were associated with ROR crashes. Nighttime, combined with rural area and wet pavement, was also associated with ROR crashes (rule 118, lift = 3.18). Noteworthy, male drivers and single vehicle were associated with ROR crashes (rule 139, lift = 1.75).

As far as head-on crashes (Table 6), most rules involved rural area (e.g., rule 144, lift = 1.48) or road types in rural area (rural autonomous communities, rural national, and rural provincial), curve alignment (e.g., rule 150, lift = 1.45), car involvement (rule 147, lift = 3.60), and male PTW drivers (rule 148, lift = 5.78).

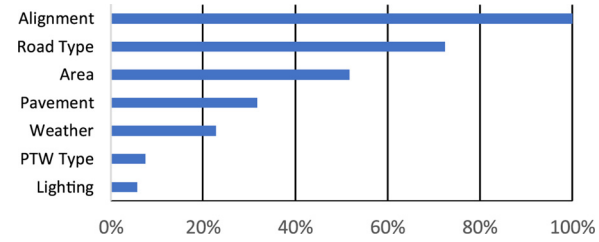
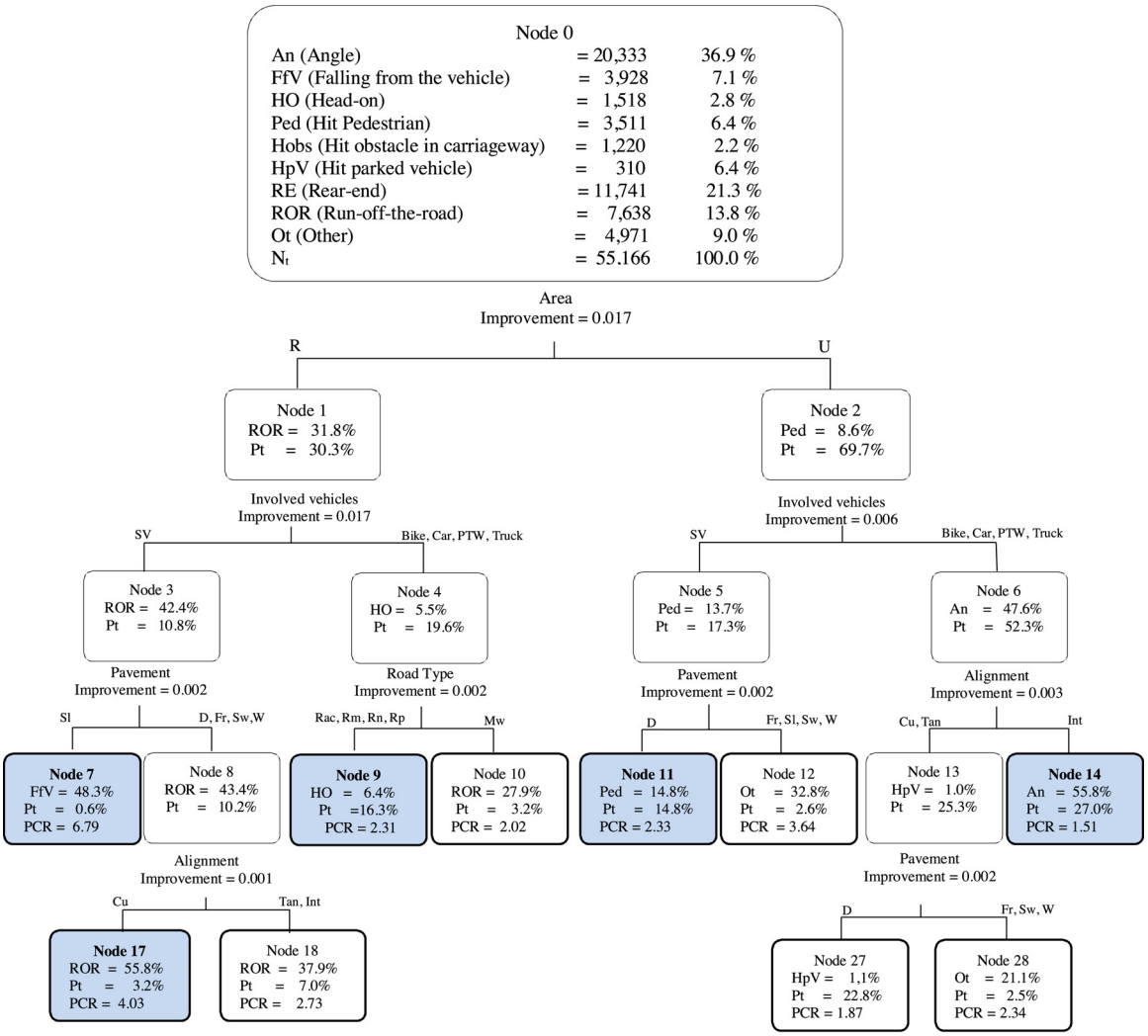


Fig. 6. Importance of the variables for crash type.



An = Angle; Bike = PTW-bicycle; Car = PTW-car; Cu = Curve; D = Dry; FfV = Falling from the vehicle; Fr = Frozen; HO = Head-on; Hobs = Hit obstacle in carriageway; HpV = Hit parked vehicle; Int = Intersection; Mw = Motorway; Ot = Other; Ped = Hit pedestrian; PTW = PTW-PTW; R = Rural; Rac = Rural autonomous communities; RE = Rear-end; Rm = Rural municipal; Rn = Rural national; ROR = Run-off-the-road; Rp = Rural provincial; Sl = Slippery; SV = PTW Single Vehicle; Sw = Snowy; Tan = Tangent; Truck = PTW-truck; U = Urban; Uac = Urban autonomous communities; Um = Urban municipal; Un = Urban national; Up = Urban provincial; W = Wet.

Fig. 5. Classification tree with crash type as response variable.

Table 4
Rules with Falling from the Vehicle as consequent.

Rule ID	Association rule Antecedent	Consequent	S %	C %	Lift	LIC
45	Pavement = Sl	FfV	0.3	26.4	3.79	n.a.
46	Pavement = Sl & Involved vehicles = SV	FfV	0.2	30.6	4.40	1.16
47	Pavement = Sl & Area = R	FfV	0.1	28.6	4.11	1.08
48	Road Type = Rm	FfV	1.7	16.9	2.43	n.a.
49	Road Type = Rm & Pavement = W	FfV	0.3	25.7	3.70	1.52
50	Road Type = Rm & Involved vehicles = SV	FfV	0.7	23.8	3.42	1.41
51	Road Type = Rm & Involved vehicles = SV & PTW driver age = 46-65	FfV	0.2	26.9	3.87	1.13
52	Road Type = Rm & PTW driver age = 46-65	FfV	0.4	19.1	2.75	1.14
53	Involved vehicles = SV	FfV	3.4	12.5	1.79	n.a.
54	Involved vehicles = SV & Pavement = W	FfV	0.5	17.2	2.47	1.38
55	Involved vehicles = SV & Pavement = W & PTW driver gender = Ma	FfV	0.4	21.6	3.11	1.26
56	Involved vehicles = SV & Pavement = W & PTW driver gender = Ma & Lighting = Day	FfV	0.3	23.4	3.36	1.08
57	Involved vehicles = SV & Pavement = W & Lighting = Day	FfV	0.4	18.4	2.64	1.07
58	Involved vehicles = SV & PTW driver age = 46-65	FfV	0.8	16.6	2.39	1.34
59	Involved vehicles = SV & Weather = Rn	FfV	0.4	16.4	2.36	1.32
60	Involved vehicles = SV & Weather = Rn & PTW Type = Moped	FfV	0.1	23.0	3.31	1.40
61	Involved vehicles = SV & Weather = Rn & PTW driver gender = Ma	FfV	0.3	20.3	2.92	1.24
62	Involved vehicles = SV & Weather = Rn & PTW driver gender = Ma & Lighting = Day	FfV	0.2	24.6	3.54	1.21
63	Involved vehicles = SV & Weather = Rn & Lighting = Day	FfV	0.3	18.3	2.63	1.11
64	Involved vehicles = SV & PTW driver age = 0-18	FfV	0.2	16.3	2.34	1.31
65	Involved vehicles = SV & PTW driver age = 0-18 & Lighting = Day	FfV	0.1	18.7	2.69	1.09
66	Involved vehicles = SV & PTW driver gender = Fe	FfV	0.5	15.3	2.21	1.23
67	Involved vehicles = SV & PTW driver gender = Fe & PTW Type = Moped	FfV	0.3	23.8	3.42	1.38
68	Involved vehicles = SV & PTW driver gender = Ma	FfV	2.7	14.6	2.10	1.17
69	Involved vehicles = SV & PTW driver gender = Ma & Road Type = Mw	FfV	0.3	19.6	2.81	1.34
70	Pavement = W	FfV	1.1	11.1	1.60	n.a.
71	Pavement = W & PTW driver age = 46-65	FfV	0.3	17.6	2.53	1.58
72	Pavement = W & PTW driver gender = Ma	FfV	0.8	13.2	1.90	1.19
73	PTW driver age = > 65	FfV	0.2	10.9	1.56	n.a.
74	Weather = Rn	FfV	0.8	10.9	1.56	n.a.
75	Weather = Rn & PTW driver age = 46-65	FfV	0.2	18.0	2.58	1.65
76	Area = R	FfV	3.2	10.6	1.53	n.a.
77	Area = R & Involved vehicles = SV	FfV	1.7	16.0	2.29	1.50
78	Area = R & Involved vehicles = SV & Pavement = Sl	FfV	0.1	31.1	4.48	1.96
79	Area = R & Involved vehicles = SV & PTW driver age = 46-65	FfV	0.4	21.1	3.04	1.33
80	Area = R & Involved vehicles = SV & PTW driver gender = Ma	FfV	1.4	19.1	2.75	1.20
81	Area = R & Involved vehicles = SV & PTW driver gender = Ma & Road Type = Mw	FfV	0.2	28.2	4.05	1.47
82	Area = R & Involved vehicles = SV & PTW driver age = 26-45	FfV	0.8	18.5	2.66	1.16
83	Area = R & PTW driver age = 46-65	FfV	0.8	14.6	2.09	1.37
84	Area = R & PTW driver gender = Fe	FfV	0.4	13.7	1.98	1.29
85	Area = R & PTW driver gender = Fe & PTW Type = Moped	FfV	0.2	17.5	2.52	1.27
86	Area = R & PTW driver gender = Fe & Lighting = Day	FfV	0.3	14.9	2.14	1.08
87	Area = R & PTW driver gender = Ma	FfV	2.4	12.7	1.83	1.20
88	Area = R & PTW driver age = 26-45	FfV	1.4	12.3	1.77	1.16
T2_7	Area = R & Involved vehicles = SV & Pavement = Sl/Ot	FfV	0.3	48.3	6.79	1.14

α_{crit} = 0.05; Fe = Female; Ma = Male; Mw = Motorway; Nt = Night; R = Rural; Rm = Rural municipal; Rn = Rainy; Sl = Slippery; SV = PTW single vehicle; W = Wet.

As far as the other crash types (Table 7), the a priori algorithm identified 54 rules (7 rules with hit obstacle as consequent, 2 rules with rear-end as consequent, 14 rules with hit pedestrian as consequent, 31 rules with angle crashes as consequent). Most rules with hit obstacle as consequent involved nighttime, with the strongest rule involving also single vehicle, moped PTW type and dry pavement (rule 157, lift = 4.40). Motorway and rainy weather were associated with rear-end crashes (rules 161 and 162). Most rules with hit pedestrian as consequent involved single vehicle, urban area or urban municipal road type, dry pavement and very young or young PTW drivers. All rules with angle crash as consequent involved crashes occurred at intersections. Additionally, most of these rules involved also urban area or urban municipal road type, car involvement, and very young (e.g. rule 178, lift = 1.51) or older PTW drivers (e.g. rule 185, lift = 1.49).

5. Discussion and conclusions

The fundamental contribution of this study is the identification of several combinations of road, environmental and drivers' characteristics associated with severity and typology of PTW crashes in Spain.

Study results showed that old (> 65), old middle-aged (46–65), and young middle-aged (26–45) PTW drivers were associated to an increase in crash severity, with the greatest effect for old drivers, followed by old-middle-aged drivers. This finding is consistent with results of U.S. studies (Savolainen and Mannering, 2007; Schneider and Savolainen, 2011) and has various potential explanations, such as degradations in riding abilities or reaction time with age, as well as physiological differences that may make older riders more susceptible to injuries. Although older riders may tend to ride at lower speeds, once in a crash they are more vulnerable to severe injury. On the other hand, very young (< 18) and young (19–25) PTW drivers were associated with hit pedestrian and angle crashes. As far as gender, PTW male drivers were associated with greater crash severity and this finding could reflect riding style and behaviour relative to males (Savolainen and Mannering, 2007).

Almost 75% of the PTW crashes occurred on municipal roads, both urban and rural, but all the other road types (motorways, rural autonomous, rural national, rural provincial, urban autonomous, urban national, and urban provincial) contributed to the increase of the proportion of severe injury and fatal crashes. This results mainly depends

Table 5
Rules with Run-off-the-road as consequent.

Rule ID	Association rule Antecedent	Consequent	S %	C %	Lift	LIC
89	Road Type = Rac	ROR	2.6	40.2	2.89	n.a.
90	Road Type = Rac & Alignment = Cu	ROR	1.1	54.6	3.92	1.36
91	Road Type = Rac & Alignment = Cu & PTW driver age = 26-45	ROR	0.5	62.2	4.47	1.14
92	Road Type = Rac & Alignment = Cu & PTW driver gender = Ma	ROR	0.8	58.5	4.21	1.06
93	Road Type = Rac & Weather = Rn	ROR	0.3	54.2	3.90	1.35
94	Road Type = Rac & Involved vehicles = SV	ROR	1.4	53.2	3.82	1.32
95	Road Type = Rac & Involved vehicles = SV & PTW driver gender = Ma	ROR	1.1	59.3	4.26	1.12
96	Road Type = Rac & Pavement = W	ROR	0.4	49.1	3.53	1.22
97	Road Type = Rp	ROR	1.8	34.4	2.47	n.a.
98	Road Type = Rp & Pavement = W	ROR	0.3	48.6	3.49	1.41
99	Road Type = Rp & Pavement = W & Involved vehicles = SV	ROR	0.1	55.8	4.01	1.18
100	Road Type = Rp & Involved vehicles = SV	ROR	0.9	48.5	3.49	1.41
101	Road Type = Rp & Alignment = Cu	ROR	0.6	44.2	3.17	1.28
102	Road Type = Rp & Alignment = Cu & PTW driver age = 26-45	ROR	0.3	48.0	3.45	1.09
103	Road Type = Rn	ROR	0.9	33.0	2.37	n.a.
104	Road Type = Rn & Pavement = W	ROR	0.2	49.3	3.54	1.49
105	Road Type = Rn & Involved vehicles = SV	ROR	0.5	47.1	3.39	1.43
106	Road Type = Rn & Involved vehicles = SV & PTW driver gender = Ma	ROR	0.3	53.1	3.82	1.13
107	Road Type = Rn & Alignment = Cu	ROR	0.3	45.7	3.28	1.38
108	Road Type = Rn & Alignment = Cu & PTW driver age = 26-45	ROR	0.1	56.2	4.04	1.23
109	Road Type = Rn & Alignment = Cu & PTW driver gender = Ma	ROR	0.2	49.4	3.55	1.08
110	Area = R	ROR	9.8	32.5	2.34	n.a.
111	Area = R & Alignment = Cu	ROR	2.6	45.1	3.24	1.39
112	Area = R & Alignment = Cu & Involved vehicles = SV	ROR	1.9	57.9	4.16	1.29
113	Area = R & Alignment = Cu & PTW driver age = 26-45	ROR	1.2	51.0	3.67	1.13
114	Area = R & Pavement = W	ROR	1.6	45.5	3.27	1.40
115	Area = R & Weather = Rn	ROR	1.0	44.0	3.16	1.35
116	Area = R & Involved vehicles = SV	ROR	4.5	43.4	3.12	1.33
117	Area = R & Lighting = Nt	ROR	2.4	37.1	2.66	1.14
118	Area = R & Lighting = Nt & Pavement = W	ROR	0.4	44.3	3.18	1.19
119	Alignment = Cu	ROR	3.0	27.0	1.94	n.a.
120	Alignment = Cu & Pavement = W	ROR	0.4	32.5	2.34	1.21
121	Alignment = Cu & PTW driver age = 26-45	ROR	1.3	32.2	2.32	1.19
122	Alignment = Cu & PTW driver gender = Ma	ROR	2.1	29.9	2.15	1.11
123	Road Type = Rm	ROR	2.7	26.8	1.92	n.a.
124	Road Type = Rm & Pavement = W	ROR	0.4	35.9	2.58	1.56
125	Road Type = Rm & Alignment = Cu	ROR	0.3	34.6	2.49	1.28
126	Road Type = Rm & Involved vehicles = SV	ROR	1.1	34.4	2.47	1.54
127	Road Type = Rm & Involved vehicles = SV & PTW driver gender = Ma	ROR	0.9	36.3	2.61	1.49
128	Pavement = Sl	ROR	0.3	25.5	1.83	n.a.
129	Road Type = Mw	ROR	1.7	24.4	1.75	n.a.
130	Road Type = Mw & Pavement = W	ROR	0.4	40.2	2.89	1.74
131	Road Type = Mw & Weather = Rn	ROR	0.3	39.5	2.84	1.62
132	Road Type = Mw & Involved vehicles = SV	ROR	0.7	28.9	2.07	1.29
133	Road Type = Mw & Involved vehicles = SV & Pavement = W	ROR	0.1	36.7	2.64	1.42
134	Pavement = W	ROR	2.2	23.1	1.66	n.a.
135	Pavement = W & Involved vehicles = SV	ROR	0.8	25.9	1.86	1.16
136	Pavement = W & Involved vehicles = SV & Road Type = Rm	ROR	0.1	41.7	2.99	1.61
137	Pavement = W & Involved vehicles = SV & Weather = Cl	ROR	0.3	33.9	2.44	1.31
138	Involved vehicles = SV	ROR	6.2	22.3	1.60	n.a.
139	Involved vehicles = SV & PTW driver gender = Ma	ROR	4.5	24.4	1.75	1.09
140	Weather = Rn	ROR	1.5	20.8	1.50	n.a.
T2_17	Area = R & Involved vehicles = SV & Pavement = W/D/Fr/Sn & Alignment = Cu	ROR	0.9	32.8	3.64	1.29

α_{crit} = 0.05; Cl = Clear; Cu = Curve; Ma = Male; Mw = Motorway; Nt = Night; R = Rural; Rac = Rural autonomous communities; Rm = Rural municipal; Rn = Rainy; Rn = Rural national; Rp = Rural provincial; Sl = Slippery; SV = PTW single vehicle; W = Wet.

on the higher operating speeds on the other road types and is consistent with the results of previous studies which found that speed is the predominant factor contributing to injury severity in a motorcycle crash (Nazemetz et al., 2019; Savolainen and Mannering, 2007; Shankar and Mannering, 1996; Shneider and Savolainen, 2011; Xin et al., 2017).

Curve alignment, especially if combined with rural area, urban provincial, rural national, rural provincial and rural autonomous roads, was associated to a large increase in crash severity. The same result was found in previous studies in Italy (Cafiso et al., 2012; Montella et al., 2012) and in the United States (Nazemetz et al., 2019; Savolainen and Mannering, 2007; Schneider and Savolainen, 2011). Moreover, the study results showed that the curve alignment, alone and combined with several other items, was associated with head-on and run-off-the-road PTW crashes which, in turn, were associated to an increase in crash severity, consistently with previous studies (Savolainen and Mannering, 2007; Schneider and Savolainen, 2011).

Head-on crashes occur when one vehicle leaves its path and comes into the path of the oncoming vehicle. Generally, the two vehicles collide at an angle approaching 180 degrees, where their relative speeds are at a maximum and produce a high energy impact. In this case, the involvement of a PTW, whose driver and passenger are the most vulnerable road users because of the slower mass compared to the other vehicle, significantly increases the proportion of severe and fatal injuries. In this study, in crashes occurring on curves in rural areas, involving cars, with male PTW drivers, the number of head-on crashes resulted 5.78 times the expected number of head-on crashes if the factors were independent, thus showing that the items rural area, car involvement and male drivers further increased the propensity toward

Table 6
Rules with Head-On as consequent.

Rule ID	Association rule Antecedent	Consequent	S %	C %	Lift	LIC
141	Road Type = Rp	HO	0.3	6.1	2.38	n.a.
142	Road Type = Rn	HO	0.1	5.4	2.14	n.a.
143	Road Type = Rac	HO	0.3	4.4	1.72	n.a.
144	Area = R	HO	1.1	3.8	1.48	n.a.
145	Area = R & Alignment = Cu	HO	0.3	5.0	1.95	1.32
146	Area = R & Alignment = Cu & PTW Type = Moped	HO	0.1	7.7	3.02	1.55
147	Area = R & Alignment = Cu & Involved vehicles = Car	HO	0.2	9.2	3.60	1.85
148	Area = R & Alignment = Cu & Involved vehicles = Car & PTW driver gender = Ma	HO	0.1	14.7	5.78	1.61
149	Area = R & Involved vehicles = Car	HO	0.7	4.6	1.80	1.22
150	Alignment = Cu	HO	0.4	3.7	1.45	n.a.
151	Alignment = Cu & Involved vehicles = Car	HO	0.2	5.9	2.31	1.59
152	Alignment = Cu & Involved vehicles = Car & PTW driver gender = Ma	HO	0.2	9.1	3.59	1.55
153	Alignment = Cu & PTW Type = Moped	HO	0.1	5.2	2.06	1.42
T2_9	Area = R & Involved vehicles = Car/Truck/PTW/Bike/Ot & Road Type = Rp/Rac/Rn/Rm/Ot	HO	1.0	6.4	2.31	1.16

α_{crit} = 0.05; Car = PTW-car; Cu = Curve; Ma = male; R = Rural; Rac = Rural autonomous communities; Rn = Rural national; Rp = Rural provincial.

head-on crashes. Interestingly, in crashes occurring on curves, the involvement of mopeds instead of motorcycles increased 1.42 times (1.55 in rural area) the proportion of head-on crashes, probably because of the lower conspicuity of mopeds.

Run-off-the-road crashes involve vehicles that leave the travel lane and encroach onto the shoulder and beyond and hit one or more of any number of natural or artificial objects, such as bridge walls, poles, embankments, guardrails, parked vehicles, and trees. Hence, the severity of these crashes is frequently very high. Motorcycle crashes against safety barriers and roadside objects are widely recognized as a major safety issue. Roadside barriers present a substantial danger to PTW riders, causing serious lower extremity and spinal injuries as well as serious head injuries (ACEM, 2004), and the encroachment in the posts of guardrails gives rise to a high fatality risk of motorcyclists (Gabler, 2007). In our study, curve alignment increased 1.94 times the proportion of ROR crashes. Furthermore, motorways, rural autonomous, rural national, rural provincial road types, and young middle-aged (26–45) PTW drivers were associated with ROR crashes. Also, slippery and wet pavements were associated with ROR crashes. Indeed, contrary to cars and other four-wheeled vehicles, a PTW has only two points of contact with the surface and the consistency of the grip of the tyres on the surface is critical for the stability (ACEM, 2004; IHE, 2018). To negotiate a curve, motorcyclists lean over at an angle whose acuteness is related to speed, pavement friction and radius of the curve, and any friction deficiency, such as deficiencies on slippery and wet pavements, can destabilize the motorcycle. Consistently with previous studies carried out in Italy (Montella et al., 2012) and in the United States (Nazemetz et al., 2019), nighttime in rural area was associated with ROR crashes. A possible explanation of this issue is that in nighttime, without artificial lighting, the visibility of the road alignment is reduced and there is a greater chance to leave the travel lane.

Falling from the vehicles crash type showed several significant associations. As expected, pavement conditions played a key role. Slippery pavement increased 3.79 times the proportion of FfV crashes whereas wet pavement increased 1.60 times FfV crash proportion. Because of the degradation in riding ability, old PTW drivers were associated with falling from the vehicle.

Moreover, the study results provided non-trivial and unsuspected relations in the data. As an example, intersection and young ($\leq$25) or old ($>$ 65) PTW drivers were associated with angle crashes (rules 179, 186, and 189). Hence, the relationship between PTW driver age and propensity towards angle crashes at intersections was not monotonic. Both young (lift equal to 1.69 for age $\leq$ 18, lift equal to 1.54 for age between 19 and 25) and old drivers (Lift equal to 1.67) were associated with angle crashes. A possible explanation of this result is that young drivers have lower propensity to give way to the other vehicles while old drivers have interaction difficulty, combined with the lower tendency of cars to give way to PTWs (ACEM, 2004). As far as falling from the vehicle crashes is concerned, several rules involved PTW driver gender and provided unsuspected relations. In rural area, female PTW drivers and mopeds were strongly associated with falling from the vehicle (rule 85, lift equal to 2.52). Even stronger was the association of single vehicle crashes, female PTW drivers and mopeds with falling from the vehicle (rule 67, lift equal to 3.42). Conversely, male PTW drivers were associated with falling from the vehicle in risky conditions, such as wet pavement (rule 72, lift equal to 1.90), rainy (& single vehicle & day, rule 62, lift equal to 3.54), or motorways (& single vehicle & rural area, rule 81, lift equal to 4.05).

Several countermeasures may be implemented to solve or mitigate the safety issues identified in our study. To reduce curve, ROR and HO crashes, associated with high severity, improvement in design consistency of high-speed roads is strongly recommended. Especially, based on the expected safety consequences of the highway geometric design (Cafiso et al., 2007, 2011; Montella and Imbriani, 2015), it is recommended to minimise the difference between friction demand and friction supply on horizontal curves, to minimise the operating speed reduction from tangents to horizontal curves, and to avoid low-operating speed curves following long tangents.

Based on study results, roadside improvements have great potential for the reduction of PTW crash severity. The roadside should be made as forgiving as possible, considering the following options in decreasing priority order (La Torre, 2018): remove the hazard, redesign the hazard to be safely traversable, relocate the hazard further away from the road, make the hazard passively safe, install a vehicle restraint system (safety barrier or crash cushion), delineate the hazard. Most road restraint systems are not designed to protect PTW drivers and passengers and the use of motorcycle road restraint systems which reduce the impact severity of motorcyclist collisions with safety barriers, tested according to the CEN technical specifications 1317-8 (CEN, 2012), should be considered. In order to minimise the consequences to a rider who impacts a barrier directly, it may be necessary to fit a barrier with a specific PTW rider protection system. Alternatively, a barrier might specifically incorporate characteristics limiting the consequences of a PTW rider impact.

To reduce slippery pavement, wet pavement, run-off-the-road and falling from the vehicle crashes, a significant improvement in both pavement skid resistance (e.g., high-friction surface treatments) and pavement unevenness (e.g., pothole repair, partial depth removal and resurfacing, in-place recycling, full depth reclamation) is required. Indeed, PTWs are more sensitive to roadway surface conditions than other motorized vehicles (ITF, 2015) and recent studies showed that pavement friction improvements significantly reduced wet and run-off-

Table 7
Rules with Angle, Hit obstacle, Hit parked Vehicle, Hit Pedestrian, and Rear-End as consequent.

Rule ID	Association rule Antecedent	Consequent	S %	C %	Lift	LIC
154	Lighting = Nt	Hobs	1.0	4.2	1.93	1.00
155	Lighting = Nt & Involved vehicles = SV	Hobs	0.4	6.4	2.90	1.50
156	Lighting = Nt & Involved vehicles = SV & PTW Type = Moped	Hobs	0.2	8.7	3.96	1.37
157	Lighting = Nt & Involved vehicles = SV & PTW Type = Moped & Pavement = D	Hobs	0.2	9.7	4.40	1.11
158	Lighting = Nt & Area = R	Hobs	0.4	6.4	2.89	1.50
159	Involved vehicles = SV	Hobs	1.0	3.6	1.63	1.00
160	Involved vehicles = SV & PTW driver age = 19-25	Hobs	0.2	5.3	2.40	1.47
161	Road Type = Mw	RE	2.2	32.0	1.52	1.00
162	Road Type = Mw & Weather = Rn	RE	1.4	35.3	1.68	1.11
163	Involved vehicles = SV	Ped	2.7	9.9	1.45	1.00
164	Involved vehicles = SV & Road Type = Um	Ped	2.4	16.0	2.36	1.63
165	Involved vehicles = SV & Road Type = Um & Pavement = D	Ped	2.2	17.4	2.56	1.08
166	Involved vehicles = SV & Road Type = Um & Pavement = D & PTW driver age = 19-25	Ped	0.3	18.7	2.74	1.07
167	Involved vehicles = SV & Road Type = Um & PTW driver age = 19-25	Ped	0.3	17.0	2.50	1.06
168	Involved vehicles = SV & Area = U	Ped	2.6	14.8	2.17	1.50
169	Involved vehicles = SV & PTW driver age = 0-18	Ped	0.1	14.5	2.13	1.47
170	Involved vehicles = SV & PTW driver age = 0-18 & Pavement = D	Ped	0.1	15.5	2.28	1.07
171	Involved vehicles = SV & Pavement = D	Ped	2.5	10.6	1.56	1.08
172	Involved vehicles = SV & Pavement = D & Area = U	Ped	2.3	15.9	2.33	1.49
173	Involved vehicles = SV & Pavement = D & PTW driver age = 19-25	Ped	0.3	12.9	1.89	1.21
174	Involved vehicles = SV & Pavement = D & PTW Type = Moped	Ped	0.8	12.0	1.76	1.13
175	Road Type = Um	Ped	6.1	9.4	1.38	1.00
176	Area = U	Ped	6.4	9.2	1.35	1.00
177	Alignment = Int	An	19.3	45.3	1.23	1.00
178	Alignment = Int & PTW driver age = 0-18	An	1.2	55.8	1.51	1.23
179	Alignment = Int & PTW driver age = 0-18 & Road Type = Um	An	0.9	62.5	1.69	1.12
180	Alignment = Int & PTW driver age = 0-18 & Area = U	An	1.0	62.0	1.68	1.11
181	Alignment = Int & PTW driver age = 0-18 & Area = U & Involved vehicles = Car	An	0.8	67.7	1.83	1.09
182	Alignment = Int & PTW driver age = 0-18 & Involved vehicles = Car	An	1.0	60.6	1.64	1.09
183	Alignment = Int & PTW driver age = 0-18 & Involved vehicles = Car & Road Type = Um	An	0.8	68.3	1.85	1.13
184	Alignment = Int & PTW driver age = 0-18 & Pavement = D	An	1.1	59.1	1.60	1.06
185	Alignment = Int & PTW driver age = > 65	An	0.4	55.0	1.49	1.21
186	Alignment = Int & PTW driver age = > 65 & Area = U	An	0.3	61.6	1.67	1.12
187	Alignment = Int & PTW driver age = > 65 & Pavement = D	An	0.4	58.6	1.59	1.07
188	Alignment = Int & PTW driver age = 19-25	An	3.1	52.2	1.41	1.15
189	Alignment = Int & PTW driver age = 19-25 & Road Type = Um	An	2.7	56.8	1.54	1.09
190	Alignment = Int & PTW driver age = 19-25 & Involved vehicles = Car	An	2.4	57.9	1.57	1.11
191	Alignment = Int & PTW driver age = 19-25 & Involved vehicles = Car & Area = U	An	2.1	61.8	1.67	1.06
192	Alignment = Int & PTW driver age = 19-25 & Area = U	An	2.7	56.6	1.53	1.09
193	Alignment = Int & Road Type = Um	An	15.5	51.6	1.40	1.14
194	Alignment = Int & Road Type = Um & Involved vehicles = PTW	An	0.8	56.1	1.52	1.09
195	Alignment = Int & Road Type = Um & PTW driver gender = Ma	An	11.5	54.3	1.47	1.05
196	Alignment = Int & Road Type = Um & Pavement = D	An	14.4	54.1	1.47	1.05
197	Alignment = Int & Road Type = Um & Pavement = D & PTW driver gender = Fe	An	2.4	57.1	1.55	1.05
198	Alignment = Int & Area = U	An	16.2	51.4	1.39	1.13
199	Alignment = Int & Area = U & Involved vehicles = Car	An	12.2	56.8	1.54	1.11
200	Alignment = Int & Involved vehicles = Car	An	14.5	50.8	1.38	1.12
201	Alignment = Int & PTW driver gender = Ma	An	14.4	48.6	1.32	1.07
202	Alignment = Int & PTW driver gender = Ma & Pavement = D	An	13.5	51.4	1.39	1.05
203	Alignment = Int & PTW driver gender = Fe	An	3.0	48.4	1.31	1.07
204	Alignment = Int & PTW driver gender = Fe & Pavement = D	An	2.8	51.5	1.39	1.06
205	Alignment = Int & Pavement = D	An	18.0	48.1	1.30	1.06
206	Alignment = Int & Pavement = D & PTW driver age = 26-45	An	8.5	51.0	1.38	1.06
207	Alignment = Int & PTW driver age = 26-45	An	9.1	48.0	1.30	1.06
T2_14	Area = U & Involved vehicles = Car/Truck/PTW/Bike/Ot & Alignment = Int/Ot	An	2.2	14.8	2.33	1.17
T2_11	Area = U & Involved vehicles = SV & Pavement = D	Ped	0.9	27.9	2.02	1.08

α_{crit} = 0.05; Car = PTW-car; D = Dry; Fe = Female; Hit obstacle = Hit obstacle in carriageway; Int = Intersection; Ma = Male; Mw = Motorway; Nt = Night time; PTW = PTW-PTW; R = Rural; Rn = Rainy; SV = PTW single vehicle; U = Urban; Um = Urban municipal.

the-road crashes (Geedipally et al., 2017; Lyon et al., 2018) and that maintenance treatments that affect the pavement surface to improve the International Roughness Index significantly reduced crashes (Nemtsov et al., 2019).

As far as PTW drivers are concerned, study results provide noteworthy suggestions. Women have higher propensity to fall from the vehicle and more training to improve everyday riding skills would be helpful to reduce this crash type. Males have higher propensity to severe crashes and more education and safety campaigns to raise awareness of the risks and consequences associated with riding a PTW, as well as highlighting the benefits of wearing appropriate protective clothing, would be helpful to improve their riding style and behaviour. Furthermore, old, old middle-aged, and young middle-aged PTW drivers were associated to an increase in crash severity. Crash studies have provided strong evidence that effective helmets and motorcycle personal protective clothing fitted with impact protection can prevent or reduce many of the injuries most commonly sustained by riders in motorcycle crashes (de Rome, 2018). Use of protective clothing, which includes jackets, pants, gloves and boots, is generally very low and the implementation of motorcycle protective clothing rating schemes, such as the MotoCAP (Motorcycle Clothing Assessment Program) developed in Australia, would provide incentive for industry to improve the

performance of protective clothing (PPE) and would increase the usage and effectiveness of the PPE available by enabling riders to make well-informed purchasing decisions. Unfortunately, Spanish crash data do not contain information on PTW protective clothing and the use of updated electronic forms and software for crash data collection, processing and analysis is strongly recommended (Montella et al., 2013, 2017a, 2017b). Finally, very young and young PTW drivers were associated with hit pedestrian and angle crashes. Both crash types strongly depend on the give way behaviour of young drivers. Safety campaigns for young drivers would raise awareness and improve behaviour while engineering countermeasures, such as build-outs pavement on pedestrian crossings, may be considered to increase pedestrian visibility. Moreover, high visible clothing aiming to improve conspicuity during daytime (fluorescent or bright clothing, vest, helmet, etc.) or during night time (reflective parts incorporated in the jacket or vest) would increase the visibility by the other road users and reduce angle crashes.

From the methodological point of view, study results show that both the classification trees and the a priori algorithm were effective in providing non-trivial and unsuspected relations in the data. Classification trees structure allowed a simpler understanding of the phenomenon under study while association discovery provided new information which was previously hidden in the data. Given that the results of the two different techniques were never contradictory, we recommend using classification trees and association discovery as complementary approaches since their combination is effective in exploring data providing meaningful insights about PTW crash characteristics and their interdependencies.

From the methodological point of view, it is noteworthy to highlight that both classification trees and association rules suffer from an extreme risk of type I error, that is of finding patterns that appear due to chance alone. We overcame this problem by randomly splitting the study data in two data sets and by using a validation procedure to minimize the risk of type I error in the classification trees and hypothesis testing to reduce the error rate in the association rules discovery. Furthermore, since the crash data were not balanced as far as the variable severity and this issue affected the performance of the model, we reduced the unbalance by combining the categories fatal injuries and severe injuries. We also optimized the performance of the data mining algorithms, introducing the posterior classification ratio in the classification trees and the lift increase criterion in the association rules. Study results show that both the classification trees and the a priori algorithm were effective in providing non-trivial and unsuspected relations in the data. Classification trees structure allowed a simpler understanding of the phenomenon under study while association discovery provided new information which was previously hidden in the data. Given that the results of the two different techniques were never contradictory, we recommend using classification trees and association discovery as complementary approaches since their combination is effective in exploring data providing meaningful insights about PTW crash characteristics and their interdependencies.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors express their gratitude to the Spanish General Directorate of Traffic (DGT) for supporting this research.

References

ACEM, 2004. MAIDS—In Depth Investigations of Accidents Involving Powered Two Wheelers. Available at:. . https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/projects_sources/maids_report_1_2_september_2004.pdf.

ACEM, 2010. The Motorcycle Industry in Europe. Available at:. . https://www.acem.eu/images/stories/doc/publications/ACEM_REPORT.pdf.

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD 207–216.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, CA.

Cafiso, S., La Cava, G., Montella, A., 2007. Safety index for evaluation of Two-lane rural highways. Transp. Res. Rec. 2019, 136–145. https://doi.org/10.3141/2019-17.

Cafiso, S., La Cava, G., Montella, A., 2011. Safety inspections as supporting tool for safety management of Low-volume Roads. Transp. Res. Rec. 2203, 116–125. https://doi.org/10.3141/2203-15.

Cafiso, S., La Cava, G., Pappalardo, G., 2012. A logistic model for powered Two-wheelers crash in Italy. Proc. Soc. Behav. Sci. 53, 880–889. https://doi.org/10.1016/j.sbspro.2012.09.937.

CEN, 2012. CEN/TS 1317-8 Road Restraint Systems–Part 8: Motorcycle Road Restraint Systems Which Reduce the Impact Severity of Motorcyclist Collisions With Safety Barriers.

Chang, F., Li, M., Xu, P., Zhou, H., Haque, Md., Huang, H., 2016. Injury severity of motorcycle riders involved in traffic crashes in hunan, China: a mixed ordered logit approach. Environmental research and public health. Traffic Saf. Inj. Prev. 13, 714. https://doi.org/10.3390/ijerph13070714.

Cunto, F., Ferreira, S., 2016. An analysis of the injury severity of motorcycle crashes in Brazil using mixed ordered response models. J. Transp. Saf. Secur. 9, 33–46. https://doi.org/10.1080/19439962.2016.1162891.

Das, S., Dutta, A., Avelar, R., Dixon, K., Sun, X., Jalayer, M., 2019. Supervised association rules mining on pedestrian crashes in urban areas: identifying patterns for appropriate countermeasures. Int. J. Urban Sci. 23, 38–40. https://doi.org/10.1080/12265934.2018.1431146.

De Oña, J., De Oña, R., Calvo, F., 2012. A classification tree approach to identify key factors of transit service quality. Expert Syst. Appl. 39, 11164–11171. https://doi.org/10.1016/j.eswa.2012.03.037.

De Oña, J., López, G., Abellán, J., 2013. Extracting decision rules from police accident reports through decision trees. Accid. Anal. Prev. 50, 1151–1160. https://doi.org/10.1016/j.aap.2012.09.006.

De Oña, J., De Oña, R., López, G., 2015. Transit service quality analysis using cluster analysis and decision trees: a step forward to personalized marketing in public transportation. Transportation 43, 725–747. https://doi.org/10.1007/s11116-015-9615-0.

De Rome, L., 2018. Stars or Standards? A Review of Motorcycle Protective Clothing from the Southern Hemisphere. Available at:. https://www.racfoundation.org/wp-content/uploads/Stars_or_standards_a_review_of_motorcycle_protective_clothing_Liz_de_Rome_December_2018.pdf.

DGT – Directorate General of Traffic, 2018. Principales Cifras de Sinistralidad. Available at:. http://www.dgt.es/es/seguridad-vial/estadisticas-e-indicadores/publicaciones/principales-cifras-siniestralidad/.

ERF – European Union Road Federation, 2018. Improving Infrastructure Safety for Powered Two-Wheelers. Available at. http://erf.be/press-releases/improving-infrastructure-safety-for-powered-two-wheelers.

European Commission, 2010. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Available at:. https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/road_safety_citizen/road_safety_citizen_100924_en.pdf.

European Commission, 2018a. EU Transport in Figures 2018. Available at:. https://ec.europa.eu/transport/facts-fundings/statistics/pocketbook-2018_en.

European Commission, 2018b. Strategic Action Plan on Road Safety. Available at. https://eur-lex.europa.eu/resource.html?uri=cellar%3A0e8b694e-59b5-11e8-ab41-01aa75ed71a1.0003.02/DOC_2&format=PDF.

Gabler, H.C., 2007. The emerging risk of fatal motorcycle crashes with guardrails. 86th TRB Annual Meeting.

Geedipally, S.R., Pratt, M.P., Lord, D., 2017. Effects of geometry and pavement friction on horizontal curve crash frequency. J. Transp. Saf. Secur. 1–22. https://doi.org/10.1080/19439962.2017.1365317.

Hahsler, M., Buchta, C., Gruen, B., Hornik, K., 2018. Arules: Mining Association Rules and Frequent Itemsets. Available at:. http://cran.r-project.org/web/packages/arules/arules.pdf.

Hidalgo-Fuentes, S., Sospedra-Baeza, M., 2018. Gender and age distribution of motorcycle crashes in Spain. Int. J. Inj. Control Saf. Promot 26, 108–114. https://doi.org/10.1080/17457300.2018.1482927.

IHE – Institute of Highway Engineers, 2018. Guidelines for Motorcycling. Available at. http://www.motorcycleguidelines.org.uk/.

IRTAD – International Transport Forum and International Traffic Safety Data and Analysis Group, 2018. Road Safety Annual Report 2018. Available at:. . https://www.itf-oecd.org/sites/default/files/docs/irtad-road-safety-annual-report-2018_2.pdf.

ITF – International Transport Forum, 2015. Improving Safety for Motorcycle, Scooter and Moped Riders. OECD Publishing, Paris. https://doi.org/10.1787/9789282101964-en.

Ivers, R., Sakashita, C., Senserrick, T., Elkington, J., Lo, S., Boufous, S., Rome, L., 2016. Does an on-road motorcycle coaching program reduce crashes in novice riders? A randomised control trial. Accid. Anal. Prev. 86, 40–46. https://doi.org/10.1016/j.aap.2015.10.015.

Kashani, A., Rabieyan, R., Besharati, M., 2014. A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers. J. Saf. Res. 51, 93–98. https://doi.org/10.1016/j.jsr.2014.09.004.

La Torre, F., 2018. Controlled access facilities (freeways). Transp. Sustain. 11, 107–126. https://doi.org/10.1108/S2044-994120180000011006.
Li, Y., Yamamoto, T., Zhang, G., 2018. Understanding factors associated with misclassification of fatigue-related accidents in police record. J. Saf. Res. 64, 155–162. https://doi.org/10.1016/j.jsr.2017.12.002.
López, G., Abellán, J., Montella, A., de Oña, J., 2014. Patterns of single-vehicle crashes on two-lane rural highways in Granada Province, Spain: in-depth analysis through decision rules. Transp. Res. Rec. 2432, 133–141. https://doi.org/10.3141/2432-16.
López, G., de Oña, J., 2017. Extracting crash patterns involving vulnerable users on two-lane rural highways. Secur. Vialis 9, 1–13. https://doi.org/10.1007/s12615-016-9088-8.
Lyon, C., Persaud, B., Merritt, D., 2018. Quantifying the safety effects of pavement friction improvements–results from a large-scale study. Int. J. Pavement Eng. 19 (2), 145–152. https://doi.org/10.1080/10298436.2016.1172709.
Mannering, F., Shankar, V., Bhat, C., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Meth. Accid. Res. 11, 1–16. https://doi.org/10.1016/j.amar.2016.04.001.
Milton, J., Shankar, V., Mannering, F., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accid. Anal. Prev. 40, 260–266. https://doi.org/10.1016/j.aap.2007.06.006.
Montella, A., 2011. Identifying crash contributory factors at urban roundabouts and using association rules to explore their relationships to different crash types. Acc. Anal. Prev. 43, 1451–1463. https://doi.org/10.1016/j.aap.2011.02.023.
Montella, A., Aria, M., D'Ambrosio, A., Mauriello, F., 2011. Data-mining techniques for exploratory analysis of pedestrian crashes. Transp. Res. Rec. 2237, 107–116. https://doi.org/10.3141/2237-12.
Montella, A., Aria, M., D'Ambrosio, A., Mauriello, F., 2012. Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accid. Anal. Prev. 49, 58–72. https://doi.org/10.1016/j.aap.2011.04.025.
Montella, A., Andreassen, D., Tarko, A., Turner, S., Mauriello, F., Imbriani, L., Romero, M., 2013. Crash databases in australasia, the European union, and the United States. Transp. Res. Rec. 2386, 128–136. https://doi.org/10.3141/2386-15.
Montella, A., Imbriani, L.L., 2015. Safety performance functions incorporating design consistency variables. Accid. Anal. Prev. 74, 133–144. https://doi.org/10.1016/j.aap.2014.10.019.
Montella, A., Chiaradonna, S., Criscuolo, G., De Martino, S., 2017a. Perspectives of a web-based software to improve crash data quality and reliability in Italy. 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS) 451–456. https://doi.org/10.1109/MTITS.2017.8005714.
Montella, A., Chiaradonna, S., Criscuolo, G., De Martino, S., 2017b. Development and evaluation of a web-based software for crash data collection, processing and analysis. Accid. Anal. Prev. https://doi.org/10.1016/j.aap.2017.01.013. in press.
Moral-García, S., Castellano, J.G., Mantas, C.J., Montella, A., Abellán, J., 2019. Decision tree ensemble method for analyzing traffic accidents of novice drivers in Urban areas. Entropy 21 (4), 360. https://doi.org/10.3390/e21040360.
Nazemetz, J.W., Bents, F.D., Perry, J.G., Thor, C., Mohamedshah, Y.M., 2019. Motorcycle Crash Causation Study: Final Report. Report FHWA-HRT-18-064. Available at:. . https://www.fhwa.dot.gov/publications/research/safety/18064/18064.pdf.
Nemtsov, I., Jafari, A., Persaud, B., Lindley, I., 2019. Safety effects of pavement maintenance treatments for two-lane rural roads: insights for pavement management. 98th TRB Annual Meeting.
Pande, A., Abdel-Aty, M., 2009. A novel approach for analyzing severe crash patterns on multilane highways. Accid. Anal. Prev. 56, 95–102. https://doi.org/10.1016/j.aap.2009.06.003.
Perez-Fuster, P., Rodrigo, M., Ballestar, M., Sanmartin, J., 2013. Modeling offenses among motorcyclists involved in crashes in Spain. Accid. Anal. Prev. 49, 44–49. https://doi.org/10.1016/j.aap.2013.03.014.
Savolainen, P., Mannering, F., 2007. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. Acc. Anal. Prev. 39, 955–963. https://doi.org/10.1016/j.aap.2006.12.016.
Savolainen, P., Mannering, F., Lord, D., Quddus, M., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Acc. Anal. Prev. 43, 1666–1676. https://doi.org/10.1016/j.aap.2011.03.025.
Schneider, W., Savolainen, P., 2011. Comparison of severity of motorcyclist injury by crash types. Transp. Res. Rec. 2265, 70–80. https://doi.org/10.3141/2265-08.
Shankar, V., Mannering, F., 1996. An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. J. Saf. Res. 27 (3), 183–194. https://doi.org/10.1016/0022-4375(96)00010-2.
Webb, G.I., 2007. Discovering significant patterns. Mach. Learn. 68, 1–33. https://doi.org/10.1007/s10994-008-5045-y.
WHO – World Health Organization, 2017. Powered Two- and Three Wheeler Safety. Available at:. http://apps.who.int/iris/bitstream/handle/10665/254759/9789241511926-eng.pdf;jsessionid=36B10DCC389E19027B4A341C83E1AD1D?sequence=1.
Xin, C., Wang, Z., Lee, C., Lin, P.S., 2017. modeling safety effects of horizontal curve design on injury severity of single-motorcycle crashes with mixed-effects logistic model. Transp. Res. Rec. 2637, 38–46. https://doi.org/10.3141/2637-05.
Ye, F., Lord, D., 2014. Comparing three commonly used crash severity models on sample size requirements: multinomial logit, ordered probit and mixed logit models. Anal. Meth. Acci. Res. 1, 72–85. https://doi.org/10.1016/j.amar.2013.03.001.