ENSIMAG / Grenoble INP

Modélisation de la sévérité et de la fréquence des accidents routiers

Rapport du projet Mathématiques Appliquées

Réalisé par :
Younes Chennouf
Ilyes Kebairi
Muhamad Zaiinizee Bin Daivin
Kylian Anaclet

Encadré par :
M. Christophe Dutang

Année universitaire 2025-2026

Date du rendu : 18 Decembre 2025

Table des matières

1 Introduction et Problématique

La sécurité routière constitue un enjeu majeur de santé publique. Si les décennies passées ont vu le nombre de tués sur les routes diminuer drastiquement depuis les années 1980 grâce aux progrès technologiques et réglementaires, cette tendance positive s'essouffle aujourd'hui.

Face à cette évolution complexe, l'analyse des données de l'ONISR (Observatoire National Interministériel de la Sécurité Routière) pour l'année 2021 vise à dépasser le simple constat descriptif. Ce projet s'articule ainsi autour de la problématique suivante :

> *« Quels modèles statistiques permettent d'expliquer la gravité des accidents, d'en identifier les typologies récurrentes et de prédire leur fréquence ? »*

Notre démarche méthodologique, réalisée sous R, s'articule ainsi en trois temps : la modélisation linéaire de la sévérité via la construction d'un score dédié, l'identification de profils types d'accidents par des méthodes de partitionnement, et enfin l'estimation de leur fréquence (lois Binomiale ou Poisson) en tenant compte de l'exposition au trafic.

2 Modèle de Régression Linéaire

2.1 Définition du Score de Sévérité

L'objectif du modèle de régression est de déterminer les facteurs qui ont le plus d'impact sur la sévérité d'un accident de la route en France en 2021.
Notre définition du Score de Sévérité est une pondération du nombre de personnes indemne, blessé, hospitalisé et tué lors de l'accident. Concrètement :

$$Score\_Severite = \alpha \times nb\_Tue + \beta \times nb\_Hospitalise + \gamma \times nb\_Blesse + \delta \times nb\_Indemne.$$

Avec α, β, γ, δ des coefficients choisis de manière à refléter la gravité relative des blessures. Pour ce faire, nous nous sommes inspiré de l'échelle AIS (Abbreviated Injury Scale) qui quantifie la gravité des lésions. On obtient ainsi : $\alpha = 6$, $\beta = 3$, $\gamma = 1$, $\delta = 0$. Ces coefficients attribuent un poids 6 fois plus important aux décès qu'aux blessés légers, et 3 fois plus aux blessés hospitalisés. Nous considérons que le nombre de personnes indemne ne doit pas avoir d'influence sur ce score car selon nous, son nombre n'est pas lié à la sévérité d'un accident.

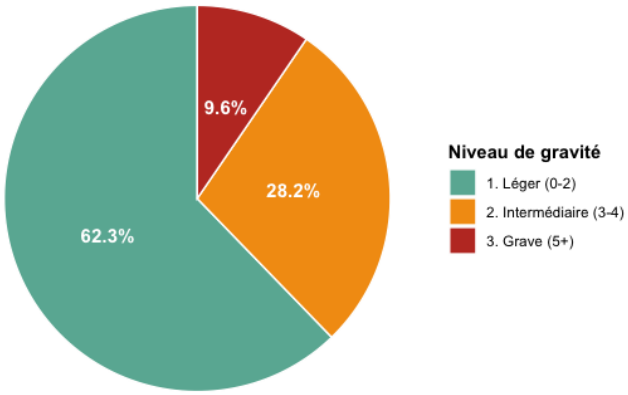

FIGURE 1 – Répartition des accidents par Score de Sévérité (Données ONISR 2021)

La Figure 1 met en évidence une distribution asymétrique du score de sévérité. En effet, 62.3% des accidents ont un score de sévérité compris entre 0 et 2 ; et 9.6% des accidents ont un score de sévérité

supérieur ou égal à 5. De même, selon le résumé de cette variable sur R, 50% des accidents ont un score de sévérité inférieur ou égal à 1 et le score de sévérité moyen des accidents est de 2. Cette caractéristique du score de sévérité motivera l'utilisation d'une transformation logarithmique sur celui-ci dans la suite de l'analyse.

2.2 Nettoyage des données

Dans la base de données ONISR 2021, un certain nombre de variables ne sont pas pertinentes pour notre modèle de régression pour différentes raisons. Les variables d'identification (comme *'Num_Acc'*, identifiant unique de l'accident) n'apportent aucune information explicative. Certaines variables possèdent un nombre trop élevé de modalités, comme *'voie'* qui correspond au numéro de la route, ce qui compliquerait l'interprétation et la stabilité du modèle. Enfin, d'autres variables présentent une proportion importantes de valeurs manquantes, notamment *'lartpc'* et *'larrrout'*, correspondant aux largeurs du terre-plein central et de la chaussée.
Ces variables ont donc été exclues de l'analyse. Pour les variables présentant trop de modalités, des regroupements ont été effectués afin de réduire la dimension du problème. Les départements ont ainsi été agrégés en régions (14 modalités, dont une pour les Outre-Mer), et les heures ont été regroupées en moments de la journée (Nuit, Matin (Pointe), Journée (Creuse), Soir (Pointe), Soirée).
Les variables qualitatives sont traitées comme des facteurs dans R, chaque modalité étant comparée à une modalité de référence. Le choix de cette référence est essentiel pour l'interprétation des coefficients. Nous avons sélectionné des modalités correspondant à des situations standards ou fréquentes, supposées peu spécifiques en termes de gravité des accidents. Par exemple, la modalité 'Plein Jour' est choisie comme référence pour les conditions d'éclairage (*'lum'*), et la modalité 'Normale' pour les conditions atmosphériques (*'atm'*).

2.3 Choix du Modèle de Régression

Nous avons construit trois modèles de régression successifs. Le premier modèle explique le score de sévérité en fonction de 21 variables explicatives, telles que la vitesse maximale autorisée, la situation de l'accident ou encore le nombre de véhicules impliqués. Ce modèle permet d'expliquer environ **19,5%** de la variabilité du score de sévérité.
Cependant, les hypothèses du modèle linéaire ne sont pas entièrement vérifiées. Le QQ-plot des résidus montre un écart à la normalité dans les quantiles extrêmes, et le graphique des résidus en fonction des valeurs ajustées met en évidence une dispersion plus importante pour les grandes valeurs prédites. Par ailleurs, le graphique *Residuals vs Leverage* fait apparaître un point à fort levier. Après vérification, ce point correspond à un accident de sévérité faible (score égal à 1) et son retrait n'a pas d'impact sur le coefficient de détermination R^2. Afin d'améliorer l'adéquation du modèle, nous avons appliqué une transformation logarithmique à la variable réponse dans un second modèle :

$$\log(\text{Score\_Severite} + 1).$$

le terme +1 permettant d'éviter les problèmes liés aux valeurs nulles. Cette transformation réduit l'influence des valeurs extrêmes et rapproche la distribution de la variable réponse d'une loi normale. Le coefficient de détermination augmente alors à $\boldsymbol{R^2 = 23{,}6\%}$, soit une amélioration de 3,1 points par rapport au premier modèle. Les hypothèses de normalité des résidus apparaissent également mieux respectées au vu du QQ-plot.
Enfin, dans un troisième modèle, nous avons retiré trois variables (*'surf', 'prof' et 'circ'*) dont les coefficients n'étaient pas statistiquement significatifs au seuil de 5% selon les tests de Student (hypothèse nulle de coefficient nul non rejetée). Ce choix permet de simplifier le modèle sans dégrader ses performances, le coefficient de détermination restant proche de celui du second modèle, avec $\boldsymbol{R^2 = 23{,}3\%}$.

2.4 Interprétation des résultats

2.4.1 Les facteurs les plus influents

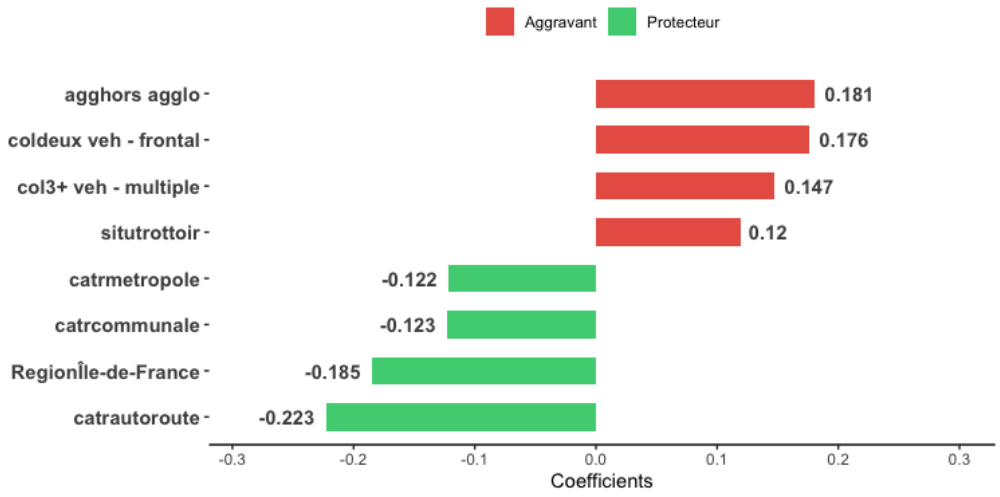

FIGURE 2 – Facteurs les plus influents sur la gravité d'un accident

Selon la Figure 2, nous observons que l'autoroute présente l'un des effets protecteurs les plus marqués sur la sévérité des accidents : toutes choses égales par ailleurs, les accidents sur autoroute sont en moyenne 20% [1] moins graves que ceux sur une route départementale. Cet apparent paradoxe peut s'expliquer par la conception sécurisée des autoroutes : absence d'intersections, séparation des voies, trafic plus homogène et comportement plus vigilant des usagers.
La région Île-de-France présente également un effet protecteur important, les accidents y étant en moyenne 17% moins graves que ceux survenus en Auvergne–Rhône–Alpes, région de référence. Cela peut s'expliquer par des vitesses de circulation plus faibles, une urbanisation plus dense ou encore un accès plus rapide aux services de secours.
Par ailleurs, les accidents hors agglomération sont en moyenne 20% [2] plus graves que ceux survenus en agglomération. Cet effet aggravant, l'un des plus importants parmi les variables considérées, peut notamment s'expliquer par des vitesses de circulation plus élevées. Le type de collision joue également un rôle significatif : une collision frontale entre deux véhicules augmente le score de sévérité d'environ 19% par rapport aux accidents sans collision.
Enfin, bien qu'il n'apparaisse pas dans la Figure 2 car il s'agit d'une variable quantitative (et non catégorielle), son impact reste significatif : la présence d'un camion supplémentaire impliqué dans un accident augmente le score de sévérité d'environ 10%.
Important : *Ces interprétations sont effectuées toutes choses égales par ailleurs et la gravité de l'accident est évaluée uniquement à partir du Score de Sévérité défini en Section 1.1.*

2.4.2 Limites

Finalement, bien que nous arrivons à mettre en évidence les facteurs les plus influents dans le cadre de nos données, notre modèle présente un pouvoir explicatif modeste ($\boldsymbol{R^2 = 23{,}6\%}$), reflétant la complexité multifactorielle de l'accidentologie. Ainsi, dans le cadre de notre modèle, plusieurs limites sont à prendre en compte.
Premièrement, les données ONISR ne contiennent pas d'information sur l'alcoolémie, l'usage du téléphone, ou la fatigue qui peuvent être des facteurs primordiaux à prendre en compte dans la sévérité d'un accident. Ensuite, notre modèle identifie des associations, mais ne peut établir de relations causales

1. Calcul : $(e^{-0.223} - 1) \times 100 \approx -20\%$
2. Calcul : $(e^{0.181} - 1) \times 100 \approx 20\%$

strictes entre les variables et la sévérité d'un accident. Enfin, bien que la transformation logarithmique améliore l'adéquation du modèle aux hypothèses de régression linéaire, la normalité des résidus ainsi que sa variance constante ne sont pas totalement vérifiés, pouvant altérer les tests statistiques.

3 Méthode de partitionnement : K-prototypes

Nous allons désormais nous concentrer sur la méthode de partitionnement [1] dans le but de regrouper les profils d'accidents. L'algorithme K-prototypes combine la distance euclidienne pour les variables numériques et une mesure de dissimilarité pour les variables catégorielles. Il minimise la fonction objective suivante :

$$\min_{c_1,\ldots,c_K} \sum_{k=1}^{K} \sum_{i,G(i)=k} \left[d_{num}(x_i^{num}, c_k^{num}) + \gamma \cdot d_{cat}(x_i^{cat}, c_k^{cat}) \right]$$

où d_{num} et d_{cat} représentent respectivement les distances pour les variables numériques et catégorielles, et γ équilibre leur importance.

L'algorithme répète trois étapes jusqu'à convergence. Chaque observation est assignée au cluster le plus proche. Les prototypes sont recalculés avec la moyenne pour les variables numériques et le mode pour les catégorielles. Le processus s'arrête lorsque les clusters ne changent plus.

3.1 Préparation des données

Afin d'optimiser le temps de calcul, l'analyse repose sur un échantillonnage aléatoire de n=5000 observations et une sélection rigoureuse de variables contextuelles (localisation, conditions d'éclairage, conditions atmosphériques, type de collision, vitesse maximale autorisée), excluant les indicateurs de gravité ou de localisation. Les variables numériques ont ensuite été standardisées par centrage-réduction pour assurer une contribution équilibrée dans les mesures de distance, tandis que les variables catégorielles ont été conservées telles quelles.

3.2 Diagnostic du regroupement et détermination du nombre optimal de clusters

3.2.1 Évaluation de la tendance au regroupement

Avant le partitionnement, nous devons vérifier que les données ont naturellement tendance à former des groupes plutôt qu'à être distribuées uniformément. Cette vérification utilise la statistique de Hopkins [3] :

$$H = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

où x_i mesure la distance entre un point réel et son plus proche voisin, et y_i la distance entre un point généré aléatoirement et son plus proche voisin. Les scores obtenus, très proches de 1, confirment que les données présentent une structure de regroupement marquée.

3.2.2 Choix du nombre de clusters k

Dans cette étude, nous analysons la variabilité du nombre d'accidents corporels par unité d'exposition afin d'identifier des facteurs de risque structurels. Pour cela, nous comparons deux modèles de fréquences issus des données de l'ONISR, qui diffèrent par leur échelle géographique et le facteur d'exposition utilisé. La détermination du nombre optimal de clusters s'est appuyée sur la méthode du coude [3] et l'analyse de la silhouette. L'étude conjointe de ces indicateurs a conduit au choix de $k = 3$ classes, un compromis idéal assurant une distinction nette des profils sans segmentation excessive (stabilisation de l'inertie et coefficient de silhouette satisfaisant).

Note : *Bien que la composition exacte des groupes puisse légèrement varier selon l'échantillon, les profils principaux restent stables, ce qui confirme la pertinence du découpage obtenu.*

3.3 Résultats et caractérisation des clusters

3.3.1 Distribution des observations

Le partitionnement final montre une répartition inégale entre les trois clusters, reflétant la fréquence des différents types d'accidents. Cette distribution permet d'identifier à la fois les accidents les plus courants et ceux qui présentent des caractéristiques moins fréquentes.

3.3.2 Profils types identifiés

Cluster 1 : Accidents Routiers Hors Agglomération Diurnes : Ce groupe rassemble les accidents hors agglomération survenant principalement de jour ou la nuit sans éclairage public, en conditions atmosphériques normales. Les accidents présentent des configurations variées (collisions latérales, par l'arrière, ou sans collision) sur des routes à vitesse élevée.

Cluster 2 : Accidents Routiers à Haute Cinétique : Ce groupe concentre les accidents hors agglomération sur routes rapides et autoroutes à vitesse maximale élevée (87 km/h en moyenne), survenant fréquemment la nuit sans éclairage ou en plein jour. Les configurations d'accidents sont diverses, incluant des collisions en chaîne et par l'arrière, caractéristiques des voies à forte circulation.

Cluster 3 : Accidents Urbains Nocturnes : Ce groupe, majoritairement en agglomération, regroupe les accidents survenant principalement la nuit en zone éclairée, à vitesse modérée (42-49 km/h). Les collisions latérales dominent, typiques des intersections urbaines, dans des conditions atmosphériques généralement normales.

3.3.3 Qualité et validation du clustering

L'évaluation quantitative du modèle, avec un coefficient de silhouette moyen de 0.2677, atteste de la pertinence de la segmentation malgré la complexité intrinsèque des données d'accidentologie. Ce résultat positif confirme que le modèle capture une structure réelle, même si la faible valeur de l'indice de Dunn (0.001133) rappelle que les frontières entre les types d'accidents sont naturellement fluides plutôt que strictes. Dans le détail, la meilleure performance du Cluster 2 (0.3107) valide notre capacité à isoler nettement le profil routier à haute cinétique sur les voies rapides hors agglomération. Le Cluster 1 présente un score intermédiaire (0.2861), reflétant la diversité des configurations d'accidents hors agglomération en conditions diurnes. Le Cluster 3, avec le score le plus faible (0.2277), traduit la complexité du milieu urbain nocturne où les contextes d'accidents aux intersections se chevauchent naturellement, sans nuire à la cohérence de l'interprétation globale.

3.3.4 Visualisation et séparation

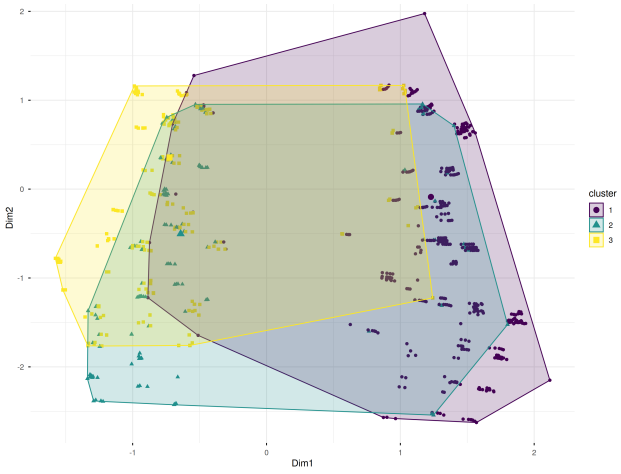

FIGURE 3 – Visualisation des clusters obtenus

La projection des clusters (Figure 3) en deux dimensions par ACP permet de visualiser le partitionnement. Les chevauchements observés sont dus à la réduction dimensionnelle, mais dans l'espace d'origine, les clusters sont bien mieux séparés.

4 Analyse sur la fréquence des accidents

4.1 Une double approche de la modélisation

Dans cette étude, nous analysons la variabilité du nombre d'accidents corporels par unité d'exposition afin d'identifier des facteurs de risque structurels. Pour cela, nous comparons deux modèles de fréquences issus des données de l'ONISR, qui diffèrent par leur échelle géographique et le facteur d'exposition utilisé.

Nous avons d'abord mis en œuvre une **approche généraliste** à l'échelle départementale (France entière), où le facteur d'exposition retenu est la population issue des données de l'INSEE [2] (dataset `Df_Pop_brut`). Cette méthode, correspondant au fichier `Frequence.Rmd`, permet d'analyser un risque d'accident rapporté à la population, relevant ainsi d'une lecture de type santé publique. Parallèlement, nous avons développé une **approche de précision** à l'échelle communale, restreinte à la région Auvergne-Rhône-Alpes. Ici, le facteur d'exposition est le Trafic Moyen Journalier Annuel (TMJA) [4] issu des comptages routiers (dataset `Df_TMJA_brut`). Cette seconde approche (`Frequence_TMJA.Rmd`) vise à estimer un risque rapporté à l'usage réel de l'infrastructure routière, offrant une lecture plus fine du risque routier.

À l'issue de ces deux approches, les données sont structurées sous la forme d'un jeu de données final (`df_final`), regroupant les informations de l'ONISR ainsi que le nombre d'accidents normalisé par le facteur d'exposition retenu, servant de base à la modélisation statistique.

4.2 Méthodologie : de Poisson à la Binomiale Négative

4.2.1 Modèles linéaires généralisés et variables explicatives

Pour modéliser la variable de comptage Y (nombre d'accidents), nous utilisons le cadre des Modèles Linéaires Généralisés (GLM). Ce type de modèle relie l'espérance de la variable cible $E(Y)$ à une combinaison linéaire des variables explicatives X_i via une fonction de lien logarithmique. L'équation générale s'écrit [5] :

$$\log(E(Y)) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \text{offset}.$$

Dans cette formulation, β_0 représente la constante (*intercept*) correspondant au niveau de risque de base, tandis que chaque coefficient β_i mesure l'influence spécifique de la variable explicative X_i associée (telle que la pluie ou le type de route) sur la fréquence attendue des accidents. L'offset est un terme dont le coefficient est contraint à 1 ; il est fondamental ici car il permet de passer d'une modélisation en « nombre brut » à une modélisation en « taux ». Ainsi, selon l'approche retenue, l'exposition est représentée soit par la population (log(Pop)), soit par le trafic (log(TMJA)).

4.2.2 Justification du modèle : De Poisson à la Binomiale Négative

Dans un premier temps, nous avons modélisé la fréquence des accidents par une régression de Poisson. L'évaluation de la pertinence de ce modèle repose sur deux indicateurs statistiques extraits des résultats. Le premier est la Déviance [5], qui quantifie l'écart entre les prédictions du modèle et les observations réelles pour mesurer l'ajustement global. Le second est l'AIC (Akaike Information Criterion), un critère de comparaison permettant de sélectionner le modèle offrant le meilleur compromis entre précision et complexité ; plus l'AIC est faible, meilleur est le modèle.

La loi de Poisson repose sur une hypothèse restrictive d'équidispersion, imposant que la variance soit égale à l'espérance ($Var(Y) = E(Y)$). Or, nos données présentent une variance nettement supérieure

à la moyenne, caractérisant une surdispersion qui suggère que des facteurs hétérogènes influencent fortement l'accidentologie. L'utilisation stricte de Poisson risquant de sous-estimer le risque d'erreur, nous avons opté pour la Loi Binomiale Négative [5]. Ce modèle généralise celui de Poisson en ajoutant un paramètre de dispersion θ, permettant de capter correctement cette variabilité excessive.

4.3 Résultats et Interprétation

4.3.1 Approche Généraliste (Population) : La preuve par la Déviance

Dans cette pemière approche, nous avons pris comme variable explicative : la pluie, les routes hors agglomération ainsi que les autouroutes et les départementales, les virages, la vitesse moyenne sur l'ensemble des départements en utlisant la population comme offset.

Les résultats obtenus avec la loi de Poisson révèlent une inadéquation majeure avec les données. La déviance résiduelle atteint une valeur critique de **6695**, bien supérieure au nombre de degrés de liberté. Bien que le modèle Poisson réussisse à expliquer les accidents dans certains territoires, il échoue significativement dans les grandes métropoles telles que Paris (75), Lille (59) ou Marseille (13). Ce phénomène s'explique par la surdispersion, illustrée par le graphique des résidus ci-dessous.

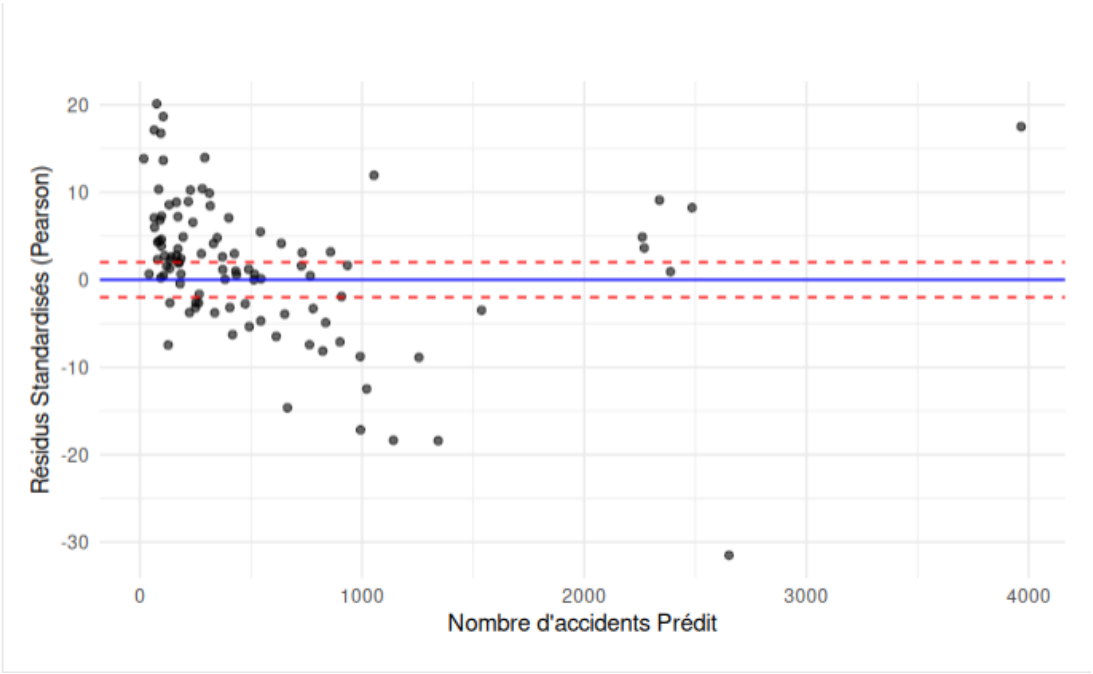

FIGURE 4 – Diagnostic des Résidus : Mise en évidence de la structure en entonnoir

La figure 4 met en évidence que la majorité des points extrêmes se situent hors de la zone de confiance (entre les lignes rouges). Cela confirme que la variance augmente avec la moyenne, justifiant notre choix d'opter pour la méthode Binomiale Négative.

Le passage à la loi Binomiale Négative permet de corriger ce biais structurel. En introduisant le paramètre de dispersion θ, la déviance chute spectaculairement à environ **102**. Cette forte réduction de la déviance confirme que la loi Binomiale Négative est largement plus adaptée pour modéliser l'hétérogénéité des données d'accidents à cette échelle. En effet, concrètement, on trouve que l'analyse des ratios de taux d'incidence révèle que les virages et autoroutes multiplient par six le nombre d'accidents attendus, tandis que la pluie semble diviser ce risque par deux, ce qui suggère une adaptation comportementale et une vigilance accrue des usagers par temps dégradé.

4.3.2 Approche de Précision (TMJA) : Résultats et Interprétation

Dans cette seconde approche, les variables explicatives ne sont pas les mêmes que celle pour l'offset Population. En effet, étant donné que le dataset TMJA était très lourd, on s'est notamment penché ici à certains facteurs, à savoir la Pluie, la Neige, la Nuit, les Virage le tout avec offset qui est égal à log(TMJA Moyen).

Nous appliquons ici la même méthodologie en restreignant l'analyse à la région Auvergne-Rhône-Alpes et en définissant l'offset par le logarithme du trafic moyen (log(TMJA)). L'intégration de cet offset et le passage à la loi Binomiale Négative améliorent drastiquement la qualité du modèle, l'AIC chutant de 11 887 à **3969**.

L'analyse des coefficients met en évidence une dualité marquée. D'une part, la Nuit (coef 1.64) apparait comme un facteur aggravant majeur qui augmente le risque de 64% à trafic constant, en raison de la visibilité réduite et de la fatigue. D'autre part,la Neige (0.61) et les Virages (0.23) présentent des coefficients inférieurs à 1, ce qui suggère un effet "protecteur" contre-intuitif.

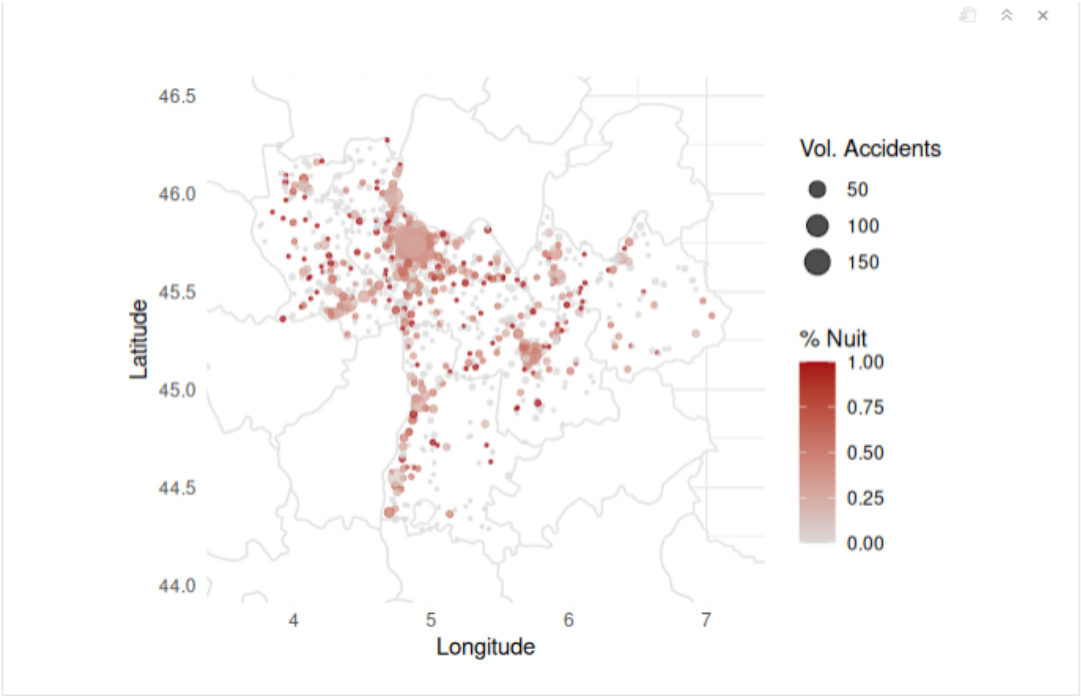

FIGURE 5 – Visualisation de l'accidentologie : Répartition spatiale et impact de la Nuit

Comme le montre la figure 5, il y a une forte concentration du volume d'accidents au sein des pôles urbains majeurs tels que Lyon et Grenoble, contrastant nettement avec les zones rurales. Alors que le modèle identifie un effet « protecteur » de la pluie lié à la prudence des usagers, cette carte rappelle que la densité du trafic en métropole reste le facteur prédominant du nombre total d'accidents, particulièrement en période nocturne.

4.4 Limites de nos modèles de fréquences

Malgré la robustesse statistique de l'approche par la loi Binomiale Négative, cette étude présente certaines limites inhérentes à la nature des données. Premièrement, le biais d'agrégation induit par le regroupement des accidents par commune lisse l'information et risque de masquer des "points noirs" très localisés, qui se retrouvent noyés dans la moyenne globale de la ville. Deuxièmement, la nature statique du TMJA constitue une approximation, car ce trafic moyen annuel ne capture pas la réalité dynamique (congestion, heures de pointe) alors que le risque diffère radicalement entre une circulation fluide et un trafic saturé. Enfin, le facteur humain reste une variable inobservée majeure ; nos modèles

structurels ne contiennent aucune donnée sur l'état du conducteur (fatigue, alcoolémie, distraction), qui demeure pourtant le déterminant principal de l'accidentologie.

5 Conclusion et perspectives

Ce projet a permis d'analyser l'accidentologie sous trois angles complémentaires : la sévérité, expliquée par le modèle linéaire ; la typologie, qui a révélé trois profils d'accidents (dont un majoritaire urbain assez hétérogène) ; et la fréquence, directement corrélée au volume de trafic. Ces travaux confirment que la compréhension des accidents nécessite de croiser des variables contextuelles variées plutôt que de s'appuyer sur une cause unique.

Nos résultats restent toutefois conditionnés par plusieurs limites techniques et matérielles. L'intégration de données comportementales (alcool, inattention) aurait ainsi permis d'expliquer la part d'accidents que la seule configuration technique ne suffit pas à justifier. Par ailleurs, disposer d'une puissance de calcul supérieure nous aurait affranchis des contraintes d'échantillonnage, rendant possible le traitement exhaustif de la base pour détecter des phénomènes plus rares. Enfin, étendre l'analyse à l'ensemble de l'historique temporel aurait offert la possibilité de lisser les effets conjoncturels d'une année unique, garantissant ainsi une meilleure généralisabilité des conclusions.

Références

[1] F. Bach, *Apprentissage Statistique (Cours)*, École Normale Supérieure (ENS), Paris, 2010. Disponible sur : `https://www.di.ens.fr/~fbach/courses/fall2010/cours3.pdf`

[2] INSEE, *Statistiques et études nationales*, Institut National de la Statistique et des Études Économiques. Disponible sur : `https://www.insee.fr/fr/statistiques/7739582?sommaire=7728826`

[3] A. Kassambara, *Practical Guide To Cluster Analysis in R : Unsupervised Machine Learning*, STHDA, 2017.

[4] Ministère de la Transition Écologique, *Trafic moyen journalier annuel sur le réseau routier national*, Plateforme data.gouv.fr. Disponible sur : `https://www.data.gouv.fr/datasets/trafic-moyen-journalier-annuel-sur-le-reseau-routier-national/`

[5] V. Monbet, *Modèles Linéaires Généralisés (GLM)*, Université de Rennes 1. Disponible sur : `https://perso.univ-rennes1.fr/valerie.monbet/GLM/GLMpharma.pdf`