IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received September 3, 2020, accepted September 24, 2020, date of publication October 1, 2020, date of current version October 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3028043

Driver Pattern Identification in Road Crashes in Spain

ALMUDENA SANJURJO-DE-NO[1], BLANCA ARENAS-RAMÍREZ[1], JOSÉ MIRA[2], AND FRANCISCO APARICIO-IZQUIERDO[1]

[1] Instituto Universitario de Investigación del Automóvil Francisco Aparicio Izquierdo (INSIA-UPM), Escuela Técnica Superior de Ingenieros Industriales (ETSII), Universidad Politécnica de Madrid (UPM), 28006 Madrid, Spain

[2] Statistics Department, Escuela Técnica Superior de Ingenieros Industriales (ETSII), Universidad Politécnica de Madrid (UPM), 28006 Madrid, Spain

Corresponding author: Almudena Sanjurjo-de-No (almudena.sanjurjo.no@gmail.com)

This work was supported in part by the "AUTOMOBILE VEHICLE SAFETY, for smart, sustainable, safe and inclusive mobility" - SEGVAUTO 4.0-CM (P2018/EMT-4362) Program of the Community of Madrid which has contributed to its development and dissemination and in part by the Instituto Universitario de Investigación del Automóvil Francisco Aparicio Izquierdo (INSIA UPM), which provided, by means of a Grant, the funding for the Ph.D.

ABSTRACT Extracting driver collision patterns by gender and age regarding offences, collision type and injury severity is very useful in road safety, providing a better understanding on behavior of the different driver groups. Self-Organizing Map (SOM) is the tool proposed for distributing and projecting 145,904 drivers according to 8 offence variables on a 2D map. Thus, drivers who are close in the original 8D space (one dimension per offence variable), will remain so in the projected one (2D). Multivariate driving and collision patterns are explored to support the development of future measures to improve road safety. Tests of proportions are used for shedding light on clusters where driver offence is present. Finally, the SOM results were compared for validation with those of the standard K-Means clustering technique. The results show that the characteristics of road crashes and the severity of injuries depend jointly, i.e., in multivariate (pattern) terms, on gender, age, type of collisions and offences. There are relevant multivariate driver behavior differences in both the type of collisions (and therefore their severity) and the type and number of offences with regard to gender and age of the driver. This research unveils different multivariate driver behavior patterns, providing information about their relative importance (proportion), which helps in road policy decision making in terms of development of prevention measures. The results help in decision making through a potentially better allocation of resources as carried out by road safety regulating offices such as the Spanish Traffic General Directorate (Dirección General de Tráfico, DGT).

INDEX TERMS Age, driver behavior, gender, pattern recognition, self-organizing maps (SOM).

I. INTRODUCTION

In the literature, extracting vehicle collision patterns among different groups of drivers mainly concerning gender, age, both combined and regarding to driver offences, type of collisions or injury severity, has been a purpose of many researchers in recent years.

In the past few decades, the presence of women on the road has increased notably compared to men drivers [1]–[4], and [5]. Therefore, the number of female drivers involved in vehicle collisions has also increased in this period [2], [4], and [5].

Regarding gender, the majority of works have found significant differences between both genders in aspects such as crash rates, injury severity, the type of offence committed and the perception of driving skills. In general, male drivers are more involved in vehicle collisions (especially fatal ones) than females [1], [3]–[8], and [9].

Men have higher crash rates and greater exposure on the road than females [1], [7], [10], and [11]. Notwithstanding, exceptions to this generalization have been found, such as those observed by [12], which is striking given the use of the number of miles traveled in the denominator of the rate, versus other exposure measures.

As for age, researchers highlight that crash rates are higher among younger [8], [9], [11], and [13] and both younger and older drivers [1], [10], [12], and [14], although this also depends on the exposure [15].

The analysis by gender and age shows their joint influence on crash rates and their consequences as pointed out by [1],

The associate editor coordinating the review of this manuscript and approving it for publication was Xiangxue Li.

and [8]. In [8] it has been observed that gender differences in crash rates tended to disappear or even to be reversed (females commit more offences than males) among younger drivers.

The differences observed in the characteristics of drivers by gender and age also depend on factors such as the type of collision [8], [10], and [16].

Regarding the injury severity of drivers, researchers have found that women are more vulnerable to vehicle collisions than men [3], [17], and [18] and injury severity depends on both gender and age. In [19] it has been observed that the risk of injury, which goes from mild to severe, also varies according to age and thus it was observed that among young drivers, men present a higher risk, whereas among older drivers, this happens in women. Moreover, [17] concluded that the mortality risk of women was higher than that of men in ages between 20 and 35 years, stating that women have higher probability of dying from physical impact.

Concerning the perception of driver skills and type of offence committed by gender and age, men tend to perceive a lower level of risk in most situations (not only in driving environments) [7], and [20]. Thus, male drivers take more risks than women, especially younger drivers, who, in general, seek more excitement, drive more aggressively and are more inexperienced [1], [7]–[10], and [11]. Moreover, men tend to underestimate the degree of severity in the different dangerous driving actions, such as the influence of alcohol [21], and this increases the likelihood that they will exhibit higher risk behaviors [1], and [7]. In addition, men, especially younger ones, tend to overestimate their driving skills [1], [5], [11], and [21], whereas women present more skills that reflect their positive attitude towards safety and traffic regulations [5], and [11]. However, they are more likely to be distracted and commit more perceptual errors [1], and [5].

In [22] it has been concluded that women stop driving before men, since they recognize and accept the decay from aging in their driving skills, so it could be expected that crash rates of older drivers would be influenced by gender.

To summarize, in the literature reviewed it is generally highlighted that vehicle collisions are higher in male drivers and among the group of younger (18-29 years) and older drivers (from 75 on). On the basis of the review literature found, driver behavior is not a trivial problem and the human factor is one of the main issues which contribute to vehicle crash occurrence. Therefore, there are multivariate features in the data and insights need to be gained on the phenomenon, analyzing the variables jointly.

The objective of this research is to extract driver behavior patterns in collisions regarding gender and age and considering offences committed, type of collisions and injuries, as well as the relative importance of these patterns (proportions). Patterns are multivariate features of the data, which may not be obvious a priori and can thus be unveiled by sophisticated machine learning tools, such as Self-Organizing Maps (SOM), which would imply a thorough methodological contribution. Thus, to this end, a joint analysis of a large number of driver-related variables will be carried out through the SOM methodology, which aims to provide more relevant and complex results than univariate (or bivariate) analyses, given that, as mentioned above, there are behaviors or patterns that only come to light when several variables are studied together. Additionally, a disaggregated analysis of only the most common types of offences was performed by means of a test of proportions in order to enhance the pattern identification process provided by SOM. Finally, the results obtained with SOM were compared for validation with those of the standard K-Means clustering technique.

II. DATABASE

Road crash data analysis is one of the major tasks in collision research and each country works to have a strong data system. In Spain, the road crash database is maintained by the Spanish Traffic General Directorate (Dirección General de Tráfico, DGT) since 1983, and contains data collected by police at the scenes of road crashes with casualties.

At present, the DGT keeps two road crash databases: the first one includes vehicle collisions from 1993 to 2013 (General Road Crashes database), and the second one (the ARENA database) from 2014 to 2018, which is a very short period for analysis. There exist some differences between them: in their procedure and the new variables collected and also, so far, that standardization of both databases has not been completed. Due to these important issues, the General Road Crashes database was selected to create the database for this research.

The database with records of all collisions between two vehicles in Spain in the period 2004-2013 includes, initially, data from 836,598 drivers, both on their characteristics (gender, age, disability, psychophysical circumstances…) and offences (speed offences, non compliance of the STOP signal…), as well as collision and environmental variables (type, location, day of the week…) and vehicle characteristics (color, year of registration…). Each record in the database corresponds to a single driver and the *ad-hoc* database has two records per collision, one for each driver.

To carry out this research the initial database was filtered to only maintain vehicle collisions between two passenger cars (head-on, off-set frontal, side and read-end collisions) in interurban areas, as shown in phase (I) of Fig. 1. Thus, the database was reduced to 146,162 drivers.

Subsequently, a process of debugging (phase II) the above mentioned filtered database was carried out with the aim of deleting erroneous records, as well as those involving drivers where information on the other driver was not found. The resulting final *ad-hoc* database has a total of 145,904 drivers. The complete treatment of the database (Fig. 1) was carried out with the R program, which is a free software environment for statistical computing and graphics [23].

From within all the variables in the ad-hoc database, all the offences and the ‘‘unfavorable conditions for driving’’ variables have been selected (potentially relevant variables), given that patterns by gender, age, type of collision and

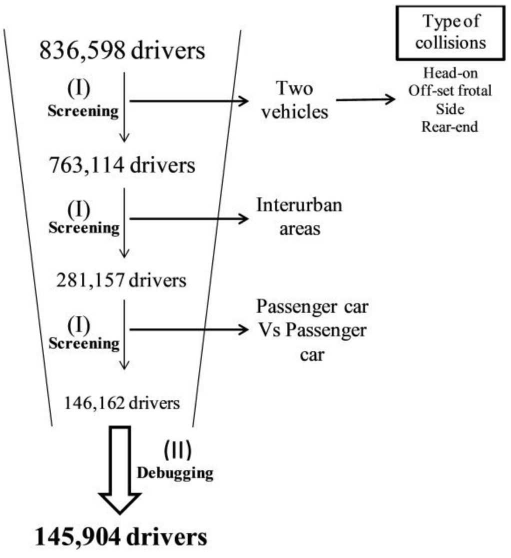


FIGURE 1. Screening and debugging procedure.

injury severity regarding the offences had to be obtained. The variables selected to build the SOM are shown in Table 1.

TABLE 1. Potentially relevant variables.

Potentially relevant variables	Types (description)
Driver offence	Distracted driving, Partial invasion of opposing lane, Illegal turn, Illegal passing maneuver, …
Speed offence	Inadequate speed under existent conditions, driving above the speed limits, and too slow driving disturbing traffic.
Administrative offence	Invalided driver license, Expired driver license, Not passing the roadworthiness test (MOT), …
Disability	Sights, Hearing, Lower - Upper limbs, …
Vehicle defect	Very worn out tyres, Missing tyre, Deficient front of rear lights, Deficient brakes, …
Alcohol/drugs use	These variables indicate: not respecting the limits of alcohol/drug during driving
Drowsiness	This variable indicates if the driver has or not drowsiness, fatigue or concern and it has been named "Drowsiness".
Sudden illness	Sudden illnesses may be defined as those which appeared unexpectedly and usually cause loss of standard abilities. Examples of sudden illnesses: passing out, epileptic seizure, heart attack, anxiety attack, …

As pointed out above, these variables will be related to gender, age, type of collision and injury severity, in order to obtain different driver patterns and their relative importance (proportion).

III. METHODOLOGY

To carry out this research, the Self-Organizing Maps (SOM) methodology has been applied to the data in order to perform a joint analysis of the variables which measure driver behavior and, thus, to obtain patterns by gender, age, type of collision and injury severity, as well as their relative importance (proportion). Additionally, a test of proportions has been carried out for the most common types of driver offences in order to shed light on clusters where this offence is present, thus as a complementary statistical tool, which strengthens the SOM pattern identification. Finally, the results obtained with SOM were compared for validation with those of the standard K-Means clustering technique.

The Self-Organizing map was developed by Kohonen around 1982 [24]; it is a very popular neural clustering technique, which can be included within Machine Learning. SOM uses an unsupervised training algorithm and goes through a process of self-organization, which is a competitive learning method that reduces data dimensionality [25], and [26] and the different nodes (clusters) in the map compete for the data assignment [27]. The map is initialized at random so that no prior knowledge is imposed on the mapping.

The aim of the SOM "hard" clustering methodology is to represent and cluster multidimensional data sets in a much smaller space, typically 2D. The SOM technique produces a 2 or 3D map, with pairs or triplets of integers defining the map nodes, e.g. (2, 3). All sample points from the original data, in a much larger number of dimensions, are allotted to a specific map node. A so-called weight, which is a vector in the original space, is assigned to each node as its centroid.

As pointed out by [28], SOM is a dimensional reduction mapping in the sense that it quantifies and represents a high dimensional space on a discrete map of a low dimension, preserving as much as possible the initial topology of the data. Thus, points that are close in the original space will remain so in the reduced dimension one.

The great advantage of dimensionality reduction is to produce a clustering which, due to its 2 or 3D visualization, is very rapidly assimilated by the analyst, which in turn allows for identifying patterns more clearly and rapidly [25], [28]–[32], and [33].

The SOM algorithm is developed in four stages: initialization, competition, cooperation and adaptation [31], and [34]. In a nutshell, the algorithm which produces the SOM maps is sequential: in each iteration, a new sample point is allotted to the closest point in the map as measured in the original space, after which the weights of the winning node and those of its neighborhood are updated. The update process, or cooperative learning step, is essential to preserve the topology of the high dimensional data. This does not exist in K-Means clustering, where no dimensional reduction is carried out.

The SOM clustering technique has been applied, as observed in the literature, to different fields. In road collisions, although their applications are more limited, a few relevant works have been found. In the research of [35], SOM has been applied with the purpose of unveiling specific regional traffic patterns. An analytical model has been developed by [32] to learn about the assignment of road collision liability in Taiwan. Finally, one should mention the work by [36] who studied pedestrian crashes applying clustering techniques with the purpose of identifying patterns which would help to design preventive measures.

On the other hand, K-Means is the best known non-hierarchical clustering methodology, which belongs to unsupervised learning techniques and whose aim is to group the data into a number of clusters K previously specified by the researcher [37]. K-Means is an iterative algorithm [31], which starts with randomly assigning a centroid to each cluster. Once the full sample ''assignment to cluster'' process has been carried out for the first time, the process is repeated using the centroids obtained from the first full sample assignment as starting values. This full sample assignment is repeated again a third, fourth...time until the assignment of observations to clusters in the last iteration is the same as the one in the iteration before last.

More information regarding the K-Means and SOM methodology can be found in [31], and [34].

The choice of SOM is justified because, with this clustering, more information has been provided for a better understanding of the multivariate data, so driver patterns will be obtained for a better deeper insight on their behavior. This is an important methodological contribution because, as far as the authors know, such multivariate driver pattern identification has not been used in the literature and can be useful for decision making when their (pattern) proportions are considered.

IV. VALUES FOR CATEGORICAL VARIABLES

In order to apply the SOM methodology, all the variables selected have been transformed to binary or ternary versions, which indicate absence, presence or unknown of the corresponding offences or defects. Thus, by consensus, the value 0 was assigned to indicate the absence of the offence or unfavorable condition and the value 2 to indicate the presence of the offence. The two discrete values taken for non-offence or offence (0 and 2) are irrelevant, given that they are the same for all variables.

The problem with this transformation is found for cases in which it is unknown if one or more of the analyzed offences are present or not in the drivers. This lack of knowledge is generated when the policeman reporting in situ on the collision, either does not know the status of that variable in the driver, or does not record such information in the collision report. These cases could not be included but, as pointed out by [38], many important methodological issues remain relating, among others, to missing data, so it has been considered important to take into account in some manner, as is described below, these records, so the results are, at least, less biased.

To assign a numerical value to unknown cases, two hypotheses were established: (I) the value assigned must lie between 0 (absence of offences or unfavorable conditions for driving) and 2 (presence of offences or unfavorable conditions for driving) and (II) it is considered that if the police do not fill out the form or if they do not know this information, then it is more likely that either this offence or unfavorable driving condition were not present. Therefore, it is believed that the value that should be assigned to unknown variables should be closer to 0 than to 2.

To assess this issue and to study the sensitivity to these unknown values, a set of previous studies have been carried out, which have consisted on monitoring the effect on SOM results when different values are taken in cases in which the value of a variable is unknown and keeping the rest of variables fixed. Thus, the values 0.25, 0.5 and 1 have been tested for cases in which the value of the variable is unknown.

The comparison of the SOMs will be carried out two at a time and (taking into account that SOM is a clustering technique) two SOMs can be considered to be equal if the relative positions of the data are the same for both maps [39], that is, if the Euclidean distance between two drivers located on the first map is equal to the one between the same drivers on the other map.

The differences obtained from the three comparisons made (SOM0.25-SOM0.5 / SOM0.25-SOM1 / SOM0.5-SOM1) were standardized with the variability of the algorithm resulting from random initialization.

The above mentioned process, which has not been detailed in this article since it is not its main purpose, concluded that the choice of the value of 0.25, 0.5 or 1 for cases in which the value of the variable is unknown is not significant. Therefore, a value of 0.25 was taken.

The variable categorizations are shown in Table 2.

TABLE 2. Variable values.

Categories	Values
No offence / No defect	0
Unknown	0.25
Offence or defect	2

V. RESULTS

In this section it is explained how the SOM has divided the driver data in nodes (clusters) according to multivariate offence data, which the drivers have or not committed, so that driver records with similar multivariate characteristics (according to the offences) are shown in the same node or in close ones of the map.

The main aim of this research is to extract the maximum information in the 8 dimensional multivariable space about the drivers analyzed regarding the aforementioned factors (gender, age, type of collision and injury severity). By means of the SOM methodology, interesting complex patterns as well as relevant information about the relative importance (proportion of sample drivers) of these patterns are extracted and analyzed. This could be used to draw attention on important patterns on which it will be interesting to apply statistical inference in future works. The ‘‘hard’’ SOM approach applied here can also be described within a ‘‘descriptive statistics’’ framework which can nonetheless be highly sophisticated, as it is stated in well-known multivariate analysis references such as [40]. The only exception, in this article, to this non inferential approach is a set of hypothesis tests for proportions carried out to shed more light into some specific SOM nodes.

In the SOM only the variables regarding driver offences and defects have been included. Subsequently, a joint and sequential analysis with the variables gender, age, type of collision and injury severity will be performed. This is the strategy chosen for extracting the driver behavior patterns.

The distribution of the 145,904 drivers along the offences SOM is shown in Fig. 2.

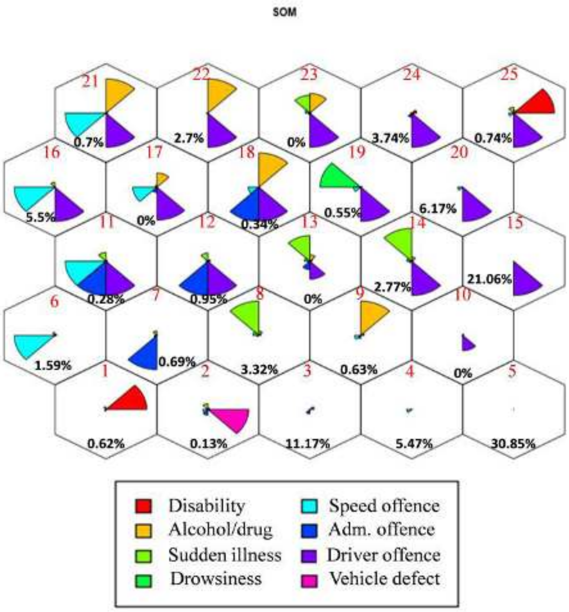


FIGURE 2. Offences SOM.

As shown in Fig. 2, the number of each cluster is indicated in red and the percentage of drivers that belong to each of them is indicated in black. The drivers were distributed over 25 clusters. The size of the SOM map (25 clusters) is decided as it is in most clustering procedures, by expert judgment-based sequential trial and error, in such way that the most useful patterns are more clearly identified. Therefore, it is determined empirically, reaching a trade-off between size (number of clusters), clarity and sample size per cluster. In this direction it is very important to take into account that with a very large map size it is not possible to extract patterns due to too few drivers in each cluster (the extreme case is that there is a cluster per driver) and with an excessively small map size the clusters are extremely heterogeneous and, therefore, adequate patterns cannot be extracted either.

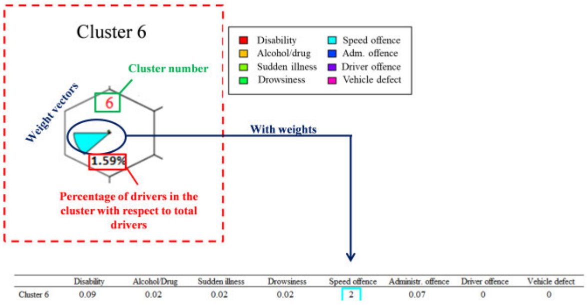
In the SOM map of offences, one may observe the driver characteristics while, as mentioned in the methodology section, preserving as much as possible the original space topology (8 dimensional in our case). A different color is used per (original) variable, as shown in Fig. 2.

Each circular sector within a cluster, which represents each of the variables introduced in the SOM with a different color, will be more or less large (in radius) depending on the average value (so-called weights) of the variable it represents, for all the drivers in the cluster. The radius will be maximum when either (a) the value of the variable in question for all drivers is 2, which means that all the drivers in the cluster have this offence or defect or (b) the average of the variable in this cluster is larger than any other one (cluster), whereas it will be minimal when the average is 0 (the circular sector is not represented for that variable) and, therefore, no driver in the cluster will have committed that offence or present the defect that the variable indicates. Table 3 shows the weights of all nodes, with the exception of nodes 10, 13, 17, and 23, which have no data points. The reason for maintaining the latter nodes in the map is that, although empty, they provide topological information to preserve distances. Note that the weights are already illustrated in Fig. 2 but in Table 3 their exact values are given.

TABLE 3. Average value (weight) of each variable in the different clusters.

SOM

	Disability	Alcohol/Drug	Sudden illness	Drowsiness	Speed offence	Administr. offence	Driver offence	Vehicle defect	Number of drivers
Cluster 1	2	0.02	0.02	0.02	0	0.09	0	0	906
Cluster 2	0.10	0.10	0.05	0.06	0.25	0.33	0	2	194
Cluster 3	0.22	0	0	0	0	0.25	0	0	16306
Cluster 4	0	0	0	0	0.25	0.03	0	0	7980
Cluster 5	0	0	0	0	0	0	0	0	45006
Cluster 6	0.09	0.02	0.02	0.02	2	0.07	0	0	2321
Cluster 7	0.02	0.03	0.03	0.07	0.14	2	0	0	1007
Cluster 8	0.10	0.24	0.27	0.29	0	0.03	0	0	4850
Cluster 9	0.17	2	0	0	0.32	0.18	0	0	925
Cluster 11	0.07	0.07	0.07	0.08	2	2	2	0.07	413
Cluster 12	0.13	0.05	0.06	0.08	0	2	2	0.05	1389
Cluster 14	0.09	0.25	0.26	0.25	0	0.04	2	0	4048
Cluster 15	0	0	0	0	0	0	2	0	30729
Cluster 16	0.02	0.04	0.04	0.04	2	0.03	2	0.03	8031
Cluster 18	0.05	2	0	0	0.52	2	2	0.04	497
Cluster 19	0.07	0	0	2	0.36	0.02	2	0.02	809
Cluster 20	0	0	0	0	0.25	0.02	2	0	9001
Cluster 21	0.09	2	0	0	2	0.03	2	0.02	1021
Cluster 22	0.03	2	0	0	0	0.04	2	0.03	3940
Cluster 24	0.25	0	0	0	0	0.24	2	0.12	5450
Cluster 25	2	0.17	0.04	0.02	0.25	0.02	2	0	1081



	Disability	Alcohol/Drug	Sudden illness	Drowsiness	Speed offence	Administr. offence	Driver offence	Vehicle defect
Cluster 6	0.09	0.02	0.02	0.02	2	0.07	0	0

FIGURE 3. Cluster 6.

In Fig. 3, one of the nodes of the SOM is zoomed out (cluster 6), representing a cluster / node of the map, to better illustrate the concepts explained above, which will facilitate the interpretation of the results obtained in each of the SOM clusters.

In Fig. 3 cluster 6 can be observed, where only speed offences have a significant importance (the average value equals 2).

As mentioned above, SOM provides information about the relative importance (proportion of sample size) of the different patterns identified. For example, driver offences, which include 52.51% of all drivers analyzed, appear in almost all clusters, but alone (without any other offence) only in clusters 15, 20 and 24, which account for 30.96% of all drivers. Therefore there is an important proportion of drivers for which driver offence is accompanied by another offence or defect (14.55% of all drivers), specially speed offences and alcohol or drugs use. These clusters, where more than one offence appear together, should be subject to special attention because some driver behaviors are unveiled only if many (8 in this case) offences or defects are jointly analyzed.

A. PATTERN IDENTIFICATION

The SOM was then applied to determine collision and offence patterns depending on gender (Fig. 4), age (Fig. 5), type of collision (Fig. 6) and injury severity (Fig. 7), by means of the disaggregated analysis of these factors along the SOM offence map.

The pattern identification process is going to be carried out sequentially, i.e: first only taking into account the SOM variables and gender, after that the above mentioned variables plus age, etc. This analysis will be presented sequentially for clarity, given that just showing the final step would be overwhelming for the reader.

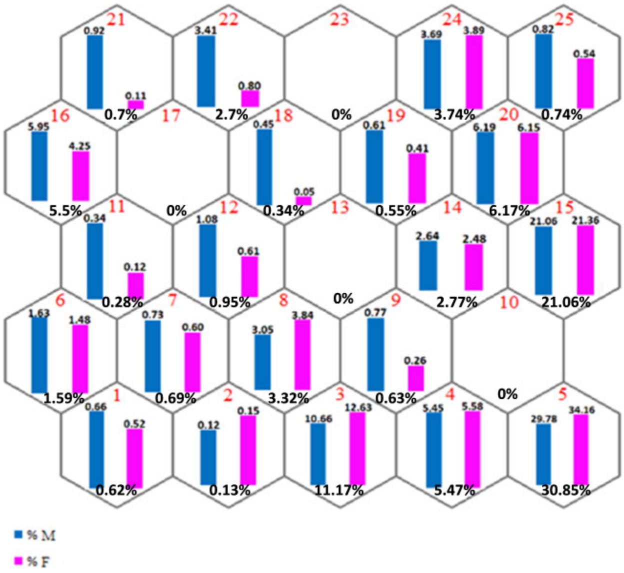


FIGURE 4. Drivers' distribution in the SOM map by gender.

1) PATTERN IDENTIFICATION BY GENDER

Fig. 4 shows the percentage of men and women who fall in each of the SOM clusters. It is very important to bear in mind that these percentages have been obtained regarding the total number of male and female drivers analyzed. As mentioned above, the percentages shown below within each node (cluster) are the drivers in each cluster (with respect to the total number of drivers) and add up to a hundred.

Based on the joint analysis of Fig. 2 and Fig. 4, it was concluded that males are more predisposed towards committing offences than females. In particular, the presence of male drivers is noticeable when compared to those of women, when more than one of the offences analyzed in the SOM occur simultaneously. Moreover, it is observed that female drivers are more represented in no-offence clusters, such as clusters 3 or 4 and in the 5th cluster, which includes all drivers who have not committed any offence. Therefore it seems that women commit fewer offences than men, as already pointed out by [8], and [10], although it is necessary to consider additional information specially his/her exposure.

When speed offences are taken into account, the percentage of males increases when compared to female's, as pointed out by [1]. These differences are significantly accentuated when the latter offences appear together with any others, such as driver offences or alcohol/drug use (clusters 11, 16 and 21).

A similar situation is found when alcohol/drug use is analyzed, where this condition appears much more frequently in males than in females (clusters 9, 18, 21 and 22), as revealed by authors like [21]. In addition, it has been observed that, as was the case with speed offences, these differences are accentuated when this condition is present jointly with other offences, especially driver and speed (clusters 21 and 22). Therefore, according to the results, it can also be concluded that the joint occurrence of multiple offences is much more frequent among male drivers.

Regarding physical defects (clusters 1 and 25) it has been observed that they are slightly more present among male than female drivers. This could be because men tend to overestimate their driving skills, as pointed out by authors such as [22], which would imply that, despite the physical defects they presented or the cognitive deterioration that they suffered as a consequence of age, they had been driving for a greater number of years than women.

Regarding driver offences, in Fig. 2 and Fig. 4 it has been observed that if they appear together with any other offences or defects, then there are more males than females and the differences in their behavior are, in general, larger than when driver offences appear alone. The only clusters in which they appear alone (without any other offence being present) were clusters 15, 20 and 24, where the percentage of women was the same, or even higher, than that of men (Fig. 4). With this information alone it is difficult to draw strong conclusions, since there are 22 different types of driver offences and in this SOM disaggregated data is not used because that would add variables to the analysis making it impossible to visualize. Therefore, to better analyze driver offences, a test of proportions [41] has been carried out for the most common or representative types of driver offences with the aim of observing if the differences between men and women are statistically significant or not, regarding the different types of driver offences analyzed and with respect to the total number of offences. This would shed light on clusters where this offence is present.

Among driver offences, the most clearly significant are distracted driving, non-compliance of the STOP signal, partial invasion of opposing lane and not maintaining the safety interval, which represent 30.34%, 12.01%, 11.67% and 9.99% of the total driver offences, respectively. Therefore, they jointly encompass around 64% of these offences.

To carry out these tests, the test statistic applied is (equivalent to using the X^2 distribution):

$$R=(\hat{p}_M-\hat{p}_F)\Big/\sqrt{\hat{p}_M(1-\hat{p}_M)/n_M+\hat{p}_F(1-\hat{p}_F)/n_F}, \quad (1)$$

where $\hat{p}_M$ and $\hat{p}_F$ are the proportions of males and females, respectively, who have committed the driver offence which is being analyzed and n_M and n_F are the sample sizes of all offences of males and females, respectively.

To conclude if the test is significant it is necessary to compute R and to fix the boundary value z_α. Thus, if $|R| \geq z_\alpha$ the test of proportions will be significant. Then the confidence level of the test α must be established. This is usually set at $\alpha = 0.95$. For this value $z_\alpha = 2$. Thus if the absolute value of the test statistic, R, is greater than or equal to 2, the test of proportions will be significant.

The higher the absolute value, the more significant the statistic is, although that does not mean that the proportion difference is higher as well.

The results of the test of proportions for the most common driver offences are shown in Table 4.

The results show that all tests are significant regarding the total number of offences committed by gender. In particular, it is observed that whereas females commit more distraction offences, more non-compliance of the STOP signal and more not maintaining the safety interval, male drivers commit more partial invasion of the opposing lane. Some authors, such as [1], already pointed out that some offences, such as distracted driving, were more frequent among women than among male drivers.

On the other hand, analyzing the rest of the driver offences, it has been observed that their occurrence is more frequent among males. However, the fact that 3 (Distracted driving, non-compliance of the STOP signal and not maintaining the safety interval) of the 4 above mentioned (most common) driver offences are more frequent among females could explain why the proportion of men and women seems matched, in the clusters in which only driver offences appear (clusters 15, 20 and 24).

TABLE 4. Proportions tests for the 4 driving offences more frequent by gender.

	Gender	Commit offences (number of drivers)	Test statistic
Distracted driving	Male	15,166	-9.25
	Female	4,883	
Non compliance of the STOP signal	Male	5,9323	-7.50
	Female	2,020	
Partial invasion of opposing lane	Male	6,129	3.49
	Female	1,567	
Non compliance of the headway distance	Male	4,868	-8.12
	Female	1,722	

We consider that the combination of SOM and hypothesis tests shown here is an interesting illustration of combined sequential methodology. First, the SOM has drawn attention on the a priori balance between male and female drivers in driver offences. Subsequently, inference is applied on disaggregated offence data to shed more light into the process with the added value of providing statistical significance to the results.

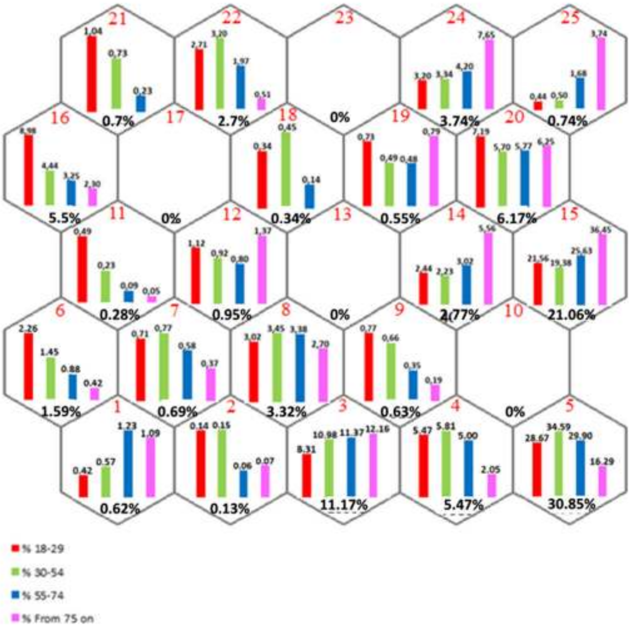


FIGURE 5. Drivers' distribution in the SOM map by age.

2) PATTERN IDENTIFICATION BY AGE

The age of drivers was recorded as categorical and segmented into the groups 18-29, 30–54, 55–74, and >75 years. This segmentation is the one used by DGT in its collision databases and will be adopted in this research.

In this subsection, driver patterns will be extracted by taking into account their age group, but also their gender, that is, considering the results obtained from the analysis of the map of Fig. 4.

In Fig. 5 the percentages of drivers, by age groups, who belong to each SOM cluster, are shown.

From the joint analysis of the SOM (Fig. 2) and the results shown in Fig. 4 and Fig. 5, it can be observed that younger and older drivers, especially males, commit more offences, although their patterns are different.

On the one hand, it can be observed that older drivers (from 75 on) commit more driver offences (cluster 15), which are the most frequent ones and, therefore, they are less represented in the 5th cluster, which includes those who have not committed any offences and, thus, are clearly non-at-fault. In addition, these drivers present more disabilities as a result of age, especially if they are linked to other offences (cluster number 25). However, this group has the lowest alcohol/drug use and speed offences rate. Among these drivers, there are no major differences between men and women.

On the other hand, regarding younger drivers (up to 54 years old), alcohol/drug use is higher. It has been observed that this feature, in general, decreases with age and is practically non-existent among drivers older than 75.

However, a slightly different behavior is observed when alcohol/drug use appears jointly with other offences, in which case (clusters 18 and 22) the group of drivers between 30-54 years becomes important, being more represented than younger drivers (up to 29 years old). This is an illustrative example of the non -a priori- evident multivariate patterns unveiled. However, when alcohol / drug use appears along with speed offences, the presence of drivers between 18-29 years is the most prominent. In both cases, males are more represented (Fig. 4). Finally, regarding speed offences, it is observed that they are clearly more frequent among younger drivers (18-29), mainly males. In addition, these differences tend to increase if other offences come into play, such as those of the driver or alcohol / drug use. Statistical inference should be carried out to test the hypotheses obtained with these results.

3) PATTERN IDENTIFICATION BY TYPE OF COLLISIONS AND TAKING INTO ACCOUNT OFFENCES, GENDER AND AGE

In this subsection, an analysis will be carried out with the purpose of extracting patterns about the type of collision depending on the different types of offences, jointly with gender and age.

In Fig. 6, the percentage of drivers that fall into each of the SOM clusters is shown, depending on the type of collision.

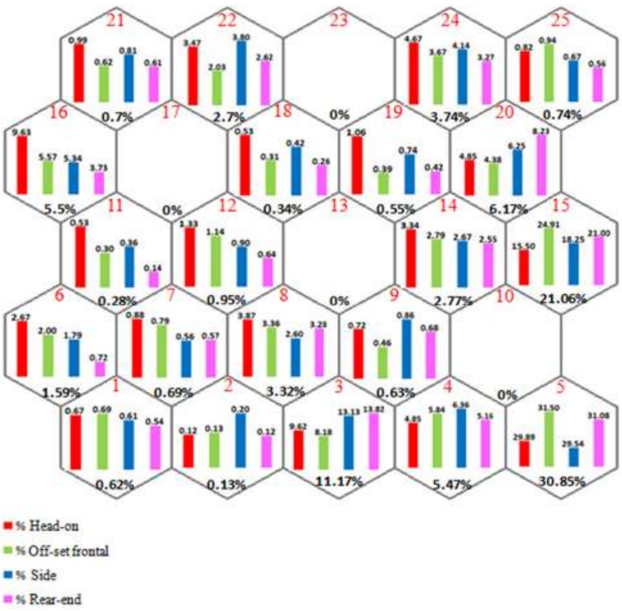


FIGURE 6. Drivers' distribution in the SOM map by type of collisions.

From the analysis of Fig. 6 together with the map in Fig. 2, Fig. 4 and Fig. 5, the clusters where there are more male drivers (11, 12, 16, 18, 21 and 22) are characterized by having more head-on collisions and, to a lesser extent, rear-end collisions. As for female drivers (mainly clusters 2, 3 and 4), they show up more in rear-end collisions.

In most head-on collisions a driver offence has been committed. The most frequent driver offences for this type of collision are partially about invading the opposite lane and, to a lesser extent, distracted driving. The presence of head-on collisions also becomes significant when speed offences and alcohol/drug use are present, especially when these offences appear jointly with the driver offences mentioned above. In this type of collisions, male, especially younger drivers, are more represented. This can be due to younger drivers, especially males, taking more risks [1], [10], and [11].

Side collisions seem to be more frequent among younger drivers (up to 54 years). These types of collisions are more frequent when alcohol/drugs use is present, as was the case with the head-on collisions mentioned above.

Off-set frontal and rear-end collisions seem to affect older drivers more, although no clear patterns have been identified.

As for old drivers (from 75 on), there is no clear type of collision pattern, given that this, mostly, depends on gender and offence type.

It can be concluded that driver behavior based on age is more related to gender than to type of collision.

4) PATTERN IDENTIFICATION BY INJURY SEVERITY TAKING INTO ACCOUNT OFFENCES, GENDER, AGE AND TYPE OF COLLISION

In this subsection, driver patterns regarding their injuries will be extracted based on the offences committed (Fig. 2), as well as on the variables previously analyzed: gender (Fig. 4), age (Fig. 5) and type of collision (Fig. 6).

Fig. 7 shows the percentage of drivers disaggregated according to the injury in the collision, which are included in each of the SOM clusters.

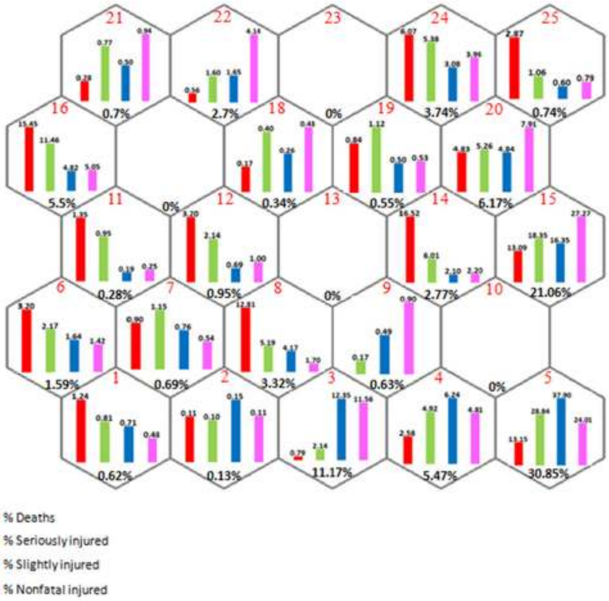


FIGURE 7. Drivers' distribution in the SOM map regarding the injury severity.

It is observed that the youngest drivers (18-29) and the eldest (from 75 years on) are the ones who suffer, regardless the rest of the factors, the most severe injuries.

The types of collisions that generate the greatest severity (death or seriously injured) are the off-set frontal and the head-on collisions, especially the latter, whereas the side and rear-end collisions are those that cause lighter injuries. The profiles of drivers, who mainly suffer the most severe collisions, are the youngest and oldest males.

In turn, it can be observed that the commission of offences influences the severity of driver injuries. Thus, according to the results obtained, drivers who have not committed any offence are those who present the less severe injuries which could explain why women suffer fewer injuries, although they are more vulnerable than men. Statistical inference should be carried out to test the hypotheses obtained with these results. Additionally, it has been observed that speed offences, the presence of physical defects or suffering from some type of sudden illness increase the severity of the injuries, which, as above mentioned, also depends on gender and age of the driver, male drivers, younger and older ones, being again the most affected. On the other hand, alcohol/drug use does not seem to increase the severity of the collision. This could be due to the fact that this type of offence favors the occurrence of side type collisions, as observed in Fig. 6, which are related to a lower injury severity.

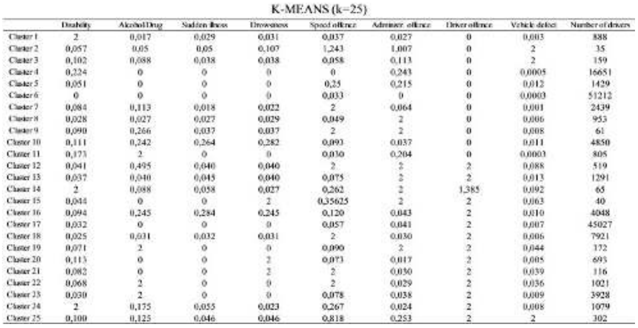
B. VALIDATION OF SOM METHODOLOGY: COMPARISON WITH K-MEANS METHOD

In this subsection the results of SOM are compared with those of another standard clustering technique (K-Means) which involves no lower - dimensional projections.

This comparison will allow for validating the quality of the SOM results in the original spaces. The main difference between K-Means and SOM is that the former only has to perform mono-criterion optimization because K-Means just works in the original space, distributing the data along the clusters in such way that the intra-cluster distances are minimized, which implies maximization of the distances between clusters. However, SOM implies a multi-criteria optimization because it shares the K-Means criterion and additionally it has to maximize conservation of topology. Therefore, this implies some distortion of the results because of the projection of data onto a smaller space.

The results for the cluster centers (weights for SOM and centroids for K-Means) are shown in Table 3 and Table 5, respectively.

TABLE 5. Average value (centroid) of each variable in the different CLuSTERS WITH K-MEANS.



K-MEANS (k=25)

	Disability	Alcohol/Drug	Sudden illness	Drowsiness	Speed offence	Administr. offence	Driver offence	Vehicle defect	Number of drivers
Cluster 1	2	0,017	0,029	0,031	0,037	0,027	0	0,003	888
Cluster 2	0,057	0,05	0,05	0,107	1,243	1,007	0	2	35
Cluster 3	0,102	0,088	0,038	0,038	0,058	0,113	0	2	159
Cluster 4	0,224	0	0	0	0	0,243	0	0,0005	16651
Cluster 5	0,051	0	0	0	0,25	0,215	0	0,012	1429
Cluster 6	0	0	0	0	0,033	0	0	0,0003	51212
Cluster 7	0,084	0,113	0,018	0,022	2	0,064	0	0,001	2439
Cluster 8	0,028	0,027	0,027	0,029	0,049	2	0	0,006	953
Cluster 9	0,090	0,266	0,037	0,037	2	2	0	0,008	61
Cluster 10	0,111	0,242	0,264	0,282	0,093	0,037	0	0,011	4850
Cluster 11	0,173	2	0	0	0,030	0,204	0	0,0003	805
Cluster 12	0,041	0,495	0,040	0,040	2	2	2	0,088	519
Cluster 13	0,037	0,040	0,045	0,040	0,075	2	2	0,013	1291
Cluster 14	2	0,088	0,058	0,027	0,262	2	1,385	0,092	65
Cluster 15	0,044	0	0	2	0,35625	2	2	0,063	40
Cluster 16	0,094	0,245	0,284	0,245	0,120	0,043	2	0,010	4048
Cluster 17	0,032	0	0	0	0,057	0,041	2	0,007	45027
Cluster 18	0,025	0,031	0,032	0,031	2	0,030	2	0,006	7921
Cluster 19	0,071	2	0	0	0,090	2	2	0,044	172
Cluster 20	0,113	0	0	2	0,073	0,017	2	0,005	693
Cluster 21	0,082	0	0	2	2	0,030	2	0,039	116
Cluster 22	0,068	2	0	0	2	0,029	2	0,036	1021
Cluster 23	0,030	2	0	0	0,078	0,038	2	0,009	3928
Cluster 24	2	0,175	0,055	0,023	0,267	0,024	2	0,008	1079
Cluster 25	0,100	0,125	0,046	0,046	0,818	0,253	2	2	302

The comparison between the K-Means centroids and the SOM weights, as well as the number of drivers per cluster in both methodologies leads to the equivalence table (Table 6).

As it can be observed in Table 6, establishing complete equivalence between the K-Means and SOM clusters is not possible, but in this case, one can get very close to it. Therefore, in spite of the distortion generated with SOM when projecting the data onto the 2D space, the results of K-Means and SOM in the original space are very similar.

VI. DISCUSSION

During the last decades, the vehicle collision toll reduction policies within EU have been very effective. In Spain, from 2003 through 2013, deaths per year in urban and interurban roads decreased from 5,399 to 1,680, in accordance with the data provided by the DGT. The main measures adopted were: reduction of "alcohol in blood" thresholds, implementation of the Penalty Point System, legislation changes and improvement in surveillance, control and penalty systems which affected all drivers. From 2014 on, certain stagnation was observed in the decrease of the number of victims. Achieving

TABLE 6. Equivalence table between K-means and SOM.

K-Means	SOM
Cluster 1	Cluster 1
Cluster 2 and 3	Cluster 2
Cluster 4	Cluster 3
Cluster 5 and 6	Cluster 4 and 5
Cluster 7	Cluster 6
Cluster 8 and 9	Cluster 7
Cluster 10	Cluster 8
Cluster 11	Cluster 9
Cluster 12	Cluster 11
Cluster 13, 14 and 15	Cluster 12
Cluster 16	Cluster 14
Cluster 17	Cluster 15, 20 and 24
Cluster 18	Cluster 16
Cluster 19	Cluster 18
Cluster 20 and 21	Cluster 19
Cluster 22	Cluster 21
Cluster 23	Cluster 22
Cluster 24	Cluster 25
Cluster 25	It is not clear

new reduction targets may require specific measurers depending on the different driver groups with different behaviors which could influence the occurrence of collisions, their types and their severity. These measures may be fundamentally oriented to information and education, by means of appropriate campaigns.

The main aim of this research is to extract the maximum information in the 8 dimensional multivariable space of driver offences and conditions and relating them to 4 additional ones (gender, age, type of collision and injury severity) as a contribution to road safety research.

The results of SOM were compared with those of another clustering technique (K-Means) in order to evaluate the extent of the distortion of the SOM resulting from also optimizing conservation of topology.

The driver behavior patterns identified are more clearly observed (visually) in SOM than in other clustering tools, such as K-Means, because the map provides a more visual presentation of the results, albeit paying a price in terms of cluster homogeneity, given that conservation of topology is also optimized.

Some findings identified coincide with existing literature. However, with this research it has been possible to extract additional patterns related to the drivers. The main contribution of this research is the use of the SOM methodology in order to find out more relevant and complex results than with univariate or bivariate analysis, given that some patterns only come to light when more than two variables are studied together.

Regarding multivariate patterns, special attention is required when driver offences are analyzed because it is observed that driver behavior patterns are different depending on whether this type of offence appears alone or not. In clusters where driver offences appear alone, the differences between males and females are smaller, as explained in detail through the hypothesis test results. In addition, the presence of older drivers is larger than in other clusters. However, when driver offences appear jointly with other offences or defects (especially speed offences and alcohol/drugs use), which occurs for an important proportion of all drivers (14.55%), it is mainly due to young males.

Another important multivariate pattern unveiled is that the group of drivers between 30-54 years becomes important when alcohol/drug use appears jointly with driver or administrative offences. However, when speed offences also appear, the youngest drivers (up to 29 years old) are more represented.

This research is relevant because it brings to light different driver behavior patterns and provides the key added value of their relative importance (proportions). This allows for focusing on the development of prevention measures: the results can be useful in decision making in the sense that the information of their relative importance of each pattern will help towards an optimal allocation of resources as carried out by road safety regulating offices such as the DGT.

VII. CONCLUSION

The Self-Organizing Maps (SOM) methodology has been chosen because a multivariate analysis of the characteristics of the drivers is pursued, given that there are patterns that can only be identified when many of the variables of interest are analyzed together. In this research, it has been analyzed in the first place how the SOM has divided the set of drivers in nodes (clusters), according to the multivariate offence data because the full aim of this work is to extract driver behavior patterns in collisions by offences committed, gender, age, type of collision and injuries, as well as the relative importance of these patterns (proportions). This implies an important methodological contribution. In addition, focusing on 3 clusters in the map, where only driver offences appeared, a disaggregated analysis of only the most common types of offences was performed by means of a test of proportions. This strengthened the pattern identification provided by SOM. Finally, for validation of the results obtained with SOM, they were compared with those of K-Means clustering technique.

When the SOM multivariate analysis of offences is related to additional variables, such as gender, age, type of collision and injury severity, different driver behavior patterns have been identified. It is observed that male drivers of younger and older ages are more represented than female drivers and drivers of other age groups in clusters where drivers commit offences. Thus, it can be concluded that the commission of multiple offences is much more frequent in men than in women and special attention is required when driver offences appear jointly with other types of offences or defects, especially speed offences and alcohol/drug use. The corresponding clusters, where more than one offence appear together, are also characterized (besides having more male and young drivers) by involving more dangerous crashes. This last case is an illustrative example of the multivariate patterns unveiled.

Moreover, it has been observed that younger drivers, especially males, commit more speed offences, more driver offences within which partial invading the opposite lane should be highlighted as well as higher consumption of alcohol/drugs. In general, young and male drivers seem to be more involved in head-on collisions, which are the most severe ones, although it has been observed that driver injuries also depend on the offences committed. Thus, if the driver has committed an offence, it is more likely that, keeping the rest of factors fixed, the severity of his/her injuries will be greater.

On the other hand, regarding older drivers, no clearly different behaviors have been identified between males and females. It has been observed that, in general, older drivers commit more driver offences, present also more physical defects as a consequence of age and appear to be more involved in off-set frontal and read-end collisions than drivers belonging to other age groups. However, clear patterns regarding the type of collision have not been identified. In addition, older drivers also seem to suffer higher severity of injuries, although, as noted above, this also depends on whether or not the driver has committed some type of offence.

In future works, statistical inference should be carried out to test the hypotheses obtained with these results. SOM is a powerful tool to identify complex driver behavior patterns so it could be useful to identify, for example, recurrent offenders.

Finally, it is important to take into account the relative importance (proportions) of the different patterns because they will help towards an optimal allocation of resources as carried out by road safety regulating offices such as the DGT, so these findings are intended as a contribution to the field of road safety for different driver groups.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Dirección General de Tráfico (Spanish Traffic General Directorate) for the access to the raw data used in this study. They also thank the ‘‘AUTOMOBILE VEHICLE SAFETY, for smart, sustainable, safe and inclusive mobility’’ - SEGVAUTO 4.0-CM (P2018/EMT-4362) program of the Community of Madrid for its partial funding which, has contributed to its development and dissemination.

REFERENCES

[1] D. L. Massie, K. L. Campbell, and A. F. Williams, ‘‘Traffic accident involvement rates by driver age and gender,’’ *Accident. Anal. Prevention*, vol. 27, no. 1, pp. 73–87, Feb. 1995, doi: 10.1016/0001-4575(94)00050-V.

[2] D. R. Mayhew, S. A. Ferguson, K. J. Desmond, and H. M. Simpson, ‘‘Trends in fatal crashes involving female drivers, 1975-1998,’’ *Accident. Anal. Prevention*, vol. 35, no. 3, pp. 407–415, May 2003, doi: 10.1016/S0001-4575(02)00019-2.

[3] D. Bose, S. M. Segui-Gomez, and J. R. Crandall, ‘‘Vulnerability of female drivers involved in motor vehicle crashes: An analysis of US population at risk,’’ *Amer. J. Public Health*, vol. 101, no. 12, pp. 2368–2373, Dec. 2011, doi: 10.2105/AJPH.2011.300275.

[4] M. Durán Segura, D. Cantón Cortés, and C. Castro Ramírez, ‘‘Changing patterns in women's driving,’’ *Int. J. Psychol. Res.*, vol. 2, no. 1, pp. 54–66, Jun. 2009, doi: 10.21500/20112084.878.

[5] S. Laapotti, E. Keskinen, and S. Rajalin, ‘‘Comparison of young male and female drivers' attitude and self-reported traffic behaviour in finland in 1978 and 2001,’’ *J. Saf. Res.*, vol. 34, no. 5, pp. 579–587, Jan. 2003, doi: 10.1016/j.jsr.2003.05.007.

[6] A. H. Al-Balbissi, ‘‘Role of gender in road accidents,’’ *Traffic Injury Prevention*, vol. 4, no. 1, pp. 64–73, Mar. 2003, doi: 10.1080/15389580309857.

[7] C. Harris and M. Jenkins, ‘‘Gender differences in risk assessment: Why do women take fewer risks than men,’’ *Judg. Dec. Mak.*, vol. 1, no. 1, pp. 48–63, Jul. 2006.

[8] J. J. Jiménez, P. Lardelli, J. D. Luna, M. García, A. Bueno, and R. Gálvez, ‘‘Efecto de la edad, el sexo y la experiencia de los conductores de 18 a 24 años sobre el riesgo de provocar colisiones entre turismos,’’ *Gac. Sanit.*, vol. 18, no. 3, pp. 166–176, Feb. 2004, doi: 10.1016/S0213-9111(04)71829-4.

[9] C. Turner and R. McClure, ‘‘Age and gender differences in risk-taking behaviour as an explanation for high incidence of motor vehicle crashes as a driver in young males,’’ *Injury Control Saf. Promotion*, vol. 10, no. 3, pp. 123–130, Aug. 2010, doi: 10.1076/icsp.10.3.123.14560.

[10] P. Lardelli, J. D. Luna, J. Jiménez, A. Bueno, M. García, and R. Gálvez, ‘‘Age and sex differences in the risk of causing vehicle collisions in Spain, 1990 to 1999,’’ *Accident. Anal. Prevention*, vol. 35, no. 2, pp. 261–272, Mar. 2003, doi: 10.1016/S0001-4575(02)00004-0.

[11] T. Özkan and T. Lajunen, ‘‘What causes the differences in driving between young men and women? The effects of gender roles and sex on young drivers' driving behaviour and self-assessment of skills,’’ *Transp. Res. F, Traffic Psychol. Behaviour*, vol. 9, no. 4, pp. 269–277, Jul. 2006, doi: 10.1016/j.trf.2006.01.005.

[12] A. F. Williams, ‘‘Teenage drivers: Patterns of risk,’’ *J. Saf. Res.*, vol. 34, no. 1, pp. 5–15, Jan. 2003, doi: 10.1016/S0022-43752)00075-0.

[13] S. Regev, J. J. Rolison, and S. Moutari, ‘‘Crash risk by driver age, gender, and time of day using a new exposure methodology,’’ *J. Saf. Res.*, vol. 66, pp. 131–140, Sep. 2018, doi: 10.1016/j.jsr.2018.07.002.

[14] S. Doroudgar, H. M. Chuang, P. J. Perry, K. Thomas, K. Bohnert, and J. Canedo, ‘‘Driving performance comparing older versus younger drivers,’’ *Traffic Injury Prevention*, vol. 18, no. 1, pp. 41–46, Jan. 2017, doi: 10.1080/15389588.2016.1194980.

[15] J. F. Antin, F. Guo, Y. Fang, T. A. Dingus, M. A. Perez, and J. M. Hankey, ‘‘A validation of the low mileage bias using naturalistic driving study data,’’ *J. Saf. Res.*, vol. 63, pp. 115–120, Dec. 2017, doi: 10.1016/j.jsr.2017.10.011.

[16] N. Stamatiadis and J. A. Deacon, ‘‘Quasi-induced exposure: Methodology and insight,’’ *Accident Anal. Prevention*, vol. 29, no. 1, pp. 37–52, 1997, doi: 10.1016/S0001-4575(96)00060-7.

[17] L. Evans, ‘‘Female compared with male fatality risk from similar physical impacts,’’ *J. Trauma, Injury, Infection, Crit. Care*, vol. 50, no. 2, pp. 281–288, Feb. 2001, doi: 10.1097/00005373-200102000-00014.

[18] S. Islam and F. Mannering, ‘‘Driver aging and its effect on male and female single-vehicle accident injuries: Some additional evidence,’’ *J. Saf. Res.*, vol. 37, no. 3, pp. 267–276, Jan. 2006, doi: 10.1016/j.jsr.2006.04.003.

[19] E. Santamariña-Rubio, K. Pérez, M. Olabarria, and A. M. Novoa, ‘‘Gender differences in road traffic injury rate using time travelled as a measure of exposure,’’ *Accident Anal. Prevention*, vol. 65, pp. 1–7, Apr. 2014, doi: 10.1016/j.aap.2013.11.015.

[20] Ö. imeko lu, T. Nordfjærn, M. F. Zavareh, A. M. Hezaveh, A. R. Mamdoohi, and T. Rundmo, ‘‘Risk perceptions, fatalism and driver behaviors in turkey and iran,’’ *Saf. Sci.*, vol. 59, pp. 187–192, Nov. 2013, doi: 10.1016/j.ssci.2013.05.014.

[21] D. M. Dejoy, ‘‘An examination of gender differences in traffic accident risk perception,’’ *Accident. Anal. Prevention*, vol. 24, no. 3, pp. 237–246, Jun. 1992, doi: 10.1016/0001-4575(92)90003-2.

[22] L. A. D'Ambrosio, L. K. M. Donorfio, J. F. Coughlin, M. Mohyde, and J. Meyer, "Gender differences in self-regulation patterns and attitudes toward driving among older adults," *J. Women Aging*, vol. 20, nos. 3–4, pp. 265–282, Aug. 2008, doi: 10.1080/08952840801984758.

[23] R Development Core Team, "R: A language and environment for statistical computing," Computing RFfS, Vienna, Austria, ed., 2013. [Online]. Available: http://www.R-project.org

[24] T. Kohonen, "The SOM methodology," in *Visual Explorations in Finance*, 1st ed. London, U.K.: Springer-Verlag, 1998.

[25] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, nos. 1–3, pp. 1–6, 1998, doi: 10.1016/S0925-2312(98)00030-7.

[26] A. C.-D. Lee and C. Rinner, "Visualizing urban social change with self-organizing maps: Toronto neighbourhoods, 1996–2006," *Habitat Int.*, vol. 45, pp. 92–98, Jan. 2015, doi: 10.1016/j.habitatint.2014.06.027.

[27] D. E. Rumelhart and D. Zipser, "Feature discovery by competitive learning," *Cognit. Sci.*, vol. 9, no. 1, pp. 75–112, Jan. 1985, doi: 10.1207/s15516709cog0901_5.

[28] N. M. Allinson and H. Yin, "Self-Organising Maps for pattern recognition," in *Kohonen Maps*, E. Oja and S. Kaski, Eds. Amsterdam, The Netherlands: Elsevier, 1999, pp. 111–120, doi: 10.1016/B978-044450270-4/50008-5.

[29] K. Lagus, "Text retrieval using self-organized document maps," *Neural Process. Lett.*, vol. 15, no. 1, pp. 21–29, 2002, doi: 10.1023/A:1013853012954.

[30] T. Kohonen, "Essentials of the self-organizing map," *Neural Netw.*, vol. 37, pp. 52–65, Jan. 2013, doi: 10.1016/j.neunet.2012.09.018.

[31] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. London, U.K.: Pearson, 2009.

[32] P. Liu, "A self-organizing feature maps and data mining based decision support system for liability authentications of traffic crashes," *Neurocomputing*, vol. 72, nos. 13–15, pp. 2902–2908, Aug. 2009, doi: 10.1016/j.neucom.2008.06.032.

[33] H. Yin, "The self-organizing maps: Background, theories, extensions and applications," in *Computational Intelligence: A Compendium. Studies in Computational Intelligence*. Berlin, Germany: Springer, 2008, pp. 715–762.

[34] M. M. Van Hulle, "Self-organizing maps," in *Handbook of Natural Computing*. Berlin, Germany: Springer, 1998, pp. 585–622.

[35] Y. Chen, Y. Zhang, J. Hu, and D. Yao, "Pattern discovering of regional traffic status with self-organizing maps," in *Proc. IEEE Intell. Transp. Syst. Conf.*, Sep. 2006, pp. 647–652, doi: 10.1109/ITSC.2006.1706815.

[36] C. G. Prato, V. Gitelman, and S. Bekhor, "Mapping patterns of pedestrian fatal accidents in israel," *Accident Anal. Prevention*, vol. 44, no. 1, pp. 56–62, Jan. 2012, doi: 10.1016/j.aap.2010.12.022.

[37] K. Kim and E. Y. Yamashita, "Using a k-means clustering algorithm to examine patterns of pedestrian involved crashes in honolulu, hawaii," *J. Adv. Transp.*, vol. 41, no. 1, pp. 69–89, Sep. 2007, doi: 10.1002/atr.5670410106.

[38] F. L. Mannering and C. R. Bhat, "Analytic methods in accident research: Methodological frontier and future directions," *Analytic Methods Accident Res.*, vol. 1, pp. 1–22, Jan. 2014, doi: 10.1016/j.amar.2013.09.001.

[39] S. Kaski and K. Lagus, "Comparing self-organizing maps," in *Proc. Int. Conf. Artif. Neural Netw.*, Berlin, Germany, 1996, pp. 809–814.

[40] L. Lebart, A. Morineau and K. M. Warwick, *Multivariate Descriptive Statistical Analysis*. New York, NY, USA: Wiley, 1984.

[41] L. Wassennan, "Hypothesis testing and p-values," in *All of Statistics. A Concise Course in Statistical Inference*. New York, NY, USA: Springer, 2004, pp. 149–173.

ALMUDENA SANJURJO-DE-NO was born in Madrid, Spain, in 1991. She received the master's degree in industrial engineering in 2016. She is currently pursuing the Ph.D. degree with the Instituto Universitario de Investigación del Automóvil (INSIA), Universidad Politécnica de Madrid (UPM), dissertation on machine learning techniques applied to road safety.

BLANCA ARENAS-RAMÍREZ was born in Salta, Argentina, in 1960. She received the Ph.D. degree in civil engineering from the Universidad Politécnica de Madrid (UPM), in 2008. She is currently the Head of the Transportation and Vehicle Environmental Impact Unit, Instituto Universitario de Investigación del Automóvil (INSIA), UPM. She is also an Associate Professor of transport engineering, scientific safety research, and mobility and transport in several Master programs at UPM and several South America Universities. Her current research interests include modeling and data analysis for road safety, vehicle environmental, and sustainability.

JOSÉ MIRA was born in Madrid, Spain, in 1960. He received the master's degree in nuclear engineering and the Ph.D. degree in applied statistics from the Universidad Politécnica de Madrid (UPM). He is currently an Associate professor of statistics with UPM. His current research interests include machine learning and Monte Carlo simulations with applications to road safety and the electricity market.

FRANCISCO APARICIO-IZQUIERDO was born in Guadix, Spain, in 1945. He received the Ph.D. degree in mechanical engineering from the Universidad Politécnica de Madrid (UPM), in 1978. He was a Professor from 1981 to 2015. He is currently a Professor Emeritus at UPM. In 1993, he founded the Instituto Universitario de Investigación del Automóvil (INSIA), UPM, and has been its Director for 22 years. He currently holds the position of the President of INSIA. He has been the Principal Investigator for more than 100 research projects on road and vehicle safety and environmental impact. He has published in highly ranked journals.