République Tunisienne
Ministére de l'Enseignement Supérieur et de la Recherche Scientifique
Université de Carthage - Ecole Supérieure de la statistique et de l'analyse de l'Information

Rapport de Projet de Fin d'Études présenté pour l'obtention du

Diplôme National d'Ingénieur en Statistique et Analyse de l'Information

Mariem Landoulsi

Analyse des facteurs de risque de mortalité par accident de la route en Tunisie

---

M.BELMUFTI Ghazi	Encadrant universitaire
M.Med Amine SOUGUIR	Encadrant
M.Med Mouloud HADDAK	Encadrant

Stage de Fin d'études effectué à

(Organisme d'accueil)

LISTE DES ABRÉVIATIONS

ONSR : l'Observatoire Nationale de la Sécurité Routiére

CAH : Classification Ascendante Hiérarchique

ACM : Analyse des Correspondances Multiples

SMOTE : Synthetic Minority Oversampling Technique

ROSE : Random Oversampling Examples

AUC : Area Under the Curve

AIC : Akaike Information Criterion

BIC : Bayesian Information Criterion

Résumé

Le bilan des accidents routiers en Tunisie montre une situation critique de la sécurité routière dans notre pays surtout lorsqu'elle est comparée aux autres pays.
Cette situation revient évidement à plusieurs facteurs qui se divisent en :facteurs humaines, c'est-à-dire les comportements des conducteurs et des utilisateurs de la route (excès de vitesse, défaut d'attention, dépassement interdit...),facteurs mécaniques (panne technique), facteurs géographiques liées à la région et la zone de déroulement de l'accident, facteurs liés à l'infrastructure du transport public, la prise en charge des victimes dans les hôpitaux, la qualité des routes...
Notre objectif est de faire une analyse approfondit de ces facteurs. La démarche suivie dans ce projet est de s'attacher, en premier temps, à décrire le jeu de données utilisé pour cette analyse.
Dans un deuxième temps, une analyse exploratoire (analyse des correspondances multiples ACM et classification hiérarchique ascendante CAH) aide à réduire le nombre de variables que nous avons utilisé par la suite pour la modélisation du risque de mortalité lié aux accidents de la route et donc identification des principaux facteurs liées à ce risque en utilisant un modèle de régression logistique. On a utilisé pour ce travail les logiciels R et Microsoft Excel.

Mots clés : sécurité routières, statistique descriptive, ACM, CAH, régression logistique, R, Excel.

Table des matières

Introduction

Le phénomène des accidents de la route constitue un enjeu majeur pour la santé publique et entraine des problèmes économiques et sociaux.
Ces accidents sont considérés comme l'une des causes principales de la mort dans le monde entier. En effet d'après l'OMS (l'Organisation Mondiale de la santé) plus de 1,3 million de personnes perdent la vie dans un accident de la circulation et de 20 à 50 millions de blessés sont enregistrées.
Bien que ce problème commence à être résolu dans la plupart des pays développés qui ont réussi à diminuer le nombre et la gravité des accidents de la route (par exemple, en France le nombre des décès à diminuer de 20% en 2020 par rapport à 2019), les pays en développement n'ont pas réussi à limiter l'insécurité routière.
En Tunisie, les accidents de la circulation représentent aujourd'hui la cinquième cause de décès, soit 3,3% de l'ensemble des décès constatés qui est de 50% supérieur à la moyenne européenne.
Au but d'améliorer ces chiffres et de lutter contre les dégâts des accidents routiers, l'Observatoire Nationale de la Sécurité Routière (ONSR) conçoit des programmes et des politiques pour la sensibilisation des usagers de la route et pour communiquer avec les responsables directes qui peuvent aider à résoudre les problémes de la sécurité routiére.
Mais, pour réussir ces stratégies il faut commencer par bien analyser les données disponibles et essayer de bien interpréter les éléments qui interviennent au risque de gravité des accidents.
C'est dans ce cadre que ce situe mon stage de fin d'étude effectué au sein de l'ONSR (l'Observatoire Nationale de la Sécurité Routière) en vue de l'obtention du titre d'Ingénieur en Statistique et Analyse de l'information.

Mon projet consiste à analyser les facteurs de risque de mortalité des accidents de la route en Tunisie en déterminant les principales causes de l'insécurité routière. Pour ce fait, une analyse approfondie de l'accidentologie routière en Tunisie sera faite en se basant sur des méthodes et des modèles statistiques.
Les données utilisées dans ce projet sont des données des années 2017,2018 ,2019 et 2021 fournis auprès de l'ONSR.
Nous essayerons donc de participer à l'amélioration de la politique de prévention des accidents de la route en Tunisie.

Chapitre 1

Présentation Générale

Avant de passer à l'élaboration de notre projet, Il est important de définir le cadre général de mon stage. Dans ce chapitre on va définir l'organisme d'acceuil,l'objectif du stage, la problématique et bien sur la démarche suivie durant le travail.

1.1 Organisme d'accueil

1.1.1 l'Observatoire National de Sécurité Routiére

Qu'est-ce que la sécurité routière ?

La sécurité routière est l'ensemble des règles et services qui ont pour but d'assurer la sécurité des usagers de la route : piétons, automobilistes, motards, cyclistes, etc...
Le concept de sécurité routière concerne la prévention d'accidents sur la route dans le but de protéger la vie des personnes.
Face à la recrudescence des accidents de la circulation routière, la Tunisie a été précurseur en mettant en place une stratégie nationale de sécurité routière dont l'objectif est de réduire d'une manière durable et continue le nombre de tués et de blessés graves. Cette stratégie est tridimentionnelle : infrastructurelle, législative et socioculturel.
Pour garentir tous les attributs de la sécurité aux usages de la route, la Tunisie a

consenti des efforts considérables à travers le ministère de l'Intérieur représenté par l'Observatoire national de la sécurité routière qui est responsable de l'éducation, de l'information, de la sensibilisation et du recueil des données sur les accidents de la route et grâce aussi aux associations pour la sécurité et la prévention routiére.

1.1.2 Définition

Un observatoire est un dispositif d'observation réalisé par un ou plusieurs organismes, pour suivre la progression d'un phénomène dans un domaine dans le temps et dans l'espace.
Une étude peut alors être déployée sur les données ou charger des fichiers préexistants, les calculer et générer des synthèses.

1.1.3 les services offferts

L'observatoire est responsable des tâches suivantes :

- Observer les conditions de sécurité routière et collecter les informations associées sur le plan national, les analyser et les répertorier dans une banque ou une base de données créée à cet effet.

- Mener des recherches et des évaluations nationales sur la sécurité routière et ses perspectives d'avenir.

- Publie des revues à court terme et périodiques liées au domaine de la sécurité routière.

- Coopérer avec des différents intervenants dans le domaine de la recherche routière.

- Concevoir des programmes et des politiques pour développer le domaine de la sécurité routière et recommander des mesures appropriées et développer la communication et la sensibilisation.

- Organiser des séminaires et des journées d'apprentissage et de formation.

1.1.4 Organigramme de l'observatoire National de sécurité Routiére

La figure ci-dessous présente l'organigramme de l'observatoire nationale de la sécurité routière constituée de deux divisions administratives « et du « Conseil scientifique » dont nos travaux mettent principalement en œuvre la cellule de données et de recherche et de coordination et de communication.

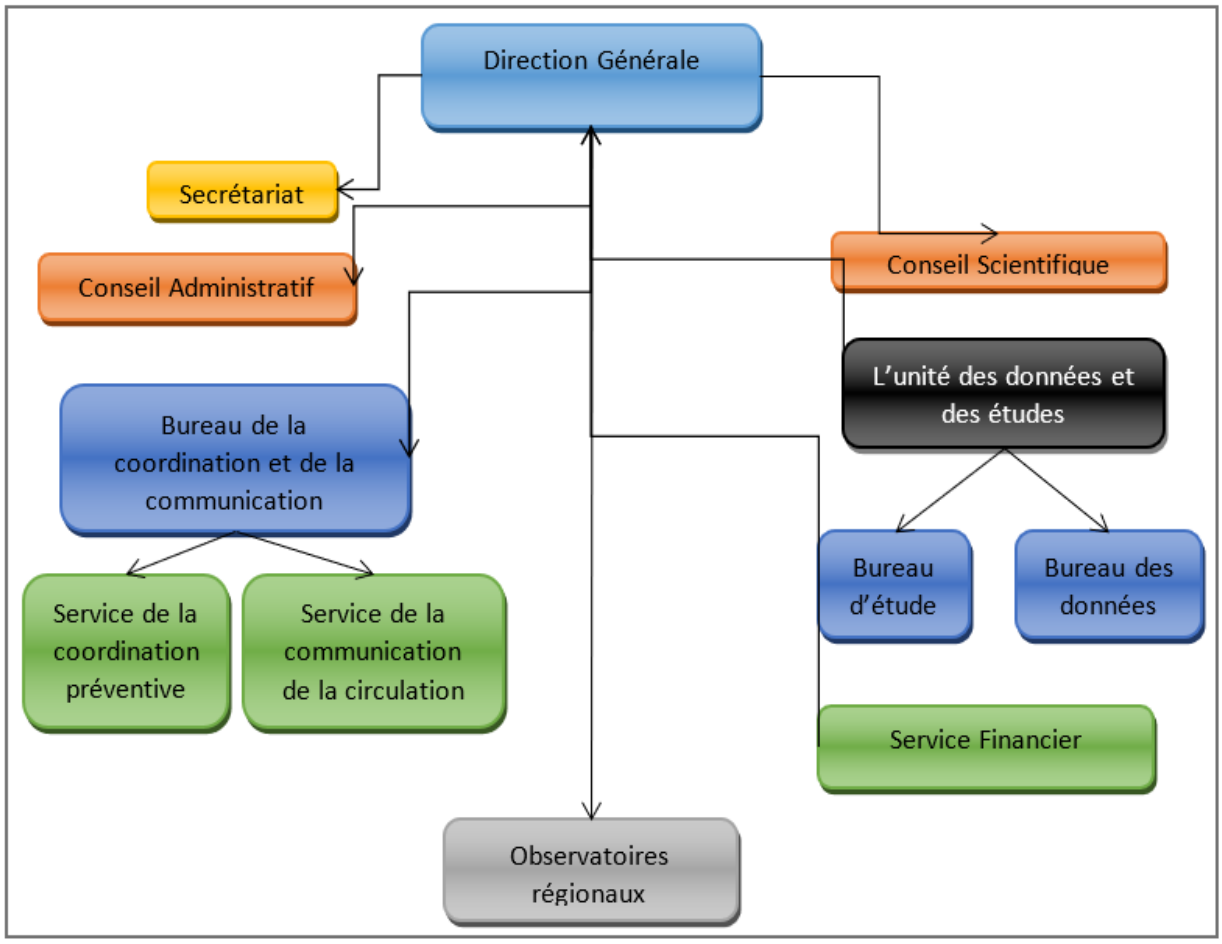

figure1 :Organigramme de l'observatoire Natinoal de Sécurité Routiére

1.1.5 les différents bureaux de l'observatoire

Bureau de la coordination et de la communication

Ce bureau est chargé de coordonner les actions des différentes administrations et associations impliquées dans le domaine de la prévention routière.
Ainsi que l'élaboration d'une stratégie d'information sur la sécurité routière, en

assurant le suivi de sa mise en œuvre.
Ce bureau est composé des services de dispatching et des services de circulation routière.

L'unité des données et des études

La tâche d'une telle unité est de collecter et de classer les données. Établir une base de données numérique et géographique sur la sécurité routière et être responsable de l'intégration de l'échange d'informations routières dans les réseaux nationaux et mondiaux. Ses tâches importantes comprennent la préparation de rapports réguliers sur la sécurité routière et la conduite de recherches et de recherches scientifiques sur des sujets liés aux organisations routières. Afin d'effectuer toutes les opérations ci-dessus, l'unité doit être composée de deux bureaux, un pour les données et un pour la recherche. A noter que la Veille Nationale de Sécurité Routière dispose d'un système d'information géographique qui permet d'organiser et de présenter des données alphanumériques, ainsi que de réaliser des plans et des cartes.

Le conseil scientifique

Le Comité Scientifique est présidé par le Directeur Général de l'Observatoire, qui assiste Dans le cadre de la recherche documentaire, il Questions scientifiques et techniques dans le champ d'application Observatoire et revue des plans annuels des séminaires et conférences scientifiques Et les recherches de l'observatoire. Ses responsabilités comprennent également l'évaluation et L'orientation du précédent plan de développement des données de l'observatoire Continuer sa sortie. Par conséquent, il est le président du comité scientifique, et l'autre Sous sa direction, des membres nommés par arrêté du ministre de l'Intérieur.

Le conseil administrative

Le directeur général est assisté par le conseil pour gérer l'observatoire Agences administratives qui expriment des avis sur les règlements et les organisations internes Observatoire, marché, contrat et convention seront conclus par l'observatoire Accepter les dons et toute question relative aux activités de l'observatoire, et Le directeur

général doit être effectivement soumis au conseil d'administration. Le comité administratif est Composé du président qui est le directeur général de l'observatoire et autres La répartition des membres est la suivante :

- Trois représentants du ministère de l'Intérieur

- Représentant du Ministère de la Justice et des Droits de l'Homme

- Représentant du Ministère des Transports

- Représentant du ministère de l'Éducation et de la Formation

- Représentant du Ministère de l'Equipement, du Logement et de l'Aménagement Territorial

- Représentant du Ministère de la Santé Publique Les données sont collectées avec deux principaux acteurs : la police nationale et la Garde Nationale.

1.2 Présentation du sujet

1.2.1 Objectif du stage

L'objectif de ce travail est d'analyser les facteurs qui causent la mort par accident de la route en Tunisie. En effet, on a l'habitude de dire que la majorité des accidents arrivent dans des circonstances très ordinaires à des gens ordinaires. Mais attention en reprenant ce constat, cela peut conduire à banaliser le risque. Beaucoup ont tendance à penser qu'il y aura toujours un résidu de risque qui serait la conséquence inévitable de tout déplacement.
Surtout dans un système non professionnel donc peu organisé, composé d'acteurs multiples aux caractéristiques et aux objectifs différents, interagissant sur un réseau dont les fonctions sont elles-mêmes multiples. Le diagnostic n'est pas faux, néanmoins les attitudes individuelles et collectives peuvent fortement moduler ce risque. Premiérement,nous allons étudier le risque d'accident corporel en définissant les différentes relations existantes entre les variables. Deuxièmement,l'analyse consiste à étudier la gravité des accidents mortels à travers differents conditions telle que le type des routes, les conditions de l'accident,les indicateurs de trafique...

1.2.2 Etat de l'art

Les accidents de la route constituent toujours un grave problème de santé publique aux niveaux mondial, régional et national. Bien que des mesures soient prises dans de nombreux pays pour améliorer la sécurité routière, il reste encore beaucoup à faire si l'on veut que le nombre de décès cesse d'augmenter.
Ces dernières années, diverses organisations ont adopté un certain nombre de méthodes pour estimer le nombre de décès consécutifs sur les routes dans le monde.
On trouve dans la littérature liée à ce projet, des nombreux travaux qui mettent en évidence les risques socioéconomiques et démographiques liées à la mortalité dans un accident.
Les études de l'accidentologie ont révélé que le phénomène de l'insécurité routière est le résultat du dysfonctionnement du système de circulation. Ce système est composé d'au moins trois éléments principaux : l'usager, le véhicule et la route[12]
L'étude de l'histoire des modèles de sécurité humaine se caractérise par un bilan incomplet, difficile du fait de la multitude des directions de recherche. Un certain nombre de travaux ont été développés par des auteurs de divers domaines de recherche tels que la psychologie, l'ingénierie, l'économie et l'économétrie. Chacun étudie et modélise l'accidentologie à sa manière.[12]
Les tentatives de modélisation économétrique des accidents ont cherché à expliquer l'évolution du nombre d'accidents de la route ou des décès qui ont eu lieu en fonction des variables agrégées de type macroéconomique. Plusieurs études ont été menées dans ce sens intégrant leurs modèles agrégés avec diverses variables explicatives : revenu moyen par habitant, consommation d'alcool par habitant, pourcentage de la population, taux de motorisation, croissance démographique, prix de l'essence, nouvelles réglementations, distance parcourue.[13]
Il existe aussi une étude faite par Peltzman [1975] qui a construit un modèle d'équations simultanées qui permet de distinguer la fréquence des accidents en fonction de leur gravité. Il a explique ces variables par la nouvelle réglementation des autorités américaines.
Selon cet auteur, ces pratiques réglementaires peuvent mieux réduire le nombre d'accidents que le nombre de décès.
Il y a des études qui disent que l'augmentation du nombre de véhicules motorisés

est l'un des facteurs qui contribuent à l'augmentation du nombre d'accidents de la circulation dans le monde.
Ces études ont prouvé qu'il existe une corrélation entre l'augmentation du nombre des véhicules automobiles et celle des accidents de la circulation et des traumatismes qui en résultent.
L'industrie automobile et l'augmentation subséquente du nombre de voitures à mesure que le développement des infrastructures routières profitent à la société, mais ce n'est pas sans les coûts que les accidents de la circulation contribuent.
C'est pourquoi plusieurs études attirent l'attention sur la nécessité d'une réflexion et d'une planification des mouvements de circulation, compte tenu de l'augmentation de la motorisation dans différentes parties du monde [14],(N2).
Les fluctuations de la taille relative des différents groupes de population affecteront Beaucoup d'informations sur le nombre de morts sur les routes. Par exemple, dans les pays industrialisés, les jeunes conducteurs et les jeunes motocyclistes qui sont plus susceptibles d'être impliqués dans des collisions sont actuellement trop nombreux dans les données sur les victime des accidents de la route[15].

1.2.3 Problématique

Le nombre des accidents routiers et des victimes de ces accidents n'ont pas cessé d'augmenter dans toutes les villes du mondes. Divers facteurs ont contribué à cette évalution ,parmi lesquels nous pouvons citer L'alcool,la drogue ou la fatigue. Cependant,il existe d'autres facteurs qui influencent clairement le risque d'accident mais on ne connait pas précisement leurs effets. C'est le cas des types de routes, du jour de la semaine, de l'heure de l'accidents...
Ce travail a donc pour objectif de traiter et analyser les facteurs de risques routiers les moins connus.

1.2.4 Conclusion

Dans ce chapitre nous avons procéder à une définition de l'organisme d'accueil du stage(ONSR). Par la suite, nous avons élaborer la problématique ainsi que les ob-

jectifs du travail.
Afin de poursuivre un enchaînement logique dans ce rapport, une étude préliminaire sera faite dans le chapitre suivant.

Chapitre 2

Etude préliminaire

Dans ce chapitre nous allons essayer d'organiser, comprendre et transformer les données brutes dont on dispose avant leur traitement et analyse. Il s'agit d'une étape importante avant le traitement proprement dit, qui implique souvent de corriger les données en utilisant des outils statistiques.

2.1 outils statistiques utilisés

Pour la mise en oeuvre de cette étude, nous avons utilisé le logiciel r et Microsoft Excel pour le nettoyage et la préparation de la table de données ainsi que l'élaboration des statistiques descriptives.

2.2 Présentation des données

La mission d'élaborer les bilans et de collecter les données des accidents de la circulation en Tunisie est confiée au ministère de l'Intérieur à travers sa structure l'Observatoire National de la Sécurité Routière (ONSR). En effet la collecte des données est journaliére et réalisée par la police de la circulation, qui ont la charge des accidents urbains et et par la garde nationale de la circulation qui ont la charge des accidents rurales.
Les rapports récapitulatifs élaborés par ces derniers sont transmis ,chaque semaines, respectivement au district régional de la police et au district de la Garde nationale

pour accumuler les données reçues et envoyer un rapport monsuel automatiquement à l'ONSR. L'étape suivante est le regroupement des données mensuelles par l'ONSR qui établit par la suite les statistiques nationales des accidents routiéres.
Ces données ne contiennent que les accidents posédents des vitimes c'est à dire les accidants ayant au moins un blessé car les forces de l'ordre ne sont pas toujours appelées e cas d'accident non mortel. Nous avons commencer à travailler sur les données annuelles fournits par l'ONSRT pour les années 2017, 2018, 2019, 2020.
Nous avons regroupé ces données pour obtenir notre jeu de données initiale qui comporte 23603 observations (accidents) décrits chacune par 11 variables qui sont :l'année de l'accident, le jour de l'accident,le gouvernorat conserné,le lieu de l'accident, le type de l'accident c'est à dire le type de collusion(les impliqués), la cause de l'accident, l'heure de l'accident, le nombre de tués et le nombre de blessés .
Ces variables sont accompagnées par d'autres variables binaires qui contiennent plus de détailles sur l'accident et les impliqués par exemple le type de l'impliqué, s'il est un motocycliste, un cyliste, un piéton...

2.3 Préparation des données

Pour la préparation des données nous avons rassembler, combiner, structurer, organiser et nettoyer les données afin de pouvoir les analyser.
En effet, notre jeu de données avait des incohérences d'où la nécessité d'un nettoyage pour améliorer la cohérence, fiabilité et valeur des données.
Premiérement, nous avons commencé à traduire nos données de l'arabe en français.Deuxièmement,nous avons choisi de travailler sur les victimes c'est à dire sur les accidentés dans le but d'éxploiter le maximum d'information ,donc nous avons construit à partir de la table initiale une nouvelle table de données où les observations sont les victimes , au cours de cette étape nous avons utilisé principalement les variables binaires de la table des accidents(table initiale) et ceci en utilisant le logiciel R. Le résultat obtenu est un jeu de données de 39283 victimes(lignes) présenté chacun par les variables suivantes :

— **Etat victime :** c'est une variable binaire qui décrit l'état de victime en indiquant si l'accidenté est vivant ou pas.

0 :blessé
1 :tué

— **Jour :**c'est une variable indiquant le jour de l'accident.
=> du lundi au Dimanche.
— **Type jour :**c'est une variable binaire indiquant si le jour est ouvré ou pas.
0 :ouvré => du lundi au vendredi
1 :non ouvré => samedi, dimanche et les jours fériés.
— **Type victime :** variable donnant l'information sur le type de victime c'est à dire le moyen de déplacement de l'accidenté.
Ses modalités sont :Conducteur véhicule,Motocycliste,Cycliste,Piéton,Passager.
— **gov :** variable crée à partir de la variable gouvernorat de la table initiale en regroupant les gouvernorats selon leurs emplacement.
Nord-Est=>Tunis,Bizerte,Ariana,Manouba,Zaghouan,Ben Arous,Nabeul.
Nord-Ouest=>Beja,Jendouba,Kef,Siliana.
Sud-Est=>Gabes,Sfax,Medenine,Tataouine.
— **Ramadhan :** variable binaire qui indique si l'accident a eu lieu au moins de ramadhan ou pas.
— **Gouvernorat :** variable indiquant la gouvernorat dans laquelle l'accident a eu lieu 24 modalités :
*Ariana
*Béja
*Ben Arous
*Bizerte
*Gabès
*Gafsa
*Jendouba
*Kairouan
*Kasserine
*Kébili
*Kef
*Mahdia

*Manouba
*Médenine
*Monastir
*Nabeul
*Sfax
*Sidi Bouzid
*Siliana
*Sousse
*Tataouine
*Tozeur
*Tunis
*Zaghouan

— **Lieu accident :** variable crée à partir de la variable Lieu de l'accident de la table initiale en regroupant les modalités pour obtenir 5 nouvelles modalités :
*Agglomération
*Autouroute
*Route locale
*Route Nationale
*Route Régionale

— **Impliqués :**Variable donnant l'information sur le type de l'accidenté de la maniére suivante :
-Véhicule léger :
-Véhicule lourd :
-Moto :
-Vélo :
-Piéton :

— **Cause :** C'est une variable construite à partir de la variable cause de la table initiale en regroupant les modalités selon le mécanisme de l'accident.
Ces modalités sont :
-défaut d'attention de Piéton
-défaut d'attention du conducteur
-dépassement interdit

-défaut technique
-collusion par l'arriére
-défaut d'attention lors de monter ou de déscendre
-non respect des régles de priorité
-Excès de vitesse
-Conduir sans permis

— **Heure Accident :** variable indiquant l'heure de l'accident.Elle est créee à partir de l'ancienne base en regroupant les modalités par tranches de 2h. Exemple : 6 :7 C'est à dire l'accident a eu lieu entre 6h et 8h.

— **Saison :** variable indiquant la saison de déroulement de l'accident.
Automne
Hiver
Été
Printemps

— **Vacance :** variable binaire indiquant si l'accident s'est déroulé aux vacances ou pas selon les calendriers universitaire et scolaire.
— **impiqué plus vulnérable :** variable donnant l'impliqué le plus vulnérable c'est à dire celui le plus exposé au risque.
Piéton
Vélo
Moto
Voiture
Taxi
Camionnette
Bus
Poids lourd

2.4 Analyse descriptive

Le but de l'analyse descriptive, dans cette partie, est de structurer et de visualiser l'information contenue dans les données. Dans cette paragraphe, on va réaliser, principalement, une analyse descriptives sur les variables qu'on a déjà présentées dans la paragraphe précédente .
Puisqu'on a décidé de travailler sur les victimes (c'est à dire observation=accidenté), notre analyse sera alors bivariée et elle consiste à étudier la gravité des variables par rapport aux autres variables. Notre étude est basée sur des variables catégorielles donc nous allons utiliser des diagrammes en bâton et des tableaux croisées.

2.4.1 le facteur type d'impliqué

Nous avons commencer à étudier la gravité des accidents par rapport au type d'impliqué.

EtattypeVic	Conducteu Re-morque	Conducteu véhicule	Cycliste	Motocycliste	Passager	Piéton	Total
blessé : Fréq	69	6925	638	6823	12448	7724	34627
Pourcentage	0,18	17,63	1,62	17,37	31,69	19,66	88,15
Pct.ligne	0.2	20.0	1.8	19.7	35.9	22.3	100.0
Pct.colonne	78.4	81.7	84.7	84.4	95.8	87.0	88.1
Tué :							
Fréq	19	1549	115	1264	569	1159	4655
Pourcentage	0.05	3,94	0,29	3,22	1,4	2,95	11,85
Pct.ligne	0.4	33.3	2.5	27.2	11.8	24.9	100.0
Pct.colonne	21.6	18.3	15.3	15.6	4.2	13.0	11.9

figure1 :Gravité des accidents par rapport au type d'impliqué

On constate que la poucentage des tués pour les conducteurs des véhicules(c'est à dire les conducteurs des voitures,des camions lourds,des camionnettes) est égale à

18,3 qui est plus important que la pourcentage des décés des autres types d'impliqués ce qui peut etre expliqué par le fait que les Tunisiens utilisent plus ces types de véhicules que les autres types.
On trouve aussi des pourcentages de decés élevées pour les cyclistes, les motocyclistes et les piétons.(respectivement égalent à : 15,3% , 15,6% , 13%).

2.4.2 le facteur cause d'accident

Dans cette partie, nous allons étudier la gravité des accidents par rapport à la cause d'accident c'est à dire la gravité liée à chaque cause d'accident.

Stat
Fréq
Pourcentage
Pct.ligne
Pct.Colonne

CauseEtat	Blessé	Tué	Total
Collusion par l'arriére	3995 10.17 92.3 11.5	331 0.843 7.7 7.1	4326 11.0
Conduire sans permis	220 0.56 86.3 0.6	35 0.089 13.7 0.8	255 0.6
Défaut d'attention piéton	3978 10.13 86.9 11.5	600 1.53 13.1 12.9	4578 11.7
défaut d'attention du conducteur	10787 27.46 88.7 31.2	1373 3.5 11.3 29.5	12160 31.0
défaut d'attention lors de monter ou de déscendre	443 1.128 90.4 1.3	47 0.12 9.6 1.0	490 1.2
défaut technique	970 2.47 79.6 2.8	249 0.634 20.4 5.3	1219 3.1
dépassement interdit	3015 7.68 89.7 8.7	345 0.88 10.3 7.4	3360 8.6
excés de vitesse	6693 17.04 82.7 19.3	1402 3.57 17.3 30.1	8095 20.6

	4526	273	
non respect des régles	11.52	0.7	4799
de priorité	94.3	5.9	12.2
	13.1	5.7	
	34627	4655	
	88.15	11.85	39282
Total	100.0	100.0	100.0
	100.0	100.0	

Tableau2 :Gravité des accidents par rapport à la cause d'accident

Un accident de la route peut avoir plusieurs causes : humaines, météorologiques, techniques. Néanmoins, le facteur d'inattention : l'erreur humaine(défaut d'attention de piéton et défaut d'attention du conducteur) est prépondérant et concerne plus de 42% de la totalité d'accidents.
Aussi, on constate que le facteur technique apparaît dans plus de 20% des accidents de la route mortels.
Les causes humaines les plus fréquentes : l'excés de vitesse, le dépassement interdit et la conduite sans permis sont présentes respectivement dans 30,1%, 7,4% et 0,8% des accidents mortels.
Le non-respect des régles de priorité est responsable de 5,7% des tués et de 13,1% des blessés.
Finalement, 11% des tués sont causés par la collusion par l'arriére.

2.4.3 Le facteur Mois

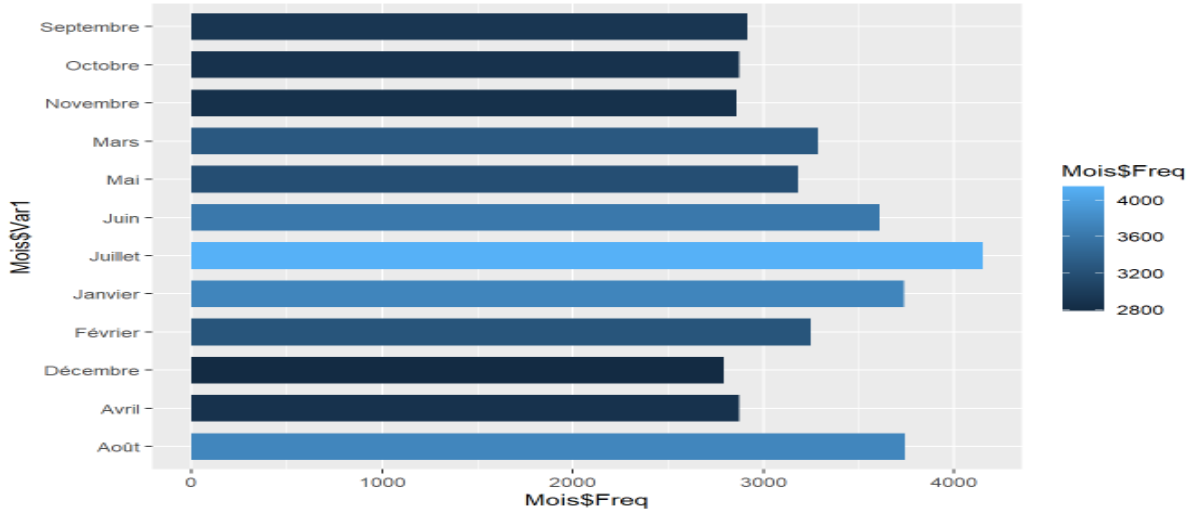

figure1 :Gravité des accidents par rapport au Mois de l'accident

D'aprés ce graphique on remarque que le nombre de victimes est trés élevé pendant les mois de Juillet et Août ce qui est expliqué par les vacances d'été puisque les Tunisiens se déplacent pendant cette période beaucoup plus qu'au cours de l'année. Aussi, le nombre d'accidents est élevé pour le mois de Janvier

2.4.4 Le facteur Source

La variable source est une variable donnat réference à la zone de l'accident c'est à dire urbaine ou rurale. Notons bien que les accidents déclarés par la police ont eu lieu au zones urbaines et ceux déclarés par la garde correspondent au zones rurales.

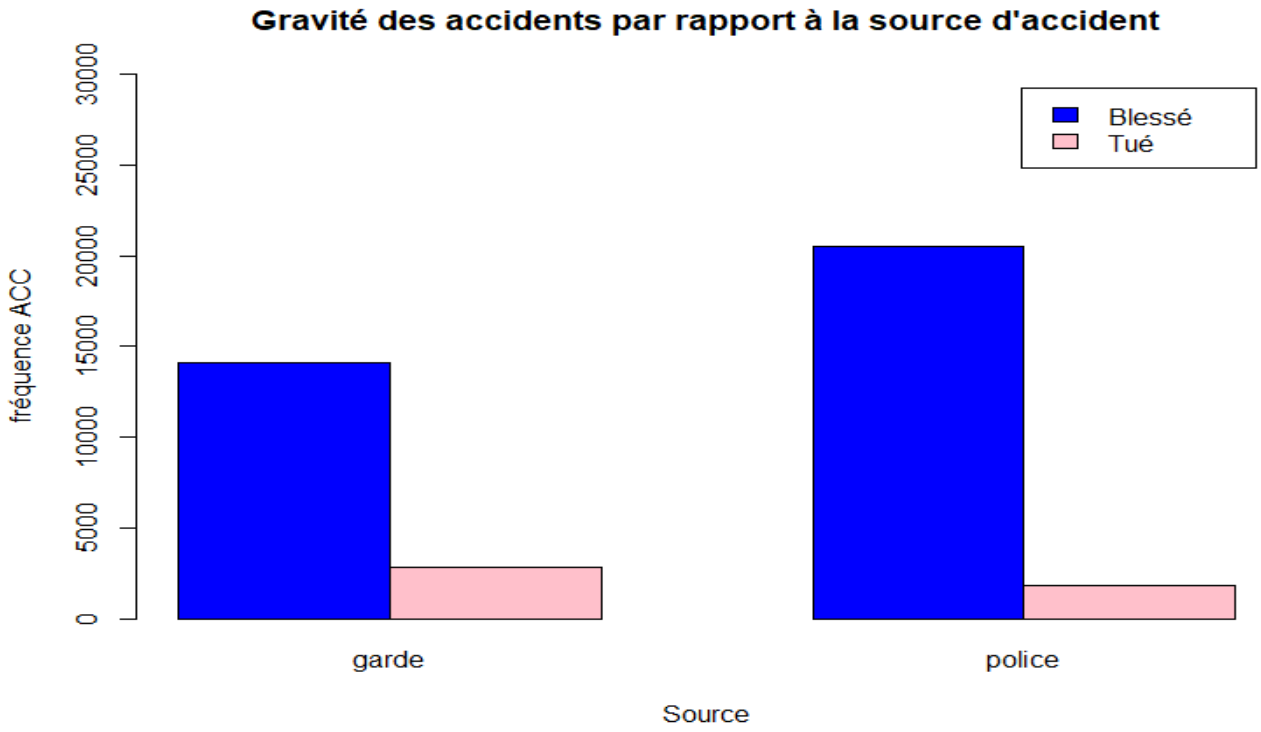

figure1 :Gravité des accidents par rapport à la source

On remarque que malgrés que le nombre de victimes est plus élevé au zones urbaines(police) la gravité est plus importante au zones rurales. En effet, le nombre de tués déclaré par la garde nationale est plus élevé que celui déclaré par la police ce qui peut être expliqué par le type des routes dans chaque zones. En fait, les zones urbaines sont plutot des agglomérations alors que les zones rurales contiennent des routes nationales, régionales.

2.4.5 Le facteur Lieu

La variable lieu d'accident est une variable donnant le type de route.Dans cette partie nous étudions la gravité des accidents par rapport au lieu de l'accident.

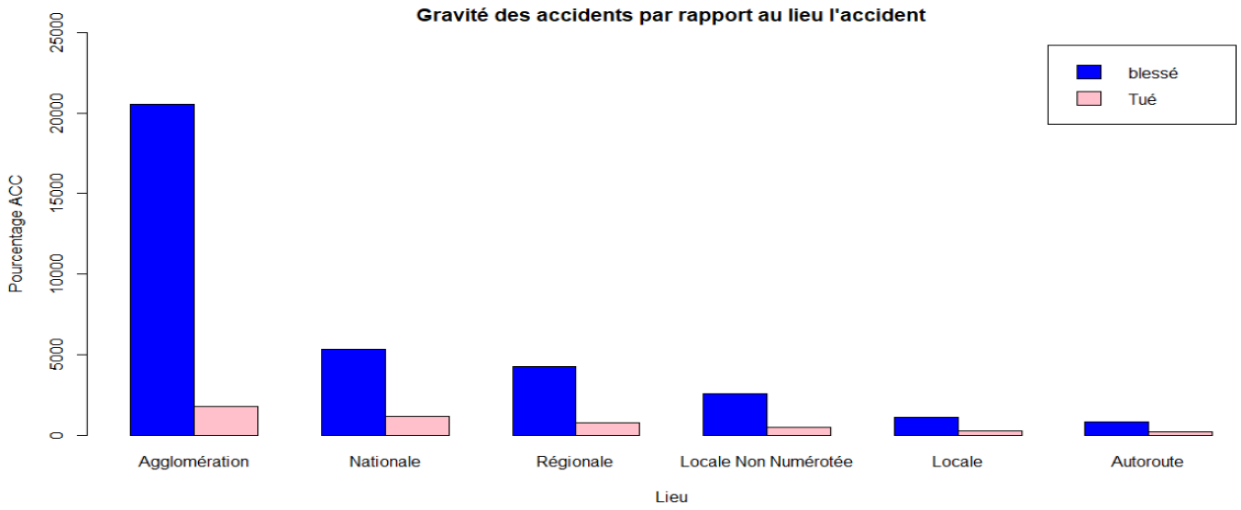

figure1 :Gravité des accidents par rapport au lieu

On remarque d'abord que le nombre d'accidents dans les agglomérations est plus élevé que dans les autres types de routes ce qui revient au fait que les déplacements sont plus élevè aux agglomérations.
Mais, en observant la gravité des accidents on trouve que les agglomération possédent le pourcentage des tués le plus bas(8% des accidents dans les agglomérations sont mortels) ce qui est expliqué par le fait que les vitesses moyennes sont moins importantes et le risque se limite généralement à de la tôle froissée.
Les accidents ayant lieu dans les routes nationales représentent 25.1% des tués et 15.5% des blessés.
Alors que les routes régionales sont réspensables de 16.8% des tués et 12.3% des blessés.
néanmoins,les accidents des routes locales et des autoroutes ont tendance à être moins graves en possédant respectivement 3.8% et 5.6% des tués.

2.4.6 Le facteur Vacances

Nous nous interessons maintenant, à la répartition des accidents en fonction de la variable vacances, c'est à dire durant les vacances et hors vacances(vacances scolaires).

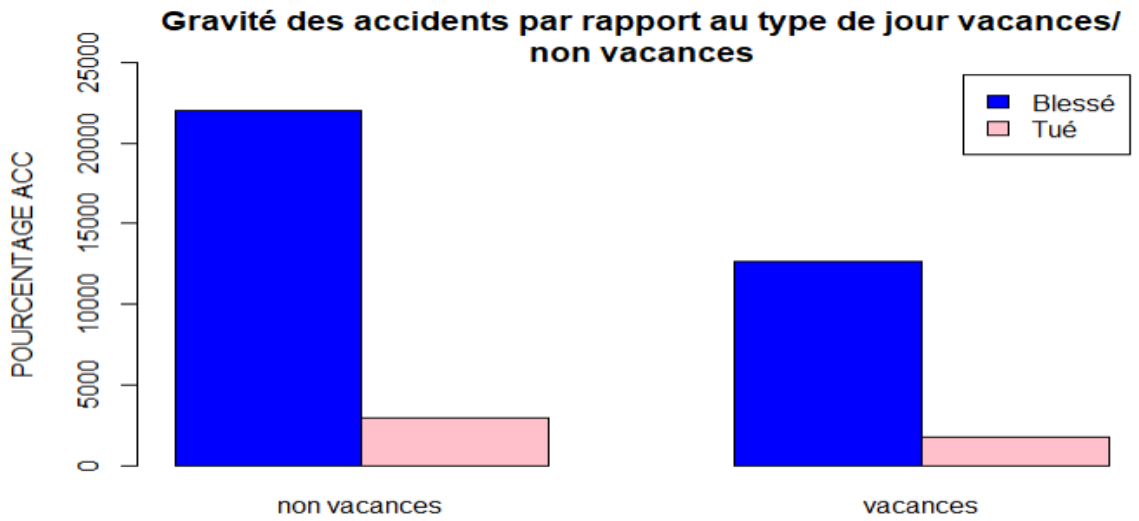

figure1 :Gravité des accidents par rapport au facteur vacances

Tout d'abord, nous remarquons que le nombre d'accidents est plus élevé au cours de l'année et hors vacances que pendant les vacances ; cela s'explique en partie par le fait que les jours des non vacances sont plus que les jours des vacances.
Pour les pourcentages et proportions, c'est difficile de les interpréter graphiquement, mais en revenant au tableau de proportions, on trouve que 64,9% des accidents ont eu lieu hors vacances scolaires.
De plus, ces accidents représentent 65% des blessés et 63,8% des Tués alors que les accidents qui ont eu lieu aux vacances sont responsables de 35% des blessés et de 36,2% des tués.
On constate aussi que la gravité est plus élevée pendant les vacances c'est à dire qu'un accident en jour de vacances aurait plus de risque de décés q'un accident en jour non vacances.
En effet, 12.2% des accidents en jours de vacances sont motels contre 11.7% pour les jours non vacances.

2.4.7 Le facteur Heure

Dans cette partie nous étudions l'évolution du nombre d'accidents par rapport au facteur heure. Rappelons qu'on a répartie la variable heure par tranche de deux heures.

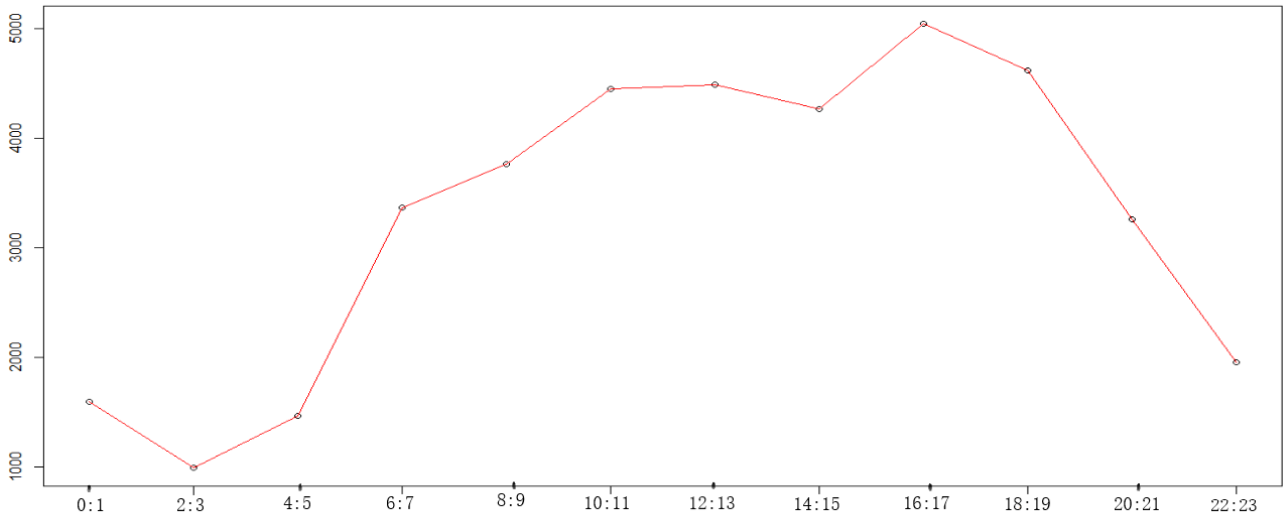

figure1 :Evolution des accidents par tranche horaires

D'aprés ce graphique, on peut remarquer que le nombre d'accidents devient de plus en plus élevé à partir du 4h du matin.
Ensuite, on peut constater clairement que ce nombre devient important entre 6h et 11h pour atteindre 4452 accidents à 11h.
Ce qui peut être expliqué par le fait que cette tranche d'heures correspond aux horaires de pointe et donc le nombre élevé d'accidents est du à la concentration du trafic.
On remarque aussi un nombre élevé d'accidents entre 12h et 13h c'est à dire durant la pause de déjeuner et donc lorsque les gens se déplace pour rentrer chez eux ou aller aux restaurents. Ce nombre est égale à 4494 accidents.
Ensuite, on observe un pic entre 16h et 17h avec un nombre d'accidents qui dépace les 5000 accidents.
Ces périodes correspondent en général au moment où la majorité des personnes actives rentrent à leur domicile.
Finallement,On note une chute du nombre d'accidents pour la période entre 18h et 23h où les déplacement sont beaucoup moins qu'au cours de la journée.

2.4.8 Le facteur impliqué plus vulnérable

Maintenant nous étudions la variable impliqué plus vulnérable. Pour cela nous allons observer cette variable en fonction des jours de la semaine.

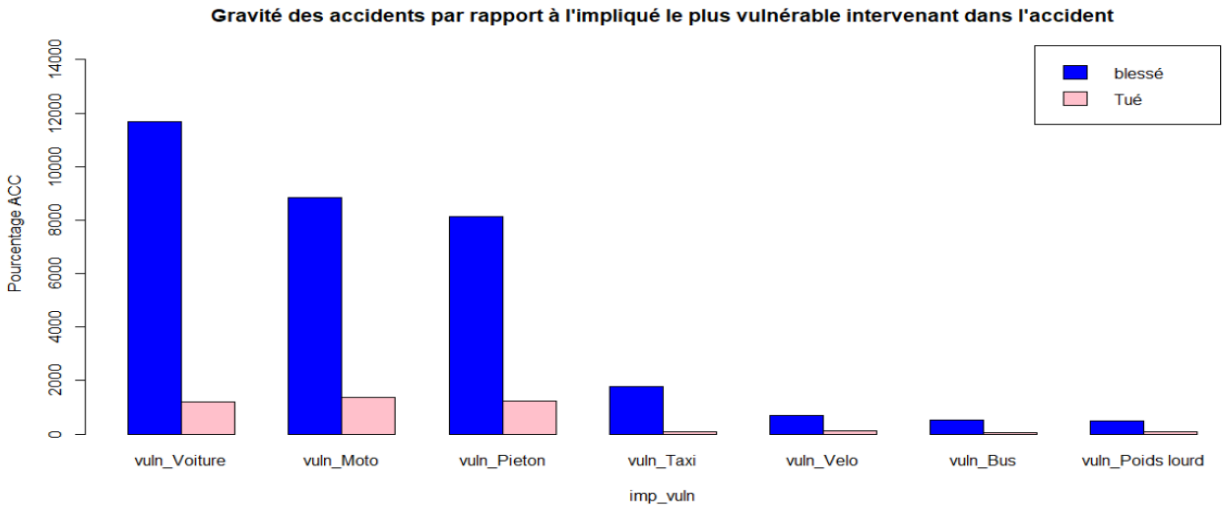

figure1 :Gravité des accidents par rapport à l'impliqué le plus vulnérable intervenant dans l'accident

2.4.9 Le facteur Jour

Maintenant nous étudions la variable jour. Pour cela nous allons observer la gravité des accidents en fonction des jours de la semaine.

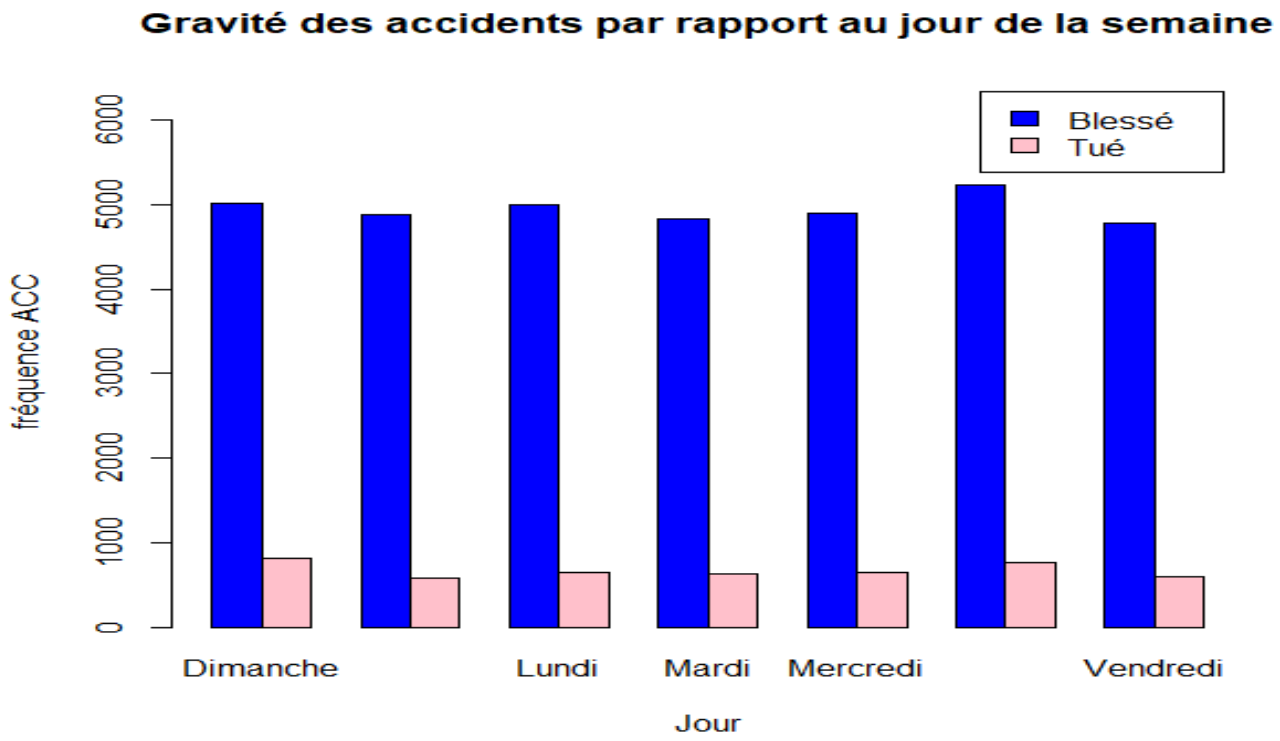

figure1 :le facteur Jour

2.4.10 Conclusion

Cette étude déscriptive nous a permis de mettre en forme et d'analyser nos données, de bien comprendre les variables ainsi que de déterminer un certain nombre de grandeurs caractéristiques de la population. Ce travail va nous aider par la suite à décider de la méthodoligie la plus adéquate pour l'analyse et la modélisation.

Chapitre 3

Analyse multivariée

L'étape de l'analyse multivariée permet de réduire le nombre de variables, qu'on va les utiliser dans la modélisation, en faite elle va éliminer les variables qui n'ont pas trop d'influence sur notre variable d'intérêt.

3.1 Analyse Factorielle des correspondances multiples

3.1.1 Fondement théorique

L'analyse des correspondances multiples (ACM) est une méthode de déscription statistique multidimensionnelle d'un tableau de données qualitatives dont l'objectif est de résumer l'information et de la synthétiser en réduisant les dimentions.
A l'instar des autres méthodes factorielles, l'ACM cherche à trouver des ressemblances,des différences, aux proximités entre individus et entre les modalités des variables qualitatives. Elle permet donc d'étudier le lien entre ces variables.
Le principe de l'ACM est de réduire l'information et de la concentrer sur les premiers axes en opérant un changement de systémes de coordonnées.
L'analyse se concentrera donc sur les premiers axes qui résume l'information, car les autres axes n'apportant qu'une faible part additionnelle d'information.
d'analyse factorielle adaptée aux données qualitatives (aussi appelées catégorielles). Elle permet d'étudier le lien entre plusieurs variables qualitatives.

En effet, l'ACM permet d'étudier le lien entre ces variables par l'intermédiaire d'un tableau disjonctif complet (TDC) ou du tableau de Burt (TB). Dans de tels tableaux de données, les individus (en lignes) sont décrits par un ensemble de variables qualitatives (en colonnes).

3.1.2 Application aux données des accidents des années 2017,2018 et 2020 en Tunisie

Test de significativité des variables

L'objectif de notre travail est d'étudier la gravité de l'accident par rapport aux différentes variables qualitatives. Donc il faut étudier l'indépendance entre la variable Etat et les variables mises en question. On formule naturellement deux hypothèses :
Hypothèse H0 : les deux variables étudiées sont indépendantes.
Hypothèse H1 :les deux variables étudiées ne sont pas indépendantes.
Cette hypothèse exprime le contraire de H0.
Nous avons commencer par faire un tableau croisé entre la variable Etat et chacune des autres variables.
Différents tests sont disponibles pour déterminer si la relation entre deux variables de tableau croisé est significative.
L'un des tests les plus courants est le khi-deux. L'un des avantages du khi-deux est qu'il est adapté à la majeure partie des types de donnée.

1. Gravité des accidents par rapport au type de jour

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  TAB1
## X-squared = 40.948, df = 1, p-value = 1.564e-10
```

figure1 :Test de khi-deux entre les deux variables Etat et type_jour

On trouve p-value = 1.564e-10 donc <0.05 d'où l'existance d'un lien statistique entre les deux variables Etat et type_jour.
Ainsi, l'hypothése H0 :les deux variable sont indépendantes est rejetée (car p-value est <0.05).

Donc, le type de jour a une influence sur l'etat de victime (c-à-d la gravité de victime).

2. Gravité des accidents par rapport à la saison de déroulement d'accident

```
##
##  Pearson's Chi-squared test
##
## data:  TAB3
## X-squared = 11.701, df = 3, p-value = 0.008481
```

figure1 :Test de khi-deux entre les deux variables Etat et saison

On contate que la p-value=0.008481>0.05.
Donc pas de lien statistique entre les deux variables Etat et saison. $\Longrightarrow$ On accepte l'hypothése H0 (car p-value est >0.05) Ce qui nous permet de conclure que la variable saison de l'accident n'influence pas l'etat de victime.

3. Gravité des accidents par rapport à la variable Ramadhan.

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  TAB6
## X-squared = 2.1163, df = 1, p-value = 0.1457
```

figure1 :Test de khi-deux entre les deux variables Etat et Ramadhan

On trouve que la p-value est égale à 0.1457>0.05 Donc il n'esiste pas un lien statistique entre les deux variables. l'hypothése H0 :les deux variables ETAT et Ramadhan sont indépendantes est accéptée (car p-value >0.05) =>la variable Ramadhan n'a pas une influence sur l'etat de victime (c-à-d la gravité de victime). La p-value est supérieure à 0.05, nous ne pouvons donc pas considérer que les variables sont liées.

4. Gravité des accidents par rapport à la variable Annee.

```
##
##  Pearson's Chi-squared test
##
## data:  TAB8
## X-squared = 1.1854, df = 3, p-value = 0.7565
```

figure1 :Test de khi-deux entre les deux variables Etat et Annee

p-value est = 0.7565=> >0.05 Puisque la p-value est supérieure à 0.05, nous ne pouvons pas considérer que les variables sont liées.

5. Gravité des accidents par rapport au autres variables.

variable	khi2
Mois	1.215929e-07
Jours	6.924502e-08
Gouvernorat	8.087167e-117
Lieu	1.493402e-159
Causes	8.976537e-122
Heure	7.187426e-66
Source	4.916405e-158
vac	1.402917e-02
type_vic	5.611839e-42
imp_vuln	1.765381e-55
region	1.718870e-40

figure1 :Test de khi-deux entre la variables Etat et autres variables

On constate que les p-values sont toutes inférieures à 0,05.Donc toutes les variables étudiées ci-dessus sont significativement liées à la gravité des accidents.

Conclusion : les variables : Saison, Ramadhan et Annee ne sont pas significatives. Alors que les variables :Type_Jour, Type_Victime, Saison, Mois, Lieu, Gouvernorat, Jours, Type_ACC, Vac, Heure, Région, Source et Causes sont significativement liées à la gravité de l'accident.

Données utilisées

L'ACM est menée sur les variables :
-Jours :Lundi,Mardi,Mercredi,Jeudi,Vendredi,Samedi,Dimanche.
-Gouvernorat
-Type_Vic
-Mois

-Lieu

-Cause

-Heure

-Source

-vacances

-imp_Pvul

La variable Etat sera la variable qualitative supplémentaire de l'ACM, donc elle ne participe pas à la construction des axes factorielles mais elle permet d'améliorer l'interprétation et ses coordonnées seront prédites.

Choix du nombre d'axes

Pour le choix du nombre d'axes on procéde au Critère de Coude dont le principe est d'observer un décrochement (coude) suivi d'une décroissance régulière sur l'eboulis des valeurs propres et de sélectionner les axes avant le décrochement.

	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	3.408492e-01	4.3698620368	4.369862
dim 2	2.267252e-01	2.9067332452	7.276595
dim 3	1.948697e-01	2.4983289307	9.774924
dim 4	1.814074e-01	2.3257355052	12.100660
dim 5	1.678319e-01	2.1516909609	14.252351
dim 6	1.431429e-01	1.8351655597	16.087516
dim 7	1.376653e-01	1.7649397681	17.852456
dim 8	1.297088e-01	1.6629328953	19.515389
dim 9	1.269688e-01	1.6278046151	21.143194
dim 10	1.242007e-01	1.5923170787	22.735511
dim 11	1.215622e-01	1.5584897849	24.294000
dim 12	1.175890e-01	1.5075512993	25.801552
dim 13	1.122301e-01	1.4388474082	27.240399
dim 14	1.119616e-01	1.4354046118	28.675804
dim 15	1.117110e-01	1.4321926763	30.107996
dim 16	1.104190e-01	1.4156288228	31.523625
dim 17	1.090539e-01	1.3981267911	32.921752
dim 18	1.085861e-01	1.3921298515	34.313882
dim 19	1.078065e-01	1.3821343656	35.696016
dim 20	1.072695e-01	1.3752493795	37.071266
dim 21	1.064008e-01	1.3641134349	38.435379
dim 22	1.062397e-01	1.3620470506	39.797426
dim 23	1.051035e-01	1.3474802796	41.144906
dim 24	1.044391e-01	1.3389627857	42.483869
dim 25	1.040204e-01	1.3335946542	43.817464

figure1 :Valeurs propres et pourcentages des variances dans l'ACM

On remarque tout d'abord que pour les premiers axes la variance expliquée est trés faible et que seulement 43% de l'information est obtenue avec 25 axes.

Donc, une correction de la pourcentage d'inertie est nécessaire.

Pourcentage d'inertie restituée – Correction de Greenacre : Pour disposer d'une indications plus réaliste sur la qualité des facteurs, nous devons utiliser un indicateur corrigé qui élimine le biais due à la redondance artificielle introduite dans les données. C'est le propos de la « correction de Greenacre ».
Principe de la correction : Cette correction s'appuie sur l'idée qu'une partie de l'information est redondante dans les données présentées à l'algorithme de l'ACM.
Elle reconsidère la proportion d'inertie portée par les facteurs. Une partie de l'information est triviale dans le tableau de Burt, il s'agit du croisement endogène de chaque variable.
Rappelons que l'inertie du tableau de Burt équivaut à la somme des valeurs propres de l'AFC qui lui est appliquée. Elle est égale aussi à la somme des carrés des valeurs propres de l'ACM sur le tableau des indicatrices.
La même quantité correspond à la moyenne des inerties des sous-tableaux qui composent le tableau de Burt.

$$I_{Burt} = \sum_{h=1}^{H_{max}} \mu_h$$

$$I_{Burt} = \sum_{h=1}^{H_{max}} \lambda_h^2$$

Si nous souhaitons nous concentrer sur l'information « utile », nous retirons les blocs diagonaux représentant les croisements endogène des variables. L'information à traiter devient :

$$S'' = \frac{p}{p-1}\left(I_{Burt} - \frac{M-p}{p^2}\right)$$

Greenacre, plutôt que de forcer artificiellement la somme cumulée des proportions à 100% sur les facteurs respectant la condition $\lambda>1$, propose de diviser la valeur propre corrigée $\lambda h'$ par (S") pour
disposer d'une vision moins optimiste de la qualité des facteurs.
Le pourcentage corrigé de variance restituée par le facteur s'écrit :

$$\tau_h'' = \frac{\lambda_h'}{S''} \times 100$$

```
f=as.vector(res.MCA$eig)[1:78]
p=11#nombre de variable
M=90#nombre de modalités
#récupérer les valeurs propres supérieur à (1/p)
lambada = f[1:57]
#appliquer la correction
lambada_prim = ((p/(p-1))*(lambada-1/p))**2
#somme corrigée de Greenacre (S'')
S2nd = (p/(p-1))*(sum(f**2)-((M-p)/(p**2)))
#pourcentage corrigé Greenacre
percent_2nd = lambada_prim/S2nd*100
```

	lambda	lambada_prim	percent	percent_2nd	cumsum.percent.	cumsum.percent_2nd.
1	0.3408492	0.0755887939	4.369862	24.3718635	4.369862	24.37186
2	0.2267252	0.0223196765	2.906733	7.1964650	7.276595	31.56833
3	0.1948697	0.0130774371	2.498329	4.2165180	9.774924	35.78485
4	0.1814074	0.0099098255	2.325736	3.1951947	12.100660	38.98004
5	0.1678319	0.0071597125	2.151691	2.3084842	14.252351	41.28853
6	0.1431429	0.0033013304	1.835166	1.0644379	16.087516	42.35296
7	0.1376653	0.0026452334	1.764940	0.8528945	17.852456	43.20586
8	0.1297088	0.0018215519	1.662933	0.5873174	19.515389	43.79318
9	0.1269688	0.0015733627	1.627805	0.5072945	21.143194	44.30047
10	0.1242007	0.0013410834	1.592317	0.4324014	22.735511	44.73287
11	0.1215622	0.0011369321	1.558490	0.3665775	24.294000	45.09945
12	0.1175890	0.0008612993	1.507551	0.2777061	25.801552	45.37715
13	0.1122301	0.0005500483	1.438847	0.1773504	27.240399	45.55451
14	0.1119616	0.0005362798	1.435405	0.1729111	28.675804	45.72742
15	0.1117110	0.0005235919	1.432193	0.1688201	30.107996	45.89624
16	0.1104190	0.0004605725	1.415629	0.1485010	31.523625	46.04474
17	0.1090539	0.0003983728	1.398127	0.1284461	32.921752	46.17318
18	0.1085861	0.0003780980	1.392130	0.1219090	34.313882	46.29509
19	0.1078065	0.0003454813	1.382134	0.1113925	35.696016	46.40648
20	0.1072695	0.0003238703	1.375249	0.1044245	37.071266	46.51091

figure1 :valeurs propres corrigés, pourcentage de l'inertie et pourcentage de l'inertie corrigé

On remarque qu'aprés la correction, les 7 premiers axes présentent un éventuel intérêt. En effet, 43% de l'information est obtenue avec ces axes. Cela signifie que la représentation des variables dans le plan factoriel ne reflète que 43% de la réalité. Donc on peut garder seulement les 7 premiers axes.

Ce résultat est récupéré aussi en observant la courbe des valeurs propres suivante :

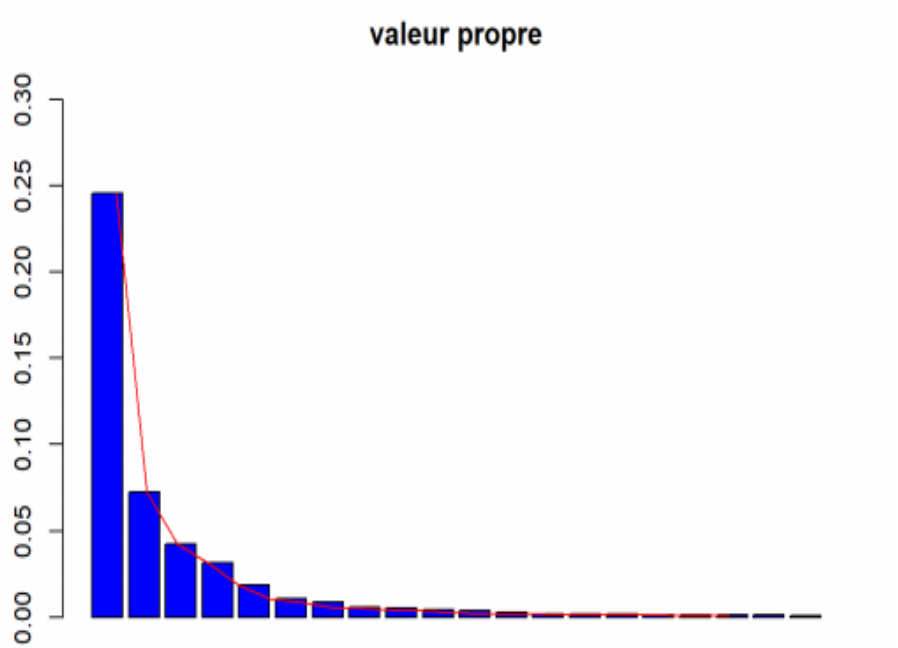

figure1 :Courbe des valeurs propres

On coupe l'éboulis des valeurs propres à l'endroit où celui-ci posséde un "coude". On observe un décrochement au 8éme axe, puis décroissance régulière à partir de cet axe. Donc, seuls les 7 premiers axes présentent un éventuel intérêt. Ce qui nous permet de conclure d'aprés ce critére que le plus pertinent est de garder les 7 premiers axes.

Pour mieux comprendre la significativité de ces axes, il est important d'identifier quelles sont les modalités qui contribuent le plus à chaque axe. Seules les modalités dont la contribution est élevée sont à considérer pour l'interprétation d'un axe c'est-à-dire celles dont la contribution est supérieure à (1/p=1/11=0,09).
On observe les modalités les plus représentatives sur chacun des axes :

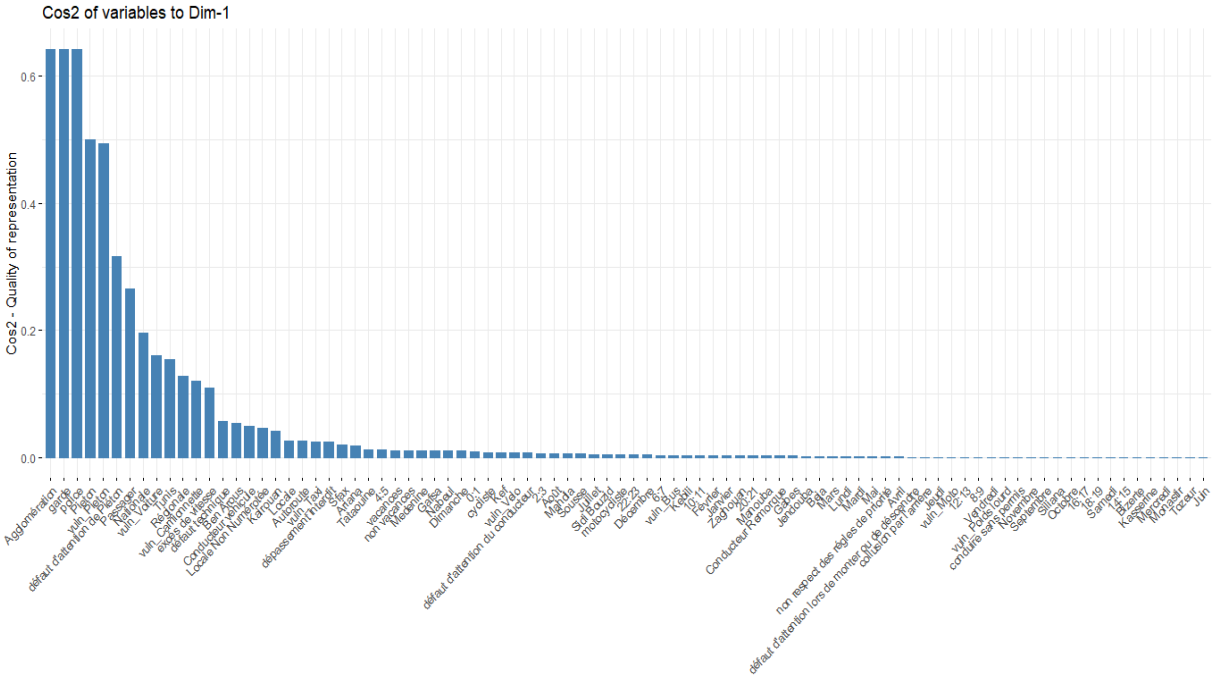

figure1 :Cos2 des modalités sur l'axe1

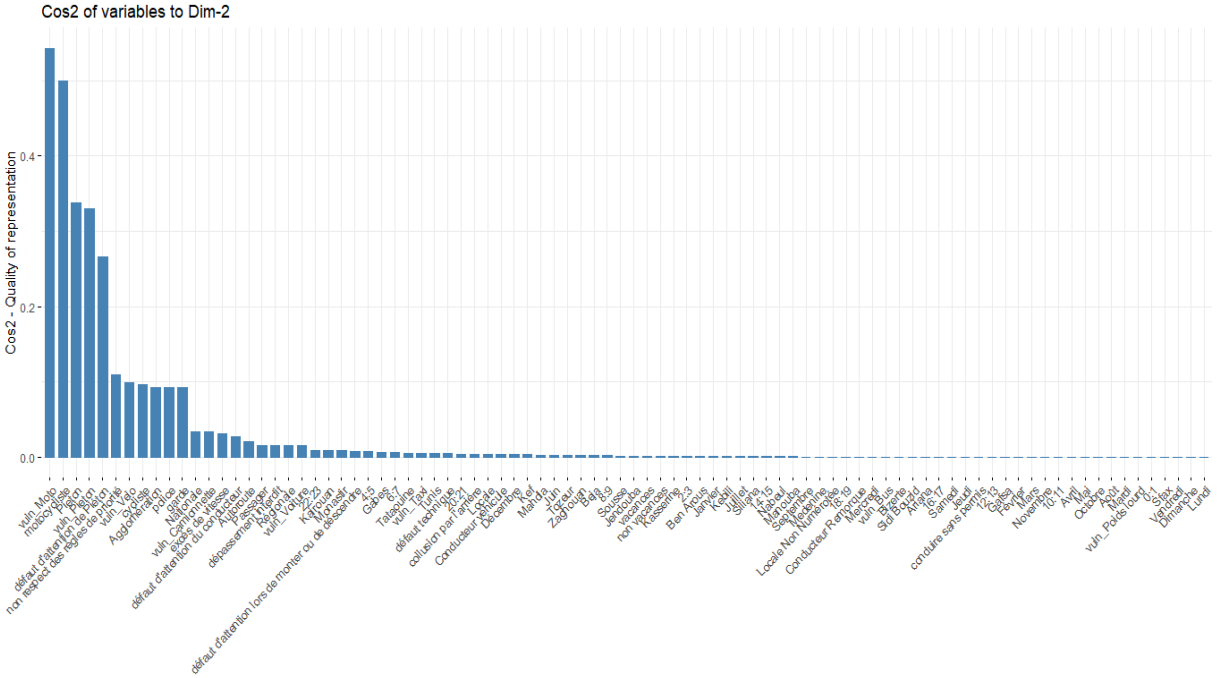

figure1 :Cos2 des modalités sur l'axe 2

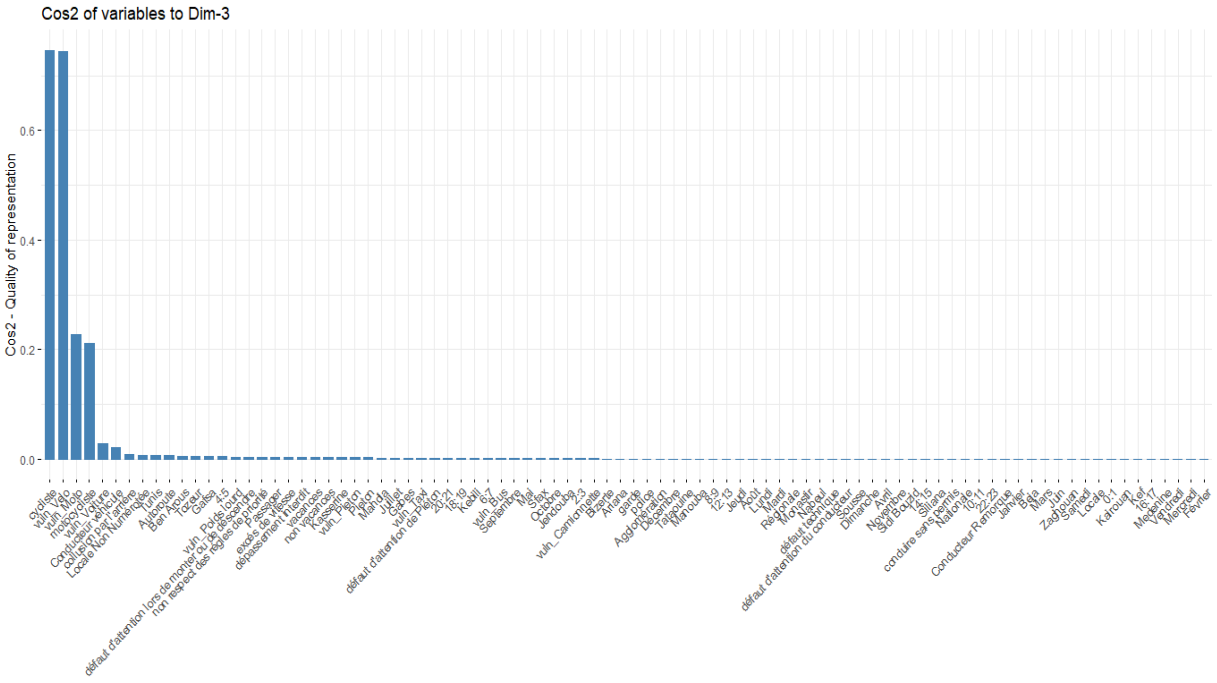

figure1 :Cos2 des modalités sur l'axe 3

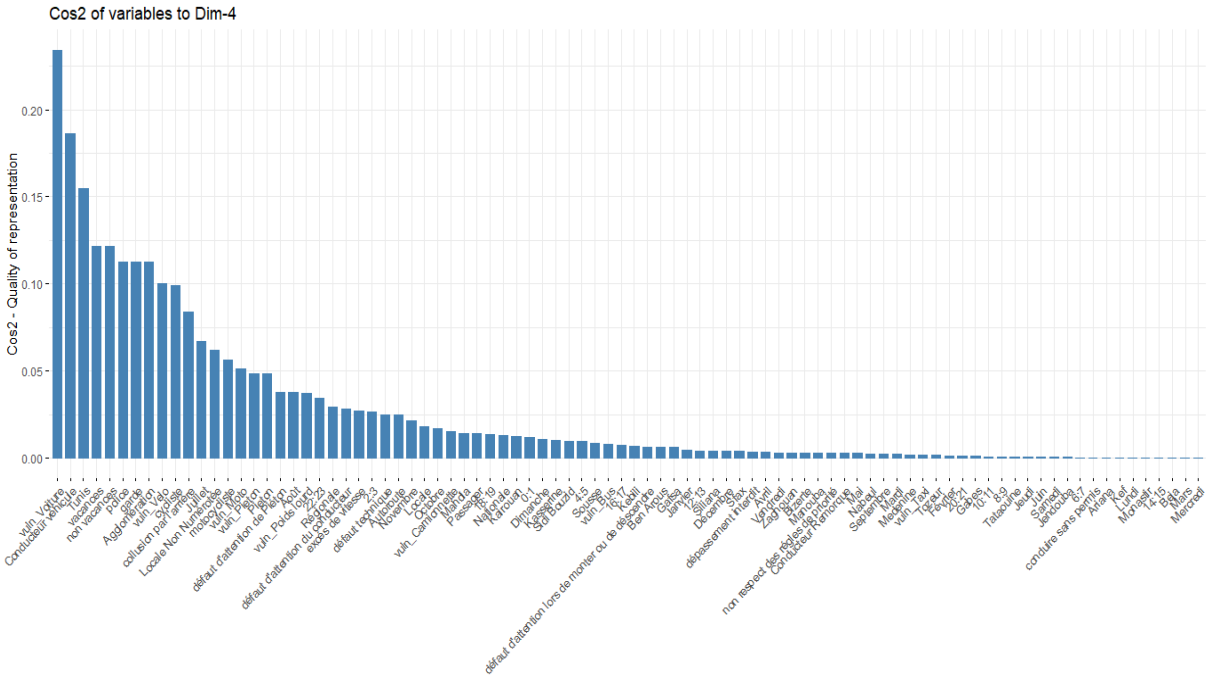

figure1 :Cos2 des modalités sur l'axe 4

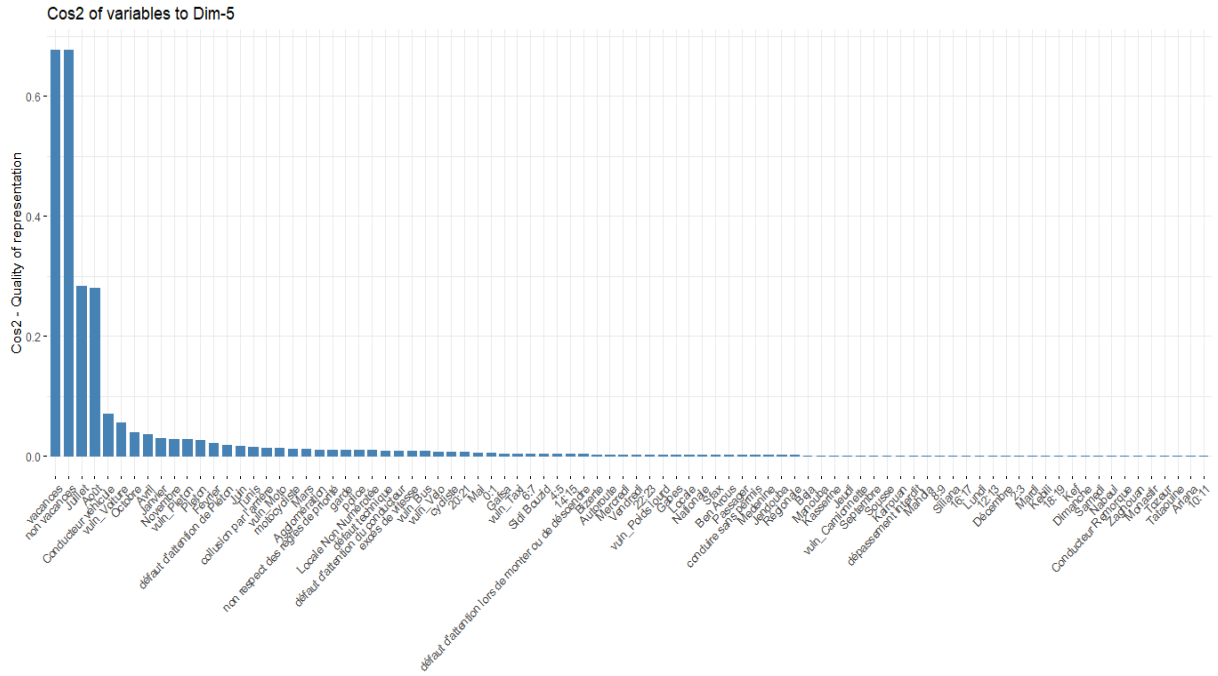

figure1 :Cos2 des modalités sur l'axe 5

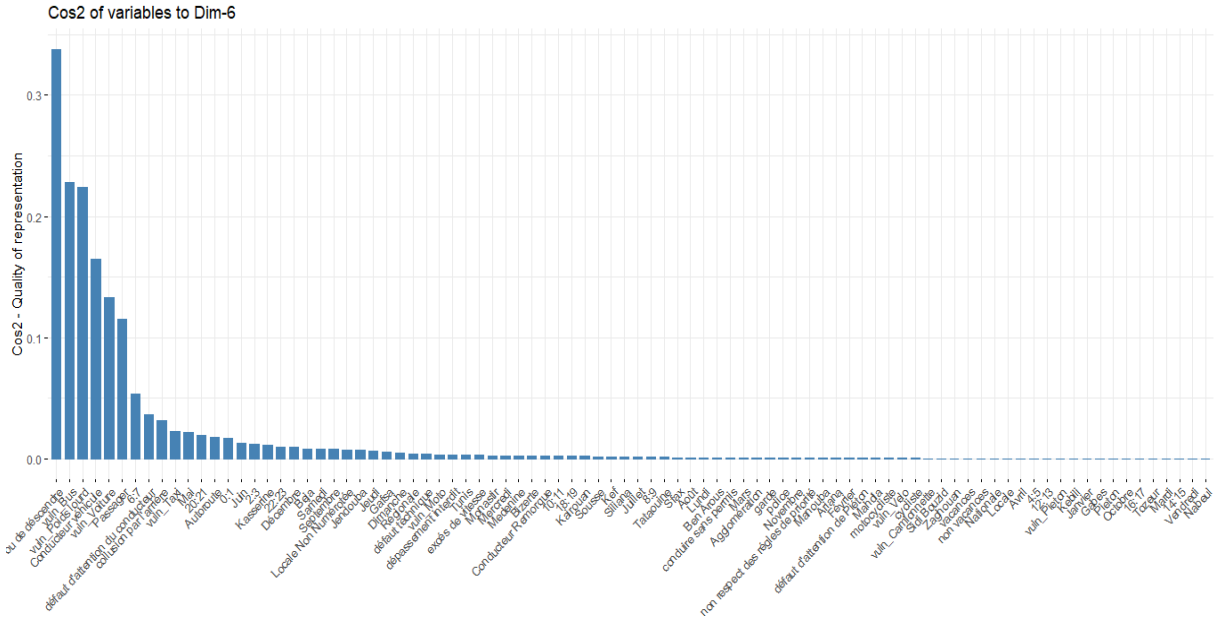

figure1 :Cos2 des modalités sur l'axe 6

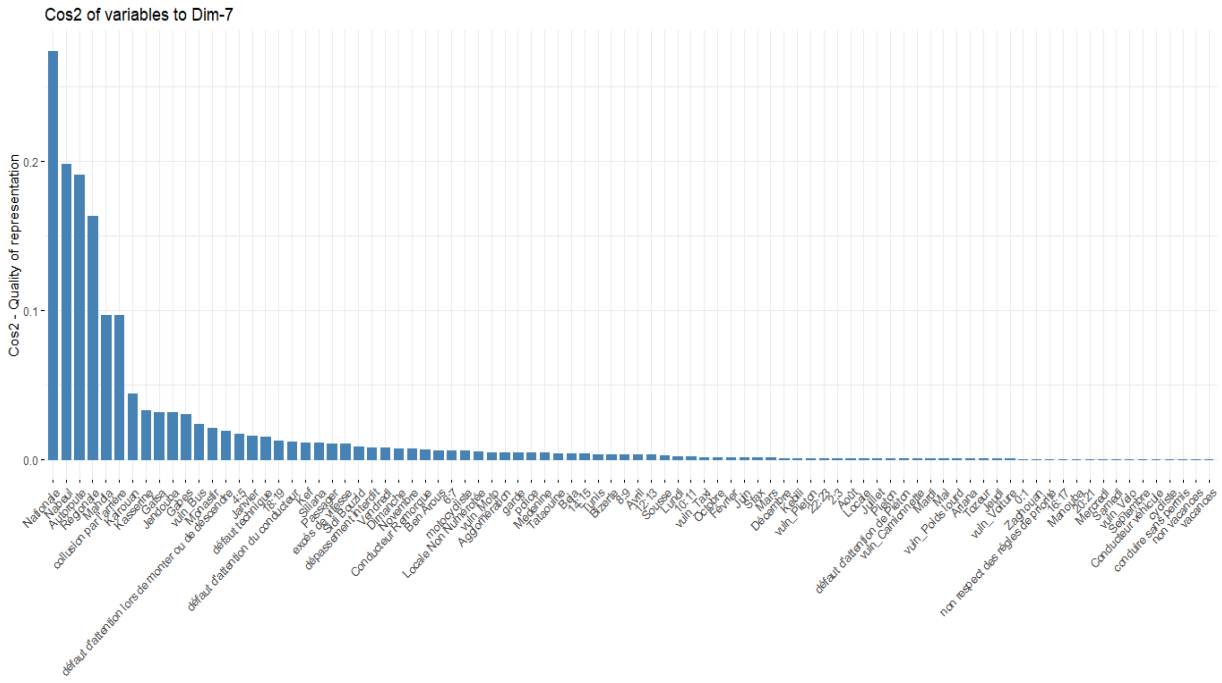

figure1 :Cos2 des modalités sur l'axe 7

Les modalités qui ont les meilleurs qualités de représentation selon chacun des deux axes factorielles cos2 vont etre retenues, en voyant les résultats précédent ,nous pouvons décider de retenir les modalités jusqu'à un seuil de 0,09 $\simeq$ 0,1.

Donc les axes sont représentés de la maniére suivante :
Axe 1 :Agglomération,Garde,Police,Piéton,vuln_Pieton,Défaut d'attention de piéton,Passager,Nationale,
Vuln_Voiture,Tunis,Régionale,Vuln_Camionnette,excés de vitesse.

Axe 2 :Vuln_Moto,Motocycliste,Piéton,Vuln_Piéton,Défaut d'attention de piéton, non respect des régles de priorité.

Axe 3 :Cycliste,Vuln_Velo,Vuln_Moto,Motocycliste.

Axe 4 :Vuln_Voiture,Conducteur véhicule,Tunis,Police,Garde,Agglomération.

Axe 5 :Non vacances,Vacances,Aout,Juillet.

Axe 6 :Défaut de monter ou de déscendre, Vuln_Bus,Vuln_Poids lourd, Conducteur_véhicule, Vuln_Voiture,Passager.

Axe 7 :Nationale,Nabeul,Autouroute,Régionale.

Interprétation des résultats graphiques de l'ACM : Le graphique par défaut de l'ACM (analyse des correspondances multiples) est un graphique "symétrique", dans lequel les lignes et les colonnes sont représentées en coordonnées principales. Nous allons interpréter les nuages des points projetés sur chacun des 7 axes.

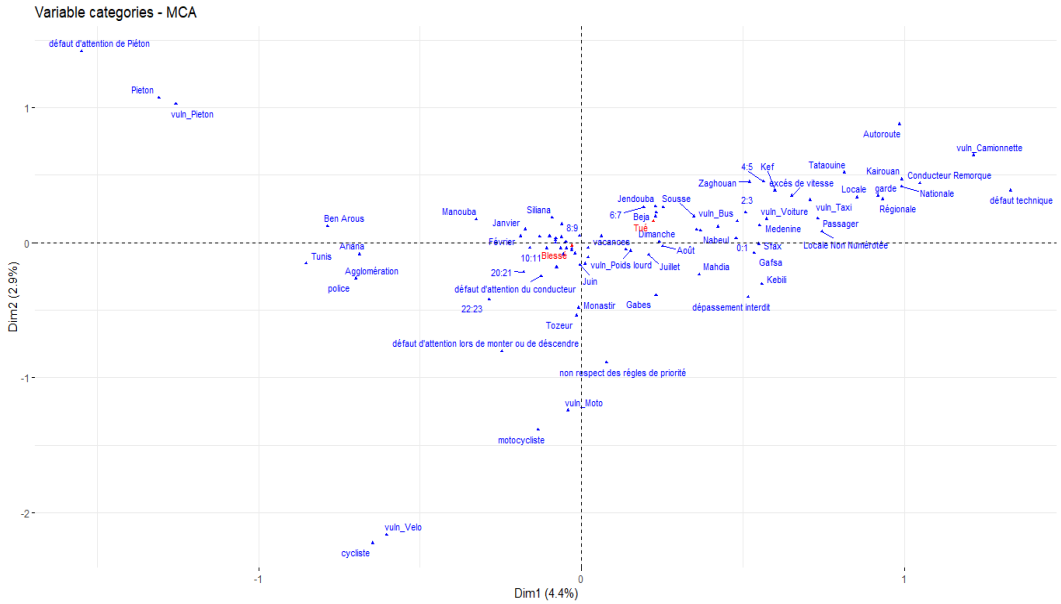

figure1 :Projection sur le plan factoriel (Axe1,Axe2)

A partir du graphe,on voit tout d'abord une opposition entre les zones urbaines et les zones rurales représentés réspectivement par les modalités police et garde de la variable source.

On constate aussi à droite qu graphe qu'il y a une forte corrélation entre les heures de nuit (entre minuit et 7h du matin) et les zones touristiques tels que Nabeul,Sousse et Mahdia.

Ce qui est expliqué par la présence des boites de nuits dans ces zones et donc on détecte plus d'accidents la nuit là bas.

De plus,le premier axe nous permet de constater que :
Les accidents mortels sont plus dans les zones rurales et les zones touristiques c'est à dire loin des agglomérations.
Aussi, On trouve que les accidents ont tendance à être plus graves aux vacances et pendant les week-ends (c'est à dire que la modalité Tué est proche des modalités vacances, Samedi et Dimanche).

Pour une interprétation meilleure, on a décidé de visualiser sur chaque axe seulement les modalités qui ont les meilleures qualités de présentation c'est à dire qui possédent une cos2 $\geq$ 0,1.

On remarque pour le premier axe une opposition entre les accidents dont les victimes sont des piétons à gauche (piéton, vuln_Pieton ,défaut d'attention de piéton) et les accidents des camionnettes à droite (vuln_Camionnette , Nationale, ,Passager).
On observe aussi une forte opposition entre les agglomérations et les routes nationales.

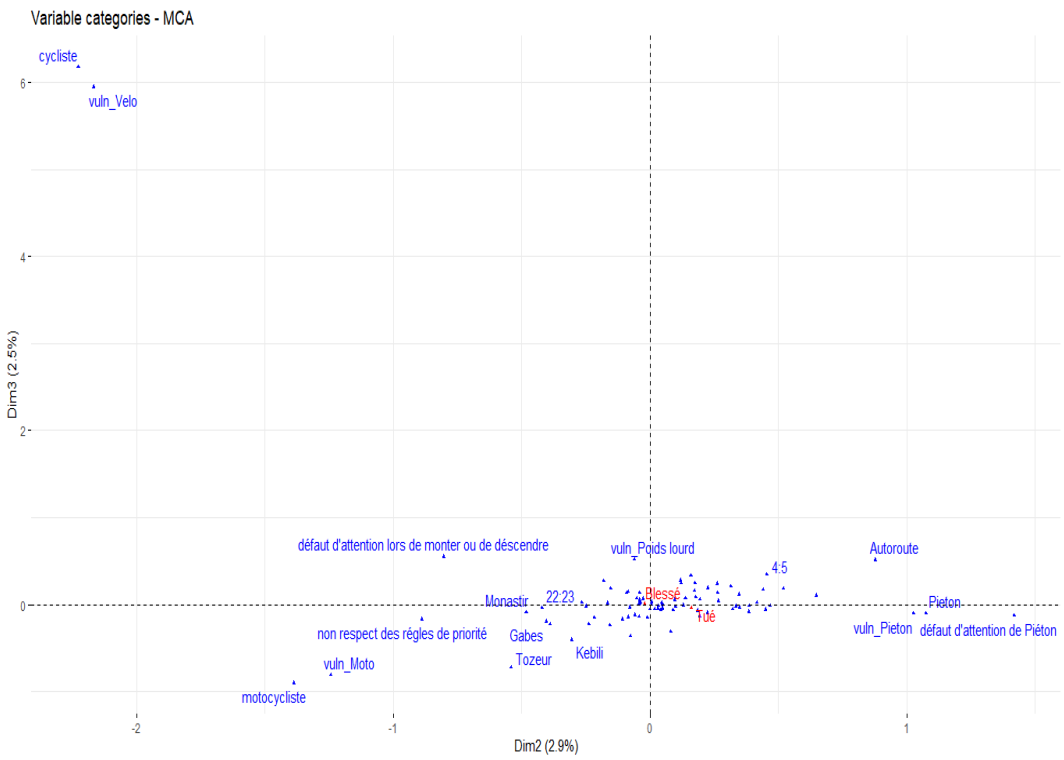

figure1 :Projection des modalités les plus contributives sur le plan factoriel (Axe2,Axe3)

Le deuxiéme axe oppose les accidentés piétons aux accidentés motocyclistes. En effet, on trouve les modalités (“défaut d’attention de piéton”, piéton, vuln_Piéton) à droite et (Vuln_Velo, Cycliste) symétriques à gauche.
On note aussi que les accidents des motos sont corrélées avec la cause "non respect des régles de priorité" ainsi que les heures 22h et 23h.
Ce qui nous permet de déduire que cette cause est fréquente pour ces types d’accidents et que ses accidents se déroulent le plus entre 22h et 23h.
De plus, on voit que les accidentés piétons sont plus soumises aux risques de mortalité car elles sont proches de la modalité "Tué" .

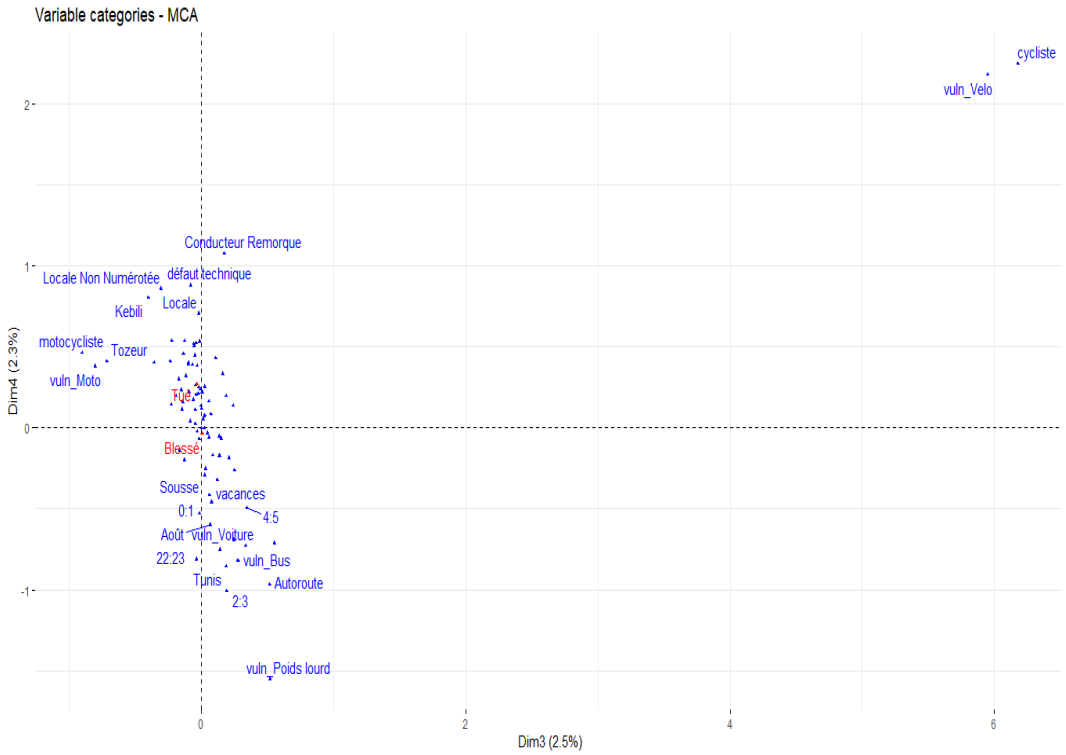

figure1 :Projection des modalités les plus contributives sur le plan factoriel (Axe3,Axe4)

Le troisiéme axe oppose les accidentés piétons au accidentés motocyclistes.
Par ailleurs, on détecte une corrélation entre les accidents des motos et les routes locales et locales non numérotés et précisement à Kebili et Tozeur.
Sinon, cet axe n’apporte pas vraiment beacoup d’informations car la pluspart des modalités des différentes variables sont mal projetées sur lui.
Néanmoins, Nous pouvons constater que les accidents des voitures qui se déroulent

pendant les vacances et plus précisément au mois d'août sont surtout aux autoroutes de Tunis et la nuit c'est à dire pour les heures entre 22h et 5h.

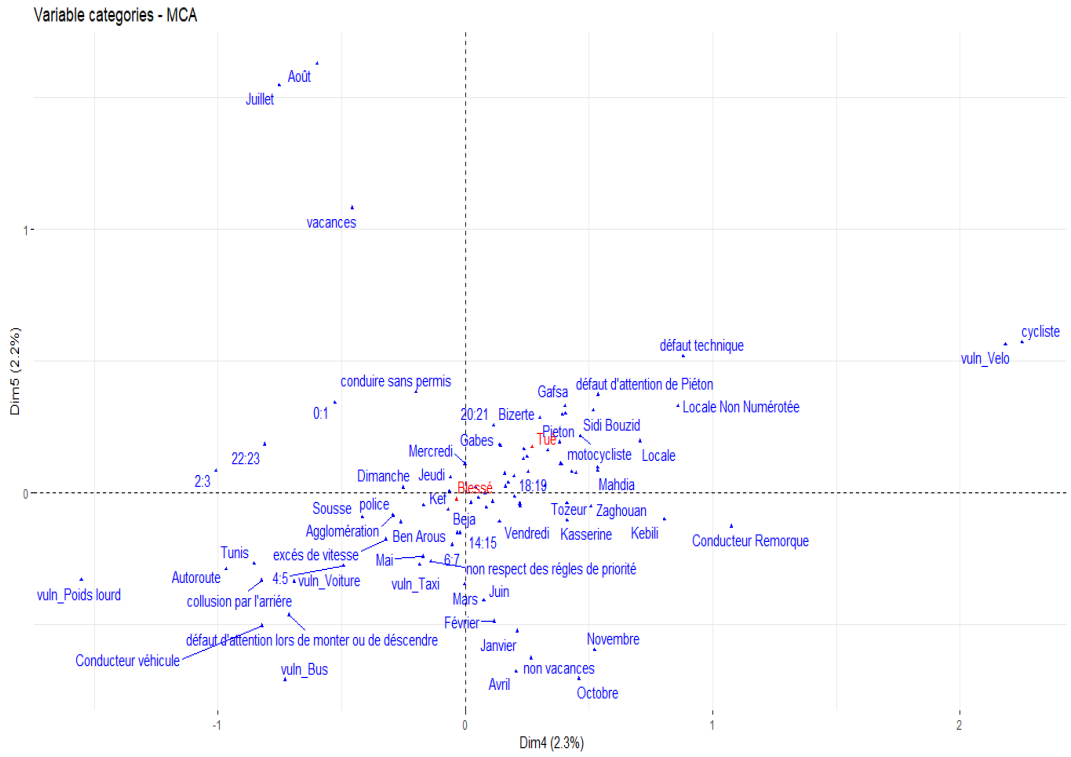

figure1 :Projection des modalités les plus contributives sur le plan factoriel (Axe4,Axe5)

En regardant bien le graphique cidessus, on peut remarquer que le quatriéme axe projette bien les véhicules impliqués dans les accidents.
En effet : -Les Bus sont bien corrélés avec la cause "défaut d'attention lors de monter ou de déscendre".
Ce qui nous permet de déduire que l'inattention quand un passager d'un bus vient de monter ou de déscendre de ce moyen de transport public est trés dangereux et présente la cause la plus fréquente pour les accidents de Bus.
-Les voitures sont trés corrélées avec la cause collision par l'arriére surtout dans les agglomération de Tunis et Ben Arous. Ce qui revient au fait que les agglomération sont embouteillées ce qui entraine les Choc arrière entre les véhicules.
-Pour les accidents des taxis les causes les plus fréquentes sont :"non respect des régles de priorités" et "excès de vitesse". C'est à dire les attitudes le plus souvent suicidaires de la part des conducteurs des Taxis.

Par ailleurs, les accidents mentionnées ci-dessus sont le plus souvent aux agglomérations et surtout à Tunis et Ben Arous.

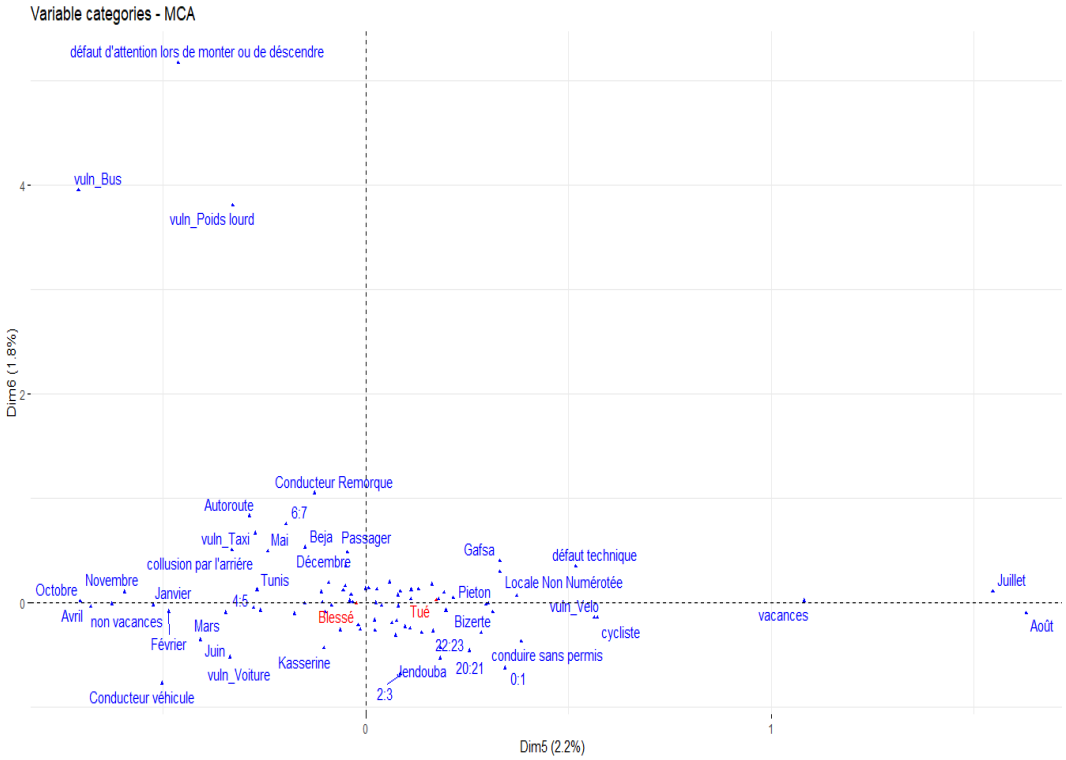

figure1 :Projection des modalités les plus contributives sur le plan factoriel (Axe5,Axe6)

Le 5éme axe oppose les accidents qui ont eu lieu pendants les vacances et ceux hors les vacances.

Les accidentés de la période des vacances sont surtout des piétons et des cyclistes alors que les accidents hors les vacances sont des accidents de voitures et Taxis.

Ce qui revient au fait qu'aux vacances les gens se déplacent à pieds et aux vélos pour se promener plus qu'au cours de l'année où chacun s'occupe de ses affaires et il y a moins de déplacement.

Le 5éme axe montre aussi que les accidentés piétons et cyclistes sont plus soumis au risque de mort (c'est à dire plus grave) car ils sont plus proches de la modalité Tué.

À partir du même graphique on constate que :

L'axe 6 oppose les accidents des poids lourd et Bus causés par un défaut d'attention lors de monter ou de descendre et les accidents des voitures.

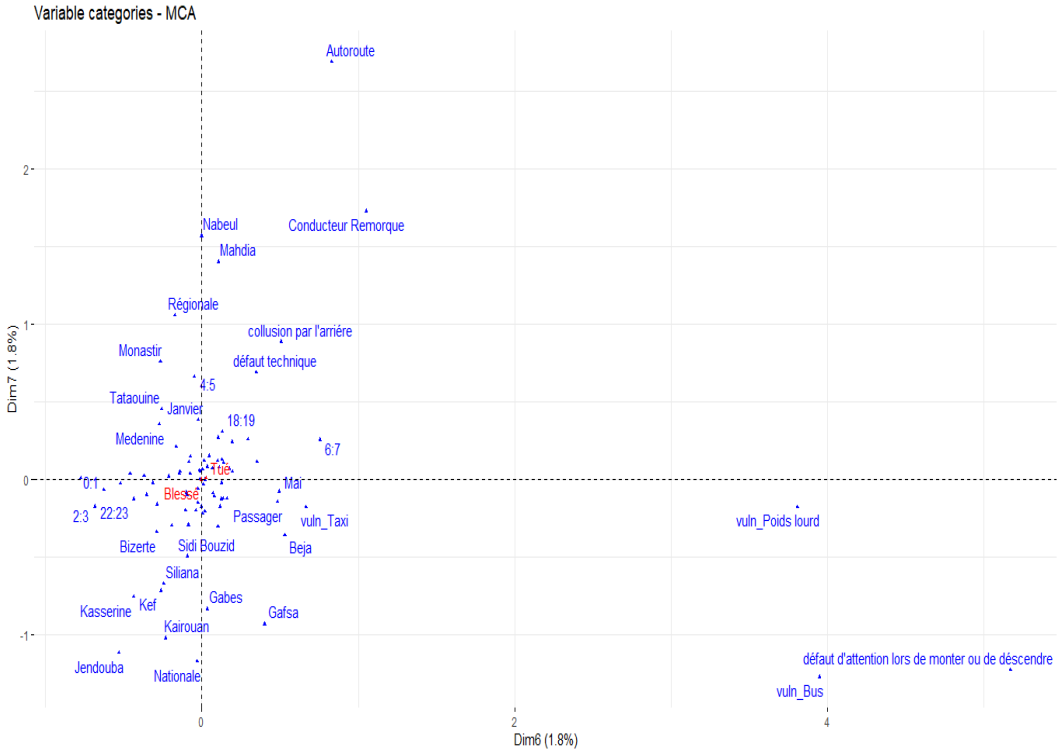

figure1 :Projection des modalités les plus contributives sur le plan factoriel (Axe6,Axe7)

Le 7éme axe oppose les accidents des autoroutes de Nabeul et Mahdia aux Pour le 7éme axe, on observe une opposition entre les accidents ayant lieu aux autoroute et aux routes régionales précisement à Nabeul et Mahdia et les accidents des routes nationales (à Bizerte, Gafsa,Sidi Bouzid, Siliana, Kef, Kasserine, Kairouan, Gabés et Béja).
En plus, on voit une forte corrélation entre les accidents des poids lours et la cause “défaut d’attention lors de monter ou de descendre”.

3.2 Classification hiérarchique ascendante

La classification hiérarchique ascendante (CAH) est une technique statistique d’analyse largement utilisé qui vise à partitionner une population en différentes classes ou sous-groupes. L’idée est de construire un dendrogramme ou arbre de classification des données qui successivement fusionne des groupes d’individus similaires. La visualisation de cet arbre fournit un résumé utile des données.

3.2.1 Fondement théorique

La CAH consiste à regrouper progressivement les individus les plus semblables dans une meme classe (homogénéité intra-classe) et les classes construites par cette méthode sont les plus dissemblables possibles (homogénéité inter-classe).
L'objectif de la CAH est de produire une arbo-rescence qui met en évidence les liens hiérarchiques entre individus caractérisés par un certain nombre de variables qui sont divisées en modalités. L'arbo-rescence permet en fait de detecter le nombre de classes au sein de la population.

3.2.2 Principe de la classification

Pour réaliser une classification, il faut définir une mesure de ressemblance entre individus. Cette mesure s'exprimera sous la forme d'une matrice de distances qui exprime la distance existante entre chaque individu pris deux à deux. Plus les deux observations seront dissemblabes , plus la distance sera importante.
Cette méthode se construit étape par étape en partant du niveau le plus fin où chaque individu est seul dans son groupe jusqu'au niveau le plus agrégé où tous les individus sont dans le même groupe.
En fait, l'aggrégation commence par les individus qui ont plus de similarités entre elles, ensuite les observations un peu moins similaires jusqu'au regroupement trivial de l'ensemble des individus à classer. La classification est dite ascendante car elle part des observations individuelles. elle est dite aussi hiérarchique car elle produit des classes de plus en plus vastes(larges), incluant des sous-groupes en leur sein. En découpant cet arbre à une certaine hauteur choisie, on produira la partition désirée.

3.2.3 Algorithme

Entrée : tableau de données : les axes d'ACM dans notre étude.
Sortie : Indicateur de partition des individus
Etape 0 : Initialisation

On considére que chaque élément est isolé et seul dans son groupe : n élément $\Longleftrightarrow$ n groupes.
Etape 1 : Dans cette étape, on calcule la matrice M_d^n des distances d entre éléments.(Plusieurs distances peuvent etre utilisées comme la distance Euclidienne, de Manhattan...). On fusionne les 2 groupes les plus proches et on les relie dans le dendogramme.

Le trait permettant de relier les deux groupes dans le dendogramme est d'autant plus long que la distance entre les groupes est élevée.

Répéter : -Détecter les deux groupes les plus proches.
-Aggréger ces groupes pour n'en former qu'un seul.

Jusqu'il ne réste plus qu'un seul et unique groupe réunissant tous les individus.

On identifie par la suite le nombre adéquat de groupes et on procéde au partitionnement.

3.2.4 Réflexion pré-algorithme

-Il est nécessaire de définir une distance entre les individus. Ce choix dépend des données étudiées et des objectifs.
-Dans notre étude on a choisit la distance Euclidienne qui est le type de distance le plus couramment utilisé. Il s'agit d'une distance géométrique dans un espace multidimensionnel :

$$\text{distance(x,y)}=(\sum_i (x_i - y_i)^2)^(1/2)$$

-Par la suite, on choisit l'indice d'agrégation : Le regroupement des éléments se fait en minimisant l'indice d'agrégation. Il existe encore différent méthodes, mais la méthode la plus connu est la méthode de ward qu'on a utilisé. L'objectif de cette méthode est de ganier le minimum d'inertie intra-classe à chaque agrégation et de perdre d'inertie interclasse due à cette agrégation. Le calcul de l'inertie pour cette méthode se fait par la formule suivante :

$$\Delta I_{min} = \left(\frac{m_h m_{h'}}{m_h + m_{h'}}\right) d^2(g_h, g'_h)$$

avec m_h et $m_{h'}$ sont les masses respectives des classes h et h' qui possédent les centres de gravité respectifs g_h et $g_{h'}$ et $d(g_h, g'_h)$ est la distance euclidienne entre les deux centres g_h et $g_{h'}$.

3.2.5 application

Nous avons effectué la CAH sur les 7 axes d'ACM et nous avnos obtenue le dendogramme suivant :

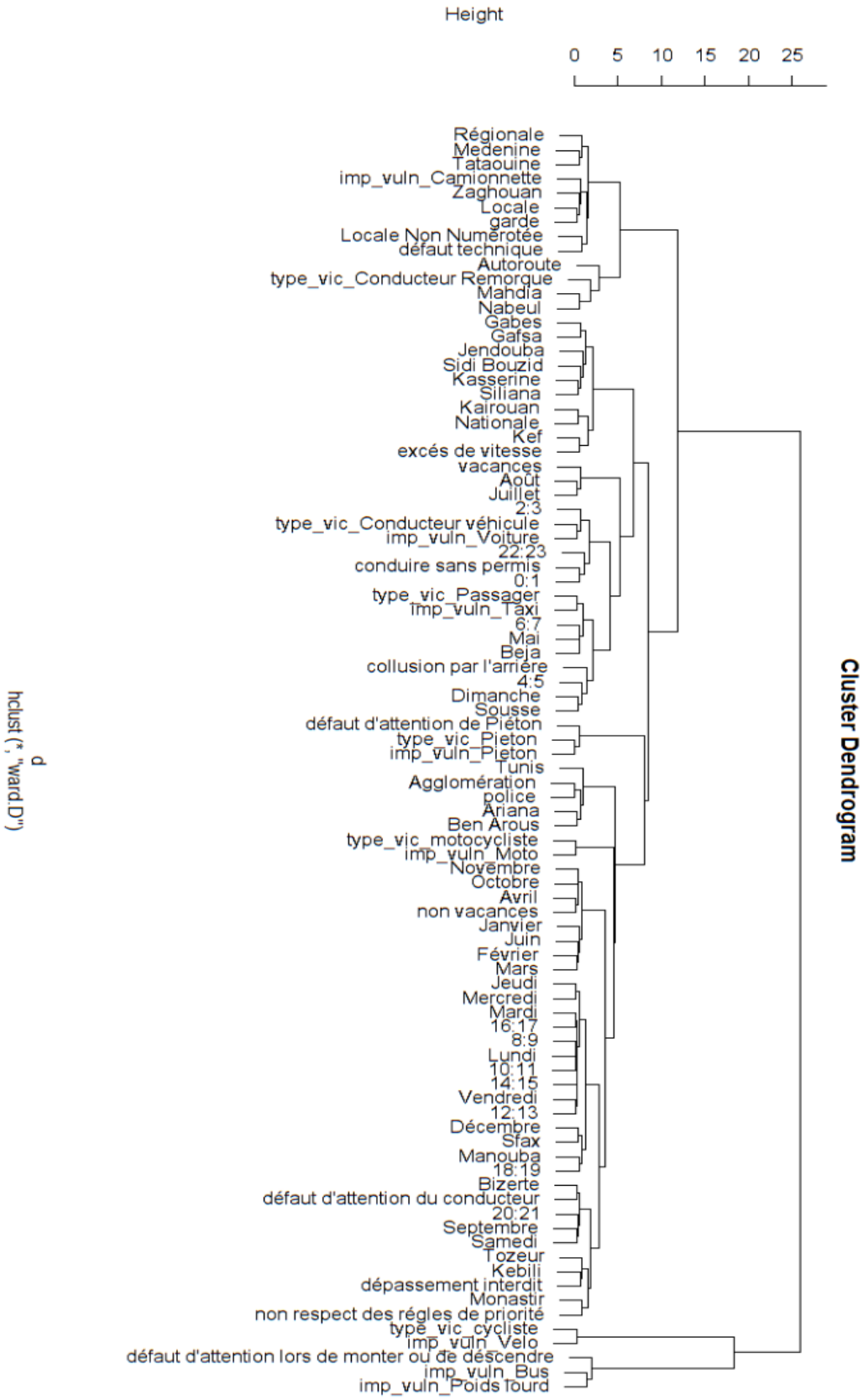

figure1 :Dendogramme de la CAH

C'est le résultat de la classification hiérarchique ascendante menée sur les premiers 7 axes principaux issus de l'ACM et principalement sur les coordonnées des modalités sur ces axes qui présentent 43,2% de l'inertie.

Choix du nombre d'axes

Nous déciderons de retenir la partition qui semble la meilleure, généralement. Pour obtenir cette partition, nous pouvons représenter les sauts d'inertie du dendrogramme selon le nombre de classes retenues.

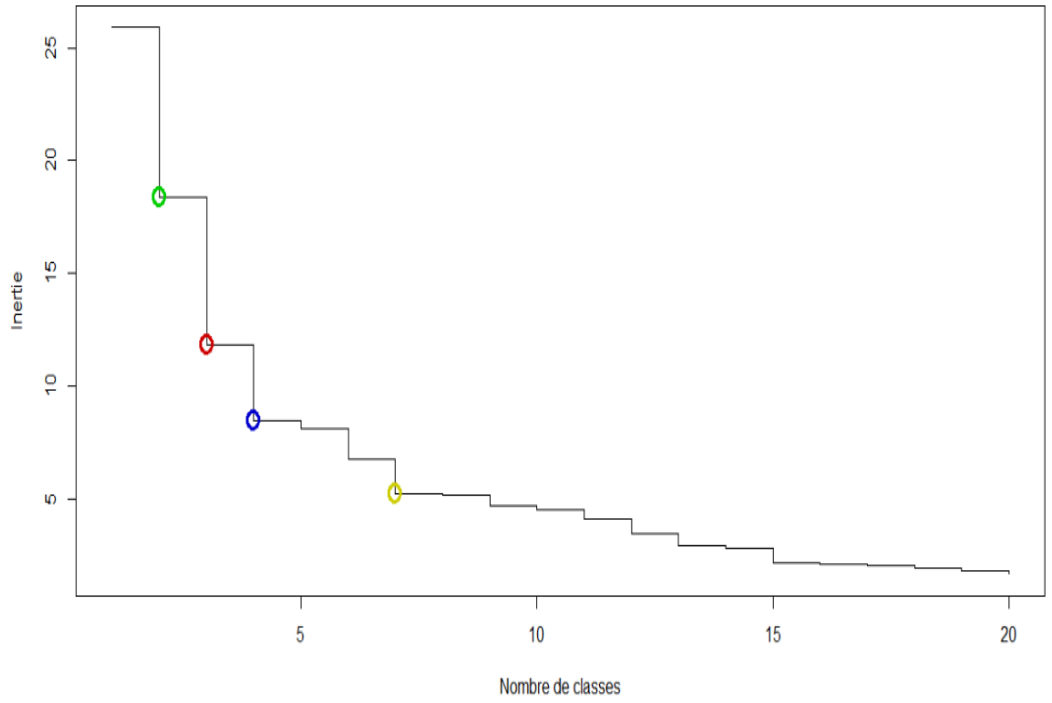

figure1 :nombre de classes selon l'inertie

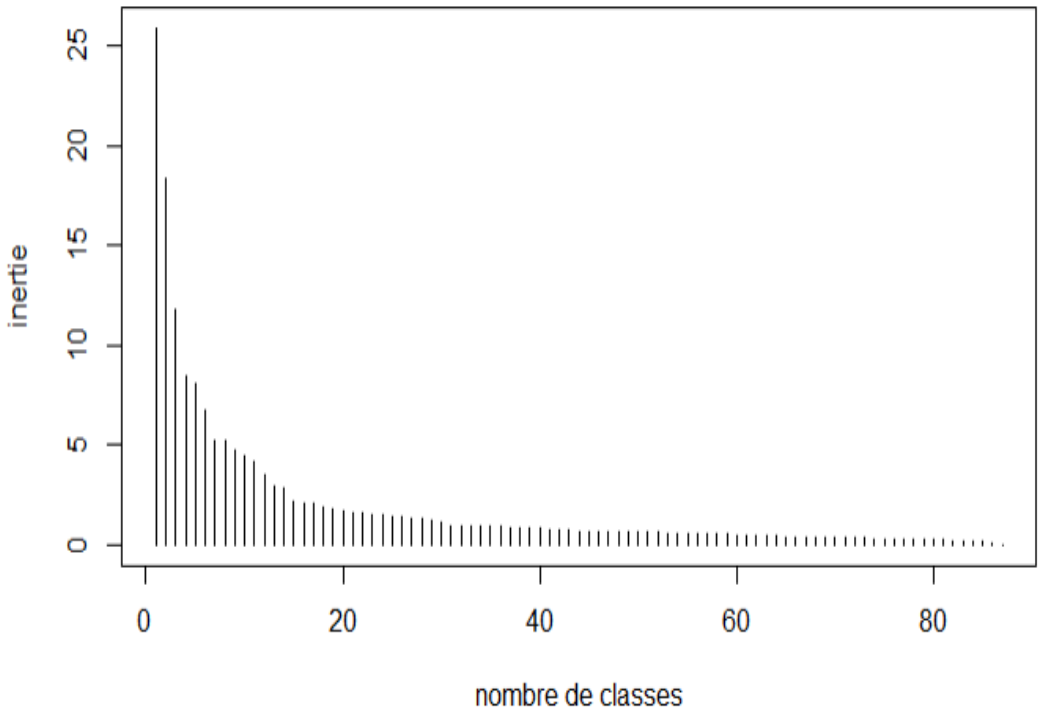

figure1 :Inertie du dendogramme selon le nombre de classes

D'après le premier graphe, On voit quatre sauts assez nets à 2, 3, 4 et 7 classes. Et on remarque bien dans le deuxiéme graphe qu'on a une concentration de l'inertie dans les 7 premiers axes donc on peut procéder à une coupure en 7 classes.On constate aussi que la perte d'inertie en passant de 7 à 8 classes est très faible. Alors une coupure en 7 classes semble judicieuse.

Pour bien aboutir à la meilleur partition, on commance par interpréter la division en 2 classes et on ajoute les classes une par une en interpretant à chaque fois les classes qui se détachent jusqu'à arriver à la répartition en 7 classes.

Partition en 2 classes :

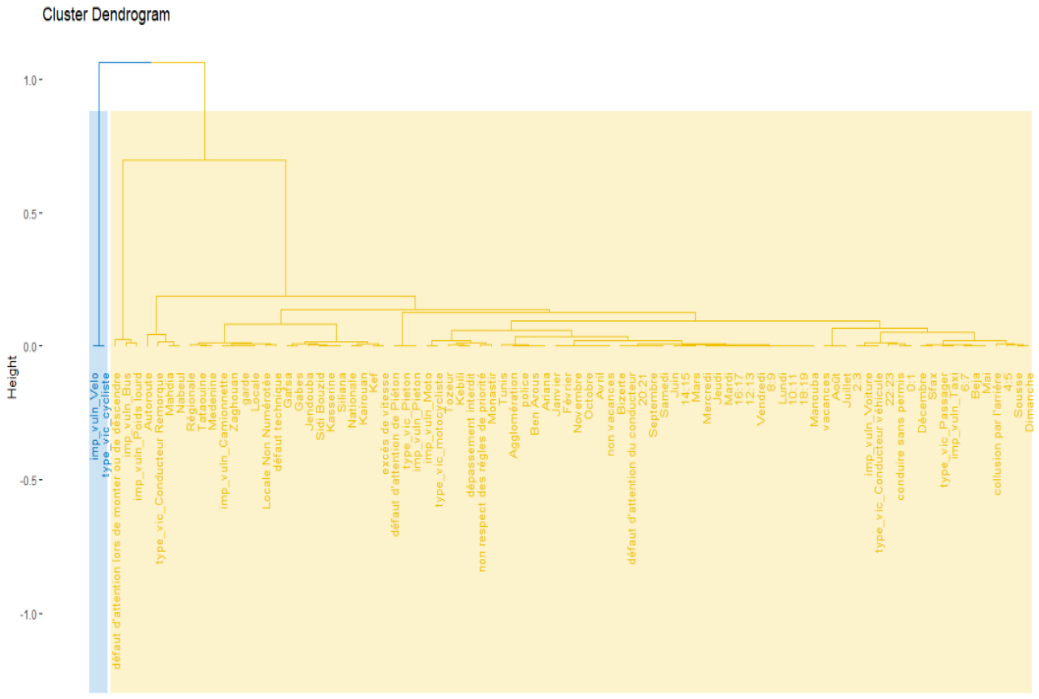

figure1 :répartition en 2 classes

La 1ére classe : La premiére partition fait apparaitre le détachement de la classe des cyclistes par rapport aux autres types d'accidents.
Ce qui prouve que la classe qui présente les cyclistes est trés dissemblables par rapport aux autres caractéristiques des accidents.

Partition en 3 classes :

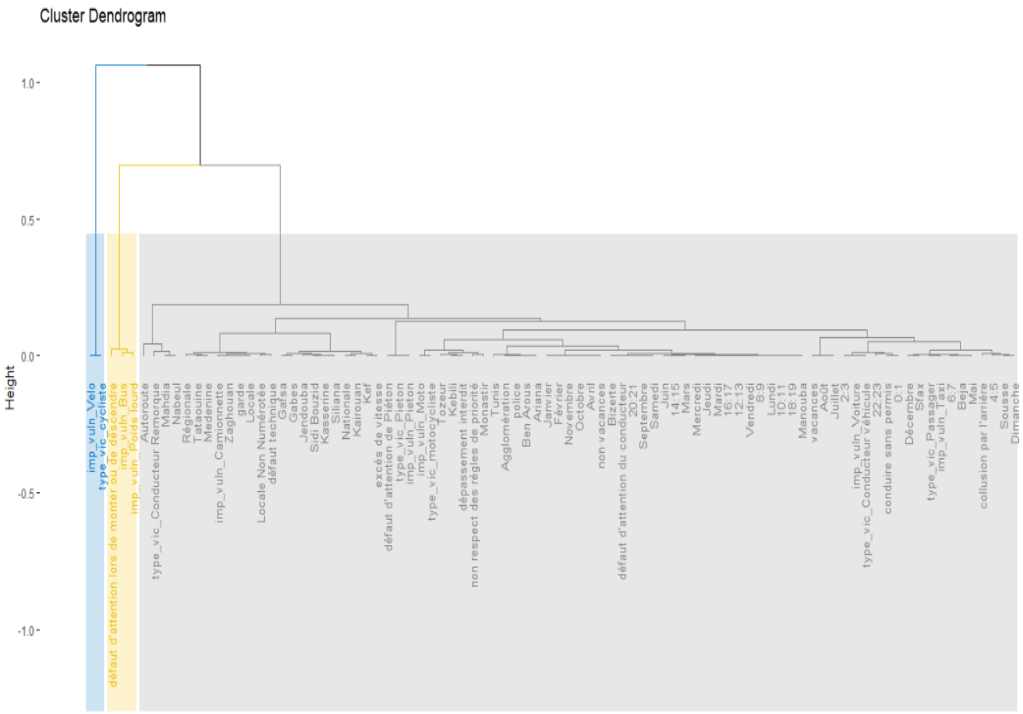

figure1 :répartition en 3 classes

La 2éme classe : On voit aprés cette partition l'émergence de la classe dans laquelle les poids lourd et des Bus sont concernés. La cause la plus fréquente est le défaut d'attention lors de monter ou de déscendre.

Partition en 4 classes :

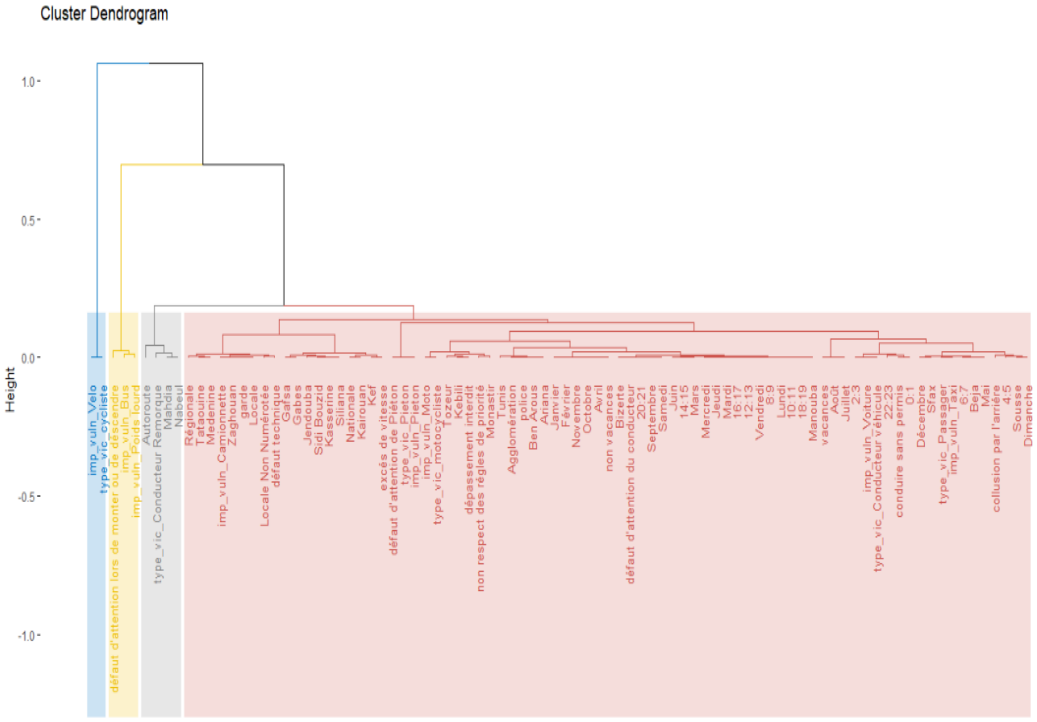

figure1 :répartition en 4 classess

La 3éme classe qui apparait est la classe des accidents des autoroutes de nabeul et Mahdia. Cela peut être expliqué par la violence des chocs due aux excès de vitesse sur les autoroutes. Le district de Nabeul est l'une des zones les plus touchées de la Tunisie surtout que l'autoroute A1 assure le déplassement aux différentes régions comme Sousse, Mahdia, Monastir, Sfax, Gabes, Médenin, Tataouine ...

Partition en 5 classes :

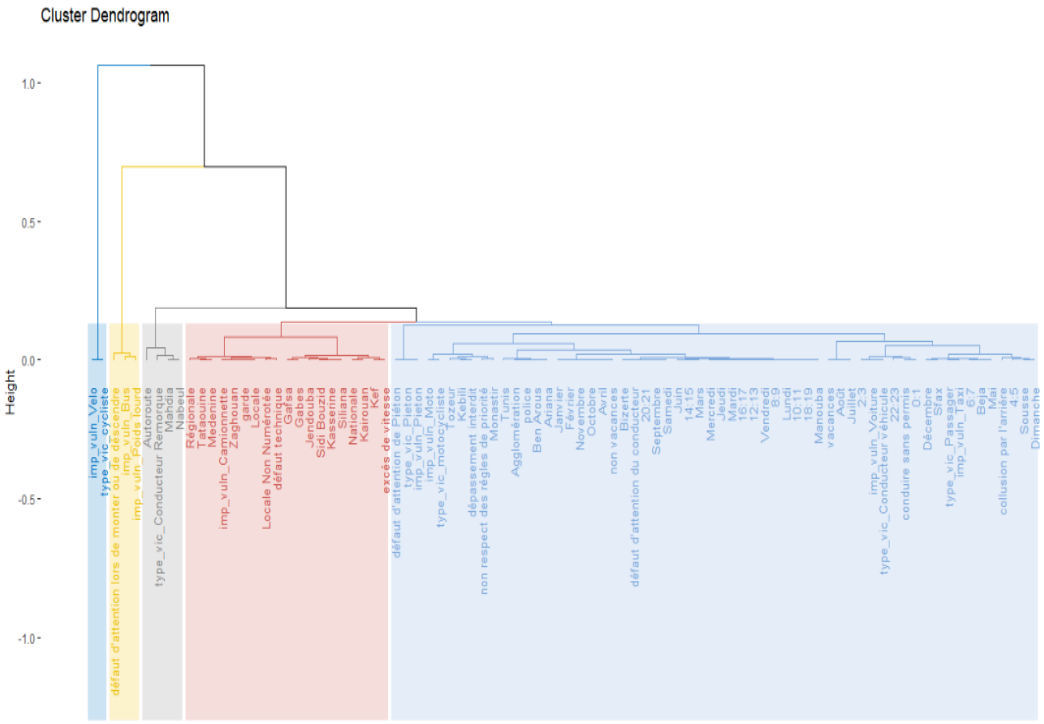

figure1 :répartition en 5 classes

La 4éme classe : La partition suivante donne naissance à une nouvelle classe qui caractérise les accidents qui sont produits aux zones rurales à l'éxtérieur des villes. Ceci est défini par la présence des routes nationales, régionales, locales et locales non numérotés et plus précisément aux gouvernorats du sud-Est et les zones frontiéres entre la Tunisie et la Libye (Médenine,Tataouine) et la Tunisie et l'Algérie (Kef, Jendouba, Gafsa).
Ceci peut-être expliqué par l'activité Commerciale et économique faite surtout par les camionnettes. Nous constatons aussi que ces accidents sont causés par l'éxcés de vitesse et les pannes téchniques.

Partition en 6 classes :

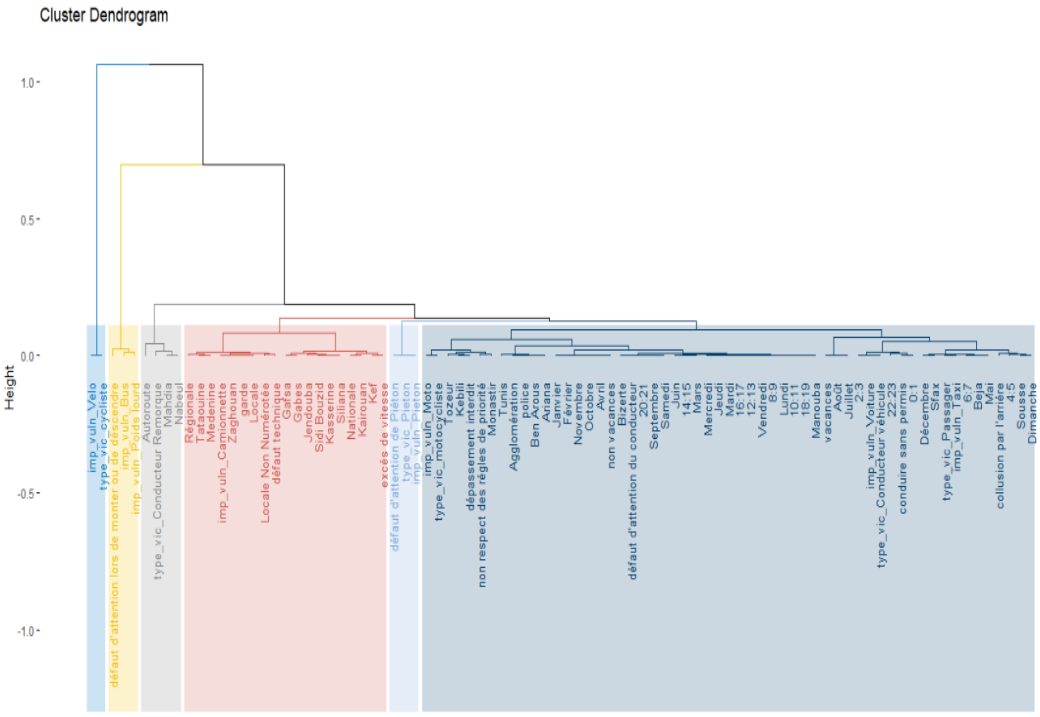

figure1 :répartition en 6 classes

La 5ème classe : englobe les accidentés qui sont les piétons. La cause de ces accidents est le défaut d'attention de piéton.

Partition en 7 classes :

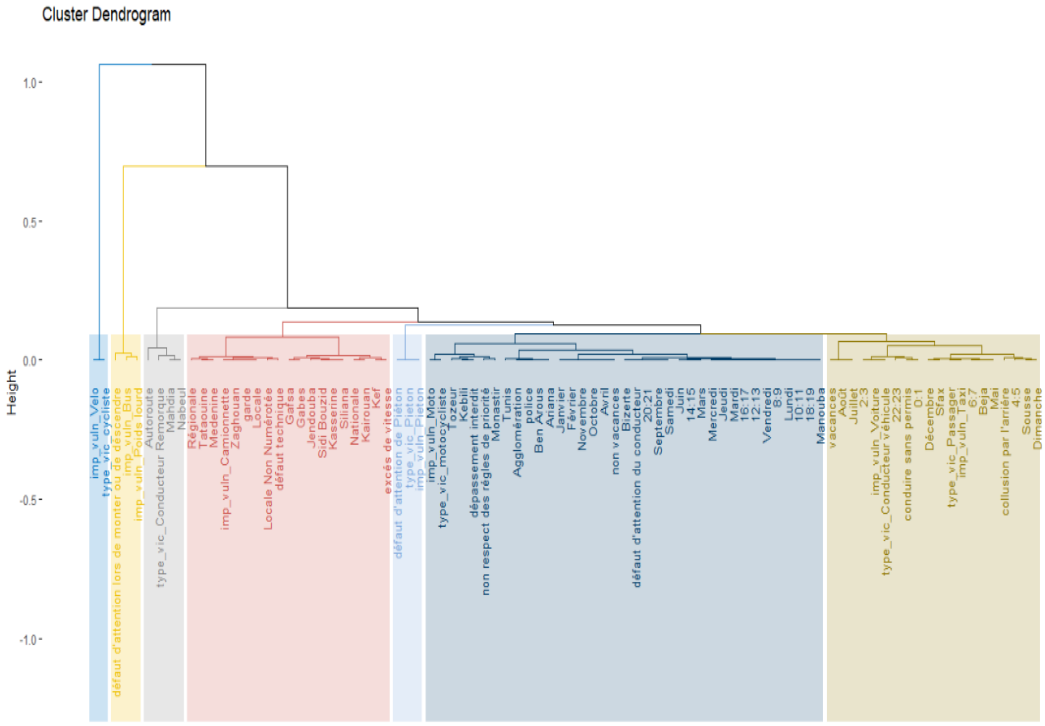

figure1 :répartition en 7 classes

La 6éme classe est la classe des accidents entre où les victimes sont des cyclistes, ces accidents ont eu lieu au cours de l'année dans les jours non-vacances pendant les heures de sorties.

Le fait que ces accidents sont dans les zones urbaines et principalement les agglomérations peut être un facteur trés important sur cette classe puisqu'en tunisie les routes ne sont clairement pas faites pour les cyclistes et surtout aux agglomérations où les routes sont parfaitement occupés par les voitures, les camionnettes et les poids lourd en général.

Ces accidents sont causés par les conducteurs que ce soit des vélos ou des autres véhicules appartenants à l'accident(défaut d'attention de conducteur,dépassement interdit et non respect des régles de priorité).

La 7éme classe regroupe les accidents des voitures ayant lieu aux vacances (les vacances d'été : Juillet,Aout et les vacances d'hiver : Décembre et en Mai pendant les vacances de l'Aiid). Ces accidents sont précisement le Dimanche, la nuit aux gouvernorats de Sfax,Béja et Sousse. La cause la plus fréquente pour ces accidents sont : la collusion par l'arriére et la conduire sans permis.

Conclusion

Tout au long de cette partie, notre objectif était de former des groupes de variables pour structurer notre ensemble de données. Par conséquent, nous avons construit des groupes homogènes, c'est-à-dire des classes qui contiennent des groupes similaires. Dans ce cadre nous avons utilisé la classification hiérarchique ascendante qui est une méthode de classification non supervisé.
Dans le chapitre suivant et dans le but d'approfendir nos analyses nous procéderons à la modélisation à l'aide d'un modéle de classification supervisé qui est la régression logistique.

Chapitre 4

Modélisation

L'objectif de cette partie est d'expliquer notre variable cible qui est une variable dichotomique (binaire) : accidenté vivant VS accidenté décédé, c'est pour ça que le modéle qui semble le plus intéressant est le modéle de régression logistique.
On va donc modéliser la probabilité de décès pour connaître les facteurs participants à la gravité des accidents.
Dans ce chapitre, tous les tests d'hypothése et les intervalles de confiance seront faits avec un risque de premiére espéce α=0.05.

4.1 Fondement théorique

La régression logistique est une technique prédictive. Elle constitue un cas particulier de modèle linéaire généralisé qui vise à construire un modèle permettant de prédire c'est à dire expliquer les valeurs prises par une variable cible qualitative Y (dans notre cas binaire, on parle alors de régression logistique binaire) à partir d'un ensemble de variables explicatives quantitatives ou qualitatives X. La régression logistique est une technique très populaire, elle est largement répandue dans de nombreux domaines.

Notation Dans le cas d'une régression logistique binaire, la variable Y prend deux modalités possibles $\{1, 0\}$.
Les variables X_j sont exclusivement qualitatives ou quantitative.

Soit :

- Y la variable expliquée (variable à prédire)
- $X = (X_1, X_2, ..., X_J)$ les variables explicatives (variables prédictives).
- Ω un ensemble de n échantillons, comportant n_1 (resp. n_0) observations correspondant à la modalité 1 (resp. 0) de Y
- P(Y=1) (resp. $P(Y = 0)$) est la probabilité a priori pour que $Y = 1$ (resp. $Y = 0$). Pour simplifier, cela sera par la suite noté $p(1)$ (resp. $p(0)$)
- p(X| 1) (resp. $p(X|0)$) est la distribution conditionnelle des X sachant la valeur prise par Y.
- La probabilité a posteriori d'obtenir la modalité 1 de Y (resp. 0) sachant la valeur prise par XX est notée $p(1|X)$ (resp. $p(0|X)$)[17]

Le but de de cette méthode est la modélisation de l'espérance conditionnelle :

$$E(Y|(X = x)) = P(Y = 1|(X = x))$$

L'astuce de la régression logistique consiste à modéliser la probabilité que la variable qualitative Y se réalise. En effet, le modèle logistique permet une expression non linéaire, variant de façon monotone entre 0 et 1, de cette probabilité en fonction des variables explicatives $X = (X_1, X_2, ..., X_J)$.

le logit de la probabilité (P(1| X)) de la réalisation de la variable à expliquer (Y) est exprimé en fonction d'un intercept (ou ordonnée à l'origine) β_0, des variables explicatives (Xi) rattachées à leurs coefficients β_j.

L'expression du modéle logit est la suivante :

$$ln(\frac{P(1|X)}{(1-P(1|X))}) = ln(\frac{P(1|X)}{P(0|X)}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_J X_{iJ}$$

après transformation de l'équation ci-dessus, nous obtenons :

$$p(1|X) = \frac{e^{b_0+b_1x_1+...+b_Jx_J}}{1+e^{b_0+b_1x_1+...+b_Jx_J}}[16]$$

4.2 Traitement des classes déséquilibrées

Il est commun que les classes d'une variable binaires soient déséquilibrées.En effet, dans notre cas le nombre des tués, ceux que l'on cherche à identifier justement, sont rares par rapport aux blessés . Ce probléme influence négativement le modéle de régression logitique car le degrés de biais est lié au nombre d'observations de la classe minoritaire.

Ce qui peut mener à un biais dans les résultats et une mal estimation de la probabilité de réalisation de la modalité minoritaire.
Pour surmonter cet écueil, nous avons pensé à utiliser l'rééchantillonnage stratifié de notre variable cible en modifiant la base de données initiale.

On distingue plusieurs méthodes de rééchantillonnage stratifié selon la variable réponse. Ces méthodes sont divisés entre méthodes classiques et méthodes hybrides :

— Les méthodes classiques sont : le **sous-échantillonnage** (undersampling) qui consiste à rééquilibrer le jeu de données en diminuant le nombre d'observations de la classe majoritaire et le **sur-échantillonnage** (oversampling) qui fonctionne en augmentant le nombre d'instances de la classe majoritaire . Ces deux méthodes visent à obtenir un ratio :($\frac{classemajoritaire}{classeminoritaire}$)satisfaisant et conduisent à diminuer le déséquilibre entre les classses pour obtenir une base

où la classe minoritaire est mieux représentée.

— Les méthodes hybrides : ont trouve la méthode both sampling qui posséde deux algorithmes : **SMOTE** et **ROSE** qui se basent sur une combinaison des deux approches sous-échantillonnage et sur-échantillonnage.[18]

Pour corriger le déséquilibre de classes dans nos données nous avons procédé à un sous-échantillonnage qui nous a donné un échantillon de 9310 observations qui sont regroupées selon la variable cible "Etat" avec des proportions égales(c'est à dire : le nombre de blessé égale au nombre de Tué).

4.3 Modélisation

4.3.1 Choix de variables

La sélection de variables est une étape clé de la modélisation. Cette étape est utilisée dans le cas où le nombre des variables explicatives est important car certaines d'entre elles peuvent être significatives et liées à la variable à expliquée, d'autres n'ont aucun rapport avec cette variable.

Nous avous décider de passer par cette étape pour faciliter l'interprétation dans l'étape suivante et améliorer les performances du modéle.
Dans cette partie, nous avons utilisé la sélection par optimisation qui consiste à trouver le sous-ensemble de variables prédictives qui minimise un critère.
Ce dernier équilibre la réduction de la déviance, prend en considération la qualité de l'ajustement, par un indicateur qui comptabilise la complexité du modèle. En rajoutant des variables pertinentes, le critère doit continuer à décroître et en rajoutant des variables qui ne sont pas en rapport avec la prédiction ou qui sont redondantes par rapport aux variables déjà choisies, il doit augmenter.
Un des critéres qui garatie ces spécification est :

— Le critère AIC (critére d'Akaike)

$$AIC = (-2LL) + 2(J+1)$$

avec :
-2LL est la déviance
(J+1) est le nombre de paramétres à estimer

Pour ce critére l'idée est d'optimiser c'est à dire minimiser l'AIC.
Donc, nous allons procéder comme suit : On va évaluer des successions de modèles emboîtés :

- En les ajoutant les variables au fur et à mesure :FORWARD
- En retirant les variables au fur et à mesure :BACKWARD
- STEPWISE : En alternant FORWARD / BACKWARD c.-à-d. vérifier que chaque ajout de variable ne provoque pas la sortie d'une autre variable.

Règle d'arrêt : l'adjonction ou le retrait d'une variable n'améliore plus le critère.[20]

FORWARD : sélection pas à pas ascendante

L'idée c'est de partir d'un modèle initial avec aucune variable et d'ajouter une à une les variables qui sont suffisamment liées à notre variable cible et qui contribuent le maximum à notre modèle en prenant en considération les variables sélectionnées précédemment.

— **En utilisant le critére AIC :**

Etape	variable	DDL	AIC	khi2	Pr>khi2
1	Lieu	5	12514.89	403.509269	< 2.2e-16
2	imp_vuln	7	12180.16	348.729606	< 2.2e-16
3	Causes	7	11876.13	318.034277	< 2.2e-16
4	Heure	11	11754.89	143.237871	< 2.2e-16
5	Gouvernorat	23	11642.22	158.669727	< 2.2e-16
6	type_vic	5	11587.56	64.664719	< 2.2e-16
7	Source	1	11587.44	2.118859	< 2.2e-16

figure1 :les étapes de sélection pas à pas Forward avec le critére AIC

critére	constante uniquement	constante et covariable
DF	9309	9250
AIC	12908.4	11587
-2Log L	12906	11467

figure1 :Statistiques d'ajustement du modéle FORWARD avec le critére AIC

BACKWARD : sélection pas à pas descendante

L'idée c'est de partir d'un modèle initial avec toutes les variables et de rejeter une à une les variables qui ne sont pas suffisamment liées à notre variable cible et qui ne contribuent pas à notre modèle en prenant en considération les variables sélectionnées précédemment.

— **En utilisant le critére AIC :**

Etape	variable	DDL	AIC	khi2	Pr>khi2
1	Mois	11	11594.43	8.85044268	0.17100700
2	Jours	6	11589.35	6.92596836	0.08370691
3	vac	1	11587.44	0.08191197	0.73354520

figure1 :les étapes de sélection pas à pas BACKWARD avec le critére AIC

critére	constante uniquement	constante et covariable
DF	9309	9250
AIC	12908.4	11587
-2Log L	12906	11467

figure1 :Statistiques d'ajustement du modéle BACKWARD avec le critére AIC

STEPWISE : sélection pas à pas mixte

L'idée c'est de partir d'un modèle initial avec aucune variable et d'ajouter une à une les variables qui sont suffisamment liées à notre variable cible mais à chaque étape il est possible de rejeter une variable si elle contribue au modèle de la même manière qu'une combinaison de nouvelles variables.

— **En utilisant le critére AIC :**

Etape	variable	DDL	AIC	khi2	Pr>khi2
1	Lieu	5	12514.89	403.509269	< 2.2e-16
2	imp_vuln	7	12180.16	348.729606	< 2.2e-16
3	Causes	7	11876.13	318.034277	< 2.2e-16
4	Heure	11	11754.89	143.237871	< 2.2e-16
5	Gouvernorat	23	11642.22	158.669727	< 2.2e-16
6	type_vic	5	11587.56	64.664719	< 2.2e-16
7	Source	1	11587.44	2.118859	< 2.2e-16

figure1 :les étapes de sélection pas à pas STEPWISE avec le critére AIC

critére	constante uniquement	constante et covariable
DF	9309	9250
AIC	12908.4	11587
-2Log L	12906	11467

figure1 :Statistiques d'ajustement du modéle STEPWISE avec le critére AIC
Nous obtenons des résultats similaires avec les trois méthodes En effet, en utilisant le critère AIC les 3 variables Mois, Jours et Vac sont rejetées et on reste donc avec les 7 variables suivantes : Gouvernorat, Lieu, Causes, Heure, type$_v ic, imp_v ulnetSource$. L'étape suivante est de définir les modalités de référence pour chaque variable sélectionnée. Cette étape est très importante car tous les coefficients dans le modèle sont calculés par rapport à la modalité de référence. Donc, il est important de choisir une modalité de référence qui fasse sens afin de faciliter l'interprétation.
Ci-dessous les variables sélectionnés ainsi que leurs modalités de référence :
imp_vuln : (Modalité de référence : Voiture)

Causes : (Modalité de référence : défaut d'attention du conducteur)

Heure : (Modalité de référence :18 :19)

Gouvernorat : (Modalité de référence :Tunis)

type_vic : (Modalité de référence :Passager)

Source : (Modalité de référence :Police)

Lieu : (Modalité de référence :Agglomération)

4.3.2 Construction du modéle :

On commence par comparer le modèle qui contient toutes les variables (toute l'information) au modèle trivial (modèle nul avec aucune variable).On va donc voir si notre modélisation de la probabilité de décès apporte plus d'information que le pire modèle (modèle Nul).
On test alors :
H0 :$\forall$ i,j, $(\beta_i)_j$ =0 Les variables n'influent pas la probabilité de décès.(C'est à dire mon modèle fait mieux que le modèle nul)
H1 :$\exists$ i,j, $(\beta_i)_j$ $\neq$0 au moins une modalité correspondante à une variable influe le modèle avec $(\beta_i)_j$ le coefficient corréspendant à la variable i quand elle prend pour valeur la modalité j. On utilise pour ça le test de rapport de vraissemblance :

khi2	DDL	Pr>khi2
1379.935	57	< 2.2e-16

figure1 :Test de significativité globale du modéle

On constate que la P-value est inférieure à 2.2e-16 donc est inférieur à 0.05.
Conclusion : Il y a au moins une variable qui influe la probabilité de décès dans le modèle.
On teste maintenant la significativité des variables explicatives une à une sachant que les autres sont incluses dans le modèle :
H0 :β_1=0 / β_2,β_3,...=0
H1 :β_1 $\neq$0 β_2,β_3,...=0
On effectue le test de khi2 de Wald sous H0 et on obtient les résultats suivants :

variable	LR	DF	P-Value
Lieu	29.70189	4	5.628220e-06
Gouvernorat	126.74028	23	2.512321e-16
Cause	272.65151	7	4.143083e-55
Heure	148.78669	11	2.637166e-26
type_vic	59.42651	5	1.596693e-11
imp_vuln	160.38437	7	2.662630e-31
Source	137.43289	1	4.527633e-25

figure1 :Test de significativité des variables

On voit que les variables renvoient toutes une p-value <0.05. On rejéte donc l'hypothése H0.
Ce qui veut dire que toutes les variables sont significatives et contribuent à la construction du modèle. L'étape suivante est de voir quelles sont les variables qui influent notre modèle d'une manière significative positive et celles qui agissent d'une manière significative négative sur le risque de décès. Ci-dessous un tableau récapitulatif qui regroupe les Odds ratio (rapport des cotes) du modèle.
A travers ce modèle nous allons interpréter l'effet des facteurs. L'interprétation se fait sous les règles suivantes :

- Si 1 n'est pas présent dans l'intervalle de confiance alors le rapport de côte est significative.
- Si toutes les valeurs de l'intervalle sont en dessous de 1 la modalité correspondante est significativement moins dangereuse que la modalité de référence.
- Si toutes les valeurs de l'intervalle sont en dessus de 1 alors la modalité observée est significativement plus dangereuse que la modalité de référence.

4.4 application

.

Bibliographie

[1] M.Másilková "Health and social consequences of road traffic accidents",Kontakt,2017.

[2] ONISR 2017 : Bilan de l'accidentalité en France (métropole et outre-mer).

[3] Bull, J.P., Roberts, B.J., 1973. Road accident statistics—a comparison of police and hospital information. Accident, Analysis and Prevention 5, 45–53.

[4] Haynes, R., Jones, A., Harvey, I., Jewell, T., Lea, D., 2005. Geographical distribution of road traffic deaths in England and Wales : place of accident compared with place of residence. Journal of Public Health 27 (1), 107–111.

[5]Leo Gimenez,2021.6 étapes pour le nettoyage des données et pourquoi c'est important.

[6]Gilles Malaterre,2000.Risque et sécurité sur la route.

[7]Claire Dupuy,2020."Dictionnaire des politiques territoriales" :Inégalités territoriales.

[8]François Husson,2009, 224 p."Analyse des données avec R, Presses Universitaires de Rennes" :Inégalités territoriales.

[9]Brigitte Escofier et Jérôme Pagès,2008, 318 p."Analyses factorielles simples et multiples" :objectifs, méthodes et interprétation, Paris, Dunod, Paris.

[10]Jérôme Pagès,2013."Analyse factorielle multiple avec R , Les Ulis, EDP sciences, Paris" :Inégalités territoriales.

[11]Ricco RAKOTOMALALA,2021."Classification automatique, typologie, clustering".

[12]Foued Aloulou, Sana Naouar([2016], p.1211-1212)."Analyse microéconométrique des accidents routiers en Tunisie".

[13]Gaudry et De Lapparent ([2008], p. 35) présentent un tableau récapitulatif des divers travaux menés sur la sécurité routière exploitant les méthodes de régression.

[14]Foued Aloulou, Sana Naouar([2016], chapitre3)."Facteurs de risque".

[15]Hakamies-Blomqvist L. Ageing Europe : the challenges and opportunities for transport safety [5e conférence européenne sur la sécurité des transports].Bruxelles (Belgique), European Transport SafetyCouncil, 2003.

[16]M. El Sanharawi "Comprendre la régression logistique",2013.

[17]Wikipédia "Régression logistique".

[18]ORNELLIA DJOFFON, "Modélisation de la survenance d'un sinistre dans le cas d'une asymétrie des classes et utilisation dans le cadre d'un modèle interne partiel"(p. 6)

[19]Ricco Rakotomalala(2017,p.115-117) "Pratique de la Régression Logistique :Régression Logistique Binaire et Polytomique(Version 2)".

[20]Ricco Rakotomalala "Régression logistique : Une approche pour rendre calculable P(Y/X)"

(N1)www.ibm.com
(N2)https ://www.who.int/
(N3)http ://www.etsc.be/eve.htm