Abstract

One of the main objectives of all public administrations is reducing traffic crashes. To this end, Road Safety Inspections (RSI) stand out as a key measure. Signaling roads is one of the foremost tasks of RSI. A road that is improperly or poorly signaled can lead to incorrect placement or maneuvers of vehicles and ambiguous situations that can increase the risk of crashes. This paper analyses the relationship between road crashes in two-lane rural highways and certain deficiencies in signaling. The results show that deficiencies such as “incomplete removal of road works markings” or “no guide sign or in incorrect position” are the ones associated with a higher probability of crashes in two-lane rural highways. In view of these results, governmental agencies should verify that the original conditions of a highway are re-established after any construction work is completed. They should also continuously follow up on the signaling of this type of highway in order to maintain them in optimal conditions.

Keywords. Traffic crashes; road safety inspections; sign and marking; Decision Trees; Decision rules

1. Introduction

Traffic accidents are complex events involving the interaction of different contributory factors, including road, driver, and vehicle. While it is well known that the human factor is the main cause of traffic crashes, present in nearly 90% of them (Siskind et al., 2011), previous studies have shown that the infrastructure also plays a significant role. Nearly 28% of crashes are due to infrastructure and, in most cases a combination of human and road factors forms a major contribution in the road crashes (Odgen, 1996).

In the literature, the crash contribution from human factors is usually analyzed in the context of driver errors. The human error most often identified in the crashes is related to the perception and processing of information presented by the road or traffic environment. Situations that cause problems with road user perception, interpretation or judgment stages may lead to driver error or loss of control (Croft and Schnerring, 2009). An estimated 30% of driver-distracted crashes derive from diverse sources outside the vehicle (Regan et al., 2009). Hence, it is crucial to maintain the road features in optimal conditions so that they have the least possible impact on the driver´s performance.

Reducing highway crashes is one of the main aim of the Administrations. One means of reducing them is to detect and correct roadway deficiencies. Road Safety Inspections (RSI) were established for this purpose, they are an effective tool for the management of safety on existing roads. The European Directive on Road Infrastructure Safety Management (EC, 2008) defines RSI as "an ordinary periodical verification of the characteristics and defects that require maintenance work for reasons of safety". Following the principle "Prevention is better than cure", the RSI are used to evaluate existing road traffic facilities and to improve road safety performance (Alfredas et al., 2012). While some RSI treatments will have a greater impact than others, as underlined by Elvik (SETRA 2008), significant reductions in crashes can be expected as a result of a RSI and associated interventions.

After some years of experience with RSI, it is broadly recognized as one of the most important and effective engineering tools available to improve road safety (Antov, 2011). This is why the European Union makes RSI mandatory for trans-European Road Networks and they are recommended for the rest of the transport infrastructures (EC, 2008). These inspections should be undertaken after establishing a series of criteria to be articulated by means of checklists. The checklists are ordered lists used to cover the most important issues which should be inspected during the RSI. The detected hazards will be identified like Road Safety Deficiency (RSD).

Some aspects of a RSD that can be analyzed include those related with signaling. They are easier to correct and involve a lower cost than other measures, such as the design of the road itself. It is important for the highway to be properly signaled, and that the information provided is clear and concise.

Several authors stress the importance of correct signaling: Miller (1992) reported that existing longitudinal pavement markings reduce crashes by 21%, and edge lines on rural two-lane highways reduce crashes by 8%; and Cho et al. (2012) suggested that pavement markings provide guidance to road travelers. An alteration of pavement color and/or texture or incomplete removal of pavement markings during construction projects could confuse individuals driving through the construction work zones. To make matters worse, under certain lighting and weather conditions the supposedly removed markings may become more visible than the new ones. Antov (2011) highlighted as common problems the missing, contradictory or incomprehensible signs/marking. Croft and Schnerring (2009) pointed out that incorrect or poorly maintained pavement marking can lead to undue placement or maneuvers of vehicles, thus increasing the risk of crashes. They also showed the influence of delineation devices in road safety—poorly placed or missing delineation devices can transmit a false picture of the way ahead, contributing to driver error (Croft and Schnerring, 2009). The methodological approach for safety evaluation of two-lane rural highways segments put forth by Cafiso et al. (2007) served to establish that daytime delineation of a road can be effectively accomplished with pavement markings, whereas nighttime and rainy conditions may require a different approach to provide long-range delineation of the roadway alignment. Supplementary delineation is an important safety factor in any condition; but it may prove critical on horizontal curves, especially on isolated curves with a short radius. Croft and Schnerring (2009) also indicated that signs poorly located/incorrectly situated can cause confusion, increasing crash risk, just as excessive signing can increase potential risk for road users. Montella (2005) described a systematic process to determine which road features should be investigated and how each should be evaluated during RSI. Accordingly, a safety improvement index was calculated and compared with the expected collision frequency, and this procedure was carried out in 406 km of rural two-lane rolling highways in Italy. The study revealed that for missing or ineffective curve warning signs on severe curves, the relative risk factor could be assumed equal to 10% (Montella, 2005).

This study analyzes the relationship between crashes and certain deficiencies in signaling identified by a previous RSI. The RSI was performed on two-lane rural highways in Andalusia (Spain). From the Road Safety standpoint, it is vital that two-lane rural highways be studied, as they are the scenario of most crashes. In Spain, 70% of crashes occur on this type of roads (Ministerio del Interior, 2013).

The analysis uses a data mining technique. This technique has been widely used in the road safety field in recent years, giving satisfactory results (Kuhnert et al., 2000; Sohn and Shin, 2001; Abdel Wahab and Abdel-Aty, 2001; Chang and Wang, 2006; De Oña et al., 2011; Kashani et al., 2011; Pakgohar et al., 2010; Chang and Chien, 2013; De Oña et al., 2013a; De Oña et al., 2013b; López et al., 2014). The main aim of this technique is the extraction of knowledge from large amounts of previously unknown and indistinguishable data. In this case, Decision Trees (DTs) are employed. DTs are appropriate for studying crashes because they are non-parametric techniques that do not require prior probabilistic knowledge of the study phenomena. Further advantages of DTs with respect to other methods having similar aims reside in the extraction of Decision Rules (DRs) (De Oña et al., 2013a). Although each crash is the result of a unique chain of events, some specific factors are common to several crash circumstances, and DRs can be used to identify these factors and their interdependences (Montella et al., 2011). Safety analysts could use these rules to understand the events leading up to an accident, and prioritize certain elements for actions intended to improve road safety.

In this paper, therefore, DRs extracted from DTs are used to analyze the relationship between the actual occurrence of traffic crashes on two-lane rural highways and the deficiencies in roadway signaling previously detected by means of RSI.

The paper is organized as follows: Section 2 presents a description of the data used, and also describes the procedures for building DTs, extract DRs, and deriving the final rule set. Section 3 presents the Results and a Discussion thereof. Finally, the last section succinctly presents some Conclusions.

2. MATERIALS AND METHODS

2.1 Description of the data

The data come from two different sources. The Andalusian Regional Government provided the Road Inventory database and the Road Safety Inspections database, while the Spanish General Directorate of Traffic provided the Spanish Road Crashes database.

The Road Inventory database contains a list of road sections with their geometrics and equipment characteristics. Two-lane rural highways from the Complementary Road Network of Andalusia were selected for this study. Urban segments, junctions and segments with road work places were removed from the study; because the factors related to crashes taking place on these sections are different, they should be analyzed separately (Moore et al., 2010). The total length of the investigated road network is 1,635 km.

The Road Safety Inspections database contains information about a RSI developed on two-lane rural highways in the Complementary Road Network in Andalusia. In this RSI some risks associated with RSD were identified. These risks were defined as Road Safety Deficiency Elements (RSD-E). The risks related with the vertical signs and pavement markings identified during the RSI are denominated Signaling Elements from Road Safety Inspection (RSD-SE). The main aim of this study is to investigate the influence of RSD-SE on road crashes.

The Spanish Road Crashes database contains a description of the location and type of crashes that occurred on Spanish roads. Information about the crashes in two-lane rural highways in the Complementary Road Network of Andalusia was extracted from this database. The period of study is three years (2006-2008), and during this period the total number of crashes with victims in these segments was 1,454.

A global database with information about crashes, road characteristics, and RSD-SE was built using the three databases. The following analysis is based on seven variables related to geometric and environmental road characteristics (see Table 1) and eight RSD-SE (see Table 2).

[Insert here Table 1]

[Insert here Table 2]

The following criteria were corroborated in order to identify the RSD-SE:

- RSD-SE1: The length of the passing zone was at least the minimum indicated by Spanish National Standards. For a speed of 100 km/h, the minimal distance is 250 meters. For 60 km/h the minimum is 75 meters. For speeds in-between, intermediate distances are established.
- RSD-SE2: The regulatory signs are present and correctly positioned (e.g., speed limit or no passing zone).
- RSD-SE3: Signs indicating danger/precaution are present and correctly positioned (e.g., road narrows, dangerous curve, animal crossing, etc.)
- RSD-SE4: Guide signs are properly situated.
- RSD-SE5: Road markings are clear and visible.
- RSD-SE6: There are no contradictions between vertical signs and road markings at a given point.
- RSD-SE7: This deficiency is considered to exist in segments where road markings have not been adequately eliminated.
- RSD-SE8: This deficiency is considered to exist in segments where the road width is greater than 7 meters and there are no post-mounted delineators or they present damage amounting to over 50%.

[Insert here Figure 1]

2.2. Classification and Decisions Trees (CART)

DTs are one of the most widely used data mining techniques for classifying and predicting class variables. When the target variable is discrete, a classification tree is developed, whereas a regression tree is developed for continuous variables. CART can be used for both kinds of target variables. In this study, the target variable is the occurrence of the accident (ACC: YES or NO) and, therefore, a classification tree is developed.

A DT is an oriented graph formed by a finite number of nodes departing from the root node. DTs are built recursively, following a descending strategy, starting with the full data set (made by the root node). Using specific split criteria, the full set of data is then split into even smaller subsets. Each subset is split recursively until all of them are pure (when all cases in each subset present the same class) or their "purity" cannot be increased. Thus the tree's terminal nodes are formed, obtained according to the answer values of the target variable (De Oña et al., 2013a).

The CART method is a particular methodology for building binary Decision Trees in which the Gini Index is used as the splitting criterion. The development of a CART model generally consists of three steps: (1) growth of the tree; (2) the pruning process; and (3) selection of an optimal tree from the pruned trees. Tree growing entails recursive partitions of the target variable to maximize ''purity'' in the two subsequent child nodes. By definition, the terminal nodes present a lower degree of impurity compared to the root node. In tree growing, predictors generate candidate partitions (or splits) at each internal node of the tree, so that a suitable criterion needs to be defined in order to choose the best partition (or the best split) of the objects. The Gini reduction criteria is applied to measure the ''worth'' of each split in terms of its contribution toward maximizing the homogeneity through the resulting split. If a split results in the splitting of one parent node into B branches, the “worth” of that split may be measured as follows:

$$\text{Worth} = \text{Impurity (Parent node)} - \sum_{n=1}^{N} P(n) * \text{Impurity}(n), \tag{1}$$

where Impurity (Parent node) denotes the Gini measure for the impurity (i.e., non-homogeneity) of the parent node, and P(b) denotes the proportion of observations in the node assigned to branch b. The impurity measure, Impurity (node), may be defined as follows:

$$\text{Impurity (node)} = 1 - \sum_{i=1}^{I} \left(\frac{\text{number of class i cases}}{\text{all cases in the node}}\right)^{\wedge 2}. \tag{2}$$

When a node is ‘pure’, Eq. (2) gives the minimum value, and its value will be higher for less homogeneous nodes. If one considers the definition of “worth” according to Eq. (1), a split resulting in more homogeneous branches (Child nodes) will have more “worth”.

While developing a CART this criterion is applied recursively to the descendants to achieve Child nodes having maximum worth which, in turn, become the parents to successive splits, and so on. The splitting process goes on until there is no (or less than a pre-specified minimum) reduction in impurity and/or the limit for a minimum number of observations in a leaf node is reached. Following this process, a saturated tree is obtained. The saturated tree provides the best fit for the used database, but overfits the information contained within the database, and this overfitting does not help in classifying other databases. Therefore, when developing a CART model data is usually divided into two subsets: one for learning (or training) and the other for testing (or validation).

The learning sample is used to split nodes, while the testing sample is used to compare the misclassification. The saturated tree is obtained from the learning data. Overly large trees could result in higher misclassification when applied to classify new databases. To decrease its complexity, the tree is pruned in a second step according to the cost-complexity algorithm, which is based on removing the branches that add little to the predictive value of the tree. The cost-complexity measure combines the precision criteria as opposed to complexity in the number of nodes and processing speed, searching for the tree that obtains the lowest value for this parameter. The final step gives rise to the optimal tree. A more detailed description of the CART method can be found in Breiman et al. (1984).

2.3. Decision Rules (DRs)

The DT's structure can be transformed into rules in order to extract its potentially useful information. A DR is a logical, conditional structure of the type if *A*->*B*, in which *A* is the antecedent of the rule and *B* is the consequent, with all the splits of the parent nodes being the antecedent and the class of the terminal node being the consequent.

Each rule starts at the root node and each variable that is included in tree division makes an IF of the rule, which ends in terminal node with a value of THEN (which is associated with the state resulting from the terminal node). The class of a node is the status that shows the highest number of cases. Thus, a priori, the number of rules can be identified with the number of terminal nodes in the tree.

Due to the fact that the occurrence of crashes is infrequent in comparison with the non-occurrence of crashes, the class of the terminal node —and therefore the class resulting from the rule— will usually be the non-occurrence of an accident (ACC=NO). Notwithstanding, from the road safety perspective, the rules of interest are the rules involving crashes. To identify this type of rule, and following previous studies (Montella et al., 2012; López et al., 2014), we use the posterior classification ratio (PCR) in order to re-assign a response class (the consequent) to each rule extracted. PCR compares the classification of the terminal nodes of the tree with the classification of the root node (Eq. 3):

$$PCR(j|t) = \frac{p\ (j|t)}{p(j|t_{raiz})} \quad (3)$$

where:

p (j|t) = Proportion of observations in node “t” that belong to the class “j”, where class “j” is “YES”; t_{root} = Root node of the tree.

The assignment of the class to each rule was performed selecting the class j* with the greatest value of PCR. In addition, we will analyze only rules in which the consequent of the class variable (ACC) is the accident occurrence (YES). For each rule, then, two parameters are calculated: Support and Probability of accident (in three years´ time).

Support: The support of the rule (*S*) is the percentage of the data set for which both *A* (antecedent) and *B* (consequent) appear, that is, the number of cases in which the following rule is fulfilled:

$$S(A \rightarrow B) = \frac{|(A \rightarrow B)_t|}{N} \quad (4)$$

Where $(A \rightarrow B)_t$ is the number of crashes for which both conditions A and B are verified; N is the total number of crashes.

Probability of accident (in 3 years): Indicates the probability that an accident will occur in three years´ time as a consequence of the circumstances given in the rule.

$$\text{Prob. accident 3 years} = \text{PCR} * \text{Prob. global acc} \quad (5)$$

Where $\text{Prob. global acc} = \frac{\text{crashes}}{km\ network}$

Because of the large number of patterns considered, DTs may suffer from an extreme risk of Type I error, that is, of finding patterns that appear only by chance to satisfy constraints on the sample data (Webb, 2007). To reduce the risk of Type I error, and following other authors (Montella et al., 2012; Kashani and Mohaymany, 2011), the dataset was split randomly in two parts: a training set (70%) and a testing set (30%). The rules extracted on the training set were validated using the testing set. The application of the tree structure obtained in the training set to the testing sample produced the testing tree that was used for validation. To reduce the risk that results were overfitted to the sample, at each node of the testing tree the assignment of the class was compared with the assignment performed in the training tree. As a result, only nodes with the same class in both the training and the testing trees were validated.

2.4. Decision Rules obtained from a Decision Tree: The global DRs set

The extraction of knowledge with DRs extracted from a DT has some limitations. The rules depend on the DT's structure because they are extracted from each tree branch from the root node to the terminal node. Therefore, knowledge is extracted only in the direction dictated from the root node to the terminal node even if other possible important rules could exist. To extract all the possible patterns from a particular data set, Abellán et al. (2013) proposed a method called information root node variation (IRNV). The main characteristic of the IRNV method is that a set of DTs is built by varying the root node. Thus, every possible set of DRs is obtained from each tree, providing a set of rules with potentially useful information.

The first step in order to obtain DRs from the different DTs built varying root node was to randomly split the dataset into the training set (70% of the data) and the testing set (30%). Then, based in the IRNV method, a total of 15 models of DTs varying the root node are developed (i.e., a different model for each one of the seven variables and the eight RSD-SE considered for studying). All the rules in which the consequent is the occurrence of the accident are extracted from these models. The main problem with this method is that most rules are extracted from DTs to which the root node has been imposed, and this node could not be essential for the pattern that describes the rule. To overcome this issue a procedure of verification of the root node is performed (López et al., 2014) to determine whether the rule should be simplified.

DR is the rule extracted from a DT in which the root node is imposed (called in this study the extended rule); and DR^{-} is the rule without the root node (called in this study the simple rule); A is the antecedent of the DR and is formed by n variables (X'1, X'2, …, X'n); A^{-} is the antecedent of the DR^{-} and is formed by n-1 variables (X'2, …, X'n). In this way, we have to compare DR: A (X'1, X'2, …, X'n) → B vs. DR^{-}: A^{-} (X'2, …, X'n) → B, where B is the consequent. The extended rule (rule with n items) is selected over a simple rule (rule with n-1 items) if it verifies two conditions (López et al., 2014):

$$\text{Condition 1: } \frac{PCR(A \to B)}{PCR(A^{-} \to B)} \geq 1.03 \tag{6}$$

$$\text{Condition 2: } \frac{S(A \to B)}{S\,(A^{-} \to B)} \geq 0.2 \tag{7}$$

Condition 1 establishes that the increase of PCR in the DR should be over 3%; and Condition 2 indicates that the support of the DR with respect to the DR^{-} should be, at least, 20%. Thus, the global DRs set is formed by extended rules (DRs) when conditions 1 and 2 are verified simultaneously, or simple rules (DRs^{-}) if one of the conditions is not verified.

Once the simple rule or the extended rule has been selected, the chosen rules are validated in the testing set. The PCR is calculated again, and the rules fulfilling PCR ≥1 (the rules whose consequents are the occurrence of the accident) are the validated rules. The rules that are validated become part of the final set of RDs and should be analyzed from the road safety standpoint.

3. RESULTS

In the first step, the dataset was randomly split into training and testing sets: 1,174 km formed the training set, having 738 road sections with crashes and 10,989 road sections without crashes.

The different models of DTs are built varying the root node using the training set. DT_1 is the model obtained directly, without imposing the root node, whereas DT_2 to DT_{15} are the models obtained varying the root node. Table 3 shows the main results for the 15 models. 106 rules were extracted from the different models. From these rules, only rules with PCR≥1 in the training set (i.e., rules whose consequent is the occurrence of the accident) are selected (62 rules).

[Insert here Table 3]

[Insert here Table 4]

In following, the root node is verified. This verification is only necessary for patterns obtained from DT_2 to DT_{15} (DTs in which the root node was imposed). Rules for DT_1 do not call for such verification because they do not come from a DT whose root node is imposed. The procedure compares the extended rule (DR) and the simple rule (DR^{-}). Altogether, 62 rules were analyzed: 20 rules are *DR* (verify conditions 1 and 2 simultaneously), and all the others (42) are DR^{-} (do not verify one condition).

Finally, the rules were validated in the testing set and a total of 61 rules were obtained. Given that most of the rules are simplified (they are DR^{-}) some appear more than once. After this process, only 17 rules remain forming the global DR set.

Table 4 shows the rules grouped in four sets to show their common patterns. In the first group (three rules), the rules only have one RSD-SE as a deficiency; in the second group (five rules), the rules are formed by RSD-SE with some deficiencies and geometric or environmental variables; in the third group (six rules), the rules are formed by RSD-SE without deficiencies (RSD-SE=NO) and geometric or environmental variables; and in the fourth group (three rules), the rules are formed by RSD-SE (with or without deficiencies) and geometric or environmental variables.

Table 4 shows the values of PCR and the probability of accident (in three years) in the training and testing set (for each rule). The average probability of accident in the network is 6.28%.

Rules in the first group show a direct relationship between some signaling deficiencies and the occurrence of crashes. Rule 1 shows that the incomplete removal of road works markings (RSD-SE 7=Y) presents a probability of accident between 20% (value in the training set) to 22% (values in the testing set). This translates as an increased probability of the order 220% to 254% (see value of PCR) with regard to the mean values of accident probability in the network analyzed (6.28%).

Table 4 shows that RSD-SE7 is present in the rules entailing a greater probability of accident (rules 1, 4, 5, 8 and 16). Rules 4 and 5, in which there are incomplete removal of road works markings, on roads with AADT less than or equal to 5000 veh/day, have similar values for probability of accident (between 18% and 23%). With the same values of AADT, if the terrain is flat or rolling, the probability of accident is between 21% and 25% (see Rule 8). Rule 16 reflects another pattern for roads with the RSD-SE7. Although it does not involve deficiencies with warning signs (RSD-SE3=N), the values of probability of traffic crashes are also high, varying between 20% and 25%. Some researches (Cafiso et al., 2007; Miller, 1992) have described the involvement of deficient road marking in crashes, showing that their improvement is likely to be cost-effective. For example, on roads with edge lines missing, a relative increase in injury accident risk of 8% could be assumed; and when the center line is missing the risk increases to 13% (Safety Audits of Existing Roads, 2003). Ellis and Pyeon (2006) indicated that pavement work markings not properly removed may confuse or distract drivers. Alteration of pavement color and/or texture, as well as incomplete removal of pavement markings, has been identified as a particular problem for motorists; they can be mistaken for navigable lanes through construction work zones. Because motorists or drivers heavily rely on pavement markings for roadway guidance, it is imperative to remove old markings to reduce crashes owing to lane confusion (Cho et al., 2012).

Rule 2 shows that lack of correspondence between vertical signs and road markings (RSD-SE6) presents a direct relationship with crashes. When this deficiency appears on the analyzed roads, the probability of accident reaches 14% to 21%. This means an increased probability that is 127% to 235% greater than the mean values for accident probability. According to the results, several road safety problems identified with the lack of correspondence between vertical signs and road markings can be tied to accident risk. The study by Antov (2011) highlights problems stemming from missing, contradictory or difficult to read signs/marking, but further typical deficiencies are incomplete or misguiding signs/road markings, or an “overload” of information.

Rule 3 shows a direct relationship between crashes and instances when the guide sign does not exist or it is in an incorrect position (RSD-SE4). In this case, the probability of accident is similar to Rule 2 (between 15% and 21%). This stands as an increased probability ranging from 135% to 238% beyond the mean values. Some investigations reveal the importance of vertical signs on traffic crashes. Cafiso et al. (2007) evaluated two-lane rural highways and established that regulatory signs, such as speed limits, could affect road safety by conveying essential information on safe behavior. For missing or ineffective signs, the relative risk factor was assumed as equal to 20%. Croft and Schnerring (2009) likewise established that poorly located or incorrect signs could lead to a confusing and ambiguous situation, increasing crash risk.

Rules 6 and 7 show the relationship between deficiencies in RSD-SE 2 and crashes on road with AADT higher than 5,000 veh/day, and with road width between 5 and 6.5 meters. Rule 7 shows that the probability of accident varies from 20% to 21%, meaning respective increased probabilities of the order 215% to 237% over the mean values. The influence of regulatory signs on accident occurrence has been investigated in two-lane rural highways by some researchers. Cafiso et al. (2007) reported explanations of the relative increase in accident risk for some safety issues as vertical sign, determining that regulatory signs such as for speed limits could affect road safety by conveying essential information on safe behavior.

Rule 6 the same pattern, adding deficiencies in road markings (not exist or were deleted - RSD-SE5). In this case, the probability of an accident in 3 years is around 20%. This would be an increased probability between 214% 217% greater than the average values for probability of accident in the network. Previous studies have shown that incorrect or poorly maintained pavement markings can lead to incorrect placement or maneuvers of vehicles, and increase the risk of crashes (Croft and Schnerring, 2009).

In the third group, the rules are formed by RSD-SE without deficiencies and geometric or environmental variables. In Rule 14 the probability of accident is similar to the probability of accident of the network analyzed (6.28%). The values of probability in Rules 9, 11 and 12 range between 6.49% and 14.89%. Only in two rules (Rules 10 and 13) the probabilities are greater, increasing to values of 17%-21%; these rules are identified on roads with high values of AADT (>5,000 veh/day) and roadway width between 5 and 6.5 meters. Such findings underline that two-lane rural highways with major traffic flow (AADT) entail an increased risk of accident.

In the fourth group, the rules are formed by RSD-SE (with or without deficiencies) and geometric or environmental variables. As seen for Rules 15 and 17, when there are no deficiencies with the RSD-SE7 (incomplete removal of road work markings) on roads with values of AADT between 1,000 to 5,000 veh/day, even if other elements fail, the probability of accident is low, very similar to the probability of the network.

4. CONCLUSIONS

This paper presents an analysis of deficiencies in signaling with regard to crashes on rural highways. In addition, some variables related to geometric and environmental road characteristic were used, and the Data Mining technique of Decision Trees was applied. In order to derive all the information possible from the database analyzed, different DT models were built, varying the root node, and from each of the models the DRs of interest were extracted (rules whose consequence is the occurrence of an accident). As a result, 62 rules were obtained, and 61 of them were validated. After elimination of the rules that were the same, a total of 26 rules made up the final set.

In order to perform a safety analysis of the rules, they are grouped in four sets: rules directly relating crashes with signaling deficiencies; rules relating crashes with deficiencies in signaling and roadway characteristics; rules that do not involve deficient signaling or highway characteristics, but under certain geometric and/or environmental conditions bear a relation with crashes; and rules that present deficiencies in some elements, in others no, and have geometric and/or environmental variables present.

In general, the element RSD-SE7 (incomplete removal of road works markings) appears in the rules with the greatest probability of accident in 3 years (Rules 4, 5, 8 and 16), producing in turn a greater probability of accident. RSD-SE7 appears with AADT less than or equal to 5,000 veh/day, which may indicate that the pavement markings are not properly re-established in this type of roadway. This result shows that the government agencies or local administration should verify that after construction is finished, the original conditions of a roadway must be re-established quickly and efficiently.
Deficiencies in RSD-SE4 (sign does not exist or it is not correctly situated) are also associated with a greatly increased probability of accident in the network analyzed. This finding serves to accentuate the importance of maintaining signaling. Indeed, it is recommended that administrations make vigilance and follow up of roadway signs and signals a priority, ensuring that they are in optimal conditions.

Acknowledgments

The authors would like to acknowledge FEDER funding by the European Union for financial support via project “Análisis de la relación entre Elementos Susceptibles de Mejora, Accidentes y TCA” of the “Programa Operativo FEDER de Andalucía 2007-2013”. We also thank the Public Works Agency and Regional Ministry of Public Works and Housing of the Regional Government of Andalusia. The authors are grateful to the Spanish General Directorate of Traffic (DGT) for providing the data necessary for this research. Griselda López wishes to express her acknowledgement to the regional ministry of Economy, Innovation and Science of the regional government of Andalusia (Spain) for a scholarship to train teachers and researchers in Deficit Areas.

REFERENCES

Abdelwahab, H., Abdel-Aty, M. (2001). Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transportation Research Record: Journal of the Transportation Research Board, (1746), pp. 6-13.

Abellán, J., G. López, J. de Oña (2013). Analysis of Traffic Accident Severity Using Decision Rules via Decision Trees. Expert Systems with Applications, 40, pp.6047-6054.

Alfredas Laurinavičius , Lina Juknevičiūtė-Žilinskienė , Kornelija Ratkevičiūtė , Ineta Lingytė , Laura Čygaitė , Vytautas Grigonis , Rasa Ušpalytė-Vitkūnienė , Dago Antov , Tiit Metsvahi , Zsuzsanna Toth-Szabo, András Várhelyi (2012) Policy instruments for managing road safety on EU-roads, Transport, 27:4, 397-404, DOI: 10.3846/16484142.2012.751934

Antov, D. (2011). Road Safety Inspection Guidelines and Checklists. BALTRIS, 43 pp.

Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984). Classification and Regression Trees. Belmont, CA: Chapman and Hall.

Cafiso, S., Cava, G., Montella, A. (2007). Safety index for evaluation of two-lane rural highways. Transportation Research Record: Journal of the Transportation Research Board, 2019, pp. 136-145.

Chang, L. Y., Wang, H. W. (2006). Analysis of traffic injury severity: An applicationof non-parametric classification tree techniques. Accident Analysis and Prevention, 38, pp.1019–1027.

Chang, L. Y., Chien, J. T. (2013). Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model.Safety science,51(1), pp. 17-22.

Croft, P., Schnerring, F. (2009). Guide to Traffic Management. Part 13: Road Environment Safety. Sydney, Australia: Austroads.

DaCoTA (2012) Roads, Deliverable 4.8q of the EC FP7 project DaCoTA

De Oña, J., López, G., Abellán, J. (2013a). Extracting Decision Rules from police accident reports through Decision Trees. Accident Analysis and Prevention, 50, pp. 1151–1160.

De Oña, J., López, G., Mujalli, R. O., Calvo, F. J. (2013b). Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. Accident Analysis and Prevention, 51, pp. 1–10.

De Oña, J., Mujalli, R. O., Calvo, F. J. (2011). Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. Accident Analysis & Prevention, 43(1), pp. 402-411.

EC. 2008. Directive 2008/96/EC of the European Parliament and of the Council of 19 November 2008 on Road Infrastructure Safety Management. European Commission. Available from Internet: http://eur-lex.europa.eu/LexUriServ/LexUriServ.

Ellis, R., Pyeon, J. (2006). Development of improved procedures for removing temporary pavement markings during highway construction. Proc., Transportation Research Board 85th Annual Meeting, Transportation Research Board, Washington, DC.

Kashani, A. T., Mohaymany, A. S. (2011). Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Safety science,49, pp. 1314-1320.

Kuhnert, P.M., Do, K.A., McClure, R. (2000). Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Computational Statistics & Data Analysis 34, pp. 371–386.

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. Transportation Research Record: Journal of the Transportation Research Board, 2432, pp. 133-141.

Miller, T.R. Benefit–cost analysis of lane marking. In Transportation Research Record, No. 1334, TRB, National Research Council, Washington, D.C., 1992, pp. 38- 45.

Ministerio del Interior (2013). Las principales cifras de la Siniestralidad Vial en España 2013. Dirección General de Tráfico, Madrid. Disponible en: http://www.dgt.es/Galerias/seguridad-vial/estadisticas-e-indicadores/publicaciones/principales-cifras-siniestralidad/Siniestralidad_Vial_2013.pdf

Montella, A., Aria, M., D'Ambrosio, A., Mauriello, F. (2012). Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. Accident Analysis and Prevention, 49, pp. 58–72.

Moore, D.N., Schneider IV, W.H., Savolainenb, P.T., Farzaneh, M. (2010). Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and nonintersection locations. Accident Analysis and Prevention, Accident Analysis and Prevention 43(3), pp. 621-630.

Ogden, KW 1996, Safer roads: a guide to road safety engineering, Avebury Technical, Aldershot, UK

Pakgohar, A., Tabrizi, R.S., Khalilli, M., Esmaeili, A. (2010). The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. Procedia Computer Science 3, pp. 764–769.

Regan, M.A., Trent, W.V., Lee, J.D., Young, K.L. (2009) Driver distraction injury prevention countermeasures – Part 3: Vehicle, technology and road design. In: Regan, M.A., Lee, J.D., Young, K.L. (Ed.), Driver Distraction: Theory, Effects and Mitigation, pp. 579-601. CRC Press, Boca Raton, Florida.

SETRA. 2008. Methodological Guide: Road Safety Inspections. Service d'etudessur les transports, les routes et leursa menagements (SETRA). 70 p. Available from Internet: http:// www.setra.equipement.gouv.fr/IMG/pdf/US_0928A_RSI_-_Road_Safety_Inspections.pdf

Siskind, V., Steinhardt, D., Sheehan, M., O'Connor, T., Hanks, H. (2011). Risk factors for fatal crashes in rural Australia. Accident Analysis & Prevention,43(3), pp. 1082-1088.

Sohn, S.Y., Shin, H.W. (2001). Data mining for road traffic accident type classification. Ergonomics 44, pp. 107–117.

Webb, G. I. (2007). Discovering significant patterns. Machine g significant patterns. Machine Learning, 68(1), 1-33 pp.

List of Tables:

List of Figures: