Multivariate Statistical Analysis for Exploring Road Crash Related Factors in the Franche-Comté region of France

Cécile Spychala, Joël Armand, Clément Dombry, Camelia Goga

Multivariate Statistical Analysis for Exploring Road Crash Related Factors in the France Franche-Comté region of France

Cécile Spychala[(1)]*, Joël Armand[(2)], Clément Dombry[(1)] and Camelia Goga[(1)]
[(1)] LMB, Université de Bourgogne Franche-Comté, Besançon, FRANCE
[(2)] Gendarmerie Nationale de Besançon, Besançon, FRANCE
cecile.spychala@univ-fcomte.fr, joel.armand@gendarmerie.interieur.gouv.fr
clement.dombry@univ-fcomte.fr, camelia.goga@univ-fcomte.fr

May 17, 2020

Abstract

Understanding and modelling road crash data is crucial in fulfilling safety goals by helping national authorities to take necessary measures to reduce crash frequency and severity. This work aims at giving a multivariate statistical analysis of road crash data from the French region of Franche-Comté with special attention to road crash gravity. The first step for this multivariate analysis was to perform Multiple Correspondence Analysis in order to assess associations between the road crash injury and several important accident-related factors and circumstances. Log-linear models are used next in order to detect associations between road crash severity and related factors such as alcohol/drug consumption or spatial crash locations. The effects of each factors have been also evaluated on the road crash gravity by using ordinal logistic regression. Data used in this study are extracted from BAAC files, the French census of road crashes.

Key words: geometric data analysis, multiple correspondence analysis, ordinal logistic regression, log-linear model, road crash severity.

1 Introduction

Over the last decade, the number of road crashes has continuously been decreasing in France. Indeed, 61 224 accidents have been recorded in 2017 instead of 58 352 in 2018, a decrease of 4,7% (ONISR, 2019). However, road accidents still happen and important efforts and means are developed to prevent them. Among these, modern statistical methods are efficient prevention tools used to describe and model accident data. This paper is concerned about road accidents that occurred in the Franche-Comté region of France (see Fig. 1). This region from the east of France is split up into four departments called Doubs, Jura, Haute-Saône and Teritoire de Belfort. Regarding to the mortality rate from 2017 to 2018, this rate has globally decreased for this region. However, the situation is quite different within each

*C. Spychala's work was supported by a grant of Grand Besançon.

department. Indeed, the death rate has increased by 3% from 2017 in the Doubs department while it has decreased in the Haute-Saône, Jura and Territoire de Belfort departments by 45%, 65% and respectively by 50% (ONISR, 2019). Understanding and modelling accident data is crucial in fulfilling safety goals by helping national authorities to undertake necessary measures to reduce crash frequency and severity.

This paper focuses on accidents in Franche-Comté involving casualties. An accident refers to a road crash with casualty needing hospital care and can involve several cars and several people. One of the main goals of the National Gendarmerie of Besançon (Doubs, France) is to reduce the number of accidents in Franche-Comté. More precisely, the National Gendarmerie of Besançon plans to be able in the near future to anticipate road crashes by using time and spatial modelling of accident data. This study aims at giving a multivariate statistical analysis of the road crashes in Franche-Comté. A first multivariate descriptive study of French accident data was conducted by Bièvre (2017) in an unpublished technical report. We intend in this work to give a deeper analysis of Franche-Comté accident data.

The main goal of this research work is to explain the variable giving the severity or the gravity of the accidents by using several covariates such as spatial location, time period, weather conditions, road type, alcohol/drug consumption... Our multivariate statistical analysis starts with a Multiple Correspondence Analysis (MCA). The MCA as suggested by Benzécri (Benzécri, 1973, 1982) is the generalization of the Correspondence Analysis (CA) for analyzing jointly more than two categorical variables. This method is widely used in categorical data analyses because it allows detecting similarities between individuals and assessing associations between categories. Geometric representations of data clouds in smaller dimension spaces allow identifying clusters of similar individuals and of associated categories or variables. Many applications of MCA and related methods in various fields such as social, demographic, economic are given in Greenacre and Blasius (2006). The goal here is to determine the accident factors mostly related to road crash severity. In the literature concerning the accident analysis and prevention, several studies used MCA in various contexts but different from our framework. For example, Das and Sun (2015) used eight years of pedestrian crash data and MCA to identify key associations between risk factors and Das and Sun (2016) used MCA to identify crash-prone factors producing fatal run-off-road crashes; Das et al. (2018) investigated the wrong way driving crash patterns by using MCA while Fort et al. (2019) tried to explain working conditions and risk exposure of employees whose occupations require driving on public roads.

The MCA analysis conducted on the Franche-Comté accident data set allows us to identify several variables associated with the road crash severity. A more in-depth analysis of these variables is next considered by log-linear modelling (Agresti, 2013). The log-linear model belongs to the class of generalized linear model (McCullagh and Nelder, 1989). In the case of categorical data, the cell counts of the contingency table are modeled by a Poisson distribution and a log link function is used for the mean. More precisely, the log-linear model specifies how the expected counts depend on the levels of the categorical variables and it allows to quantify the associations and interactions between those variables. Unlike MCA, log-linear models allow getting insight into complex dependence patterns such as conditional

or marginal dependence which may exist between several categorical variables. In our framework, we will use log-linear models in order to detect conditional or marginal associations between road crash severity and other variables such as alcohol/drug consumption and spatial location. In a similar way, Abdel-Aty et al. (1998) used log-linear models to explain associations between the driver age and several important factors and circumstances related to the accident. Also, Yannis et al. (2005) performed a log-linear analysis in order to test the significance of first- and second-order effects among various combinations of driver age and engine size categories in relation to two-wheeler accident severity and at-fault risk rates. Then, Abdel-Aty and Abdelwahab (2000) used log-linear models to investigate whether there are associations between the different driver characteristics and alcohol involvement and also in order to identify the high-risk group within each driver factor.

Finally, we propose ordinal logistic regression (Agresti, 2013) to model the gravity level probabilities as a function of explicative covariates such as alcohol/drug consumption, time period and spatial locations. This method widely used in accident data analysis is a popular supervised learning method for analyzing dependencies between a binary or multiclass response categorical variable and several explanatory variables. It allows in particular to separate and identify the effects of each explanatory variable on the response variable. Rezapour and Ksaibati (2018) used ordinal logistic regression to investigate the contributory factors that increased the odds of severe single-truck and multiple-vehicle crashes such as characteristics related to driver or vehicle for instance. Then, Mekonnen (2018) has also performed ordinal logistic regression in order to identify the risk factors among driver age, speed record or alcohol consumption for example for severity levels of road traffic accident.

The paper is structured as follows. We first describe in Section 2 our data set as well as the analysis methods: MCA is described briefly in Section 2.2.1, log-linear modelling in Section 2.2.2 and ordinal logistic regression in Section 2.2.3. Section 3 contains the main results of our study and, lastly, Section 4 concludes and proposes several recommendations and perspectives.

2 Material and methods

2.1 Franche-Comté accident data

Data used in this study concern the Franche-Comté road crashes between 2005 to 2018 which are extracted from the French national analysis bulletin of road traffic injury accidents called BAAC [1] (*Bulletin d'Analyse des Accidents Corporels*). The BAAC data are filled in by the security forces present on the accident scene and next, data are treated, analyzed and put online by the national interdepartmental observatory of road safety (*Observatoire National Interministériel de la Sécurité Routière*).

The BAAC files contained more than 50 variables from which 15 new categorical variables have been created and/or reclassified. The Franche-Comté accident dataset has 11 776 casualties registered in 4 950 accidents. The study focuses only on the accident itself and

[1] The reader can find the BAAC open data on the government website https://www.data.gouv.fr/fr/

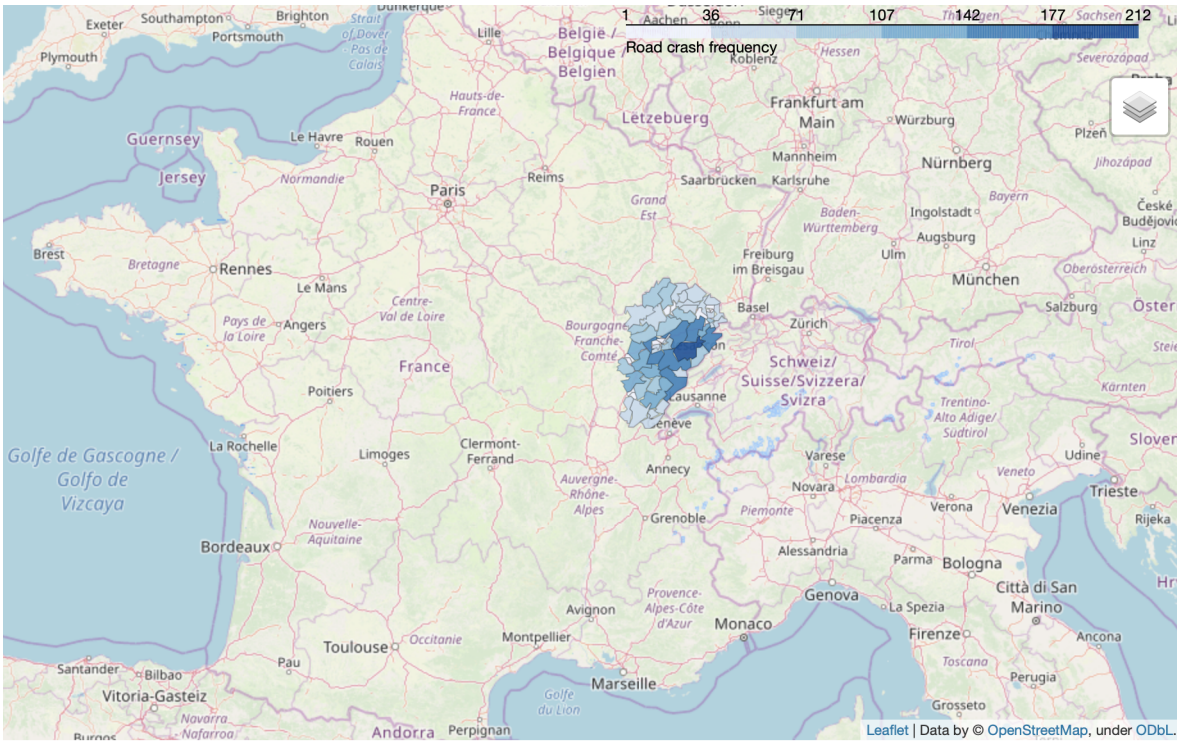

FIG 1: France map with road crash frequency of Franche-Comté region. Each small division corresponds to a canton.

not on each casualties. The region Franche-Comté is situated in the east of France and neighboring Switzerland as we can see from FIG. (1). The counties situated on the west of Franche-Comté are mountainous and entirely deserved by national and departmental roads. A daily intensive border activity between Switzerland and France is also present in these counties.

The analysis emphasizes the accident severity, denoted by *type_acc*, classified into three ordered levels: "slight_safe", "serious" and "fatal". An accident is considered as "slight_safe" (11,47% of accidents) if all passengers were safe or had minor injuries; the label "serious" was attributed to accidents involving at least one casualty needing hospital care for more than 24 hours (69,82% of accidents) and lastly, an accident is considered as "fatal" (18,71% of accidents) if at least one casualty involved died.

The alcohol/drug consumption by car drivers is one of the main accident causes and has a great impact on their severity. The categorical variable *substance* describing the alcohol/drug consumption by the drivers involved in a road crash has the following levels:

- "alcohol_drug" when at least one of the involved drivers has consumed both alcohol and drugs (2.69%);
- "drug" when at least one of the involved drivers has consumed drugs (2.73%);
- "alcohol" when at least one of the involved drivers has consumed alcohol but not drugs (16.22%);

- "none" is associated with accidents involving only sober drivers (78.36%).

If the accident involves only one driver, the variable *substance* concerns the unique driver.

As mentioned above, the goal of this study is the statistical analysis of accidents and an accident may involve several drivers and casualties. Variables such as *age* or *sex* refer to individuals and are not straightforward to recode for an accident involving several persons. For this reason, *age* or *sex* do not appear in our multivariate analysis.

In order to conduct the temporal analysis of the Franche-Comté road crashes, we used the following categorical variables related to the time period when the accident occurred:

- *season* with four categories: spring, summer, autumn and winter;
- *week* with two categories: weekday and week_end;
- *daytime* with two categories: day and night;
- *time* with five categories: 7am_10am, 11am_3pm, 4pm_7pm, 8pm_11pm and midnight_6am. Note that the category 7am_10am means from 7:00 am to 10:59 am. It is also the case for the other categories of *time*.

In our accident data, each accident is located by the *commune* (town or village) and the *department* (Doubs, Jura, Haute-Saône, Territoire de Belfort) where the accident took place. The variable *commune* was used to build the variable *canton* (district) by regrouping the 1176 communes into 50 cantons. In fact, the region of Franche-Comté is splitted up into 62 cantons, however, some cantons have been grouped together as for instance "Belfort-1", "Belfort-2" and "Belfort-3" into "Belfort". This reclassification allows to smooth the variability of *cantons* categories. Hence, Jura department is divided into 15 cantons, Haute-Saône and Doubs both into 14 cantons and Territoire de Belfort into 7 cantons. The categorical variables *department* and *canton* have been used for the spatial analysis, whereas the variable *commune* was dropped due to too many categories. We give in FIG. (1) the division of Franche-Comté into cantons with their road crash frequencies; Jura department is the department with the highest road crash frequency.

In order to give a more thorough statistical analysis, we considered further 7 categorical variables giving supplementary information about the weather, the type of the road and of the collision:

- *weather* with two categories: normal and other kind (such as rainy, cloudy or snowy weather);
- *area* with two categories: unurban and urban;
- *intersection* with two categories: intersection and out_of_intersection;
- *obstacle* corresponding to a mobile obstacle with four categories: vehicle, pedestrian, other_kind (such as animals) and none;
- *shape_road* with two categories: curve and straight;

- *collision* with three categories: usual (such as frontal or rear-end collisions), other_kind and none;
- *type_road* with five categories: communal, departmental, national, highway and other_kind (such as parking).

TAB 1 gives the cross-tabulation of the accident severity (*type_acc*) with the different categorical variables.

TAB 1: Contingency tables crossing the accident severity *type_acc* with the other categorical variables. The variable *canton* has too many levels and only "Arbois" and "Villersexel" are described. The numbers in parentheses are, for each categorical variable, the relative frequencies of its different levels and, within each level, the distribution of the accident severity.

Attribute		Type of accident			Attribute		Type of accident		
		slight_safe (11,47%)	serious (69,81%)	fatal (18,71%)			slight_safe (11,47%)	serious (69,81%)	fatal (18,71%)
substance					*department*				
alcohol_drug	133 (2,69%)	10 (7,52%)	61 (45,86%)	62 (46,62%)	Doubs	2060 (41,62%)	255 (12,38%)	1433 (69,56%)	372 (18,06%)
drug	135 (2,73%)	8 (5,93%)	86 (63,70%)	41 (30,37%)	Haute_Saone	1210 (24,44%)	138 (11,40%)	859 (70,99%)	213 (17,60%)
alcohol	803 (16,22%)	5 (7,22%)	542 (67,50%)	203 (25,28%)	Jura	1373 (27,74%)	103 (7,50%)	972 (70,79%)	298 (21,70%)
none	3879 (78,36%)	192 (12,68%)	2767 (71,33%)	620 (15,98%)	Terr_Belfort	307 (6,20%)	72 (23,45%)	192 (62,54%)	43 (14,01%)
season					*canton*				
spring	1199 (24,22%)	151 (12,59%)	810 (67,56%)	238 (19,85%)	Arbois	131 (2,65%)	11 (8,40%)	99 (75,57%)	21 (16,03%)
summer	1580 (31,92%)	167 (10,57%)	1142 (72,28%)	271 (17,15%)	...	...	...	...	...
autumn	1223 (24,71%)	119 (9,73%)	874 (71,46%)	230 (18,81%)	Villersexel	93 (1,88%)	10 (10,75%)	60 (64,52%)	23 (24,73%)
winter	948 (19,15%)	131 (13,82%)	630 (66,46%)	187 (19, 73%)					
					obstacle				
week					vehicle	2484 (42,10%)	321 (12,92%)	1734 (69,81%)	429 (17,27%)
weekday	3133 (63,29%)	395 (12,61%)	2155 (68,78%)	583 (18,61%)	pedestrian	410 (8,28%)	58 (14,15%)	280 (68,29%)	72 (17,56%)
week_end	1817 (36,71%)	173 (9,52%)	1301 (71,60%)	343 (18,88%)	other_kind	146 (2,95%)	21 (14,38%)	98 (67,12%)	27 (18,49%)
					none	1910 (38,59%)	168 (8,80%)	1344 (70,37%)	398 (20,84%)
daytime									
day	3634 (73,41%)	440 (12,11%)	2581 (71,02%)	613 (16,87%)	*shape_road*				
night	1316 (26,59%)	128 (9,73%)	875 (66,49%)	313 (23,78%)	curve	1996 (40,32%)	183 (9,17%)	1427 (71,49%)	386 (19,34%)
					straight	2954 (59,68%)	385 (13,03%)	2029 (68,69%)	540 (18,28%)
time									
7am_10am	785 (15,86%)	79 (10,06%)	566 (72,10%)	140 (17,83%)	*collision*				
11am_3pm	1408 (28,44%)	171 (12,14%)	992 (70,45%)	245 (17,40%)	usual	2995 (60,51%)	407 (13,59%)	2056 (68,65%)	532 (17,76%)
4pm_7pm	1587 (32,06%)	212 (13,36%)	1132 (71,33%)	243 (15,31%)	*other_kind*	1271 (25,68%)	106 (8,34%)	887 (69,79%)	278 (21,87%)
8pm_11pm	618 (12,48%)	63 (10,19%)	410 (66,34%)	145 (23,46%)	none	684 (13,82%)	55 (8,04%)	513 (75,00%)	116 (16,96%)
midnight_6am	552 (11,15%)	43 (7,79%)	356 (64,49%)	153 (27,72%)					
					type_road				
weather					communal	503 (10,16%)	63 (12,52%)	378 (75,15%)	62 (12,33%)
normal	3599 (72,71%)	409 (11,36%)	2550 (70,85%)	640 (17,78%)	departmental	3563 (71,98%)	400 (11,23%)	2488 (69,83%)	675 (18,94%)
other_kind	1351 (37,29%)	159 (11,77%)	906 (67,06%)	286 (21,17%)	national	667 (13,47%)	66 (9,90%)	449 (67,32%)	152 (22,79%)
					highway	92 (1,86%)	20 (21,74%)	51 (55,43%)	21 (22,83%)
area					other_kind	125 (2,53%)	19 (15,20%)	90 (72,00%)	16 (12,80%)
unurban	3460 (69,90%)	337 (9,74%)	2355 (68,06%)	768 (22,20%)					
urban	1490 (30,10%)	231 (15,50%)	1101 (73,89%)	158 (10,60%)	*intersection*				
					intersection	575 (11,62%)	76 (13,22%)	421 (73,22%)	78 (13,57%)
					out_of_intersection	4375 (88,38%)	492 (11,25%)	3035 (69,37%)	848 (19,38%)

2.2 Statistical analysis

2.2.1 Multiple Correspondence Analysis

The Multiple Correspondence Analysis (MCA) is an efficient unsupervised method for exploring multivariate categorical data. The aim of MCA is to study the similarities between the individuals, to assess the relationships between the variables and to examine the associations between the categories. For a thorough description of MCA as well as of related methods, the reader is referred to the book of Greenacre and Blasius (2006). This method allows, if appropriate, to corroborate a strong link between categorical variables. In some cases, MCA enables to cluster categories and to reduce the data dimension allowing multivariate data to be analyzed more easily. Indeed, a graphical representation of individuals and variables is built in an orthogonal system similarly as in Correspondence Analysis (CA). This statistical tool is powerful for understanding, visualizing and simplifying the data.

MCA can be derived in several ways. One way is to apply CA on the indicator matrix $\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2 \quad \dots \quad \mathbf{X}_p]$ derived from the original data $Individuals \times Categorical\ variables$ of p categorical variables recorded on n individuals. Each indicator matrix $\mathbf{X}_j$ is obtained by column concatenation of K_j dummy variables where K_j is the number of categories of the jth categorial variable, $j = 1, \dots, p$. Hence, $\mathbf{X}$ is a respondents-by-categories matrix having n rows, corresponding to individuals, and $K = \sum_{j=1}^{p} K_j$ columns, corresponding to variable categories. An element of this table, denoted by x_{ik}, is equal to 1 if the individual i has the category k and 0 otherwise, $i = 1, \dots, n$ and $k = 1, \dots, K$. The indicator matrix $\mathbf{X}$ has row sums equal to the constant p and column sums equal to n_k, the marginal frequency of the kth category, namely the number of individuals having the category k.

This kind of data implies the study of three kinds of objects: the individuals, the variables but also their categories. The scheme of MCA is to compare individuals and evaluate variables characteristics by providing row typologies, column typologies and the relationships between these typologies (Escofier and Pagès, 2008).

From a technical point of view, MCA uses as CA the χ^2 distance in order to assess similarity or dissimilarity between different columns or lines contained in $\mathbf{X}$. The indicator matrix $\mathbf{X}$ is transformed in order to obtain row profiles by dividing each element of a row by the row frequency as well as column profiles by dividing each element of a column by its frequency. In the case of MCA, row and column profiles are very simple. The ith row profile is given by $(x_{ik}/p)_{k=1}^{K}$: the elements of a row profile have only zero and $1/p$ values, the non-zero value being recorded if the individual i possesses the category k. So, row profiles will be different only for ith and i'th individuals having mismatching category levels. The kth column profile is given by $(x_{ik}/n_k)_{i=1}^{n}$: the elements of the column profile are zero and $1/n_k$ values.

The χ^2-distance between two individuals i and i' is a weighted sum of squared distances between the ith and i'th row profiles with weights given by the inverse of the average row profile given by $(n_k/np)_{k=1}^{p}$:

$$d^2_{i,i'} = \sum_{k=1}^{K} \frac{np}{n_k} \left(\frac{x_{ik}}{p} - \frac{x_{i'k}}{p} \right)^2 = \frac{n}{p} \sum_{k=1}^{K} \frac{(x_{ik} - x_{i'k})^2}{n_k}, \qquad 1 \leq i, i' \leq n. \tag{1}$$

Hence, the terms from the above sum will be all zero for coincident zero values and coincident $1/p$ values meaning that these squared differences will not contribute to the distance measure. Only differences between noncoincident categories will contribute to the distance $d^2_{i,i'}$ and this contribution is proportional to $(1/p)^2$ with weight equal to the inverse of the marginal frequency n_k. The χ^2 distance between row profiles can be interpreted as a weighted mismatching dissimilarity coefficient: small distance $d_{i,i'}$ means that individuals i and i' have many categories in commun, so they are very similar and on the contrary, large distance $d_{i,i'}$ means that i and i' have few categories in commun, so they are very different. Moreover, a rare category (small n_k) has a large contribution to the final distance and moves its owner or owners far away from the others individuals.

While the interpretation of the χ^2 distance between individuals is similar to the one given in the CA, the χ^2 distance interpretation for variable analysis is quite different and more difficult to justify (Greenacre, 2006). Information contained in a variable can be studied through its categories, thus, MCA focuses mostly on variable categories. As for row-profiles, the distance between categories k and k' is defined as the weighted sum of squared distances between the kth and k'th column profiles with weights given by the inverse of the average column profile which has in this case all elements equal to $1/n$:

$$d^2_{k,k'} = n \sum_{i=1}^{n} \left(\frac{x_{ik}}{n_k} - \frac{x_{ik'}}{n_{k'}} \right)^2 = \frac{1}{p_k} + \frac{1}{p_{k'}} - \frac{2p_{kk'}}{p_k p_{k'}}, \qquad 1 \leq k, k^{'} \leq K, \tag{2}$$

where $p_k = n_k/n$ is the relative frequency of the category k and $p_{kk'}$ the relative frequency of occurence of categories k and k'. If k and k' are different categories of the same variable, then $p_{kk'} = 0$. As it is defined, the distance between column profiles is a decreasing function with respect to the relative frequencies p_k and joint relative frequencies $p_{kk'}$. Two categories are close one to each other with respect to this χ^2 distance if they have many individuals in common. Again, rare categories are far away from the others. In brief, it is important to take the frequency of each category into account. However, as remarked by Greenacre (1989) and Greenacre (2006), the terms $1/p_k$ present in the χ^2 distance are hard to interpret.

Once that distances between objects (individuals and variables) have been defined, the next step in a MCA is to represent individuals and variables in new orthogonal systems and to make the geometric data analysis on smaller dimension sets (Le Roux and Rouanet, 2004). As in principal component or correspondence analysis, new orthogonal systems are built such that they maximise the projected inertia of the individual cloud or variables on these new orthogonal axis, the inertia being defined as usual as the weighted sum of squared distance of individuals or variables to their barycenter. Each axis represents a certain percentage from the total inertia. However, these percentages in MCA are lower than in CA and more dimensions are needed to interpret properly the analysis. Transition relations link the cloud of individuals with the cloud of categories and a biplot representation is usually used as a joint map of individuals and variable categories. The contribution of each individual to each axis as well as the quality of its representation on each axis are obtained in a similar way to CA. For more details about the graphical representation and all matters connected therewith, see for example Greenacre (2006), Escofier and Pagès (2008, chapter 4), Husson

et al. (2016, chapter 3).

2.2.2 Log-linear model

Multivariate categorical data as multidimensional contingency tables (with an order greater than two-way) display relationships between categorical variables. This kind of data can me model by a log-linear model, that is a generalized linear model for Poisson regression. The Poisson distribution is the simplest distribution for count data. The model describes association and interaction among categorical variables and its purpose is to establish dependence patterns between variables. There is no distinction between explanatory or response variables since only the cell counts are considered. The reader may find a comprehensive description in Agresti (2013, chapter 9).

For the sake of simplicity, we present the method for three categorical variables X_1, X_2 and X_3 respectively with K_1, K_2 and K_3 categories. The most general log-linear model for the three-way table $K_1 \times K_2 \times K_3$ is written as

$$\log \mu_{k_1k_2k_3} = \lambda + \lambda_{k_1}^{X_1} + \lambda_{k_2}^{X_2} + \lambda_{k_3}^{X_3} + \lambda_{k_1k_2}^{X_1X_2} + \lambda_{k_1k_3}^{X_1X_3} + \lambda_{k_2k_3}^{X_2X_3} + \lambda_{k_1k_2k_3}^{X_1X_2X_3}, \tag{3}$$

where $\mu_{k_1k_2k_3}$ is the expected frequency of the cell with $X_1 = k_1$, $X_2 = k_2$ and $X_3 = k_3$. The model-parameters are interpreted as follows: λ is the overall effect; $\lambda_{k_j}^{X_j}$ is the effect of the level $X_j = k_j$, $j = 1, 2, 3$; $\lambda_{k_jk_{j'}}^{X_jX_{j'}}$ is the interaction effect of levels $X_j = k_j$ and $X_{j'} = k_{j'}$, $1 \leq j, j' \leq 3$; finally $\lambda_{k_1k_2k_3}^{X_1X_2X_3}$ is the interaction effect between the levels $X_1 = k_1$, $X_2 = k_2$ and $X_3 = k_3$. The model (3) is called the saturated model, it includes all possible main effects and interactions between the variables. Some constraints between the parameters ensure model identifiability and the number of free parameters in the saturated model is equal to the number of cells $K_1K_2K_3$, which is why the saturated model fits the data perfectly. It reproduces exactly the observed cell frequencies and does not provide much relevant information.

The aim is to find the simplest model that fits the data adequately, that is, a more parsimonious model with less parameters. An unsaturated model is obtained by imposing the nullity of some coefficients in (3) and may be more appropriate due to simpler interpretations. Validation is performed thanks to goodness-of-fit assessment comparing the expected cell frequencies to the observed frequencies. The goodness-of-fit can be tested with the likelihood-ratio statistic:

$$G^2 = 2 \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{k_3=1}^{K_3} n_{k_1k_2k_3} \log \left(\frac{n_{k_1k_2k_3}}{\hat{\mu}_{k_1k_2k_3}} \right),$$

where $n_{k_1k_2k_3}$ and $\hat{\mu}_{k_1k_2k_3}$ are respectively the cell frequencies and the fitted values from model (3) taking into account the nullity constraint (Agresti, 1990). The G^2 statistic is used to determine the rejection or acceptance of a model. The larger the value of G^2, the more evidence there is against that the related model does fit the data adequately, hence it should not be kept.

Different types of unsaturated log-linear models correspond to different type of dependence between the variables X_1, X_2 and X_3. We will consider here only hierarchical models,

TAB 2: Different structures of log-linear models corresponding to different dependence structures. The third column "Symbol" corresponds to model notations, that is, the higher-order model term represented of each variable used in the model.

Log-linear model	Interpretation	Symbol
$\lambda + \lambda_{k_1}^{X_1} + \lambda_{k_2}^{X_2} + \lambda_{k_3}^{X_3}$	mutual independence	$(X_1,\ X_2,\ X_3)$
$\lambda + \lambda_{k_1}^{X_1} + \lambda_{k_2}^{X_2} + \lambda_{k_3}^{X_3} + \lambda_{k_2k_3}^{X_2X_3}$	independence of X_1 and (X_2, X_3)	$(X_1,\ X_2X_3)$
$\lambda + \lambda_{k_1}^{X_1} + \lambda_{k_2}^{X_2} + \lambda_{k_3}^{X_3} + \lambda_{k_1k_3}^{X_1X_3} + \lambda_{k_2k_3}^{X_2X_3}$	independence of X_1 and X_2 given X_3	$(X_1X_3,\ X_2X_3)$
$\lambda + \lambda_{k_1}^{X_1} + \lambda_{k_2}^{X_2} + \lambda_{k_3}^{X_3} + \lambda_{k_1k_2}^{X_1X_2} + \lambda_{k_1k_3}^{X_1X_3} + \lambda_{k_2k_3}^{X_2X_3}$	homogeneous association	$(X_1X_2,\ X_2X_3,X_1X_3)$

meaning that if variables are involved in high order interactions, all the lower-order interaction term must also appear. For example, if the model contains $\lambda_{k_1k_2}^{X_1X_2}$, then it also must contain $\lambda_{k_1}^{X_1}$ and $\lambda_{k_2}^{X_2}$. Table 2 summarizes the different types of resulting models which are ordered with increasing complexity. The simplest model, noted (X_1, X_2, X_3), assumes the nullity of all the interaction effects and corresponds to the mutual independence of X_1, X_2 and X_3. The model with no interaction of order 3 and no interaction of second order between X_1, X_2 and X_1, X_3 is noted (X_1, X_2X_3) and corresponds to the independence of X_1 and (X_2, X_3). The model with no interaction of order 3 and no interaction of order 2 between X_1 and X_2 is noted (X_1X_3, X_2X_3) and corresponds to the conditional independence of X_1 and X_2 given X_3. Finally, the model (X_1X_2, X_2X_3, X_1X_3) has all interactions of order 2 but no interaction of order 3 and corresponds to homogeneous association that we will explain below. One goal of the analysis of the log-linear model is to find out which is the simplest model suitably fitting the data.

We now discuss marginal and conditional association of variables. A two-way contingency table can be obtained by marginalizing out the third variable, obtaining the so-called marginal table. Associations in this table are summarized by the marginal odd ratios. The marginal odds ratio of a 2×2 table (of X_1 and X_2) is defined by

$$\theta_{X_1X_2} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}}.$$

where $\mu_{ij+} = \sum_{k_3} \mu_{ijk_3}$ are the expected marginal frequencies with $i, j = 1, 2$ and k_3 a fixed category of X_3.

The distribution of the two variables X_1 and X_2 can be displayed conditionally on different levels of X_3 using cross sections of the three-way contingency table. The associations in these cross-sections (also called partial tables) are called conditional associations and summarized by conditional odds ratios: for instance the ratio of the odds of a $2 \times 2 \times K_3$ table is defined by

$$\theta_{X_1X_2(k_3)} = \frac{\mu_{11k_3}\mu_{22k_3}}{\mu_{12k_3}\mu_{21k_3}}.$$

On the other hand, the absence of interaction of order 3 in the model (X_1X_2, X_2X_3, X_1X_3) implies that the conditional odds ratios do not depend on the category of the third conditioning variable (Agresti, 2013). This property explains the term homogeneous association.

2.2.3 Ordinal regression model

The logistic regression is a popular supervised learning method for analysing dependencies between a response categorical variable Y (binary or multiclass) and explanatory variables denoted by $\mathbf{X} = (X_1, \ldots, X_p)$. More precisely, the logistic regression is used in order to separate the effects of each variable, that is, identify the effects of an explanatory variable X_j, $j = 1, \ldots, p$, on the response variable Y. The logistic regression for a binary or multiclass response variable will be presented briefly below, for more details see for example (McCullagh and Nelder, 1989, chapter 5), (Agresti, 1990, chapter 9) or (Hothorn and Everitt, 2014, chapter 7).

Let $Y \in \{0, 1\}$ be a binary response variable. The logistic regression model is written as

$$\mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) = F(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}),$$

where $\mathbf{x} \in \mathbb{R}^p$, $\beta_0 \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^p$ and $F(t) = e^t/(1 + e^t)$, $t \in \mathbb{R}$, is the inverse logistic link function. The coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are estimated by maximum likelihood method. Equivalently, the log odds of the event $\{Y = 1\}$ given $\mathbf{X} = \mathbf{x}$ is linear in $\mathbf{x}$:

$$\log \operatorname{odds}(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \log \frac{\mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x})}{1 - \mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x})} = \boldsymbol{\beta}^T \mathbf{x}.$$

Finally, a variable X_j reveals to have an effect on the response variable if the result of the nullity coefficient test for β_j is significant, that means, β_j not equal to 0 (several nullity tests exist such as Wald test for instance).

Now, in this study, the focus lies on a categorical variable with more than two categories. Let Y be a multiclass response variable. The logistic regression for a multiclass response variable is an extension of the logistic regression for a binary one. When the categories $\{m_1, \ldots, m_q\}$ of the response variable Y are hierarchically ordered as $m_1 \prec \ldots \prec m_q$, a way to model Y is to suppose that there exists a latent unobserved continuous variable denoted $Y^* \in \mathbb{R}$, with logistic distribution F, such that

$$Y = m_k \quad \text{if and only if} \quad c_{k-1} < Y^* \leq c_k,$$

where $-\infty = c_0 < c_1 < \ldots < c_{q-1} < c_q = +\infty$ and $k = 1, 2, \ldots, q$. Then, the ordinal regression model is written as

$$\mathbb{P}(Y \preceq m_k \mid \mathbf{X} = \mathbf{x}) = F(c_k - \boldsymbol{\beta}^T \mathbf{x}), \tag{4}$$

where $k = 1, 2, \ldots, q-1$. Note that the general intercept β_0 is replaced by the set of ordered intercept paramaters c_k mentioned before. The unknown coefficients $c_1, \ldots, c_{q-1}, \beta_1, \ldots, \beta_p$ are estimated by maximimum likelihood.

Model (4) is also called the proportional-odds model due to the following property: the log odds ratio of $\{Y \preceq m_k\}$ at $\mathbf{X} = \mathbf{x_1}$ and $\mathbf{X} = \mathbf{x_2}$ is given by

$$\log \frac{\operatorname{odds}(Y \preceq m_k \mid \mathbf{X} = \mathbf{x_1})}{\operatorname{odds}(Y \preceq m_k \mid \mathbf{X} = \mathbf{x_2})} = -\boldsymbol{\beta}^T (\mathbf{x_1} - \mathbf{x_2}),$$

and does not depend on the category m_k.

3 Results

This section aims at giving a multivariate analysis of Franche-Comté road crash data by using the above described methods. Our analysis begins by performing MCA analyses (temporal and spatial) on road crash variables and provides insights multivariate road crash related variables by using geometric data visualization. This method is a powerful tool to distinct non-trivial category associations if it is the case. Log-linear models have been fitted in second time to describes the association and interaction patterns among the set of categorical variables related to road crash gravity. Finally, the ordinal regression model allows to quantify the effects of each explanatory variable on the ordered response variable road crash gravity.

This study used open-source R software packages `FactoMineR` (Lê et al., 2008) and `factoextra` (Kassambara and Mundt, 2019) to perform MCA, `glm` function to perform log-linear models, then packages `MASS` (Venables and Ripley, 2002) and `ordinal` (Christensen, 2019) to perform ordinal logistic regression. Graphics were plotted with `ggplot2` package (Wickham, 2016).

3.1 MCA of road crash temporal variables

We conducted a MCA temporal analysis by considering all variables described in Section (2.1) except the spatial variables *department* and *canton*. Fig (2) gives the percentages of variance explained by each of the first ten axes built by the MCA analysis. The first three-factorial axes explain 22,09% of the total variance and only these axes were kept for further analysis. Two-dimensional geometrical representations are given in Fig (3)-(5) and interpreted below. Each time, only the 25 best represented categories have been plotted.

The two-dimensional map in Fig (3) gives the representation of categories on the plane made by axis 1 and axis 2 and it accounts for 16,07% of the total inertia. The more categories a variable has, the more it contributes to the inertia. The variables *season* and *weather* are not represented on the first factorial plane since they are very poorly represented on this plane. Next, categories with the greatest contribution to the axis 1 are "night" (13,26%), "none" from *obstacle* (10,57%) and "midnight_6am" (9,26%) and respectively, "pedestrian" (28,45%), "urban" (19,04%) and "other_kind" from *collision* (11,46%) for axis 2.

In this first factorial plane, axis 1 shows the contrast between weekday accidents (accidents occurring during the week) and weekend ones (accidents occurring during the weekend). Weekday accidents are more frequent during the day and mostly around lunch time, in urban areas, on communal roads and are not associated to alcohol or drug consumption. These accidents are more likely to happen on straight roads, at intersections and caused by collisions between several vehicles. Weekend accidents, instead, are more frequent during the night and mostly between 8 pm and 6 am, outside urban areas and involve more frequently drug consumers. These accidents occur mainly on curve roads and no external factors seem to impact (out of intersections, no bumped mobile obstacles or no collisions). To sum up, this first factorial axis is related to fatal accidents.

Axis 2 provides a similar information as axis 1: it opposes accidents occurred during

weekday time, in urban area, at intersection and on straight road to accidents occurred during weekend time, outside urban area, out of intersection and on curve road. However, axis 2 stresses the fact that fatal road crash are more likely to occur between midnight and 6 am especially when alcohol and drugs have been consumed.

Geometrical representations derived in MCA plot closely associated categories and unassociated ones further apart. The first factorial plane reveals several strong associations among categories: categories "straight", "intersection", "weekday", "substance_none", "11am_3pm", "collision_usual" are strongly associated to categories "obstacle_vehicle"; categories "curve", "out_of_intersection, "week_end", "unurban" are strongly associated to "fatal"; "night", "8pm_11pm", "midnight_6am", "obstacle_none", "alcohol" and "alcohol_drug" in the same way.

On the other hand, this factorial plane also shows that some categories are far from the others, this results from their lower frequencies. Indeed, as it is given in Tab 1, "obstacle_pedestrian" and "type_road_other_kind" represents respectively only 8,28% and 2,53% of road accidents.

Fig (4) gives the two-dimensional map of axis 1 and axis 3 and it explains $15,32\%$ of the total inertia. The variable *area* has been omitted from this geometrical representation due to its poor representation quality. Categories that contribute the most to axis 3 are categories "night" (11,82%), "none" (11,08%) of the variable *collision* and category "winter" (10,96%) of the variable *season*.

In this factorial plane, axis 3 suggests that summer accidents tend to differentiate from winter ones. Summer accidents are more likely to occur during the day, around lunch time and on week-end time. They are globally associated to no alcohol or drug consumption, happening on curve roads, out of intersections and other kind of collision. Winter accidents instead occur more frequently during the night, the week time and on national roads. They are also mostly associated with substances consumed, happening on straight roads, at intersections and collisions with vehicles. The associations with straight roads, intersections and collisions with vehicles seem to be caused by weather ("other_kind") which is generally snowy in winter. In addition, the axis 3 specifies that winter accidents are more likely to be fatal.

Two groups of strongly associated categories stand out in the second factorial plane: the group formed by "obstacle_vehicle", "collision_usual", "intersection", "straight", "weekday", "substance_none", "day", "11am_3pm" and the other group formed by "summer"; "night", "8pm_11pm", "midnight_6am", "alcohol_drug". These groups resonates to those mentioned previously.

Fig (5) gives the two-dimensional map of axis 2 and axis 3 and it explains 12,79% of the total inertia. From the thirteen variables used, the variables *substance*, *week* and *intersection* are very poorly represented on this map and are omitted from this geometrical representation. This plot emphasizes the differences between serious and fatal accidents which tend to be strongly associated with lunch time and respectively night time. We distinguish two groups of close categories: "winter", "night", "fatal", "8pm_11pm", "obstacle_vehicle", "collision_usual" and "weather_other_kind"; "type_road_departmental",

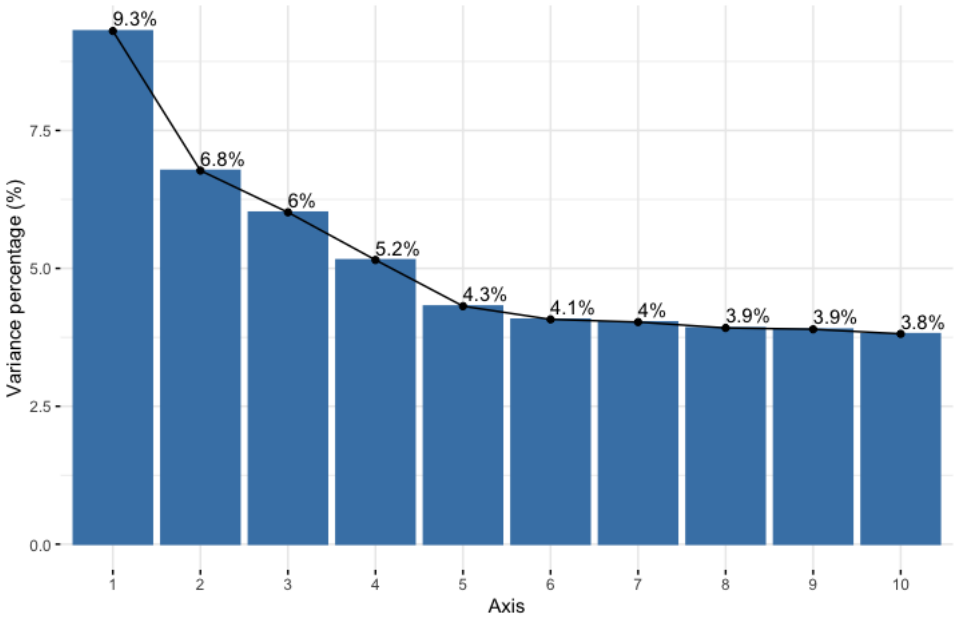

FIG 2: MCA temporal analysis: variance percentage explained by the first 10 axes.

"serious", "weather_normal", "spring", "curve", "day", "summer", "11am_3pm" and "obstacle_none".

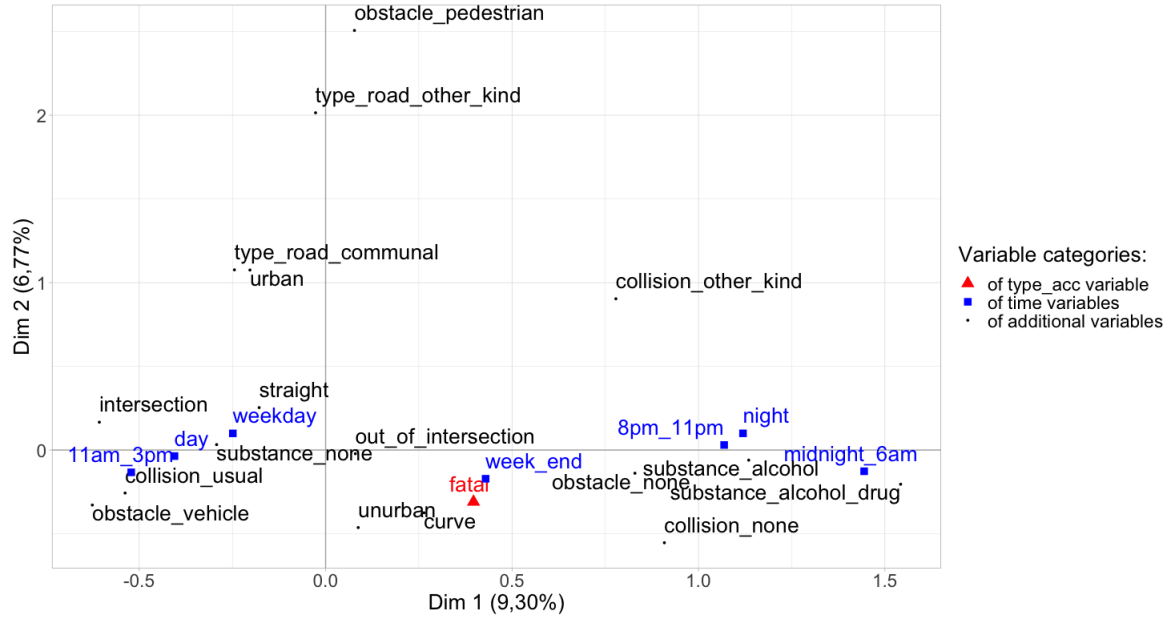

FIG 3: MCA temporal analysis: factorial plane made by axis 1 and axis 2.

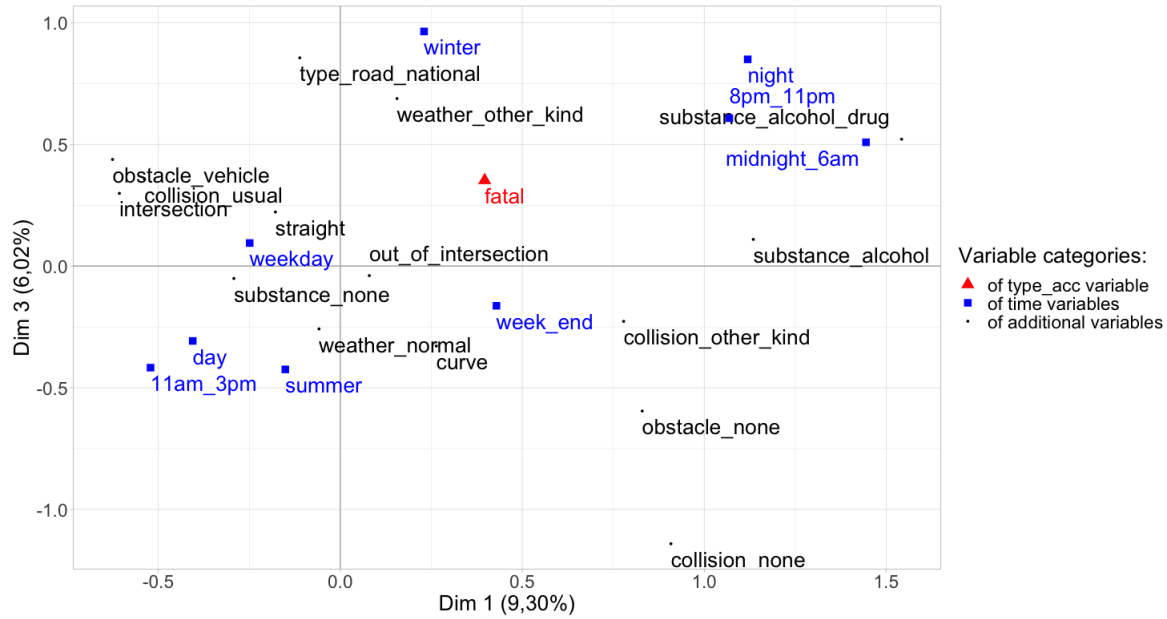

FIG 4: MCA temporal analysis: factorial plane made by axis 1 and axis 3.

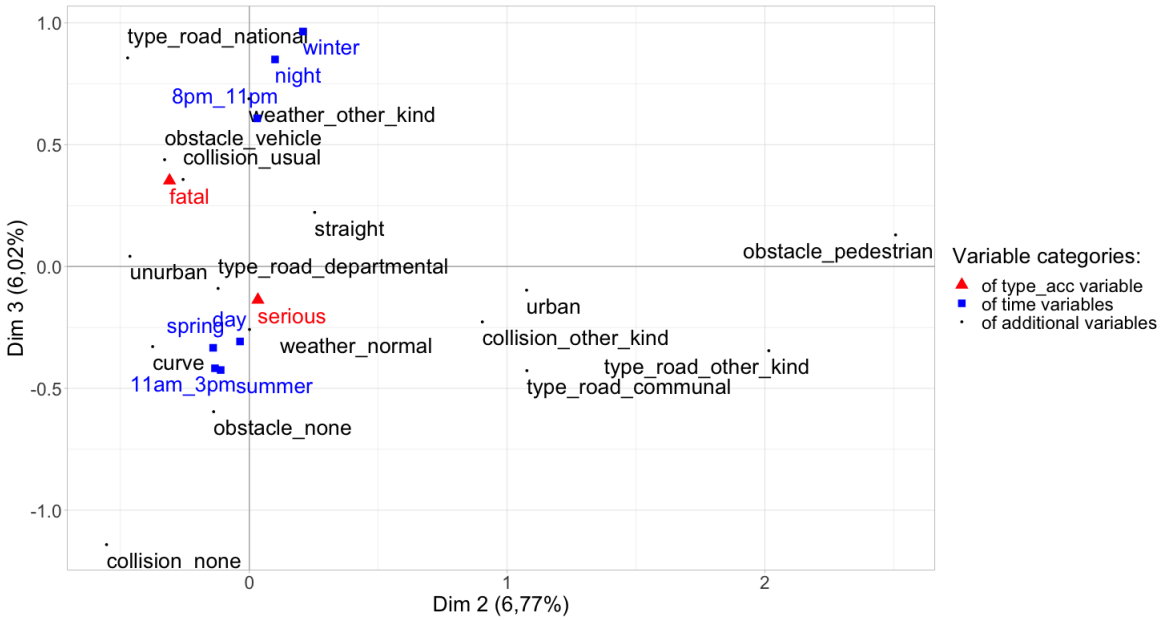

FIG 5: MCA temporal analysis: factorial plane made by axis 2 and axis 3.

3.2 MCA of road crash spatial variables

Due to the several categories of the variable *canton*, the spatial analysis of Franche-Comté has been splitted into two analyses corresponding to Doubs and Jura departments. To conduct this spatial analysis, all the variables have been used except temporal ones (*season*, *week*, *daytime* and *time*) for this spatial analysis. We will interpret only relationships from categories situated close one to another in the plot.

3.2.1 Doubs department

For the spatial analysis of Doubs department, the first four factorial axes explain 21,22% of the total inertia. Note that only the first 25 categories with the most important representation qualities were plotted.

Plot made by axis 1 and axis 2 given in FIG (7) explains 12,74% of the total inertia. Not all the variables are plotted, the variable *weather* is less well represented than the other categories and it does not figure on the plot. Categories which contribute the most for axis 1 are "none" (20,77%) and "vehicle" (14,06%) from *obstacle*, then "usual" from *collision* (12,88%); and for axis 2"pedestrian" (22,53%), "urban" (18,92%) and "other_kind" from *collision* (9,73%).

As mentioned before, associations can be highlighted by the proximity of categories on the factorial plot. Two groups of close categories with spatial connotations stand out: "type_road_other_kind", "type_road_communal", "Bethoncourt", "urban" and "Valentigney"; "Besançon", "collision_usual" and "obstacle_vehicle". The first group emphasizes that Bethoncourt and Valentigney accidents are more frequent in urban areas. The second one strongly insists that the canton of Besançon is more conducive to collisions with vehicles.

Additional plots made by combinations of axis 1, 2, 3 and 4 explain between 8,48% and

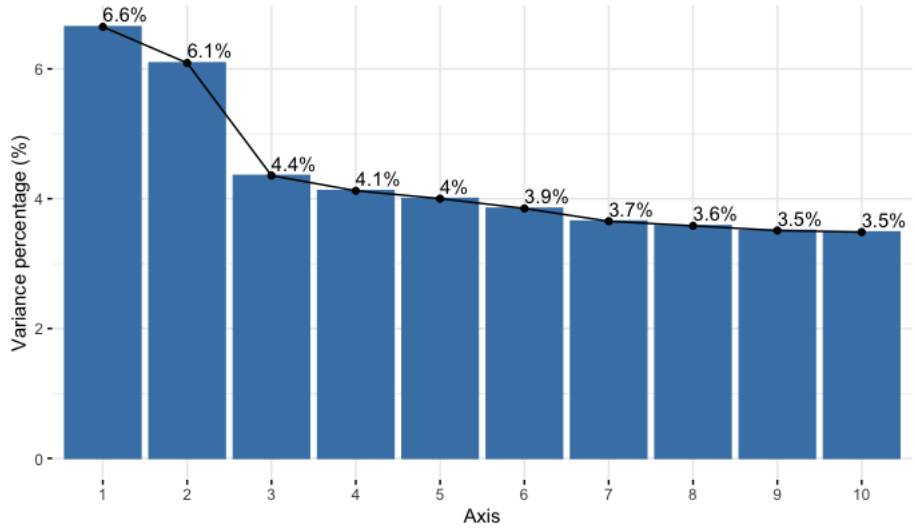

FIG 6: MCA spatial analysis: variance percentage on the first 10 axes, Doubs department.

11,01% of the total inertia, it should be noted that compared to each other the associations do not differ. These plots indicates, in addition to what was said before, that Besançon accidents tend to be fatal when the weather is bad and more characterized by bumped pedestrians; Bethoncourt accidents are mostly associated with substance consumption and very strongly associated to "drug"; Baume-les-Dames accidents are more likely to be fatal; Besançon, Saint-Vit and Ornans accidents are mainly similar and more frequent on communal and national roads; most of Maîche accidents are not caused by collisions.

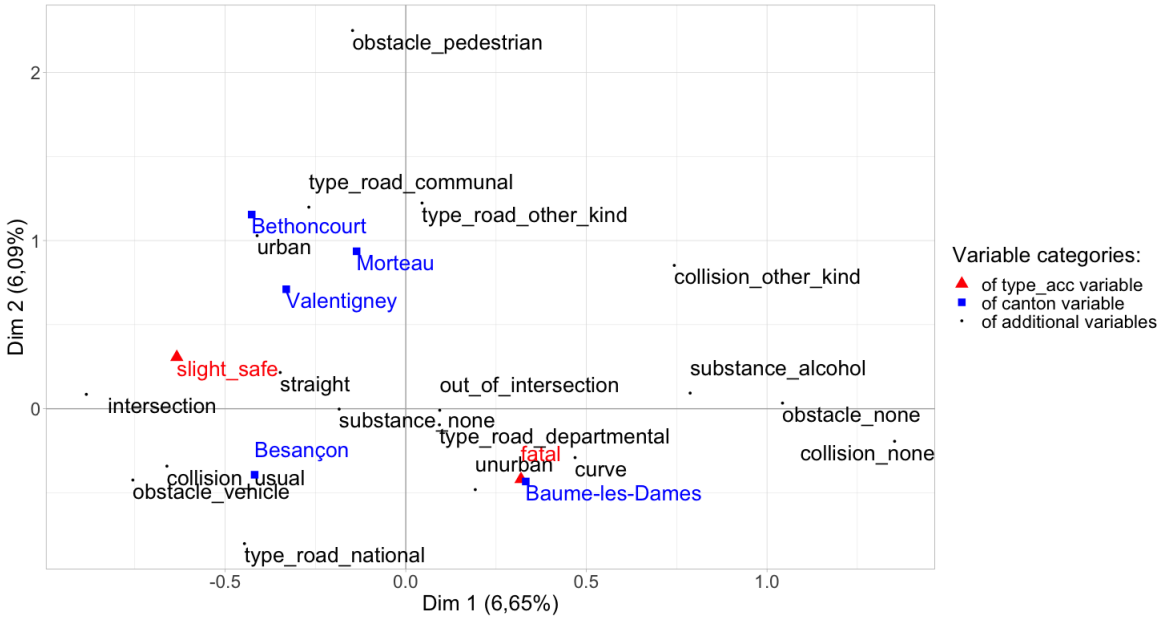

FIG 7: Factorial plane made by axis 1 and axis 2.

3.2.2 Jura department

For the spatial analysis of Jura department, the first four factorial axes explain 21,88% of the total inertia. Note that only the first 25 categories with the most important representation qualities were plotted.

Plot made by axis 1 and axis 2 explains 13,08% of the total inertia and is given in Fig (9). Not all the variables are plotted, the variable *weather* is less well represented than the other categories and it does not figure on the plot. Categories which contribute the most for axis 1 are "none" (18,34%) and "vehicle" (13,79%) from *obstacle*, then "usual" from *collision* (11,94%); and for axis 2 "pedestrian" (24,89%), "urban" (18,69%) and "other_kind" from *type_road* (11,23%).

We can distinguish three groups of close categories, with components of the *canton* variable that stand out in this factorial plane: "Saint-Lupicin", "fatal", "drug", "alcohol_drug", "alcohol" and "obstacle_none"; "Authume", "Saint-Laurent-en-Grandvaux", "un-urban", "type_road_departmental" and "out_of_intersection"; "Champagnole", "urban" and "type_road_communal". The first group highlights that accidents in the canton of Saint-Lupicin are more likely to be fatal and associated to alcohol and drug consumption. The second one tells that accidents happening in Authume and Saint-Laurent-en-Grandvaux are more frequent in non urban areas, on departmental roads. Then, the third group of close categories shows that accidents in the canton of Champagnole are occuring more commonly in urban areas and on communal roads. Finally, the structure of this factorial plane tells that accidents happening in Authume, Saint-Laurent-en-Grandvaux and Saint-Lupicin are more likely to be fatal.

Additional plots made by combinations of axis 1, 2, 3 and 4 explain between 8,81% and 11,24% of the total inertia, it should be noted that compared to each other the associations do not differ. The cantons of Champagnole, Morez and Saint-Laurent-en-Grandvaux have been associated to each other in many factorial planes, it seems that accidents are more likely to happen in these cantons when the weather is qualified as "other_kind" (cloudy, rainy or snowy). This is opposed to accidents happening in Authume and Bletterans where accidents are more frequent when the weather is "normal". Many cantons have been related to alcohol or drug consumption: accidents occurring in the Dole canton are more commonly associated to drug, Moirans-en-Montagne and Saint-Claude cantons are instead matched to alcohol. Finally, the canton where accidents happen in higher proportion on highway is Dole, it is also the canton where accidents are more likely to be fatal, and finally, the canton where pedestrians are bumped in much higher amounts is Champagnole.

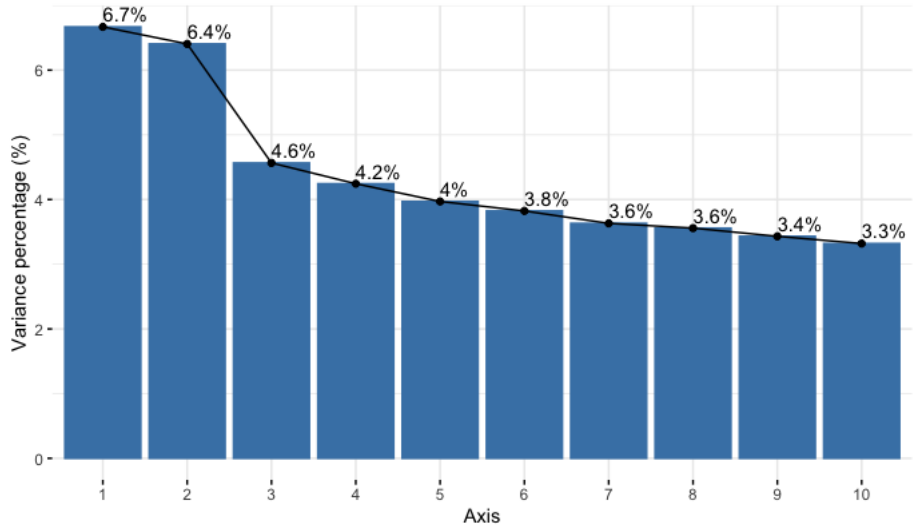

FIG 8: MCA spatial analysis: variance percentage on the first 10 axes, Jura department.

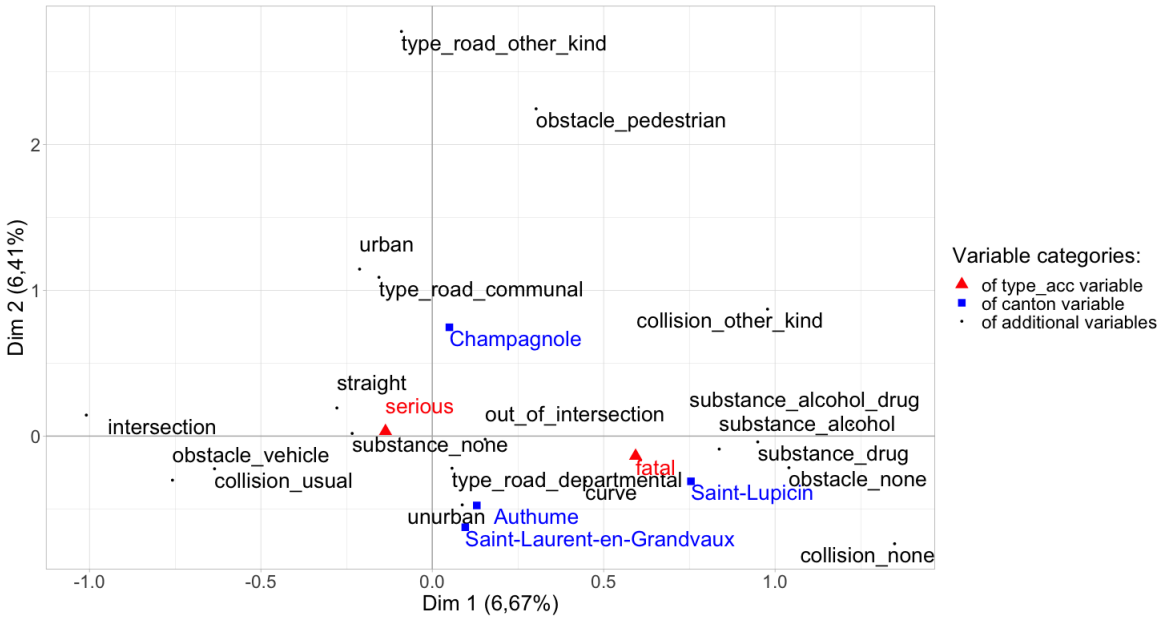

FIG 9: Factorial plane made by axis 1 and axis 2.

3.3 Log-linear modelling

The MCA performed in sections 3.1 and 3.2 reveals that there are associations between the gravity of the accidents (*type_acc*) and the drug/alcohol consumption (*substance*). Moreover, we could see during the spatial analysis that these associations are observed within each department. In order to describe more thoroughly the association patterns between these categorical variables and the variable *department*, several hierarchical log-linear models have been fitted on the related three-way contingency table (corresponding to the column "Observed values" of the table TAB 3).

TAB 4 gives the likelihood-ratio statistic G^2 of each hierarchical model and the 5%-level

associated significance test (p-value). A model fits the data well if the null hypothesis of the goodness-of-fit test is accepted. The p-values in TAB 4 show that all the models fit the data poorly except (TS, TD, SD) which is close to the observed data (corresponding to the column "Fitted values" of the table TAB 3). This unsaturated hierarchical final model has been kept as the objective was to find the simplest model that fits the data adequately. It is written as

$$\log \mu_{kk'k''} = \lambda + \lambda_k^T + \lambda_{k'}^S + \lambda_{k''}^D + \lambda_{kk'}^{TS} + \lambda_{kk''}^{TD} + \lambda_{k'k''}^{SD}, \tag{5}$$

where k is "slight_safe", "serious" or "fatal" for the categorical variable *type_acc* (T) ; k' is "none", "alcohol", "drug" or "alcohol_drug" for the variable *substance* (S) ; and k'' is "Doubs", "Haute_Saone", "Jura" or "Terr_Belfort" for the variable *department* (D).

This is the model with no three-factor interaction. The conditional association terms appear for each pair of variables, this means that no pair is conditionally independent. The odds ratios related to this model have been calculated and are given in TAB 5. Note that the baseline categories of *type_acc*, *substance* and *department* were respectively "slight_safe", "none" and "Doubs". For instance, the odds ratio relating the level "serious" of *type_acc* and "alcohol_drug" of *substance* at the level "Doubs" of *department* is calculated as

$$\frac{25,64 \times 221,37}{1149,58 \times 3,96} = 1,25.$$

Remind that the no three-factor interaction model means that the association between two variables is identical at each level of the third variable. Hence, in the same way, calculating this odds ratio with the fitted values regarding to the levels "Haute_Saone", "Jura" or "Terr_Belfort" of *department* would have given the same result.

In general marginal odds ratios may differ from conditional ones in a no three-factor interaction model, however in this case marginal and conditional odds ratio are very close. This means that controlling or ignoring the third variable does not change significantly the association between the two variables. Only conditional odds ratio will be interpreted below as the interpretations of marginal ones are the same.

Regarding substances consumption, the odds for an accident to be serious or fatal increases when alcohol, drug or both are consumed. Indeed, the odds ratios for an accident to be serious are estimated to be respectively 1,64, 1,94 and 1,25 higher than slight when alcohol, drug and both are consumed compared with no consumption. Similarly, the odds ratios for an accident to be fatal are estimated to be respectively 2,76, 4,15 and 5,93 higher than slight when alcohol, drug and both are consumed. The highest risk for an accident to be fatal corresponds to drug and alcohol consumption, almost two times larger than alcohol consumption. Regarding the department where the accident happens, the odds to be serious or fatal decreases only for the department Territoire de Belfort. The odds ratio for an accident to be serious is estimated to be 0,48 times lower than slight when it occurs in this department compared with Doubs. Similarly, the odds for an accident to be fatal is estimated to be 0,39 times lower than slight when it occurs in Territoire de Belfort compared with Doubs. The highest risk for an accident to be fatal is when it occurs in Jura department compared with Doubs, almost four times larger than in Territoire de Belfort.

TAB 3: Three-way contingency table with *type_acc*, *substance* and *department* as categorical variables. Left side correspond to the observed values, right one is equal to the predicted frequencies from the log-linear model (TS, TD, SD).

		Observed values			Fitted values (TS, TD, SD)		
		type_acc			*type_acc*		
department	*substance*	slight_safe	serious	fatal	slight_safe	serious	fatal
Doubs	alcohol_drug	6	26	24	3,96	25,64	26,40
	drug	2	30	13	2,87	28,79	13,34
	alcohol	27	227	85	26,81	228,99	83,20
	none	220	1150	250	221,37	1149,58	249,06
Haute_Saone	alcohol_drug	2	15	20	2,48	17,60	16,92
	drug	4	21	14	2,32	25,60	11,08
	alcohol	11	165	35	15,58	145,92	49,50
	none	121	658	144	117,62	669,88	135,50
Jura	alcohol_drug	0	12	10	0,86	9,40	11,74
	drug	2	27	12	1,49	25,29	14,22
	alcohol	12	131	75	10,03	144,34	63,63
	none	89	802	201	90,62	792,97	208,42
Terr_Belfort	alcohol_drug	2	8	8	2,70	8,36	6,94
	drug	0	8	2	1,32	6,32	2,36
	alcohol	8	19	8	5,58	22,75	6,67
	none	62	157	25	62,40	154,58	27,03

3.4 Ordinal regression modelling

An ordinal regression has been fitted on the response variable *type_acc* (ordered as slight_safe, serious and then fatal). The analysis was performed with all the explanatory variables except *canton* (due to too many categories).

Note that the categorical variable *type_acc* is considered as a response variable. It is relevant because it gives the severity of accidents and the aim of this study lies on understanding how the gendarmerie can avoid serious injuries.

The initial dataset has been split randomly into two sets: a training set (seventy-five percent of the initial one) and a test set (the remaining twenty-five percent). An ordinal regression model has been fitted on the training set, the results are given in TAB ??. Only five explanatory variables reveal to have an effect on the response variable: *time*, *substance*, *department*, *collision* and *area*. This model, with full parameters, gives a misclassification error of 28,84% on the test set.

Next, a variable selection has been performed by using AIC criterion (with backward selection). The final model is composed by the categorical variables *time*, *substance*, *department*, *collision* and *area*. This model gives a misclassification error of 29,00% on the test

Tab 4: Goodness-of-Fit Tests for log-linear models relating *type_acc* (T), *substance* (S) and *department* (D).

Model	G^2	p-value
(T, S, D)	246,39	0,00
(T, SD)	218,97	0,00
(S, TD)	179,65	0,00
(D, TS)	125,36	1,09e-12
(TS, TD)	58,61	3,99e-4
(TS, SD)	97,94	6,72e-11
(TD, SD)	152,23	0,00
(TS, TD, SD)	27,38	0,07
(TSD)	0,00	–

Tab 5: Odds ratio estimated from (TS, TD, SD) log-linear model. The table is divided into two parts, which are also splitted up into two parts: conditional odds ratios in top have been calculated respectively in left and right sides controlling levels of *department* and *substance* variable, and marginal odds ratios in bottom have been calculated respectively in left and right sides ignoring *department* and *substance* variable. Each odds ratio has been calculated as each level of *type_acc*, *substance* and *department* was opposed respectively to "slight_safe", "none" and "Doubs".

Conditional odds ratios					
	serious	fatal		serious	fatal
alcohol_drug	1,25	5,93	Haute_Saone	1,10	1,02
drug	1,94	4,15	Jura	1,69	2,05
alcohol	1,64	2,76	Terr_Belfort	0,48	0,39
Marginal odds ratios					
	serious	fatal		serious	fatal
alcohol_drug	1,08	4,92	Haute_Saone	1,11	1,06
drug	1,91	4,07	Jura	1,68	1,98
alcohol	1,66	2,78	Terr_Belfort	0,47	0,41

set, a score very close to the previous one. The odds ratio of these variables are given in Fig (10). Only odds ratios with confidence intervals not containing the value 1 (represented by the vertical dotted line) are interpreted.

Tab 6: Ordinal regression model results with *type_acc* as ordered response variable. The item * means that the p-value of the nullity coefficient test is less than 0,05. The category in parentheses correspond to the baseline category of the above categorical variable.

Attribute	Ordinal regression results				Attribute	Ordinal regression results			
	Coefficients	Estimate	p-value			Coefficients	Estimate	p-value	
substance	alcohol_drug	1,16	5,19e-8	*	*department*	Haute_Saone	-0,01	0,89	
(none)	drug	0,85	4,32e-5	*	(Doubs)	Jura	0,35	5,87e-5	*
	alcohol	0,46	1,14e-5	*		Terr_Belfort	-0,58	2,67e-4	*
season	spring	0,01	0,30		*obstacle*	vehicle	-0,02	0,85	
(winter)	summer	0,11	0,27		(none)	pedestrian	0,23	0,15	
	autumn	-0,05	0,67			other_kind	0,23	0,30	
week	week_end	-0,04	0,53		*shape_road*	curve	-0,01	0,90	
(weekday)					(straight)				
daytime	day	0,08	0,51		*collision*	usual	-0,02	0,91	
(night)					(none)	other_kind	0,25	0,04	*
time	11am_3pm	-0,08	0,50		*type_road*	communal	0,43	0,14	
(7am_10am)	4pm_7pm	-0,26	0,02	*	(highway)	departmental	0,39	0,15	
	8pm_11pm	-0,10	0,52			national	0,47	0,10	
	midnight_6am	0,18	0,27			other_kind	0,34	0,34	
weather	other_kind	0,00	0,97		*intersection*	intersection	-0,12	0,31	
(normal)					(out_of_intersection)				
area	unurban	0,74	3,97e-16	*					
(urban)									

Regarding the odds ratio, the highest risk for an accident to be serious or fatal is if substances have been consumed by one of the drivers involved. Indeed, the most two important odds ratios are alcohol_drug and drug which are equal to $\exp(1,16) = 3,19$ and $\exp(0,85) = 2,34$ respectively. The risk for an accident to be serious or fatal increases if the accident happens in non urban areas (odds ratio 2,10) or in the Jura department (odds ratio 1,42 with Doubs as reference). Accident involving uncommon collision also have a slightly increased risk (odds ratio 1,28). On the opposite, two categories have a protective effect and are associated with lower risk of serious or fatal accident. This is the case for Territoire de Belfort department (odds ratio 0,56 with Doubs as reference) and for the occurrence time between 4pm and 7pm (odds ratio 0,77).

Each department odds ratio has been represented on the map in Fig 11. Remind that the odds of Haute-Saône department were not significant and hence meaningless. The odds ratio of Doubs department is equal to 0 as it was the baseline category for the *department* variable. As a symbol, Jura and Territoire de Belfort departments have been represented in red and blue respectively due to their odds ratio: the riskiest and the less risky.

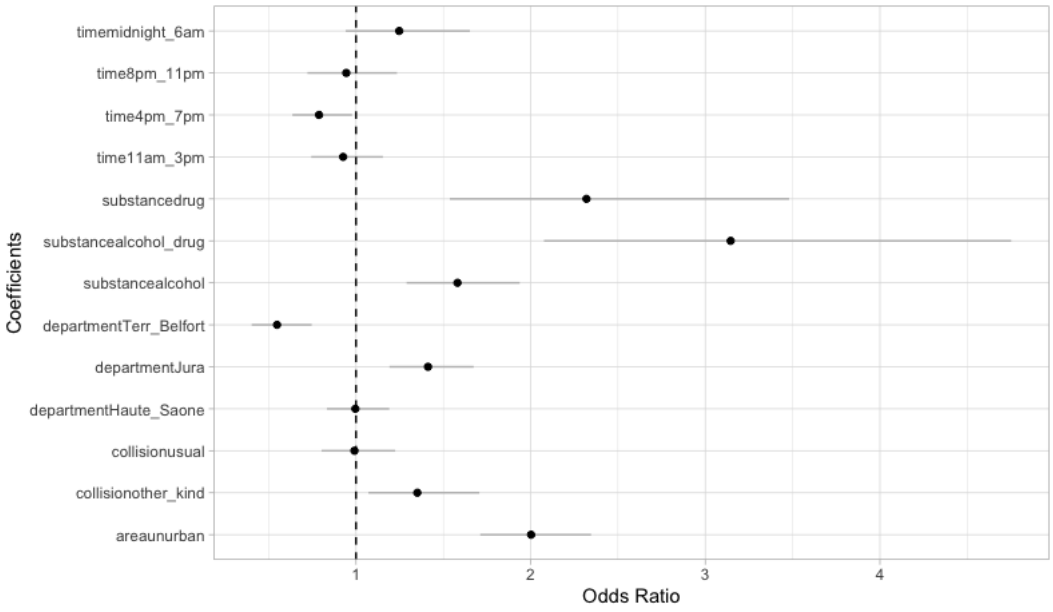

FIG 10: Odds ratios obtained by ordinal regression model. Grey lines represent the confidence intervals and black points the values of the odds ratios.

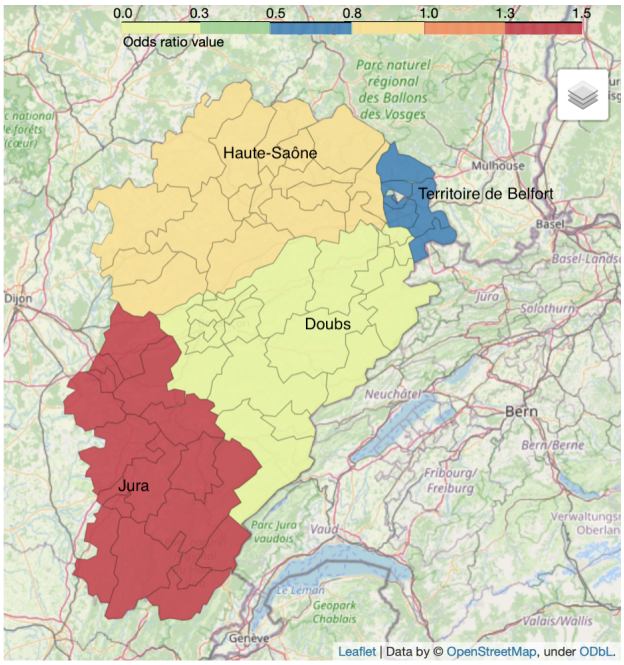

FIG 11: Franche-Comté map with department odds ratio. Each color corresponds to a department with its odds ratio value considering the Doubs department as the reference and each small division corresponds to a canton.

4 Conclusion

A study of accidents with the purpose of mitigating the crash severity is critical for the well-being of a society and the safety concern posed by road crashes. The aim of this work was to understand factors which are the most influential in road accidents from the

French region Franche-Comté. To respond to these issues, three statistical methods were used: Multiple Correspondence Analysis (MCA) and log-linear model and ordinal logistic regression.

MCA, the only unsupervised or descriptive statistical method used in this study, allowed to assess relationships between the categorical variables and examine the associations between the different categories. Geometric representations of data in smaller dimension spaces were produced and proximities between several road crash related categories have been observed. This analysis allowed to establish a global vision of the data and to draw up temporal and spatial profiles of accidents occurring in the Franche-Comté region. Regarding the MCA temporal analysis, several associations have been highlighted. We remarked that accidents occurring during the week are different from those occurring during the weekend. Indeed, weekend accidents are more likely to happen during the night, be fatal and associated to alcohol/drug consumption. There was also a contrast between summer and winter accidents. The MCA spatial analysis revealed that several cantons of the Franche-Comté region are strongly related to alcohol/drug consumption (Bethoncourt, Saint-Lupicin or Dole) or to fatal accidents (Besançon, Baume-les-Dames, Authume, Saint-Laurent-en-Grandvaux, Saint-Lupicin and Dole).Bethoncourt, Besançon and Baume-les-Dames are situated in the Doubs department and others in the Jura department.

The log-linear model was used next in order to evaluate dependencies between the road crash gravity, the alcohol/drug consumption and Franche-Comté departments. This tool models the multidimensional contingency table formed by these three categorical variables and describes associations and interactions among them. It allowed establishing patterns. The selected model concludes on no interaction between the categorical variables *type_acc*, *substance* and *department*. It corresponds to the model of homogeneous association, which means that each pair of variable were conditionally dependent. Odds ratios estimated from this model allowed to quantify the risks about alcohol/drug consumption and the department where the accident happened. We remarked that the highest risks for a serious or fatal accident to happen are if drug, alcohol or both are consumed. Conversely, the lowest risk for an accident to be serious or fatal is if it happens in the Territoire de Belfort department. The Jura department was, instead of the previous one, a location which increases this risk.

The ordinal regression allowed the study to assess each effect of road crash related factors on the road crash gravity. Eight circumstances revealed to be influential on the accident gravity: the consumption of alcohol, drug or both; the period of the day between 4pm and 7pm; the roads situated outside urban areas; the roads situated in Jura or Territoire de Belfort departments; collisions qualified as "other_kind" (not usual as frontal or rear-end for example). Odds ratios estimated from this model allowed to quantify the risks due to each of these circumstances. Similarly to the log-linear analysis, the risk of an accident being fatal is highest if alcohol and drugs are consumed and lowest if the accident happens in the Territoire de Belfort department.

The results obtained with these three methods allow us to conclude that the most important factor to take into account for road crashes in Franche-Comté is the alcohol/drug consumption. As expected, this factor strongly influences the nature of accidents. Hence,

based on results obtained with this statistical study, more efforts should be gathered by the National Gendarmerie of Besançon to prevent the alcohol/drug consumption especially in the cantons which were associated to this factor. For example, more alcohol/drug tests and driver awareness measures can be performed.

In order to be more precise for the spatial analysis, the future study would be focused on how GPS coordinate can be used to prevent accidents.

References

Abdel-Aty, M. A. and Abdelwahab, H. T. (2000). Exploring the relationship between alcohol and the driver characteristics in motor vehicle accidents. *Accident Analysis and Prevention*, pages 473–482.

Abdel-Aty, M. A., Chen, C., and Schott, J. R. (1998). An assessment of the effect of driver age on traffic accident involvement using log-linear models. *Accident Analysis and Prevention*, 30(6):851–861.

Agresti, A. (1990). *Categorical Data Analysis*. Wiley.

Agresti, A. (2013). *Categorical Data Analysis*. Wiley-Blackwell, 3nd edition.

Benzécri, J.-P. (1982). *Histoire et Préhistoire de l'Analyse des Données*. Paris: Dunod.

Benzécri, J.-P. e. a. (1973). *L'Analyse des Données: L'Analyse des Correspondences*. Paris: Dunod.

Bièvre, D. (2017). Acteurs de contingence et insécurité routière: les limites de la statistique descriptive. Technical report.

Christensen, R. H. B. (2019). ordinal: Regression models for ordinal data. R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

Das, S., Avelar, R., Dixon, K., and Sun, X. (2018). Investigation on the wrong way driving crash patterns using multiple correspondence analysis. *Accident Analysis and Prevention*, 11:43–55.

Das, S. and Sun, X. (2015). Factor association with multiple correspondence analysis in vehicle-pedestrian crashes. *Trasportation Research Record: Journal of the Transportation Research Board*, 2519:95–103.

Das, S. and Sun, X. (2016). Association knowledge for fatal run-off-road crashes by multiple correspondence analysis. *IATSS Research*, 39:146–155.

Escofier, B. and Pagès, J. (2008). *Analyses factorielles simples et multiples*. Dunod, 4ème édition edition.

Fort, E., Gadegbeku, B., Gat, E., Pelissier, C., Hours, M., and Charbotel, B. (2019). Working conditions and risk exposure of employees whose occupations require driving on public roads – factorial analysis and classification. *Accident Analysis and Prevention*, 131:254–267.

Greenacre, M. (1989). The carroll-green-schaffer scaling in correspondence analysis: a theoretical and empirical appraisal. *Journal of Marketing Research*, 26:358–365.

Greenacre, M. (2006). From simple to multiple correspondence analysis. In Greenacre, M. and Blasius, J., editors, *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall/CRC.

Greenacre, M. and Blasius, J. (2006). *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall/CRC.

Hothorn, T. and Everitt, B. S. (2014). *A Handbook of Statistical Analyses Using R*. CRC Press, third edition edition.

Husson, F., Lê, S., and Pagès, J. (2016). *Analyse des données avec R*. Presses Universitaires de Rennes, 2ème édition edition.

Kassambara, A. and Mundt, F. (2019). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.6.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, 25(1):1–18.

Le Roux, B. and Rouanet, H. (2004). *Geometric Data Analysis: From Correspondence Analysis to Structured Data*. Dordrecth: Kluwer.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall/CRC, second edition edition.

Mekonnen, B. (2018). Risk factors of road traffic accidents and its severity in north shewa zone, amhara region, ethiopia. *American Journal of Theoretical and Applied Statistics*, 7(4):163–166.

ONISR (2019). *La sécurité routière en France – Bilan de l'accidentalité de l'année 2018*. ONISR, Paris.

Rezapour, M. and Ksaibati, K. (2018). Application of multinomial and ordinal logistic regression to model injury severity of truck crashes, using violation and crash data. *Springer*.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Yannis, G., Golias, J., and Papadimitriou, E. (2005). Driver age and vehicle engine size effects on fault and severity in young motorcyclists accidents. *Accident Analysis and Prevention*, pages 327–333.