17 Janvier 2021

Analyse statistique et prédiction de la gravité d'un accident en France

REBOUILLAT Caroline - THONNEAU Laura

Remerciements

Nous remercions Patrick SEVESTRE, professeur des universités en économétrie et statistiques à l'Aix-Marseille School of Economics (AMSE), pour l'aide technique qu'il nous a apportée ainsi que pour les pistes de réflexion qu'il a pu nous suggérer.

Table des matières

Résumé

L'objectif principal de cette étude est de fournir une analyse complète des potentiels facteurs déterminant la gravité des accidents, en France, afin de mieux cibler les politiques de prévention, d'aménagement et de développement des routes. Compte tenu de ces exigences, il est nécessaire de soulever deux principales questions :

- Comment adapter les politiques de prévention en matière de sécurité routière en fonction de **l'environnement de l'accident** et des caractéristiques du véhicule ?
- Quelles **caractéristiques particulières des zones d'accident** devraient être ciblées pour adapter les politiques de développement ?

Pour répondre à ces deux grandes problématiques, notre équipe a effectué une étude basée sur des outils statistiques et économétriques.

Les données fournies par la DSR sont réparties en *4 tables* :

- Lieux
- Caractéristiques
- Véhicules
- Usagers

A partir de ces quatre tables, nous avons pu créer une base de données finale dans laquelle une ligne correspond à un véhicule accidenté en regroupant certaines informations : nous aons gardé quasiment toutes les variables des tables Lieux et Caractéristiques, une partie de la table Véhicules et deux de la table Usagers (la variable à expliquer, i.e. la gravité des blessures causées ainsi que le motif de déplacement). Réduire le nombre d'observation contenues dans ces variables pour les incorporer à la base finale a nécessité un peu plus de travail : nous avons transformé l'indicateur de gravité des blessures en indicateur de gravité de l'accident en prenant la gravité de blessure maximale dans chaque véhicule, et nous avons pris en compte le motif de déplacement du conducteur uniquement (et non ceux des passagers).

Par ailleurs, nous avons nettoyé la base de toute donnée manquante : notre parti-pris a été de supprimer les variables dans lesquelles il y avait un grand nombre de données manquantes (ces variables étaient toutes peu utiles pour notre analyse), puis de supprimer les lignes où il restait des données manquantes (il en restait peu). Enfin, nous avons recodé nos variables catégoriques : redéfinition de certaines

catégories de variables, puis transformation de toutes les variables catégoriques en variables binaires. Le travail de prétraitement est donc véritablement une réorganisation des données pour les rendre exploitables.

Dans un premier temps, nous appliquons à notre base un modèle économétrique utilisant la méthode de régression logistique. C'est un modèle approprié aux bases de données pour lesquelles nous cherchons à expliquer une variable binaire (la gravité possède 4 catégories, mais nous les avons regroupées en classe 0 (faible gravité) et en classe 1 (forte gravité)). Grâce à cette méthode, nous pouvons mieux cerner l'importance de chaque facteur sur la gravité. Nous trouvons que des éléments environnants de la route peuvent clairement aggraver un accident, comme la présence d'arbres, de bâtiments, de fossés, de supports de signalisation, etc.

Dans le cadre de la politique d'aménagement des routes, il est donc visiblement important de tenir compte de ces éléments, et de construire autant que possible des axes routiers isolés des villes, des ravins et des forêts.

Nous avons également appliqué un modèle de Machine Learning pour prédire la gravité : les arbres décisionnels. La pertinence de ce modèle est d'environ 77% contre 76% pour le premier. De plus, ses prédictions sont plus précises. Enfin, il met en valeurs d'autres facteurs importants, comme les véhicules légers, les pentes, les agglomérations, les routes bidirectionnelles et départementales, etc. Nous obtenons donc des pistes de réponse très intéressantes quant à l'affinement des campagnes de prévention, et des politiques de développement et d'aménagement des routes.

Introduction

En 2018, le nombre d'accidents de la route est de 58 352. C'est 4,7% de moins qu'en 2017 et 32,9% de moins qu'en 2005. Si la fréquence des accidents en France décroît de manière significative, ce n'est pas le cas pour la gravité des accidents qui, elle, varie d'une année à l'autre mais reste dans l'ensemble constante : sur 100 accidents corporels en 2018, on compte 6 morts contre 5,77 en 2007 et 6,37 en 2005. Nous comprenons dès lors que l'un des enjeux principaux de politiques de prévention routière, d'aménagement et de développement des routes est de viser à réduire la gravité des accidents à long terme. C'est dans cette direction que pourrait s'orienter les volontés de la Délégation à la Sécurité Routière (DSR). En effet, du fait du contexte sanitaire lié à la Covid-19, les dépenses du gouvernement augmentent de manière accrue. C'est pourquoi des plans de restrictions budgétaires ont été mis en place dans certaines branches du gouvernement, ce qui demande à la DSR plus d'efficience autour de l'analyse de la gravité des accidents.

Cette analyse est articulée autour de deux principales questions :

- Comment adapter les politiques de prévention concernant la sécurité routière en fonction des éléments environnementaux des accidents et des caractéristiques des véhicules ?
- Quelles sont les facteurs à prendre en compte dans le cadre de politiques de développement des routes ?

Avant toute chose, il est important de connaître l'état de l'art, c'est-à-dire les réponses qui ont déjà pu être apportées dans la Recherche. C'est une étape cruciale pour orienter habilement notre étude. En ce qui concerne notre problématique, il y a beaucoup de solutions apportées impliquant des modèles économétriques.

Premièrement, dans *Econometric models of road use, accidents, and road investment decisions, Volume II*, by Lasse Fridstrøm (1999), la densité du trafic est un facteur pointé du doigt concernant la gravité des accidents, mais aussi les condition météorologiques (par exemple, la chute de neige à un impact négatif sur la gravité d'un accident).

Deuxièmement, Noureddine Benlagha et Lanouar Charfeddine dans *Risk factors of road accident severity and the development of a new system for prevention: New insights from China* (2020) nous disent que les facteurs qui expliquent la gravité d'un accident dépend du niveau de gravité de l'accident. Par exemple, le genre de la personne est uniquement significatif pour le degré maximal de gravité (la mort) : les hommes ont tendance à prendre plus de risques sur la route, causant plus souvent des accidents mortels.

Troisièmement, en ce qui concerne les bus, la gravité de l'accident augmente pour les conducteurs qui ont moins de 25 ans et plus de 55 ans, pour les femmes, lorsque les limitations de vitesse sont soit très faibles (en-dessous de 30 km/h, la gravité est très élevée probablement car la personne touchée est un piéton ou un cycliste), soit très élevées (au-dessus de 100 km/h), aux intersections et lorsque la conduite est inattentive et risquée, selon Sigal Kaplan et Carlo Giacomo Prato dans *Risk factors associated with bus accident severity in the United States : A generalized ordered logit model* (2012).

Quatrièmement, dans *Risk factors affecting the severity of traffic accidents at Shanghai river-crossing tunnel* (2015), Jian John Lu, Yingying Xing, Chen Wang et Xiaonan Cai nous montrent que la gravité augmente avec le fait que le conducteur soit un homme, âgé de plus de 65 ans, entre minuit et l'aube, lors des week-ends, sur une surface mouillée, avec des véhicules de très bonne qualité, lorsque la limitation de vitesse est moyenne (entre 50 et 79 km/h) et dans des très longs tunnels (plus de 3 km de long).

Nous pouvons constater que beaucoup d'analyses dans le domaine ont déjà été effectuées, et que, donc, il y a déjà beaucoup de réponses aux problématiques posées par la DSR. Cependant, nous pouvons constater que certains résultats des études précédemment citées sont contradictoires. Cela pourrait s'expliquer par le fait que ces travaux ont été effectués sur des données qui concernent des populations différentes (accidents en Chine, aux Etats-Unis et en Norvège) ce qui pourrait impliquer qu'il y ait des différences de comportements sur la route suivant le pays. Il est à noter que des études sur les accidents de la route, en France, ont été opérées, mais il semblerait qu'elle ne concerne pas la gravité des accidents, plutôt leur fréquence.

Ainsi, mener une étude sur les facteurs aggravants des accidents en France peut aboutir à des résultats différents, d'où l'intérêt de notre démarche.

En ce qui concerne notre approche, pour répondre le plus pertinemment possible aux questions énoncées précédemment, nous nous appuyons sur les données des accidents corporels sur les routes françaises (y compris les territoires d'Outre-Mer) de 2018, disponibles sur le site du gouvernement. Les données sont réparties en 4 tables : Caractéristiques des accidents, Lieu des accidents, Véhicules impliqués et Usagers impliqués. Après avoir procédé à une réorganisation des données (notre table de données finale est composée de la variable gravité prenant la valeur 0 pour peu grave et 1 pour très grave et toutes nos autres variables sont binaires), nous avons appliqué un modèle économétrique très populaire pour quantifier l'importance des facteurs que nous analysons : la régression logistique. C'est un modèle capable de définir la probabilité d'un évènement (variable à expliquer) en fonction de facteurs (variables explicatives). Grâce au calcul des effets marginaux de chaque variable sur la gravité, i.e. le calcul de l'évolution de la probabilité d'avoir un accident grave lorsque la valeur d'une variable passe de 0 à 1, toutes les autres restant constantes, nous avons pu répondre fournir une première réponse. En effet, selon nos résultats, les éléments aggravant le plus les conséquences d'un accident sont la présence :

- d'arbres
- de bâtiments, murs ou piliers de pont
- de supports de signalisation
- de poteau
- de fossé, talus ou paroi rocheuse

Ces variables sont celles dont les effets marginaux sont les plus élevés, nous discuterons des autres dans la section 4. Finalement, pour répondre à la problématique de pouvoir obtenir un indicateur de gravité d'accident, nous avons utilisé notre modèle logistique. Sa pertinence sur la prédiction de gravités d'accidents avec lesquels il n'a pas été entraîné (nous appelons cela les prédictions hors-échantillon) est de 76%, ce qui s'avère être plutôt encourageant.

Cependant, pour fournir l'indicateur le plus précis possible, nous avons décidé d'appliquer à ce modèle d'autres techniques. Par ailleurs, nous avons également utilisé un algorithme de Machine Learning pour avoir une autre approche de prédiction : les arbres décisionnels. Il s'avère qu'il est non seulement un peu plus pertinent hors-échantillon (77%), mais plus précis également dans les prédictions. De plus, il met en lumière les éléments les plus importants dans son analyse, à savoir les véhicules légers, les pentes, les routes bidirectionnelles, les routes départementales, les agglomérations, etc.

Ce rapport, s'articule autour de 4 axes. Tout d'abord une brève partie théorique concernant la définition des modèles économétriques et des techniques statistiques utilisées. Puis, suit la présentation des données, de leur réorganisation et de leur analyse descriptive. Les résultats des modèles et outils utilisés seront ensuite dévoilés et interprétés pour aboutir à une conclusion répondant aux questions traitées. Enfin, nous proposons dans une quatrième et dernière section des pistes d'amélioration de l'analyse et de la prédiction. Les dernières pages sont consacrées, quant à elles, à l'annexe où se trouvent toutes les figures citées dans l'études ainsi qu'à la bibliographie et sitographie.

Section 1 : Méthodes statistiques et économétriques

L'économétrie est une application quantitative de modèles statistiques et mathématiques qui utilise des données pour développer des théories, confirmer ou infirmer des hypothèses économiques et prédire des tendances depuis des données historiques. Toutefois, les modèles économétriques peuvent être adaptés à toutes sortes de domaines, et pas uniquement au secteur économique.

La régression logistique

Ce modèle est utilisé pour estimer la relation entre une variable dépendante (celle que l'on cherche à expliquer, dans notre cas, la gravité), qui doit être binaire, et une ou plusieurs variables explicatives (dans notre cas, les facteurs qui pourraient expliquer la gravité, par exemple, le type de route, l'état de la chaussée, etc.). Par ailleurs, ce modèle utilise une fonction qui sert à prédire les probabilités qu'un évènement arrive (dans notre cas, que gravité = 1, i.e. qu'il y ait grave blessure ou mort). En fonction de la valeur de la probabilité prédite, la classe ou le label, est attribué à l'observation : par exemple, à partir de 0,5, on prédit niveau de gravité 1 (nous pouvons choisir de modifier ce seuil de probabilité).

Appliquée à nos données, l'équation de la régression logistique s'exprime comme suit :

$$logit(P(gravité = 1)) = ln\left(\frac{P(gravité=1)}{P(gravité=0)}\right) = b_0 + b_1*Véhicule\_léger_1 + b_2*Véhicule\_lourd_2 + \ldots + b_{107}*Ivre_{107}$$

Note : nous n'avons pas pu afficher toutes les variables explicatives car nous en avons 107, voir leur description dans la Section 2.

Le terme « logit » fait référence à la transformation logistique du rapport des probabilités (la probabilité que gravité = 1 divisée par la probabilité que gravité = 0).

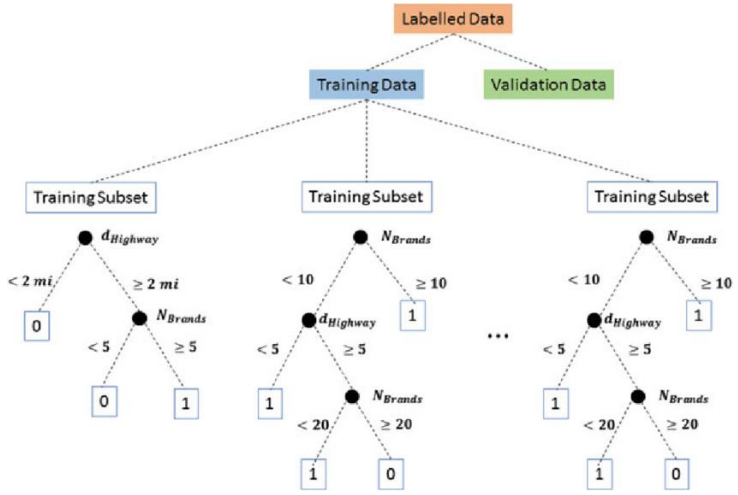
Pour pouvoir commensurer l'impact de chaque variable explicative, nous avons calculé leurs effets marginaux, c'est-à-dire leur impact sur la probabilité d'observer un accident grave lorsqu'elles passent de la valeur 0 à la valeur 1 (comme toutes nos variables sont binaires), toutes les autres variables restant constantes. Nous avons également regardé la significativité de l'effet marginal de chaque variable grâce au test de Student. En effet, cette information nous permet de savoir avec un seuil de confiance 95% si la variable a un réel impact sur la variable à expliquer ou non.

La régression logistique avec l'Elastic Net

Parfois, un modèle peut être très bien estimé, de sorte que les erreurs soient les plus faibles possibles mais devienne très peu pertinent sur des observations qui ne sont pas les mêmes que celles qui ont servi à estimer le modèle : on appelle cela le problème d'*overfitting*. En substance, cela signifie que notre modèle est trop complexe du fait du trop grand nombre de nos variables explicatives : le modèle colle presque parfaitement aux données sur lesquelles il est estimé, mais n'est pas capable d'effectuer une « généralisation » des information pour produire des prédictions pertinentes. Pour pallier ce souci, il existe des méthodes qui pénalisent le nombre de variables. Nous avons choisi l'une de ces méthodes, l'Elastic Net, car il pénalise la complexité du modèle de manière robuste à la colinéarité de certaines variables (ce qui n'est pas le cas d'autres méthodes). Concrètement, cette technique peut sélectionner les variables les plus pertinentes, s'il estime que d'autres le sont moins. De ce fait, nos prédictions peuvent être plus précises.

La technique d'arbres décisionnels

Cette technique est un algorithme de Machine Learning, c'est-à-dire que le modèle d'arbres décisionnels apprend des données, sans faire d'hypothèse initiales sur leurs relations. En effet, il améliore de manière autonome les prédictions, nous n'avons donc pas besoin d'appliquer une pénalisation comme pour le cas de la régression logistique. Ce modèle est fait d'une multitude d'arbres de décisions (voir schéma ci-dessous). Il va alors résulter de cette analyse une prédiction (dans notre cas le niveau de gravité) en fonction des valeurs des variables explicatives. Cet algorithme nous intéresse d'autant plus qu'il permet de résumer l'importance de chaque variable explicative pour la prédiction : nous pouvons alors identifier les principaux facteurs aggravants d'un accident.



La technique de Gradient Descent

Cette technique est un processus visant à réduire petit à petit la fonction de coût de notre modèle. Dans notre cas, notre régression logistique fait des prédictions, et ses erreurs de prédictions sont modélisées par une fonction de coût. Le Gradient Descent va donc essayer de minimiser cette fonction de coût en modifiant étape par étape les coefficients qui ont été attribués à chacune de nos variables explicatives, le but étant « d'apprendre de ses erreurs » et de fournir les prédictions les plus pertinentes possibles.

Section 2 : Exploration des données

1) *Présentation des données*

Notre source de données est le site du Gouvernement Français, sur lequel nous avons pu récolter les données des accidents corporels ayant eu lieu en 2018. Est qualifié d'accident corporel tout accident ayant fait au moins une victime, nécessitant des soins et étant survenu sur une voie ouverte à la circulation. Pour chacun de ces accidents, des informations sont collectées par les forces de l'ordre et l'ensemble de ces données compose ce qu'on appelle le « fichier BAAC (Base de données annuelles des accidents corporels) ». Cette base est divisée, rappelons-le, en quatre tables : Caractéristiques des accidents, Lieu des accidents, Véhicules impliqués et Usagers impliqués.

Ce fichier BAAC, comme son nom l'indique, est disponible chaque année et ce, depuis 2005 jusqu'à 2019. Cependant nous avons décidé de prendre uniquement l'année 2018 car, comme nous l'avons dit précédemment, la fréquence des accidents a fortement diminué depuis 2005 mais la gravité varie très peu. Nous voulions donc avoir les données les plus récentes pour ne pas risquer de biaiser nos résultats. A noter également, que nous avons décidé de ne pas prendre les données de l'année 2019 puisque celles-ci ont été sujet à une importante modification concernant le label de gravité de l'accident. En effet, l'autorité de la statistique publique ne labellise plus l'indicateur « blessé hospitalisé » et cet indicateur nous paraissait, à première vue, important pour notre analyse. De plus, cette base de données ne prend pas en compte certaines caractéristiques concernant les usagers, les voitures et comportements, par souci de divulgation et d'atteinte à la protection de la vie privée. C'est pourquoi nous ne disposons pas d'indicateur tel que le taux d'alcool.

De plus, la base de données étant constituée de quatre tables différentes nous avons dû les fusionner pour recueillir toutes les informations dont nous avions besoin, mais au préalable, il a fallu les réorganiser. D'une part, car certaines variables ne sont

pas utiles, au regard de notre sujet. En effet, certaines descriptions des usagers et des véhicules ne sont pas retenues puisque nous voulons surtout mettre en lumière des facteurs topographiques aggravant un accident, ce qui est dissocié du comportement des conducteurs et des caractéristiques des véhicules. D'autre part, car les tables de données n'ont pas le même nombre d'observations et qu'elles avaient des données manquantes.

Notre variable expliquée est donc la gravité de l'accident qui est codé en binaire comme nous l'avons vu précédemment. Et nous avons 19 variables explicatives dans notre modèle, toutes catégoriques (voir annexe A). Dans ces variables il y a notamment le type de véhicule, le type d'obstacle fixe ou mobile heurté, le type de manœuvre effectuée avant l'accident, le type d'intersection, le motif du trajet, les conditions lumineuses et météorologiques, le type de route, l'état de la surface, si l'accident a eu lieu en agglomération ou hors etc.

2) *Prétraitement et réorganisation*

Comme expliqué précédemment, nous avons réorganisé les quatre tables afin d'avoir une unique table qui soit exploitable pour nos analyses. Dans cette table de données, ayant 89816 observations, une ligne représente un véhicule ayant été impliqué dans un accident et chaque véhicule est associé à un identifiant d'accident. Avant de pouvoir utiliser la base de données, nous avons dû effectuer plusieurs modifications.

Tout d'abord, concernant la variable dépendante, qui possède quatre classes (Indemne, blessé léger, blessé hospitalisé et décédé), nous devions garder seulement une valeur de gravité par véhicule. Pour ce faire, nous avons testé deux méthodes : garder la gravité de blessure la plus importante parmi les usagers d'un véhicule ou bien faire la moyenne des gravités dans un même véhicule. Le choix s'est fait en comparant la méthode qui se rapprochait le plus des données initiales en termes de proportion de chaque gravité dans le jeu de données (voir annexe B) et il s'avère que

la méthode du maximum est la meilleure. De plus, pour pouvoir faire une régression logistique il faut que notre variable à prédire soit binaire, c'est pourquoi nous avons réuni les individus indemnes et légèrement blessés dans la classe 0 et ceux grièvement blessés ou décédés dans la classe 1. À noter également que, comme avec notre variable dépendante, nous avons dû regrouper prendre un motif de trajet par véhicule et nous avons choisi de prendre le motif du conducteur.

Deuxièmement, nous avons regroupé les catégories de la variable explicative « type de véhicules » en seulement 4 catégories : véhicules légers (type deux roues), véhicules légers pesant moins de 3,5 tonnes, poids lourds pesant plus de 3,5 tonnes et enfin la catégorie des trains et tramways. Il nous a semblé nécessaire de réduire les 32 catégories initiales de véhicules pour plus de facilité et d'interprétabilité. À noter que nous avons également regroupé les DOM-TOM dans une même catégorie.

Ensuite nous avons trouvé pertinent de créer deux variables et de les intégrer à notre analyse afin d'améliorer notre modèle. La première est la variable « responsabilité », cette information pourrait refléter l'importance du choc et donc de la gravité des blessures. Celle-ci a été créée grâce à la variable « choc » qui indique à quel endroit sur le véhicule le choc a eu lieu. En effet, nous considérons un véhicule ayant eu une collision à l'avant comme responsable, un véhicule ayant subi un choc à l'arrière comme non responsable et lorsque qu'il y a des chocs multiples, comme des tonneaux, la responsabilité est inconnue.

La deuxième variable que nous avons jugé bon de créer est la probabilité d'être alcoolisé, cette variable pourrait effectivement avoir un impact sur la gravité de l'accident. Pour cela nous avons croisé à la fois la variable du motif du trajet avec l'heure et le jour auxquels l'accident à eu lieu. La variable « proba_alcool » est égale à 1 si le motif du trajet est promenade ou loisirs et si l'accident a eu lieu le soir ou la nuit d'un jeudi, vendredi ou samedi, ou bien un dimanche après-midi. Si ces conditions ne sont pas remplies alors la probabilité d'être alcoolisé est moins forte et la variable prend la valeur 0.

Enfin nous avons dû transformer nos 19 variables catégoriques en variables binaires afin de pouvoir interpréter plus facilement nos résultats. Nous nous retrouvons donc avec un total 107 variables explicatives.

3) *Statistiques descriptives*

Après avoir nettoyé et organisé nos données, nous avons voulu commencer par nous familiariser avec celles-ci et avoir une première vue dessus. Nos variables étant des variables catégoriques, nous ne pouvons pas calculer des statistiques de base, comme la moyenne par exemple. Nous avons donc opté pour l'idée d'analyser les proportions de chaque valeur dans chaque variable, afin de mieux cerner nos données. Et nous avons commencé par la variable à prédire, la gravité. La classe 0 représente 75,8 % de nos observations contre 24,2 % pour la classe 1. Nos données sont donc déséquilibrées ce qui pourrait biaiser notre prédiction, c'est-à-dire bien prédire les usagers indemnes ou blessées légers mais très mal prédire la classe minoritaire (blessés hospitalisés ou tué). Nous avons donc pallié ce problème en utilisant une technique de rééquilibrage de données appelée SMOTE. Cette technique sélectionne des observations qui sont proches au niveau des caractéristiques, dessine une ligne entre ces observations et dessine un nouvel échantillon à partir de cela. En substance, elle crée de nouvelles observations en regardant les valeurs pour chaque variable d'observations proches ayant le même label.

Deuxièmement, en analysant et croisant certaines de nos variables nous avons pu observer que l'accident typique a lieu en plein jour, sans intersection, dans des conditions atmosphériques normales, sur une voie communale bidirectionnelle, en métropole, sur du plat, en ligne droite et sur une surface en état normal. Il semblerait que les accidents ont lieu le plus souvent sur des routes où les conditions sont idéales, ce qui doit sûrement faire baisser l'attention des usagers. Cependant nous ne pouvons pas tirer de conclusion trop hâtive avec cette seule analyse.

Ensuite nous avons voulu regarder comment différaient le nombre d'accidents et leur gravité selon les mois et selon les jours.

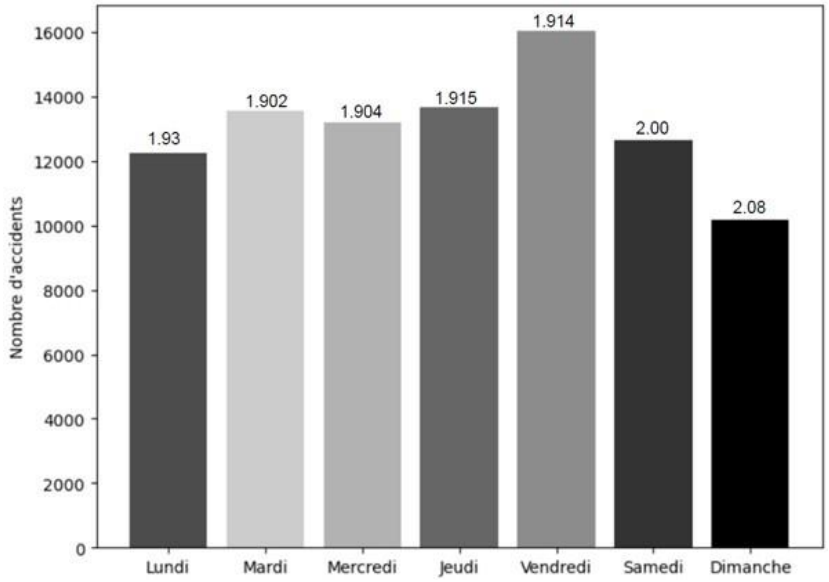


Figure 1 : Nombre d'accidents et gravité moyenne selon le jour de la semaine

Ce que nous pouvons tirer de ce graphique est qu'en moyenne, il y a beaucoup plus d'accidents le vendredi et moins le week-end. Cependant, les accidents se produisant le week-end sont en moyenne plus graves puisque, le samedi, la moyenne de gravité est de 2 et le dimanche de 2.08. Cela est peut-être dû aux sorties et loisirs plus fréquents le week-end.

Similairement, si on utilise la même technique pour les mois de l'année (voir annexe C), on observe que le mois de février a le nombre d'accidents moyen le moins élevé mais la gravité moyenne la plus élevée. Cependant pour les jours comme pour les mois, les moyennes de gravité restent très proches.

Troisièmement, nous avons voulu représenter sur une carte les différents accidents en les croisant avec certaines de nos variables explicatives. Par exemple, nous avons voulu voir si la répartition des accidents selon la saison était différente en France, ce qui ne s'avère pas très révélateur (voir annexe D). Ce qui est plutôt évident, c'est qu'il y a une concentration des accidents vers les grandes villes.

Enfin, nous avons voulu tester la corrélation entre nos variables puisque c'est une des conditions nécessaires pour avoir un modèle de régression logistique pertinent. Après vérification (voir annexe E), nous n'observons pas de corrélation très forte entre nos variables. Nous pouvons donc passer à la description et l'élaboration de nos modèles.

Section 3 : Résultats et conclusion

Les résultats de notre régression logistique mettent en lumière l'effet de variables concernant des éléments faisant partie de l'environnement de la route et de mettre l'importance de certaines variables. En effet, 23 variables ne sont pas statistiquement significatives selon le test de Student, par exemple la neige, les vents forts, la chaussée inondée, etc. Mais cela ne veut pas dire qu'elles ne sont réellement pas importantes : ces résultats peuvent être influencés par le fait qu'on observe très peu ces variables dans la base données. Par ailleurs, en calculant les effets marginaux, nous avons pu remarquer que 5 éléments se distinguaient en termes d'impact sur la gravité de l'accident :

- heurter un arbre
- heurter un bâtiment, un mur ou un pilier de pont
- heurter un support de signalisation
- heurter un poteau
- tomber dans un fossé, heurter un talus ou une paroi rocheuse

Figure 3 - Les 5 effets marginaux les plus élevés et significatifs

Variable	Effet marginal	P-value
Heurte arbre	0.3202	< 0.05
Heurte bâtiment, mur, pilier de pont	0.1011	< 0.05
Heurte support de signalisation	0.0503	< 0.05
Heurte poteau	0.0651	< 0.05
Fossé, talus, paroi rocheuse	0.2190	< 0.05

Les effets marginaux de 5 variables sont statistiquement significatifs avec un niveau de confiance de 95% (il faut se référer au p-values, si elles sont inférieures à 5%, l'effet marginal est significatif). Par ailleurs, pour interpréter les effets marginaux, il faut se référer à une variable que l'on appelle variable de référence et que l'on a choisi au préalable. Elle est la même pour les 6 variables affichée plus haut : le fait de ne heurter aucun obstacle fixe. Ainsi, nous pouvons affirmer que, selon notre modèle, lorsque l'on observe un accident, la probabilité qu'il soit grave augmente de 32,02% si le véhicule a heurté un arbre par rapport à s'il n'avait pas heurté d'obstacle fixe, toutes choses égales par ailleurs. De plus, la pertinence hors-échantillon du modèle est d'environ 76%, ce qui est plutôt bon. Cependant, ce qu'il est important de regarder pour plus de précisions, ce sont la matrice de confusion (met en perspective ce que le modèle a prédit et les classes effectives), le rapport de classification et la courbe ROC (Receiver Operating Characteristic) qui compare le taux de faux négatifs et de vrais positifs. Les faux négatifs sont les accidents que le modèle a prédit comme non graves alors qu'ils le sont, ce qui est la pire des situations dans notre cas (il est moins grave de dire qu'un accident est grave alors qu'il ne l'est pas). Les vrais positifs, eux sont les accidents prédits comme graves et qui le sont effectivement. Nous pouvons observer le nombre de chacun de ces éléments dans la figure 4, la matrice de confusion et leur rapport dans la courbe ROC (figure 5).

Figure 4 – Matrice de confusion de la Régression Logistique

Classe effective : 0	**18 489**	**2 014**
Classe effective : 1	**4 326**	**2 116**
	Classe prédite : 0	**Classe prédite : 1**

Le nombre de faux négatifs est de 4 326, ce qui est beaucoup. Il faudrait pouvoir améliorer notre modèle pour réduire ce nombre.

Figure 5 – Receiver Operating Characteristics

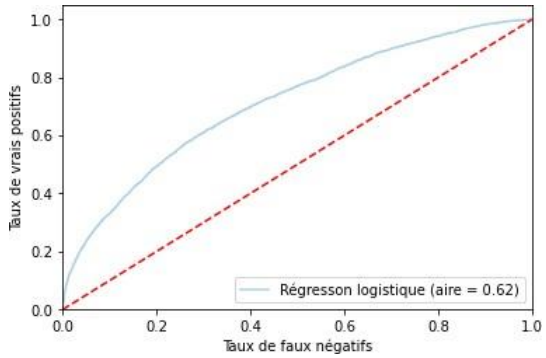


Figure 6 – Rapport de classification

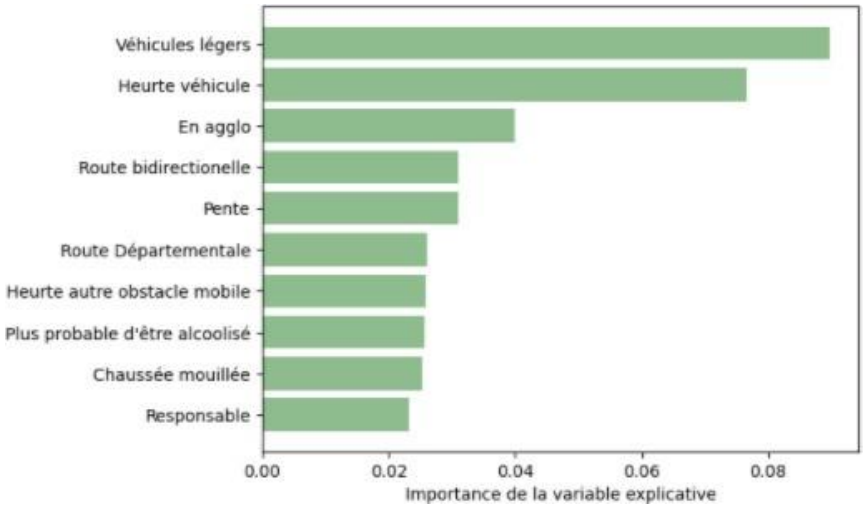
	Précision
Classe 0	0.81
Classe 1	0.51
Moyenne pondérée	0.74

L'aire sous la courbe est 62%. Cela veut dire que notre classification est loin d'être une classification aléatoire. De plus, il semblerait, selon la précision, que le modèle prédise correctement en moyenne 74% des classifications, 81% des classifications en 0 et 51% des classifications en 1. En somme, le modèle prédit beaucoup mieux les accidents non graves que graves, ce qui confirme ce qui est affiché par la matrice de confusion. Il est important de se tourner vers des techniques qui nous permettent de réduire le nombre de faux négatifs.

Concernant l'application de l'Elastic Net, aucune variable n'a été sélectionnée, c'est-à-dire qu'aucun coefficient n'a été réduit à 0 par cette technique. Donc, la pertinence de ce deuxième modèle hors échantillon est la même que celle du premier modèle. De plus, les résultats affichés par sa confusion matrix, sa courbe ROC et son rapport de classifications montrent que ce modèle ne prédit que des classes 0, ce qui le révèle comme étant un mauvais modèle car le nombre de faux négatifs explose.

Par rapport à la technique d'arbres décisionnels, nous avons obtenu un taux de pertinence hors-échantillon frôlant les 77%, ce qui est un peu mieux. L'aire en dessous de la courbe ROC est aussi plus grande (68%) ce qui signifie que cette classification s'éloigne encore plus d'une classification aléatoire. Concernant le nombre de faux négatifs, il a diminué de moitié par rapport au premier modèle : nous en avons désormais 3 149. Enfin, selon la précision, le modèle d'arbres décisionnels fournit une classification correcte à 77%. Il prédit les classes 0 correctement à 85%, et à 51% pour les classes 1 : ce modèle est donc globalement plus précis, surtout en ce qui concerne la prédiction de classes 0. Par ailleurs, cet algorithme a pu mettre en évidence les variables qu'il a considéré comme les plus importantes pour la prédiction (voir figure 7).

Figure 7 – Les 10 facteurs les plus importants pour le modèle d'arbres décisionnels



Finalement, concernant la régression logistique avec la technique de Gradient Descent, il semblerait que la pertinence hors-échantillon soit de seulement 74%, ce qui est surprenant. De plus, l'aire sous la courbe ROC montre que ce modèle se rapproche un peu plus d'une classification aléatoire (même s'il en est encore loin) que le premier modèle de régression logistique. Toutefois, grâce à la matrice de confusion, nous remarquons que ce modèle prédit moins de faux négatifs, ce qui est un élément d'amélioration. En somme, la régression logistique calculée avec la méthode du Gradient Descent prédit correctement 73% de la classification, 81% des classes 0 et 48% des classes 1.

En conclusion, au regard de nos résultats, nous suggérons à la DSR de tenir compte de la technique d'arbres décisionnels en tant qu'indicateur de gravité. Par ailleurs, pour répondre aux questions concernant l'orientation des campagnes de prévention routière, des politiques d'aménagement et de développement des routes, nous souhaiterions suggérer plusieurs choses. Premièrement, la régression logistique a permis de mettre en lumière des facteurs environnant des routes. Cela pourrait impliquer que, pour la construction future d'axes routiers, il faille se tourner, si possible, vers des zones éloignées de piliers de pont, de murs ou de bâtiments mais aussi de forêts. De plus, il serait intéressant de réfléchir à l'aménagement des supports de signalisation : par exemple, les éloigner un maximum de la chaussée pour éviter toute collision. Pour les routes déjà existantes, il serait peut-être intéressant de sécuriser les chaussées traversant une forêt et essayer d'amortir un potentiel choc avec un arbre. Enfin, il semblerait qu'orienter les campagnes de prévention vers les véhicules légers soit pertinent. Placer des rappels de vigilance dans les pentes, sur les routes bidirectionnelles et département pourrait être également une solution. Augmenter la signalisation de chaussée mouillée sur la majorité des routes semblerait également pertinent.

Mettant en parallèle les articles que nous avons cités dans notre introduction, nous voyons que nous arrivons à des résultats (sur l'importance des facteurs) en commun (chaussée mouillée, alcool) mais nous avons pu aussi mettre en lumière d'autres éléments, relevant plus de la topographie des lieux (arbres, pente, etc.).

Section 4 : Pistes d'amélioration

La base de données fournie par le gouvernement est très intéressante et peut-être analysée de différentes manières. Par exemple, il aurait été intéressant de faire des séries temporelles sur plusieurs années pour pouvoir exploiter la dimension temporelle et relier les accidents entre eux suivant cette notion (peut-être cela aurait-il mené à des prédictions plus précises). Cependant, il existe une méthode qui réunit l'analyse d'observations (ce que nous avons fait) et l'analyse temporelle : la méthode de panel data. Cela aurait pu permettre de prendre en considération le fait que des véhicules font partie des mêmes accidents et, donc, de proposer une analyse un peu plus fine. Le logiciel SAS propose des fonctions simplifiant le traitement de cette analyse.

Enfin, un modèle de Machine Learning pour être efficace en tant qu'indicateur de gravité : le modèle de réseaux neuronaux. C'est un algorithme qui apprend comme un cerveau humain, c'est-à-dire qu'il établit des liens et des connexions entre chaque variable explicative et qui apprend de ses erreurs. Généralement, ce modèle arrive aux meilleurs taux de pertinence hors-échantillon : il pourrait donc produire des prédictions plus précises, et réduire le nombre de faux négatifs.

Annexe

Annexe A : variables du modèle

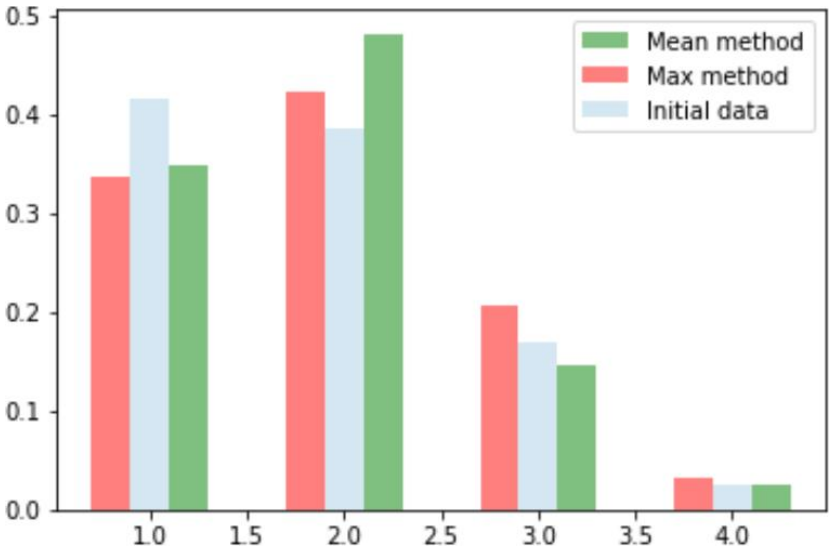
Variable	Valeurs prises
« gravite »	0 - Indemne ou blessé léger 1 - Blessé hospitalisé ou tué
« catv »	Catégorie du véhicule : 1-Véhicule léger (type deux roues) 2- Véhicule léger (< 3,5tonnes) 3- Poids lourds (>3,5 tonnes) 4- Train ou tramway
« obs »	Obstacle fixe heurté 1 - Véhicule en stationnement 2 - Arbre 3 - Glissière métallique 4 - Glissière béton 5 - Autre glissière 6 - Bâtiment, mur, pile de pont 7 - Support de signalisation verticale ou poste d'appel d'urgence 8 - Poteau 9 - Mobilier urbain 10 - Parapet 11 - lot, refuge, borne haute 12 -Bordure de trottoir 13 - Fossé, talus, paroi rocheuse 14 - Autre obstacle fixe sur chaussée 15 - Autre obstacle fixe sur trottoir ou accotement 16 - Sortie de chaussée sans obstacle
« Obsm »	Obstacle mobile heurté : 1 - Piéton 2 - Véhicule 4 - Véhicule sur rail 5 - Animal domestique 6 - Animal sauvage 9 - Autre

« manv »	**Manoeuvre principale avant l'accident** : 1 - Sans changement de direction 2 - Même sens, même file 3 - Entre 2 files 4 - En marche arrière 5 - A contresens 6 - En franchissant le terre-plein central 7 - Dans le couloir bus, dans le même sens 8 - Dans le couloir bus, dans le sens inverse 9 - En s'insérant 10 - En faisant demi-tour sur la chaussée **Changeant de file** **11** - A gauche 12 - A droite **Déporté** 13 - A gauche 14 - A droite **Tournant** 15 - A gauche 16 - A droite **Dépassant** 17 - A gauche 18 - A droite **Divers** 19 - Traversant la chaussée 20 - Manoeuvre de stationnement 21 - Manoeuvre d'évitement 22 – Ouverture deporte 23 – Arrêté (hors stationnement) 24 – En stationnement (avec occupants)
« trajet »	Motif du déplacement au moment de l'accident : 1 - Domicile – travail 2 - Domicile – école 3 - Courses – achats 4 - Utilisation professionnelle 5 - Promenade – loisirs 9 - Autre
« lum »	Lumière : conditions d'éclairage de l'accident 1 - Plein jour 2 - Crépuscule ou aube 3 - Nuit sans éclairage public 4 - Nuit avec éclairage public non allumé 5 - Nuit avec éclairage public allumé
« agg »	Localisation : 1 – Hors agglomération 2 – En agglomération

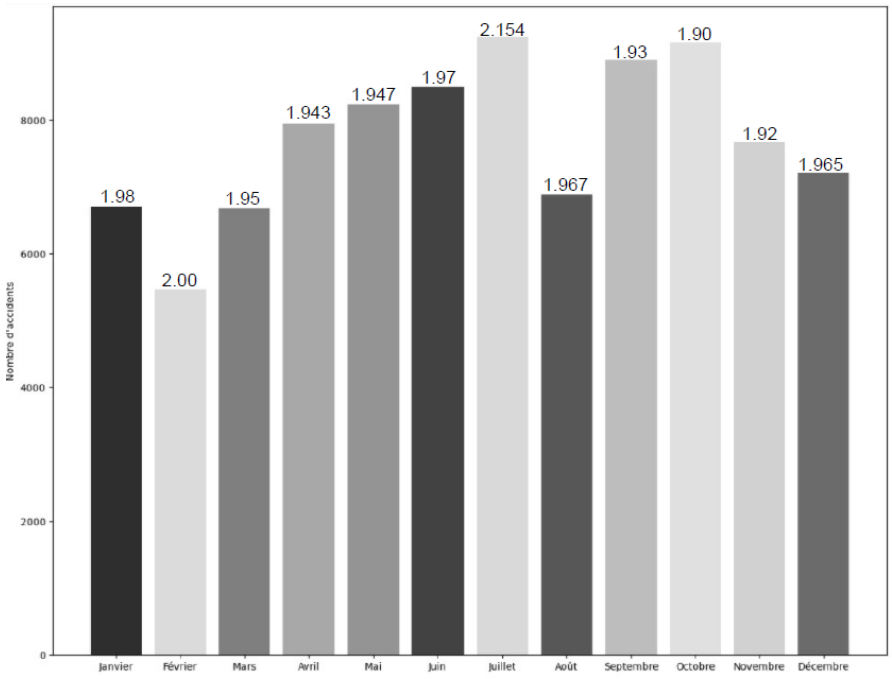
« int »	Intersection : 1 - Hors intersection 2 - Intersection en X 3 - Intersection en T 4 - Intersection en Y 5 - Intersection à plus de 4 branches 6 - Giratoire 7 - Place 8 - Passage à niveau 9 - Autre intersection
« atm »	Conditions atmosphériques : 1 - Normale 2 - Pluie légère 3 - Pluie forte 4 - Neige - grêle 5 - Brouillard - fumée 6 - Vent fort - tempête 7 - Temps éblouissant 8 - Temps couvert 9 - Autre
« catr »	Catégorie de route : 1 - Autoroute 2 - Route Nationale 3 - Route Départementale 4 - Voie Communale 5 - Hors réseau public 6 - Parc de stationnement ouvert à la circulation publique 9 - autre
« vosp »	Signale l'existence d'une voie réservée, indépendamment du fait que l'accident ait lieu ou non sur cette voie. 1 – Piste cyclable 2 – Banquecyclable 3 – Voie réservée
« circ »	Régime de circulation : 1 - A sens unique 2 - Bidirectionnelle 3 - À chaussées séparées 4 - Avec voies d'affectation variable
« plan »	Tracé en plan : 1 - Partie rectiligne 2 - En courbe à gauche 3 - En courbe à droite 4 - En « S »

« surf »	Etat de la surface 1 - normale 2 - mouillée 3 - flaques 4 - inondée 5 - enneigée 6 - boue 7 - verglacée 8 - corps gras - huile 9 - autre
« infra »	Aménagement - Infrastructure : 1 - Souterrain - tunnel 2 - Pont - autopont 3 - Bretelle d'échangeur ou de raccordement 4 - Voie ferrée 5 - Carrefour aménagé 6 - Zone piétonne 7 - Zone de péage
« gps »	Indique la région de l'accident : 1 - Métropole 2 - DOM-TOM
« prof »	Déclivité de la route à l'endroit de l'accident 1 - Plat 2 - Pente 3 - Sommet de côte 4 - Bas de côte
« choc »	Potentiel responsabilité du véhicule : 0 - Non responsable 1 - responsable
« proba_alcool »	Probabilité que l'usager soit sous l'emprise de l'alcool : 0 - Peu de probabilité 1 - plus forte probabilité

Annexe B : comparaison des techniques de traitement de la gravité



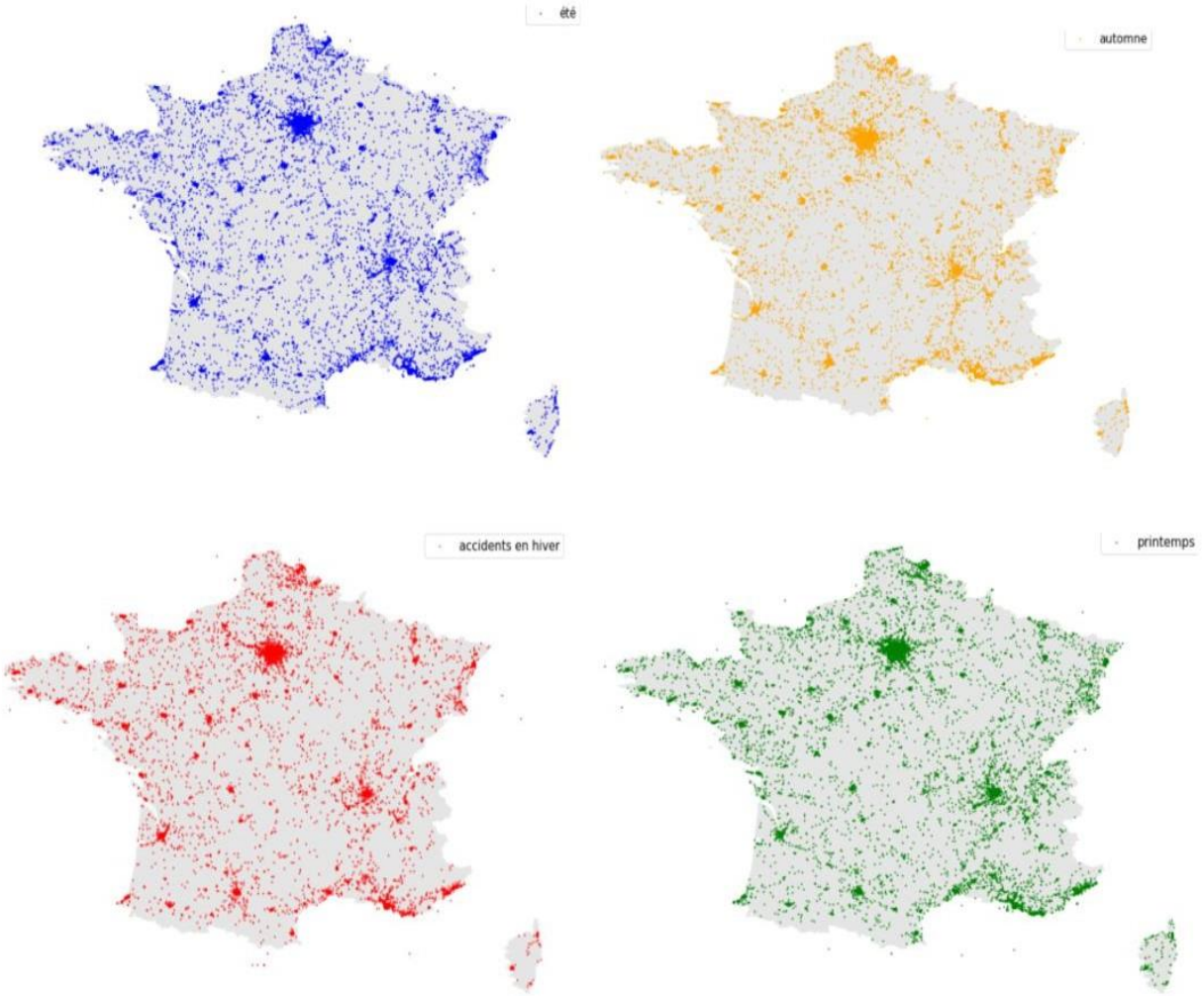
Annexe C : fréquence des accidents et gravité moyenne par jour de la semaine



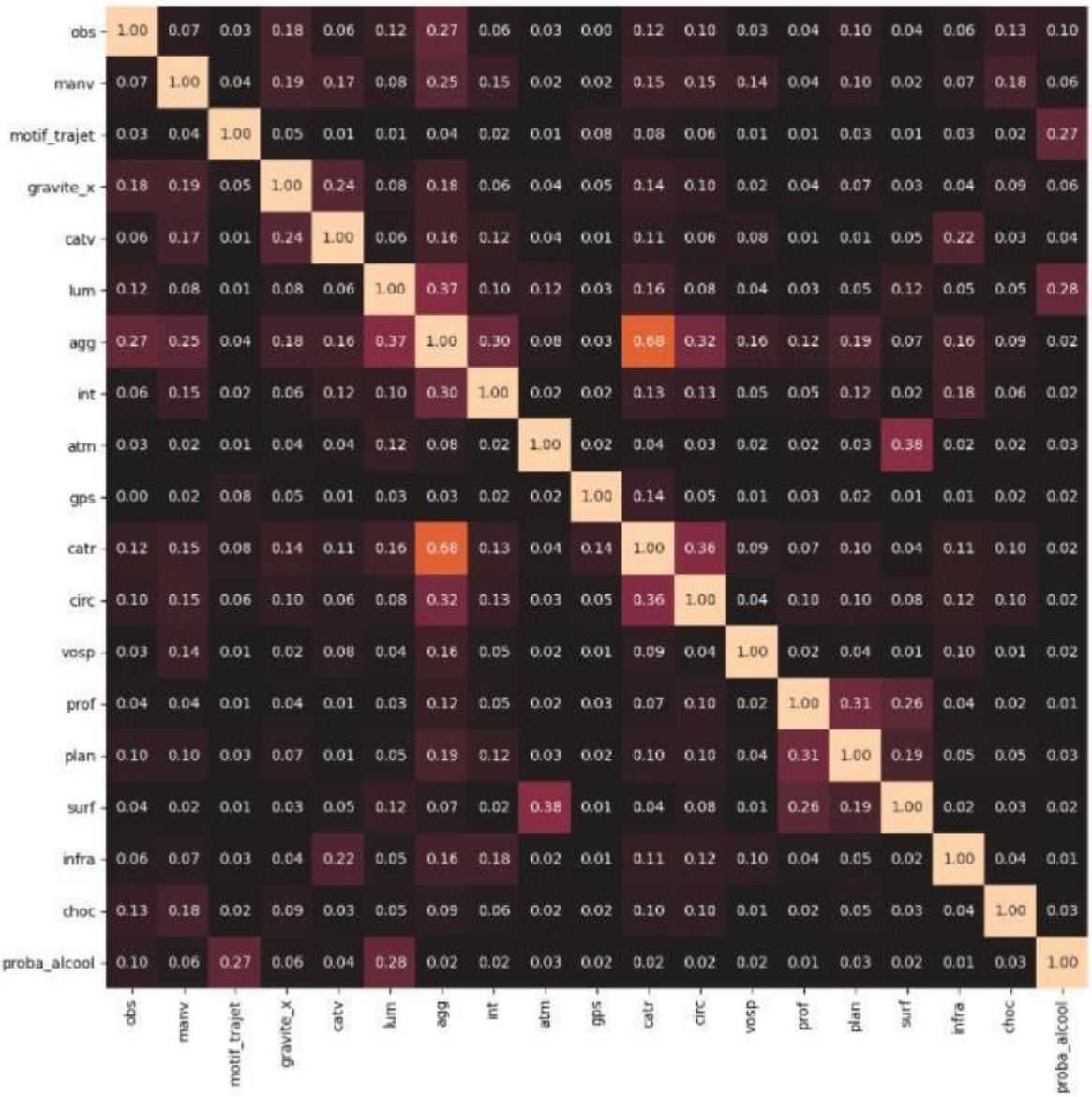
Nombre d'accidents et gravité moyenne selon le mois de l'année.

Annexe D : carte des occurrences d'accidents suivant la saison

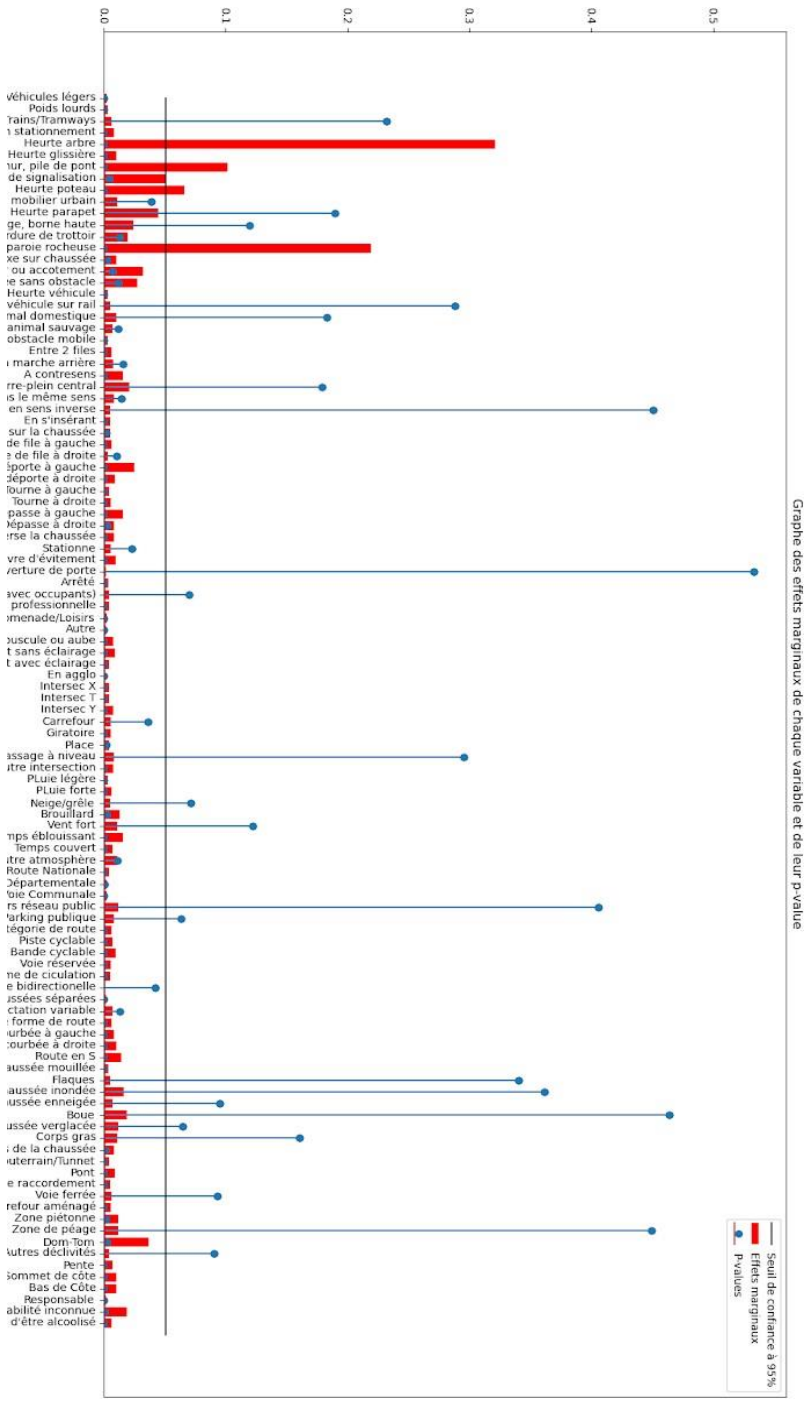
:



Annexe E : matrice de corrélation des variables explicatives



Annexe F : effets marginaux des variables explicatives et leur p-values



Sitographie

- Statistiques de début d'introduction : https://fr.wikipedia.org/wiki/Accident_de_la_route_en_France#:~:text=Les%20accidents%20lors%20de%20l%C3%A9placement%20professionnels%20ont%20tu%C3%A9,travail%20et%20un%20quart%20lors%20d%E2%80%99un%20trajet%20professionnel.
- Schéma de la technique d'arbres de décision : https://www.researchgate.net/figure/Cartoon- representation-of-a-random-forest-classifier_fig1_337361248

Bibliographie

- *Econometric models of road use, accidents, and road investment decisions, Volume II*, Lasse Fridstrøm (1999)
- *Risk factors of road accident severity and the development of a new system for prevention: New insights from China*, Noureddine Benlagha et al. (2020)
- *Risk factors associated with bus accident severity in the United States : A generalized ordered logit model*, Sigal Kaplan et al. (2012)
- *Risk factors affecting the severity of traffic accidents at Shanghai river-crossing tunnel*, Jian John Lu et al. (2015)