Extracting crash patterns involving vulnerable users on two-lane rural highways

By: Griselda López and Juan de Oña

Extracting crash patterns involving vulnerable users on two-lane rural highways / *Extracción de patrones de accidents en carreteras convencionales con usuarios vulnerables involucrados*

[1]Griselda López, [2]Juan de Oña

[1] TRYSE Research Group, Department of Civil Engineering, University of Granada, ETSI Caminos, Canales y Puertos, c/ Severo Ochoa, s/n, 18071 Granada (Spain), griselda@ugr.es

[2] TRYSE Research Group, Department of Civil Engineering, University of Granada, ETSI Caminos, Canales y Puertos, c/ Severo Ochoa, s/n, 18071 Granada (Spain), jdona@ugr.es

KEYWORDS: severity; crash analysis; vulnerable users; decision trees; CART; C4.5
(gravedad; análisis de accidentes; usuarios vulnerables; Árboles de Decisión; CART; C4.5)

ABSTRACT

This study analyzes the severity of crashes on two-lane rural highways involving vulnerable users, such as pedestrians, cyclists or motorcyclists. The main aim is to detect patterns in order to develop safety improvement strategies specifically focused on these users. Such patterns were identified using Decision Rules extracted from Decision Trees. The main findings indicate that pedestrian crashes on two-lane rural highways are associated with accidents entailing death or serious injury. To reduce the severity of traffic crashes involving cyclists, the administration should improve shoulder conditions. The patterns for motorcycle accidents show their severity to be influenced by atmospheric factors, gender and age of the driver, type of accident and the alignment of the road. The most severe patterns for motorcycles are associated with pedestrian accidents and run-off-the-road collisions when the alignment of the road is a signalized curve without speed limit.

RESUMEN

Este estudio analiza la gravedad de los accidentes que involucran a usuarios vulnerables en carreteras convencionales. El objetivo principal es detectar determinados patrones de accidentes, con el fin de desarrollar estrategias de mejora de seguridad vial, centradas específicamente en estos usuarios. Estos patrones han sido identificados utilizando Reglas de Decisión extraídas de Árboles de Decisión. Los resultados indican que los accidentes con peatones en carreteras convencionales de dos carriles se asocian con los accidentes más graves, es decir, accidentes mortales o con lesiones graves. Para disminuir la gravedad de los accidentes que involucran a ciclistas, las administraciones competentes deben mejorar las condiciones de los arcenes. Los patrones obtenidos para accidentes con motocicletas muestran que su gravedad está influenciada por factores atmosféricos, el género y la edad del conductor, el tipo de accidente y la alineación de la carretera. Además, los accidentes con motocicletas más graves son salidas de la vía, que involucran peatones y ocurren en tramos de carretera en el que la alineación es una curva señalizada y sin límite de restricción velocidad.

1. INTRODUCTION

Two-lane rural highways present higher crash incidence and crash injury rates than other types of roadways *(1*), and they represent a substantial proportion of the road network in most countries. Authors Wang et al. *(2)* highlighted that the fatal crash rate on rural highways is more than double that for urban roads, even though the rate for all rural highway crashes is just under half of that for urban highways.

In Spain, accidents on rural roads represent around 75% of the total, whereas only around 20% of accidents are produced on freeways, and the remaining 5% on motorways *(3)*. These numbers underline the necessity to improve safety on rural roads. In fact, the Spanish Road Safety Strategy 2011-2020 *(4)* identified safety improvement on two-lane rural highways as one of its priorities to reduce the socio-economic impact of highway crashes. Another highlighted priority is the protection of vulnerable users.

Numerous studies in the literature look at accidents involving motorcycle riders, who are more vulnerable than other vehicle drivers because of the lack of protection in the event of a crash *(5-9)*. In addition, pedestrian and motor vehicle crashes are a serious problem throughout the world, for which reason other studies analyze the severity of crashes with pedestrians *(10-12)*. Some studies examine the severity of cyclist involvement in traffic crashes *(13-15)*.

The present study has a dual focus, analyzing accidents that occur on two-lane rural highways and moreover involve vulnerable users. To this end Data Mining techniques have been applied, which have been receiving much attention on the part of road safety researchers *(16)*. More specifically, this paper applies Decision Trees (DTs), as they make it possible to draw crash patterns that are easy to understand for safety managers.

While DTs in general have been widely applied in road safety literature, the CART method is the one most used *(6, 12, 17-19)*. There are other algorithms, such as ID3 and C4.5, that likewise can be used to build DTs. Recently, De Oña et al. *(18)* compared the results of all these algorithms, and found that CART and C4.5 lead to better results than ID3. Therefore, in this study CART and C4.5 algorithms will be used to build DTs.

A relevant advantage of DTs as compared to other methods with similar aims is that the structure of a DT allows for the extraction of Decision Rules (DRs). DRs provide an easy manner of identifying crash patterns. In view of these patterns, safety managers can more readily establish specific countermeasures to improve road safety.

The paper is structured as follows: section 2 shows the methodology used to conduct the analysis, with a description of DTs and crash data. In section 3 the results and discussion are described, and finally, the conclusions are reported.

2. MATERIALS AND METHODS

2.1. DECISION TREES

A DT is a predictive model that can be used to represent both classifiers and regression models. DTs are popular due to their simplicity and transparency; moreover, they are usually presented graphically as hierarchical structures, making them easy to interpret.

Nodes and branches form DTs. There are three types of nodes: root nodes, decision nodes (or intermediate nodes) and leaves (or terminal nodes). A Root node contains all the data. Decision nodes gather a test of a particular attribute. Ultimately, to classify an unlabeled instance, the case is routed down the tree according to the values of the attributes tested in successive decision nodes, and when a leaf is reached, the instance is classified according to the probability distribution over all classification possibilities. Thus, leaves are the terminal nodes of the tree and they specify the ultimate decision of the tree.

The DT is typically constructed by means of a "divide-and-conquer" approach. Thus, the first step is to select an attribute that will serve as a root node of the tree. This root node splits up and divides the dataset into different subsets, one for every value of the root node. A branch specifies each value. The construction of the tree becomes a recursive problem, since the process can be repeated for every branch of the tree. It should be noted that only those instances that actually reach the branch are used in the construction of the tree. In order to determine which attribute to split on, given a set of examples with different classes, different algorithms can be used (i.e. ID3, C4.5, CART).

Any algorithm can be used to fit a tree to a sample using recursive partitioning. The sample is split into increasingly homogeneous subsets until the leaf node contains only cases from a single class, or until the stopping criterion is reached.

Sample cases are recursively subdivided into segments at each stage of subdivision, a segment is divided according to an explanatory variable that is selected based on a specific criterion. The explanatory variable giving the highest value for the criteria is chosen from among all explanatory variables at each division. Such division continues until all cases in each segment have the same class, or until the stopping criterion is reached.

2.1.1. CART

The CART method builds binary Decision Trees. This algorithm uses the Gini Index as the splitting criterion. Depending on the nature of the dependent variable, CART develops classification trees (discrete target variable) or regression trees (continuous target variable).

The development of a CART model generally comprises three steps: (1) tree growing (2) pruning process (3) optimal tree selection from the pruned trees. Tree growing consists of recursively partitioning the target variable to maximize ''purity'' in the two child nodes. By definition, the terminal nodes present a low degree of impurity compared to the root node. In the tree growing stage, predictors generate candidate partitions (or splits) at each internal node of the tree, so that a suitable criterion must be defined to choose the best split of the objects. Gini reduction criteria measure the ''worth'' of each split in terms of its contribution toward maximizing homogeneity through the resulting split. If a split results in splitting of one parent node into B branches, the ''worth'' of that split may be measured as follows:

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

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

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

When a node is “pure”, this measure (Eq. 1) will have a minimum value, and its value will be higher for less homogeneous nodes. If one considers the definition of “worth” (Eq. 2), 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. The splitting process continues until there is no reduction in impurity and/or the limit for the minimum number of observations in a node is reached. Hence, a saturated tree is obtained, providing the best fit for the data used. However, this overfitting does not help in accurately classifying another data set. Therefore, to develop a CART model, the data is usually divided into two subsets, one for training and the other one for testing. The training sample is used to split nodes, while the testing sample is used to compare any misclassification. The saturated tree is constructed from the training data.

Overly large trees could result in higher misclassification when used to classify new data sets. A tree is therefore pruned in the second step to decrease its complexity. Pruning is performed according to the cost-complexity algorithm, which is based on removing 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 last step therefore leads to the optimal tree. A more detailed description of the CART method can be found in Breiman et al. *(20)*.

2.1.2. C4.5

When using the C4.5 algorithm *(21)*, the splitting criterion is the gain ratio, a criterion based on information theory. The information conveyed by a message about an event depends on the probability of the event; it can be measured in bits as minus the logarithm to base 2 of that probability. The information within a message that a random case belongs to a certain class is given as:

$$-log_2[\frac{freq(C_i,T)}{|T|}]\ bits \tag{3}$$

where T is a sample of cases, C_i is class i, and $\text{freq}(\text{C}_\text{i}, \text{T})$ is the number of cases in T that belong to class C_i.

In this way, the expected amount of information *info(T)* from such a message pertaining to T (also called entropy) is measured as follows:

$$info(T) = -\sum_{i=1}^{K}\{\frac{freq(C_i,T)}{|T|} x log_2[\frac{freq(C_i,T)}{|T|}]\} \tag{4}$$

Entropy can also be measured after T has been partitioned into n sets using the outcome of a test carried out on attribute *X*:

$$info_x(T) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} x info(T_i) \tag{5}$$

With these two measurements, the gain criterion used in conjunction with the ID3 algorithm *(22)* can be defined as follows:

$$gain(X) = info(T) - info_x(T) \tag{6}$$

The gain criterion measures the information gained by partitioning the training set using test X. We should stress that this gain criterion has an implicit preference for splitting nominal attributes with many values. Therefore, it produces trees that discard the remaining attributes prematurely, because they soon come to branches that have only a few cases. As an improvement upon the ID3 algorithm, Quinlan *(21)* introduces the C4.5 algorithm, where the gain criterion is replaced by a Gain Ratio criterion.

The gain ratio *(22)* is obtained by normalization, where gain(X) is divided by the potential information that can be generated by division X:

$$split\ info(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} x log_2 \left(\frac{|T_i|}{|T|}\right) \tag{7}$$

Accordingly, the gain ratio is defined as:

$$gain\ ratio(X) = \frac{gain(T)}{split\ info(X)} \tag{8}$$

Any DT built following this procedure would overfit the data. To avoid overfitting, pruning strategies can be used, simplifying the tree by discarding one or more subtrees and replacing them with leaves. The algorithm incorporates pruning once a tree has been induced, by applying a hypothesis test on whether or not to expand a branch. A more detailed description of the C4.5 algorithm can be found in Quinlan *(21)*.

2.2. DECISION RULES

After the decision tree is constructed, the tree can be easily turned into a rule set by deriving a rule for each path in the tree that starts at the root and ends at the leaf node. The rules conform a logical-conditional structure of the type “IF (X) → THEN (Y)”, where X is the antecedent (formed by a set of statuses of several attribute variables); and Y is the consequent (formed by only one state of the class variable).

A priori, the number of rules can be determined by the number of terminal nodes on the tree. However, in order to extract significant rules that would provide useful information for the implementation of road safety strategies in the future, three parameters and the minimum threshold are used in each possible “X→Y” rule:

- Population (Po) is the percentage of the dataset where “X” appears.
- Support (S) is the percentage of the dataset where “X & Y” appear.
- Probability (P) is the percentage of cases in which the rule is accurate (i.e. P=S/Po expressed as percentage).
- Lift (L) 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.

The threshold values for the parameters (Po, S, P and Lift) are normally selected in light of the following characteristics: nature of the data balanced or unbalanced; significant interest in fatal crashes (rare events); and sample size —small or large datasets *(23)*. In this study, the threshold values considered are *(18, 23)*: Po≥1%; S≥ 0.6%; P≥60%; Lift≥1.2.

Due to the large number of patterns considered, DTs may run an extreme risk of type I error that is, finding patterns apparently owing to chance alone that satisfy constraints on the sample data *(24)*. To reduce the risk of type I error, the data set was randomly split into a training test and a test set *(12, 17; 19)*.

2.3.DATA

Accident data were obtained from the Spanish General Traffic Accident Directorate (DGT) for two-lane rural highways in Andalusia over a period of seven years (2003–2009). This study included only rural highways with two lanes (one for each direction) and accidents involving one vehicle and vulnerable users (pedestrians, motorcycles or bicycles). The total number of such accidents was 3,225.

The variable under study is accident severity. Following previous studies *(17; 19, 25)* severity is defined based on the worst injury sustained in the accident, and two levels are established: slight injury (SI) and accidents where persons were killed or seriously injured (KSI).

To identify the main factors that affect the accident severity of vulnerable users, 19 variables were analyzed (see Table 1). These variables describe characteristics related to the driver (age and gender); accident (month, time, day, occupants involved, accident type and cause); road (alignment, safety barriers, pavement width, lane width, shoulder type, paved shoulder, pavement markings and sight distance); vehicle (vehicle type); and context (atmospheric factors and lighting). Table 1 offers a description of the variables together with the frequency distribution.

Variables: Description	Code	Severity: KSI / SI		Total
AGE: Age	>=18	100	103	203
	[19-25]	317	375	692
	[26-45]	787	855	1642
	[46-65]	247	306	553
	>=66	59	76	135
ACT: Accident type	FO: Fixed objects collision	21	30	51
	OT: Other	49	96	145
	PED: Collision with pedestrian	489	290	779
	RO: Rollover	146	240	386
	ROR: Run-off-road without collision	372	613	985
	ROR_CO: Run-off-road with collision	433	446	879
ALI: Alignement	CH: Curve heavy	92	104	196
	CHS: Curve heavy with sign speed	152	161	313
	CHWS: Curve heavy without sign speed	248	232	480
	CS: Curve smooth	256	306	562
	INT: Intersection	177	208	385
	TG: Tangent	585	704	1289
ATF: Atmospheric factors	GW: Good weather	1420	1578	2998
	HR: Heavy rain	12	28	40
	LR: Ligth rain	42	74	116
	OT: Other	36	35	71
BAR: Safety Barriers	N: No	1025	1221	2246
	Y: Yes	485	494	979
CAU: Cause	COM: Combination of factors	105	130	235
	DF: Driver factors	1244	1345	2589
	OT: Other	141	173	314
	RF: Road factors	10	50	60
	VF: Vehicle factors	10	17	27
DAY: Day	APH: After public holiday	124	149	273
	BPH: Before public holiday	273	260	533
	PH: Public holiday	513	537	1050
	WD: Working day	600	769	1369
LAW: Lane width	MED: [3.25–3.75] m	1088	1236	2324
	THI: <3.25 m	388	423	811
	WID: >3.75 m	34	56	90
LIG: Lighting	DAY: Day	825	1101	1926
	DUS: Dusk	73	65	138
	NIL: Insufficient (night time)	122	113	235
	NSL: Suficient (night time)	81	79	160
	NWL: Without lighting (night-time)	409	357	766
MON: month	AUT: Autumn	352	378	730
	SPR; Spring	402	465	867
	SUM; Summer	380	476	856
	WIN: Winter	376	396	772
OI: ocuppant involved	[1]: 1 occupant	1199	1351	2550
	[2]: 2 occupants	225	284	509
	[>2]: >2 occupants	86	80	166
PAS: paved shoulder	N: No, non existent or impassable	556	658	1214
	Y: Yes	954	1057	2011
PAW: pavement width	MED: [6–7] m	474	605	1079
	THI: < 6 m	174	202	376
	WID: > 7 m	862	908	1770
ROM: road marking	DE: Does not exist or was deleted	83	115	198
	SLD: Separate lanes and define road margins	1198	1371	2569
	SLO: Separate lanes only	44	42	86
	SMR: Separate margins of roadway	185	187	372
GEN: gender	F: Female	133	224	357
	M: Male	1377	1491	2868
SHT: shoulder type	MED: [1.5–2.5] m	196	222	418
	NE: Non-existent or impassable	611	739	1350
	THI: <1.5 m	703	754	1457
SID: sight distance	ATM: Atmospheric	16	28	44
	BUI: Building	6	4	10
	OT: Other	48	34	82
	TOP: Topography	253	265	518
	VEG: Vegetation	9	10	19
	WR: Without restriction	1178	1374	2552
TIM: time	[0-6]	274	247	521
	(6-12]	363	444	807
	(12-18]	449	606	1055
	(18-24]	424	418	842
VEH: vehicle type	BIC: Biclycles	60	83	143
	CAR: Cars	389	221	610
	OT: Other	15	7	22
	PTW: Mortorbikes and motorcycles	1016	1382	2398
	TRU: Trucks	30	22	52

Table 1. Variables, values and classification by severity.

3. RESULTS

The model was built using Weka (26), an open source freeware available at: http://www.cs.waikato.ac.nz/ml/weka/.

The model accuracy (that is the percentage of cases correctly classified) was very similar with CART and C4.5: respectively, 59.4% and 58.1%. These accuracy values lie within the range obtained in previous studies applying classification methods with similar objectives. Abdel Wahab and Abdel-Aty (27) obtained 61% accuracy when they applied Bayesian networks and 58.1% accuracy with neural networks; De Oña et al. (25) obtained 58%, 59% and 61% applying Bayesian networks with different algorithms. In De Oña et al. (18), the accuracy was 55.8%, 54.2% and 52.7% using different algorithms to build DTs.

Figure 1 shows the DT built using the CART method. The number of nodes, the total number of accidents in each node, and the node classification based on the 2 categories (SI and KSI) are indicated for each node.

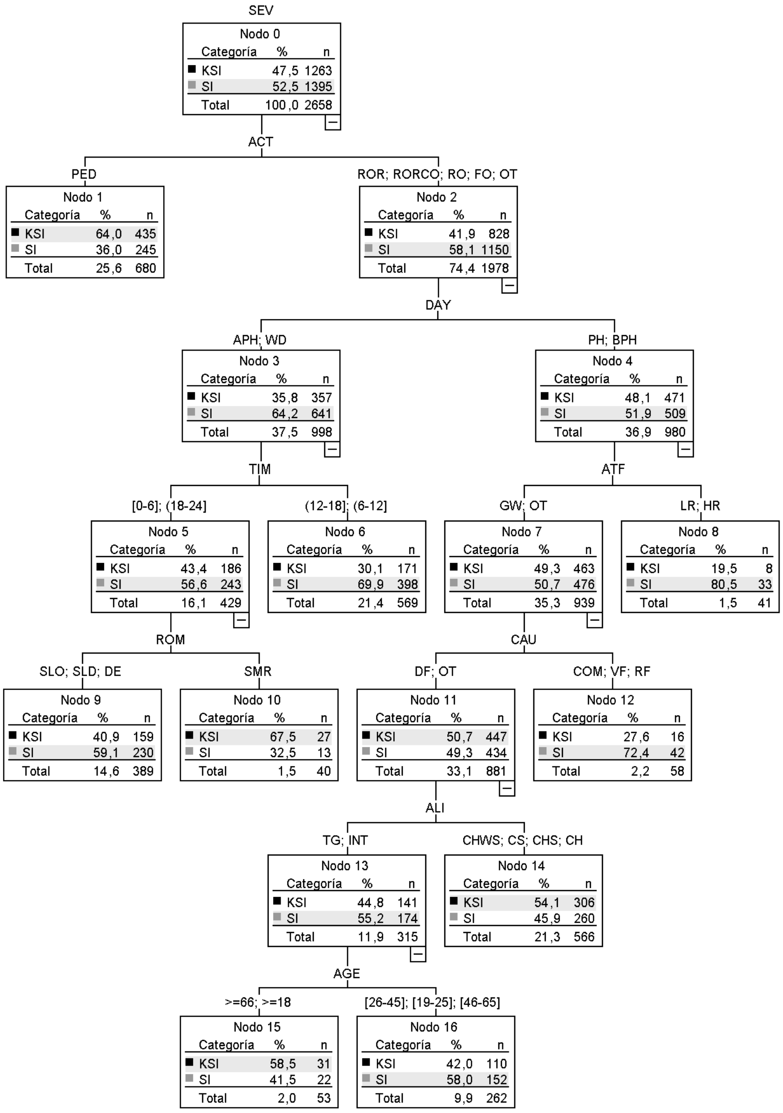

Figure 1. DT´structure using CART.

With CART, seven variables are used as predictors of the tree: accident type (ACT), day (DAY), time (TIM), atmospheric factors (ATF), cause(s) of the accident (CAU), alignment of the road (ALI) and age (AGE). Sixteen nodes form the tree, nine of them terminal nodes. The first splitting variable is ACT.

The root node splits into two branches (nodes 1 and 2). Node 1 shows accidents with pedestrians, classified as KSI with 64% probability. For all the other accidents, the tree is divided by DAY. If DAY is a working day or just after a public holiday, the tree splits using TIM. Node 6 shows SI accidents between 6 a.m. and 6 p.m. with a probability of 69.9%. When accidents occur during the rest of the day (6 p.m. – 6 a.m.), and depending on ROM, accidents are classified according to the degree of severity (see Figure 1): if road markings separate margins of the road, accidents are KSI, with a probability of 67.5% (node 10); in any other case accidents are SI, with a probability of 59.1% (node 9).

When DAY is a public holiday or the day before a public holiday, the tree splits by ATF. When ATF are bad (light or heavy rain), accidents are SI, with a probability of 80.5% (node 8). With good atmospheric factors or others, the tree divides by CAU. One terminal node appears when CAU is a combination of vehicle or road factors (node 12). In this case, accidents are SI with a probability of 72.4%. For other causes the tree splits by ALI.

When ALI is a curve (node 14), accidents are KSI (54.1% of probability). However, when alignment is a tangent or intersection, and depending on the driver´s age, two terminal nodes appear with different degrees of severity: node 15 identifies KSI accidents for very young and old drivers (with a probability of 58.5%), whereas node 16 shows SI accidents for drivers between 19 and 65 years of age (probability is 58%).

Figure 2 shows the structure of the DT built for C4.5: there are 39 nodes from the DT, 30 of them being terminal nodes. In this case, the DT predictors are nine variables: vehicle type (VEH), paved shoulder (PAS), atmospheric factors (ATF), gender (GEN), accident type (ACT), sight distance (SID), age (AGE), safety barriers (BAR) and alignment of the road (ALI). In this case, the first splitting variable is VEH.

The root node is divided into five nodes (nodes 1 to 5, see Figure 2). When VEH is a car, truck or other type, accidents are classified as KSI with probabilities of 65%, 82.4% and 58% (nodes 2, 3 and 5, respectively). When the vehicle involved is a bicycle, the severity depends on the variable PAS. In that case, two terminal nodes are obtained: SI accidents on a shouldered road with 61.1% probability (node 7), and KSI accidents on road having no shoulder, or an impassable road, with 58% of probability (node 6).

Most of the tree is formed by accidents involving motorcycles (node 1). Thus, depending on ATF, four nodes are created and three of them are terminal nodes. Under bad weather conditions, nodes 9, 10 and 11 are created. All of them involved SI accidents, with respective probabilities of 74.3%, 75.8% and 52.3%. When weather conditions were good the tree splits by GEN. For females, accidents are classified as SI (55.2%, node 12). For males, the tree grows depending on ACT. Then, four terminal nodes are obtained: accidents with pedestrians are classified as KSI with 60.8% of probability (node 14), whereas rollover (node 17), collision with fixed objects (node 18) and other types of accidents (node 16) are SI. For run-off-the-road, with or without collision, the tree keeps on growing.

When accidents are run-off-the-roads without collision, and depending on SID, the following nodes are obtained: four terminal nodes involve KSI accidents when SID is restrained by building, atmospheric factor or other (nodes 25, 23 and 22, respectively). In contrast, regardless of whether SID is restrained by topography or not restrained, accidents are SI (nodes 24 and 20). In the case of SID restrained by topography, the severity of the accident depends on safety barriers. If there are safety barriers, accidents are more severe than if no barriers are present (node 26 vs. node 27).

When accidents are run-off-the-roads with collision, and depending on AGE, the results are the following: KSI accidents for young drivers (node 30) entail a probability of 60.4% and drivers between 26 and 45 years (55.3% of probability in node 29); SI accidents for drivers aged 46-65 (node 32) and older drivers (node 28) have probabilities of 59.7% and 52%. Finally, for drivers between 19 and 25 years of age, the tree split by ALI. When the alignment of the road is a tangent or a smooth curve (node 36), accidents are SI. Accidents are also SI in intersections or in heavy curves (nodes 35 and 37). This result underlines that in particular locations drivers increase their precautions. However, in signaled heavy curves, whether or not the speed limit is marked, accidents are KSI (nodes 37 and 34). In this case, drivers are more confident and the severity of the accident increases.

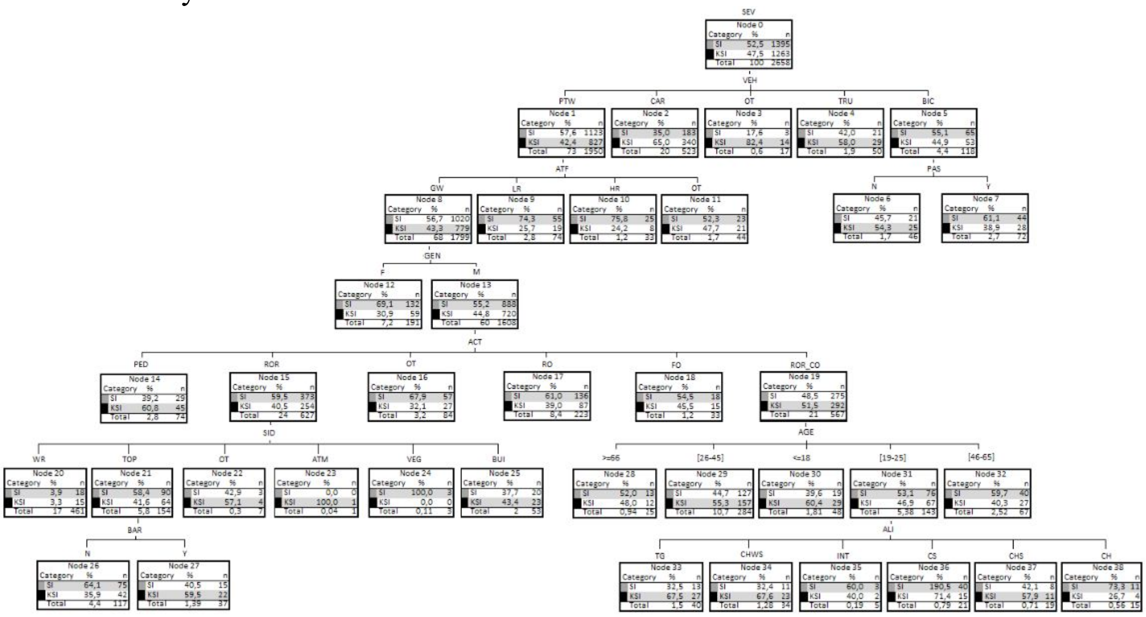
SEV
Node 0 Category % n SI 52,5 1395 KSI 47,5 1263 Total 100 2658
VEH
PTW
Node 1 Category % n SI 57,6 1123 KSI 42,4 827 Total 73 1950
CAR
Node 2 Category % n SI 35,0 183 KSI 65,0 340 Total 20 523
OT
Node 3 Category % n SI 17,6 3 KSI 82,4 14 Total 0,6 17
TRU
Node 4 Category % n SI 42,0 21 KSI 58,0 29 Total 1,9 50
BIC
Node 5 Category % n SI 55,1 65 KSI 44,9 53 Total 4,4 118
ATF
GW
Node 8 Category % n SI 56,7 1020 KSI 43,3 779 Total 68 1799
LR
Node 9 Category % n SI 74,3 55 KSI 25,7 19 Total 2,8 74
HR
Node 10 Category % n SI 75,8 25 KSI 24,2 8 Total 1,2 33
OT
Node 11 Category % n SI 52,3 23 KSI 47,7 21 Total 1,7 44
PAS
N
Node 6 Category % n SI 45,7 21 KSI 54,3 25 Total 1,7 46
Y
Node 7 Category % n SI 61,1 44 KSI 38,9 28 Total 2,7 72
GEN
F
Node 12 Category % n SI 69,1 132 KSI 30,9 59 Total 7,2 191
M
Node 13 Category % n SI 55,2 888 KSI 44,8 720 Total 60 1608
ACT
PED
Node 14 Category % n SI 39,2 29 KSI 60,8 45 Total 2,8 74
ROR
Node 15 Category % n SI 59,5 373 KSI 40,5 254 Total 24 627
OT
Node 16 Category % n SI 67,9 57 KSI 32,1 27 Total 3,2 84
RO
Node 17 Category % n SI 61,0 136 KSI 39,0 87 Total 8,4 223
FO
Node 18 Category % n SI 54,5 18 KSI 45,5 15 Total 1,2 33
ROR_CO
Node 19 Category % n SI 48,5 275 KSI 51,5 292 Total 21 567
SID
WR
Node 20 Category % n SI 3,9 18 KSI 3,3 15 Total 17 461
TOP
Node 21 Category % n SI 58,4 90 KSI 41,6 64 Total 5,8 154
OT
Node 22 Category % n SI 42,9 3 KSI 57,1 4 Total 0,3 7
ATM
Node 23 Category % n SI 0,0 0 KSI 100,0 1 Total 0,04 1
VEG
Node 24 Category % n SI 100,0 3 KSI 0,0 0 Total 0,11 3
BUI
Node 25 Category % n SI 37,7 20 KSI 43,4 23 Total 2 53
BAR
N
Node 26 Category % n SI 64,1 75 KSI 35,9 42 Total 4,4 117
Y
Node 27 Category % n SI 40,5 15 KSI 59,5 22 Total 1,39 37
AGE
>=66
Node 28 Category % n SI 52,0 13 KSI 48,0 12 Total 0,94 25
[26-45]
Node 29 Category % n SI 44,7 127 KSI 55,3 157 Total 10,7 284
<=18
Node 30 Category % n SI 39,6 19 KSI 60,4 29 Total 1,81 48
[19-25]
Node 31 Category % n SI 53,1 76 KSI 46,9 67 Total 5,38 143
[46-65]
Node 32 Category % n SI 59,7 40 KSI 40,3 27 Total 2,52 67
ALI
TG
Node 33 Category % n SI 32,5 13 KSI 67,5 27 Total 1,5 40
CHWS
Node 34 Category % n SI 32,4 11 KSI 67,6 23 Total 1,28 34
INT
Node 35 Category % n SI 60,0 3 KSI 40,0 2 Total 0,19 5
CS
Node 36 Category % n SI 190,5 40 KSI 71,4 15 Total 0,79 21
CHS
Node 37 Category % n SI 42,1 8 KSI 57,9 11 Total 0,71 19
CH
Node 38 Category % n SI 73,3 11 KSI 26,7 4 Total 0,56 15

Figure 2. DT´structure using C4.5.

3.1.DECISION RULES

Decision Tree structures can be transformed into Decision Rules. To extract significant and useful rules (that might provide useful information for the future implementation of road safety strategies), parameters population, support, confidence and lift with the minimum threshold were calculated. Tables 2 and 3 respectively show the significant rules extracted from the training set in the case of CART and C4.5.

CART identifies just five significant rules: three referring to SI accidents and two with respect to KSI. The minimum probability is 64% (rule 1). Population ranges between 1.5% (rules 8 and 10) and 25.6% (rule 1). All the rules have support higher than 0.6%. Lift varies between 1.3 (rule 6) to 1.5 (rule 8).

NODE	RULES: IF…	THEN	Po(%)	S(%)	P(%)	Lift
6	ACT≠PED AND DAY=(APH OR WD) AND TIM= ((12-18] OR (6-12])	SEV= SI	21.4	15.0	69.9	1.3
8	ACT≠PED AND DAY=(BPH OR PH) AND ATF= (LR OR HR)	SEV= SI	1.5	1.2	80.5	1.5
12	ACT≠PED AND DAY=(BPH OR PH) AND ATF= (GW OR OT) AND CAU=(COM OR VF OR RF)	SEV= SI	2.2	1.6	72.4	1.4
1	ACT=PED	SEV= KSI	25.6	16.4	64.0	1.4
10	ACT≠PED AND DAY=(APH OR WD) AND TIM= ([0-6] OR (18-24]) AND ROM= SMR	SEV= KSI	1.5	1.0	67.5	1.4

Table 2. Rules extracted from tree built with CART

In turn, C4.5 provides 11 significant rules: seven of them for SI accidents and four rules for KSI accidents (see Table 3). Probability ranges between 60.4% (rule 30) and 75.8% (rule 10). All the rules have populations higher than 1%. Support varies from 0.9% (rules 10 and 34) to 12.8% (rule 2); lift ranges from 1.2 (rules 7 and 17) to 1.4 (rules 2, 9, 10 and 34).

NODE	RULES: IF…	THEN	Po(%)	S(%)	P(%)	Lift
33	VEH = PTW & ATF = GW & GEN=H & ACT=ROR_CO & AGE= [19-25] & ALI=TG	SEV= SI	1.5	1.0	67.5	1.3
17	VEH = PTW & ATF = GW & GEN=H & ACT=RO	SEV= SI	8.4	5.1	61.0	1.2
16	VEH = PTW & ATF = GW & GEN=H & ACT=OT	SEV= SI	3.2	2.1	67.9	1.3
12	VEH = PTW & ATF = GW & GEN=F	SEV= SI	7.2	5.0	69.1	1.3
9	VEH = PTW & ATF = LR	SEV= SI	2.8	2.1	74.3	1.4
10	VEH = PTW & ATF = HR	SEV= SI	1.2	0.9	75.8	1.4
7	VEH = BIC & PAS=Y	SEV= SI	2.7	1.7	61.1	1.2
14	VEH = PTW & ATF = GW & GEN=H & ACT=PED	SEV= KSI	2.8	1.7	60.8	1.3
30	VEH = PTW & ATF = GW & GEN=H & ACT=ROR_CO & AGE= <=18	SEV= KSI	1.8	1.1	60.4	1.3
34	VEH = PTW & ATF = GW & GEN=H & ACT=ROR_CO & AGE= [19-25] & ALI=CHWS	SEV= KSI	1.3	0.9	67.6	1.4
2	VEH = CAR	SEV= KSI	19.7	12.8	65.0	1.4

Table 3. Rules extracted from tree built with C4.5

The next step consists of validating these rules using the test set. Validation entails verifying the rule classification, and the parameters with their minimum threshold. Finally, taking into account the severity of the rules, eight validated rules were obtained for SI accidents (see Table 4); whereas four rules were validated for KSI accidents (see Table 5). All the rules obtained by means of CART (see Table 2) were validated using the test set, while only seven rules were validated for C4.5.

NODE	RULES: IF…	METHOD	Po(%)	S(%)	P(%)	Lift
33	VEH = PTW & ATF = GW & GEN=H & ACT=ROR_CO & AGE= [19-25] & ALI=TG	C4.5	1.5	1.0	67.5	1.3
16	VEH = PTW & ATF = GW & GEN=H & ACT=OT	C4.5	3.2	2.1	67.9	1.3
12	VEH = PTW & ATF = GW & GEN=F	C4.5	7.2	5.0	69.1	1.3
9	VEH = PTW & ATF = LR	C4.5	2.8	2.1	74.3	1.4
7	VEH = BIC & PAS=Y	C4.5	2.7	1.7	61.1	1.2
6	ACT≠PED AND DAY=(APH OR WD) AND TIM= ((12-18] OR (6-12])	CART	21.4	15.0	69.9	1.3
8	ACT≠PED AND DAY=(BPH OR PH) AND ATF= (LR OR HR)	CART	1.5	1.2	80.5	1.5
12	ACT≠PED AND DAY=(BPH OR PH) AND ATF= (GW OR OT) AND CAU=(COM OR VF OR RF)	CART	2.2	1.6	72.4	1.4

Table 4. SI rules validated.

Table 4 shows that C4.5 provides just one pattern for bicycles, the rest relating to motorcycles. Rule 7, which involved 2.7% of the population, refers to accidents with bicycles on the road with a paved shoulder; in this case, accidents are SI (with a probability of 61.1%). In contrast, for accidents produced on the road where shoulders were not paved or inexistent (node 7 in Figure 2), crashes were KSI. From a safety point of view, these results demonstrate that the severity of these accidents depends on the shoulder of the road. That is, severity is greater on roads where the shoulder is not paved or does not even exist. These findings indicate that reducing the severity of accidents involving cyclists and cars on rural road calls for intervention by the local administration to improve shoulder conditions on roads frequently used by cyclists.

Meanwhile, patterns with motorcycles are related with atmospheric factors. Rule 9 shows SI accidents give a probability of 74.3% when accidents happen with light rain. Similar results are found under other bad weather conditions (see Figure 2, heavy rain in node 9 and other in node 11). In line with De Oña et al. *(18)*, this result demonstrates that drivers try to be very careful under poor atmospheric conditions.

The rest of the patterns involving motorcycles tend to imply good weather conditions and depend on the driver´s gender. Gender is a key factor in the severity of accidents with motorcycles. Rule 12 shows SI accidents for female drivers. This rule has the highest value of population (7.2%) and support (5%), with a high probability (69.1%). For male drivers, four patterns are obtained, two of them are SI accidents. Rule 16 identifies SI accidents with motorcycles produced under good weather for male drivers when the type of accident is “other”. Rule 33 also identifies SI accidents when the accident type is a run-off-the-road with collision for drivers 19-25 years old and alignment is tangent. Good weather conditions could in fact be a contributing factor in motorcycle crashes, related with driver behavior (and driving speeds).

CART identifies SI patterns without pedestrians (run-off-road, collision with fixed objects, rollover, other types). Rule 6 identifies this kind of accident on working days, or directly after public holidays, between 6 a.m. and 6 p.m. Rules 8 and 12 identify SI accidents on public holidays or immediately before public holidays. In rule 8, SI accidents occur under bad weather conditions (light or heavy rain) with a probability of 80.5%. This result again suggests that drivers are more careful under difficult atmospheric conditions. In view of rule 12, SI accidents occur when weather conditions are good or other, and accident causes stem from the vehicle, road or a combination of factors (probability is 72.4%).

NODE	RULES: IF…	**METHOD**	Po(%)	S(%)	P(%)	Lift
14	VEH = PTW & ATF = GW & GEN=H & ACT=PED	C4.5	2.8	1.7	60.8	1.3
34	VEH = PTW & ATF = GW & GEN=H & ACT=ROR_CO & AGE= [19-25] & ALI=CHWS	C4.5	1.3	0.9	67.6	1.4
1	ACT=PED	CART	25.6	16.4	64.0	1.4
10	ACT≠PED AND DAY=(APH OR WD) AND TIM= ([0-6] OR (18-24]) AND ROM= SMR	CART	1.5	1.0	67.5	1.4

Table 5. KSI rules validated.

As seen in Table 5, each method identifies two KSI patterns. The C4.5 algorithm leads us to identify patterns involving motorcycles and male drivers. Rule 34 identifies accidents under good weather conditions when the alignment of the road involves a signalized curve without a speed limit. In this case, accidents are KSI with a probability of 67.6%. Some studies show that the severity of motorcycle accidents increases on roads with straight and curved grades *(6, 15)*.

Rule 14 shows KSI accidents for motorcycles under good weather conditions for males when accidents involve a pedestrian (probability of 60.8%). Pedestrian accidents produced by motorcycles are associated with KSI accidents, as reported in previous studies *(28, 29)*. CART identifies another KSI pattern for accidents with pedestrians (rule 1): in this case, the relationship between pedestrian and KSI accident is direct (probability of 64%).

Finally, rule 10 shows KSI accidents according to the type of accident (run-off-road, rollover, collision with fixed object, other), on working days or the day after public holidays, in the time period from 6 p.m. to 6 a.m., on roads where pavement marking separates only the margin, with a probability of 67.5%.

4. CONCLUSIONS

This study attempts to identify certain patterns on rural roads involving vulnerable users (pedestrians, cyclists and motorcycle riders). The patterns have been identified using two different types of DTs (CART and C4.5). Later, in order to identify patterns easy to understand by road safety managers, we have extracted some validated rules from the DTs.

It is known that CART and C4.5, based on different splitting criteria and respectively producing binary and non-binary trees, were successfully used in the past to analyze crash severity, providing more reliable results than methods such as ID3. The accuracy of the models obtained by means of both the above methods was very similar, and agrees with previous studies. CART provides binary DTs. Therefore, different categories of the variables are grouped in the branches, increasing node support, yet making it impossible to analyze the influence of a specific category on severity, complicating the interpretation of results overall. In turn, C4.5 generates a branch for each category, thus enabling one to look at the influence of all the variables on severity. In short, C4.5 generates DTs with more branches than CART, and it therefore produces more rules and a lower level of support. The most noteworthy conclusion, in this case, is that not all the rules are finally validated.

From a safety viewpoint, the main findings can be summed up as follows:

- Pedestrian crashes are more frequent in urban areas than in rural areas. However, this study highlights that pedestrian crashes in rural areas also call for safety measures because of their severity. One possible countermeasure to mitigate this safety issue would be to implement perceptual cues, such as gateways and traffic calming devices, in segments of rural highways where there is high pedestrian activity.
- As cyclist crashes on roads having no paved shoulder are more severe, to mitigate the severity of this type of accident, specific actions should be adopted for road margins where the volume of cyclists is high.
- The main factors associated with the severity of motorcycle accidents are atmospheric conditions, gender and age of driver, type of accident, and alignment of the road. Interestingly enough, most severe or fatal crashes

associated with motorcycles occur under good weather conditions, a pattern possibly related to driver behavior and speed. Thus, electronic speed monitors could produce significant safety benefits in the face of potential crashes. In addition, when the drivers are males, accidents are run-off-the-road with collision and they occur in curves that are signalized, but do not specify speed limit, are associated with KSI accidents. In these cases, specific road markings or signals could be implemented to reduce the severity of accidents.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Spanish Ministry of Economy and Competitiveness for his support (Research Project TRA2012-37823). The authors express their gratitude to the Spanish General Directorate of Traffic (DGT) for supporting 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 their scholarship to train teachers and researchers in Deficit Areas.

REFERENCES

[1] Peek-Asa, C., Zwerling, C., Stallones, L., 2004. Acute traumatic injuries in rural populations. American Journal of Public Health 94(10), pp. 1689−1693.

[2] Wang, Y., Nguyen, N., Levy, A., Wu, Y-J., 2008. Cost Effective Safety Improvements for Two- Lane Rural Roads. Transportation Northwest Regional Center X, T Report TNW2008-04, Seattle.

[3] DGT, 2011. Las principales cifras de la siniestralidad vial. España 2010. Traffic General Directorate, Madrid. Available in:
http://www.dgt.es/was6/portal/contenidos/es/seguridad_vial/estadistica/publicaciones/princip_cifras_siniestral/cifras_siniestralidadl013.pdf

[4] DGT, 2011. Spanish Road Safety Strategy 2011-2020. Traffic General Directorate, Madrid, 222p.

[5] De Lapparent, M., 2006. Empirical Bayesian analysis of accident severity for motorcyclists in large French urban areas. Accident Analysis and Prevention 38, 260–268.

[6] 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.

[7] Quddus, M.A., Noland, R.B., Chin, H.C., 2002. An analysis of motorcycle injury and vehicle damage severity using ordered probit models. Journal of Safety Research 33, pp. 445–462

[8] Rifaat, S.M., Tay, R., de Barros, A., 2011. Severity of motorcycle crashes in Calgary. Accident Analysis and Prevention, 49, 44-49

[9] Savolainen, P., Mannering, F., 2007. Probabilistic models of motorcyclists' injury severities in single and multi-vehicle crashes. Accident Analysis and Prevention 39, pp. 955–963

[10] Giacomo, C., Rasmussen, T., Kaplan, S., 2014. Risk Factors Associated with Crash Severity on Low-Volume Rural Roads in Denmark. Journal of Transportation Safety & Security, 6, pp.1-20

[11] Lyon, C., and B. Persaud, 2002. Pedestrian Collision Prediction Models for Urban Intersections. In Transportation Research Record: Journal of the Transportation Research Board, No. 1818, Transportation Research Board of the National Academies, Washington, D.C., pp. 102–107.

[12] Montella A., Aria M., D'Ambrosio A., Mauriello F., 2011. Data Mining Techniques for Exploratory Analysis of Pedestrian Crashes. Transportation Research Record 2237, pp. 107–116.

[13] Boufous, S., De Rome, L., Senserrick, T., Ivers, R.Q., 2012. Risk factors for severe injury in cyclists involved in traffic crashes in Victoria. Accident Analysis and Prevention 49, pp. 404–409.

[14] Kim, J.K., Kim, S., Ulfarsson, G.F., Porrello, L.A., 2007. Bicyclist injury severities in bicycle–motor vehicle accidents. Accident Analysis and Prevention 39 (2), pp. 238–251.

[15] Klop, J.R., Khattak, A.J., 1999. Factors influencing bicycle crash severity on two-lane, undivided roadways in North Carolina. Transportation Research Record 1674, pp. 78–85.

[16] Pande, A., Abdel-Aty, M., 2009. Market basket analysis of crash data from large jurisdictions and its potential as a decision supporting tool. Safety Science, 47, pp. 145–54.

[17] Chang, L.Y. and Wang, H.W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, pp. 1019–1027.

[18] De Oña, J., López, G., Abellán, J., 2013. Extracting decision rules from police accident reports through decision trees. Accident Analysis and Prevention 50, pp. 1151–1160.

[19] Kashani, A., Mohaymany, A., 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.

[20] Breiman, L., Friedman, J., Olshen, R., Stone, C. Classification and Regression Trees. Chapman & Hall, Belmont, CA, 1984

[21] Quinlan, J. R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, California.

[22] Quinlan, J.R., 1986. Induction of decision trees. Machine Learning, 1(1), pp. 81-106.

[23] Abellán, J., López, G., De Oña, J., 2013. Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Systems with Applications, 40, pp. 6047-6054.

[24] Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD 1993), pp. 207–2166.

[25] De Oña, J., Mujalli, R.O., Calvo, F.J., 2011. Analysis of traffic accident injury on Spanish rural highways using Bayesian networks. Accident Analysis and Prevention 43, pp. 402-411

[26] Witten, I.H., Frank, E. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. Morgan Kaufmann Publishers, San Francisco, CA, 2005.

[27] Abdel Wahab, H.T., Abdel-Aty, M.A., 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections.TransportationResearchRecord1746, pp. 6–13.

[28] Perandones, J.M., Molinero, A., Martin, C., Mansilla, A., Pedrero, D., 2008. Recommendations for location of motorcyclist protection devices in Spanish regional road network of Castilla y Leon. In: Presented at the 87th Annual Meeting of the Transportation Research Board, Washington, D.C.

[29] Daniello, A., Gabler, H.C., 2011. Fatality risk in motorcycle collisions with roadside objects in the United States. Accident Analysis and Prevention, 43, pp. 1167–1170.