Université de Lille

Faculté d'Ingénierie et Management de la Santé (ILIS)

Master Data Science pour la Santé

MEMOIRE DE FIN D'ETUDES

2ème ANNEE DE MASTER

Logistic Model For Analyzing Risk Factors Associated With A Road Crash Body Severity Index

Par Soumaya EL ABBOUTI

Sous la direction de Assi L. N'GUESSAN

2019/2020

<u>Composition des membres du jury :</u>

Président de jury : Benjamin GUINHOUYA

2ème membre du jury : Djamel ZITOUNI

3ème membre de jury : Mohamed LEMDANI

Date de Soutenance : 03 Octobre 2020

Faculté d'Ingénierie et Management de la Santé - ILIS

42 rue Ambroise Paré

F-59120 LOOS

Soumaya El abbouti
Master 2 Data Science pour la Santé
2019/2020

MODÈLE LOGISTIQUE D'ANALYSE DE FACTEURS DE RISQUE ASSOCIÉS À UN INDICE DE GRAVITÉ CORPORELLE DES ACCIDENTS DE LA ROUTE

***Résumé* :**
56 019 accidents de la route en 2019 ont provoqués environ 1,35 million de morts et environ 20 à 50 millions de blessés. Cette étude vise à améliorer la prise en charge des victimes d'accidents de la route dans les hôpitaux. Pour cela, identifier rapidement le niveau de gravité corporelle lors d'un accident peut être intéressant et utile.

Le principe est donc de mettre en œuvre un **modèle logistique** pour prédire et expliquer le niveau de gravité corporelle (également appelé **indice MAIS**) en fonction de plusieurs facteurs de risque d'accidents de la route. Dans ce projet, nous nous intéressons particulièrement à deux types de véhicules: les Véhicules Légers (**VL**) et les véhicules à Deux-Roues Motorisés (**2RM**).
Certaines **mesures d'évaluation** ont été utilisées pour connaître la qualité de l'ajustement des modèles et ces dernières ont été validés par l'utilisation de la courbe ROC ainsi que d'une matrice de confusion.

A partir des résultats, il a été observé, par exemple, qu'un utilisateur touché au crâne lors d'un accident d'un VL a 5 fois plus de risque d'avoir un niveau de gravité corporelle plus élevé (MAIS3+). Un usager 2RM touché au crâne, lui, a trois fois plus de risque d'être gravement blessé.

Mots-clés* : *Modèle Logistique, VL, 2RM, indice MAIS, Mesures d'Evaluation

LOGISTIC MODEL FOR ANALYZING RISK FACTORS ASSOCIATED WITH A ROAD CRASH BODY SEVERITY INDEX

***Abstract* :**
56 019 road accidents in 2019 which cause approximately 1.35 million deaths and 20 to 50 million injuries. This study aimed to improve the care of road crashes victims in hospitals. For that, identify quickly the bodily severity level in an accident is useful.

The principle is therefore to implement statistical model (**logistic regression**) to predict and explain body severity level (also named **MAIS index**) according to several risk factors for road traffic crashes. In this project, we are interested in two types of vehicles: Light Vehicles (**LV**) and Motorized Two-Wheelers (**2WM**).

Some **evaluation metrics** were used to know the model goodness of fit and models was validated by the use of the ROC curve as well as the classification table.

It has been observed, for example, that a user hit in the skull during an accident from a LV is 5 times more likely to have a higher level of bodily severity (MAIS3+). A 2WM user has three times the risk of being seriously injured.

Key Words* : *Logistic regression, LV, 2WM, MAIS index, Evaluation Metrics

TABLE OF CONTENTS

FIGURE TABLE

GLOSSARY

ACRONYM	DEFINITION
LR	Logistic Regression
2WM	2 Wheels Motor vehicles
LV	Light Vehicles
AIS	Abbreviation Injury Scale
MAIS	Maximum Abbreviation Injury Scale
API	Application Programming Interface
OR	Odd Ratio
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
DoF	Degree of Freedom
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve

INTRODUCTION

There are approximately 56 019 road accidents in 2019 which cause approximately 1.35 million deaths and 20 to 50 million injuries [1]. These accidents cause considerable economic losses for the victims, the family and the country as a whole (ONISR).

A bodily accident is a road accident that occurs on public roads involving at least one vehicle and having made a victim requiring treatment.

Several works have been developed by authors from different research fields such as psychology, epidemiology, economics, etc. Each studies and models road accidents according to their own method. A Pubmed query search with « regression road traffic crashes » keywords returned 1 301 results between 1985 and 2020.

The first attempts at modeling road accidents were based on aggregate simple and multiple regression models. The objective was to explain the evolution of the frequency of road accident cases as well as the deaths according to aggregated variables. Several studies have been carried out in this direction but by integrating explanatory variables such as for example the average salary per capita, alcohol consumption per capita, kilometers traveled, etc. [2]. However, the level of severity or the environment of the accident was not taken into account in the studies described above.

For this, Peltzman [3] built a simultaneous equation model making it possible to distinguish the frequency of accidents according to their severity (number of fatalities and injuries, material damage). He took into account the regulations of the American authorities to add explanatory variables such as wearing a seatbelt, speed limit, etc.

However, the frequency of an accident and its severity don't repeat themselves alike. They vary according to the socio-economic characteristics of the driver (sex, age, family situation, etc.), his driving behavior (speed, wearing of a seat belt, etc.), the characteristics of the vehicle (age, dimension, condition , etc.) and the specific circumstances of the accident (location, time, weather, road conditions, etc.) [2]. An accident is the interaction result of several specific factors to each situation and the combination of the components of the traffic system (driver, vehicle, road).

Subsequently, Savolainen et al. [4] explained that it is necessary to separate the factors influencing the frequency of accidents and the factors affecting the severity of the accident. Several studies have resulted from the idea cited above, i.e. studying the relationship of the level of severity and several explanatory variables (attributes of the driver, characteristics of the vehicle and characteristics of the accident, period of the accident, etc.) .

Recently, researchers have used statistical models to assess the characteristics of the factors on the severity of the accident. It was found that the majority of the most used models are logistic regressions.

For example, Bédard et al. [5] applied multinomial logistic regression to assess the independent contribution of driver, crash and vehicle characteristics to fatal injuries sustained by drivers. They found that factors such as increasing driver age, female gender, beltless drivers, driver-side impacts, travel speeds above 70km/h and older vehicles were associated with a higher probability of fatal consequences.

For Rui Garrido et al. [6], they used the ordinal logistic regression model to examine the contribution of several factors to the severity of injuries to occupants of motor vehicles involved in motor vehicle crashes. The estimated results suggest that occupants of light vehicles traveling on two-way roads and dry pavements tend to suffer more serious injuries than those traveling in heavy vehicles, on one-way roads and on wet roads. In addition, it came back that women are more likely to suffer serious or fatal injuries than men.

Finally, bodily gravity indexes exist. In fact, a study [7] aimed to investigate the association between speed impact and the risk of pedestrian fatalities in passenger vehicle collisions based on actual accident cases in China. For this, they used as a sampling criterion, the injury according to the Abbreviated Injure Scale (AIS) and other criteria. A multiple logistic regression model on the risk of injury to pedestrians AIS 3+ (severe injuries) was developed by considering age and impact speed as two explanatory variables. It has been found that the risk of pedestrian death is 26% at 50 km/h, 50% at 58 km/h and 82% at 70 km/h.

Likewise for the Stephen A. Et Al [8] study, analyzes to define how the age of occupants affects collision exposure and risk of AIS3 + injuries while taking into account comorbid factors and gender. Using Logistic Regression (LR) models, they found that age significantly increases the risk of AIS3 + injury in all types of collisions.

Our study aime to improve the care of road crashes victims in hospitals. This improvement will make it possible to quickly identify the bodily severity level in an accident.

The principle is therefore to implement statistical model to predict and explain body severity level according to several risk factors for road traffic crashes. Throughout the project, we will mainly use the R language (in RStudio Software), as well as the Excel software for descriptive statistics in particular.

MATERIALS AND METHODS

STUDY DESCRIPTION

In this project, we will use the Versailles/Paris database between January 2015 and January 2016. This database contains 1 646 observations and 627 quantitative and qualitative variables (available in **Annex 1**). Moreover, in our database, there are 1 237 accidents recorded in 2015 and 409 in 2016.

An observation represents a row in our database representing a person involved in an accident on a specific date and location. Indeed, for each bodily injury, an information sheet describing the accident is entered by different units. Here, there are two units : the Paris fire brigade and the Yvelines departmental fire and rescue service. Each observation is uniquely identified by a variable called IDUSA.

In addition, a map of accident frequency and body severity level rate (classified as free / moderate / severe) by location was created. To do that, we used the addresses of each accident, and from the Google Maps API, the longitudes and latitudes were extracted to allow the points to be projected on the map. The aggregated data used and the Map are presented in **Annex 1.** They are deployed from RShiny (which is an R package that allows you to build an interactive web application) and are available on the following link: : **https://elabboutisoumaya.shinyapps.io/MapTrafficAccidents/**. For example, in Figure 1, we observe that accidents located in Highway 13 (A13) is very frequent (N> 10). In addition, 78% of victims come out unharmed, 17% suffered minor injuries and 6% had serious injuries (according to the AIS classification)

The variables present in the database correspond to different groups of variables. The groups of variables are the following :

- the user : his age, state of health, transfer to hospital, bodily gravity, etc.,
- the vehicle: type of vehicle (light vehicle, two-wheeled motor vehicles, etc), condition after accident, general information, etc.,
- the user in his vehicle corresponding to the information relating to the user located in the vehicle during the accident.

The variables to be explained correspond to the body severity index linked to the user's condition (AIS, MAIS, ISS, RTS, etc.). Here, we will use the AIS. below, some explanations of these indexes:

AIS (Abbreviated Injure Scale)

To assess body severity, an injury score called the AIS (Abbreviated Injure Scale) will be used. This index is used to rank and compare injuries by level of severity (rated in order of severity from 1 to 6, see Table 1) as well as to list the injuries (concerning 9 areas of the human body: head, face, neck, thorax, abdomen, spine, upper limbs, lower limbs external surface). The MAIS index represents the Maximum AIS index and will be, in our study, the variable to be explained [9,10].

Table 1 : AIS score reference

AIS INDEX	INJURY SEVERITY LEVEL
0	ABSENCE
1	MINOR
2	MODERATE
3	SERIOUS
4	SEVERE
5	CRITICAL
6	MAXIMAL (untreatable)

Others indexes : Injury Severity Score (ISS), Revised Trauma Score (RTS), Trauma Injury Severity Score (TRISS) and the Glasgow score.

The ISS score is derived from the AIS score previously described. The ISS (Injury Severity Score) and the RST (Revised Trauma Score) are used to determine the severity of a multiple trauma patient. The ISS takes into account the 3 most serious injuries. [11,12,13,14]

The Glasgow scale is calculated by adding the values of three components : motor, verbal and ocular. Different combinations with identical values can give profiles with different forecasts. For a long time, the motor component of the score seemed sufficient to establish a patient's prognosis. Assessment of the verbal component of the scale is impossible in drunk patients, for example. A score less than 8 defines severe head trauma [15].

DATA

- ***DATA DESCRIPTION***

A pre-selection of the variables was carried out by the accident specialists. Finally, we have 331 variables out of the 627 variables initially present in the database (see **Annex 2**).

Descriptive analysis allows variables to be summarized and described using statistics. In other words, to observe the correct filling of the variables according to the modalities of the variables. Subsequently, this will allow us to analyze the rate of missing values, to clean the variables if necessary and to recode the modalities by grouping them, for example.

In **Annex 3**, we find some examples of descriptive statistical analysis of the variables about users :

- In Figure 1, the variable « USEXE » represents the gender of the user. Here, we see that in our data, we have more men than women with respective numbers of 50% and 30%. We observe a non-information rate concerning the user of 18%.
- Figure 2a, we observe that the average age of users is around 36 years. For this variable « UAGE », there are 265 observations not filled in, which represents approximately 16% of all our observations. In Figure 2b, we see that 575 (or 35%) users are between 26 and 35 years old.
- Figure 3, regarding the variable to be explained, i.e. the MAIS index, we can see that in the majority (1 080 observations or 65.61% of total observations) of cases the MAIS is equal to 0.
- Finally, in Table 1, we observe that about 48% of the observations are uninjured or injured people (with a percentage of about 52%). The rate of missing data is very low: around 0.12% (0.06 + 0.06). People who have suffered an accident are generally of « normal » body according to the Body Mass Index (BMI). Almost all users (99.39%) received an estimate of the severity of injuries almost nonexistent by the first doctor. 88.76% of people were not injured in the skull in the accident. There is almost equality between people who were transported to hospital : 51.64% were not transported against 48.36% who were. And finally, the two types of user who represent the different means of transport of the person. We observe that the 2 most represented modalities are Light Vehicles (LV) (74.1%) and 15.8% 2-Wheeled Motor vehicles (2WM).

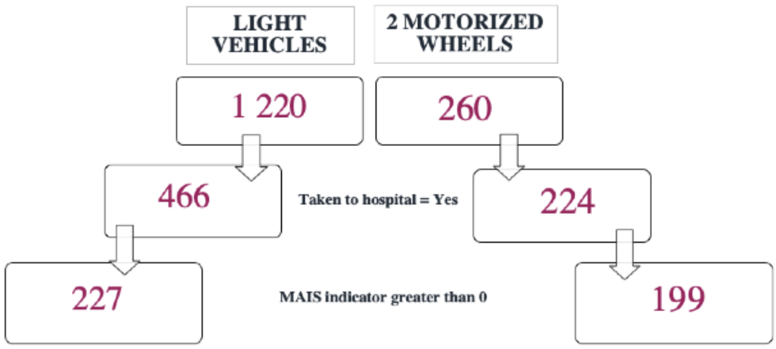
- ***DATA PRE-PROCESSING***

After carrying out the overall descriptive analysis of our data, we will base ourselves on two types of transport which are best represented, ie Light Vehicles and 2WM. For each type of transport, we will do a cleaning (missing, incorrectly filled in, recoding) of the specific variables (see **Annex 4**).

As a reminder, we have 3 « groups » of variables : those concerned with the user himself, the vehicle as well as the user within the vehicle. Regarding the explanatory variables linked to the 1 220 Light Vehicles, there are a total of 26. For the explanatory variables concerning the 260 2WM, there are a total of 21. All explanatory variables are categorical. The descriptive statistics about these two types of users (LV and 2WM) are presented in **Annex 5**.

We want to model the body severity of road accident victims, we will therefore filter the observations to keep the users who have been transported to the hospital and having an index MAIS > 0. We can summarize the process to obtain our final samples for the modeling with a diagram:

Figure 1. Construction of 2WM and LV samples for modeling



As a reminder, we had first 1 646 users. Then, we isolated the LV as well as the 2WM and we obtained 1220 LV and 260 2WM users. We use the variable « TRHOP », which signifies whether the user has been transported to the hospital, to have the « real » MAIS zero (because it is possible that users who have not been transported to the hospital, have received a score set to 0 by default and which may not be truly zero). We therefore obtain 466 LV users and 224 2WM users at this level. Finally, given that we want to model body gravity, we will therefore filter to have users with a score MAIS > 0, ie 227 LV users and 199 2WM users.

The aim of the modeling is to estimate the probability of having an overall severity MAIS greater than or equal to 3 (MAIS ≥ 3) depending on the variables. Indeed, a person with a MAIS score = 3 or > 3 is clinically classified as seriously injured [9]. For this, it is necessary to recode the variable « MAISGLOB » in two modalities : greater than or equal to 3 and strictly less than 3. The recoding is represented below by tables :

Figure 2. Recoding of the MAISGLOB variable for 2WM and LV users

MAISGLOB	Modalities description
0	**Absence**
1	**Minor**
2	**Moderate**
3	**Serious**
4	**Severe**
…	**….**

Distribution of the MAISGLOB variable for LV users (N=227) :

MAISGLOB	Modalities description	Freq.	%
1	MAIS < 3	201	88,5
2	MAIS ≥ 3	26	11,4

Distribution of the MAISGLOB variable for 2WM users (N=199) :

MAISGLOB	Modalities description	Freq.	%
1	MAIS < 3	165	82,9
2	MAIS ≥ 3	34	17

MODEL DESCRIPTION

- ***LOGISTIC MODEL (LOGIT MODEL)***

Logistic Regression (LR) is widely used in several domains. Especially in epidemiology, this method is used to identify the factors associated with a particular pathology. It makes it possible to carry out a reasoning said all things being equal: that is to say, isolate the effects of each variable and therefore identify the residual effects of an explanatory variable according to a variable to be explained in the model.

The principle of a so-called binary logistic regression is to explain a binary variable (for example: Yes/No, True/False) from explanatory variables of the quantitative or qualitative type. The results of the analysis are in the form of an Odds Ratio (OR). The mathematical expression of the logistic function is shown below :

$$logit(p) = log(\frac{p}{1-p}) = \beta_0 + \sum_{j=1}^{n} \beta_j \, X_{ij}$$

- with X_{ij} :explanatory variables,
- β_j the estimated coefficients of the predictor variables,
- β_0 corresponds to the constant/intercept

The objective of the modeling is to estimate the probability of having an overall severity MAIS (body severity index) greater than or equal to 3 (MAIS ≥ 3 or MAIS3+) according to the explanatory variables which are categorical type. Given that our variable to be explained has two modalities (<3 and ≥ 3), this implies that we must use a binary type logistic modeling.

The modeling will be done in two stages : a modeling concerning the 227 light vehicle type users and another concerning the 199 two-wheeled motor vehicles.

To do this, we used the function glm (*generalized linear models*) on R which allows to calculate several statistical models. Binary logistic regression corresponds to the logit model (part of the family of binomial models) [16].

The main link function that is privileged in this study is the binary logistic link. This function requires a relevant choice of the reference modality with respect to each explanatory variable.

This choice is important for reading the results. The choice of reference modality does not affect the quality of the prediction. Whatever the reference modality adopted for each categorical explanatory variable, we will obtain exactly the same prediction when the model is applied to an individual in the population.

For the reference modality, R chooses by default the last modality or the most represented as the reference modality. But, it is possible to modify the latter. This choice was modified in our case, we preferred to choose a reference modality that is extreme to the variable to be explained (that is to say, high bodily gravity during a road accident). Thus, the coefficients associated with the modalities linked to their explanatory variable will all have the same sign.

The choices of the reference modalities for the two types of vehicles (2WM and LV) are presented in **Annex 4** (Figures 1-2, columns « *reference* »).

- ***INTERPRETING THE ODDS RATIO (OR)***

We know that the term $\frac{p}{1-p}$ is an OR. In other words, this term refers to the probability that a user is in the body severity group (MAIS3+) of severe type over the probability that he does not have it (MAIS < 3). It is from this term that we can measure the relationship between our explanatory variables and the variable to be explained. So our coefficients of the explanatory variables resulting from the logistic modeling correspond to the log OR :

$$\log(\frac{p}{1-p}) = e^{\beta j} \qquad [17]$$

When an independent variable X_i increases by one unit (X_i+1), with all other factors remaining constant, the odds of the dependent variable increase by a factor $e^{\beta j}$ (or OR) and ranges from zero (0) to positive infinity.

We interpret an OR of a qualitative variable only from the reference modality. If the value is significantly greater than 1 (risk effect), then the user is more likely to be part of the MAIS3+ group (severe bodily severity) and vice versa if the value is significantly less than 1 (protective effect). On the other hand, an OR which is not significantly different from 1, means no effect. In addition, the OR of the significant modalities go in the same direction as the signs of the estimated coefficients.

- ***EVALUATION OF THE MODEL***

The goodness-of-fit for the LR model can be assessed in several ways. First, is to assess the overall model (relationship between all of the independent variables and dependent variable). Second, the significance of each of the independent variables needs to be assessed.

- **Goodness of Fit Metrics**

  - ***AIC and BIC***

The AIC (Akaike Information Criterion) criterion is a maximum likelihood method. The AIC criterion is defined by :

$$\text{AIC} = \text{Deviance} + 2*\text{k} \qquad [18]$$

With Deviance = -2 log of the maximum likelihood penalized by 2 times the number of parameters k.

Several models were tested and thanks to the step function on R, we were able to build our final model. This function makes it possible to gradually remove the less significant variables and also makes it possible to obtain a better AIC criterion. The lower the AIC, the better the model. The model will be more parsimonious.

The Bayesian information criterion BIC is defined by :

$$\text{BIC} = \text{Deviance} + \text{k}*\log(\text{n})$$

It is more parsimonious than the AIC criterion because it penalizes more the number of variables (k) present in the model. Moreover, Ripley [19] emphasizes that AIC retains the most relevant variables and that the BIC criterion selects the statistically most significant variables.

  - ***The likelihood ratio test***

The overall fit of a model shows how strong a relationship between all of the independent variables, taken together, and dependent variable is. It can be assessed by comparing the fit of the two models with and without the independent variables. A LR model with the k independent variables is said to provide a better fit to the data if it demonstrates an improvement over the model with no independent variables (the null model). The overall fit of the model with k coefficients can be examined through a likelihood ratio test, which tests the null hypothesis.

To do this, the deviance with just the intercept (−2 log likelihood of the null model) is compared with the deviance when the k independent variables have been added (−2 log likelihood of the given model). The difference between the two yields a goodness of fit index G, $\chi 2$ statistic with k degrees of freedom (DoF). This is a measure of how well all of the independent variables affect the outcome or dependent variable.

$$G = \chi^2 = -2 \ln(\frac{L0}{L1}) \qquad [20]$$

With L0 = likelihood of the null model and L1= likelihood of the given model

Where, the ratio of the maximum likelihood is calculated before taking the natural logarithm (ln) and multiplying by −2. The term “likelihood ratio test” is used to describe this test. If the *p*-value for the overall model fit statistic is less than the significance level of the test, conventionally 0.05 ($P < 0.05$), then H0 is rejected, with the conclusion that there is evidence that at least one of the independent variables contributes to the prediction of the outcome.

~ *Hosmer-Lemeshow test*

The Hosmer-Lemeshow test is used to examine if the observed proportions of events are similar to the predicted probabilities of occurrence in sub-groups of the model population. The Hosmer-Lemeshow test is performed by dividing the predicted probabilities into deciles (10 groups based on percentile ranks) and then computing a Pearson's Chi-square ($\chi 2$) that compares the predicted to the observed frequencies in a table. The value of the test statistics is expressed as :

$$H = \sum^{10}{}_{g=1} \frac{Og - Eg}{Eg}$$

Where, O_g and E_g denote the observed and expected events for the gth decile group. The test statistic asymptotically follows a χ^2 distribution with 8 (number of groups minus 2) DoF.

Small values (with large *P*-value closer to 1) indicate a good fit to the data, therefore, good overall model fit. Large values (with $P < 0.05$) indicate a poor fit to the data [21].

- **Statistical Test for Individual Regression Coefficients : Wald Test**

Statistical tests of significance can be applied to each variable's coefficients. For each coefficient, the null hypothesis that the coefficient is zero is tested against the inverse where the coefficient is not null using a Wald test, W_j. A Wald test can also be used to compare a full model containing all the predictor variables with a reduced model with coefficients set to zero.

So the Wald statistic can be used to assess the contribution of individual predictors or the significance of individual coefficients in a given model [22].

The Wald statistic is the ratio of the square of the regression coefficient to the square of the Standard Error (SE) of the coefficient. The Wald statistic is distributed as a $\chi 2$ distribution:

$$W_j = \frac{\beta^2{}_j}{SE^2\beta_j} \quad [23]$$

Each Wald statistic is compared with a $\chi 2$ critical value with 1 DoF.

- ***VALIDATION OF PREDICTED VALUES***

The predictive accuracy or discriminating ability of the model needs to be evaluated.

~ *Confusion Matrix (or Classification Table)*

The classification table or confusion matrix (Table 2) is a method to evaluate the predictive accuracy of the LR model [24]. In this table, the observed values for the dependent outcome and the predicted values (defined cutoff value) are cross-classified. For example, if a cutoff value is 0.5, all predicted values above 0.5 can be classified as predicting an event, and all below 0.5 as not predicting the event. Then a table of data can be constructed with dichotomous observed outcomes and dichotomous predicted outcomes. The table has the following form :

Table 2. Sample classification table

	Observed	
Predicted	True	False
Positive	**TP**	**FP**
Negative	**FN**	**TN**

Where TP = True Positive, TN = True Negative , FN = False Negative, FP = False Positive.

From this matrix, we can calculate evaluation metrics of the model dependent on each other :

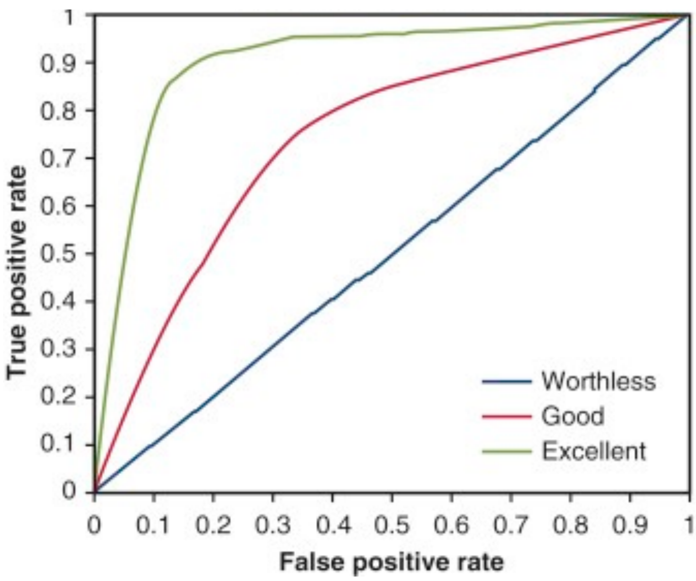
- **Accuracy** (which is the good classification rate) = $\frac{TP+TN}{Total}$
- **Sensitivity** or Recall (true positive rate, proportion of correctly classified events) = $\frac{TP}{TP+FN}$
- **Specificity** (true negative rate, proportion of correctly classified nonevents) = $\frac{TN}{TN+FP}$
- **PPV** (Positive Predictive Value) or Precision = $\frac{TP}{TP+FP}$
- **NPV** (Negative Predictive Value) = $\frac{TN}{TN+FN}$

~ *Receiver Operating Characteristic Curve (ROC)*

The ROC (R*eceiver Operating Characteristic*) curve is a way to visualize the performance of a binary classifier. This curve (Figure 3) represents the sensitivity or true positive rate according to 1-specificity or false positive rate. If we vary the threshold from which we judge that an event should be considered positive, the sensitivity and specificity vary. The blue curve (first bisector) corresponds to what we would obtain with a random model. A model close to the blue curve is therefore inefficient and is no better than a simple random draw. A model below this curve would be catastrophic because it would be worse than chance.

The area under the curve is an index calculated for ROC curves [25]. AUC is the conversion for a positive event to have a higher response (given by the model) than a negative event. For an ideal model, on an AUC = 1, for a random model, on an AUC < 0.5.

Figure 3. Example of ROC Curve



RESULTS

INTERPRETING MODEL RESULTS FOR LIGHT VEHICLES (LV)

A step-by-step top-down selection technique based on a quality criterion is used to improve the model. Thanks to the step function on R, we keep only the variables having a significant effect on the variable to be explained (to increase efficiency of estimation and refer to such a model as the reduced model). In the reduced model, variables with *P*-value less than or equal to α-value are treated as statistically significant.

The most widely used quality criterion is the Akaike Information Criterion (also called AIC). The lower the AIC, the better the model. Here, we go from an AIC = 138.04 to an AIC = 111.8 after applying this technique. Finally, we have chosen this following model :

$$ln\left(\frac{p(MAISGLOB \geq 3)}{1 - p(MAISGLOB \geq 3)}\right) = 1.50 + 1{,}62*CRANE\_Oui - 0.12*CLASSE\_AGE[19\text{-}25] +$$
2.11*CLASSE_AGE[26-35] + 2.27*CLASSE_AGE[36-45] - 1.75*CLASSE_AGE[46-59] + 2.77*CLASSE_AGE[>60] - 2.49*UVINTRU_Non + 3.29*UVLOPDC_Autre + 1.41*UVLOPDC_Distant + 0.88*UVAIRBL_Autre + 2.06*UVAIRBL_Non + 0.41*UVAIRBL_Oui_ndepl + 2.03*VLCHOCLG_Oui - 4.23*VLOUTIL_Non - 3.24*VLENCAS_Non + 2.22*DAY_Night

This model takes into account 9 factors (explanatory variables). Indeed, the factors linked to bodily gravity during a light vehicle accident are : a skull impact (*CRANE*), the driver's age (*CLASSE_AGE*), if there has been an intrusion into the vehicle (*UVINTRU*), if the impact was localized far away or otherwise (*UVLOPDC*), if the vehicle suffered a left side impact (*VLCHOCLG*), if tools were used (*VLOUTIL*), if the accident took place during the day or at night (*DAYS*), if it there was no embedding (*VLENCAS*) and if the side airbag has been deployed (*UVAIRBL*).

The following types of information are presenting the logistic regression results. Table 3 presents the logistic regression model with statistical significance of individual regression coefficients (estimate) tested using the Wald $\chi 2$ statistic. In **Annex 6** (Figures 1-2), we have another representation of the OR results of the variables.

Table 4 illustrates the overall evaluation of the logistic model and the goodness-of-fit statistics. Finally, Table 5 and Figure 4 represents an assessment of the predicted probabilities (classification table and ROC curve).

If we refer to **Table 3**, on the row corresponding to the variable CRANE, we can see that the log OR = 1.62, that is to say an OR = $e^{1.62}$ = 5.035. This means that if the user was hit in the skull during the accident, the risk of being in the MAIS3+ group increases 5 times more than if he had not been hit in the skull (moreover, the sign of the estimated coefficient is positive, and therefore the OR > 1).

Moreover, the associated *p*-value is < 0.05. It can be concluded that there is a significant relationship between belonging to the MAIS3+ group and being affected in the skull with a risk threshold of 1%. Being hit in the skull during the accident is therefore a risk factor for the risk of belonging to the MAIS3+ group.

For the UVINTRU_No modality (i.e. if there was no intrusion into the vehicle), we have an OR significantly (*p*-value <0.05) less than 1 (OR = 0.08289). We can therefore conclude that the fact that there is no intrusion has 91.7% more chance that the user is not in the MAIS3 + group than if there was an intrusion. In addition, the sign of the estimated coefficient is negative so we have a protective effect.

The UVLOPDC, VLCHOCLG and DAY were estimated to be a significant predictor for the event. Indeed, compared to references of specific variable, if the impact is in other to the occupant location (OR = 26.992, $P < 0.01$), if there is a left side impact (OR = 7.6297, $P < 0.05$) and if the accident took place on the night (OR = 9.2191, $P < 0.01$) are characterized by significantly higher probability to belong to the MAIS3+ group.

On the other hand, if there has been no intrusion (OR = 0.08289, $P < 0.05$), unused of tools (OR = 0.01459, $P < 0.001$) or vehicle recessed (OR = 0.03930, $P < 0.05$), it shows that the probability to have a low body severity index is higher.

Table 3. Result of the logistic model, N=227
(estimated coefficients, standard error, p-value, odds ratio and their confidence intervals)

						Confidence Intervals	
Explanatory Variable	**Estimate**	**Standard Error**	**Wald Test W_j**	***P*-value**	**Odds Ratio**	**2.5 %**	**97.5 %**
Intercept	1.50	2.54	0.59	0.55	4.48	0.024	672.46
SKULL							
Yes	1.62	0.67	2.40	*0.01630 **	5.035	1.3721	20.0125
CLASSE_AGE							
[19-25]	-0.12	1.79	-0.069	0.94505	0.8833	0.03599	54.1086
[26-35]	2.11	1.63	1.296	0.19510	8.2514	0.59592	411.1729
[36-45]	2.27	1.67	1.36	0.17382	9.7148	0.55141	471.8459
[46-60]	-1.75	2.81	-0.62	0.53237	0.17288	0.000687	28.3902
>60	2.78	1.69	1.64	0.10126	16.094	0.94067	877.9991
UVINTRU							
No	-2.49	1	-2.48	*0.01314 **	0.08289	0.00965	0.5365
UVLOPDC							
Other	3.29	1.23	2.69	*0.00716 ***	26.992	2.5392	340.3005
Distant	1.41	0.87	1.61	0.10669	4.1050	0.76409	25.4513
UVAIRBL							
Other	0.88	1.63	0.54	0.58825	2.4135	0.10358	79.9981
No	2.06	1.33	1.543	0.12290	7.8445	0.80986	182.758
Yes_ndepl	0.41	1.45	0.280	0.77959	1.502	0.10151	39.5128
VLCHOCLG							
Yes	2.03	0.89	2.27	*0.02293 **	7.6297	1.3941	49.5942
VLOUTIL							
No	-4.23	1.01	-4.17	*3.06e-05 ****	0.01459	0.00163	0.0937
VLENCAS							
No	-3.24	1.50	-2.15	*0.03127 **	0.03930	0.00184	0.7775
DAY							
Night	2.22	0.79	2.80	*0.00507 ***	9.2191	2.1203	50.1346

*Significance threshold of p-values: "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1.*
The lower is the p-value, more value is significant.

From **Table 4,** two inferential statistical tests for overall model evaluation : the likelihood ratio and Wald tests, are shown. All two tests yield similar conclusions for the given data set. It could be noticed from the results of the likelihood ratio test and the Wald test presented in Table 4 that the logistic model with independent variables was more effective than the null model ($P < 0.01$).

Table 4 also presents the Hosmer-Lemeshow goodness-of-fit test. This statistical test measures the correspondence of the actual and predicted (expected) values of the dependent variable (MAIS3+). A better model fit is characterized by insignificant differences between the actual and expected values. It tests the hypothesis H_0, there is no difference between the predicted and actual values against H_1, there is difference between the predicted and actual values. At *p*-value of 0.9249, the null hypothesis is accepted and we conclude that insignificant differences remain between the actual and expected values, suggesting that the model fitted the data well.

Table 4. Model Evaluation Metrics

Test	χ^2	DoF	*P*-value
Likelihood Ratio Test	83.778	16	3.445e-11 ***
Wald Test	34.8	16	0.0043
Hosmer and Lemeshow Test	3.1449	8	0.9249
	Likelihood	**AIC**	**BIC**
Model Summary	77.8	111.8	170.024

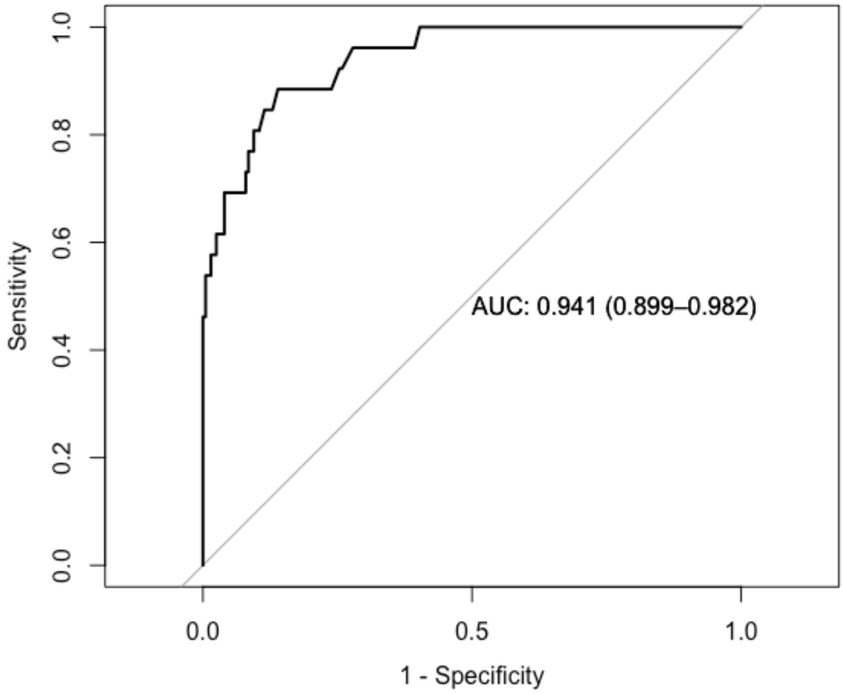
Table 5 presents the degree to which predicted probabilities agree with actual outcomes in a classification table. The accuracy (overall correct prediction), 93.8% shows an improvement over the chance level which is 50%. In other words, if a user will have the characteristics of the model, they will be classified in the group of users with an overall severity MAIS greater than or equal to 3 with a good rank rate of 93.8%.

Figure 6 shows that AUC is 0.941 (95% IC = 0.899-0.982), so it can be considered to be a very good discriminant.

Table 5. Result of Classification Table

		OBSERVED		
		MAIS3+		**% CORRECT**
PREDICTED		**True**	**False**	
MAIS3+	**Positive**	15	3	83.3
	Negative	11	198	94.7
% CORRECT		57.7	98.5	
			ACCURACY (%)	93.8

Figure 4. ROC Curve



INTERPRETING MODEL RESULTS FOR TWO MOTORIZED WHEELS

Like the light vehicles, after using the step function, we go from an AIC = 196.71 to an AIC = 172.93. The final model for 2WM is as follows:

$$ln\left(\frac{\text{p(MAISGLOB} \geq 3)}{1 - \text{p(MAISGLOB} \geq 3)}\right) = 0.60 + 1.12*\text{CRANE\_Oui} - 2.59*\text{CLASSE\_AGE[19-25]} - 1.30*\text{CLASSE\_AGE2[26-35]} - 1.59*\text{CLASSE\_AGE[36-45]} - 0.58*\text{CLASSE\_AGE[>45]} + 2.72*\text{UDTROT\_Oui} - 1.33*\text{UDBOTTES\_Non} + 0.72*\text{DRTYPOBS\_Autre} - 1.22*\text{DRDEFORM\_Autre} - 1.01*\text{DRDEFORM\_Avant} + 0.72*\text{DRDEFORM\_Partout}$$

This model takes into account 6 factors (explanatory variables). Indeed, the factors related to bodily gravity during a two-wheel motorized accident are : a skull impact (*CRANE*), the driver's age (*CLASSE_AGE*), if there was an impact with the sidewalk (*UDTROT*), if the user was wearing boots (*UDBOTTES*), if the type of obstacle is other than a light vehicle (*DRTYPOBS*) and finally if the deformation of the vehicle is located ahead, everywhere or elsewhere (*DRDEFORM*).

In **Table 6**, concerning the variable UDTROT, we observe an OR = 15.115. This means that if the user has touched the sidewalk is 15 times more likely to be in the MAIS3+ group than if they had not touched ($P < 0.01$). It can be concluded that there is a significant relationship between belonging to the MAIS3+ group and touching the sidewalk with a threshold of 1%.

Compared with users aged under 18 years old and, all other variables held constant, the relative probability of not belonging MAIS3+ group decreases by 92,5% for users aged between 19 and 25 years old ($P < 0.01$).

Table 6. Result of the logistic model, N=199
(estimated coefficients, standard error, p-value, odds ratio and their confidence interval)

Explanatory Variable	Estimate	Standard Error	Wald Test W_j	*P*-value	Odds Ratio	Confidence Intervals 2.5 %	97.5 %
Intercept	0.6044	0.9625	0.628	0.52999	1.8302	0.2679	12.115
SKULL							
Yes	1.1232	0.4842	2.32	*0.02035* *	3.0746	1.187	8.0527
CLASSE_AGE							
[19-25]	-2.5877	0.9358	-2.765	*0.00569* **	0.0751	0.0106	0.4427
[26-35]	-1.3072	0.7487	-1.746	*0.08083* .	0.2705	0.0614	1.2099
[36-45]	-1.5957	0.8013	-1.991	*0.04645* *	0.2027	0.0402	0.9763
>45	-0.5803	0.7046	-0.824	0.41015	0.5597	0.1419	2.3491
UDTROT							
Yes	2.7157	0.9744	2.787	*0.00532* **	15.115	2.2937	115.011
UDBOTTES							
No	-1.3242	0.5419	-2.443	0.01455 *	0.2660	0.0906	0.7766
DRTYPOBS							
Other	0.7172	0.4737	1.514	0.13006	2.0486	0.8030	5.2277
DRDEFORM							
Other	-1.2246	0.6784	-1.805	*0.07105* .	0.2938	0.0762	1.1343
Front of	-1.0051	0.6344	-1.584	0.11308	0.3659	0.1064	1.3239
Everywhere	0.7172	0.7341	0.977	0.32959	2.048	0.4950	9.0976

*Significance threshold of p-values: "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1.*
Lower is the p-value, more value is significant.

The likelihood ratio test statistic is 33.051 (distributed Chi-squared), see in **Table 7**, with 2 DoF. The associated *p*-value, which is $P < 0.001$, indicating that the model with all predictors fits significantly better than the model without predictors.

We see the Chi-squared value generated by the Wald test, as well as the *p*-value associated with a chi-squared of 24.3 with 11 DoF. The *p*-value is less than the generally used criterion of 0.05, so we are able to reject the null hypothesis, indicating that the coefficients are not simultaneously equal to zero. Because including statistically significant predictors should lead to better prediction (i.e., better model fit),we can conclude that including all predictors results in a statistically significant improvement in the fit of the model.

From the Hosmer & Lemeshow test, the *p*-value of 0.09949. So, the null hypothesis is accepted and we suggest that the model fitted the data well.

Table 7. Model Evaluation Metrics

Test	χ^2	DoF	*P*-value
Likelihood Ratio Test	33.051	11	0.0005162 ***
Wald Test	24.3	11	0.011
Hosmer and Lemeshow Test	13.378	8	0.09949
	Likelihood	**AIC**	**BIC**
Model Summary	148.93	172.93	212.449

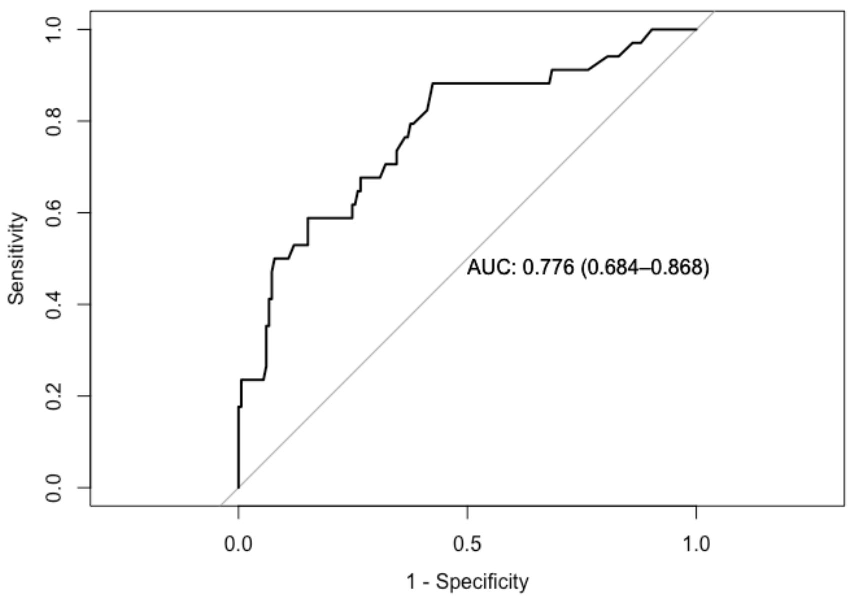
The good classification rate is therefore 86.4% (**Table 8**). The good classification rate therefore represents the correct predictions. In other words, if a user will have the characteristics of the model, he will be classified in the group of users with an overall severity MAIS greater than or equal to 3 with a good classification rate of 86.4%

Regarding the ROC curve shown in **Figure 5**, AUC amounts to 0.776 (95% CI = 0.684-0.868), so we can consider it to be a good discriminant.

Table 8. Result of Classification Table

		OBSERVED		
		MAIS3+		**% CORRECT**
PREDICTED		**True**	**False**	
MAIS3+	**Positive**	8	1	88.9
	Negative	26	164	86.3
% CORRECT		23.5	99.4	
			ACCURACY (%)	86.4

Figure 5. ROC Curve



DISCUSSION AND CONCLUSION

An accident is an involuntary and random phenomenon from which no one who uses the road is safe. The probability of having an accident exists as soon as a person uses the road to get from one place to another for one reason or another. It can be explained by several factors such as the characteristics of the vehicle, the socio-economic categories of the drivers, etc.

We cannot avoid this phenomenon, but we can reduce its severity through preventive security measures.

The objective of this report is to implement a logistic model in order to predict and explain the level of bodily severity according to the different characteristics of accidents. Several factors were used such as the characteristics of the user, the vehicle and the user in the vehicle.

However, this model poses a problem of direct interpretation of the estimated parameters. To overcome this limit, we will have to set a reference situation which must be compared with the observed situations and measure the probability ratio of bodily severity of an accident according to these two situations.

We applied a logistics model on two samples composed of 199 motorized two-wheeled vehicles (2WM) and 227 light vehicles (LV) collected by two units: the Paris fire brigade and the Yvelines departmental fire and rescue service in France.

These models clearly show a level of explanation of the factors (explanatory variables) of the bodily severity of a road accident. These results may be of great practical interest to firefighters affected by this phenomenon. In fact, improved care will allow faster identification of the level of severity of the accident as well as more efficient allocation of injured persons to hospitals.

From these results, we were able to observe that the most representative risk factor (whether for Light Vehicles or 2WM) is the skull of the user affected during the accident. This is because the fact that the skull is affected increases the likelihood of having a high level of body severity.

Concerning road accidents involving light vehicles, the risk factors for belonging to the MAIS3+ class are as follows: a left side impact, another location (other than remote / adjacent) of the impact depending on the occupant of the vehicle as well as the accident taking place overnight. And for 2WM, the main factor is the fact that the user of the vehicle touches the sidewalk.

Finally, note that this work has some limitations which may reduce the significance and relevance of our results, but which also offer opportunities for future research. In fact, our study remains static and timeless. It took into account neither the evolution of the number of victims or the frequency of occurrence of accidents. In addition, the small number of observations processed can limit the quality of measurement and fit of ours models.

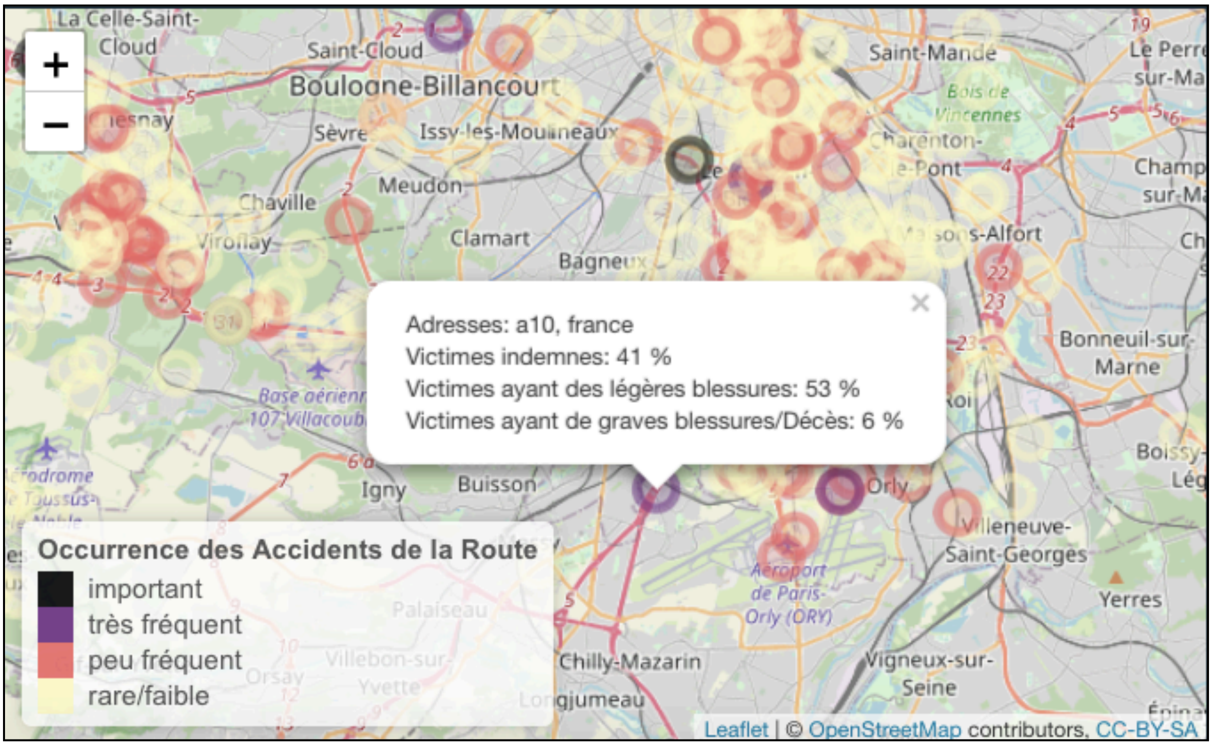
ANNEX

ANNEX 1 : CONSTRUCTION OF THE HOTSPOT MAP

Table 1. ***Data used for the construction of the map (N = 543, aggregated data)***

geoAddress	Freq	faible/modérée	grave/mortelle	indemne	lon	lat	AIS_index
1 avenue sadi lecointe, 78140 vélizy-villacoublay, fra...	3	0	0	3	2.184830	48.78289	peu fréquent
1 bis rue françois arago, 78190 trappes, france	1	0	0	1	1.973784	48.76085	rare/faible
1 boulevard vauban, 78180 montigny-le-bretonneux,...	2	1	0	1	2.045459	48.78449	rare/faible
1 place de la porte de versailles, 75015 paris, france	1	0	0	1	2.287037	48.83006	rare/faible
1 place saint-michel, 75005 paris, france	2	1	0	1	2.344660	48.85353	rare/faible
1 quai de la corse, 75004 paris, france	2	0	0	2	2.346756	48.85602	rare/faible
1 rue camille groult, 94400 vitry-sur-seine, france	2	1	0	1	2.402490	48.78982	rare/faible
1 rue de bicètre, 94240 l'haÿ-les-roses, france	2	1	0	1	2.349665	48.77139	rare/faible
1 rue du marché, 94150 rungis, france	3	0	0	3	2.350313	48.74698	peu fréquent
1 rue du maréchal devaux, 94390 paray-vieille-poste...	3	1	0	2	2.369075	48.74538	peu fréquent
1 rue royale, 78000 versailles, france	2	0	0	2	2.127977	48.79903	rare/faible
10 avenue du général de gaulle, 78000 versailles, fra...	1	1	0	0	2.128735	48.80105	rare/faible
10 rue charlot, 75003 paris, france	1	0	0	1	2.361202	48.86137	rare/faible
10 rue marguerite chapon, 94800 villejuif, france	2	1	0	1	2.355264	48.80227	rare/faible
10 rue nationale, 75013 paris, france	4	1	0	3	2.368967	48.82270	peu fréquent

Figure 1. ***Hotspot map : Occurrence of road traffic crashes in Île-de-France (2015-2016)***



ANNEX 2 : TABLE SHOWING THE LIST OF ALL THE VARIABLES OF THE INITIAL DATABASE (pre-selected by the accident specialists)

Highlighted in Yellow represents the variables selected to build our models

Nom	Description	Nom	Description
IDUSA	Identifiant usager	HEURSMUR	Heure SMUR sur place (hhmm)
INTERV	N° Intervention	HEURSMUR_ET	Etat heure SMUR
UNITE	Unité	HEURDEP	Heure départ des lieux (hhmm)
DAACC	Date Accident (aaaammjj)	HEURARR	Heure d'arrivée à l'hôpital (hhmm)
HEURD	Heure intervention (hhmm)	HEURARR_ET	Etat heure d'arrivée
LIACC	Lieu de l'accident	MEDSMUR	Médecin du SMUR
UNOM	Nom	MEDSMUR_ET	Etat médecin
UPREN	Prénom	FICHEQ	Fiche remplie par
NAISS	Date de naissance (aaaamm)	VIENCEINTE	Grossesse > 20 SA
UAGE	Age (année)	VIANTICOA	Anticoagulants
USEXE	Sexe	VIANTIAGR	Antiagrégants
APPAR	Apparence	REINTASY	TA systolique initiale (mmHg)
ENNBM	Si femme enceinte nb mois	REINITFC	FC initial (/ mn)
TYPUS	Type d'usager	REHEMOC	Hémocue (g/dl)
IMMAT	Immatriculation	RESPO2	SpO2 (%)
NUSAG	N° usager	RESOUSO2	Sous O2 (l/mn)
UGRAV	Gravité présumée	REFIO2	FI O2 (%)
EQMED	Présence équipe médicale sur les lieux	RECTRIST	Cristalloïdes (ml)
TRHOP	Transport à l'hôpital	RECOLLO	Colloïdes (ml)
MOYTR	Moyen de transport utilisé	RECATHEC	Cathécolamines
HOPIT	Hôpital 1iere accueil	REINITACR	ACR initial
UENFA	L'usager est un enfant (<= 12 ans)	REVENTMEC	Ventillation mécanique
TYDRE	Type de retenue enfant utilisé	CRANE	Crâne
EATTA	Enfant était-il attaché ?	CRGLASGO	Glasgow (0 à 5)
DREAT	LE DRE était-il attaché au véhicule	CRANISO	Anisocorie
VITEST	Estimation de la vitesse par les pompiers	CRMYDRI	Mydriase
UPOIDS	Poids (kg) ou NR si non renseigné	CRTC	Traumatisme craniene
UTAILLE	Taille (cm) ou NR si non renseigné	CRPCI	Perte de conscience initiale
NBVICT	Nombre total de victimes	CRLOCAL	Localisation
CENTREG	Centre de régulation médicale	CRPLAIE	Plaie au scalp
TRAMED	Transport médicalisé	CREMBAR	Embarrure / Ottoragie
TRANOMED	Transport non médicalisé	CRAGITE	Agité
SURPLACE	Laissé sur place	CRSSH	Serum salé
REFUSTR	Refus de transport	CRMANNIT	Mannitol
DCD	Décédé	FACIAL	Maxillo-facial
DESTHOP	Destination Hôpital	FANORMAL	Maxillo-facial Normal
DESTCLI	Destination Clinique	RACHIS	Rachis
STRACC	Structure d'accueil hospitalière	RNEURO	Neuro
ESTGRAV	Estimation de la gravité jugée par le 1er médecin (0 à 10)	RNiVEAU	Neuro Niveau
HEURAPP	Heure d'appel (hhmm)	RNCOMP	Neuro Complet
HEURVSAV	Heure VSAV sur place (hhmm)	THORAX	Thorax

Nom	Description	Nom	Description
THASYM	Thorax : Asymétrie auscultatoire	PLAQPRE	Plaquettes - Préadmission
THDRAINE	Thorax : Drainé	PLAQHOP	Plaquettes - Hopital
THEXSUF	Thorax : Exsufflé	TPPRE	TP - Préadmission
THVOLET	Thorax : Volet	TPHOP	TP - Hopital
THDETRES	Thorax : Détresse	TCAPRE	TCA - Préadmission
ABDOM	Abdomen	TCAHOP	TCA - Hopital
ANORMAL	abdomen Normal	FIBPRE	Fibrinogène - Préadmission
BASSIN	Bassin	FIBHOP	Fibrinogène - Hopital
BNORMAL	Bassin Normal	ACIDPRE	Acide tranéxamique
MEMBRES	Membres	ACIDHOP	Acide tranéxamique - Hopital
MEFRACTURE	type de blessure : Fracture	CATPRE	Catécholamines - Préadmission
MEFOUVERTE	Type de blessure : Ouverte	CATHOP	Catécholamines - Hopital
MEPLAIE	Type de Blessure : Plaie	VENPRE	Ventilation mécanique - préadmission
MEAMPUT	Type de blessure : Amputation	VENHOP	Ventilation mécanique - Hopital
MEISCHEM	Type de blessure : Ischémie	CGRPRE	Transfusion CGR - Préadmission
MELUXA	Type de blessure : Luxation	CGRPRETX	Taux CGR préadmission
MEBRAS	localisation blessure : Bras	CGRHOP	Transfusion CGR - Hopital
MEAVBRAS	Localisation blessure : Avant-bras	CGRHOPTX	Taux - CGR Hopital
MEMAIN	Localisation blessure : Main	PFCPRE	Transfusion PFC - Préadmission
MECUISSE	Localisation blessure : Cuisse	PFCPRETX	Taux PFC préadmission
MEJAMBE	Localisation blessure : Jambe	PFCHOP	Transfusion PFC - Hopital
MEPIED	Localisation blessure : Pied	PFCHOPTX	Taux PFC Hopital
HEMOR	Hémorragie	CUPPRE	Transfusion CUP - Préadmission
HEMLOCA	Localisation de l'hémorragie	CUPPRETX	Taux CUP préadmission
ADMISS	Admission	CUPHOP	Transfusion CUP - Hopital
FCPRE	FC - Préadmission	CUPHOPTX	Taux CUP Hopital
FCHOP	FC - Hopital	ECHOPRE	Echographie pré-hospitalière
TASYPRE	TA systolique - Preadmission	ECHOINT	Echographie intra-hospitalière
TASYHOP	TA systolique - Hopital	RADIONOR	Radiographie du thorax : normale
SPOPRE	SpO2 - Préadmission	HEMOP	Hémopéritoine
SPOHOP	SpO2 - Hôpital	HEMOT	Hémothorax
FRPRE	FR - Préadmission	PNEUM	Pneumothorax
FRHOP	FR - Hôpital	EPANCH	Epanchement péricardique
GLASPRE	Glasgow - Préadmission	DATEEV	devenir : Date évènement (aaaammjj)
GLASHOP	Glasgow - Hôpital	EVDECES	devenir : décès
HEMOPRE	Hémocue Préadmission	EVDECJO	devenir : Décès survenu à J
HEMOHOP	Hémocue - Hôpital	EVVIVANT	devenir : Vivant à 30j
ARCAPRE	Arrêt cardiaque - Préadmission	EVDURHOP	devenir: vivant - Durée hospitalisation
ARCAHOP	Arrêt cardiaque - Hôpital	REANIM	devenir : Réanimation
LACTPRE	Lactates - Préadmission	HEURDECH	devenir : réanimation – heure de sortie du décochage (hhmm)
LACTHOP	Lactates - Hôpital	READUR	devenir : Durée réanimation (j)

Nom	Description	Nom	Description
VENDUR	devenir : Durée ventilation (j)	PIPOSPLA	Position piéton/Véhicule (latérale)
EERDUR	devenir : Durée EER (j)	PITOBST	Type d'obstacle principal rencontré
CATDUR	devenir : Durée catécholamines (j)	PIAUTOB	Autre obstacle
DEVSALLE	devenir : salle	PIFRANCH	Le véhicule a t-il franchi le piéton ?
DEVDOMIC	devenir : domicile	PIAOBST	Autres obstacles fixes impactés
DEVTRANS	devenir : transfert	PILEQUEL	Si oui, lesquels
MAISTEC	MAIS Tête + Cou (1 à 6)	Ai, Bi ou Pi	Déformations
MAISFAC	MAIS Face (1 - 6)	UDIDUSA	Identifiant usager
MAISMEB	MAIS Membres + Bassin (1 - 6)	UDNUSAG	N° usager
MAISTHO	MAIS Thorax (1 - 6)	UDIDV2R	Identifiant véhicule
MAISABD	MAIS Abdomen + Pelvis (1 - 6)	UDIMMAT	Immatriculation
MAISPEAU	MAIS Peau + tissus sous cutanés (1 - 6)	UDPLACE	Place occupée
MAISGLOB	MAIS Global (1 - 6)	UDCASQUE	Port d'un casque ?
ISS	Score ISS	UDCINTEG	Casque intégral ?
SCGLASGOW	Score Glasgow	UDGANTS	Port de gants ?
PASSCORE	Score Pres. Art. Systolique	UDBLOUS	Port d'un blouson moto ?
FRSCORE	Score Fréquence respiratoire	UDPANTAL	Port pantalon moto ?
RTS	Score RTS	UDBOTTES	Port de bottes moto ?
TRISSF	Score TRISS lésion fermée	UDEQAUT	Autre(s) équipement(s) moto
TRISSP	Score TRISS lésion pénétrante	UDSOL	Sol
IDUSA	Identifiant usager	UDTROT	trottoir
IDLESION	Identifiant lésion	UDVEHIC	Véhicule
UELENU	Numéro de la lésion	UDMUR	Mur
UELDES	Description de la lésion	UDPOTEAU	Poteau / Arbre
UELRTC	Territoire corporel	UDGLISS	Glissière
UELTSA	Structure Anatomique	UDANIMAL	Animal
UELSAS	Structure anatomique spécifique	UDAUTROB	Autre obstacle ?
UELNAL	Type d'atteinte lésionnelle	UDAUTREOB	Autre(s) obstacles ?
UELAIS	AIS	UDDISPRJ	Distance de projection depuis le point d'impact (m)
UELCAIS	Code AIS de la blessure	DRIDV2R	Identifiant 2roues
IDAIS	Identifiant CODEAIS	DRIMMAT	Immatriculation
TERISS	Territoire ISS	DRINTERV	N° intervention
PIIDUSA	Identifiant usager	DRINCEN	Véhicule incendié
NUMUSA	Numéro Usager	DRTYP2R	Type de 2 roues
PINUSAG	Identifiant piéton	DRTYCHOC	Type de Choc
PIIPARE	Pare-brise impacté	DRCHUTE	Chute
PIICAPO	Capot impacté	DRIMMER	Immersion
PIITOIT	Toit impacté	DRRAYURE	Rayures
PIICOFF	Coffre impacté	DRDEFORM	Déformations
PIDISPR	Distance du projection du piéton (m)	DRTYPOBS	Type d'obstacle principal
PIPOSPLO	Position piéton/Véhicule (longitudinale)	DRAUTOBS	Autre(s) obstacle(s)

Nom	Description	Nom	Description
DRPOSFI	Position finale du véhicule	QCHOCAR	Choc Arrière
DRDISTPDC	Distance de projection du véhicule (m)	QCHOCTO	Choc Toit
UPIDUSA	Identifiant usager	QCHOCTN	Tonneau ou renversement
UPNUSAG	N° Usager	QCHOCCH	Chute
UPIDVPL	Identifiant PL	QCHOCIM	Immersion
UPIMMAT	Immatriculation	QTYPOBS	Type d'obstacle principal
UPPLACE	Place occupée	QAUTOBS	Autre obstacle
UPCEINT	Occupant ceinturé ?	QCEINT	Personne ceinturée ?
UPAIRBAF	Airbag frontal ?	VLIDVVL	Identifiant VL
UPEJECT	Personne éjectée ?	VLIMMAT	Immatriculation
UPLOCCHO	Localisation du point de choc	VLINTERV	N° Intervention
UPINTRUS	Intrusion directe sur occupant	VLANCIE	Ancienneté
UPAUTDEC	Autre personne décédée dans le véhicule	VLMOTOR	Motorisation
PLIDVPL	Identifiant PL	VLINCEN	Véhicule incendié ?
PLIMMAT	Immatriculation	VLOBSPR	Type obstacle principal
PLINTERV	N° intervention	VLAUTOB	Autre(s) obstacle(s)
PLINCEND	Véhicule incendié ?	VLENCAS	Encastrement ?
PLOUTILS	Utilisation d'outils ?	VLOUTIL	Utilisation d'outils ?
PLDEFCAB	Déformation de la cabine ?	VLCHOCFR	Choc Frontal
PLRENVER	Renversement ?	VLCHOCLG	Choc Latéral Gauche
PLTYPOB	Type d'obstacle principal	VLCHOCLD	Choc Latéral Droit
PLAUTOB	Autre obstacle	VLCHOCAR	Choc Arrière
PLCHOCFR	Choc Frontal	VLCHOCTN	Tonneau ou renversement
PLCHOCLG	Choc Latéral Gauche	VLCHOCAU	Aucune déformation
PLCHOCLD	Choc Latéral Droit	VLCHOCNR	Choc Non Renseigné
PLCHOCAR	Choc Arrière	Lij	Déformations vue latérale
PLCHOCTN	Tonneau ou renversement	Dij	Déformation Vue dessus
Fij	Déformation Vue face	UVIDUSA	Identifiant usager
Lij	Déformation Vue Latérale	UVNUSAG	N° Usager
UQIDUSA	Identifiant usager	UVIDVVL	Identifiant VL
UQNUSAG	N° Usager	UVIMMAT	Immatriculation
UQIDVQU	Identifiant Quad	UVCEINT	Personne ceinturée ?
UQIMMAT	Immatriculation	UVAIRBF	Airbag frontal
UQCEINTUR	Personne ceinturée ?	UVAIRBL	Airbag latéral
UQEJECT	Personne éjectée ?	UVLOPDC	Localisation du choc / occupant
UQLOCCHO	Localisation du point de choc	UVINTRU	Intrusion sur occupant
UQAUTDEC	Autre personne décédée dans le véhicule	UVOBJLO	La victime a-t-elle subit une charge par un objet lourd
QIDVQU	Identifiant Quad	UVVPIEG	Victime piégée dans le véhicule ?
QIMMAT	Immatriculation	UVAVPIE	Autre victime piégée dans le véhicule
QINTERV	N° Intervention	UVEJECT	Victime éjectée ?
QINCEND	Véhicule incendié ?	UVAPDEC	Autre personne décédée dans le véhicule
QCHOCFR	Choc Frontal	UVPLACE	Place occupée
QCHOCLA	Choc Latéral		

ANNEX 3 : DESCRIPTIVE STATISTICAL ANALYSIS OF VARIABLES

Figure 1. Gender distribution of users (coded USEXE) N=1646

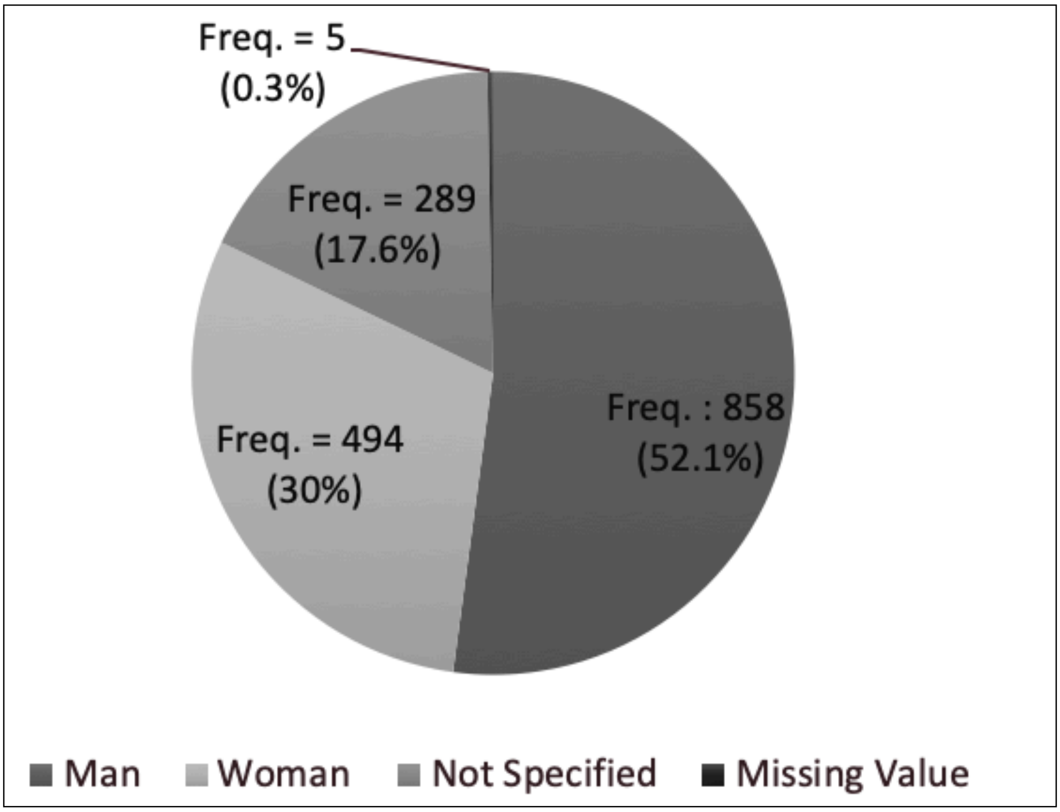


Figure 2a. Boxplot of accident users age (coded UAGE) N = 1646

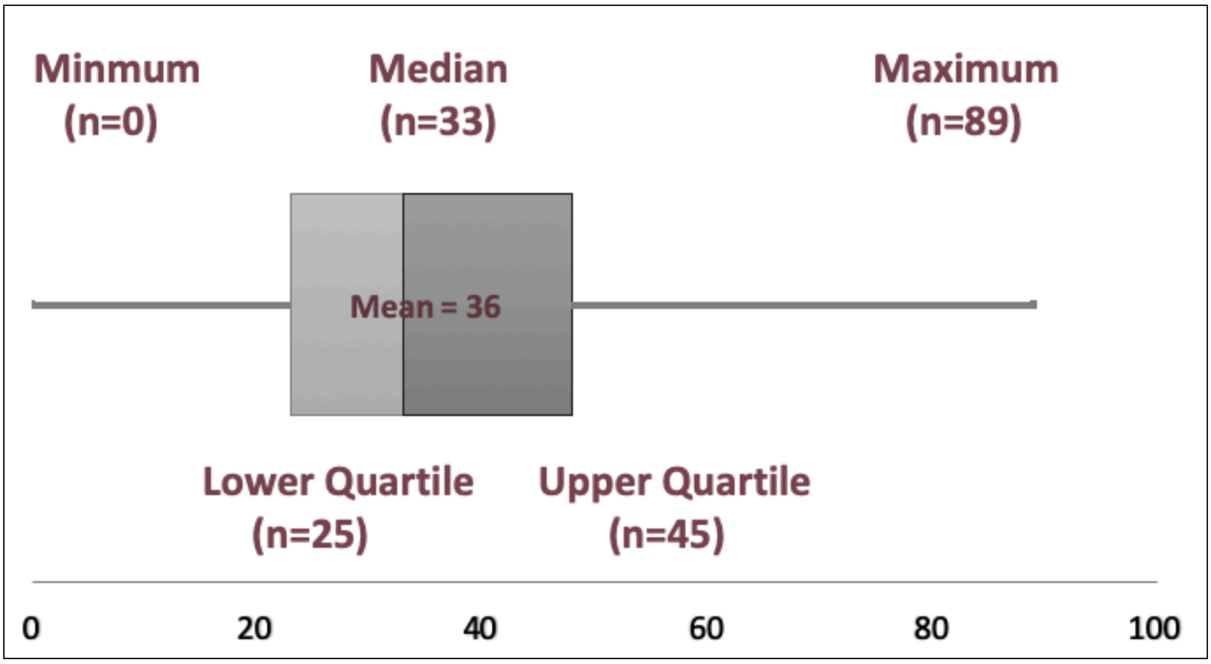


Figure 2b. Age Pyramid of users (N = 1646)

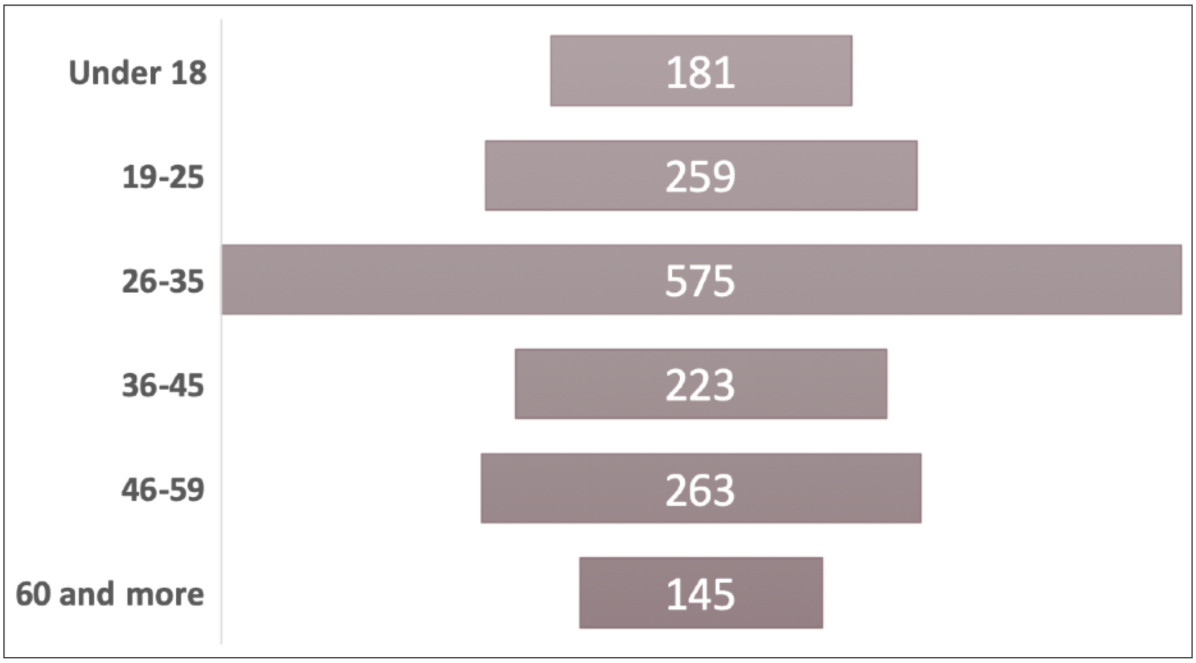


Figure 3. Bar chart of the Maximum Body Severity Index Global (coded MAISGLOB) according to the number of accident users (N=1646)

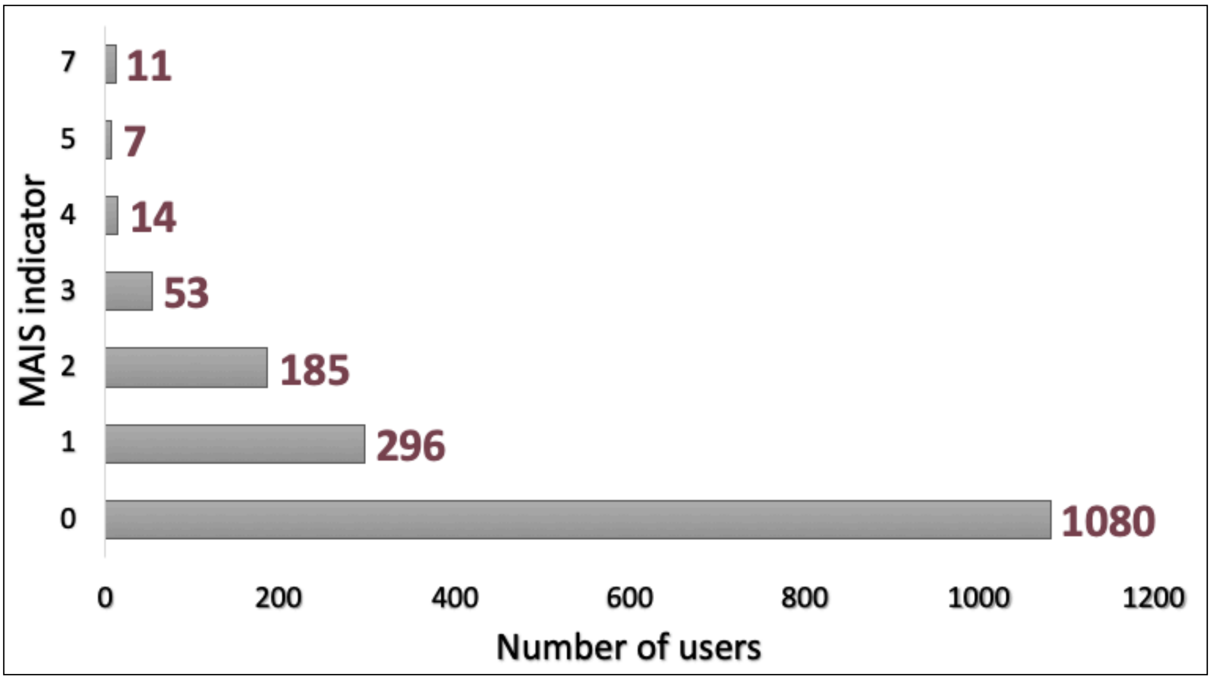


Table 1. Results of descriptive statistics of variables (N=1646)

Variables	Modalities	Modality Description	Freq.	%
« UGRAV » : presumed seriousness of the accident	1	Unscathed	783	47.57
	2	Deceased	11	0.67
	3	Injured	850	51.64
	NR	Not Specified	1	0.06
	NA	Missing Value	1	0.06
« APPAR » : appearance of the accident victim	1	Skinny	60	3.64
	2	Normal	1179	71.63
	3	Overweight	108	6.56
	4	Pregnant	7	0.43
	9	Not Specified	290	17.62
	NA	Missing Value	2	0.12
« ESTGRAV » : estimate of the severity of injuries noted out of 10 of the accident victims	0	Almost non-existent severity	1636	99.39
	1	Low severity	1	0.06
	5	Medium severity	1	0.06
	10	High severity	8	0.49
« CRANE » : if the user was hit in the skull, the appearance of the accident victim	0	No	1461	88.76
	1	Yes	185	11.24
« TRHOP » : user transported to hospital	1	Yes	796	48.36
	2	No	850	51.64
« TYPUS » : user type	1	Pedestrian	84	5.1
	2	Bike	25	1.6
	3	2 Motorized Wheels (2MD)	260	15.8
	4	Light Vehicle (LV)	1220	74.1
	5	Heavy goods vehicles	55	3.3
	6	Quad	2	0.1

ANNEX 4 : EXAMPLE OF RECODING OF THE MODALITIES

EXEMPLE WITH « APPAR », « UAGE » AND « HEURD » VARIABLES :

APPAR	Modalities Description
1	**Skinny**
2	**Normal**
3	**Overweight**
4	**Pregnant**
9	**Not Specified**
NA	**Missing Value**

→

APPAR	Modalities Description
1	Skinny
2	Normal
3	Overweight
4	Other

UAGE
1
20
48
…
80
NR

→

CLASSE_AGE	Modalities Description
1	under 18
2	19-25
3	26-35
…	…
6	>60
NR	Not Specified (replaced by the median)

HEURD
11 (am)
7 (am)
23 (pm)
…
NR

→

DAY	Modalities Description
1	Day
2	Night

The « HEURD » variable was in « date » format containing the hours of the accident. The transformation was carried out to have the night (> 19 hours) and day (> 7 hours) modalities.

A few comments :
- There have been cases in which certain variables with several modalities have a modality which was the most represented (such as for example the variable “ESTGRAV”). This is why these variables were not retained for the rest of the project.
- Other problems arise with "not applicable", "not informed" terms. In some variables, they are numerous and sometimes represent around 20 to 30% of the observations. That's why, we kept one more modality called "other" to avoid influencing other modalities.

ANNEX 5 : DESCRIPTIVE STATISTICAL ANALYSIS ABOUT LIGHT VEHICLES AND MOTORIZED 2-WHEELS VEHICLES USERS

Figure 1. Descriptive statistical analysis about Light Vehicle Users (N=1220)

Variable Groups	Variables/Description	Modalities	Reference	Freq.	%
ABOUT USERS	**CLASSE_AGE** : User age group	1: «under18 » 2: « 19-25 » 3: « 26-35 » 4: « 36-45 » 5 : « 46-60 » 6: « >60 »	1: «under18»	120 190 446 156 196 112	9,8 15,6 36,6 12,8 16 9,2
	CRANE : If there is a skull injury	1: Yes 0: No	0: No	1138 82	93,3 6,7
	TRHOP : Taken to hospital	1: Yes 2: No	/	466 754	38,2 61,8
	UENFA : Presence of children under 12 years old	1: Yes 2: No	1: Yes	50 1170	4,1 95,9
	APPAR : Physical appearance	1: « skinny » 2: « normal » 3: « overweight » 4: « other »	4: « other »	35 854 71 260	2,9 70 5,8 21,3
	DAY : Time of day of the accident	1: « day » 2 : « night »	1: « day »	973 230	79,5 18,5
ABOUT THE VEHICLE	**VLCHOCAU :** No deformation	1: Yes 0: No	0: No	20 901	1,6 73,8
	VLCHCONR : Unspecified impact	1: Yes 0: No	0: No	9 912	0,7 74,7
	VLCHOCTN : Barrel or overturn	1: Yes 0: No	0: No	46 875	3,8 71,7
	VLCHOCAR : Rear impact	1: Yes 0: No	0: No	204 717	16,7 58,8
	VLCHOCLD : Right side impact	1: Yes 0: No	0: No	126 795	10,3 65,2
	VLCHOCLG : Left side impact	1: Yes 0: No	0: No	166 755	13,6 61,9
	VLCHOCFR : Frontal impact	1: Yes 0: No	0: No	397 524	32,5 42,9
	VLOUTIL : Use of tools	1: Yes 0: No	1: Yes	20 901	1,6 73,8
	VLENCAS : Recessed vehicle	1: Yes 0: No	1: Yes	17 904	1,4 74,1

Variable Groups	Variables/Description	Modalities	Reference	Freq.	%
	VLMOTOR : Type of motorization	1: Thermal 2: Other	1: Thermal	894 27	73,3 2,2
	VLANCIE : Age of the vehicle	1: Old 2: Recent 3: Other	1: Old	201 705 15	16,5 57,8 1,2
ABOUT THE USER IN THE VEHICLE	**UVAVPIE :** Another victim trapped in the vehicle	1: Yes 2 : No	1.Yes	16 1009	1,3 82,7
	UVPLACE : Space occupied	1: Driver 2: Front-right passenger 3: Other	1: Driver	781 142 93	64 11,6 7,6
	UVVPIEG : Victim trapped in the vehicle?	1: Yes 2: No	1: Yes	22 1003	1,8 82,2
	UVOBJLO : Was the victim loaded by a heavy object?	1: Yes 2: No	1: Yes	16 1009	1,3 82,7
	UVINTRU : Intrusion on occupant	1: Yes 2: No	1: Yes	54 971	4,4 79,5
	UVLOPDC : Location of impact / occupant	1: Adjacent 2: Distant 3: Other	1: Adjacent	512 466 47	41,9 38,2 3,8
	UVAIRBL : Side airbag	1: Yes, deployed 2: Yes, not deployed 3: No 4: Other	1: Yes, deployed	68 452 396 109	5,6 37 32,5 8,9
	UVAIRBF : Front airbag	1: Yes, deployed 2: Yes, not deployed 3: No 4: Other	1: Yes, deployed	219 625 100 81	17,9 51,2 8,2 6,6
	UVCEINT : Belted person	1: Yes 2: No	1: Yes	1005 20	82,4 1,6

Figure 2. Descriptive statistical analysis about Motorized 2-Wheels Users (N=260)

Variable Groups	Variables	Modalities	Reference	Freq.	%
ABOUT USERS	**CLASSE_AGE** : User age group	1: «under18 » 2: « 19-25 » 3: « 26-35 » 4: « 36-45 » 5 : « >45 »	1: «under18»	28 55 74 51 52	10,8 21,1 28,5 19,6 0,2
	CRANE : If there is a skull injury	1: Yes 0: No	0: No	209 51	80,4 19,6
	TRHOP : Taken to hospital	1: Yes 2: No	/	224 36	86,1 13,8
	APPAR : Physical appearance	1: « skinny » 2: « normal » 3: « overweight » 4: « other »	1: « skinny »	12 211 30 6	4,6 81,1 11,5 2,3
	DAY : Time of day of the accident	1: « day » 2 : « night »	1: « day »	214 39	82,3 0,15
ABOUT THE VEHICLE	**DRPOSFI :** Vehicle final position	1: Left side 2: Right site 3: Other	1: Left side	94 93 59	36,1 35,8 22,7
	DRTYPOBS : Main obstacle type	1: LV 2: Other	1: LV	173 73	66,5 28
	DRDEFORM : Deformation	1: Front of 2: Everywhere 3: None 4: Other	3: None	96 28 34 88	36,9 10,8 13 33,8
	DRRAYURE : Scratch	1: Yes 0: No	0: No	147 99	56,5 38
	DRTYCHOC : Impact type	1: Frontal 2: Side 3: Back of 4: Other	1: Frontal	93 95 10 48	35,8 36,5 3,8 18,5
	UDPOTEAU : Post / tree	1: Yes 0: No	0: No	9 243	3,5 93,5
	UDVEHIC : Vehicle	1: Yes 0: No	0: No	119 133	45,8 51,1
	UDTROT : Pavement	1: Yes 0: No	0: No	10 242	3,8 93
	UDSOL : Ground	1: Yes 0: No	0: No	166 86	6,1 33
	UDBOTTES : Motorcycle boots wearing	1: Yes 2: No	1: Yes	46 206	17,7 79,2

Variable Groups	Variables	Modalities	Reference	Freq.	%
ABOUT THE USER IN THE VEHICLE	**UDPANTAL :** Motorcycle pants wearing	1: Yes 2: No	1: Yes	37 215	14,2 82,7
	UDBLOUS : Wearing a motorcycle jacket	1: Yes 2: No	1: Yes	108 144	41,5 55,4
	UDGANTS : Wearing gloves	1: Yes 2: No	1: Yes	172 80	66,1 3
	UDCINTEG : Full face helmet	1: Yes 2: No	1: Yes	187 65	71,9 25
	UDCASQUE : Helmet wearing	1: Yes 2: No	1: Yes	244 8	93,8 3
	UDPLACE : Occupied place	1: Driver 2: Other	1: Driver	239 13	91,9 5

ANNEX 6 : ODDS RATIO REPRESENTATION FOR THE 2 TYPES OF VEHICLE

Figure 1. Odds Ratio Representation (about LV)

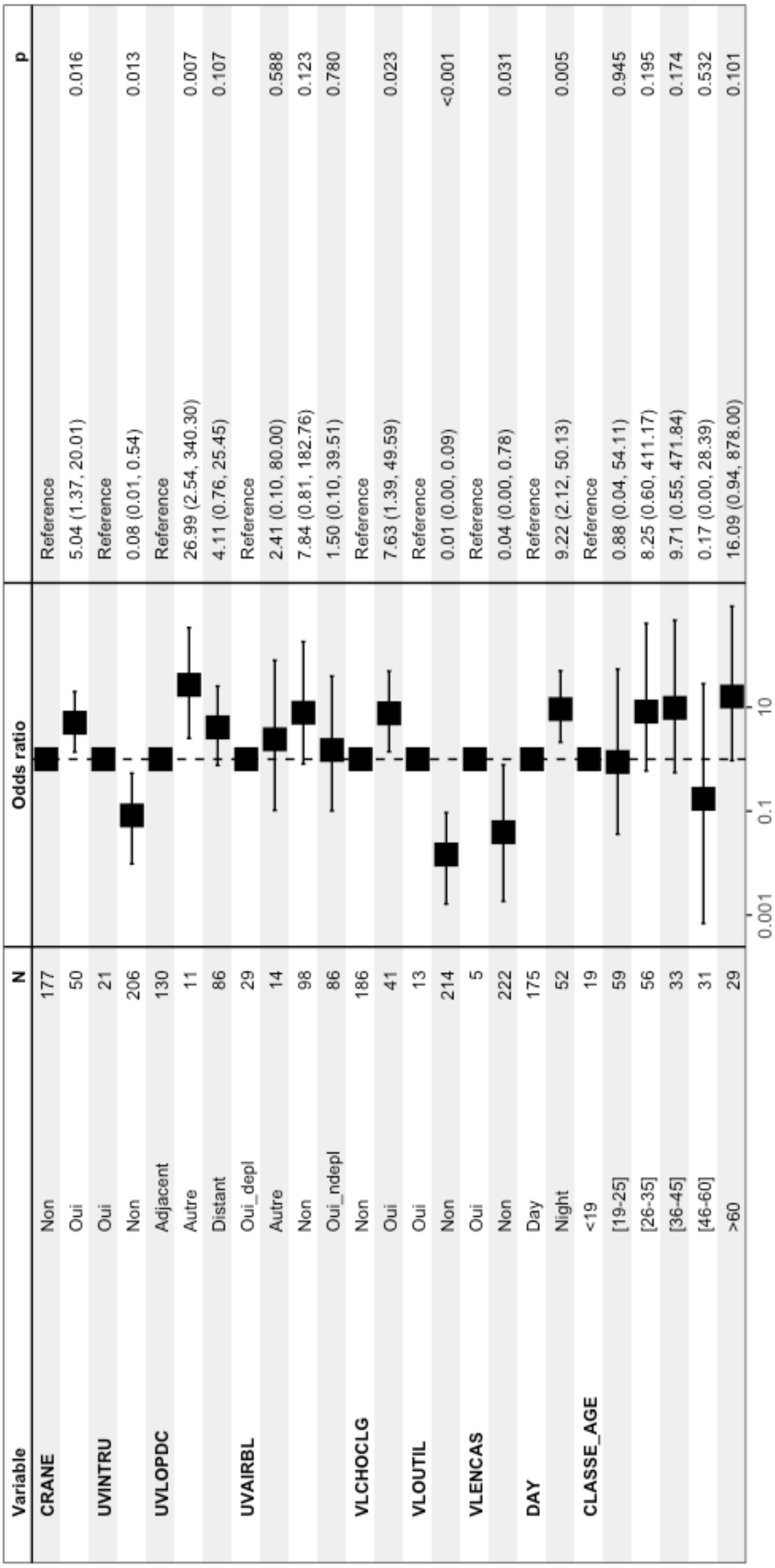
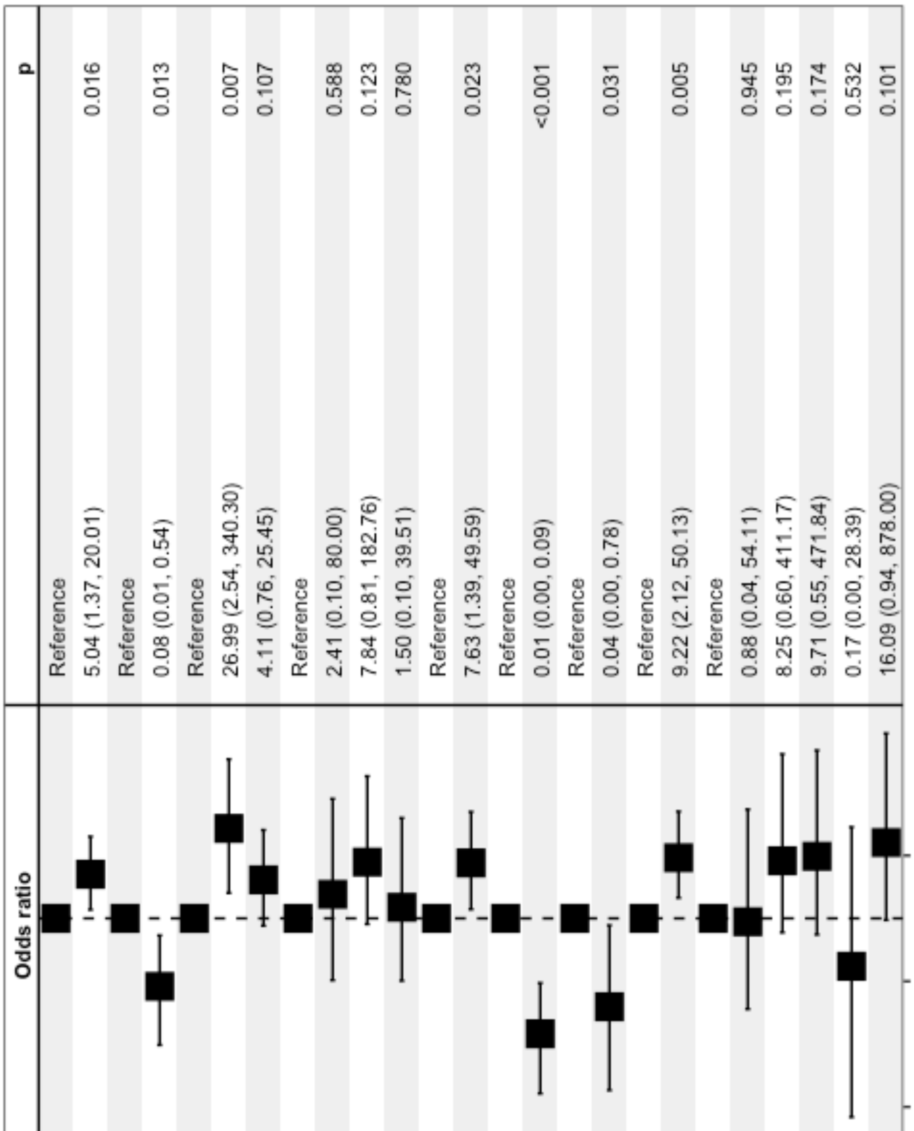


Figure 2. Odds Ratio Representation (about 2WM)



REFERENCES

1. **ONISR,** 2020, *Bilan 2019 de la sécurité routière,* Available online : <https://www.onisr.securite-routiere.gouv.fr/etat-de-l-insecurite-routiere/bilans-annuels-de-la-securite-routiere/bilan-2019-de-la-securite-routiere>.
2. **M.J.I GAUDRY ET M. DE LAPPARENT**, 2010, *La modélisation structurelle des bilans nationaux de l'insécurité routière: un état de l'art*. p35.
3. **S. PELTZMAN**, 1975, *The effects of Automobile Safety Regulation.*
4. **T. SAVOLAINEN ET AL.**, 2011, *The statistical analysis of highway crash-injury severities : a review and assessment of methodological alternatives.* Available online : <https://pubmed.ncbi.nlm.nih.gov/21658493>.
5. **BEDARD ET AL.**, 2002, *The independent contribution of driver, crash and vehicle characteristics to driver fatalities*. pp717-727. Available online : <https://www.sciencedirect.com/science/article/abs/pii/S0001457501000720>.
6. **R. GARRIDO ET AL.**, 2014, *Prediction of road accident severity using the ordered Probit model*. Available online : <https://www.sciencedirect.com/science/article/pii/S2352146514002701>.
7. **C. KONG ET AL**., 2010, *Logistic Regression Analysis of Pedestrian Casualty Risk in Passenger Vehicle Collisions in China.* Available online : <https://pubmed.ncbi.nlm.nih.gov/20441804/>.
8. **STEPHEN A. RIDELLA1, JONATHAN D. RUPP AND KRISTIN POLAND**, 2012, *Age-Related Differences in AIS 3+ Crash Injury Risk, Types, Causation and Mechanisms. Available online :* <http://www.ircobi.org/wordpress/downloads/irc12/pdf_files/14.pdf>
9. **D. LLOYD,** 2015, *Estimating clinically seriously injured (MAIS3+) road casualties in the UK.* Available online : <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/556648/rrcgb2015-03.pdf>.
10. **B, VIVIEN ET AL.,** 2008, *Critères et scores de gravité,* Available online : <https://sofia.medicalistes.fr/spip/IMG/pdf/scores_et_criteres_de_gravite_urgences_2008_vivien_riou_carli_.pdf>.
11. **BAKER SP ET AL.**, 1974, *The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care.* The Journal of Trauma.
12. **CHAMPION HR ET AL.,** 1989, *A Revision of the Trauma Score*. The Journal of Trauma.
13. **BOYD CR ET AL.**, 1987, *Evaluating Trauma Care : The TRISS Method*. The Journal of Trauma.
14. **CHAMPION HR ET AL.**, 1995, *Editorial Comment (Coefficients update)*. The Journal of Trauma.
15. **RUTLEDGE R ET AL.,** 1996, *Appropriate use of the Glasgow Coma Scale in intubated patients: a linear regression prediction of the Glasgow verbal score from the Glasgow eye and motor scores.* The Journal of Trauma.
16. **LEEPER, T.J.**, 2017. *Interpreting regression results using average marginal effects with R's margins*.
17. **KING, G. AND ZENG, L**., 2001, *Logistic Regression in Rare Events Data.* Political Analysis. Available online : <https://doi.org/10.1093/oxfordjournals.pan.a004868>.
18. **RENAUD LANCELOT,MATTHIEU LESNOFF,** 2005, *Selection de modèles avec l'AIC et critères d'information dérivé*s, Version 3. Available online : <http://www.cef-cfr.ca/uploads/Reference/alncelotLesnoff.pdf>.
19. **RIPLEY, B. D.**, 2003. *Model selection in complex classes of models*. Available online : <http://web.maths.unsw.edu.au/~inge/statlearn/ripley1.pdf>

20. **BEWICK, V., CHEEK, L. AND BALL, J**, 2005, *Statistic Review 14 : Logistic Regression. Critical Care*, Statistics Review 14, pp.112-118. Available online : <https://ccforum.biomedcentral.com/articles/10.1186/cc3045>.
21. **HOSMER D.W. AND LEMESHOW, S.,** 2000, *Applied Logistic Regression*. 2nd Edition, Wiley, New York. Available online : <https://doi.org/10.1002/0471722146>.
22. **HOSMER, D.W., LEMESHOW, S. AND STURDIVANT, R.X**., 1989, *The Multiple Logistic Regression Model*. *Applied Logistic Regression*, pp 25-37.
23. **PENG, C.Y.J. AND SO, T.S.H**., 20*02, Logistic Regression Analysis and Reporting: A Primer. Understanding Statistics: Statistical Issues in Psycholog*y, Education, and the Social Sciences, pp 31-70. Available online : <https://doi.org/10.1207/S15328031US0101_04>.
24. **DAVID M W**, 2011. *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation.* Journal of Machine Learning Technologies. Available online : <https://www.researchgate.net/publication/228529307_Evaluation_From_Precision_Recall_and_F-Factor_to_ROC_Informedness_Markedness_Correlation>.
25. **BEWICK, V., CHEEK, L. AND BALL, J.**, 2004*, Statistics Review 13: Receiver Operating Characteristic Curves.* pp 508-512. Available online : <https://doi.org/10.1186/cc3000>.