Rapport de projet

ING3 - Filière Geo Data Science - UE2 Analyse de données

Analyse des accidents de la circulation : tendances et facteurs déterminants (2019-2023)

Laurie Behloul, Rieulle Brusq, Baptiste Delaunay & Joanna Gosse

13 Janvier 2025

Sommaire

1. Introduction

1.1 Contexte et enjeux des accidents routiers

Les accidents routiers sont un enjeu majeur de santé publique et de sécurité. En 2023, 3 398 personnes ont perdu la vie sur les routes de France. Le nombre de blessés graves est estimé à 16 000 et le nombre de blessés légers ou modérés à 219 000. Au-delà des conséquences humaines et sociales, ces accidents sont synonymes de coûts économiques importants liés aux soins médicaux, à la perte de productivité et aux nombreux dommages matériels.

En France, le suivi et l'analyse des accidents de la route sont indispensables pour comprendre les causes sous-jacentes et identifier les leviers d'amélioration. Ce suivi est assuré par l'observatoire national interministériel de la sécurité routière [1] (**ONISR**). De nombreux efforts ont été déployés pour réduire la fréquence et la gravité de ces accidents, notamment à travers des campagnes de sensibilisation, des politiques de sécurité routière et des innovations dans la conception des infrastructures et des véhicules. Le gouvernement souhaite réduire le nombre de morts et de blessés sur les routes de 50% d'ici 2030.

Dans ce contexte, les données fournies par le **fichier BAAC** (Bulletin d'Analyse des Accidents Corporels) jouent un rôle clé. Ces données sont collectées par les forces de l'ordre après chaque accident corporel et nous offrent une vision détaillée des circonstances, des lieux, des usagers et des véhicules impliqués. Une analyse approfondie de ces données nous a permis d'identifier les facteurs de risque, d'évaluer les mesures de prévention existantes, et de proposer des stratégies adaptées pour améliorer la sécurité routière.

Aujourd'hui lorsqu'un accident corporel a lieu, les pompiers sont envoyés sur le site de l'accident afin d'assurer la prise en charge des blessés, suite à l'appel d'un témoin ou d'une victime. Après avoir évalué la gravité de l'accident, ils décident de solliciter ou non, l'aide de personnels plus adaptés, tels que le SAMU (Service d'Aide Médicale Urgente) par exemple. Nous savons qu'une prise en charge rapide des victimes garantit à celles-ci de plus grandes chances de survie. Ainsi, en déterminant, au moment de l'appel d'urgence, la gravité de l'accident en fonction des caractéristiques décrites, les secours supplémentaires nécessaires pourraient être envoyés plus tôt et le nombre de décès sur les routes pourrait être considérablement réduit. Notre projet s'inscrit donc dans cette volonté d'optimiser la prise en charge des victimes d'accidents corporels dès la réception des appels d'urgence.

1.2 Objectifs de l'étude

L'objectif principal de cette étude est d'analyser les accidents routiers en France et de prédire leur gravité. Nous nous pencherons sur une période s'étendant de 2019 à 2023 afin de comprendre les tendances, d'identifier les facteurs associés à leur fréquence et à leur gravité, et de proposer des recommandations pour réduire ces incidents.

Notre étude vise à:

- Décrire et visualiser les tendances temporelles et spatiales des accidents.
- Identifier les corrélations entre les variables explicatives (conditions météorologiques, éclairage, type de route, etc.) et la gravité des accidents.

- Évaluer les évolutions au fil des années pour détecter les changements significatifs dans les caractéristiques des accidents.
- Utiliser des techniques statistiques et cartographiques pour proposer des analyses précises et exploitables.
- Donner des éléments permettant de prédire la gravité d'un accident en fonction des caractéristiques de celui-ci.
- À terme, fournir un outil d'aide à la décision permettant de déterminer si l'envoi de secours supplémentaires doit être effectué dès l'appel d'urgence.

2. Présentation des Données

2.1 Sources de données

Les analyses réalisées dans cette étude reposent sur les fichiers BAAC administrés par l'ONISR. Ces fichiers couvrent les accidents corporels de la circulation routière entre 2019 et 2023. Ils sont répartis en quatre tables distinctes:

- **Caractéristiques** : ce fichier contient des informations générales sur chaque accident, comme la date, l'heure, les conditions de lumière (jour/nuit) ou encore les conditions atmosphériques (pluie, neige, etc.).
- **Lieux** : cette table fournit des informations sur la localisation géographique des accidents et les caractéristiques du lieu, en spécifiant la catégorie de route, le nombre de voies, les caractéristiques liées à l'environnement, etc.
- **Véhicules** : ce fichier détaille les caractéristiques des véhicules impliqués, comme leur catégorie (voiture, moto, vélo), l'obstacle heurté et le type de motorisation.
- **Usagers** : ce fichier décrit les personnes impliquées dans les accidents, en indiquant leur rôle (conducteur, passager, piéton), leur âge, leur sexe et la gravité des blessures (indemne, blessé, tué).

Il est possible d'effectuer des croisements entre les données grâce au numéro unique d'identification des accidents (**Num_Acc**).

Jointure des tables

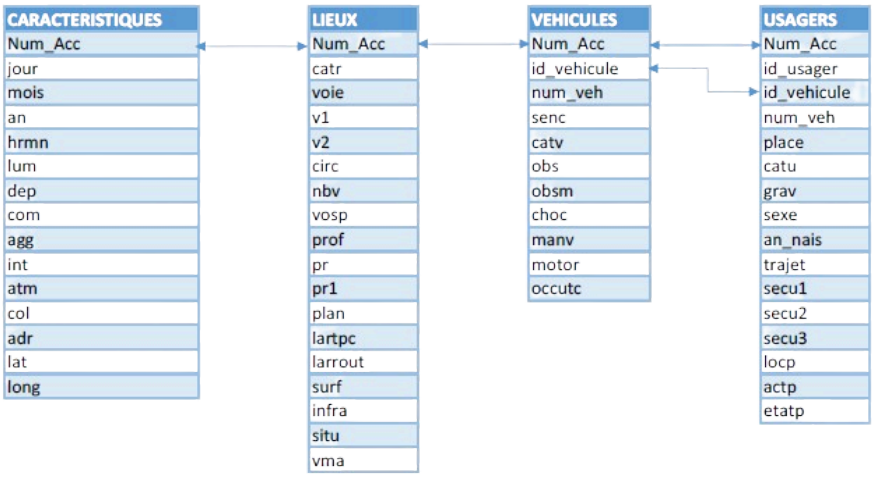


Table 1 : Structure des Tables de Données BAAC

2.2 Description des variables principales

Les données utilisées dans cette étude comportent de nombreuses variables, mais certaines jouent un rôle central dans les analyses. En voici certaines:

- **Variables temporelles** :
  - ***jour*** : jour de l'accident.
  - ***mois*** : mois de l'accident.
  - ***an*** : année de l'accident.
  - ***hrmn*** : heure et minute précises de l'accident.
- **Variables environnementales** :
  - ***lum*** : conditions d'éclairage (plein jour, crépuscule, nuit, etc.).
  - ***atm*** : conditions atmosphériques (pluie, neige, brouillard, etc.).
  - ***agg*** : localisation de l'accident (en agglomération ou hors agglomération).
  - ***catr*** : catégorie de la route (autoroute, route nationale, départementale, etc.).
- **Variables liées aux usagers** :
  - ***catu*** : catégorie de l'usager (conducteur, passager, piéton).
  - ***sexe*** : sexe de l'usager (masculin ou féminin).
  - ***an_nais*** : année de naissance de l'usager.
  - ***grav*** : gravité des blessures (indemne, blessé léger, hospitalisé, tué).
- **Variables liées aux véhicules** :
  - ***catv*** : type de véhicule impliqué (voiture, moto, vélo, etc.).
  - ***motor*** : type de motorisation (hydrocarbure, électrique, etc.).
  - ***vma*** : vitesse maximale autorisée sur le lieu de l'accident.

Ces variables sont à la tête de nos analyses exploratoires, temporelles, spatiales et statistiques pour mieux comprendre les facteurs associés aux accidents graves et leur évolution au fil des années.

La documentation complète de la base de données se trouve dans le dépôt Git : il s'agit du fichier *"Description des bases de données annuelles des accidents corporels de la circulation routière - Années de 2005 à 2023"*.

2.3 Collecte des données

Pour collecter les données des accidents routiers, nous avons développé un script Python pour automatiser le téléchargement des fichiers relatifs au fichier BAAC, disponibles sur le site data.gouv.fr. Ces fichiers comprennent les informations sur les caractéristiques des accidents, les lieux, les véhicules impliqués et les usagers, pour chaque année de 2005 à 2023. Nous nous concentrerons ici sur les données à partir de 2019 car les conditions de renseignement des données du fichier BAAC ont été modifiées en 2018. Un traitement pourrait permettre de les exploiter mais celui-ci est chronophage et donc irréalisable dans le temps du projet.

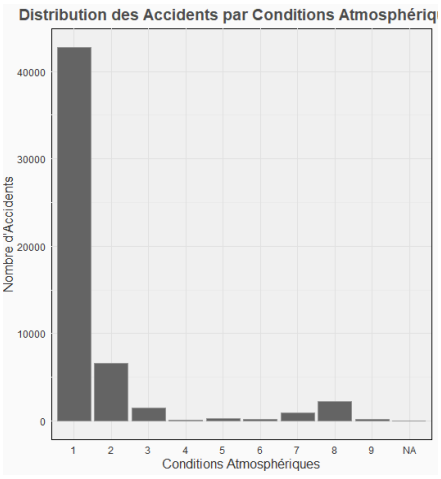
Le script de téléchargement des données repose sur les étapes suivantes :

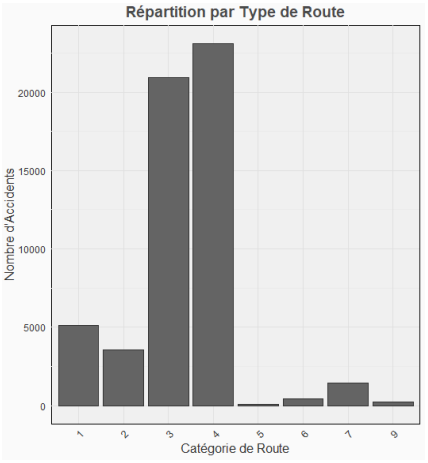
1. **Création d'un répertoire de stockage :**
   Un dossier *data* est créé dans le répertoire courant si celui-ci n'existe pas déjà. Ce dossier est destiné à stocker les fichiers CSV des données téléchargées.
2. **Liste des liens de téléchargement :**
   Les URLs des fichiers à télécharger sont spécifiées dans un fichier CSV local nommé *download_links.csv*. Ce fichier contient une liste des liens associés à chaque année et chaque catégorie de données (caractéristiques, lieux, véhicules, usagers).
3. **Téléchargement automatique :**
   Le script parcourt les URLs du fichier *download_links.csv* et télécharge chaque fichier en vérifiant si celui-ci existe déjà localement. Si un fichier est déjà présent dans le dossier `data`, il est ignoré pour éviter les téléchargements redondants.
4. **Structure du fichier de stockage :**
   Chaque fichier est sauvegardé sous la forme *annee_nom_de_table.csv* (par exemple, *2018_caracteristiques.csv*), garantissant une organisation claire et cohérente des données.
5. **Gestion des erreurs :**
   En cas d'indisponibilité des fichiers ou d'éventuels problèmes liés au téléchargement, un message d'erreur est affiché pour signaler l'anomalie.

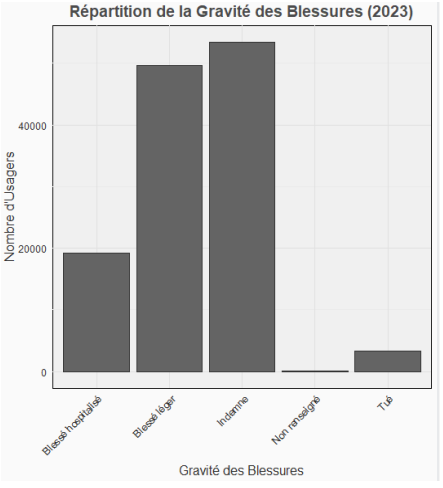
3. Analyse Descriptive

3.1 Statistiques globales des accidents (2019-2023)

Notre objectif est d'analyser les principales caractéristiques des accidents pour avoir une vue globale de la situation routière en France. Pour ce faire, nous avons réalisé une fusion des tables "*caractéristiques*", "*lieux*", "*véhicules*", et "*usagers*" via l'identifiant unique `Num_Acc`. De plus, nous avons agrégé les variables clés à savoir: "gravité des blessures", "conditions atmosphériques", "type de route", "nombre d'usagers" et "nombre de véhicules impliqués" et créé un colonne "Mortalité" indiquant si l'accident est mortel ou non.







Les visualisations ci-dessus offrent un aperçu des données sur les accidents ayant eu lieu en 2023, permettant de mieux comprendre les facteurs clés et les tendances majeures.

Les valeurs prises par l'indice des conditions atmosphériques correspondent aux conditions suivantes :

- NA : non renseigné ;
- 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.

Les valeurs prises par l'indice des catégories de route correspondent aux conditions suivantes :

- 1 : Autoroute ;
- 2 : Route nationale ;
- 3 : Route départementale ;
- 4 : Voie communales ;
- 5 : Hors réseau public ;
- 6 : Parc de stationnement ouvert à la circulation publique ;
- 7 : Routes de métropole urbaine ;
- 9 : Autre.

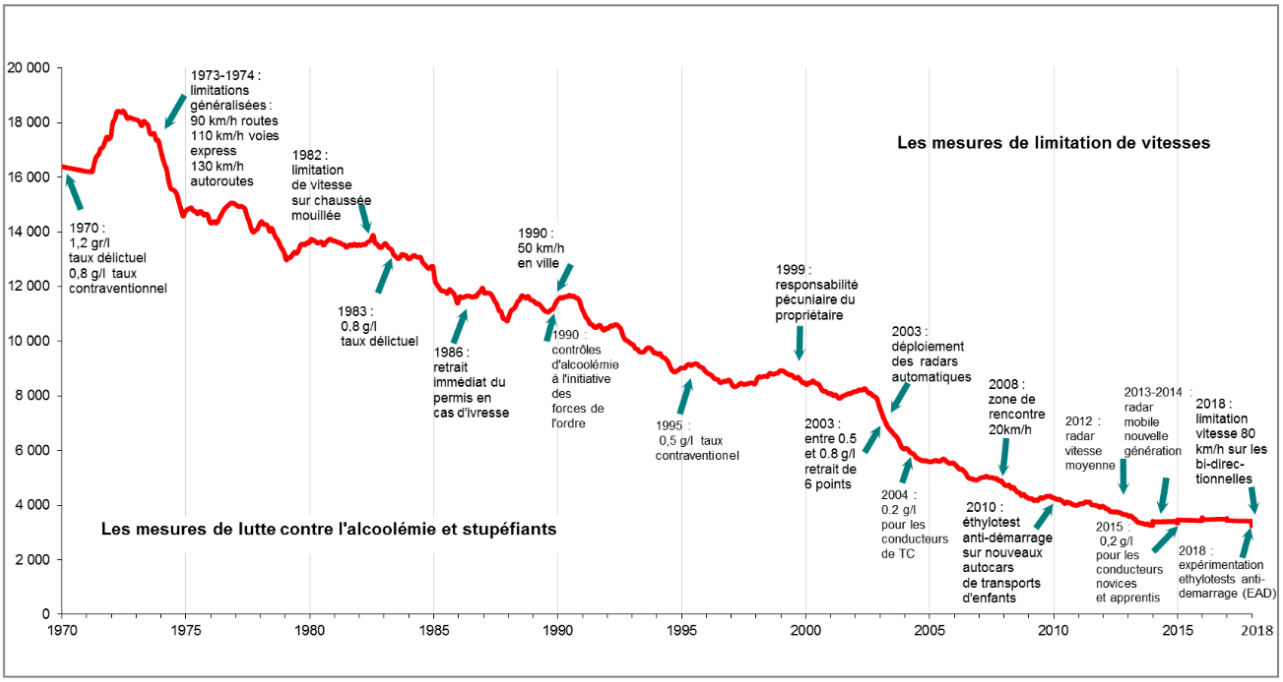
On relève que la majorité des accidents en 2023 ont eu lieu sous de bonnes conditions atmosphériques et sur des routes très fréquentées (types 3 et 4). La plupart des usagers impliqués s'en sont sortis indemnes ou légèrement blessés, les cas graves et mortels restent présents mais minoritaires.

3.2 Tendances temporelles : évolution annuelle et saisonnière

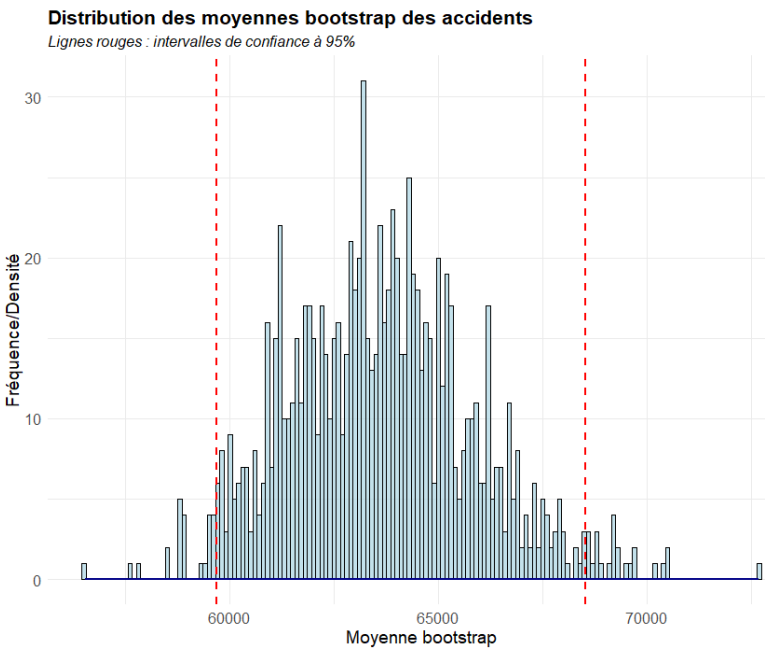
L'analyse temporelle vise à examiner l'évolution du nombre d'accidents au fil des ans et à identifier les schémas saisonniers récurrents.

Ici, nous nous sommes penchés sur le nombre d'accidents routiers entre 2006 et 2023 comme le montre le graphique *Évolution annuelle du nombre d'accidents (2006-2023*) ci-dessous et nous avons relevé une tendance générale à la baisse sur la période avec une diminution progressive du nombre d'accidents notamment à partir de 2007 ce qui peut s'expliquer par de nouvelles réglementations de sécurité routière entrées en vigueur à ce moment là,probablement fin 2007 ou bien 2008 comme on peut le voir sur le graphique ci-dessous avec une nouvelle réglementation à 20 km/h en zone de rencontre. Comme on pouvait s'y attendre, on observe une chute notable en 2020, probablement en raison des restrictions de mobilité liées au COVID-19.

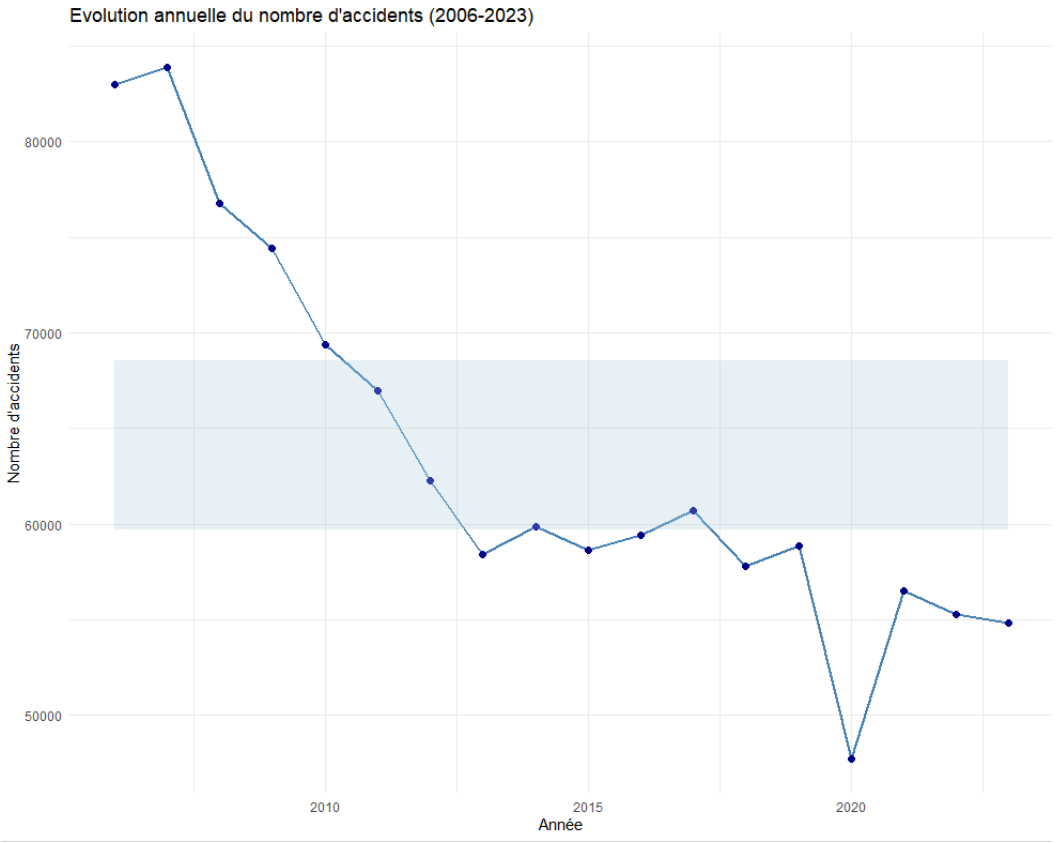
Évolution de la mortalité routière en France métropolitaine et les mesures prises en matière de sécurité 1970 - 2018 (moyenne glissante sur 12 mois)



Source: Site de l'ONISR



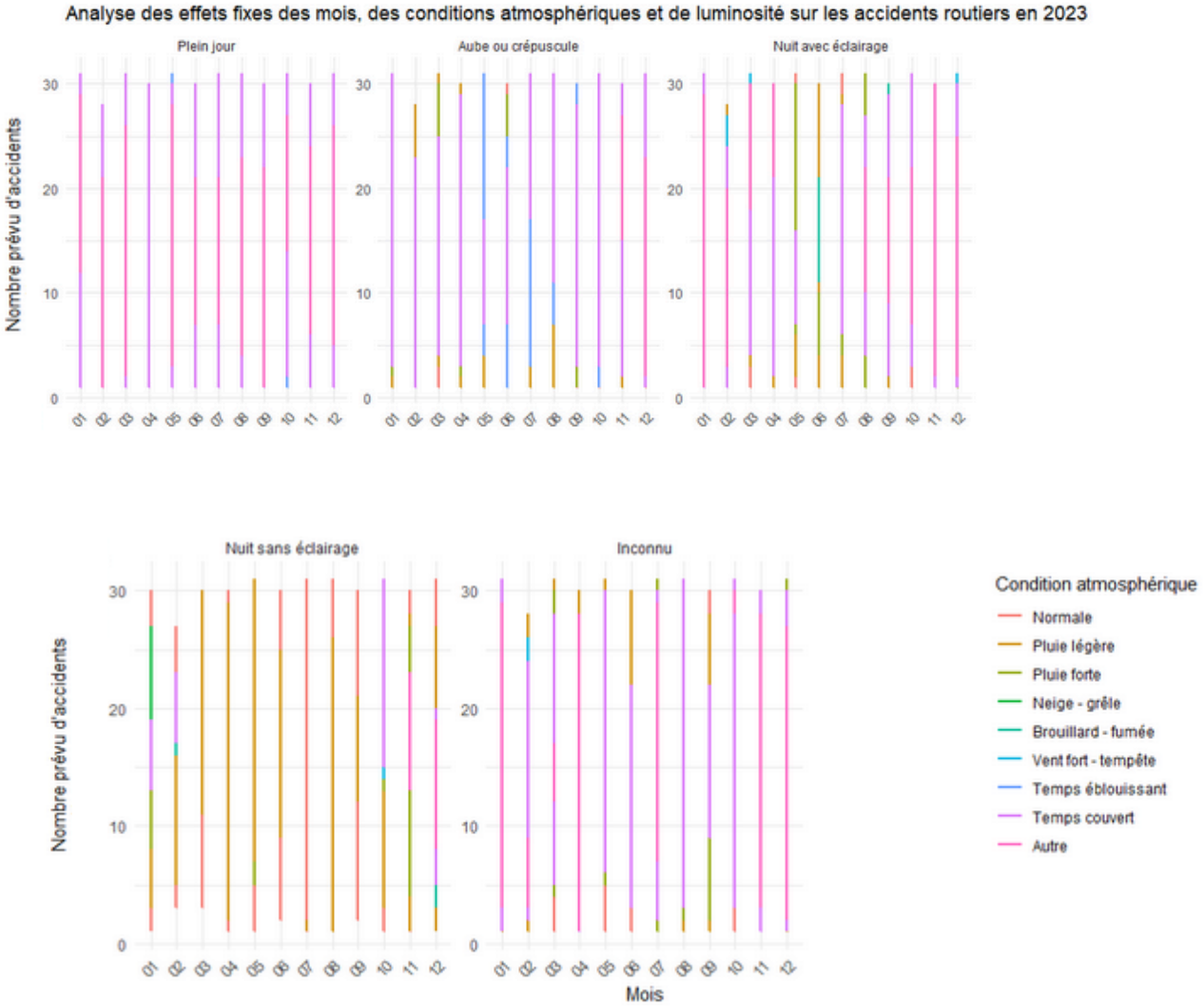
Des bandes de confiance ont été calculées à l'aide de la méthode bootstrap avec 1000 simulations. Ces bandes montrent que la variabilité interannuelle reste relativement limitée. Le pic inversé de 2020 est visuellement distinct mais n'est pas statistiquement significatif selon les tests effectués (p-value = 1). Malgré une baisse marquée, elle ne s'écarte pas de manière significative de la variabilité attendue dans les données donc la diminution générale reflète surement des améliorations structurelles en matière de sécurité routière et l'impact ponctuel de 2020 peut être attribué à un facteur exceptionnel (COVID).



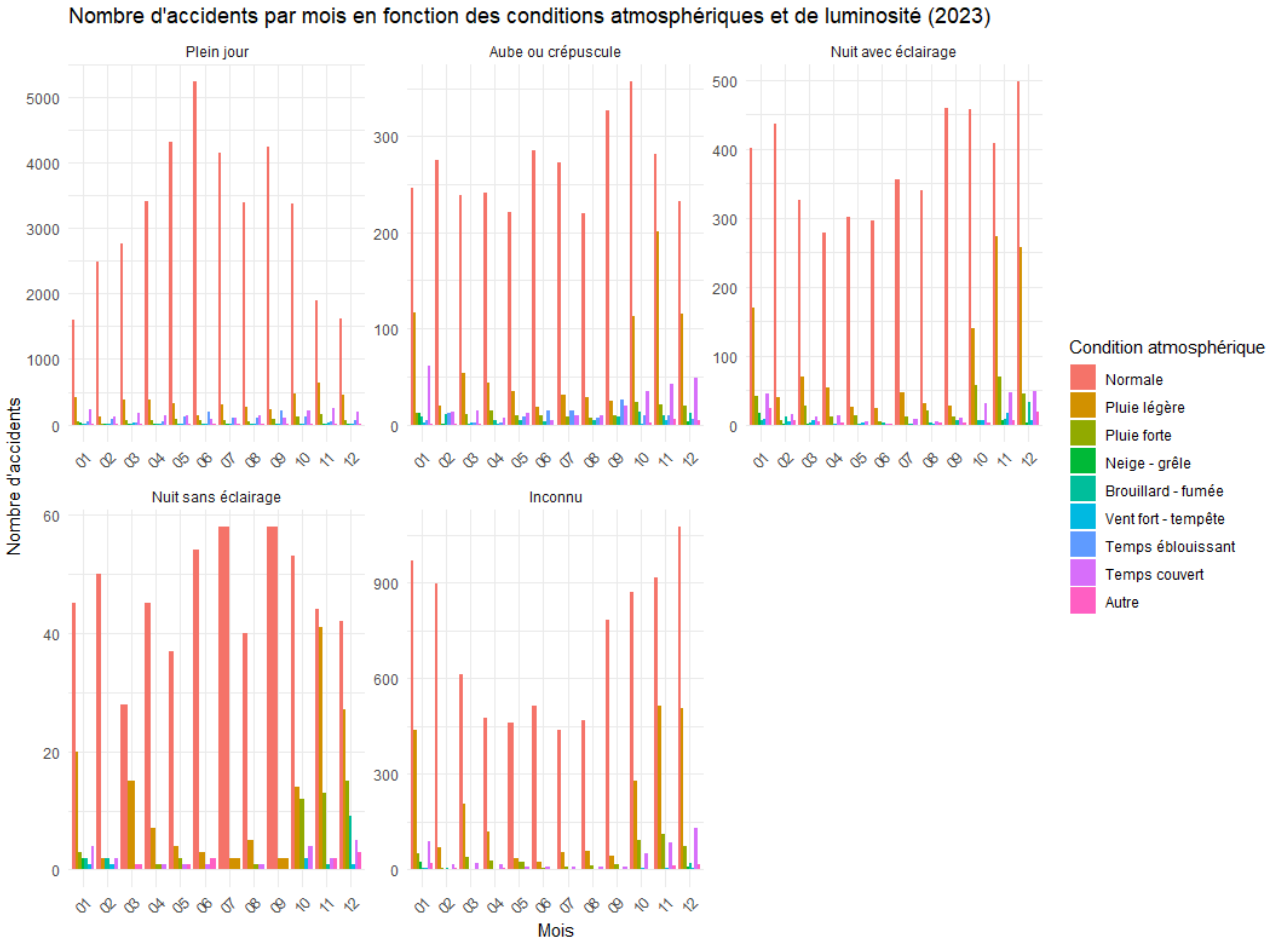
De plus, nous avons mené une nouvelle analyse afin d'explorer l'impact des mois, des conditions atmosphériques (*atm*) et des conditions de luminosité (*lum*) sur les accidents routiers en 2023. Deux approches ont été réalisées : une modélisation des effets fixes avec un modèle mixte et une exploration des distributions par mois.

Pour ceux qui touchent à l'analyse statistique, le modèle fixe (LME) qui comprend les variables "mois", "lum" et "atm" nous a permis de prédire les accidents par jour en tenant compte des caractéristiques atmosphériques et de luminosité. Les estimations des effets fixes montrent une variation importante en fonction des mois avec un pic important en été (de juin à août) et une baisse en hiver. Les conditions "Temps couvert" et "Pluie légère" augmentent le risque d'accident ainsi que "Nuit sans éclairage". En revanche "Temps éblouissant" paraît protecteur!

- Le graphique des effets fixes montre que les mois influencent différemment les accidents en fonction des conditions de luminosité et des conditions atmosphériques. Les accidents sont plus fréquents en plein jour avec une condition atmosphérique normale.



- Pour le graphique des occurrences par mois, on remarque qu'il présente des prépondérances des accidents en plein jour et en conditions normales, mais il y a tout de même des pics sous certaines conditions comme la pluie ou le brouillard.

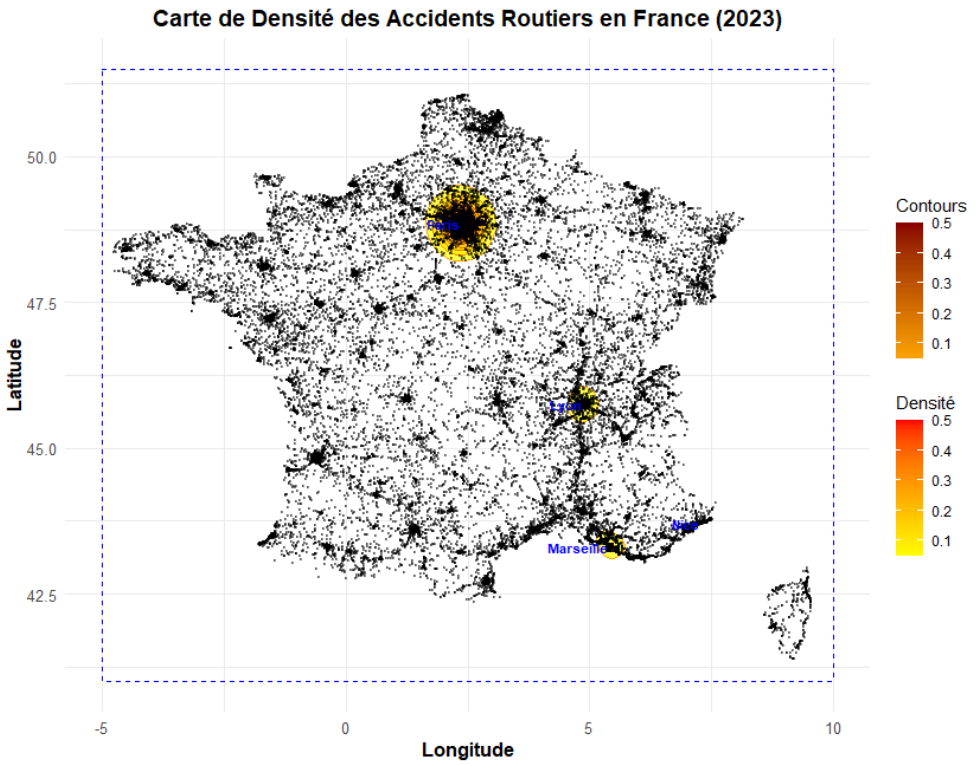


On relève que les effets des paramètres *atm* et *lum* ne sont pas statistiquement significatifs pour expliquer les variations du nombre d'accidents.

3.3 Analyse spatiale : localisation et densité des accidents

La localisation et la densité des accidents est un point essentiel que nous avons choisi d'aborder à travers une **e**stimation par noyau (KDE) pour l'année 2023. Les bornes géographiques sont les suivantes: latitude : 41°-51.5° N et longitude : -5°-10° E pour limiter l'analyse des territoires (France). Les résultats montrent des zones de concentration élevée (zones rouges) notamment autour de Paris, qui constituent un point chaud majeur. D'autres pôles de densité, moins marqués, sont observés à proximité de Lyon et Marseille. Les zones rurales et moins peuplées affichent une faible densité d'accidents et sont représentées par des contours noirs espacés et des zones non colorées.

Nous avons fait le choix d'ajouter une visualisation des contours pour une lecture précise des niveaux de densité. La palette de couleurs (jaune à rouge) hiérarchise visuellement les zones d'intensité. Cette carte permet de repérer des zones critiques en sécurité routière.

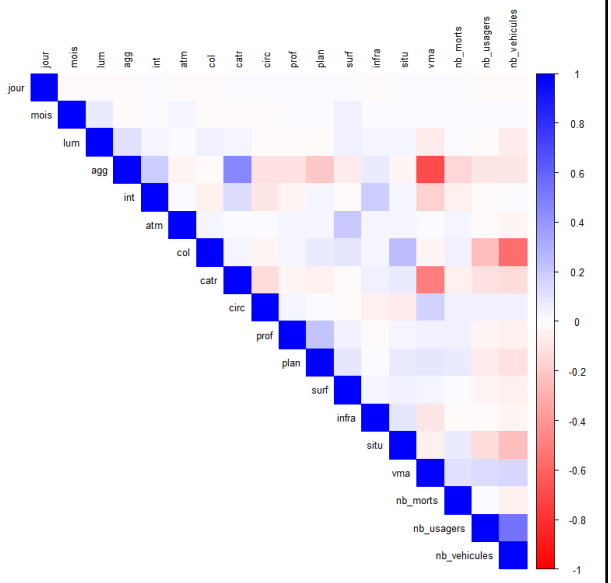
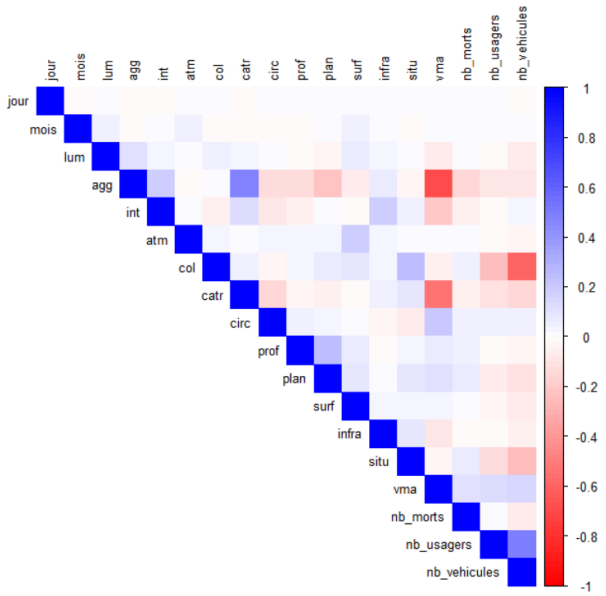


4. Analyse statistique et modélisation

4.1 Analyse des corrélations entre variables

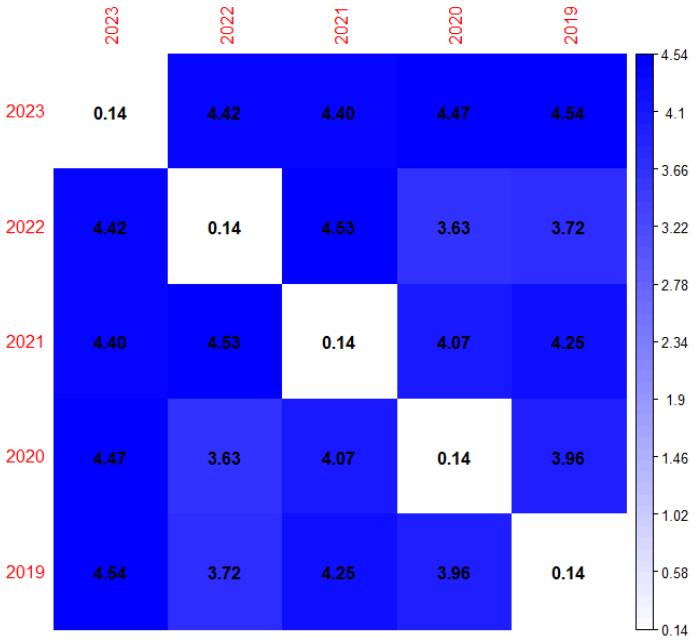
Nous avons calculé une matrice de corrélation pour chaque année (à retrouver **Annexe [A]**) afin d'identifier les associations significatives, les variables redondantes et d'observer la stabilité de ces relations dans le temps.
Les relations entre les variables telles que la luminosité, les caractéristiques de l'accident, les conditions météorologiques, la vitesse maximale autorisée et le nombre de victimes ont été explorées.
Cela nous a permis de mettre l'accent sur :

1. Corrélations importantes : Les variables liées aux caractéristiques des accidents (par exemple, le type de route – *catr*, et le contexte urbain ou non urbain – *agg*) montrent de fortes corrélations avec le nombre de véhicules impliqués (*nb_vehicules*).
2. Relations consistantes : Les relations entre la vitesse maximale autorisée (*vma*) et le nombre de victimes (*nb_morts*) ou d'usagers impliqués (*nb_usagers*) montrent l'impact de la vitesse sur la gravité des accidents.
3. Variations temporelles : Certaines corrélations fluctuent légèrement entre les années, comme celles impliquant les conditions atmosphériques (*atm*) et les infrastructures routières (*infra*).

Analyse des corrélations entre variables (2022)	Analyse des corrélations entre variables (2023)
	

Pour ce qui est de l'analyse globale, nous nous sommes penchés sur la distance de Frobenius qui évalue les différences structurelles entre les matrices de corrélation des données d'accidents routiers de 2019 à 2023. De plus, nous avons réalisé un bootstrap sur 1 000 itérations pour estimer la robustesse des résultats en générant des intervalles de confiance à 95 %.

Les résultats montrent que les distances moyennes sur la diagonale sont faibles (~0.14). La quasi identité des matrices avec elles-mêmes est confirmée, les valeurs ne sont pas totalement nulles en raison du bruit aléatoire du bootstrap. De plus, les distances entre des années adjacentes comme 2023 vs 2022 (4.42) ou 2020 vs 2019 (3.96) sont faibles, donc on obtient une forte stabilité structurelle des relations entre variables sur des périodes proches. Enfin , les distances augmentent pour des années plus éloignées, par exemple, entre 2023 et 2019 (4.54) ce qui indique des différences dans les relations entre variables sûrement dus à des évolutions contextuelles.



L'intervalle de confiance autour des distances moyennes (par exemple, 4.42 entre 2023 et 2022, IC : [4.39, 4.45]) montre que les variations observées ne sont pas issues des fluctuations aléatoires, mais plutôt des différences entre les matrices de corrélation.

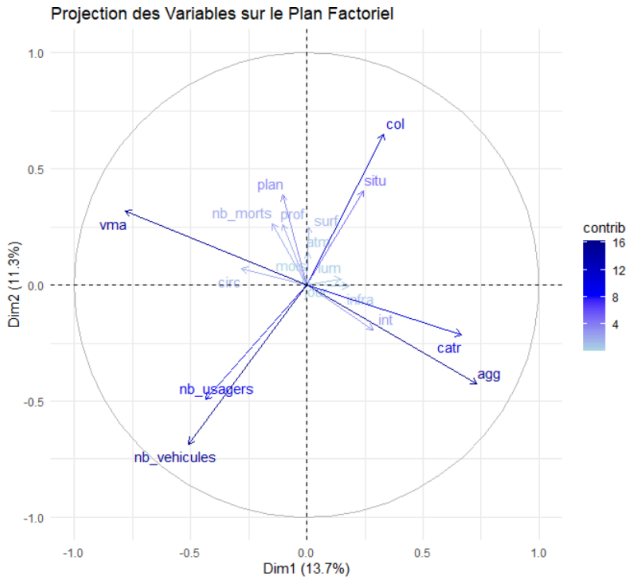
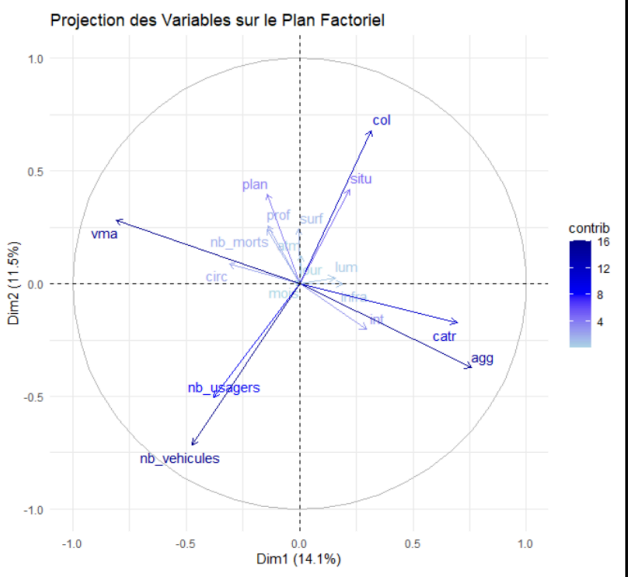
4.2 Analyse en Composantes Principales (ACP)

Nous avons réalisé une analyse en composantes principales pour réduire la dimensionnalité des données tout en préservant un maximum d'information (dont une partie est en **Annexe [B]**). Notre objectif est de relever les variables contribuant le plus à la variabilité des données et de comprendre leur évolution dans le temps.

Nous avons observé les éléments suivants:

- ➢ Variance expliquée :
  - La première composante principale (CP1) explique entre 13.7% (2022) et 14.1% (2023) de la variance selon l'année. Les deux premières composantes cumulées expliquent environ 25% de la variance sur l'ensemble des années.
  - Ces résultats indiquent qu'aucune variable unique ne domine totalement l'explication des données, ce qui reflète la nature multivariée des accidents de la route.
- ➢ Contributions des variables :
  - CP1 : La vitesse maximale autorisée (*vma*), le contexte urbain ou non urbain (*agg*), et le type de route (*catr*) sont les principales variables contributrices, expliquant ensemble environ 30% de la variance sur cette dimension.
  - CP2 : Le nombre de véhicules impliqués (*nb_vehicules*), le type de collision (*col*) et le nombre d'usagers impliqués (*nb_usagers*) dominent sur la deuxième dimension, reflétant des caractéristiques structurelles des accidents.
- ➢ Projection factorielle :
  - Les variables comme *vma* et *nb_vehicules* sont projetées loin du centre sur le plan factoriel, illustrant leur contribution dominante à la variance. D'autres variables, comme *atm* et *surf*, restent plus proches du centre, indiquant une contribution plus faible.
- ➢ Variations temporelles :
  - En comparant les années, les contributions des variables majeures (par exemple, *vma* et *nb_vehicules*) restent stables, mais des légères différences apparaissent dans l'importance relative des variables comme *atm* (conditions atmosphériques) et *col* (type de collision).

Les contributions des variables et des axes principaux montrent une stabilité générale dans les dimensions clés expliquant les accidents de la route, avec seulement de légères variations dans les variables secondaires.

ACP (2022)	ACP (2023)
	

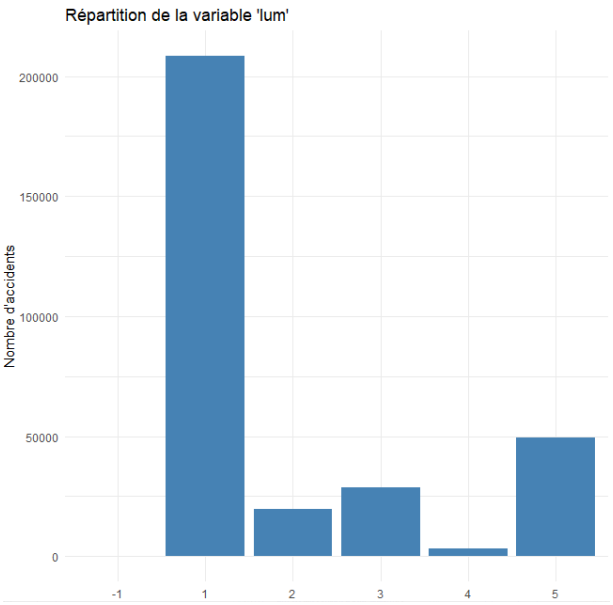
4.3 Étude des relations entre la gravité des accidents et les facteurs explicatifs

L'objectif de cette partie est de s'intéresser aux relations entre la gravité des accidents et les différents facteurs explicatifs à travers différentes analyses statistiques. Nous passerons par deux approches: des tests statistiques pour détecter des relations significatives entre les variables (Chi², Fisher, ANOVA) et une modélisation par régression logistique pour prédire la gravité des accidents.

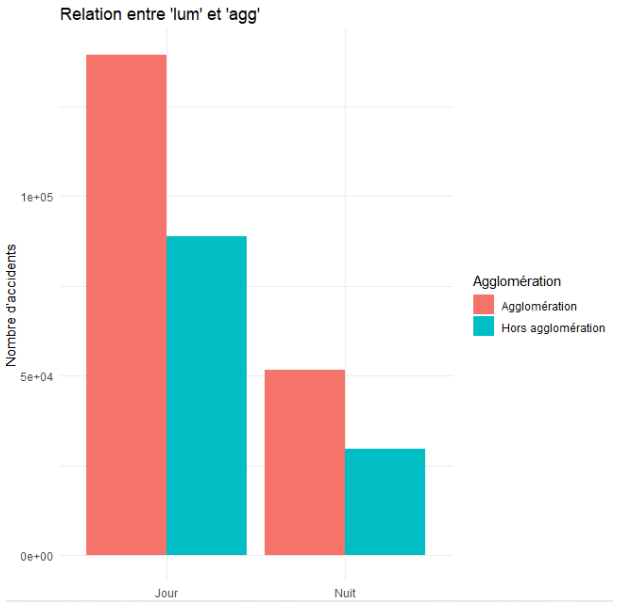
4.3.1 Analyses statistiques

➢ **Test du Chi²**

Le test du Chi² a été utilisé pour examiner la répartition de la luminosité (*lum*) dans les données d'accidents. Les résultats montrent un Chi² = 605009 avec un degré de liberté de 5 et une p-value < 2.2e-16. La distribution de la variable *lum* n'est pas uniforme, certaines conditions de lumière sont significativement plus associées aux accidents.



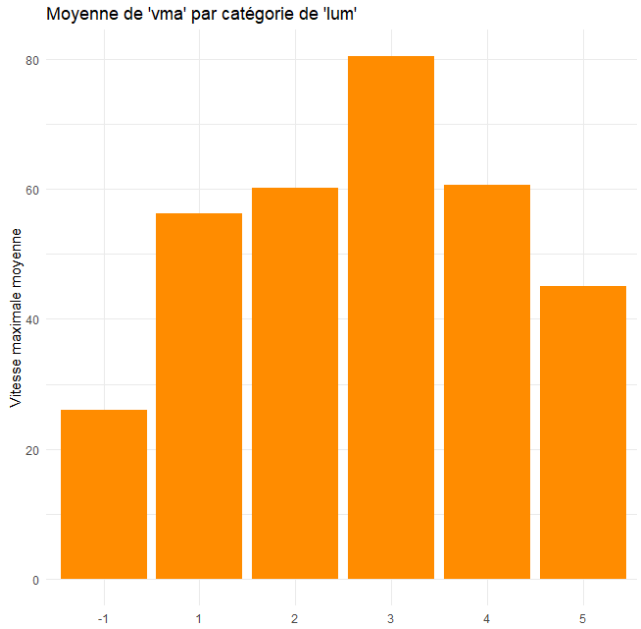
➢ Test de Fisher



Le test exact de Fisher a été réalisé pour analyser la relation entre les conditions de lumière regroupées (Jour/Nuit) et le type d'agglomération (Agglomération/Hors agglomération). La p-value est inférieur à 2.2e-16 donc il y a une association entre ces deux variables. De plus, l'odds ratio estimé à 0.90 (IC 95% : [0.88, 0.92]) suggère qu'il est légèrement moins probable qu'un accident survienne dans une agglomération pendant la nuit.

➢ Analyse de la variance (ANOVA)

L'ANOVA a permis d'évaluer l'impact de la luminosité (*lum*) sur la vitesse maximale autorisée (*vma*). Les résultats montrent un effet significatif avec une p-value inférieur à 2e-16. Les moyennes de « vma » varient selon les catégories de lum, avec par exemple une vitesse maximale moyenne plus élevée dans les conditions de Nuit sans éclairage public *(80 km/h)* comparée aux autres conditions. Les conditions de lumière influencent indirectement les conditions des routes où les accidents se produisent.



Catégorie de lumière	Vitesse maximale moyenne
-1	26 km/h
1	56.3 km/h
2	60.1 km/h
3	80.4 km/h
4	60.6 km/h
5	45.1 km/h

4.3.2 Régression logistique pour prédire la gravité des accidents

Nous avons réalisé une régression logistique pour prédire la gravité des accidents (*grave*, binaire : 1 pour grave, 0 pour non grave) en fonction des facteurs explicatifs tels que *lum*, *atm*, *agg*, *catr*, et *vma*.

➢ Résultats du modèle :
- Variables significatives :
    - Luminosité (*lum*) : p-value = 2.86e-13, effet positif.
    - Type d'agglomération (*agg*) : p-value = < 2e-16, effet négatif (les accidents hors agglomération sont moins graves).
    - Type de route (*catr*) : p-value = 2.40e-14, effet positif.
    - Vitesse maximale autorisée (*vma*) : p-value = 0.091, effet marginalement significatif.

Les variables atmosphériques (*atm*) n'ont pas d'effet significatif sur la gravité des accidents dans ce modèle.

➢ **Évaluation des performances :**
- **Matrice de confusion** :
    - Précision globale : **54.01 %**
    - Sensibilité (capacité à détecter les accidents graves) : **0.30 %**
    - Spécificité (capacité à détecter les accidents non graves) : **99.75 %**

Le modèle montre des performances globales limitées en raison de la faible sensibilité pour identifier les accidents graves. Cela peut s'expliquer par un déséquilibre dans les classes de la variable cible et par des facteurs explicatifs insuffisants pour capturer la complexité de la gravité des accidents.

5. Prédiction de la gravité

5.1 Random Forest

Dans le cadre de notre analyse, nous avons trouvé intéressant d'entraîner un algorithme random forest pour essayer de prédire la gravité d'un accident et plus précisément sa mortalité. De cette façon, les secours pourraient, avant même d'arriver sur les lieux, avoir une idée de la gravité de l'accident pour ainsi savoir s'il serait utile de déployer des moyens supplémentaires ou non (comme un appel anticipé au SAMU). Pour ce faire, nous avons utilisé 23 paramètres qui selon nous, semblaient potentiellement corrélés avec la gravité des accidents:

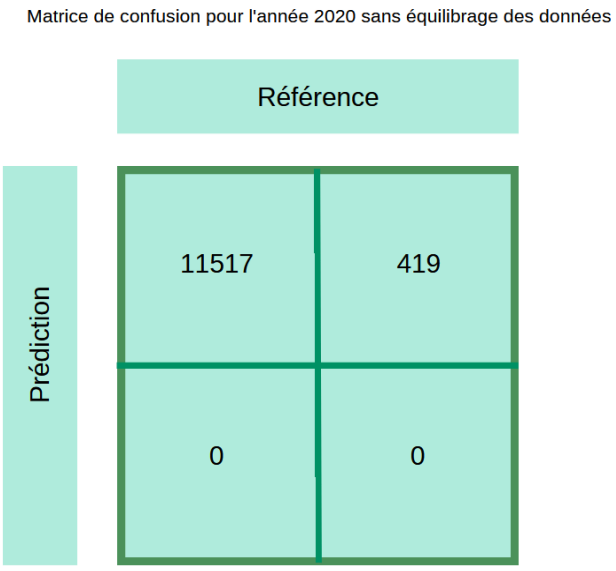
- **Variables temporelles** :
    - ***jour*** : jour de l'accident.
    - ***mois*** : mois de l'accident.
    - ***an*** : année de l'accident.
    - ***hrmn*** : heure et minute précises de l'accident.
- **Variables environnementales** :
    - ***lum*** : conditions d'éclairage (plein jour, crépuscule, nuit, etc.).
    - ***atm*** : conditions atmosphériques (pluie, neige, brouillard, etc.).

- ***agg*** : localisation de l'accident (en agglomération ou hors agglomération).
- ***catr*** : catégorie de la route (autoroute, route nationale, départementale, etc.).
- ***com*** : commune où l'accident a eu lieu.
- ***dep*** : département où l'accident a eu lieu.
- ***int*** : type d'intersection.
- ***col*** : type de collision.
- ***adr*** : adresse du lieu de l'accident
- ***lat***: latitude
- ***long***: longitude
- ***circ*** : voie et régime de circulation
- ***prof*** : profil en long décrit la déclivité de la route à l'endroit de l'accident
- ***plan*** : tracé en plan
- ***surf*** : état de la surface
- ***infra*** : aménagement - infrastructure
- ***situ*** : situation de l'accident
- ***nb_vehicules*** : nombre de véhicules
- ***vma*** : vitesse maximale autorisée sur le lieu de l'accident.

Pour chaque exécution de l'algorithme random forest, nous avons choisi d'utiliser 500 arbres avec une profondeur maximale de 5. Ce choix de paramètres est relativement classique pour cet algorithme. L'objectif était de d'utiliser les bons paramètres afin de ne pas sur-entraîner le modèle et ainsi fausser les résultats.

Pour commencer, nous voulions observer les résultats d'un tel algorithme sur une seule année. Nous avons essayé sur 2020 et sur 2022 en gardant en tête que l'année 2020 pouvait avoir des résultats différents dû à la pandémie du Covid19.

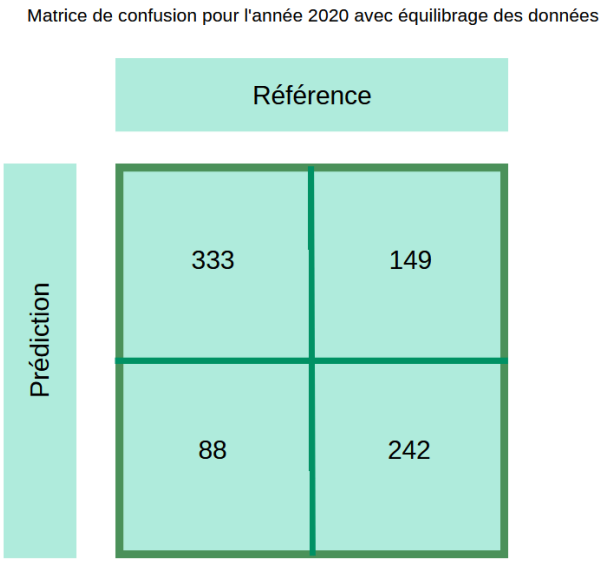
➢ **Année 2020**



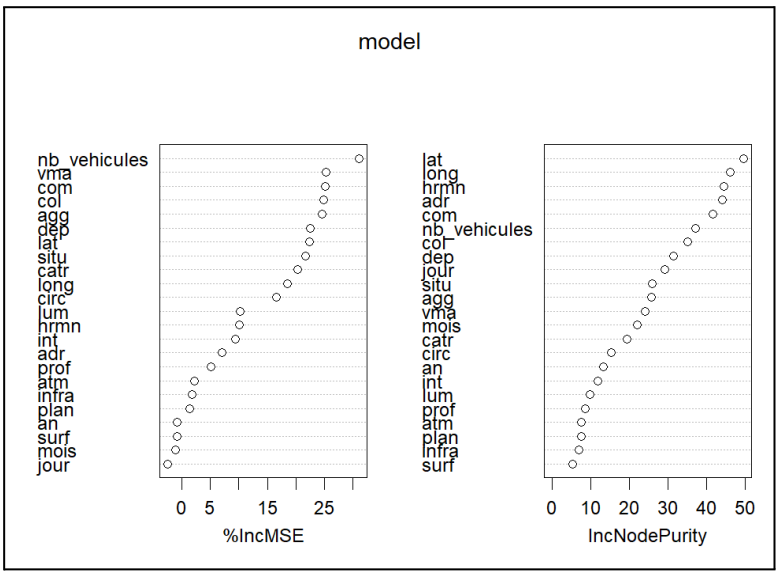
<u>Sans équilibrage des données:</u>

On remarque qu'en effet l'accuracy est d'environ 96% et que le F1 score est extrêmement élevé avec une valeur de 98%. Cependant, lorsqu'on observe plus précisément, on peut voir que la matrice de confusion est plutôt mauvaise. De fait, la sensibilité est excellente (100%) mais la spécificité est de 0 ! Cela est dû au fait que les données ne sont pas du tout équilibrées avec les données des accidents non mortels qui sont beaucoup plus nombreuses que les données des accidents mortels. On aurait aussi pu voir que la valeur de p-value était de 0.513 ce qui nous indiquait que ces résultats ne sont pas statistiquement significatifs et donc ne sont pas dignes de confiance.

Avec équilibrage des données



On peut voir que le F1 score est plus faible à 74% mais le modèle est beaucoup plus fiable avec une valeur de p-value inférieure à 2.2e-16.
Voici l'importance des variables lors de l'exécution de l'algorithme:

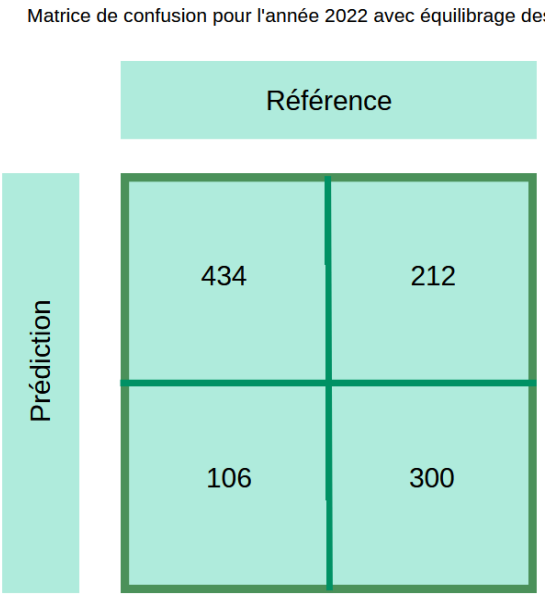


Sur ces deux graphiques on mesure l'impact des variables sur les prédictions, %IncMSE (Pourcentage d'augmentation de l'erreur quadratique moyenne) correspond à la mesure la diminution de la précision du modèle lorsqu'une variable est permutée aléatoirement. Si la permutation d'une variable entraîne une augmentation significative de l'erreur, cela signifie que cette variable est importante pour le modèle, car sa perturbation affecte fortement la performance du modèle. Quant à IncNodePurity, cette mesure évalue l'importance d'une variable en fonction de l'amélioration de la pureté des nœuds des arbres de décision. Plus une variable apparaît dans des splits (division de l'algorithme) qui améliorent fortement la pureté des nœuds, plus son importance est élevée.

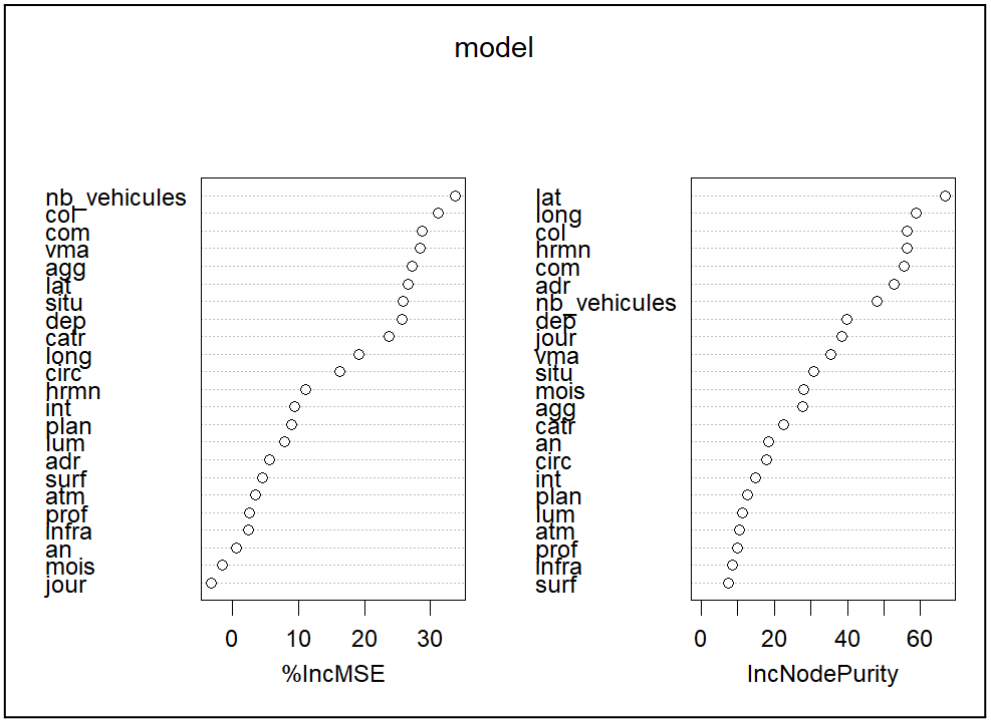
Dans notre cas, on sera plus intéressés par la valeur %IncMSE. Ici, on remarque que ce sont les variables correspondant au nombre de véhicules, à la vitesse maximale autorisée, à la commune où a eu lieu l'accident, au type de collision et à la localisation de l'accident (agglomération ou non) sont celles qui ont le plus influencées la prédiction, avec pour le nombre de véhicules une valeur d'environ 32% d'augmentation de l'erreur quadratique si cette variable était permutée qui est énorme ! On peut par ailleurs noter que ces variables sont les mêmes que celles qui semblaient avoir le plus d'importance lors de l'analyse de notre graphique d'ACP.

- **Année 2022**

Afin de voir s'il y a de grosses différences en fonction des années, nous avons aussi réalisé un random forest pour l'année 2022 (en équilibrant les données).



On obtient un F1 score de 73% ce qui est plutôt proche des résultats qu'on avait obtenu pour l'année 2020. De même que précédemment, la p-value est inférieure à 2.2e-16 ce qui nous permet d'avoir confiance en nos résultats.



On voit que les facteurs déterminants sont les mêmes que pour 2020 sans grande surprise. Cela nous montre que les résultats sont similaires au fil des années. Ainsi, nous avons décidé de rassembler les années 2019, 2020, 2021, 2022 et 2023 afin d'avoir des données plus nombreuses et plus variées.

➢ **Années 2019 à 2023**

Pour ce faire, nous avons entraîné un modèle random forest sur des données d'accident que nous avons récupérées. Malheureusement pour l'étude, et heureusement pour nous, les données ne sont pas équilibrées étant donné qu'il y a très peu d'accidents mortels comparé au nombre total d'accidents. Ainsi dans cette étude, nous avons sélectionné aléatoirement 19168 accidents entre 2019 et 2023 de façon à ce qu'il y ait autant d'accidents mortels que légers. En effet, dans le cas contraire, et nous en avons fait l'expérience au-dessus, la quantité d'accidents non mortels écrase ceux qui le sont et ainsi semble donner de très bons résultats sans que ce ne soit réellement le cas.

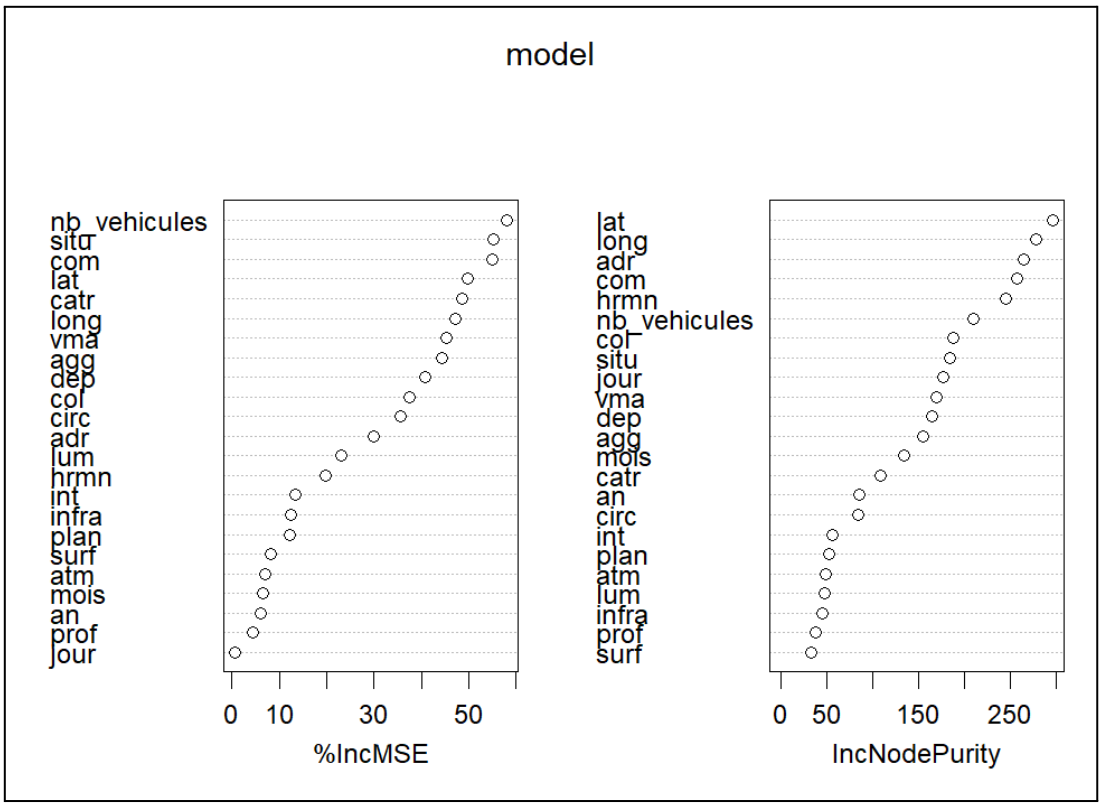
Résultat pour un jeu équilibré de 19168 accidents entre 2019 et 2023 :

Ainsi, on obtient un F1 score de 0.69 et une p-value inférieure à 2.2e-16 ce qui n'est pas trop éloigné de ce qu'on avait obtenu pour les années 2020 et 2022 individuellement. De plus, ces résultats semblent cohérents avec les études que nous avons pu trouver sur ce sujet [1][2].

On remarque que les variables les plus importantes pour ce modèle sont similaires même si ici, le nombre de véhicules, la situation de l'accident et l'identifiant de la commune ressortent vraiment. C'est plutôt curieux pour la commune, cela encouragerait à croire qu'il existe une forte corrélation entre lieu de l'accident et mortalité de celui-ci. Une étude pourrait être réalisée pour vérifier cette hypothèse.

Matrice de confusion avec données de 2019 à 2023 avec équilibrage des données

Prédiction \ Référence		
	1493	441
	900	1958



5.2 Validation croisée

Pour évaluer la performance du modèle et tester le modèle sur plusieurs sous-ensembles de données afin de réduire les biais liés à une seule division des données, nous avons réalisé une validation croisée.

Résultats validation croisée pour une division en 8:

Précision moyenne: 0.703881469115192
Écart-type de la précision: 0.0098740741010412

F1 score moyen: 0.654846303950644
Écart-type du F1 score: 0.0133312915574584

Ces résultats semblent cohérents avec nos résultats précédents et donc valide la stabilité du modèle notamment au vu des faibles écart-type.

5.3 Bootstrap

Le bootstrapping permet d'estimer la variabilité d'un estimateur en créant de multiples échantillons à partir d'un ensemble de données d'origine. Cela aide à comprendre comment un estimateur peut varier en fonction des données échantillonnées, ce qui est crucial pour évaluer la précision de notre modèle. Nous avons choisi le nombre de 25 répétitions.

Moyenne des F1-scores : 0.6684043
Écart-type des F1-scores : 0.008155665
Quantiles des F1-scores :

0%	25%	50%	75%	100%
0.6556319	0.6640647	0.6668521	0.6740454	0.6899945

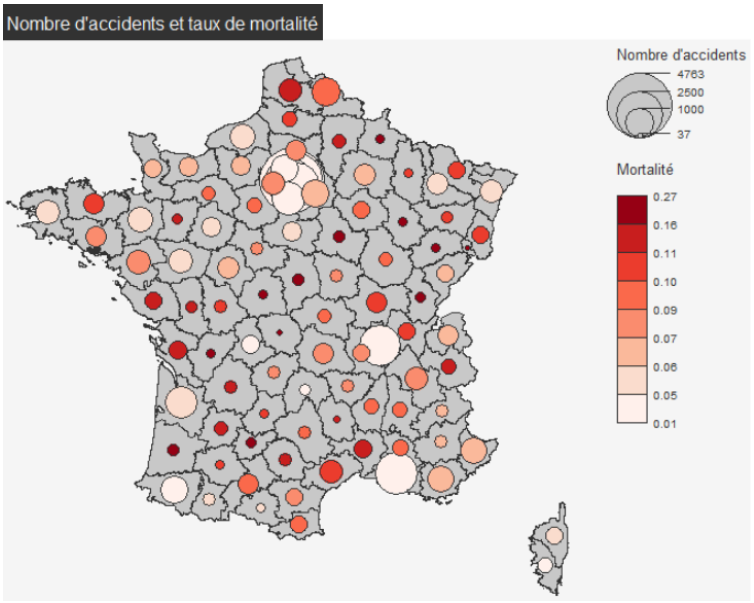
La valeur à 0.67 de la moyenne des F1 scores suggère une performance modérée du modèle et semble plutôt bien correspondre à nos résultats précédents. De plus, le faible écart-type indique une bonne stabilité du modèle à travers les échantillons, de même que la faible différence de résultats entre les quantiles.

6. Visualisation cartographique de la donnée

6.1 Méthode naïve

Dans un premier temps, nous avons calculé des indicateurs tels que la mortalité des accidents et le taux d'accidents se déroulant en agglomération. Ces résultats ont été affichés sur une carte, comme l'on peut le voir ici avec le taux de mortalité et le nombre d'accidents par département.

6.2 Régression géographique pondérée



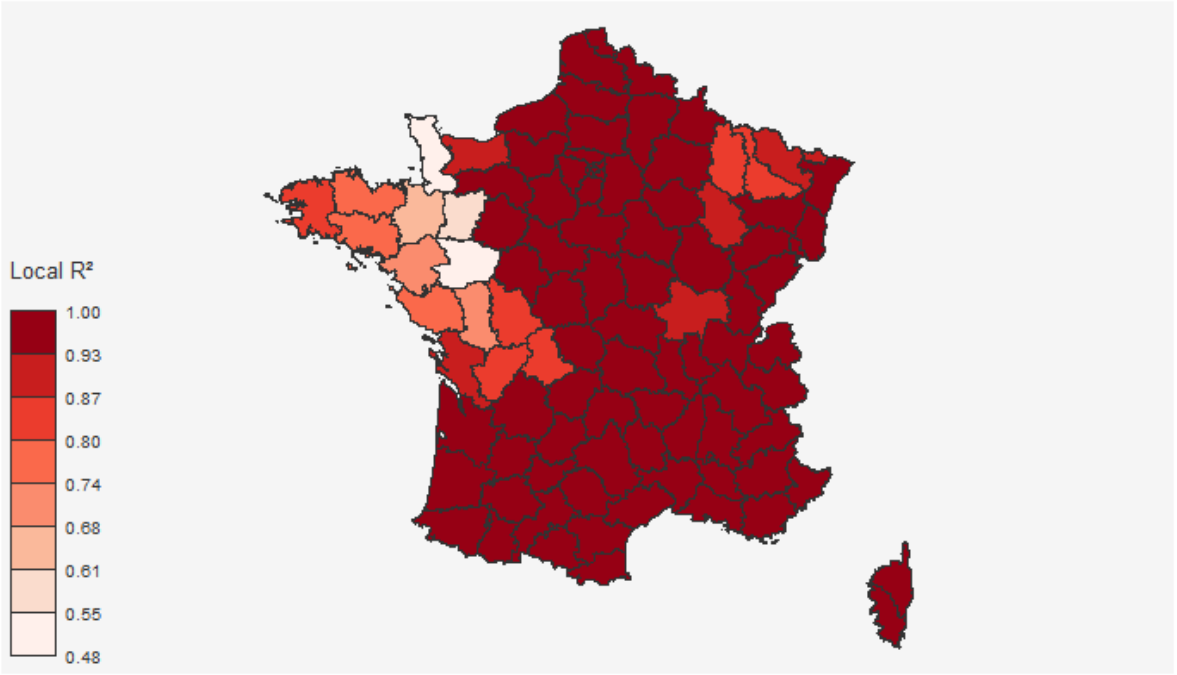
Cependant, cette méthode ne permet pas une analyse très approfondie des variations spatiales de la donnée, car elle repose sur une approche qui ne prend pas en compte les interactions géographiques. La régression géographique pondérée (GWR) est une méthode statistique qui convient mieux pour analyser les relations entre une variable dépendante et une ou plusieurs variables indépendantes tout en tenant compte des variations géographiques. Contrairement à une régression classique où l'on suppose que les relations sont homogènes sur l'ensemble du territoire, la GWR considère que ces relations peuvent différer en fonction des localisations géographiques. Ainsi, chaque observation est pondérée en fonction de sa position géographique, ce qui permet d'obtenir des coefficients de régression spécifiques à chaque zone géographique, cela rend l'analyse plus précise**.**

Dans notre cas, nous avons choisi de comparer la densité des départements français à différentes données (le nombre d'accidents, leur mortalité, le taux d'accidents en agglomération et le nombre de véhicules impliqués).

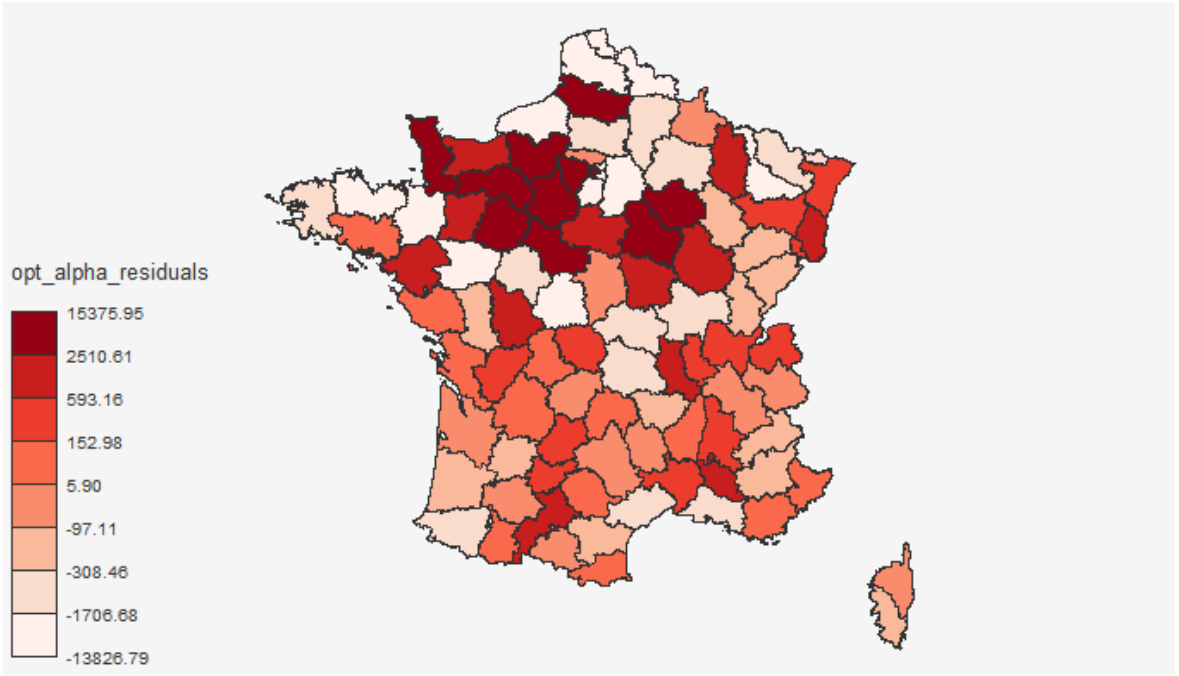
Premièrement, nous avons calculé la valeur optimisée de la bandwith. Nous avons obtenu une bandwith de 38, ce qui signifie que le modèle prendra en compte les 38 plus proches voisins et une valeur d'AICc de 1936.225. Pour être sûr que cette valeur est pertinente, il faudrait la comparer à d'autres modèles.

Ensuite, on utilise la méthode "forward selection" qui permet de conserver uniquement les variables pertinentes. Ainsi, parmi les quatres candidates, le taux d'accidents en agglomération a été supprimé.

Nous obtenons donc une carte qui montre une certaine corrélation (valeur proche de 1) entre la densité de population et le nombre d'accidents, le nombre de véhicules impliqués et la mortalité des accidents. De plus, la carte des résidus a des valeurs qui sont éparpillées (hormis dans le nord-ouest ou un phénomène n'a peut-être pas été pris en compte).



Ces résultats montrent une bonne corrélation entre les les évènements pris en compte et la densité de population départementale. La réalisation d'un modèle de prévision est donc justifiée.



7. Limites et perspectives

Au cours de ce projet, certains éléments ont freiné notre analyse et ont rendu celle-ci moins réaliste :

- Tous les accidents ne sont pas correctement géolocalisés de façon précise. A minima, la commune de l'accident est fournie mais ceci réduit la précision spatiale des analyses.
- Les fichiers BAAC utilisés peuvent contenir des erreurs. La base de données fournie est brute et les erreurs de saisie font l'objet de corrections ultérieurs lorsque des utilisateurs remontent, par courriel, des anomalies qu'ils auraient relevé au cours de l'exploitation des données.
- Les valeurs sous forme d'indice des différents paramètres ne permettent pas de calculer des moyennes ou des écarts-types : ces données ne permettent pas d'estimer les conditions moyennes des accidents car un indice comme 1.5 par exemple ne correspond à aucun paramètre de la documentation.

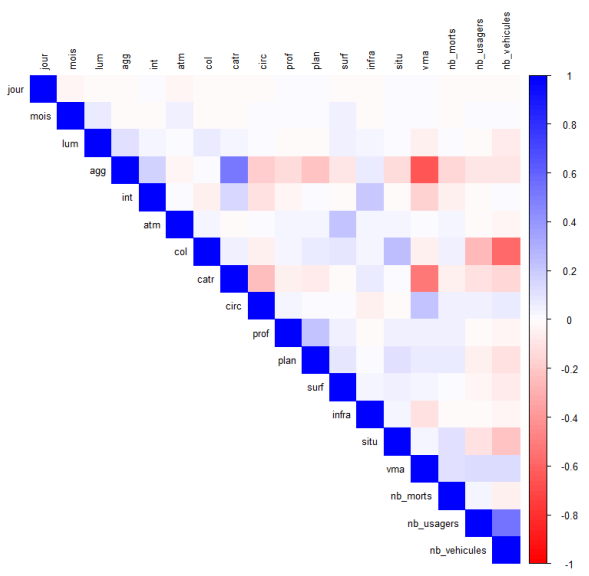
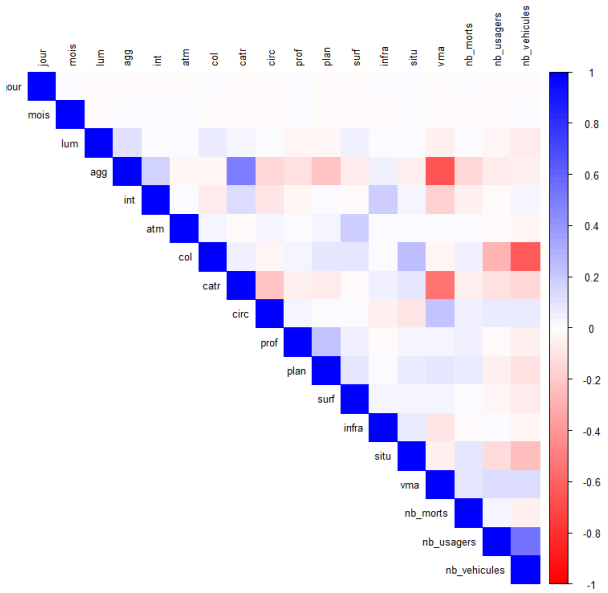
Il aurait aussi été intéressant de refaire tourner un algorithme random forest, seulement avec les variables les plus significatives qui étaient révélées par l'ACP afin de voir si nous aurions obtenu des résultats similaires ou bien beaucoup moins bons.

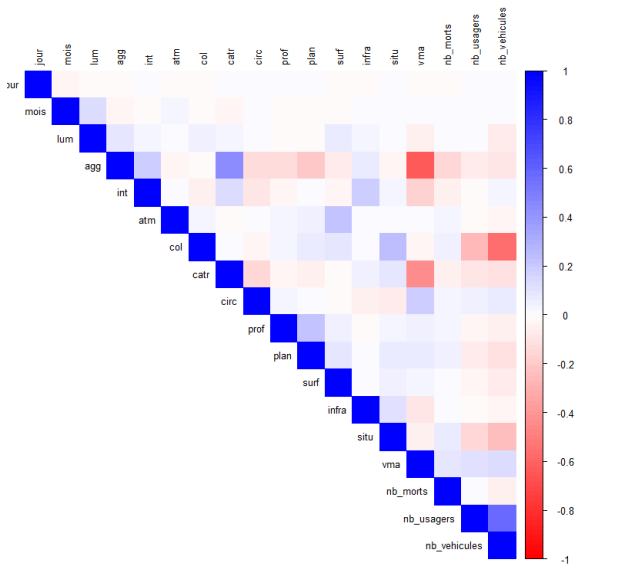
8. Conclusion

Cette étude a permis d'identifier les principaux facteurs influençant la gravité des accidents de la route en France et de développer des modèles prédictifs exploitables. Des variables comme le nombre de véhicules, la vitesse autorisée et le type de route se sont révélées déterminantes. Malgré les limites liées à la qualité et à l'équilibrage des données, les résultats montrent un potentiel prometteur pour améliorer la prise en charge des victimes et la sécurité routière. Les travaux futurs devraient se concentrer sur l'enrichissement des données et l'optimisation des modèles afin de renforcer leur précision et leur utilité pratique.

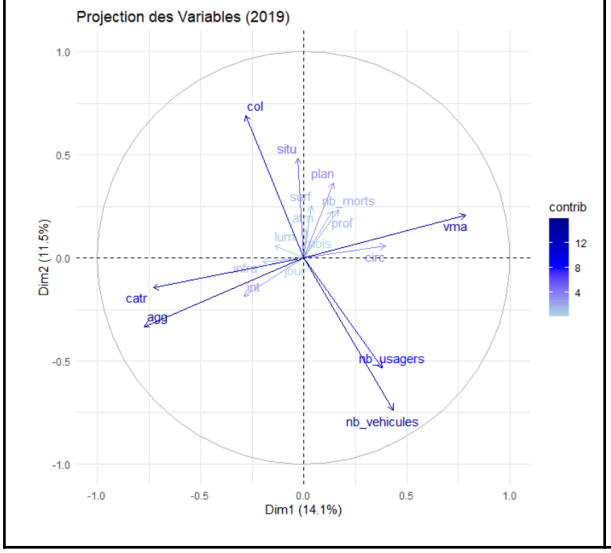
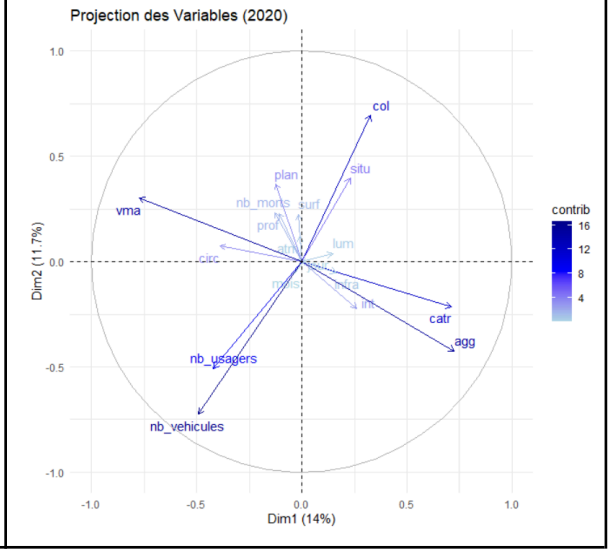
9. Annexes

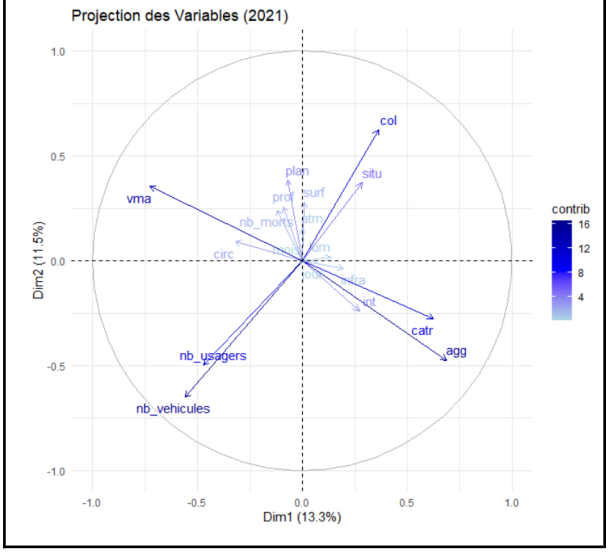
[A]: Analyse des corrélations entre variable 2019 - 2021

Analyse des corrélations entre variables (2019)	Analyse des corrélations entre variables (2020)
	

Analyse des corrélations entre variables (2021)


[B] : ACP de 2019 à 2021

ACP (2019)	ACP (2020)
	

ACP (2021)


Remerciements

Nous tenions à vous adresser nos remerciement les plus sincères pour votre soutien tout au long de ce projet. Vos conseils, votre bienveillance, et votre capacité à nous guider avec gentillesse ont été d'une aide précieuse à chaque étape.

Merci à vous deux d'avoir été là pour nous encourager, nous challenger et partager vos idées. Ce projet a été une expérience enrichissante et mémorable grâce à votre présence et à vos encouragements. Vous avez été des mentors formidables, et nous vous en sommes très reconnaissants.

Nous souhaitons un merveilleux congé paternité à Yann Meneroux– que cette période soit remplie de bonheur et de moments précieux en famille. Et à Juste Raimbault, nous adressons nos vœux de réussite pour la suite, en espérant avoir l'occasion de vous recroiser!

Références

[1]Al-Mamlook, R., Abouchabaka, J., & Al-Gadhib, A. (2020).
Comparaison des algorithmes d'apprentissage automatique pour la prédiction de la gravité des accidents de la route.
Disponible sur ResearchGate.

[2] Bulut, A. (2020).
Prédiction de la gravité des accidents de la route à l'aide d'algorithmes d'apprentissage automatique.
DergiPark Journal. Disponible sur DergiPark.

[3] Cybergeo: Revue européenne de géographie (2018).
Statistiques spatiales des accidents de la route.
Disponible sur Cybergeo.

[4] Université Gustave Eiffel, LASTIG. (2025).
Projet d'analyse des données – ENSG Geo Data Science UE2.
Dépôt disponible sur GitHub.