UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MODÈLES STATISTIQUES DE SURVEILLANCE DE MALADIES INFECTIEUSES APPLIQUÉS AUX VOLS D'AUTOMOBILES

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

MARCO LAVOIE

NOVEMBRE 2024

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.12-2023). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

REMERCIEMENTS

Je tiens à remercier mon directeur de recherche Jean-Philippe Boucher de m'avoir pris sous son aile. Ses connaissances et ses conseils m'ont permis de grandir et de me développer dans le monde de la recherche. Je ne peux passer sous silence sa patience ainsi que sa disponibilité qui furent très appréciées.

Je tiens également à remercier ma femme qui m'a soutenu tout au long de mon parcours universitaire. Sans sa présence et ses sacrifices, je n'aurais pas pu consacrer tout le temps souhaité et nécessaire à ce mémoire. Merci également pour sa patience et ses encouragements.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

ACRONYMES

UQAM Université du Québec à Montréal.

GLR Generalized Likelihood Ratio.

VR Valeur de Référence.

GLM Generalized Linear Model.

MLE Maximum Likelihood Estimation.

CSP Contrôle Statistique des Processus.

LASSO Least Absolute Shrinkage and Selection Operator.

OLS Ordinary Least Squares.

ARIMA AutoRegressive Integrated Moving Average.

CUSUM Cumulative Sum.

IWLS Iterated Weighted Least Squares.

NB2 Negative Binomial 2.

ASFC Agence des Services Frontaliers du Canada.

RÉSUMÉ

Ce mémoire explore l'application de modèles statistiques de surveillance des maladies infectieuses aux vols de voitures dans le but de fournir des outils analytiques pour les compagnies d'assurance. Les modèles Farrington et GLR (Generalized Likelihood Ratio) sont examinés pour détecter des anomalies dans les données de vols de véhicules, permettant ainsi une gestion proactive de l'assureur pour limiter les vols. Le modèle Farrington utilise un sous-ensemble de données historiques pour prédire les occurrences et fixer des seuils d'alerte, tandis que le modèle GLR, plus récent, intègre toutes les données disponibles dans un modèle paramétrique qui capture la saisonnalité. De plus, pour améliorer la qualité de prédiction, des variables explicatives propres aux assurés sont ajoutées dans le modèle GLR. Les résultats montrent que les deux modèles sont efficaces pour détecter des valeurs aberrantes, mais que l'approche GLR est plus robuste et adaptative. En intégrant ces modèles dans leurs systèmes de gestion des risques, les compagnies d'assurance peuvent améliorer leur réactivité et développer des stratégies pour minimiser les pertes dues aux vols de voitures. Cette étude souligne l'importance de l'adaptation continue des modèles aux spécificités des données et des tendances de marché pour optimiser leur utilisation en assurance automobile.

Mots-clés : IARD, modèle Farrington, modèle GLR, LASSO, GLM, vols de voitures, actuariat, covariables, quasi-Poisson, binomiale négative, saisonnalité, tendance, seuil, exposition.

INTRODUCTION

La surveillance des maladies infectieuses a toujours été un domaine fondamental de la santé publique. Traditionnellement, la surveillance épidémiologique emploie des méthodes permettant de détecter rapidement les épidémies émergentes, facilitant ainsi des interventions efficaces. Ce mémoire explore l'application de modèles statistiques aux vols de voitures pour fournir des informations utiles aux compagnies d'assurance.

Le vol d'automobiles demeure un problème constant qui affecte les sociétés modernes, avec des répercussions économiques, sociales et sécuritaires significatives. Ce mémoire se penche sur diverses facettes du vol d'automobiles, notamment ses causes, les techniques utilisées par les voleurs, son impact économique et social, ainsi que des mesures préventives efficaces. Le vol d'automobiles est défini comme l'appropriation illégale d'un véhicule à moteur. Il s'agit d'un phénomène mondial, bien que l'incidence varie selon les régions et les pays. Il est crucial de comprendre les facteurs qui contribuent au vol d'automobiles ainsi que les stratégies mises en oeuvre pour les prévenir, afin de développer des politiques efficaces. Plusieurs études ont souligné le lien entre la pauvreté, le chômage et le vol d'automobiles. Les individus vivant dans des conditions économiques difficiles peuvent être plus enclins à se tourner vers des activités illégales, telles que le vol, pour subvenir à leurs besoins. (Clarke et Harris, 1992) notent que la motivation économique constitue souvent un moteur principal pour les voleurs d'automobiles. Les caractéristiques environnementales telles que l'urbanisation et la densité de la population jouent également un rôle significatif dans l'incidence du vol d'automobiles. Les zones urbaines densément peuplées tendent à avoir des taux de vol plus élevés, en raison de la facilité d'accès et de l'anonymat qu'elles offrent aux criminels. (Braga et Clarke, 2014) a démontré que la concentration de crimes est souvent plus élevée dans les milieux urbains en raison de ces facteurs.

En outre, les avancées technologiques ont entraîné des méthodes de vol d'automobiles plus sophistiquées. Les voleurs utilisent désormais des dispositifs électroniques pour contourner les systèmes de sécurité avancés des véhicules modernes. (Farrell *et al.*, 2011) identifient une tendance croissante à l'utilisation de la technologie dans les crimes liés au vol de véhicules, soulignant l'adaptation continue des voleurs aux nouvelles technologies. Le cambriolage traditionnel, impliquant le bris de fenêtres et le forçage des serrures, reste encore courant. Cependant, le piratage électronique a gagné en popularité, les voleurs utilisant des dispositifs pour intercepter les signaux des clés électroniques, comme le note (Ayres et Levitt, 1998). Les vols impliquant la fraude, tels que les fraudes d'identité pour obtenir des prêts de véhicules, augmentent également. Les voleurs peuvent louer des véhicules avec de fausses identités et ne jamais les restituer, selon

(Webb et Laycock, 1992) et (Tilley et Laycock, 2018).

Le vol d'automobiles a des conséquences économiques importantes, coûtant des milliards de dollars chaque année aux compagnies d'assurance, aux propriétaires de véhicules et aux gouvernements. Le vol d'automobiles augmente aussi le coût des primes d'assurance et entraîne des pertes financières pour les propriétaires. Une étude de (Levi et Maguire, 2004) estime que le coût total du vol de véhicules aux États-Unis représente plusieurs milliards de dollars annuellement. Sur le plan social, les communautés où le vol d'automobiles est courant peuvent voir une détérioration de la qualité de vie. Le sentiment d'insécurité peut entraîner des comportements de repli et une diminution de l'engagement communautaire, comme l'ont souligné (Welsh et Farrington, 2003) dans leur étude sur les effets de l'éclairage public amélioré sur la criminalité.

Diverses stratégies ont été développées pour prévenir le vol d'automobiles, allant de l'utilisation de la technologie aux politiques publiques. Par exemple, l'installation de dispositifs antivol, tels que les alarmes, les systèmes de suivi GPS et les dispositifs d'immobilisation, s'est révélée efficace pour dissuader le vol. Les politiques publiques, telles que l'amélioration de l'éclairage public et l'augmentation de la surveillance policière, peuvent également réduire les opportunités de vol. En 2024, le gouvernement du Canada (Arcand, 2024) a mis en place des mesures concrètes pour réduire le nombre de véhicules volés en instaurera un plan d'action de 28 millions de dollars. Les principales mesures incluent des peines plus sévères, l'amélioration de la communication entre les forces de police, une meilleure capacité d'intervention sur le terrain et l'Agence des services frontaliers du Canada (ASFC) utilisera l'intelligence artificielle pour retrouver les véhicules volés. Comme le rapporte (Ekblom, 1997), des programmes communautaires visant à sensibiliser le public aux risques de vols d'automobiles ont également été efficaces dans ses recherches sur les cadres dynamiques pour aider les concepteurs à se maintenir au niveau des criminels dans un monde en évolution. Les initiatives communautaires, comme les groupes de surveillance de quartier, jouent un rôle clé dans la prévention du vol d'automobiles. En encourageant la coopération entre les résidents et les forces de l'ordre, ces programmes peuvent créer un environnement moins accueillant pour les criminels. (Tilley et Laycock, 2018) ont exploré l'importance de la mise en oeuvre de la prévention du crime d'un point de vue international et ont également constaté que l'implication communautaire est cruciale pour le succès de ces initiatives.

En somme, le vol d'automobiles est un problème complexe qui nécessite une approche multidimensionnelle pour être traité efficacement. Les stratégies de prévention doivent être adaptées aux contextes locaux et

doivent inclure des mesures technologiques, des politiques publiques et la participation communautaire. La recherche continue sur les tendances et les méthodes de vols d'automobiles est essentielle pour développer des stratégies de prévention plus efficaces et mieux ciblées.

Toutes les références citées fournissent des ressources supplémentaires pour approfondir la compréhension du sujet. (Ayres et Levitt, 1998) discutent des externalités positives découlant des précautions invisibles prises par les victimes, en utilisant Lojack comme exemple. (Braga et Clarke, 2014) analysent les concentrations de crimes à haut risque dans les villes, en explorant la désorganisation sociale et les opportunités criminelles. (Clarke et Harris, 1992) offrent un aperçu des mesures de prévention du vol d'automobiles, tandis que (Ekblom, 1997) propose un cadre pour aider les concepteurs à faire face aux criminels. (Farrell *et al.*, 2011) examinent l'impact de la sécurité automobile sur le crime, tandis que (Welsh et Farrington, 2003) évaluent les effets de l'éclairage public amélioré sur la criminalité. (Levi et Maguire, 2004) critiquent les approches fondées sur des preuves pour réduire la criminalité organisée. Enfin, (Webb et Laycock, 1992) explorent l'ampleur et la nature du vol d'automobiles dans le cadre de l'unité de prévention du crime, et (Tilley et Laycock, 2018) soulignent l'importance de la prévention du crime à l'échelle internationale. Pour résumer, le vol d'automobiles reste un défi majeur pour la sécurité publique et l'économie mondiale. La compréhension approfondie de ses causes, de ses méthodes et de son impact, ainsi que la mise en œuvre de mesures de prévention bien conçues, sont essentielles pour atténuer ce problème persistant.

Tout comme en contexte d'épidémie, le nombre de vols de voitures connaît une hausse notable au Canada, avec des taux particulièrement élevés au Québec. En 2022, environ 9500 véhicules ont été volés à Toronto, et une augmentation de 50 % a été signalée au Québec pour la même année (Gérard, 2023). Il est donc pertinent de faire un parallèle entre l'augmentation du nombre de vols flagrants et la surveillance épidémiologique.

Dans ce mémoire de recherche, deux modèles généralement utilisés pour la surveillance épidémiologique seront appliqués aux vols de voitures, la méthode Farrington (Farrington *et al.*, 1996) et le modèle de rapport de vraisemblance généralisé (GLR) (Höhle et Paul, 2008). Le coeur de l'étude repose donc sur ces deux modèles initialement utilisés en contexte épidémique. La méthode Farrington, qui utilise un historique de valeurs passées similaires pour prédire les occurrences et fixer des seuils d'alerte, est comparée au modèle GLR. Ce dernier, développé plus récemment, intègre toutes les données historiques disponibles et un modèle paramétrique pour capturer la saisonnalité, offrant ainsi une approche robuste pour détecter des

valeurs aberrantes. Le modèle GLR se distingue par sa capacité à maximiser le rapport de vraisemblance pour déterminer les points de changement dans les données, s'inscrivant dans un cadre de contrôle statistique des processus (CSP).

Cette recherche a pour but d'adapter ces modèles au contexte des assurances pour identifier rapidement les anomalies dans les statistiques de vols de voitures. L'objectif est de permettre aux compagnies d'assurance de réagir proactivement aux valeurs aberrantes, limitant ainsi les pertes potentielles. Les modèles, en tenant compte de la saisonnalité et de la dynamique des données, offrent des outils puissants pour la surveillance des vols et l'élaboration de stratégies de prévention efficaces.

La structure de ce mémoire sera la suivante. Nous commencerons, au chapitre 1, par introduire quelques concepts théoriques essentiels à la science actuarielle de l'assurance de dommages. Au chapitre 2, nous ferons une analyse descriptive des données utilisées pour la recherche. Puis, au chapitre 3, une sélection de variables sera effectuée à partir d'outils statistiques. Au chapitre 4, les modèles Farrington et GLR seront décrits et appliqués sur les données de vols. Finalement, au chapitre 5, nous développerons une approche que nous appellerons GLR-étendu qui correspond à l'ajout de covariables au modèle GLR. Ce nouveau modèle sera aussi appliqué aux données d'assurance et les résultats obtenus seront comparés avec ceux obtenus par les autres modèles.

Finalement, il est à noter que les calculs effectués pour ce mémoire ont été faits avec le logiciel R. La notation mathématique des chiffres et des nombres est en anglais, ainsi un point est utilisé pour identifier les décimales au lieu d'une virgule comme en français.

CHAPITRE 1

LES CONCEPTS ESSENTIELS

1.1 Définition des termes clés utilisés

Il est essentiel de commencer par établir les définitions des concepts fondamentaux qui seront utilisés tout au long de ce document.

Définition 1.1 (Exposition au risque) *L'exposition au risque se réfère à la durée pendant laquelle un assuré est couvert par un contrat d'assurance, mesurée en unités de temps (jours, heures, etc.). Dans le cas des assurances automobiles, les contrats sont généralement souscrits pour une durée d'un an et il est habituel de voir l'exposition au risque comme une proportion d'année. Néanmoins, dans le cas de ce mémoire, l'exposition sera définie en fonction du nombre de jours assurés.*

Exemple 1 (Unité de mesure de l'exposition) *Un véhicule assuré du 1er janvier 2023 au 1er juillet inclusivement est assuré pendant 182 jours. Sous une mesure d'exposition annuelle, l'exposition de cet assuré est d'environ 0.5 année, mais elle est de 182 jours sous une mesure d'exposition définie en nombre de jours assurés.*

Définition 1.2 (Fréquence de réclamations) *La fréquence de réclamations représente le nombre de sinistres enregistrés par unité d'exposition au risque. Formellement, nous avons ainsi :*

$$\textit{Fréquence} = \frac{\textit{Nombre de réclamations}}{\textit{Exposition totale}}. \tag{1.1}$$

*Il est à noter que selon le choix de mesure de l'exposition, nous pouvons avoir différentes formes de fréquence, et ainsi avoir par exemple une fréquence **annuelle** de sinistres, ou encore une fréquence **quotidienne**.*

Exemple 2 (Calcul de la fréquence) *L'assuré de l'exemple précédent est victime d'un vol de voiture. Si nous utilisons une mesure d'exposition annuelle, cet assuré présentera une fréquence de réclamation de* $\frac{1}{0.5} =$

200% *alors que sous une base d'exposition par jour assuré, il présentera une fréquence de réclamations de* $\frac{1}{182} = 0.55\%$.

Définition 1.3 (Sévérité des réclamations) *La sévérité des réclamations désigne le coût moyen associé à un sinistre, calculé en divisant le coût total des sinistres par le nombre de sinistres sur une période déterminée. Nous avons ainsi :*

$$\textit{Sévérité} = \frac{\textit{Sommes des pertes}}{\textit{Nombre de réclamations}}. \tag{1.2}$$

Contrairement à la fréquence, la sévérité ne dépend pas de l'unité de mesure choisie pour l'exposition.

Définition 1.4 (Charge pure) *La charge pure des réclamations correspond au coût moyen d'assurance par unité d'exposition, calculé en multipliant la fréquence et la sévérité. Elle s'exprime comme :*

$$\begin{aligned} \textit{Charge pure} &= \frac{\textit{Sommes des pertes}}{\textit{Exposition totale}} \\ &= \underbrace{\frac{\textit{Sommes des pertes}}{\textit{Nombre de réclamations}}}_{\textit{Sévérité}} \times \underbrace{\frac{\textit{Nombre de réclamations}}{\textit{Exposition totale}}}_{\textit{Fréquence}} \\ &= \textit{Sévérité} \times \textit{Fréquence}\ . \end{aligned} \tag{1.3}$$

La charge pure mentionnée dans l'équation (1.3) est souvent désignée par le terme anglais loss cost. Cette charge pure indique le montant moyen des sinistres payés pour chaque unité d'exposition.

Comme pour la fréquence, selon le choix de mesure de l'exposition, nous aurons une charge pure annuelle (qui se compare ainsi à la prime annuelle payée par un assuré) ou une charge pure quotidienne.

Exemple 3 (Calcul de la charge pure) *La perte estimée du véhicule volé de l'assuré des deux exemples précédents est de 65 000$. Sous une base d'exposition annuelle, la charge pure de cet assuré sera de* $\frac{65000}{0.5} = 130000\$$ *, alors que sous une base quotidienne la charge pure sera de* $\frac{65000}{182} = 357\$$.

1.2 Les modèles linéaires généralisés

Contrairement aux modèles de régression linéaire, qui sont limités aux variables dépendantes continues suivant une distribution normale et utilisant généralement une fonction de lien identitaire, les modèles linéaires généralisés (GLM) élargissent cette approche (de Jong et Heller, 2008). Les GLM permettent de traiter une variété de types de données en utilisant des distributions appropriées et des fonctions de lien adaptées à chaque cas.

1.2.1 Famille exponentielle linéaire

Pour une variable aléatoire Y, la fonction $f(y)$ est utilisée pour décrire soit la fonction de densité d'une variable aléatoire continue, soit la fonction de probabilité d'une variable aléatoire discrète. Les distributions appartenant à la famille exponentielle linéaire peuvent être exprimées sous la forme :

$$f(y) = \exp\left(\frac{y\theta - a(\theta)}{\phi} + c(y,\phi)\right), \tag{1.4}$$

où θ est le paramètre canonique et ϕ est généralement le paramètre de dispersion. Les distributions dont la fonction de densité ou de probabilité peut être écrite sous cette forme sont classées comme membres de la famille exponentielle linéaire.

Pour une variable aléatoire Y de la famille exponentielle linéaire, les moments suivants sont donnés par les expressions suivantes :

$$\mathbb{E}(Y) = a'(\theta) \quad \text{et} \quad \mathrm{Var}(Y) = a''(\theta)\phi.$$

On peut ainsi voir que la variance de Y se décompose en deux parties :

- La première partie, $a''(\theta)$, dépend uniquement du paramètre θ et est appelée fonction de variance.
- La seconde partie dépend seulement du paramètre ϕ et est indépendante de θ. La fonction de variance peut ainsi être exprimée en fonction de μ.

En notant $\mu = \mathbb{E}(Y)$, on établit que le paramètre θ est lié à la moyenne μ. La fonction de variance peut donc être écrite en fonction de μ comme suit :

$$\mathrm{V}(\mu) = a''\left([a']^{-1}(\mu)\right).$$

Quelques fonctions de variance sont illustrées à la table 1.1. La fonction de variance caractérise entièrement la loi au sein de la famille exponentielle linéaire. Chaque loi de cette famille est associée à une fonction de lien spécifique, appelée fonction de lien canonique, qui relie μ au paramètre canonique θ via $g(\mu) = \theta$. Ainsi, $\mu = a'(\theta)$ implique que $g(\cdot) = (a')^{-1}(\cdot)$. La fonction de lien $g(\mu) = \mathbf{X}^T\beta$ est une transformation de la moyenne μ et est linéairement liée aux variables explicatives contenues dans le vecteur $\mathbf{X}$.

Loi de probabilité	V(μ)
Normale	1
Poisson	μ
Gamma	μ^2
Inverse gaussienne	μ^3
Bernouilli	$\mu(1-\mu)$

Table 1.1 – Quelques fonctions de variance des distributions membres de la famille exponentielle linéaire

1.2.2 Estimation du maximum de vraisemblance

Le modèle linéaire généralisé (GLM) repose sur l'idée de traiter un ensemble de variables aléatoires indépendantes $(Y_1, \ldots, Y_n)$, chacune suivant une distribution de la famille exponentielle linéaire. La log-vraisemblance générale pour ces variables est obtenue à partir de la fonction de densité décrite dans l'équation (1.4) :

$$\begin{aligned}
\ell(\beta, \phi) &= \sum_{i=1}^{n} \ln f(y_i; \beta, \phi) \\
&= \sum_{i=1}^{n} \left(\ln(c(y_i, \phi)) + \frac{y_i\theta_i - a(\theta_i)}{\phi} \right) \\
&= \frac{1}{\phi} \sum_{i=1}^{n} [y_i\theta_i - a(\theta_i)] + \sum_{i=1}^{n} \ln(c(y_i, \phi)),
\end{aligned}$$

où $\theta_i = \mathbf{X}_i^T \beta = g(\mu_i)$ pour les variables $y_i, i = 1, \ldots, n$ indépendantes. En utilisant la règle de dérivation en chaîne, les estimateurs du maximum de vraisemblance (MLE) pour β_j sont obtenus comme suit :

$$\begin{aligned}\frac{\partial \ell}{\partial \beta} &= \sum_{i=1}^{n} \frac{\partial \ln f(y_i; \beta, \phi)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta} \\ &= \sum_{i=1}^{n} \frac{1}{\phi} \left[y_i - a'(\theta_i) \right] \frac{\partial \theta_i}{\partial \beta}.\end{aligned}$$

Sachant que $E[Y_i] = \mu_i = a'(\theta_i)$ et que $\mathrm{Var}[Y_i] = \phi a''(\theta_i)$, nous obtenons :

$$\begin{aligned}\frac{\partial \mu_i}{\partial \beta} = a''(\theta_i) \frac{\partial \theta_i}{\partial \beta} &= \frac{1}{\phi} \mathrm{Var}[Y_i] \frac{\partial \theta_i}{\partial \beta} \\ \Rightarrow \frac{\partial \theta_i}{\partial \beta} &= \frac{\partial \mu_i}{\partial \beta} \frac{\phi}{\mathrm{Var}[Y_i]}.\end{aligned}$$

Dès lors :

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \frac{y_i - a'(\theta_i)}{\mathrm{Var}[Y_i]} \frac{\partial \mu_i}{\partial \beta}.$$

En appliquant de nouveau la règle de dérivation en chaîne :

$$\frac{\partial \mu_i}{\partial \beta} = \frac{\partial \mu_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta},$$

où :

$$\begin{aligned}\frac{\partial \theta_i}{\partial \beta} &= \mathbf{X}_i, \\ \frac{\partial \mu_i}{\partial \theta_i} = \left(\frac{\partial \theta_i}{\partial \mu_i} \right)^{-1} &= \left(g'(\mu_i) \right)^{-1}.\end{aligned}$$

En conséquence, la condition de premier ordre se formule comme suit :

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{g'(\mu_i)\,\mathrm{Var}[Y_i]} X_{ij} = 0. \tag{1.5}$$

En général, l'équation (1.5) n'a pas de solutions explicites et doit donc être résolue numériquement, souvent à l'aide de techniques telles que la méthode de Newton-Raphson, qui sera décrite dans la sous-section suivante.

1.2.3 La méthode Newton-Raphson

La méthode de Newton-Raphson, nommée en partie d'après le mathématicien anglais Isaac Newton, est une technique efficace pour trouver les racines des équations non linéaires. Elle repose sur l'idée que, à chaque point, les premières et deuxièmes dérivées de la fonction à maximiser peuvent être évaluées de manière précise. En utilisant ces dérivées, la méthode effectue une approximation quadratique de la fonction, puis maximise cette approximation. Les nouvelles valeurs obtenues sont utilisées pour calculer de nouvelles approximations, et ce processus se répète jusqu'à convergence. La méthode converge généralement rapidement vers un maximum.

L'idée principale de la méthode Newton-Raphson est de déterminer la tangente de la fonction au point actuel, de suivre cette tangente jusqu'à ce qu'elle croise l'axe des abscisses, et d'utiliser ce point d'intersection comme nouvelle estimation. Si la fonction à ce nouveau point n'est pas suffisamment proche de zéro, la procédure recommence à partir de ce point.

Supposons que ϕ soit connu et que $\ell(\beta)$ soit la log-vraisemblance en fonction du vecteur de paramètres inconnus β. L'approximation quadratique de la série de Taylor au point β est donnée par :

$$\ell(\beta + \delta) \approx \ell(\beta) + \ell'(\beta)\delta + \frac{\delta^2}{2}\ell''(\beta), \tag{1.6}$$

où $\ell'(\beta)$ est le vecteur des dérivées partielles $\partial \ell / \partial \beta_j$ (appelé vecteur score), et $\ell''(\beta)$ est la matrice des dérivées partielles croisées $\partial^2 \ell / \partial \beta_j \partial \beta_k$ (appelée matrice hessienne).

En dérivant l'expression de l'équation (1.6) par rapport à δ et en l'égalant à zéro, on obtient :

$$\ell'(\beta) + \delta\ell''(\beta) = 0 \quad \Rightarrow \quad \delta = -\{\ell''(\beta)\}^{-1}\ell'(\beta). \tag{1.7}$$

Connaissant β et δ, on met à jour β en utilisant :

$$\beta^{(t+1)} = \beta^{(t)} - \left\{\ell''\left(\beta^{(t)}\right)\right\}^{-1}\ell'\left(\beta^{(t)}\right), \tag{1.8}$$

où $\beta^{(t)}$ représente la valeur de β à l'itération t.

Les itérations de cette mise à jour convergent souvent rapidement vers une solution optimale. Les valeurs initiales sont choisies arbitrairement pour commencer l'itération. Un critère d'arrêt est utilisé pour arrêter les itérations, soit :

$$\frac{\ell(\beta^{(t+1)}) - \ell(\beta^{(t)})}{\ell(\beta^{(t)})} \leq \varepsilon, \tag{1.9}$$

où ε est une petite tolérance.

Il est nécessaire que la matrice hessienne soit définie négative pour garantir un maximum local strict. Si elle est définie positive, nous aurions un minimum local strict. La procédure d'évaluation répétée du vecteur score et de la matrice hessienne pour mettre à jour les estimations est appelée l'itération de Newton-Raphson.

1.3 Estimation par noyau

L'estimation par noyau est une méthode non paramétrique pour évaluer la densité de probabilité d'une variable aléatoire à partir d'un ensemble de données. Contrairement aux approches paramétriques, elle ne repose sur aucune hypothèse préalable concernant la forme de la distribution sous-jacente. Cette technique est particulièrement utile lorsque la forme de la distribution des données est inconnue ou difficile à modéliser avec une fonction spécifique.

Cette méthode généralise l'estimation par histogramme en lissant les contributions de chaque donnée pour

produire une estimation continue de la densité. Elle utilise une fonction de noyau, qui est symétrique autour de zéro, pour attribuer des poids aux données voisines à chaque point d'estimation.

L'estimateur de la densité par noyau est donné par :

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1.10}$$

où :

- $\hat{f}_h(x)$ représente l'estimation de la densité en x ;
- n est le nombre total d'observations ;
- h est le paramètre de lissage, ou largeur de bande ;
- K est la fonction de noyau ;
- X_i sont les points de données observés.

Le choix de la fonction de noyau K et de la largeur de bande h est crucial pour la qualité de l'estimation. Les noyaux fréquemment utilisés incluent le noyau gaussien, le noyau épanechnikov et le noyau uniforme.

Bien que le choix du noyau ait un impact relativement faible, le paramètre de largeur de bande est déterminant. Un h trop petit peut introduire des détails artificiels, tandis qu'un h trop grand peut lisser excessivement les données, effaçant des caractéristiques importantes. Il est donc essentiel de sélectionner soigneusement h pour obtenir une estimation optimale.

En somme, l'estimation par noyau est une méthode flexible et efficace pour estimer des densités de probabilité sans hypothèses strictes sur la distribution des données. Son efficacité dépend de la sélection appropriée de la largeur de bande et de la fonction de noyau, visant à équilibrer le biais et la variance de l'estimation. Lorsqu'elle est correctement appliquée, elle peut offrir une vision précise de la structure sous-jacente des données, même dans des contextes complexes où les méthodes paramétriques peuvent échouer. On peut se référer à (Deheuvels, 1977) pour plus de détails à propos des densités à noyau.

CHAPITRE 2
LES DONNÉES

Les données utilisées dans ce projet de maîtrise proviennent d'une compagnie d'assurance canadienne, une coopérative canadienne majeure dans le domaine des assurances et des services financiers. Elles couvrent tous les contrats d'assurance automobile pour la protection en assurance de vols de voitures survenus en Ontario et au Québec entre 2013 et 2023, à l'exclusion des vols de motos. Pour le Québec, les données ne sont accessibles qu'à partir de 2015. Chaque enregistrement dans la base de données inclut la date de début et de fin du contrat, la date de survenance du sinistre avec les coûts associés et des informations sur les assurés sont également disponibles.

Une analyse sommaire des données est essentielle pour comprendre les tendances et les dynamiques sous-jacentes au phénomène étudié. Ce chapitre vise à analyser les données utilisées afin de dresser un portrait juste de la situation. L'exposition utilisée sera toujours définie comme étant le nombre de jours assurés.

2.1 Sinistralité quotidienne

Le tableau 2.1 présente ainsi un résumé de la sinistralité quotidienne, pour les vols automobiles :

- On peut voir qu'en moyenne, à chaque jour, un peu moins de 500 000 véhicules étaient couverts pour la protection en assurance de vols de voitures pour la base de données étudiée ;
- En moyenne, presque deux véhicules par jour sont volés dans le portefeuille étudié, avec un sommet à 17 véhicules pour une seule journée en 2023 ;
- Pour la période étudiée, la sévérité moyenne est d'un peu moins de 66 000$, avec une perte maximale supérieur à 700 000$.

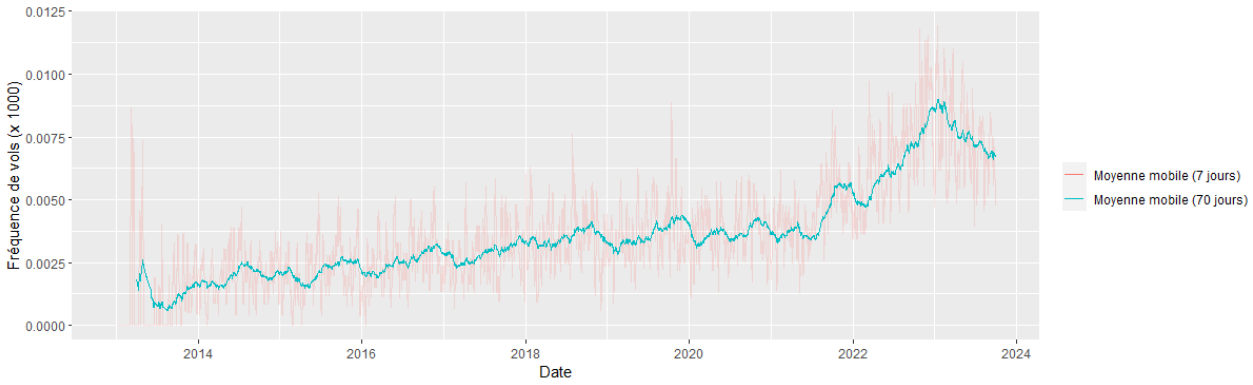
Statistiques	Exposition quotidienne (en milliers de véhicules par jour)	Nombre de vols	Fréquence quotidienne (par 1000 véhicules)	Sévérité
1er quartile	361.075	0.000	0.000%	10 048
Médiane	462.026	1.000	0.292%	31 126
Moyenne	464.048	1.914	0.358%	65 374
3e quartile	616.407	3.000	0.573%	84 334

Table 2.1 – Statistiques descriptives de la sinistralité quotidienne

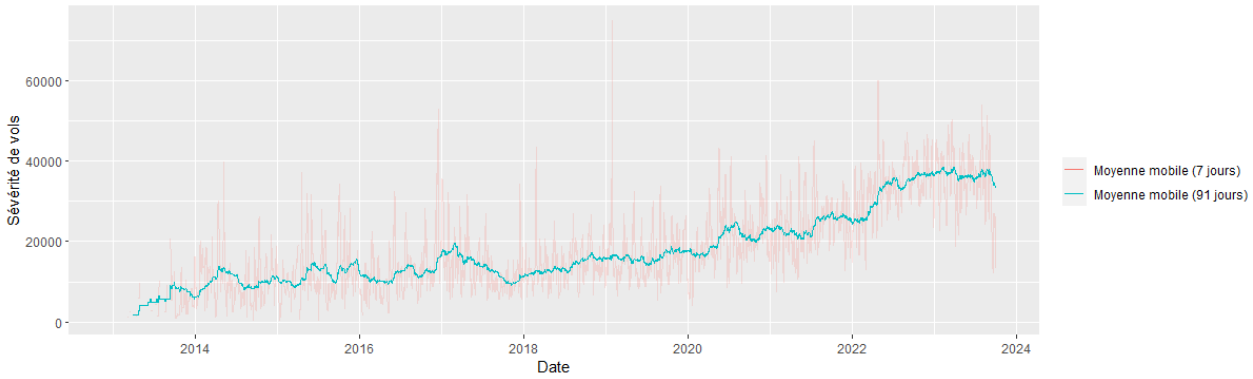
La fréquence de vols (par 1000 véhicules assurés), la sévérité des vols et la charge pure des vols, comme spécifié dans les équations (1.1), (1.2) et (1.3), ont été utilisées pour la figure 2.1. Pour les trois statistiques illustrées, les figures révèlent des éléments intéressants pour mieux comprendre l'épidémie de vols observée au Canada depuis 2020 :

- La moyenne mobile sur 70 jours de la fréquence indique une légère augmentation entre 2014 et 2022, suivie d'une forte montée jusqu'en 2023, avant une baisse ultérieure ;
- La moyenne mobile sur 91 jours de la sévérité diminue entre 2017 et 2018, puis montre une tendance à la hausse jusqu'en 2024 ;
- La moyenne mobile sur 91 jours de la charge pure est similaire à ce que nous observons pour la fréquence.

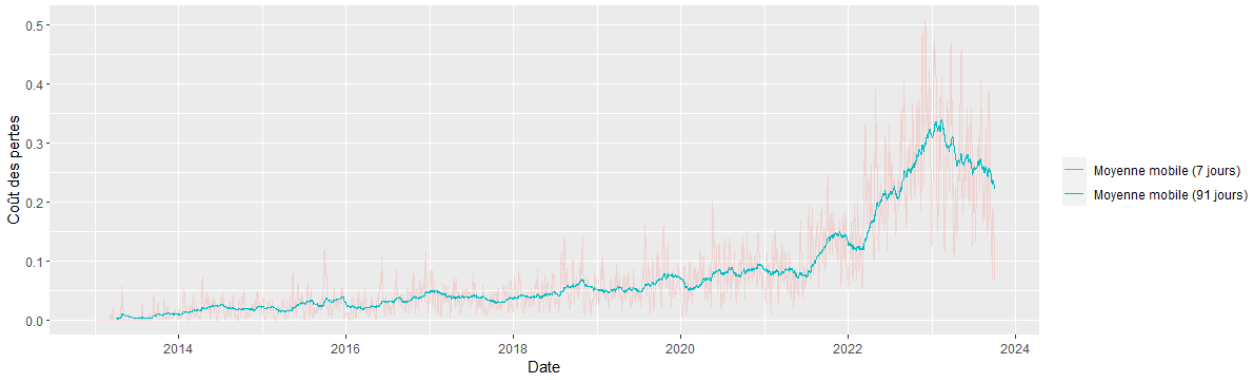
Dans tous les cas, une tendance générale à la hausse est remarquée pour les trois mesures au début de 2022.



(a) Fréquence quotidienne de vols



(b) Sévérité quotidienne de vols



(c) Charge pure quotidienne de vols

Figure 2.1 – Moyenne mobile quotidienne de la sinistralité

2.2 Caractéristiques des contrats d'assurance

Dans cette section, nous présenterons un résumé des données analysées, en mettant l'accent sur diverses caractéristiques des contrats d'assurance. Plus précisément, nous examinons les données en fonction des caractéristiques des assurés, telles que le sexe de l'assuré et sa région de résidence, ainsi que les caractéristiques des véhicules assurés, comme la marque ou le type de véhicule. Ce résumé fournira un aperçu des principaux facteurs influençant les sinistres et les vols de véhicules.

La table 2.2 définit certaines variables disponibles dans la base de données ainsi que leurs divisions en catégories ou modalités.

Nom de la variable	Description	Séparation
Chevaux	Puissance du véhicule	0-190 ch ou > 190 ch
Empat	L'empattement du véhicule	0-2745 mm ou > 2745 mm
Type	Type de véhicule	Camion ou pas des camions
Marital	État matrimonial légal	Séparé, célibataire, marié ou divorcé
Vehage	L'âge du véhicule	-2-5 ou 6-15 ou 16-64 ou > 64 ans
Genre	Sexe de l'assuré	Homme ou femme
Province	La province de l'assuré	Québec ou Ontario
Cat	Catégorie du véhicule	Camion, SUV, sedan ou autre
Poids	Le poids du véhicule	0-1611 kg ou > 1611 kg
Prix	Le prix d'achat du véhicule	0$-23 498$ ou > 23 498$
Saison	Les saisons	Été, printemps, hiver et automne
Marque	La marque du véhicule	Séparé par marque
Modele	Le modèle du véhicule	Séparé par modèle
Region	L'endroit où le véhicule a été volé	Séparé par région

Table 2.2 – Quelques caractéristiques du risque disponibles dans la base de données

Le choix des modalités a été fait selon l'approche suivante :

— Dans le cas des variables quantitatives, la création des modalités de chaque variable a été choisie par

rapport à leurs médianes respectives [1]. Au final, seule la variable quantitative Vehage a été séparée de façon à obtenir 4 groupes représentant les véhicules très récents, un peu moins récents, les vieux véhicules et les véhicules anciens ;

— Pour les variables qualitatives, toutes auront deux modalités, sauf la variable Marital qui en aura quatre et la variable Cat qui en aura trois.

2.2.1 Saisons

La table 2.3 présente une analyse des vols de voitures selon les saisons. Encore une fois, la fréquence (par millier de véhicules assurés), la sévérité des vols, de même que la charge pure sont indiquées. Il est à noter que la charge pure correspondra toujours au coût moyen d'assurance par jour. Ainsi, il faudrait multiplier cette valeur par 365 pour obtenir le coût de la protection annuelle.

On peut ainsi voir :

— Fréquence de vols : On observe que l'automne enregistre la fréquence la plus élevée des vols de voitures, à 0.441%, suivi par l'été, le printemps, et enfin l'hiver avec la fréquence la plus basse à 0.377%. Cette tendance suggère que les vols de voitures sont plus fréquents pendant les mois les plus chauds, probablement en raison de l'augmentation des déplacements et de l'utilisation des véhicules ;

— Sévérité des incidents : La sévérité des vols, mesurée par les pertes moyennes, ne suit pas nécessairement la même tendance que la fréquence. L'été enregistre la sévérité la plus élevée avec 25 535$, tandis que l'hiver, malgré une fréquence plus basse, a une sévérité de 23 883$;

— Charge pure : La charge pure, qui représente une combinaison des deux précédents indicateurs, est la plus élevée en été (0.11$) et la plus basse en hiver (0.09$). Cela reflète un risque global plus important en été.

1. Nous aurions pu utiliser la moyenne dans la création des modalités, mais la moyenne est influencée par les valeurs extrêmes et pourrait donner une image trompeuse de la tendance centrale. De plus, si les données ont une queue longue ou si elles sont asymétriques, la moyenne pourrait créer des catégories déséquilibrées.

Saison	Fréquence quotidienne (par 1000 véhicules)	Sévérité	Charge pure quotidienne
Automne	0.441%	23 359	0.10
Été	0.424%	25 535	0.11
Printemps	0.406%	25 138	0.10
Hiver	0.377%	23 883	0.09

Table 2.3 – L'analyse des vols de véhicules par saison

2.2.2 Marques de véhicules

La table 2.4 présente les dix marques de véhicules les plus volées possédant une exposition quotidienne supérieure à 4 millions, classées en ordre décroissant en fonction des fréquences de vol (par 1000 véhicules assurés). Les colonnes indiquent la marque du véhicule, l'exposition quotidienne (en milliers de véhicules), le nombre de vols, la fréquence de vols quotidienne (par 1000 véhicules), la sévérité du vol (mesurée en termes de coût moyen des sinistres) et la charge pure (la multiplication de la fréquence quotidienne et de la sévérité).

Marque du véhicule	Fréq. quotidienne (par 1000 véhicules)	Sévérité	Charge pure quotidienne
LAND ROVER	4.648%	69 920	3.25
LEXUS	2.536%	43 045	1.09
ACURA	0.778%	25 440	0.20
JEEP	0.769%	38 813	0.30
GMC	0.670%	15 529	0.10
HONDA	0.610%	23 378	0.14
MERCEDES-BENZ	0.594%	31 842	0.19
BMW	0.593%	18 941	0.11
CADILLAC	0.562%	29 766	0.17
PORSCHE	0.562%	42 945	0.24

Table 2.4 – Les dix voitures les plus volées (entre 2013 et 2023) selon la marque du véhicule en ordre décroissant des fréquences de vol (pour les véhicules ayant une exposition quotidienne totale supérieure à 4 000 000)

D'après les données de la table 2.4, Land Rover est la marque la plus fréquemment volée, avec une fréquence de 4.648%. La sévérité des vols pour Land Rover est supérieure à celle des autres marques à 69 920$, suivie de Lexus à 43 045$ et Porsche à 42 945$. La sévérité varie considérablement entre les différentes marques, avec des valeurs allant de 15 529$ pour GMC à plus de 69 920$ pour Land Rover.

La colonne Charge pure met en évidence le risque relatif des différentes marques en combinant fréquence et sévérité. Land Rover, avec une charge pure de 3.25$, est la marque associée au risque le plus élevé, suivie de loin par Lexus avec 1.09$. À l'autre extrémité du spectre, les marques comme GMC (0.10$) et BMW (0.11$) présentent les risques les plus faibles.

Ce tableau met en lumière la variabilité des risques de vol associés à différentes marques de véhicules. Cette information est cruciale pour les compagnies d'assurance qui doivent ajuster leurs primes en fonction des risques spécifiques à chaque marque de véhicule.

2.2.3 Régions

La table 2.5 présente une analyse des régions canadiennes les plus touchées par le vol d'automobiles. Cette analyse met en évidence les fréquences de vol (par 1000 véhicules assurés), la sévérité des pertes et la charge pure, soit une mesure combinée des deux premières métriques. Les données sont classées par ordre décroissant de fréquence, offrant une perspective claire sur les zones les plus vulnérables.

Emplacement	Fréq. quotidienne (par 1000 véh.)	Sév.	Charge pure quotidienne
Grand Toronto	0.953%	32 018	0.31
Grand Montréal	0.864%	29 259	0.26
Centre de l'Ontario	0.465%	27 743	0.13
Ouest du Québec	0.424%	32 957	0.14
Sud-Ouest de l'Ontario	0.382%	13 729	0.05
Est de l'Ontario	0.264%	23 933	0.06
Nord de l'Ontario	0.189%	17 349	0.03
Est du Québec	0.129%	33 317	0.04

Table 2.5 – Les régions les plus volées en ordre décroissant des fréquences de vol

La région du Grand Toronto semble la plus dangereuse avec une fréquence de vols à 0.953% suivi par le Grand Montréal à 0.864%. L'est du Québec affiche la sévérité la plus élevée à 33 317$, soulignant des pertes financières importantes par incident. De manière générale, les régions urbaines, telles que le Grand Montréal et le Grand Toronto, semblent subir des pertes financières plus lourdes.

2.2.3.1 Carte géographique des vols

La figure 2.2 présente une carte de chaleur des vols de voitures à travers une vaste zone géographique du Canada. Les régions avec les couleurs les plus intenses (du bleu au rouge) indiquent les zones où les vols de voitures sont les plus fréquents. Par exemple, les régions autour de Toronto, Ottawa et Montréal apparaissent comme particulièrement touchées, ce qui suggère une concentration plus élevée des vols de voitures dans ces zones. Conformément à l'introduction, (Braga et Clarke, 2014) ont démontré que la concentration de crimes est généralement plus élevée dans les milieux urbains.

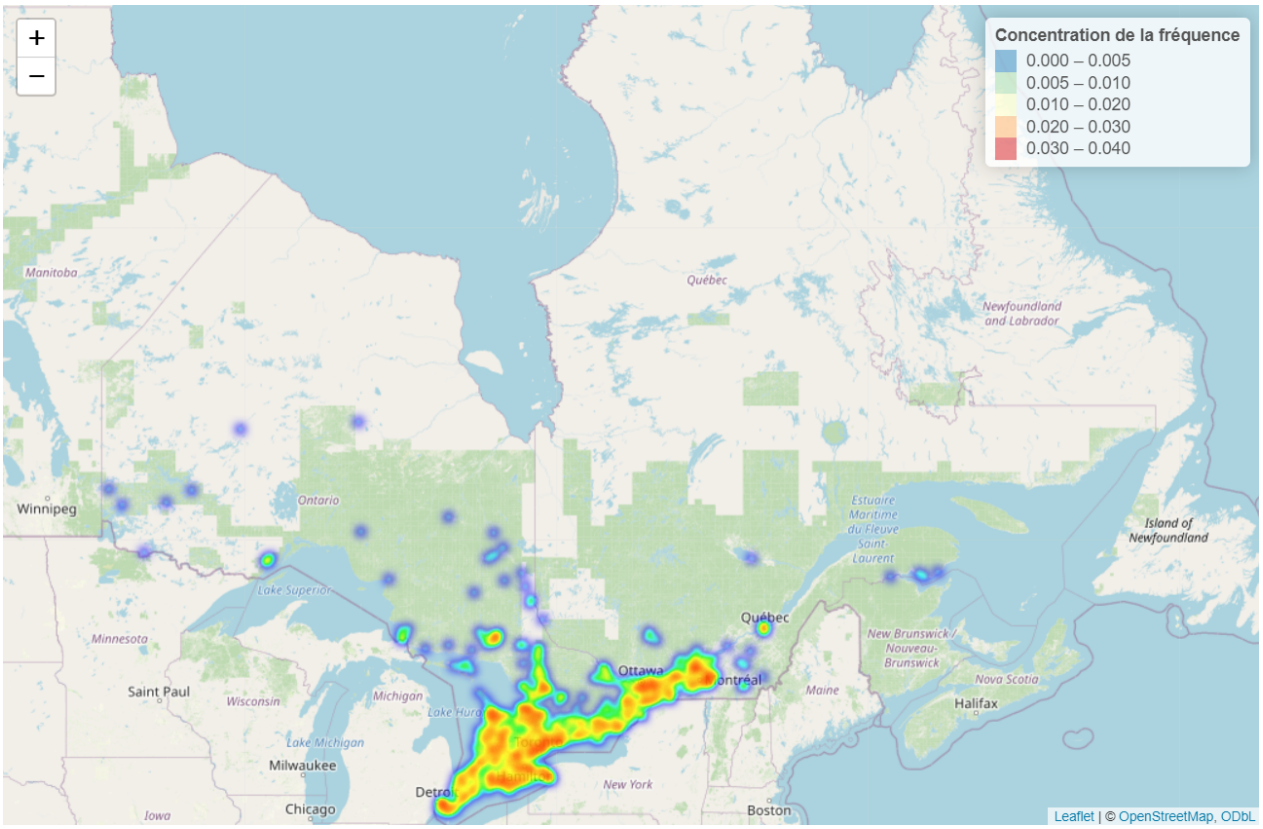


Figure 2.2 – Représentation cartographique des vols de voitures de 2013 à 2023

2.2.4 Sexe des assurés

La table 2.6 compare les vols de véhicules segmentés par le genre du conducteur principal associé au véhicule assuré. Ce tableau, divisé entre les femmes et les hommes, présente encore une fois des statistiques sur la fréquence (par 1000 véhicules assurés), la sévérité et la charge pure des vols.

L'analyse détaillée des données montre des tendances intéressantes. Pour les femmes, la fréquence des vols débute à 0.095% en 2013 et connaît une augmentation significative jusqu'à atteindre la valeur de 0.606%. Du côté des hommes, la fréquence commence à un niveau légèrement supérieur à celui des femmes en 2013, avec 0.190%, et suit une tendance ascendante pour atteindre 0.858% en 2023.

En termes de sévérité, les femmes voient leurs chiffres grimper régulièrement de 7 507$ en 2013 à un sommet de 33 964$ en 2023, avec des hausses annuelles presque constantes. Pour les hommes, la sévérité des vols augmente également de manière régulière, passant de 6 394$ en 2013 à 36 831$ en 2023. L'année 2021

se distingue par un saut significatif en termes de coût par incident, reflétant une aggravation continue de la situation.

En conclusion, l'analyse des données révèle une augmentation de la fréquence, de la sévérité et de la charge pure des vols de véhicules pour les deux genres. On constate que les véhicules appartenant à des hommes représentent des risques plus élevés que les voitures appartenant à des femmes.

	Femme			Homme		
Année	Fréq. quotidienne (par 1000 véh.)	Sévérité	Charge pure quotidienne	Fréq. quotidienne (par 1000 véh.)	Sévérité	Charge pure quotidienne
2013	0.095%	7 507	0.01	0.190%	6 394	0.01
2014	0.153%	12 252	0.02	0.253%	9 560	0.02
2015	0.188%	11 426	0.02	0.250%	12 916	0.03
2016	0.220%	14 425	0.03	0.303%	11 913	0.04
2017	0.208%	12 637	0.03	0.380%	12 935	0.05
2018	0.277%	14 203	0.04	0.425%	14 188	0.06
2019	0.319%	15 396	0.05	0.422%	17 516	0.07
2020	0.266%	20 180	0.05	0.463%	21 871	0.10
2021	0.363%	24 163	0.09	0.523%	24 429	0.13
2022	0.533%	32 927	0.18	0.819%	35 359	0.29
2023	0.606%	33 964	0.21	0.858%	36 831	0.32

Table 2.6 – L'analyse des vols de véhicules selon le genre du conducteur principal et pour chacune des années

2.2.5 La puissance du moteur

La table 2.7 compare les vols de véhicules segmentés par la puissance du véhicule assuré. Elle compare les véhicules ayant une puissance $\leq$ 190 chevaux avec ceux dépassant 190 chevaux.

Tout d'abord, en ce qui concerne la fréquence des vols (par 1000 véhicules assurés), on observe une augmentation progressive pour les deux catégories de véhicules. Pour les véhicules de moins de 190 chevaux, la fréquence des vols est passée de 0.112% en 2013 à 0.297% en 2023. Cette tendance à la hausse est particulièrement marquée avec un pic à 0.419% en 2022. Pour les véhicules de plus de 190 chevaux, la fréquence

des vols est constamment plus élevée, débutant à 0.183% en 2013 pour atteindre 1.193% en 2023, avec une augmentation significative à partir de 2021.

	Chevaux $\leq$ 190			Chevaux > 190		
Année	Fréq. quotidienne (par 1000 véh.)	Sévérité	Charge pure quotidienne	Fréq. quotidienne (par 1000 véh.)	Sévérité	Charge pure quotidienne
2013	0.112%	5 125	0.01	0.183%	8 057	0.01
2014	0.147%	6 867	0.01	0.275%	12 986	0.04
2015	0.148%	7 825	0.01	0.308%	14 911	0.05
2016	0.181%	10 028	0.02	0.363%	14 700	0.05
2017	0.179%	9 365	0.02	0.438%	14 521	0.06
2018	0.234%	10 358	0.02	0.496%	16 300	0.08
2019	0.247%	11 655	0.03	0.520%	19 434	0.10
2020	0.250%	11 306	0.03	0.511%	26 852	0.14
2021	0.291%	16 236	0.05	0.624%	28 516	0.18
2022	0.419%	21 672	0.09	0.970%	40 361	0.39
2023	0.297%	17 631	0.05	1.192%	40 326	0.48

Table 2.7 – L'analyse des vols de véhicules par la puissance du véhicule et pour chacune des années

Ensuite, la sévérité des sinistres montre également une tendance à la hausse, reflétant l'augmentation des coûts moyens par sinistre au fil des années. Pour les véhicules de puissance inférieure ou égale à 190 chevaux, la sévérité passe de 5 125$ en 2013 à 17 631$ en 2023, avec un pic notable à 21 672$ en 2022. Pour les véhicules de plus de 190 chevaux, la sévérité des sinistres est bien plus élevée, évoluant de 8 057$ en 2013 à 40 326$ en 2023. L'augmentation est particulièrement prononcée de 2021 à 2022, atteignant son apogée à 40 361$.

Enfin, la charge pure suit une tendance à la hausse pour les deux catégories de véhicules. Pour les véhicules de moins de 190 chevaux, la charge pure croît graduellement correspondant à l'augmentation de la fréquence et de la sévérité des sinistres observées durant cette période. Concernant les véhicules de plus de 190 chevaux, la charge pure montre une forte croissance sur l'ensemble de la période, atteignant 0.48$ en 2023. Cela indique une augmentation significative du coût total des sinistres.

Globalement, à partir de 2017, les véhicules avec une puissance moteur supérieure à 190 sont toujours au minimum 3 fois plus risqués que les autres véhicules. La différence de risque entre ces deux modalités est extrêmement significative.

2.2.6 L'empattement du véhicule

La table 2.8 fait l'analyse des vols de véhicules en fonction de l'empattement du véhicule (distance entre les essieux). Deux catégories d'empattement sont comparées, les véhicules avec un empattement $\leq$ 2 745 mm et ceux avec un empattement > 2 745 mm.

Concernant la fréquence de vols (par 1000 véhicules assurés), on observe une augmentation progressive pour les deux catégories de véhicules. Pour les véhicules avec un empattement de 2 745 mm ou moins, la fréquence de vols est passée de 0.115% en 2013 à 0.357% en 2023. Cette tendance à la hausse est particulièrement marquée à partir de 2018, atteignant son sommet en 2022 à 0.438%. Pour les véhicules avec un empattement supérieur à 2 745 mm, la fréquence des vols est systématiquement plus élevée. Elle passe de 0.173% en 2013 à 1.181% en 2023, avec une croissance notable à partir de 2020.

La sévérité des sinistres montre également une augmentation considérable au fil des années. Pour les véhicules avec un empattement inférieur ou égal à 2 745 mm, la sévérité est passée de 5 101$ en 2013 à 23 873$ en 2023, avec un pic en 2022 à 24 451$. Cette augmentation est particulièrement prononcée entre 2021 et 2022. En ce qui concerne les véhicules avec un empattement supérieur à 2 745 mm, la sévérité des sinistres est plus élevée, passant de 7 911$ en 2013 à 39 850$ en 2023. Une forte augmentation est observée à partir de 2020, atteignant presque 40 000$ en 2022 et 2023.

	Empat $\leq$ 2745			Empat > 2745		
Année	Fréq. quotidienne (par 1000 véh.)	Sévérité	Charge pure quotidienne	Fréq. quotidienne (par 1000 véh.)	Sévérité	Charge pure quotidienne
2013	0.115%	5 101	0.01	0.173%	7 911	0.01
2014	0.137%	7 850	0.01	0.274%	11 915	0.03
2015	0.154%	10 617	0.02	0.288%	13 236	0.04
2016	0.178%	11 117	0.02	0.354%	13 916	0.05
2017	0.175%	9 848	0.02	0.431%	14 150	0.06
2018	0.207%	12 357	0.03	0.517%	14 997	0.08
2019	0.258%	12 569	0.03	0.503%	19 034	0.10
2020	0.266%	13 232	0.04	0.494%	26 280	0.13
2021	0.295%	16 840	0.05	0.629%	28 454	0.18
2022	0.438%	24 451	0.11	0.975%	39 709	0.39
2023	0.357%	23 873	0.09	1.181%	39 850	0.47

Table 2.8 – L'analyse des vols de véhicules par l'empattement du véhicule et pour chacune des années

Enfin, la charge pure montre une croissance continue pour les deux catégories de véhicules. Pour les véhicules avec un empattement inférieur ou égal à 2 745 mm, la charge pure a augmenté régulièrement, passant de 0.01$ en 2013 à 0.09$ en 2023. Cette augmentation est particulièrement notable entre 2021 et 2022, en ligne avec la hausse des autres indicateurs. Pour les véhicules avec un empattement supérieur à 2 745 mm, la charge pure a fortement augmenté, passant de 0.01$ en 2013 à 0.47$ en 2023, reflétant les augmentations significatives de la fréquence et de la sévérité, surtout à partir de 2021.

On constate une tendance claire, les véhicules avec un empattement plus élevé sont plus souvent la cible de vols et ces vols entraînent des pertes plus importantes. De plus, les coûts associés à ces vols ont considérablement augmenté ces dernières années, particulièrement après 2021.

2.2.7 Le type de véhicule

La table 2.9 compare ainsi les vols de camions aux autres types de véhicules. On peut voir que les camions présentent une fréquence de vols (par 1000 véhicules assurés) systématiquement supérieure à celle des

autres véhicules. En 2023, par exemple, la fréquence pour les camions était de 1.031% contre 0.323% pour les autres véhicules. Pour la sévérité des pertes, on remarque que les camions subissent des pertes financières bien plus importantes. En 2023, la sévérité pour les camions atteint 38 463$, soit presque le double de celle des autres véhicules à 23 243$. De plus, la charge pure pour les camions montre une augmentation constante, culminant à 0.40$ en 2023, illustrant une escalade des coûts associés aux vols de camions au fil du temps.

	Autres			Camions		
Année	Fréq. quotidienne (par 1000 véh.)	Sévérité	Charge pure quotidienne	Fréq. quotidienne (par 1000 véh.)	Sévérité	Charge pure quotidienne
2013	0.113%	4 772	0.01	0.173%	8 048	0.01
2014	0.157%	9 458	0.01	0.252%	11 210	0.03
2015	0.155%	8 915	0.01	0.282%	14 100	0.04
2016	0.202%	9 640	0.02	0.321%	14 914	0.05
2017	0.212%	10 339	0.02	0.378%	14 157	0.05
2018	0.266%	11 424	0.03	0.435%	15 715	0.07
2019	0.288%	11 981	0.03	0.448%	19 288	0.09
2020	0.303%	13 799	0.04	0.429%	25 697	0.11
2021	0.288%	13 366	0.04	0.576%	28 640	0.16
2022	0.325%	18 442	0.06	0.952%	38 497	0.37
2023	0.323%	23 243	0.08	1.031%	38 463	0.40

Table 2.9 – L'analyse des vols de véhicules par le type de véhicule et pour chacune des années

Concrètement, les différences de risque entre les camions et les autres véhicules sont importantes. Les camions sont de loin les plus risqués.

CHAPITRE 3
SÉLECTION DE VARIABLES

3.1 Introduction du modèle LASSO

Notre objectif dans ce chapitre est d'identifier les régresseurs les plus pertinents pour décrire la variable réponse. Cette tâche est l'une des premières étapes cruciales dans l'analyse de données, car nous cherchons à limiter le plus possible le nombre de régresseurs pour comprendre le risque de vols de voitures.

Dans le domaine de la statistique et de la modélisation, plusieurs méthodes ont été développées pour optimiser et standardiser ce processus, mais cela reste un défi. Parmi ces méthodes, on retrouve la technique des forêts aléatoires (Genuer et Poggi, 2017) et la méthode *stepwise* (Wagner et Shimshak, 2007). Pour ce mémoire, nous utiliserons le modèle LASSO (*Least Absolute Shrinkage and Selection Operator*), introduit par (Tibshirani, 1996). Ce modèle est particulièrement puissant pour deux tâches principales : la régularisation et la sélection de variables.

3.1.1 Description générale de la méthode

La méthode LASSO impose une contrainte sur la somme des valeurs absolues des paramètres du modèle, qui doit rester inférieure à une valeur fixée (limite supérieure). Cela se traduit par un processus de rétrécissement (régularisation), où les coefficients des variables de régression sont pénalisés, ce qui conduit à réduire certains d'entre eux à zéro. Lors de la sélection des variables, seules celles avec des coefficients non nuls après le rétrécissement sont retenues pour le modèle. L'objectif est de minimiser l'erreur de prédiction.

En pratique, un paramètre (qui est habituellement noté λ) joue un rôle crucial en contrôlant la force de la pénalité. Un λ élevé force davantage de coefficients à être exactement égaux à zéro, ce qui permet de réduire la dimensionnalité du modèle. Plus λ est grand, plus le nombre de coefficients réduits à zéro augmente.

L'un des principaux avantages de la méthode LASSO est qu'elle offre une bonne précision de prédiction, car la réduction et l'élimination des coefficients peuvent diminuer la variance sans augmenter de manière significative le biais. Cela est particulièrement utile lorsque le nombre d'observations est faible par rapport au nombre de variables. Le paramètre λ permet de trouver un équilibre entre biais et variance, puisque

l'augmentation de λ accroît le biais mais réduit la variance, et vice versa.

De plus, le modèle LASSO améliore l'interprétabilité en éliminant les variables non pertinentes, ce qui contribue à réduire le surapprentissage. Pour des exemples ou simplement pour obtenir plus d'informations, voir (Fonti et Belitser, 2017). Ce dernier avantage est particulièrement pertinent pour ce mémoire, car nous aurons à mettre en place un modèle qui a un nombre limité de régresseurs, ou en d'autres mots, nous avons un modèle qui possède ce qu'on pourrait appeler un *budget-variable*.

3.1.2 Approche dans un contexte de régression linéaire

Pour introduire le modèle LASSO et ses différentes généralisations, nous supposons tout d'abord que nous sommes dans un contexte de régression linéaire. Cette hypothèse de travail nous permettra de mieux expliquer le fonctionnement de l'approche, ce qui sera utile lorsque l'approche LASSO sera généralisée dans un contexte de régression linéaire généralisée.

3.1.2.1 Approche LASSO

Appliqué à la regression linéaire, le modèle LASSO est une technique de régularisation et de sélection de variables qui cherche à minimiser la somme des erreurs au carré tout en imposant une contrainte sur la somme des valeurs absolues des coefficients du modèle.

Plus formellement, l'estimation LASSO est déterminée par la résolution du problème d'optimisation suivant :

$$\min\left(\frac{\|\mathbf{Y}-\mathbf{X}\beta\|_2^2}{n}\right) \text{ sous la contrainte } \|\beta\|_1 < t,$$

où t représente la limite supérieure de la somme des coefficients. Ce problème est équivalent à l'estimation des paramètres décrite par :

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}}\left(\frac{\|\mathbf{Y}-\mathbf{X}\beta\|_2^2}{n} + \lambda\|\beta\|_1\right), \tag{3.1}$$

où $\|\mathbf{Y}-\mathbf{X}\beta\|_2^2$ est la somme des carrés des résidus et $\|\beta\|_1$ est la somme des valeurs absolues des coefficients.

Comme nous l'indiquions plus tôt, le paramètre λ contrôle la force de la pénalité : plus λ est élevé, plus la réduction des coefficients est importante.

3.1.2.2 Approche Ridge

Une autre approche, la régression Ridge, est aussi une possibilité à considérer, voir (Marquardt et Snee, 1975) pour plus d'informations. Ce modèle est plus approprié lorsque les variables indépendantes sont fortement corrélées. Elle réduit l'influence de certaines variables, mais ne peut pas les éliminer complètement, contrairement à l'approche LASSO.

L'estimation du modèle Ridge est obtenue en résolvant le problème d'optimisation suivant :

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(\frac{\|\mathbf{Y} - \mathbf{X}\beta\|_2^2}{n} + \lambda \|\beta\|_2^2 \right), \tag{3.2}$$

où $\|\beta\|_2^2$ est la somme des carrés des coefficients.

3.1.2.3 Régularisation élastique

La régularisation élastique (*Elastic-Net*) combine les régularisations Ridge et LASSO pour éviter la sélectivité excessive de LASSO tout en maintenant la possibilité de traiter des variables fortement corrélées. L'estimation de la régularisation élastique est obtenue en résolvant :

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{X}_i\beta)^2 + \lambda \left(\frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right) \right], \tag{3.3}$$

avec les paramètres suivants :

λ : Un paramètre de réglage qui ajuste le niveau de régularisation ($\lambda \geq 0$) ;

α : Un hyperparamètre qui équilibre les contributions de LASSO et Ridge ($\alpha \in [0,1]$) ;

$\|\beta\|_2^2$: Terme de régularisation Ridge (ou L_2) ;

$\|\beta\|_1$: Terme de régularisation LASSO (ou L_1).

La première partie de l'équation, $\frac{1}{n}\sum_{i=1}^{n} \left(y_i - \beta_0 - x_i^T\beta\right)^2$, évalue l'adéquation du modèle aux données en mesurant la somme des carrés des écarts entre les valeurs observées et les valeurs prédites, avec l'objectif de minimiser l'erreur. La seconde partie, $\lambda \left(\frac{1-\alpha}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1\right)$, régule la complexité du modèle en pénalisant les coefficients élevés, ce qui aide à éviter le surajustement et améliore la capacité de généralisation. La composante LASSO ($\alpha\|\beta\|_1$) favorise la parcimonie en permettant à certains coefficients d'être exactement nuls, tandis que la composante Ridge ($\frac{1-\alpha}{2}\|\beta\|_2^2$) favorise des coefficients plus petits et uniformément répartis, ce qui aide à gérer la multicolinéarité.

3.1.3 Application dans un contexte de régression linéaire généralisée

Contrairement à la régression linéaire discutée dans la sous-section 3.1.2, qui pénalise la somme des erreurs quadratiques, nous supposons que la variable réponse fait partie de la famille exponentielle linéaire, et que la théorie de la régression linéaire généralisée de la sous-section 1.2 s'applique. Plus précisément, afin de pouvoir travailler avec le nombre de vols de voitures, nous supposerons une loi de Poisson avec un lien logarithmique.

La log-vraisemblance pour la distribution de Poisson est exprimée comme suit :

$$\ell(\beta) = \sum_{i=1}^{n} \left(y_i (\mathbf{X}_i \beta) - e^{\mathbf{X}_i \beta} - \ln(y_i!) \right). \tag{3.4}$$

Dans l'approche généralisant la théorie des GLM, la log-vraisemblance est utilisée à la place des moindres carrés que nous voyons dans le premier terme de l'équation (3.3). Ainsi, l'estimation de la régularisation élastique dans le cadre de la loi de Poisson est obtenue en trouvant la solution de :

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left[-\frac{1}{n} \ell(\beta) + \lambda \left(\frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right) \right]. \tag{3.5}$$

Pour le cas plus précis de la pénalisation L_1 appliquée à la loi de Poisson, comme défini par (Kondofersky et Theis, 2018), l'estimation de l'approche LASSO est obtenue en résolvant :

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left[-\frac{1}{n} \ell(\beta) + \lambda \|\beta\|_1 \right]. \tag{3.6}$$

3.1.3.1 Application aux données de vols de voitures

Tel que mentionné plus tôt, nous avons choisi de travailler avec le modèle LASSO en raison de notre problème de budget de variables. Dans le contexte des données de vols d'automobiles, avec une vaste base de données contenant de nombreuses covariables, il est crucial de réduire le nombre de variables tout en conservant les plus pertinentes.

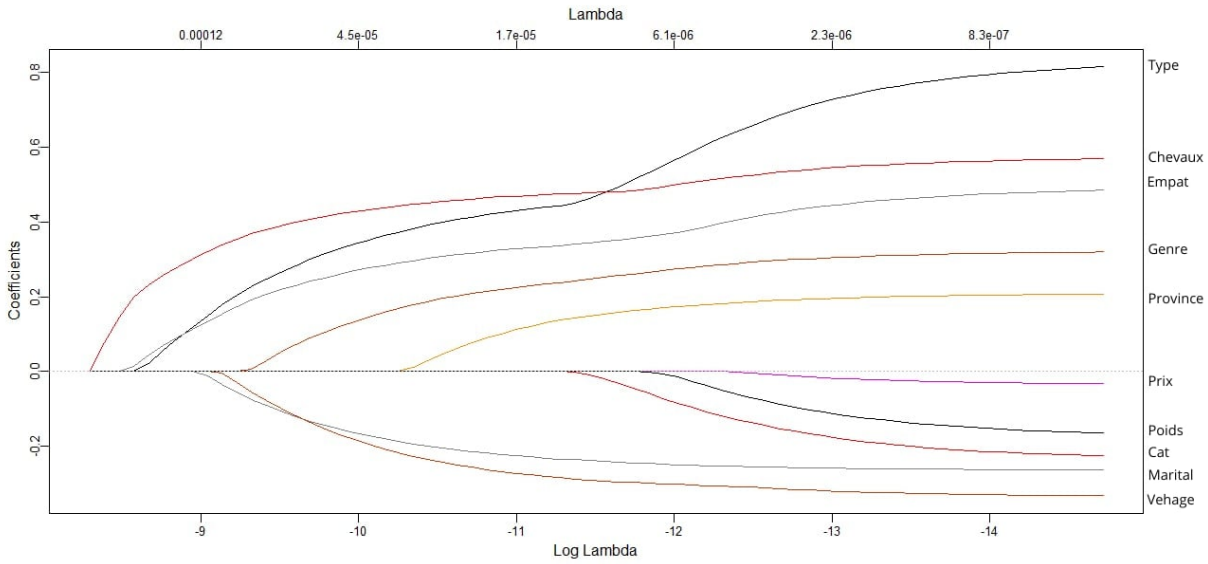


Figure 3.1 – Les variables sélectionnées par le modèle LASSO

Les saisons, la marque de véhicule, le modèle de véhicule ainsi que l'emplacement n'ont pas été ajoutés dans l'évaluation LASSO. Nous avons décidé d'exclure la covariable des saisons. Pour ce qui est des trois autres covariables, nous les avons exclues, car le nombre de modalités de chacune de ces covariables est très grand. Nous avons ainsi appliqué le modèle sur les variables restantes de la figure 2.2 de la section 2.2.

La figure 3.1 illustre le résultat obtenu. Les variables sont sélectionnées de gauche à droite en ordre d'importance. On peut faire les observations suivantes :

— L'approche LASSO a établi que la puissance du véhicule (Chevaux) est la covariable la plus pertinente, suivi de l'empattement du véhicule (Empat), le type de véhicule (Type), l'état matrimonial légal (Marital), l'âge du véhicule (Vehage), le sexe de l'assuré (Genre), la province de l'assuré (Province), la catégorie du véhicule (Cat), le poids du véhicule (Poids) et finalement, le prix d'achat (Prix) ;

— La puissance du véhicule entre dans le modèle en premier et a un effet positif et progressif sur la variable réponse ;

— L'état matrimonial légal est la quatrième variable la plus importante, entrant dans le modèle avec un effet négatif sur la variable réponse ;

— Comme le montre le graphique, la catégorie du véhicule influence la tendance du type de véhicule après son inclusion dans le modèle.

Le modèle nous permet ainsi de sélectionner les variables explicatives en fonction d'un budget de variables, tel que nous le verrons dans un prochain chapitre.

CHAPITRE 4
SURVEILLANCE

Ce chapitre a pour objectif de présenter un aperçu des méthodes de surveillance les plus pertinentes, telles que le modèle Farrington et le modèle GLR (Generalized Likelihood Ratio), afin de les adapter à l'analyse des statistiques de sinistralité en assurance, spécifiquement pour le nombre de vols complets de véhicules.

La surveillance en santé publique cherche à atténuer la charge de morbidité en identifiant rapidement les épidémies émergentes, notamment dans le cas des maladies infectieuses. D'un point de vue statistique, cela nécessite l'emploi de méthodes appropriées pour surveiller les séries chronologiques des cas agrégés et détecter les anomalies. Le paquet *surveillance* de R propose des outils pour la détection automatique des anomalies. La conception d'algorithmes pour la surveillance des maladies infectieuses s'inspire des techniques de contrôle statistique des processus (CSP).

Le contrôle statistique des processus (CSP) est une méthodologie efficace utilisant des techniques statistiques pour *surveiller*, *analyser* et *contrôler* les processus, afin d'assurer leur fonctionnement optimal. En mesurant et en comprenant les variations au sein des processus, le CSP aide à identifier et éliminer les inefficacités, améliorant ainsi la qualité des produits et services. Le CSP inclut l'utilisation d'outils statistiques pour surveiller et contrôler un processus ou une méthode de production, permettant aux organisations de suivre le comportement des processus, de détecter les problèmes internes et de résoudre les défis de production. Parmi les outils couramment utilisés dans le CSP figure la carte de contrôle, développée par Walter Shewhart dans les années 1920. Les cartes de contrôle enregistrent les données et signalent visuellement les événements inhabituels, comme des valeurs exceptionnellement élevées ou faibles par rapport aux performances normales d'un processus. Elles distinguent deux types de variations : la variation de cause commune, inhérente au processus et toujours présente, et la variation de cause spéciale, provenant de sources externes et indiquant que le processus est hors de contrôle statistique. Les outils de contrôle de qualité incluent également le diagramme de cause à effet (diagramme d'Ishikawa), la feuille de contrôle, l'histogramme, le diagramme de Pareto, le diagramme de dispersion et la stratification. L'objectif principal du CSP est d'analyser les données et d'améliorer continuellement les processus.

4.1 Historique

Les premières méthodes de surveillance, comme celles de (Stroup *et al.*, 1989) et (Farrington *et al.*, 1996), reposaient principalement sur l'utilisation répétée d'intervalles de confiance, une approche qui ne prenait pas en compte l'ensemble des données passées pour l'évaluation des statistiques. La surveillance moderne, telle que décrite dans (Lawson et Kleinman, 2005), s'appuie sur les avancées issues de la littérature sur le contrôle statistique des processus (CSP), comme le montrent (Frisén et Wessman, 1999) et (Wood, 2006). Cependant, les séries temporelles de comptages issues des données de surveillance présentent des caractéristiques particulières qui ne sont pas toujours couvertes par les méthodes CSP traditionnelles et nécessitent des solutions spécifiques. Une approche consiste à intégrer les informations sur les covariables, telles que les variations saisonnières de la moyenne, les ajustements pour les populations à risque, ou d'autres facteurs explicatifs. Cette approche repose sur des graphiques de régression utilisant des modèles linéaires généralisés (GLM).

La littérature statistique et technique propose diverses méthodes pour examiner la stabilité temporelle des relations de régression, telles que les diagrammes de régression à réponse normale (Brown *et al.*, 1975). Les tests de rapport de vraisemblance, comme ceux décrits par (Kim et Siegmund, 1989), permettent de détecter les points de changement dans la régression linéaire simple en considérant différentes alternatives. Les problèmes de détection de changements, les critères pour concevoir et analyser les performances des techniques de détection des changements, ainsi que la conception et l'étude d'algorithmes de détection des changements sont également abordés dans (Basseville et Nikiforov, 1993). De plus, une théorie unifiée de la détection séquentielle des points de changement est introduite par (Lai, 1995).

Les modèles gaussiens de séries temporelles, tels que les modèles autoregressifs (AR) et certaines généralisations comme les approches ARIMA (AutoRegressive Integrated Moving Average), supposent que les données suivent une distribution normale. Cependant, pour les maladies infectieuses rares ou lorsque la répartition des cas varie en fonction de facteurs comme l'âge, le sexe et le lieu, les séries temporelles peuvent contenir un faible nombre de cas et ne pas suivre une distribution normale. Les modèles gaussiens ne sont pas adaptés à ces situations, car ils ne capturent pas correctement les distributions discrètes. Concernant la détection des valeurs aberrantes, les modèles gaussiens utilisent souvent l'analyse des résidus pour identifier les observations atypiques, mais cette méthode peut être peu fiable lorsque les données ne suivent pas une distribution normale. Ainsi, il est préférable d'utiliser des modèles de distribution discrète, plus adaptés aux données observées, qui tiennent compte des caractéristiques spécifiques des données, comme les

faibles nombres de cas, pour offrir des prédictions plus précises.

Des exemples de graphiques de régression basés sur des GLM sont trouvés dans la littérature sur le CSP, comme le décrit (Skinner *et al.*, 2003). Cette procédure de surveillance pour les données de comptage multiples utilise la statistique du rapport de vraisemblance pour les données suivant une loi de Poisson lorsque les variables d'entrée sont mesurables. La littérature sur la surveillance, telle que (Rogerson et Yamada, 2004), présentent également l'utilisation de méthodes de somme cumulative (CUSUM) pour cumuler les écarts entre les données observées et attendues selon une loi de Poisson. Ces écarts sont ajustés pour tenir compte des variations au fil du temps, y compris les effets hebdomadaires et mensuels. Les objectifs de cette étude étaient de concevoir et d'illustrer un système de surveillance multirégional basé sur des données constituées de petites occurrences régionales, où les fréquences sont généralement inférieures ou égales à 5.

4.2 Le modèle Farrington

Le modèle de régression quasi-Poisson a été utilisé par (Farrington *et al.*, 1996) pour permettre une détection rapide des foyers de maladies infectieuses, facilitant ainsi la mise en place de mesures de contrôle efficaces. Ce modèle est généralement appliqué au nombre d'occurrences par semaine, bien qu'il puisse également être adapté pour des périodes différentes, comme le nombre d'occurrences par mois ou par jour. L'approche de Farrington repose sur la théorie des modèles linéaires généralisés (GLM) pour prédire le nombre d'occurrences d'un événement à chaque moment donné. En comparant les prédictions à un historique de valeurs jugées similaires, un seuil supérieur est déterminé pour chaque observation en fonction d'un quantile spécifique de l'intervalle de prédiction. Si l'observation dépasse ce seuil, une alerte est déclenchée pour signaler un nombre excessif de cas de maladie.

4.2.1 Historique de référence

Une des complexités de l'approche de Farrington est le choix de l'historique à utiliser pour vérifier si une observation est aberrante. Notons ainsi tout d'abord les éléments suivants :

- Nous travaillons avec $t = 1, 2, \ldots, T$ observations hebdomadaires. Ainsi, nous aurons $\{x_t\}_{t=1}^{T}$, soit une série temporelle univariée de comptage ;
- Le temps actuel de l'observation que nous cherchons à analyser est noté t_0, et la variable $x_{t_0} \equiv x_0$

représente l'observation sous surveillance ;

— L'information historique disponible pour fin de comparaison sera incluse dans la filtration $\mathscr{H}_{t_0} = \{x_t ; t \leq t_0\}$.

Un sous-ensemble des données de référence $\mathscr{H}_{t_0}$, noté $\mathscr{H}^*_{t_0}$, sera par la suite utilisé. L'utilisation d'un sous-ensemble de données a pour objectif de capturer les cycles et la saisonnalité. Les effets saisonniers sont pris en compte en basant le calcul du seuil uniquement sur des périodes comparables des années passées.

Ce sous-ensemble sera basé sur les paramètres suivants :

— Le nombre d'années passées considérées dans l'historique de comparaison, noté b ;

— Pour chacune des années passées, la taille de la demi-fenêtre, c'est-à-dire le nombre de semaines à inclure avant et après la semaine de référence t_0, noté w.

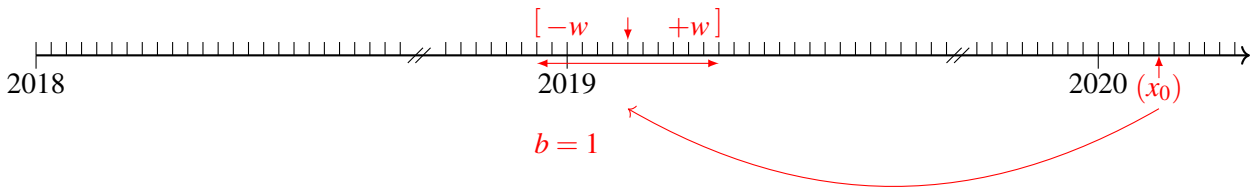
Au total, nous aurons ainsi $n = b(2w+1)$ observations de référence pour $\mathscr{H}^*_{t_0}$. L'intérêt d'une grande valeur de n pour augmenter la précision doit être balancé avec la nécessité d'une demi-fenêtre w appropriée par rapport à l'échelle de temps des variations saisonnières et le nombre d'années b sélectionné de façon à obtenir l'information passée assez récente.

Exemple 4 (Données de référence) *Illustrons par un exemple de quelle manière on peut utiliser b et w pour définir $\mathscr{H}^*_{t_0}$. Supposons que nous cherchons à analyser le nombre de cas hebdomadaires x_0 pour le temps t_0, soit la 4e semaine de l'année 2020 :*

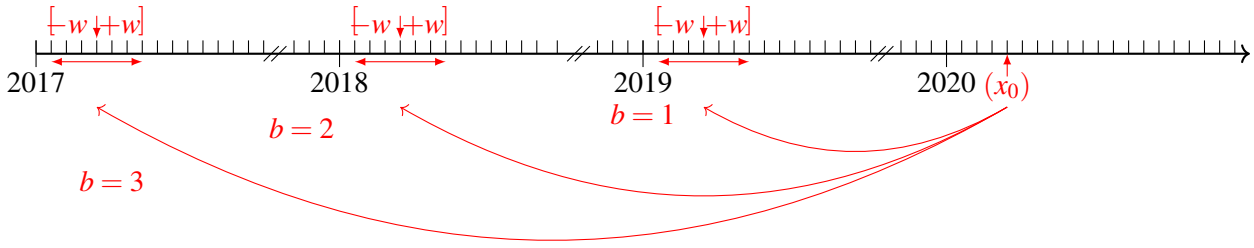
— *Si nous posons $w = 6$ et $b = 1$ comme dans la figure 4.1a, cela signifie que nous utiliserons seulement l'année précédente dans l'analyse et une demi-fenêtre de 6 semaines, avant et après la semaine équivalente à t_0 sera considérée. Donc, avec t_0 correspondant à la 4e semaine de 2020, on se basera uniquement sur les observations des semaines 50 à 52 de l'année 2018 ainsi que sur les semaines 1 à 10 de l'année 2019. Au total, nous aurons ainsi $n = b(2w+1) = 13$ observations pour vérifier si x_0 est une valeur aberrante. Ce procédé sera ensuite appliqué à chacune des autres semaines à analyser ;*

— *Si nous posons cette fois $w = 3$ et $b = 3$ comme dans la figure 4.1b, cela signifie que nous considérons les trois années précédentes à t_0 dans l'analyse et une demi-fenêtre de 3 semaines, avant et après la semaine équivalente à t_0 sera considérée. Dans notre choix de t_0, on se basera donc sur*

les observations des semaines 1 à 7 de chacune des années 2017, 2018 et 2019. Au total, nous aurons $n = b(2w+1) = 21$ semaines de référence. Encore une fois, ce procédé sera ensuite appliqué à chacune des semaines à analyser.



(a) Les valeurs utilisées (semaines de référence) pour une observation donnée sous surveillance avec $b = 1$ et $w = 6$



(b) Les valeurs utilisées (semaines de référence) pour une observation donnée sous surveillance avec $b = 3$ et $w = 3$

Figure 4.1 – Les données sélectionnées pour différentes années et différentes demi-fenêtres de l'exemple 4

4.2.2 Le modèle de régression

Le but du modèle à développer est de prédire une variable réponse de comptage. Bien que l'utilisation d'une distribution de Poisson puisse être envisagée, il est proposé d'opter pour une loi quasi-Poisson. Pour un temps donné t_0, les n observations de référence $x_i, i = 1, \ldots, n$ incluses dans $\mathscr{H}^*_{t_0}$ sont supposées indépendantes, avec une moyenne μ_i et une variance $\phi\mu_i$. Ces paramètres sont estimés à l'aide de la méthode de quasi-vraisemblance. La loi quasi-Poisson est utilisée pour modéliser le nombre de cas comme suit :

$$\mu_i = \exp(\alpha + \beta t_i + \log(d_i)),$$

où t_i est le numéro de la semaine et d_i représente l'exposition. Une fois les paramètres α et β estimés, la prédiction $\hat{\mu}_0$ peut être calculée :

$$\hat{\mu}_0 = \exp(\hat{\alpha} + \hat{\beta} t_0 + \log(d_0)), \tag{4.1}$$

où t_0 est le numéro de la semaine analysée et d_i représente l'exposition correspondant à la semaine analysée. Ici, β représente la tendance linéaire des fréquences au fil du temps. Les données de comptage peuvent parfois présenter de la sur-dispersion ou de la sous-dispersion, ce qui signifie que la variance ne sera pas égale à la moyenne. En cas de sous-dispersion, la variance sera inférieure à la moyenne, tandis qu'en cas de surdispersion, la variance sera supérieure à la moyenne. La loi quasi-Poisson prend en compte cette situation. Le paramètre ϕ, appelé paramètre de dispersion, indique la présence de surdispersion ou de sous-dispersion. Bien que l'estimation des paramètres α et β reste inchangée, leurs écarts-types seront multipliés par $\sqrt{\phi}$, ce qui pourrait rendre certains paramètres non significatifs.

L'estimation du paramètre de dispersion se fait en utilisant l'équation suivante :

$$\hat{\phi} = \max\left\{ \frac{1}{n-p} \sum_{i=1}^{n} \frac{(x_i - \hat{\mu}_i)^2}{\hat{\mu}_i}, 1 \right\},$$

où p représente le nombre de paramètres inclus dans le modèle, soit α et β dans la paramétrisation proposée. Le calcul de ϕ est crucial pour l'analyse des valeurs aberrantes. Dans ce cas, la valeur du paramètre de dispersion sera utilisée pour calculer l'intervalle de confiance, et si la donnée sous surveillance dépasse le seuil supérieur, une alerte sera déclenchée, un point qui sera discuté en détail ultérieurement. Il convient de noter que, pour éviter des extrapolations irréalistes, la tendance temporelle β n'est incluse dans le modèle que si les données historiques couvrent au moins trois ans, si elle est significative au niveau de 5%, et si $\hat{\mu}_0 \leq \max\{x_i : i = 1, 2, \ldots, n\}$, comme appliqué dans (Farrington *et al.*, 1996).

4.2.3 Correction des données passées

Une difficulté du modèle de régression de Farrington réside dans la gestion de l'influence de certaines observations historiques dans le sous-ensemble de comparaison $\mathscr{H}_t^*$. Il est possible que certaines semaines passées coïncident avec des épidémies. L'intégration de ces périodes épidémiques dans l'historique de comparaison et le calcul des seuils peut entraîner des niveaux d'alarme excessifs, ce qui diminue la capacité du modèle à détecter les véritables anomalies.

Pour pallier ce problème, une surveillance manuelle de l'historique de comparaison $\mathscr{H}_t^*$ pourrait être effectuée pour identifier et exclure les valeurs aberrantes des calculs. Cependant, cette méthode n'est pas automatique et manque de praticité. Une approche alternative consiste à appliquer une procédure de repondération pour atténuer l'influence des épidémies passées. La fonction de repondération est particulièrement utile en attribuant des poids très faibles aux observations passées présentant des résidus élevés. Bien que cette repondération réduise significativement l'effet des épidémies antérieures, elle ne l'élimine pas complètement (voir la sous-section 3.3 de (Farrington *et al.*, 1996) pour plus de détails).

Cette méthode repose sur les estimations initiales de μ_i et de ϕ. Les résidus s_i sont définis par :

$$s_i = \frac{3\left(x_i^{2/3} - \hat{\mu}_i^{2/3}\right)}{2\hat{\phi}^{1/2}\hat{\mu}_i^{1/6}\left(1-h_{ii}\right)^{1/2}},$$

où h_{ii} sont les éléments diagonaux de la matrice chapeau $H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$, avec W une matrice diagonale de poids provenant de l'itération finale de l'ajustement des moindres carrés pondérés itératifs (IWLS). Il convient de noter que pour des données issues d'une distribution de Poisson (où $\phi = 1$), les s_i correspondent aux résidus standardisés d'Anscombe (Davison et Tsai, 1992).

Les poids w_i sont ensuite définis comme suit :

$$w_i = \begin{cases} \gamma s_i^{-2}, & \text{si } s_i < 1 \\ \gamma, & \text{sinon} \end{cases} \tag{4.2}$$

où γ est une constante telle que $\sum_{i=1}^{n} w_i = n$.

Ces poids sont ensuite utilisés pour recalculer le paramètre de dispersion ϕ :

$$\hat{\phi} = \max\left\{\frac{1}{n-p}\sum_{i=1}^{n} w_i \frac{(x_i - \hat{\mu}_i)^2}{\hat{\mu}_i}, 1\right\}.$$

Si une repondération est appliquée, c'est cette valeur recalculée de ϕ qui sera utilisée pour évaluer les seuils.

4.2.4 Le calcul du seuil et du score

Pour déterminer si une observation est anormale, il est nécessaire de définir un intervalle de prédiction. Si la valeur observée dépasse la limite supérieure de cet intervalle, l'observation est considérée comme anormale. Dans (Farrington *et al.*, 1996), trois seuils différents sont proposés pour ajuster les données et obtenir des intervalles de prédiction appropriés. Dans les cas où il existe une asymétrie, une correction peut être nécessaire. Cette correction peut être réalisée en appliquant une transformation exponentielle de $\frac{2}{3}$, avec :

$$E(x_0^{2/3}) = \hat{\mu}_0^{2/3},$$

$$var(x_0^{2/3}) = \frac{4}{9}\phi\hat{\mu}_0^{1/3},$$

et

$$var(\hat{\mu}_0^{2/3}) = \frac{4}{9}\hat{\mu}_0^{-2/3}var(\hat{\mu}_0).$$

L'erreur de prédiction de la variance après la transformation exponentielle de $\frac{2}{3}$ est donnée par

$$var(x_0^{2/3} - \hat{\mu}_0^{2/3}) = \frac{4}{9}\tau\hat{\mu}_0^{1/3},$$

où

$$\tau = \phi + var(\hat{\mu}_0)/\hat{\mu}_0.$$

L'intervalle de prédiction est donc défini comme suit : pour la borne supérieure,

$$U = \left\{ \hat{\mu}_0^{2/3} + z_{1-\alpha} \sqrt{\frac{4}{9} \hat{\mu}_0^{1/3} \tau} \right\}^{3/2}, \tag{4.3}$$

et pour la borne inférieure,

$$L = \max \left\{ \left\{ \hat{\mu}_0^{2/3} - z_{1-\alpha} \sqrt{\frac{4}{9} \hat{\mu}_0^{1/3} \tau} \right\}^{3/2}, 0 \right\},$$

où $z_{1-\alpha}$ est le percentile de la distribution normale, $\hat{\mu}_0$ est la prédiction obtenue, et $var(\hat{\mu}_0)$ est la variance associée. Il est important de noter que l'erreur de prédiction de la variance après la transformation exponentielle $\frac{2}{3}$ se trouve sous la racine carrée de l'intervalle.

Deux autres seuils supérieurs sont mentionnés dans (Farrington *et al.*, 1996) pour évaluer les valeurs anormales. Il y a sans transformation et la transformation racine carrée. Les seuils supérieurs pour ces deux cas sont respectivement :

$$U = \left\{ \hat{\mu}_0 + z_{1-\alpha} \sqrt{\hat{\mu}_0 \tau} \right\},$$

$$U = \left\{ \hat{\mu}_0^{1/2} + z_{1-\alpha} \sqrt{\frac{1}{4} \tau} \right\}^2.$$

La transformation racine carrée est utile pour stabiliser la variance dans le cadre d'une distribution de Poisson. Une fois les seuils établis, les résultats peuvent être classés en fonction du score de dépassement, défini comme :

$$X = \frac{x_0 - \hat{\mu}_0}{U - \hat{\mu}_0}. \tag{4.4}$$

Pour éviter des alertes lorsque la série temporelle comporte très peu de cas, l'algorithme applique un critère heuristique (voir la section 3.8 de (Farrington *et al.*, 1996)) pour se protéger contre les faibles occurrences.

Aucune alerte n'est déclenchée si moins de 5 cas sont observés dans les 4 semaines précédant l'observation. Les valeurs du nombre de cas ainsi que le nombre de semaines peuvent être ajustées par l'utilisateur. De plus, aucune alerte n'est déclenchée pour des valeurs extrêmement élevées si les quatre semaines précédentes sont particulièrement faibles. Le score de dépassement X de l'équation (4.4) est fixé à 0 si moins de cinq cas ont été signalés au cours des quatre semaines précédentes. Les cas où la valeur de X dépasse 1 sont alors signalés pour un examen plus approfondi. Le seuil minimal de 5 cas sur les quatre semaines précédentes réduit la probabilité que des cas sporadiques soient signalés. Si le nombre total de cas n'est jamais inférieur à 5 dans les 4 semaines précédentes, on peut comparer directement les observations avec le seuil supérieur obtenu.

4.2.5 Application du modèle Farrington sur les données de vols de voitures

Nous cherchons à identifier des anomalies statistiques dans le nombre de cas hebdomadaires observés. Comme mentionné précédemment, nous allons examiner les statistiques de vols de voitures sur la période allant de la 4e semaine de 2018 jusqu'à la dernière semaine du mois de septembre de l'année 2023. La méthode Farrington peut être directement appliquée à l'aide du logiciel R en utilisant le paquet *surveillance* (voir (Salmon *et al.*, 2016) pour plus de détails). La fonction *farringtonFlexible* incluse dans ce paquet a été employée pour estimer les statistiques et générer le graphique.

Pour adapter cette méthode au contexte de l'assurance, l'exposition d_i et le nombre de vols de voitures par semaine seront utilisés. Formellement, l'ajustement du modèle sera basé sur la forme paramétrique suivante :

$$\log \mu_i = \alpha + \beta t_i + \log(d_i). \tag{4.5}$$

La procédure d'estimation du modèle de Farrington reste inchangée. En ce qui concerne les paramètres utilisés dans la méthode Farrington, nous avons choisi $b = 5$ années et une demi-fenêtre $w = 3$, soit trois semaines avant et après la semaine équivalente à t_0. Ce choix est motivé par la nécessité de conserver des observations de comparaison suffisamment récentes tout en ayant une quantité de données suffisante. Un grand nombre d'observations passées est essentiel pour bien capturer l'information dans les cas de changements majeurs et pour tenir compte de la saisonnalité. Nous avons donc opté pour cinq années

avec une fenêtre totale de sept semaines afin de disposer de suffisamment de données historiques pour analyser les valeurs aberrantes.

La repondération de l'équation (4.2) ainsi que l'évaluation de la tendance temporelle linéaire de l'équation (4.5) seront prises en compte. Pour le calcul du seuil, la transformation $\frac{2}{3}$ sera appliquée, et un niveau $\alpha = 1\%$ pour le percentile de la distribution normale de l'équation (4.3) sera utilisé.

Le graphique de la figure 4.2, illustre les résultats obtenus. Les triangles rouges représentent les valeurs où les observations de vols de voitures ont dépassé le seuil. La ligne bleue pointillée indique les seuils supérieurs obtenus, et les barres grises correspondent aux observations, c'est-à-dire aux vols de véhicules complets survenus chaque semaine durant les années surveillées.

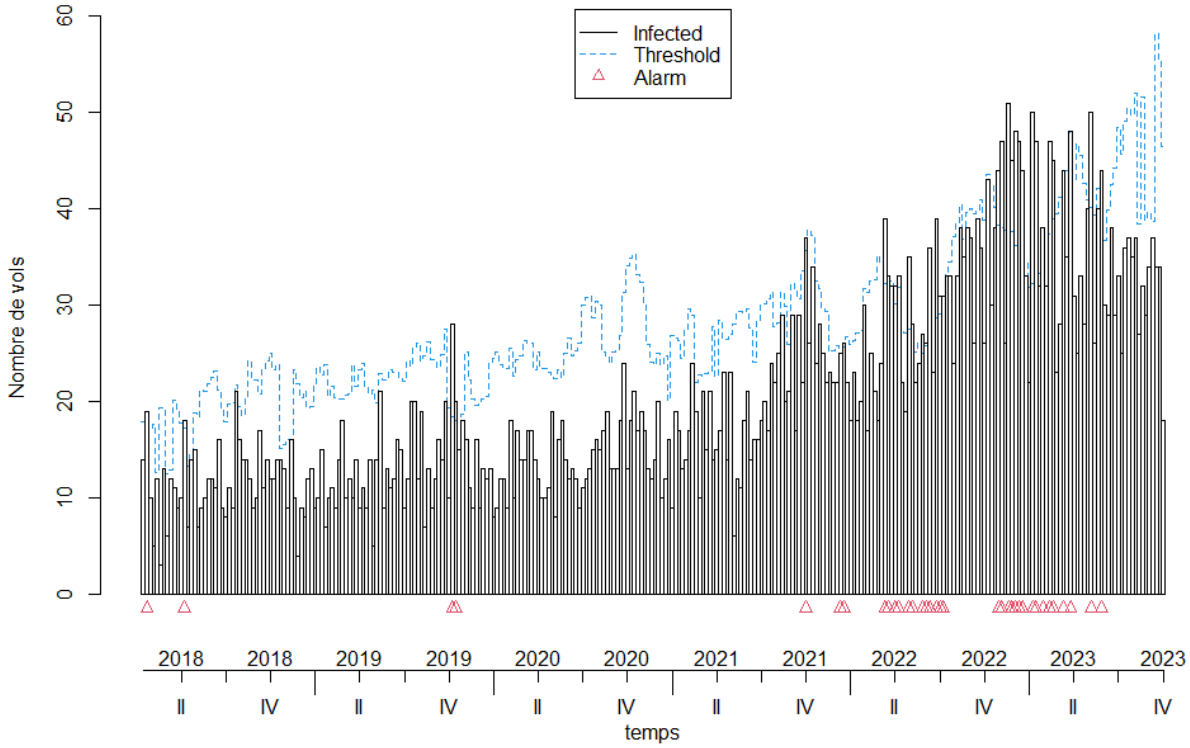


Figure 4.2 – La méthode Farrington

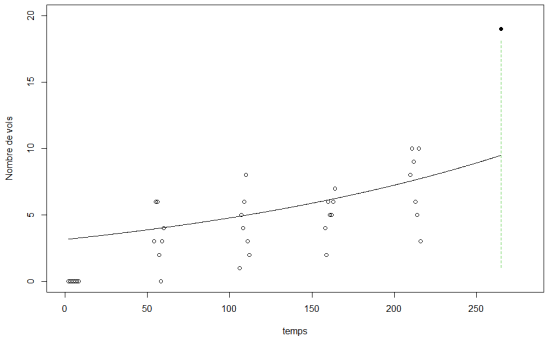
Nous pouvons faire quelques constatations provenant de la figure 4.2 :

1. Les premières alertes apparaissent lors de la 4e et de la 15e semaine de 2018, avec un nombre de vols observés de 19 et 18 respectivement. Cela est en partie dû à la présence de nombreuses semaines avec zéro vol dans l'historique de comparaison, notamment en 2013. Les baisses ultérieures semblent principalement dues à la saisonnalité. Les scores de dépassement pour ces alertes sont de 1,107 et 1,096 respectivement, légèrement au-dessus de 1. Aucun autre signe d'alerte n'est détecté avant la 41e semaine de 2019 ;
2. Deux nouvelles alertes sont notées aux semaines 41 et 42 de 2019, avec 28 et 20 vols respectivement. Le nombre de 28 vols est particulièrement élevé, surtout lorsque l'on considère que les onze semaines précédentes et suivantes ne dépassent pas 20 vols. Cette hausse peut également être liée à la saisonnalité. Les scores de dépassement pour ces alertes sont de 2,150 et 1,238 respectivement ;
3. À partir de la 10e semaine de 2022, les alertes se multiplient rapidement. De nombreux messages auraient été envoyés à l'assureur, signalant d'éventuelles anomalies. Deux périodes marquantes se distinguent, la première dès le 2e trimestre 2022, suivie d'une pause, puis une nouvelle vague d'alertes commence à la fin du 4e trimestre 2022, et se poursuit jusqu'à la fin du 2e trimestre 2023. Il est étonnant de constater cette interruption durant les 3e et 4e trimestres de 2022, d'autant plus que les observations suivant ces alertes semblent s'intensifier au lieu de diminuer. Une analyse approfondie s'impose pour comprendre ce phénomène.

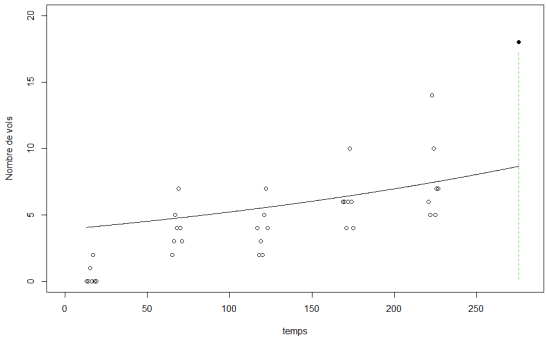
En résumé, entre la 3e semaine de 2018 et la fin de la 38e semaine de 2021, l'assurance a reçu 4 alertes sur 193 semaines, ce qui représente environ 2% des semaines dans cette période. Chaque alerte nécessite une enquête rapide par un expert pour identifier et gérer les anomalies efficacement, afin de minimiser les pertes potentielles.

Avec l'historique de comparaison $\mathscr{H}_t^*$ utilisé pour chaque observation t, il a été possible de déterminer si une observation était aberrante ou non. Des analyses plus approfondies peuvent être menées sur certaines observations pour mieux comprendre le modèle. Les trois graphiques de la figure 4.3 examinent trois observations spécifiques de la figure 4.2 : la 265e semaine (figure 4.3a), correspondant à la 4e semaine de 2018, la 276e semaine (figure 4.3b), correspondant à la 15e semaine de 2018, et la 516e semaine (figure 4.3c), correspondant à la 46e semaine de 2022. Les historiques de comparaison $\mathscr{H}_{265}^*$, $\mathscr{H}_{276}^*$ et $\mathscr{H}_{516}^*$ sont sélectionnés avec $b = 5$ et $w = 3$. Chaque observation est comparée à $b = 5$ groupes de $2w + 1 = 7$ données. Le point noir plein représente l'observation x_0 de la semaine en cours, tandis que les points ouverts montrent les valeurs de référence dans $\mathscr{H}_t^*$. La courbe noire ajoutée indique une tendance qui peut être intégrée

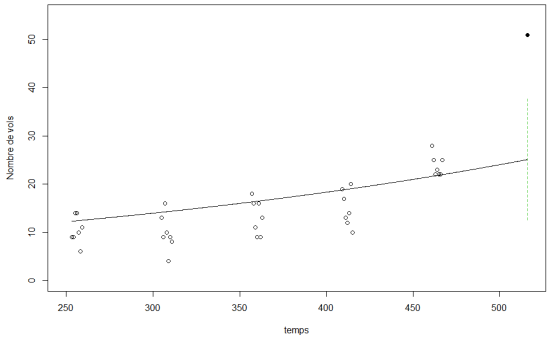
au modèle pour refléter l'évolution temporelle. L'intersection entre la ligne verte pointillée et cette courbe représente la valeur attendue du modèle, notée $\hat{\mu}_0$. Les extrémités de la ligne verte pointillée indiquent les seuils supérieur et inférieur du modèle de prédiction.



(a) Estimation de la prédiction à t_0 = 265



(b) Estimation de la prédiction à t_0 = 276



(c) Estimation de la prédiction à t_0 = 516

Figure 4.3 – L'analyse des deux premières alertes et d'une alerte en 2022

Dans les figures 4.3a et 4.3b, on observe que les valeurs sous surveillance sont nettement plus élevées que les valeurs de l'historique de comparaison, avec des intervalles de confiance très larges et une tendance générale à la hausse. Les valeurs sous surveillance dépassent le seuil supérieur, entraînant une alerte pour une investigation. Toutefois, étant donné que les valeurs sont relativement proches du seuil supérieur, il est possible que les alertes soient sans conséquences significatives.

En revanche, la figure 4.3c présente une situation légèrement différente. Les valeurs historiques y sont très proches les unes des autres, et l'intervalle de confiance est beaucoup plus étroit par rapport aux figures

précédentes. Dans ce cas, la valeur sous surveillance est loin au-dessus du seuil supérieur.

4.2.6 Commentaires à propos du modèle Farrington

Un des inconvénients de la méthode est qu'elle utilise uniquement un échantillon des observations passées, ce qui n'est pas toujours optimal. En particulier, les données de la dernière année en cours ne sont pas prises en compte pour le calcul des seuils. Cela signifie que si une modification importante se produit dans l'année en cours, la méthode Farrington ne pourra pas l'intégrer, car elle ne prend pas en compte ces observations récentes. Ainsi, cette méthode suppose une certaine stabilité dans le comportement du phénomène étudié.

De plus, le calcul des seuils ne prend pas en compte les corrélations temporelles entre les différents ensembles de comparaison $\mathscr{H}_t^*$. Pour les vols rares et sporadiques, les corrélations sont généralement faibles. Cependant, pour les vols plus fréquents, les dynamiques d'augmentation du nombre de vols et les effets résiduels de la saisonnalité peuvent engendrer des corrélations plus marquées.

Une autre limitation du modèle Farrington est le nombre restreint de semaines de référence utilisé dans le modèle. Avec les changements à long terme dans la collecte des données, il n'est pas conseillé d'augmenter le nombre d'années pour inclure davantage de données historiques. Un nombre trop élevé de périodes de référence peut rendre les observations très éloignées moins pertinentes, ce qui pourrait dégrader la performance du modèle. La détermination de la taille optimale de la demi-fenêtre est également complexe en raison des effets saisonniers parfois difficiles à discerner. De plus, pour obtenir une symétrie approximative, les prédictions sont transformées avant le calcul des seuils afin de les ajuster à une loi normale. Cependant, cette approche peut ne pas être optimale dans toutes les situations.

Le modèle Farrington prend en compte la saisonnalité en utilisant un sous-ensemble des données disponibles pour ajuster le GLM sans avoir une forme paramétrique fixe. Selon (Noufaily *et al.*, 2013), l'algorithme est plus performant lorsqu'il utilise un plus grand nombre de données historiques. Les auteurs ont introduit une spline d'ordre zéro avec 11 nœuds, représentant un facteur à 10 niveaux, comme amélioration. Plusieurs autres améliorations au modèle Farrington ont été proposées par (Noufaily *et al.*, 2013) pour les personnes intéressées.

4.3 Le modèle GLR

Une approche plus récente développée par (Höhle et Paul, 2008) pour la détection des anomalies est le GLR (Generalized Likelihood Ratio). Contrairement à la méthode de Farrington, le GLR se caractérise par :

- La modélisation paramétrique de la saisonnalité ;
- L'utilisation de l'ensemble complet des valeurs historiques disponibles $\mathscr{H}_{t_0}$, plutôt que d'un sous-échantillon $\mathscr{H}^*_{t_0}$.

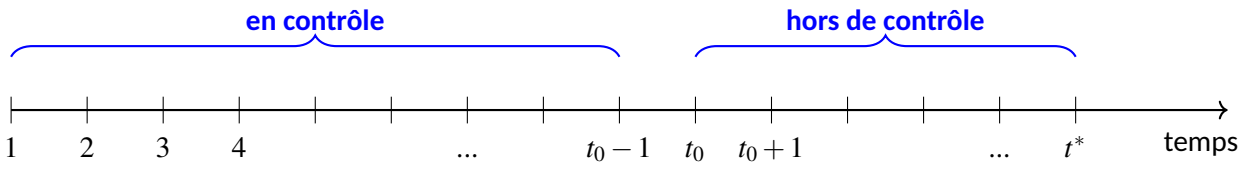
Le modèle GLR se distingue par sa capacité à modéliser les données de comptage et à traiter le problème de la détection d'anomalies dans un cadre orienté CSP (voir la sous-section 4.1 pour plus d'informations sur le CSP). En utilisant le rapport de vraisemblance, qui évalue la probabilité des données observées sous un modèle avec changement par rapport à un modèle sans changement, le GLR permet de détecter des anomalies de manière robuste, tout comme le modèle de Farrington. En maximisant ce rapport de vraisemblance, le GLR identifie les points de changement et les variations significatives dans les données. À la différence du modèle Farrington, le seuil pour les alertes est défini par l'utilisateur.

4.3.1 Données utilisées

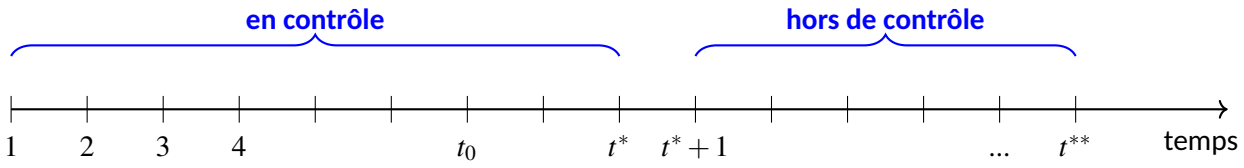
Comme pour l'approche de Farrington, il est important de clarifier la manière dont les données sont utilisées pour identifier les anomalies avec la méthode GLR. Cette approche repose sur deux états possibles : **en contrôle** et **hors de contrôle**.

La figure 4.4a présente une situation initiale où l'on cherche à détecter des anomalies à partir du moment t_0. Les observations des périodes précédant t_0 sont considérées comme étant dans l'état en contrôle. En revanche, les observations à partir de t_0 sont classées dans l'état hors de contrôle jusqu'à ce qu'une anomalie soit détectée, ici au temps t^*. Lorsqu'une anomalie est identifiée et qu'une alerte est déclenchée (au temps t^* dans cet exemple), les états en contrôle et hors de contrôle sont réévalués, comme le montre la figure 4.4b. La procédure de définition des états en contrôle et hors de contrôle est alors répétée pour chaque nouvelle alerte.

Ainsi, la première observation surveillée (à t_0) est intégrée dans le calcul de la statistique. Si aucune anomalie n'est détectée, les deux premières observations (t_0 et $t_0 + 1$) seront alors considérées comme l'échantillon



(a) Les valeurs utilisées dans la situation en contrôle et hors de contrôle



(b) Les valeurs utilisées dans la situation en contrôle et hors de contrôle suite à une alerte à t^*

Figure 4.4 – Les valeurs utilisées avant et après les alertes

hors de contrôle et ainsi de suite jusqu'à l'obtention de la première alerte.

4.3.2 Distribution du nombre de vols

Pour l'approche GLR, nous posons que les observations $x_1, x_2, \ldots$ suivent une distribution paramétrique avec une densité f_θ qui varie selon l'état :

$$x_t \mid \mathbf{z}_t \sim \begin{cases} f_{\theta_0}(\cdot \mid \mathbf{z}_t) & \text{pour } t = 1, \ldots, t_0 - 1 \text{ (en contrôle)} \\ f_{\theta_1}(\cdot \mid \mathbf{z}_t) & \text{pour } t = t_0, t_0 + 1, \ldots \text{ (hors de contrôle)} \end{cases}, \tag{4.6}$$

où t_0 marque la première observation sous surveillance et $\mathbf{z}_t$ représente les covariables connues au temps t. Pour simplifier, nous appellerons f_{θ_0} la densité de base et f_{θ_1} la densité augmentée.

Bien qu'une distribution de Poisson soit souvent utilisée pour modéliser le nombre de vols à chaque période, nous étendons l'approche en considérant une distribution binomiale négative avec un paramètre fixe α. Lorsque $\alpha \to 0$, cette distribution converge vers une distribution de Poisson.

Pour tenir compte de la saisonnalité, nous utiliserons un modèle log-linéaire pour la période en contrôle, avec la moyenne de la distribution définie comme suit :

$$\log \mu_{0,t} = \beta_0 + \beta_1 t + \sum_{s=1}^{S} (\beta_{2s} \cos(\omega s t) + \beta_{2s+1} \sin(\omega s t)) \tag{4.7}$$

où S est le nombre d'ondes harmoniques utilisées, $\omega = \frac{2\pi}{T}$ est la fréquence, et T est la période (par exemple, $T = 52$ pour des données hebdomadaires). Le modèle requiert donc $2S + 3$ paramètres : $2S + 2$ pour la moyenne et un paramètre pour la surdispersion α. D'autres approches, comme les splines T-périodiques, pourraient également être envisagées.

Les distributions f_{θ_0} et f_{θ_1} auront des moyennes respectives $\mu_{0,t}$ et $\mu_{1,t}$. Un état hors de contrôle est caractérisé par un changement multiplicatif dans les moyennes, selon la relation :

$$\mu_{1,t} = \mu_{0,t} \cdot \exp(\kappa), \tag{4.8}$$

où κ représente un changement additif sur l'échelle logarithmique. Si κ est significativement positif, cela indique une augmentation notable du nombre de cas.

Lorsqu'une alerte est détectée, par exemple au temps t_0^*, les densités en contrôle et hors de contrôle sont révisées comme suit :

$$x_t \mid \mathbf{z}_t \sim \begin{cases} f_{\theta_0}(\cdot \mid \mathbf{z}_t) \text{ pour } t = 1, \ldots, t_0^* & \text{(en contrôle)} \\ f_{\theta_1}(\cdot \mid \mathbf{z}_t) \text{ pour } t = t_0^* + 1, t_0^* + 2, \ldots & \text{(hors de contrôle)} \end{cases}, \tag{4.9}$$

avec les données historiques $\mathscr{H}_{t_0^*}$, et ce processus est répété jusqu'à la dernière observation sous surveillance.

4.3.3 Hypothèses du modèle

Nous allons maintenant examiner le rapport de vraisemblance généralisé (GLR) tel que défini par (Lai, 1995). Ce modèle généralise la méthode CUSUM en utilisant la règle d'arrêt suivante :

$$N_G = \inf\left\{ n \geq 1 : \max_{1\leq k\leq n} \sup_{\theta\in\Theta} \left[\sum_{t=k}^{n} \log\left\{ \frac{f_\theta\left(x_t \mid \mathbf{z}_t\right)}{f_{\theta_0}\left(x_t \mid \mathbf{z}_t\right)} \right\} \right] \geq c_\gamma \right\}. \tag{4.10}$$

Dans cette équation, on maximise la log-vraisemblance sur $\theta \in \Theta$ pour chaque point de changement k, où k varie entre 1 et n, couvrant toutes les valeurs sous observation. La valeur c_γ représente le seuil déterminé par l'utilisateur. Le paramètre N_G indique le moment où la statistique GLR dépasse pour la première fois ce seuil. Après chaque alerte, le système est réinitialisé et recommence à partir de $N_G + 1$, comme illustré dans la figure 4.4. Ainsi, contrairement à la méthode Farrington, les alertes pour les valeurs aberrantes surviennent lorsque la statistique GLR dépasse le seuil défini, plutôt que lorsque la valeur observée elle-même dépasse le seuil.

Le théorème de Wilks, décrit en détail dans (Wilks, 1938), stipule que pour deux modèles imbriqués, la statistique du rapport de vraisemblance suit asymptotiquement une loi du khi-deux sous l'hypothèse nulle. Si $\mathscr{L}_0$ est la vraisemblance sous le modèle restreint (hypothèse nulle H_0) et $\mathscr{L}_1$ est la vraisemblance sous le modèle général (hypothèse alternative H_1), la statistique est donnée par :

$$\lambda = -2\log\left(\frac{\mathscr{L}_0}{\mathscr{L}_1}\right).$$

Sous H_0, λ suit asymptotiquement une distribution χ^2 avec des degrés de liberté correspondant à la différence de paramètres entre les deux modèles, pour des échantillons de grande taille. Les conditions essentielles pour cette statistique incluent :

- Les modèles doivent être imbriqués ;
- La distribution doit être régulière (toute fonction continue et localement intégrable définit une distribution régulière) ;
- Les estimations doivent être obtenues par la méthode du maximum de vraisemblance (MLE).

Les hypothèses des moyennes des densités en utilisant des distributions binomiales négatives de type NB2 sont :

- Hypothèse nulle $H_0 : \mu_{0,t} = \mu_{1,t}$ pour $t = 1, 2, \ldots$;

— Hypothèse alternative H_1 : $\mu_{0,t} \neq \mu_{1,t}$.

La densité binomiale négative NB2 est définie par :

$$f_\theta\left(x_t \mid \mathbf{z}_t\right) = \frac{\Gamma(x+\frac{1}{\alpha})}{\Gamma(x+1)\Gamma(\frac{1}{\alpha})}\left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{x}.$$

Les fonctions de vraisemblance sous H_0 et H_1 sont respectivement :

$$\mathscr{L}_0 = \prod_{t=k}^{n} f_{\theta_0}\left(x_t \mid \mathbf{z}_t\right),$$
$$\mathscr{L}_1 = \prod_{t=k}^{n} f_{\theta_1}\left(x_t \mid \mathbf{z}_t\right).$$

Sous H_0, la statistique λ suit approximativement une loi de khi-deux avec dl degrés de liberté, où :

$$\lambda = -2\log\left(\frac{\mathscr{L}_0}{\mathscr{L}_1}\right) = 2\log\left(\frac{\mathscr{L}_1}{\mathscr{L}_0}\right) \sim \chi^2_{dl}, \tag{4.11}$$

et donc :

$$\log\left(\frac{\mathscr{L}_1}{\mathscr{L}_0}\right) \sim \frac{\chi^2_{dl}}{2}, \tag{4.12}$$

où dl est la différence du nombre de paramètres entre les deux modèles. Des théorèmes concernant cette relation, ainsi que des preuves, sont détaillés dans (Casella et Berger, 2024), contenant plusieurs exemples appliqués.

4.3.4 Développement des dérivées de la loi binomiale négative et de la loi de Poisson

Les dérivations sont fondées sur le cas où f représente une distribution binomiale négative, paramétrée par sa moyenne μ et son paramètre de dispersion α, de manière à ce que la variance soit $\mu + \alpha\mu^2$. Dans cette paramétrisation, lorsque $\alpha \to 0$, on obtient une distribution de Poisson de moyenne μ.

Proposition 4.1 *Lorsque $\mu_{0,t}$ et le paramètre de dispersion α, sont connus et en utilisant l'équation (4.8), le calcul du ratio de vraisemblance se définit comme :*

$$\log\left\{\frac{f_{\theta_1}(x_t \mid \mathbf{z}_t)}{f_{\theta_0}(x_t \mid \mathbf{z}_t)}\right\} = \kappa \cdot x_t + \left(x_t + \frac{1}{\alpha}\right)\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}\exp(\kappa)}\right), t = 1,2,\ldots \tag{4.13}$$

Preuve.

Avec :

$$f_{\theta_0}(x_t \mid \mathbf{z}_t) = P(X_t = x_t|\mu,\alpha) = \frac{\Gamma(x_t+\frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(x_t+1)}\left(\frac{\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}}\right)^{x_t}\left(\frac{1}{1+\alpha\mu_{0,t}}\right)^{1/\alpha},$$
$$f_{\theta_1}(x_t \mid \mathbf{z}_t) = P(X_t = x_t|\mu,\alpha) = \frac{\Gamma(x_t+\frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(x_t+1)}\left(\frac{\alpha\mu_{0,t}\exp(\kappa)}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)^{x_t}\left(\frac{1}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)^{1/\alpha},$$

nous avons ainsi :

$$\frac{f_{\theta_1}(x_t \mid \mathbf{z}_t)}{f_{\theta_0}(x_t \mid \mathbf{z}_t)} = \frac{\left(\frac{\alpha\mu_{0,t}\exp(\kappa)}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)^{x_t}\left(\frac{1}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)^{1/\alpha}}{\left(\frac{\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}}\right)^{x_t}\left(\frac{1}{1+\alpha\mu_{0,t}}\right)^{1/\alpha}}.$$

Nous obtenons alors :

$$\begin{aligned}
&\log\left(\frac{f_{\theta_1}(x_t \mid \mathbf{z}_t)}{f_{\theta_0}(x_t \mid \mathbf{z}_t)}\right)\\
&= \log\left(\left(\alpha\mu_{0,t}\frac{\exp(\kappa)}{(1+\alpha\mu_{0,t}\exp(\kappa))}\right)^{x_t}\left(\frac{1}{(1+\alpha\mu_{0,t}\exp(\kappa))}\right)^{\frac{1}{\alpha}}\left(\left(\alpha\frac{\mu_{0,t}}{(1+\alpha\mu_{0,t})}\right)^{x_t}\left(\frac{1}{1+\alpha\mu_{0,t}}\right)^{\frac{1}{\alpha}}\right)^{-1}\right)\\
&= \log\left(\left(\alpha\frac{\exp(\kappa)}{(\alpha\exp(\kappa)\mu_{0,t}+1)}\mu_{0,t}\right)^{x_t}\right) + \log\left(\left(\frac{1}{(\alpha\exp(\kappa)\mu_{0,t}+1)}\right)^{\frac{1}{\alpha}}\right) - \log\left(\left(\frac{1}{(\alpha\mu_{0,t}+1)}\right)^{\frac{1}{\alpha}}\left(\alpha\frac{\mu_{0,t}}{(\alpha\mu_{0,t}+1)}\right)^{x_t}\right)\\
&= x_t\log\left(\frac{\alpha\exp(\kappa)\mu_{0,t}}{\alpha\exp(\kappa)\mu_{0,t}+1}\right) - \frac{\log(\alpha\exp(\kappa)\mu_{0,t}+1)}{\alpha} - \left(\frac{1}{\alpha}\log\left(\frac{1}{\alpha\mu_{0,t}+1}\right) + x_t\log\left(\alpha\frac{\mu_{0,t}}{\alpha\mu_{0,t}+1}\right)\right)\\
&= x_t\log\left(\frac{\alpha\exp(\kappa)\mu_{0,t}}{\alpha\exp(\kappa)\mu_{0,t}+1}\right) - \frac{\log(\alpha\exp(\kappa)\mu_{0,t}+1)}{\alpha} - \left(x_t\log\left(\frac{\alpha\mu_{0,t}}{\alpha\mu_{0,t}+1}\right) - \frac{\log(\alpha\mu_{0,t}+1)}{\alpha}\right)
\end{aligned}$$

$$
\begin{aligned}
&= x_t \log\left(\frac{\alpha\exp(\kappa)\mu_{0,t}}{1+\alpha\exp(\kappa)\mu_{0,t}}\right) + \frac{-\log(1+\alpha\exp(\kappa)\mu_{0,t}) + \log(1+\alpha\mu_{0,t})}{\alpha} - x_t \log\left(\frac{\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}}\right) \\
&= x_t \log\left(\frac{\alpha\exp(\kappa)\mu_{0,t}}{1+\alpha\exp(\kappa)\mu_{0,t}}\right) + \frac{1}{\alpha}\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\exp(\kappa)\mu_{0,t}}\right) - x_t \log\left(\frac{\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}}\right) \\
&= \frac{1}{\alpha}\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\exp(\kappa)\mu_{0,t}}\right) + x_t \log\left(\frac{\alpha\exp(\kappa)\mu_{0,t}}{1+\alpha\exp(\kappa)\mu_{0,t}}\right) - x_t \log\left(\frac{\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}}\right) \\
&= \frac{1}{\alpha}\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\exp(\kappa)\mu_{0,t}}\right) + x_t\kappa + x_t \log\left(\frac{\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}\exp(\kappa)}\right) - x_t \log\left(\frac{\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}}\right) \\
&= \frac{1}{\alpha}\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\exp(\kappa)\mu_{0,t}}\right) + x_t\kappa + x_t \log\left(\frac{\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}\exp(\kappa)} \cdot \frac{1+\alpha\mu_{0,t}}{\alpha\mu_{0,t}}\right) \\
&= x_t\kappa + \frac{1}{\alpha}\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\exp(\kappa)\mu_{0,t}}\right) + x_t \log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}\exp(\kappa)}\right) \\
&= \kappa\cdot x_t + \left(x_t + \frac{1}{\alpha}\right)\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)
\end{aligned}
$$

□

Proposition 4.2 *Dans le cas où $\alpha \to 0$, on obtiendra la log-vraisemblance du ratio pour la loi de Poisson telle que :*

$$
\lim_{\alpha\to 0}\left[\kappa\cdot x_t + \left(x_t + \frac{1}{\alpha}\right)\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)\right] = \kappa\cdot x_t + (1-\exp(\kappa))\cdot\mu_{0,t}. \tag{4.14}
$$

Preuve.

$$
\begin{aligned}
&\lim_{\alpha\to 0}\left[\kappa\cdot x_t+\left(x_t+\frac{1}{\alpha}\right)\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)\right]\\
&=\lim_{\alpha\to 0}\left[\kappa\cdot x_t+x_t\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)+\frac{1}{\alpha}\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)\right]\\
&=\kappa\cdot x_t+\lim_{\alpha\to 0}\left[x_t\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)+\frac{1}{\alpha}\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)\right]\\
&=\kappa\cdot x_t+\lim_{\alpha\to 0}\left[x_t\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)\right]+\lim_{\alpha\to 0}\left[\frac{1}{\alpha}\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)\right]\\
&=\kappa\cdot x_t+0+\lim_{\alpha\to 0}\left[\frac{1}{\alpha}\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)\right]\\
&=\kappa\cdot x_t+\lim_{\alpha\to 0}\left[\frac{1}{\alpha}\log\left(1+\alpha\mu_{0,t}\right)-\frac{1}{\alpha}\log\left(1+\alpha\mu_{0,t}\exp(\kappa)\right)\right]\\
&=\kappa\cdot x_t+\lim_{\alpha\to 0}\left[\frac{1}{\alpha}\log\left(1+\alpha\mu_{0,t}\right)\right]-\lim_{\alpha\to 0}\left[\frac{1}{\alpha}\log\left(1+\alpha\mu_{0,t}\exp(\kappa)\right)\right]\\
&=\kappa\cdot x_t+\lim_{\alpha\to 0}\left[\frac{1}{\alpha}\log\left(1+\alpha\mu_{0,t}\right)\right]-\lim_{\alpha\to 0}\left[\frac{1}{\alpha}\log\left(1+\alpha\mu_{0,t}\exp(\kappa)\right)\right]\\
&=\kappa\cdot x_t+\lim_{\alpha\to 0}\left[\frac{\frac{\partial}{\partial\alpha}\log\left(1+\alpha\mu_{0,t}\right)}{\frac{\partial}{\partial\alpha}\alpha}\right]-\lim_{\alpha\to 0}\left[\frac{\frac{\partial}{\partial\alpha}\log\left(1+\alpha\mu_{0,t}\exp(\kappa)\right)}{\frac{\partial}{\partial\alpha}\alpha}\right]\\
&=\kappa\cdot x_t+\lim_{\alpha\to 0}\left[\frac{\partial}{\partial\alpha}\log\left(1+\alpha\mu_{0,t}\right)\right]-\lim_{\alpha\to 0}\left[\frac{\partial}{\partial\alpha}\log\left(1+\alpha\mu_{0,t}\exp(\kappa)\right)\right]\\
&=\kappa\cdot x_t+\lim_{\alpha\to 0}\left[\frac{1}{1+\mu_{0,t}\alpha}\cdot\mu_{0,t}\right]-\lim_{\alpha\to 0}\left[\frac{1}{1+\mu_{0,t}\exp(\kappa)\alpha}\cdot\mu_{0,t}\exp(\kappa)\right]\\
&=\kappa\cdot x_t+\mu_{0,t}-\mu_{0,t}\exp(\kappa)\\
&=\kappa\cdot x_t+\mu_{0,t}\left(1-\exp(\kappa)\right)
\end{aligned}
$$

□

Dans la preuve ci-haut de l'équation (4.14), nous avons utilisé la règle de l'hôpital pour les deux limites, car chacune d'elles sera de la forme $\frac{0}{0}$ de plus, $\lim_{x\to c}\frac{u(x)}{v(x)}=\lim_{x\to c}\frac{u'(x)}{v'(x)}$.

Pour calculer l'estimateur du maximum de vraisemblance (MLE) pour κ avec les observations $x_k,\ldots,x_n$ requises par le GLR de l'équation (4.10), la log-vraisemblance de la loi de probabilité binomiale négative augmentée utilisée sera :

$$l_{n,k}=\sum_{t=k}^{n}\log f_\kappa\left(x_t\mid\mathbf{z}_t\right). \tag{4.15}$$

Proposition 4.3 *La première dérivée de l'équation (4.15) sachant $\mu_{0,t}$ et le paramètre de dispersion α sera :*

$$\frac{\partial l_{n,k}}{\partial \kappa} = \sum_{t=k}^{n} \frac{x_t - \mu_{0,t}\exp(\kappa)}{\alpha\mu_{0,t}\exp(\kappa)+1}. \tag{4.16}$$

Preuve.

Avec

$$P(X_t = x_t) = \frac{\Gamma(x_t + \frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(x_t+1)} \left(\frac{\alpha\mu_{0,t}\exp(\kappa)}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)^{x_t} \left(\frac{1}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)^{1/\alpha}, \tag{4.17}$$

suivi de la vraisemblance :

$$L(\kappa) = \prod_{t=k}^{n} P(X_t = x_t) = \prod_{t=k}^{n} \frac{\Gamma(x_t + \frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(x_t+1)} \left(\frac{\alpha\mu_{0,t}\exp(\kappa)}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)^{x_t} \left(\frac{1}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)^{1/\alpha},$$

la log-vraisemblance :

$$\begin{aligned} l(\kappa) &= \sum_{t=k}^{n} \log\left[\frac{\Gamma(x_t + \frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(x_t+1)} \left(\frac{\alpha\mu_{0,t}\exp(\kappa)}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)^{x_t} \left(\frac{1}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)^{1/\alpha}\right] \\ &= \sum_{t=k}^{n} \log\left(\frac{\Gamma(x_t + \frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(x_t+1)}\right) + x_t \log\left(\frac{\alpha\mu_{0,t}\exp(\kappa)}{1+\alpha\mu_{0,t}\exp(\kappa)}\right) + \frac{1}{\alpha}\log\left(\frac{1}{1+\alpha\mu_{0,t}\exp(\kappa)}\right). \end{aligned}$$

La dérivée par rapport à κ :

$$\begin{aligned} &\frac{\partial l(\kappa)}{\partial \kappa} \\ &= \sum_{t=k}^{n} \frac{\partial}{\partial\kappa}\left(x_t \log\left(\frac{\alpha\mu_{0,t}\exp(\kappa)}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)\right) + \frac{\partial}{\partial\kappa}\left(\frac{1}{\alpha}\log\left(\frac{1}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)\right) \\ &= \sum_{t=k}^{n} x_t \cdot \frac{\partial}{\partial\kappa}\left[\log\left(\frac{\alpha\mu_{0,t}\exp(\kappa)}{\alpha\mu_{0,t}\exp(\kappa)+1}\right)\right] - \frac{1}{\alpha}\cdot\frac{\partial}{\partial\kappa}\left[\log\left(\alpha\mu_{0,t}\exp(\kappa)+1\right)\right] \\ &= \sum_{t=k}^{n} x_t \cdot \frac{\alpha\mu_{0,t}\exp(\kappa)+1}{\alpha\mu_{0,t}\exp(\kappa)} \cdot \frac{\partial}{\partial\kappa}\left[\frac{\alpha\mu_{0,t}\exp(\kappa)}{\alpha\mu_{0,t}\exp(\kappa)+1}\right] - \frac{\frac{1}{\alpha\mu_{0,t}\exp(\kappa)+1}\cdot\frac{\partial}{\partial\kappa}\left[\alpha\mu_{0,t}\exp(\kappa)+1\right]}{\alpha} \\ &= \sum_{t=k}^{n} \frac{x_t\alpha\mu_{0,t}\cdot\frac{\partial}{\partial\kappa}\left[\frac{\exp(\kappa)}{\alpha\mu_{0,t}\exp(\kappa)+1}\right]\cdot\exp(-\kappa)\cdot(\alpha\mu_{0,t}\exp(\kappa)+1)}{\alpha\mu_{0,t}} - \frac{\alpha\mu_{0,t}\cdot\frac{\partial}{\partial\kappa}\left[\exp(\kappa)\right]+\frac{\partial}{\partial\kappa}[1]}{\alpha\cdot(\alpha\mu_{0,t}\exp(\kappa)+1)} \\ &= \sum_{t=k}^{n} \frac{x_t\cdot\left(\exp(\kappa)\cdot(\alpha\mu_{0,t}\exp(\kappa)+1) - \left(\alpha\mu_{0,t}\cdot\frac{\partial}{\partial\kappa}\left[\exp(\kappa)\right]+\frac{\partial}{\partial\kappa}[1]\right)\exp(\kappa)\right)\exp(-\kappa)}{\alpha\mu_{0,t}\exp(\kappa)+1} - \frac{\mu_{0,t}\exp(\kappa)}{\alpha\mu_{0,t}\exp(\kappa)+1} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{t=k}^{n} \frac{x_t \cdot (\exp(\kappa) \cdot (\alpha\mu_{0,t}\exp(\kappa)+1) - (\alpha\mu_{0,t}\exp(\kappa)+0)\exp(\kappa))\exp(-\kappa)}{\alpha\mu_{0,t}\exp(\kappa)+1} - \frac{\mu_{0,t}\exp(\kappa)}{\alpha\mu_{0,t}\exp(\kappa)+1} \\
&= \sum_{t=k}^{n} \frac{x_t \exp(-\kappa) \cdot (\exp(\kappa) \cdot (\alpha\mu_{0,t}\exp(\kappa)+1) - \alpha\mu_{0,t} exp(2\kappa))}{\alpha\mu_{0,t}\exp(\kappa)+1} - \frac{\mu_{0,t}\exp(\kappa)}{\alpha\mu_{0,t}\exp(\kappa)+1} \\
&= \sum_{t=k}^{n} \frac{x_t \exp(-\kappa)\left[\alpha\mu_{0,t} exp(2\kappa) + \exp(\kappa) - \alpha\mu_{0,t} exp(2\kappa)\right] - \mu_{0,t}\exp(\kappa)}{\alpha\mu_{0,t}\exp(\kappa)+1} \\
&= \sum_{t=k}^{n} \frac{x_t - \mu_{0,t}\exp(\kappa)}{\alpha\mu_{0,t}\exp(\kappa)+1}
\end{aligned}
$$

□

Proposition 4.4 *Le développement de la seconde dérivée sachant $\mu_{0,t}$ et le paramètre de dispersion α sera :*

$$
\frac{\partial^2 l_{n,k}}{\partial \kappa^2} = -\sum_{t=k}^{n} \frac{\mu_{0,t}\exp(\kappa)\,(1+\alpha x_t)}{(\alpha\mu_{0,t}\exp(\kappa)+1)^2}. \tag{4.18}
$$

Preuve.

$$
\begin{aligned}
&\frac{\partial}{\partial\kappa}\sum_{t=k}^{n} \frac{x_t - \mu_{0,t}\exp(\kappa)}{\alpha\mu_{0,t}\exp(\kappa)+1} \\
&= \sum_{t=k}^{n} -\frac{\frac{\partial}{\partial\kappa}\left[\mu_{0,t}\exp(\kappa) - x_t\right] \cdot (\alpha\mu_{0,t}\exp(\kappa)+1) - (\mu_{0,t}\exp(\kappa)-x_t)\cdot\frac{\partial}{\partial\kappa}\left[\alpha\mu_{0,t}\exp(\kappa)+1\right]}{(\alpha\mu_{0,t}\exp(\kappa)+1)^2} \\
&= \sum_{t=k}^{n} -\frac{\left(\mu_{0,t}\cdot\frac{\partial}{\partial\kappa}\left[\exp(\kappa)\right] + \frac{\partial}{\partial\kappa}[-x_t]\right)(\alpha\mu_{0,t}\exp(\kappa)+1) - (\mu_{0,t}\exp(\kappa)-x_t)\left(\alpha\mu_{0,t}\cdot\frac{\partial}{\partial\kappa}\left[\exp(\kappa)\right] + \frac{\partial}{\partial\kappa}[1]\right)}{(\alpha\mu_{0,t}\exp(\kappa)+1)^2} \\
&= \sum_{t=k}^{n} -\frac{(\mu_{0,t}\exp(\kappa)+0)(\alpha\mu_{0,t}\exp(\kappa)+1) - (\mu_{0,t}\exp(\kappa)-x_t)(\alpha\mu_{0,t}\exp(\kappa)+0)}{(\alpha\mu_{0,t}\exp(\kappa)+1)^2} \\
&= \sum_{t=k}^{n} -\frac{\mu_{0,t}\exp(\kappa)\cdot(\alpha\mu_{0,t}\exp(\kappa)+1) - \alpha\mu_{0,t}\exp(\kappa)\cdot(\mu_{0,t}\exp(\kappa)-x_t)}{(\alpha\mu_{0,t}\exp(\kappa)+1)^2} \\
&= \sum_{t=k}^{n} \frac{\alpha\mu_{0,t}^2\exp(2\kappa) + \mu_{0,t}\exp(\kappa) - \alpha\mu_{0,t}^2\exp(2\kappa) + \alpha\mu_{0,t}\exp(\kappa)x_t}{(\alpha\mu_{0,t}\exp(\kappa)+1)^2} \\
&= \sum_{t=k}^{n} \frac{\alpha\mu_{0,t}^2\exp(2\kappa) + \mu_{0,t}\exp(\kappa) - \alpha\mu_{0,t}^2\exp(2\kappa) + \alpha\mu_{0,t}\exp(\kappa)x_t}{(\alpha\mu_{0,t}\exp(\kappa)+1)^2} \\
&= -\sum_{t=k}^{n} \frac{\mu_{0,t}\exp(\kappa) + \alpha\mu_{0,t}\exp(\kappa)x_t}{(\alpha\mu_{0,t}\exp(\kappa)+1)^2} \\
&= -\sum_{t=k}^{n} \frac{\mu_{0,t}\exp(\kappa)\,(1+\alpha x_t)}{(\alpha\mu_{0,t}\exp(\kappa)+1)^2},
\end{aligned}
$$

□

En général, pour résoudre les équations $\partial l_{n,k}/\partial \kappa = 0$ des équations (4.16 et 4.18), il est nécessaire d'utiliser une méthode numérique, comme l'itération de Newton-Raphson (expliqué en détail à la sous-section 1.2.3) ou des logiciels spécialisés pour les *GLM*. On peut utiliser $\hat{\kappa}_{n,k+1}$ comme point de départ pour $\hat{\kappa}_{n,k}$. Avec ce point de départ, la convergence se produit généralement après quelques itérations. Pour le cas de la loi de Poisson, une solution analytique est possible :

$$\hat{\kappa}_{n,k} = \log\left(\frac{\sum_{t=k}^{n} x_t}{\sum_{t=k}^{n} \mu_{0,t}}\right).$$

Pour garantir que $\kappa \geq 0$, on utilise $\hat{\kappa}^{+}_{n,k} = \max(0, \hat{\kappa}_{n,k})$, et on calcule :

$$r_{n,k} = \sup_{\theta\in\Theta} \sum_{t=k}^{n} \log \frac{f_\theta(x_t \mid \mathbf{z}_t)}{f_{\theta_0}(x_t \mid \mathbf{z}_t)} = \hat{\kappa}^{+}_{n,k} \sum_{t=k}^{n} x_t + \left(1 - \exp\left(\hat{\kappa}^{+}_{n,k}\right)\right) \sum_{t=k}^{n} \mu_{0,t}.$$

Quelques remarques sur l'évolution du ratio de la log-vraisemblance généralisé de l'équation (4.13) en fonction des valeurs de x et κ :

- Lorsque $x \geq 0$ et $\kappa = 0$, le ratio est nul.
- Lorsque $x = 0$ et $\kappa > 0$, le ratio est donné par $\frac{1}{\alpha}\log\left(\frac{1+\alpha\mu_{0,t}}{1+\alpha\mu_{0,t}\exp(\kappa)}\right)$.
- Lorsque $x \geq 0$ et $\kappa > 0$, si κ augmente, le ratio diminuera, tandis que si x augmente, le ratio augmentera.

4.3.5 Application du modèle GLR avec exposition

Dans un contexte actuariel, et plus précisément dans la modélisation du nombre de vols de voitures, l'exposition est un facteur clé à prendre en compte. Par contre, la fonction *glrnb* incluse dans le paquet *surveillance* (voir (Salmon *et al.*, 2016) pour plus de détails), ne permet pas de rajouter de covariables ou de prendre en compte l'exposition. Dès que nous voulons rajouter l'exposition, que ce soit avec ou sans variables explicatives, le modèle ne peut pas être utilisé et doit donc être traité sans la fonction *glrnb*.

En supposant que le modèle en situation contrôlée et non contrôlée suive toujours une loi binomiale négative et en tenant compte de l'exposition, nous obtenons :

$$\mu_{0,t} = d_t \exp\left(\beta_0 + \beta_1 t + \sum_{s=1}^{S} \left(\beta_{2s}\cos(\omega st) + \beta_{2s+1}\sin(\omega st)\right)\right) \tag{4.19}$$

où S représente le nombre d'ondes harmoniques utilisées, $\omega = \frac{2\pi}{T}$ avec T étant la division par période utilisée (par exemple, pour des données hebdomadaires, $T = 52$), et d_t désigne l'exposition.

L'algorithme s'applique de la manière suivante : tout d'abord, les paramètres β et α de l'équation (4.19) sont obtenus à partir de la situation initiale en contrôle. Il est important de noter que le paramètre α n'est estimé qu'une seule fois. Ensuite, les moyennes prévues sous surveillance $\mu_{0,t}$ sont calculées pour $t = t_0, t_0 + 1, \ldots$. Le paramètre κ est ensuite évalué en utilisant la première et la deuxième dérivée dans l'itération de Newton-Raphson (voir la sous-section 1.2.3 pour plus d'informations), selon la formule suivante :

$$\kappa_{new} = \kappa_{old} + nbScore(\kappa_{old}, x, \mu_{0,t}, \alpha, k, n)/nbFisher(\kappa_{old}, x, \mu_{0,t}, \alpha, k, n),$$

où nbScore et nbFisher représentent respectivement la première et la deuxième dérivée provenant des équations (4.16) et (4.18). La valeur initiale pour κ_{old} est fixée à 0,4. Le $max(\kappa, 0)$ est utilisé, car seules les augmentations additives de la moyenne sur l'échelle logarithmique nous intéressent. Bien qu'il soit également possible de détecter des diminutions de la moyenne, cela ne sera pas abordé dans ce mémoire.

Une fois le paramètre κ déterminé, les valeurs de $\mu_{1,t}$ de l'équation (4.8) sont calculées pour $t = t_0, t_0 + 1, \ldots$. Après avoir obtenu les moyennes, le rapport de vraisemblance généralisé de l'équation (4.13) est calculé. Si ce rapport de vraisemblance généralisé dépasse le seuil $GLR(n) \geq c_\gamma$, une alerte est déclenchée. Lorsqu'une alerte est détectée, le processus s'arrête et redémarre au temps suivant l'alerte, soit à partir de $t_0^* + 1$ (voir la sous-section 4.3.2 et la figure 4.4 pour un exemple visuel), et ainsi de suite jusqu'à la dernière observation de la période sous surveillance.

Les paramètres utilisés incluent un seuil c_γ fixé à 3.3175 et une saisonnalité annuelle avec $S = 1$. Il est important de noter qu'aucun recalcul des paramètres β n'a été effectué après les alertes. Le choix du seuil c_γ a été déterminé à partir de la théorie abordée à la sous-section 4.3.3. La ligne verte pointillée à la figure 4.5 représente le seuil qui est utilisé par le modèle GLR. La différence de paramètre entre la densité de base et de la densité augmentée est égale à un. Le paramètre α sélectionné est fixé à 1%. Pour $\alpha = 1\%$ et un degré

de liberté ($dl = 1$), la valeur critique $\chi^2_{\alpha=1\%,dl=1}$ est d'environ 6.635. Ainsi, lorsque $\log\left(\frac{\mathscr{L}_1}{\mathscr{L}_0}\right) \geq 3.3175$, nous rejetterons H_0 au niveau de signification de 1 %. Par conséquent, lorsque $GLR(n) \geq c_\gamma = 3.3175$ une alerte survient.

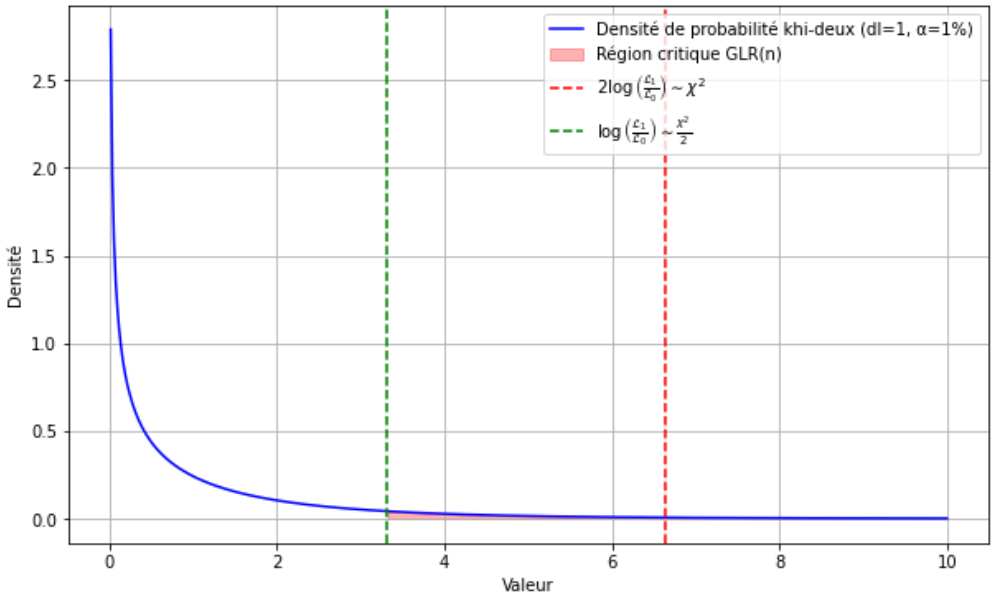


Figure 4.5 – La distribution khi-deux

La figure 4.6 contient en arrière-plan (barres grises) le nombre quotidien de voitures volées entre le 3 janvier 2018 et le 30 septembre 2023. La ligne pointillée bleue représente la statistique GLR calculée pour toutes les journées sous surveillance. La ligne verte correspond au seuil sélectionné, et les triangles indiquent les alertes. Elle illustre l'ajustement du modèle GLR avec l'exposition, sans ajout de variables supplémentaires. Les alertes se déclenchent lorsque la valeur du GLR(n) dépasse le seuil. On observe une augmentation significative du nombre d'alertes entre le 26 octobre 2022 et le 22 mars 2023, ce qui est cohérent avec les observations de la fréquence de la figure 2.1a. En effet, une forte hausse de la fréquence est visible, atteignant un pic en 2023, ce qui concorde avec les valeurs GLR(n).

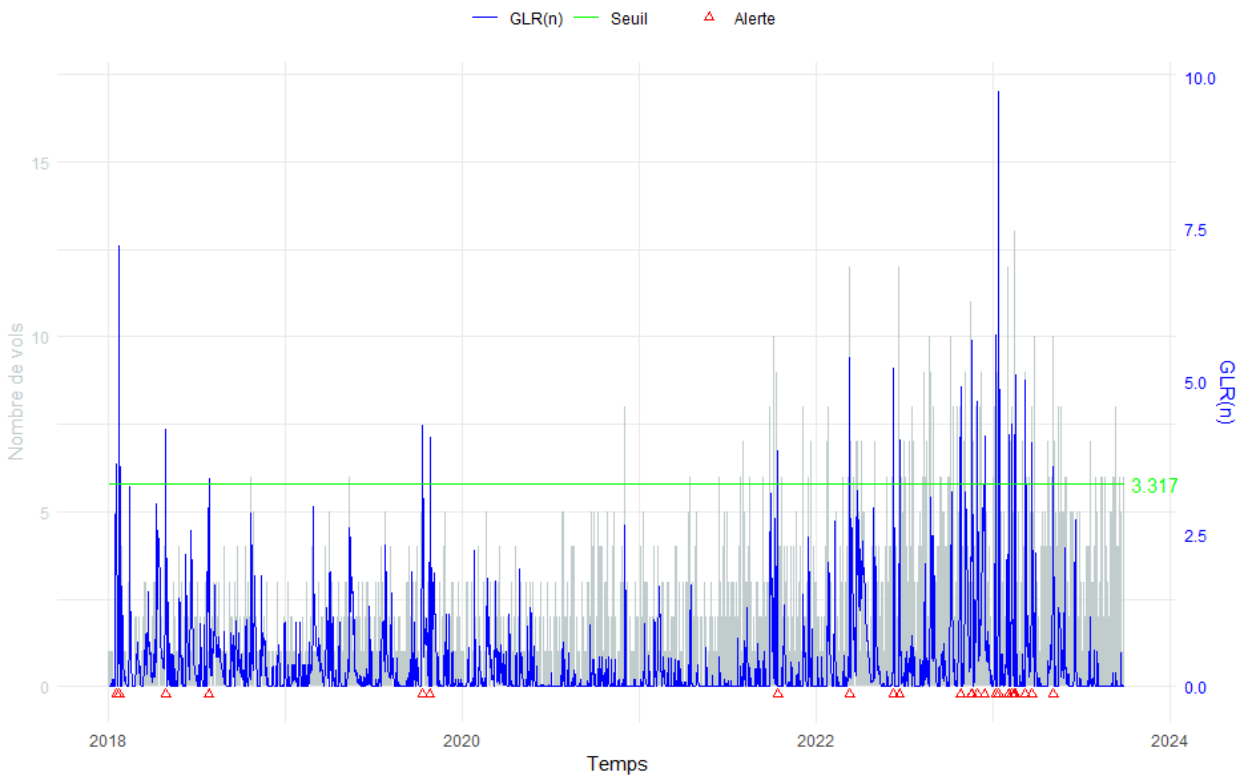


Figure 4.6 – Application du modèle GLR en prenant en compte l'exposition du 3 janvier 2018 au 30 septembre 2023

Nous pouvons tirer quelques observations de la figure 4.6 :

- Les premières alertes de 2018 apparaissent les 19 janvier, 24 janvier, 30 avril et 28 juillet, avec respectivement 4, 7, 4 et 5 vols de voitures signalés, et des statistiques de 3.65, 7.21, 4.22 et 3.4. Il semble que les baisses observées entre ces alertes soient très probablement liées à une diminution de la saisonnalité ;
- Deux autres alertes sont ensuite observées les 12 et 27 octobre 2019, avec 7 et 8 vols de voitures enregistrés, et des statistiques de 4.29 et 4.09 respectivement. Aucune autre alerte n'est détectée jusqu'au 13 octobre 2021, ce qui, là encore, est probablement dû à une baisse de la saisonnalité ;
- À partir du 26 octobre 2022, les alertes sont plus fréquentes. Un grand nombre de messages d'alerte aurait été envoyé à la compagnie d'assurance pour signaler des valeurs potentiellement aberrantes. On observe une concentration importante d'alertes entre la fin de 2022 et le 22 mars 2023. Ensuite, une seule alerte est signalée le 5 mai 2023, puis plus aucune jusqu'à la fin des observations. Une

analyse approfondie de cette concentration d'alertes est nécessaire.

Globalement, entre le 3 janvier 2018 et le 12 octobre 2021, la compagnie d'assurance a reçu six alertes sur un total de 1 379 jours, soit environ 0.44% des jours durant cette période. Chaque alerte doit être rapidement examinée par un expert afin de mettre en place des solutions pour gérer efficacement les anomalies détectées, en réduisant les pertes potentielles qui pourraient en découler.

4.4 Comparaison entre le modèle GLR et le modèle Farrington

Dans le domaine de la surveillance des séries temporelles, les modèles de détection d'anomalies jouent un rôle crucial dans la compréhension des tendances et des variations des données. Les deux modèles largement discutés jusqu'ici sont le modèle de Farrington et le modèle GLR. Bien que ces modèles soient conçus pour identifier des valeurs aberrantes, ils présentent des approches et des caractéristiques distinctes qui influencent leur performance et leur applicabilité dans différentes situations. En examinant les différences et les similitudes entre les modèles, ce mémoire vise à fournir une compréhension approfondie de leur fonctionnement et à guider les praticiens dans le choix des outils les plus appropriés pour leurs besoins en matière de détection d'anomalies dans les séries temporelles.

Le modèle Farrington a été utilisé avec des données par semaine et aucune variable explicative n'a été ajoutée au modèle. Le modèle GLR a utilisé un nombre de vols journaliers. En comparant la figure 4.2 du modèle Farrington et la figure 4.6 du modèle GLR, on observe des alertes survenant à peu près dans les mêmes périodes. En comparant les alertes des deux modèles, on remarque que le modèle GLR tend à déclencher des alertes plus fréquemment et de manière plus réactive aux variations soudaines dans les données. Le modèle Farrington en revanche, détecte des augmentations plus progressives et saisonnières. La statistique GLR(n) s'ajuste de près aux variations des données réelles et montre une capacité à suivre les fluctuations rapides et irrégulières, tandis que le seuil du modèle Farrington montre une tendance plus lisse. Cette différence illustre que le GLR peut mieux capturer les variations à court terme et les changements rapides, tandis que le modèle Farrington est plus robuste pour des tendances à long terme et des motifs saisonniers plus réguliers.

Les modèles GLR et Farrington présentent des différences significatives. Le modèle GLR offre une plus grande flexibilité et permet d'intégrer plusieurs covariables pour ajuster les prévisions. Toutefois, le GLR est plus complexe à paramétrer et à interpréter et nécessite des calculs intensifs lorsqu'il y a beaucoup de données. Comparativement au modèle GLR, le modèle Farrington utilise une méthode non paramétrique pour estimer le nombre attendu à partir d'un sous-ensemble des données historiques pour capter les variations saisonnières. La méthode GLR quant à elle, utilisera une forme paramétrique dans la somme à droite de l'équation (4.7) appelée onde harmonique. La méthode Farrington est caractérisée par sa simplicité et sa robustesse qui la rendent particulièrement efficace pour les séries temporelles avec des variations saisonnières marquées. Cependant, le modèle Farrington est moins flexible que le modèle GLR pour intégrer diverses distributions ou des covariables, et peut-être moins précis pour des séries temporelles plus compli-

quées. En comparaison, le GLR offre une plus grande flexibilité, mais au prix d'une plus grande complexité, tandis que le modèle Farrington est plus simple à utiliser, mais s'adapte moins bien à des données qui varient beaucoup.

Les ondes harmoniques des GLR utilisent des fonctions sinusoïdales qui modélisent les variations périodiques des données. Cette approche offre une précision élevée pour des motifs saisonniers réguliers et périodiques. Toutefois, elle peut devenir complexe avec l'ajout de multiples ondes harmoniques pour capturer des variations plus subtiles. Les ondes harmoniques sont idéales pour des motifs saisonniers réguliers, fournissant des modèles précis et interprétables, mais moins flexibles pour des variations non régulières. En comparaison, les méthodes non paramétriques s'adaptent mieux aux motifs saisonniers complexes et dynamiques. Les méthodes non paramétriques ne supposent pas de forme spécifique pour la saisonnalité et s'adaptent directement aux données telles que les moyennes mobiles, les splines et les techniques basées sur les fenêtres glissantes. Cette dernière méthode est utilisée par le modèle Farrington (voir la sous-section 4.2.1 pour un rappel).

CHAPITRE 5
EXTENSIONS DE L'APPROCHE GLR

5.1 Introduction

Ce chapitre se concentre sur l'application concrète du modèle GLR, avec l'intégration de certaines covariables pour en améliorer la précision. Cette phase cruciale de notre étude vise à évaluer l'efficacité du modèle GLR (*Generalized Likelihood Ratio*, du chapitre précédent) dans des contextes réels et complexes. À travers une série de tests et de comparaisons, ce chapitre démontre comment cette extension peut aider les compagnies d'assurance à gérer les risques de manière proactive. L'ajout de variables explicatives à l'approche GLR permet de transformer le modèle théorique de base en un outil puissant pour des analyses spécifiques.

Ce chapitre explore en profondeur l'intégration de variables explicatives dans le modèle de rapport de vraisemblance généralisé (GLR), initialement conçu pour surveiller les maladies infectieuses, mais ici appliqué aux vols d'automobiles. Le modèle GLR dans sa forme de base, décrit dans la sous-section 4.3, s'appuie sur des données historiques pour détecter des anomalies, sans prendre en compte des facteurs externes qui pourraient avoir une influence. Toutefois, son application dans le contexte des vols de véhicules nécessite une plus grande complexité pour mieux saisir la dynamique réelle des données. Cette complexité provient principalement du fait que de nombreux facteurs (comme les caractéristiques des assurés) peuvent expliquer les valeurs aberrantes.

Nous examinerons comment l'ajout de variables explicatives telles que les caractéristiques des véhicules et les particularités des assurés enrichit le modèle GLR, permettant ainsi une détection plus fine et précise des anomalies. L'intégration de ces variables transforme le modèle GLR en un outil capable de prendre en compte les diverses influences qui peuvent affecter les statistiques de vols. En analysant l'impact de variables comme la puissance des véhicules, la saisonnalité et les comportements des assurés, ce chapitre démontre comment le GLR ajusté peut fournir des prévisions plus robustes et adaptées aux données spécifiques. Cette démarche ne se limite pas à améliorer la précision du modèle, elle ouvre également la voie à de nouvelles perspectives pour élaborer des stratégies de prévention plus efficaces pour les compagnies d'assurance. En incorporant ces variables explicatives, le modèle GLR devient un instrument clé pour anticiper les tendances et réagir proactivement aux fluctuations des statistiques de vols de véhicules, offrant ainsi

un avantage significatif dans la gestion des risques. De plus, le modèle GLR de base nous alerte lorsqu'un nombre élevé de vols est détecté, tandis que l'extension avec des covariables pourrait permettre d'identifier les caractéristiques responsables de cette augmentation.

5.2 Ajout de covariables

Notre stratégie sera de procéder par étapes pour inclure les covariables dans le modèle GLR-étendu, en les introduisant une à une. En lien avec l'approche LASSO du chapitre 3, qui peut être interprétée comme une technique de sélection de variables basée sur un *budget*, cette méthode permet de déterminer le modèle optimal en fonction du nombre de covariables autorisées. Comme indiqué dans le chapitre 3, à la figure 3.1, nous ajouterons les covariables suivantes dans l'approche GLR, dans cet ordre (avec le nom de la variable dans la base de données entre parenthèses) :

1. Puissance du véhicule (Chevaux) ;
2. L'empattement du véhicule : la distance entre deux essieux (Empat) ;
3. Camion ou tous les autres véhicules qui ne sont pas des camions (Type) ;
4. État matrimonial légal (Marital) ;
5. L'âge du véhicule (Vehage) ;
6. Sexe de l'assuré (Genre) ;
7. La province de l'assuré (Province) ;
8. Catégorie du véhicule (Cat) ;
9. Le poids du véhicule (Poids) ;
10. Le prix d'achat du véhicule (Prix).

Pour plus de détails sur ces variables, il est possible de consulter la table 2.2.

5.2.1 GLR indépendants

Une première approche intuitive pour inclure les covariables dans l'approche GLR consiste à appliquer un modèle GLR indépendant pour chacune des classes de risque de la base de données, où une classe de risque correspond à la combinaison des modalités disponibles. Tous les GLR indépendants seront évalués avec les mêmes paramètres que les paramètres utilisés pour le GLR de base de la sous-section 4.3.5, à savoir un seuil

c_γ de 3.3175 et une saisonnalité annuelle avec $S = 1$. Un modèle linéaire généralisé avec une loi binomiale négative sera utilisé. Des données journalières seront employées.

Exemple 5 (Classes de risque) *Si nous cherchons à considérer une covariable qui ne comporte que deux modalités, nous n'aurons que deux classes de risque et l'approche proposée reviendrait à estimer deux modèles GLR distincts. Par contre, si nous voulons intégrer 3 covariables ayant chacune 2 modalités, nous aurons ainsi jusqu'à* $2^3 = 8$ *classes de risque, ce qui impliquerait d'ajuster 8 approches GLR distinctes.*

Cette approche basée sur la production de modèles GLR pour chacune des classes de risque a plusieurs avantages :

- Analyse granulaire : Cette méthode permet de capturer les différences spécifiques à chaque groupe, offrant des prédictions plus fines et adaptées ;
- Gestion de l'hétéroscédasticité : En travaillant avec des groupes homogènes, le modèle peut mieux gérer la variabilité non constante des erreurs, améliorant ainsi la fiabilité des estimations ;
- Personnalisation des prédictions : Les modèles par modalité permettent de produire des prédictions adaptées aux spécificités de chaque segment, augmentant leur pertinence ;
- Réduction de la multicolinéarité : Travailler avec des sous-ensembles de données peut aider à mieux gérer les problèmes de multicolinéarité.

Le problème de l'approche est que les GLR sont estimés indépendamment, alors qu'il semble clair que certaines tendances sur le nombre de vols observés pour l'une des classes de risque pourraient informer le nombre de vols des autres modalités. Entre autres, le contexte économique est le même pour chacune des classes de risque, on pourrait s'attendre à ce que la saisonnalité soit aussi identique, etc. D'autres désavantages de cette méthode peuvent aussi être mentionnés :

- Complexité de la segmentation : Diviser les données en groupes homogènes peut être complexe et nécessite une compréhension approfondie des différences entre les groupes ;
- Augmentation du nombre de modèles : Chaque groupe nécessitant un modèle distinct, cela augmente le nombre total de modèles à construire, valider et maintenir ;
- Risque de sous-ajustement : En segmentant les données, des informations cruciales sur les interactions entre les groupes peuvent être perdues, ce qui peut entraîner un sous-ajustement ;

— Problèmes de taille d'échantillon : Certains segments peuvent disposer de moins de données, ce qui peut affecter la robustesse et la fiabilité des estimations.

5.2.2 GLR-étendu

Une approche un peu plus efficace serait ainsi de ne faire qu'un seul modèle GLR qui intègre simultanément toutes les covariables que nous cherchons à mettre dans le modèle. L'idée est ainsi de proposer une généralisation de la moyenne du modèle exprimée par l'équation (4.7) afin d'avoir la forme suivante :

$$\mu_{0,t}^{j} = d_t \exp\left(\beta_0 + \beta_1 t + \sum_{s=1}^{S} \left[\beta_{2s}\cos(\omega st) + \beta_{2s+1}\sin(\omega st)\right] + \gamma \mathbf{X}_j^T\right), \tag{5.1}$$

où le vecteur γ représente les paramètres associés aux covariables $\mathbf{X}$ que nous souhaitons ajouter au modèle GLR de base. Les mêmes distributions que celles de la sous-section 4.3.4 peuvent être utilisées. Une fois qu'une covariable sera ajoutée à la moyenne du modèle GLR, nous l'appellerons GLR-étendu. Tous les GLR-étendu seront évalués avec les mêmes paramètres que les paramètres utilisés pour le GLR de base de la sous-section 4.3.5, à savoir un seuil c_γ de 3.3175 et une saisonnalité annuelle avec $S = 1$. Le nombre total de classes de risque sera égal au produit du nombre de modalités de chaque covariable individuelle, moins les combinaisons qui ne sont pas présentes dans les données. Un modèle linéaire généralisé avec une loi binomiale négative sera utilisé. Des données journalières seront employées.

Après avoir ajouté les variables, y compris l'exposition, nous obtenons la log-vraisemblance de la densité augmentée :

$$l_{n,k,j} = \sum_{t=k}^{n}\sum_{j=1}^{J} \log f_\kappa\left(x_{t,j} \mid z_{t,j}\right), \tag{5.2}$$

où J représente le nombre de classes de risque disponibles dans les données. La première dérivée est donnée par :

$$\frac{\partial l_{n,k,j}}{\partial \kappa} = \sum_{t=k}^{n}\sum_{j=1}^{J} \frac{x_{t,j} - \exp(\kappa)\mu_{0,t,j}}{1 + \alpha\exp(\kappa)\mu_{0,t,j}}, \tag{5.3}$$

et la seconde dérivée est :

$$\frac{\partial^2 l_{n,k,j}}{\partial \kappa^2} = -\sum_{t=k}^{n} \sum_{j=1}^{J} \frac{\exp(\kappa)\mu_{0,t,j}\,(\alpha x_{t,j}+1)}{(1+\alpha\exp(\kappa)\mu_{0,t,j})^2}. \tag{5.4}$$

Le modèle GLR-étendu intègre un plus grand nombre de variables explicatives dans l'analyse. Les principaux avantages du modèle GLR-étendu sont :

- Précision accrue : L'ajout de variables pertinentes permet de capturer plus de variabilité dans les données, conduisant à des prédictions plus précises ;
- Réduction des biais : En tenant compte de plusieurs facteurs, le modèle devient plus représentatif des phénomènes étudiés, réduisant ainsi les biais potentiels ;
- Stabilité du modèle : L'inclusion de variables explicatives pertinentes améliore la stabilité du modèle ;
- Simplification de la modélisation : Construire un seul modèle pour l'ensemble des données simplifie le processus de modélisation et d'interprétation des résultats.

Cependant, cette méthode présente aussi des désavantages :

- Complexité accrue : Plus de variables entraînent une complexité plus élevée du modèle, rendant l'interprétation des résultats plus difficile ;
- Risque de sur-ajustement : L'ajout excessif de variables peut conduire à un modèle trop adapté aux données d'entraînement, ce qui nuit à sa généralisation sur de nouvelles données ;
- Collecte et préparation des données : La nécessité de plus de variables implique souvent des coûts et une complexité supplémentaires dans la collecte et la préparation des données.

5.3 Application aux données de vols

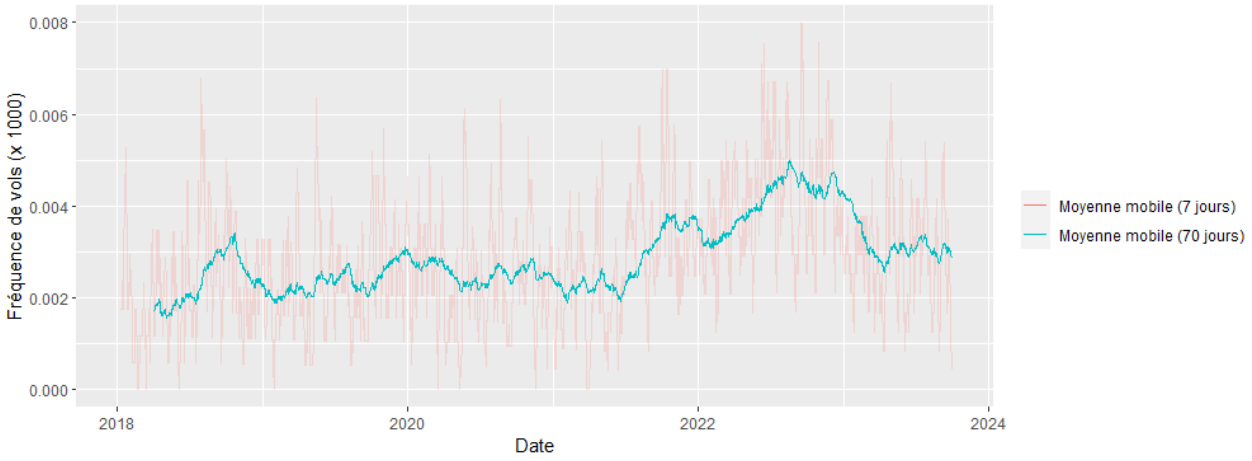
En utilisant la base de données d'assurance des vols d'automobiles, nous intégrerons une à une les covariables dans le modèle, selon l'ordre indiqué à la section 5.2.

5.3.1 Ajout du régresseur Chevaux

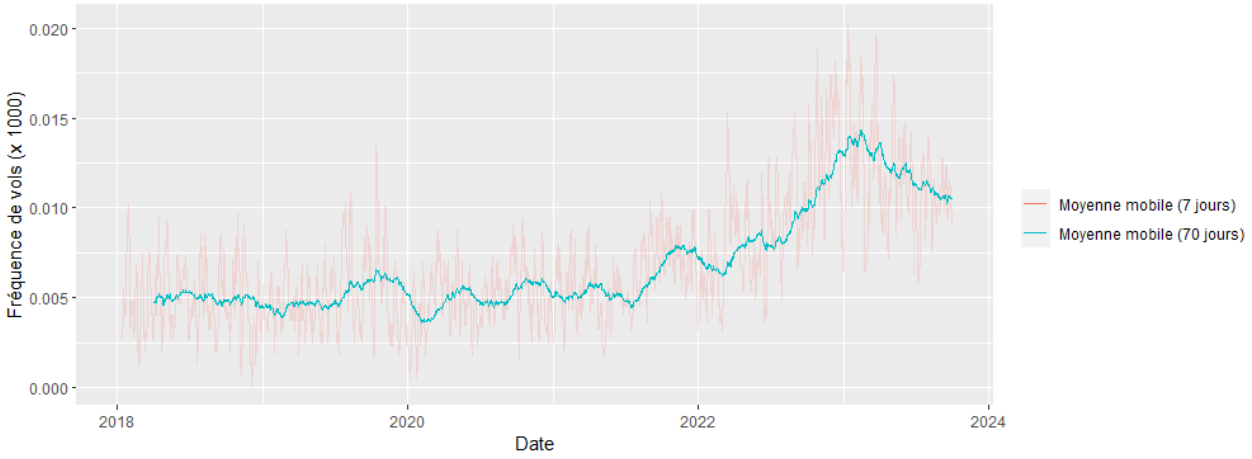
La première covariable ajoutée au modèle GLR est la puissance du véhicule (Chevaux). Cette covariable a été séparée en deux groupes : les véhicules avec un Chevaux inférieur ou égal à 190, et les autres véhicules. Plus formellement, nous avons ainsi :

$$X_1 = \begin{cases} 1, & \text{si Cheveaux} \leq 190 \\ 0, & \text{sinon.} \end{cases}$$

La figure 5.1a illustre la fréquence de vols de la variable Chevaux $\leq$ 190 et la figure 5.1b montre la fréquence de vols de la variable Chevaux > 190 entre 2018 et 2024. Pour les véhicules ayant un Chevaux supérieur à 190, on constate que la fréquence est en général plus élevée que pour ceux avec un Chevaux inférieur ou égal à 190. De plus, Chevaux > 190 se révèle généralement plus stable que Chevaux $\leq$ 190, présentant des variations moins prononcées jusqu'en 2022, avant d'atteindre un pic en 2023, pour ensuite redescendre.



(a) Cheveaux ≤ 190



(b) Cheveaux > 190

Figure 5.1 – Fréquence de vols quotidienne pour chacune des modalités de la covariable Chevaux

Pour encore mieux distinguer l'expérience de sinistralité des deux modalités de la variable Chevaux, la figure 5.2 illustre la distribution de la fréquence de vols quotidienne pour chacune des modalités. La courbe a été obtenue grâce à une estimation par noyau, basée sur la théorie présentée dans la section 1.3. Un noyau basé sur la loi gaussienne a été utilisé, ainsi qu'un paramètre de lissage accru pour adoucir la distribution.

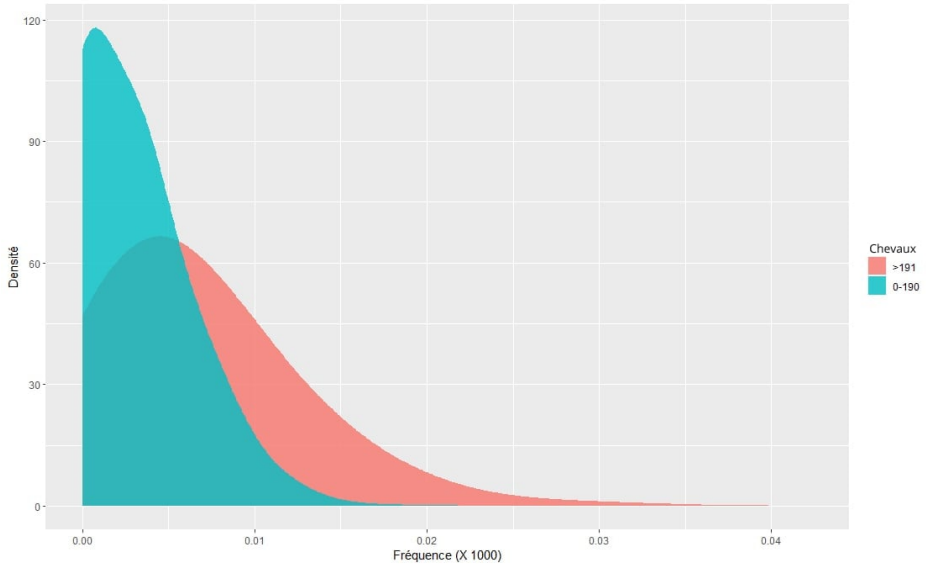
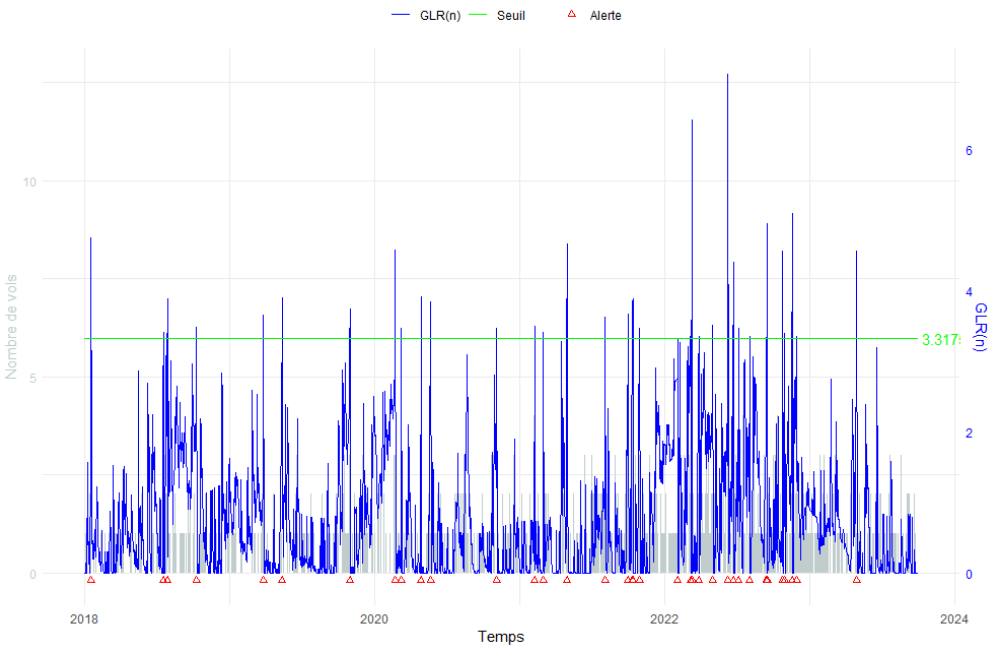


Figure 5.2 – L'estimation par noyau de la fréquence de vols quotidienne pour chacune des modalités de la covariable Chevaux

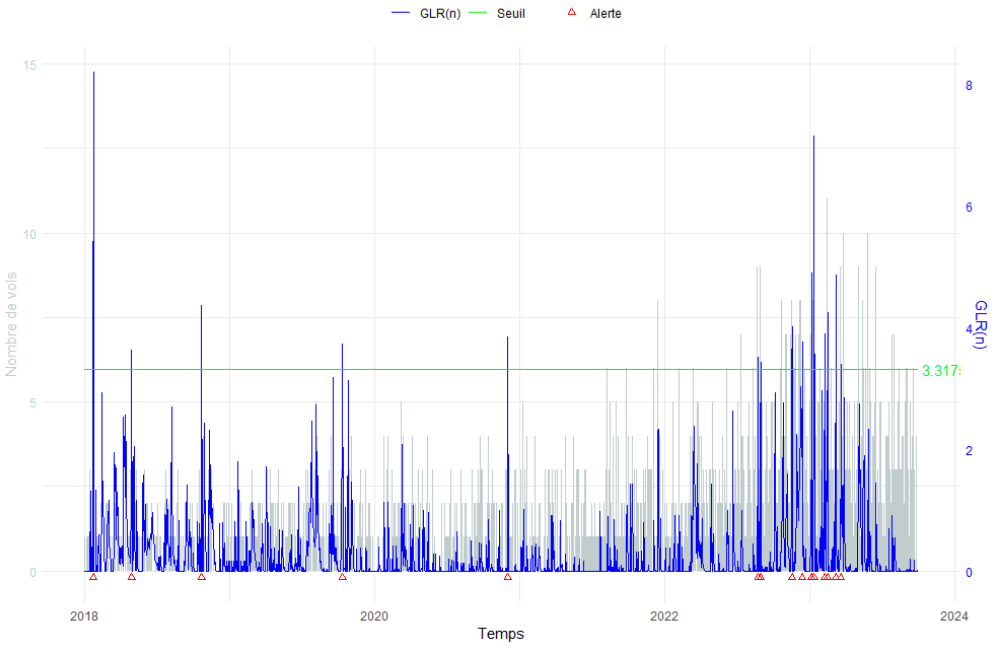
On peut voir, à la figure 5.2 que les données avec Chevaux ≤ 190, ont une concentration importante à zéro, surpassant presque toutes les autres valeurs avec une distribution très restreinte. On constate, que les données avec Chevaux > 190, ont une proportion de valeurs nulles réduite et que la répartition des données est plus équilibrée.

5.3.1.1 Différents GLR indépendants

Grâce aux analyses sommaires que nous venons de faire, nous pouvons comprendre qu'il pourrait être pertinent d'inclure la puissance du véhicule dans les modèles de détection d'épidémie de vols. Nous allons appliquer un modèle GLR indépendant pour chacune des modalités. Pour la covariable Chevaux, qui ne comporte que deux modalités, cela revient à estimer deux modèles GLR distincts. La figure 5.3 présente ainsi les résultats pour chacune des modalités.



(a) Cheveaux ≤ 190



(b) Cheveaux > 190

Figure 5.3 – Application d'approches GLR indépendantes pour les deux modalités de la variable Chevaux

Nous pouvons faire quelques observations tirées de la figure 5.3a (pour les véhicules avec Chevaux ≤ 190) :

— On remarque un nombre important d'alertes (36) et cela de manière constante tout au long des observations sous surveillance ;
— En comparant les alertes communes avec le GLR de la figure 4.6 qui évaluait la totalité du portefeuille, on observe une alerte le 13 octobre 2021, et les autres alertes se concentrent principalement en 2022, notamment aux dates suivantes : 11 mars, 9 juin, 18 novembre et 29 novembre.

Pour le modèle GLR de la figure 5.3b (pour les véhicules avec Chevaux > 190) :

— Un plus petit nombre d'alertes (15) sont présentes ;
— Les résultats obtenus ressemblent beaucoup à ceux du modèle de la figure 4.6, avec un peu moins d'alertes et celles-ci semblent survenir sensiblement aux mêmes périodes.

Il y a une seule alerte commune entre la figure 5.3b et la figure 5.3a, et ce, le 18 novembre 2022.

5.3.1.2 Approche par GLR-étendu

Tel que nous l'avons introduit dans la sous-section 5.2.2, il devient ainsi pertinent d'utiliser un modèle intégrant des variables explicatives directement dans le paramètre de moyenne de l'approche GLR. Cela permet une détection plus fine et précise des anomalies. Le modèle GLR-étendu offre des prévisions plus robustes et mieux adaptées au contexte des données.

La moyenne utilisée dans l'application du modèle GLR-étendu avec la variable Chevaux sera :

$$\mu_{0,t}^{j} = d_t \exp\left(\beta_0 + \beta_1 t + \sum_{s=1}^{S} [\beta_{2s}\cos(\omega st) + \beta_{2s+1}\sin(\omega st)] + \gamma_1 X_1\right). \tag{5.5}$$

L'estimation de la log-vraisemblance de la densité augmentée s'obtient en utilisant l'équation (5.2), ainsi que les équations (5.3) et (5.4), la première et la deuxième dérivée respectivement.

La figure 5.4 illustre les résultats obtenus pour l'approche proposée. Il est important de comparer l'effet de l'ajout de la variable Chevaux avec l'analyse sans cette variable, telle que présentée à la section précédente,

à la figure 4.6. La figure 5.4 montre les résultats du modèle GLR incluant la variable Chevaux, en utilisant les mêmes paramètres que la figure 4.6, à savoir un seuil c_γ de 3.3175 et une saisonnalité annuelle avec $S = 1$. La figure 5.4 et 4.6 affiche le même nombre d'alertes, et ce, aux mêmes dates. Par contre, l'inclusion de Chevaux permet de réduire l'ampleur entre les différentes statistiques, ce qui suggère que cette variable a contribué à lisser les valeurs du GLR et à diminuer les fluctuations extrêmes. La présence de Chevaux semble rendre le modèle plus stable, avec moins de variabilité. De plus, on remarque que les valeurs GLR(n) sont de façon générale plus grandes.

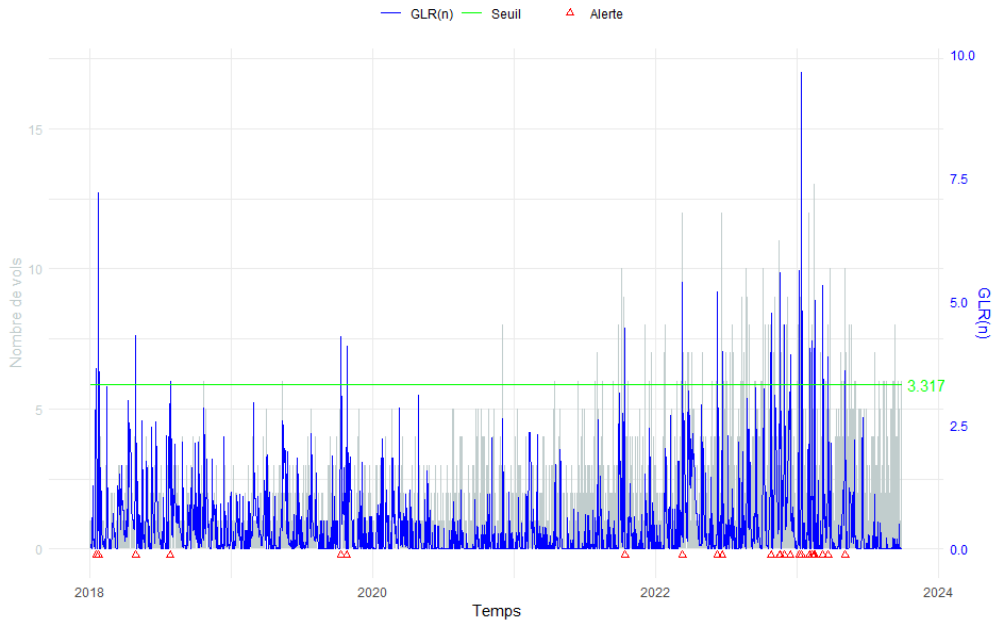


Figure 5.4 – Application du modèle GLR en prenant en compte l'exposition et la variable Chevaux

En comparant les GLR indépendants des figures 5.1a et 5.1b de la sous-section 5.3.1.1 précédente avec notre nouvelle figure 5.4, on observe une seule alerte commune, soit le 18 novembre 2022. On remarque que le GLR indépendant avec la modalité Chevaux > 190 et la figure 5.4 se ressemblent beaucoup, à la différence que la figure 5.4 contient plus d'alertes. Le nombre total d'alertes du modèle GLR-étendu est de 24, ce qui se situe entre les deux GLR par modalité, soit 36 et 15 pour Chevaux $\leq$ 190 et Chevaux > 190 respectivement. Le GLR-étendu semble faire un bon compromis entre les deux GLR par modalité.

5.3.2 Ajout du régresseur Empat

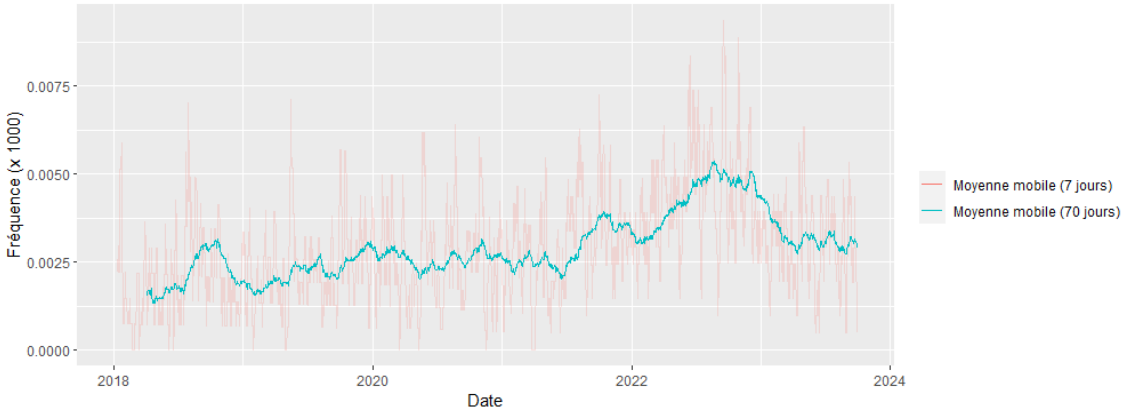
La prochaine covariable ajoutée au modèle de la sous-section précédente 5.3.1 est l'empattement du véhicule (Empat). Nous avons ainsi une nouvelle variable à créée :

$$X_2 = \begin{cases} 1, & \text{si Empat} \leq 2745 \\ 0, & \text{sinon.} \end{cases}$$

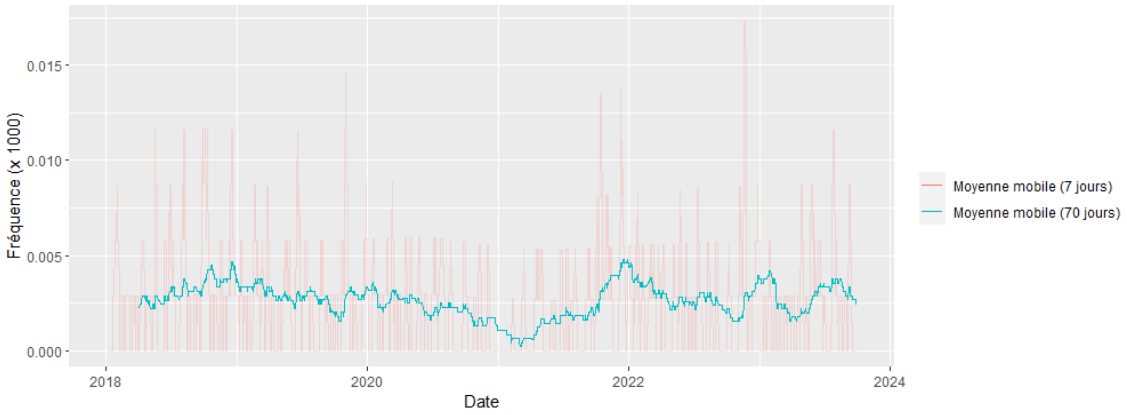
Étant donné que Chevaux et Empat ont chacun deux modalités, cela implique l'estimation de quatre approches GLR indépendantes. Pour analyser les effets des variables Chevaux et Empat dans le GLR, nous avons ainsi divisé le portefeuille d'assurés en quatre groupes selon les valeurs de Chevaux et Empat.

La figure 5.5 illustre l'évolution de la fréquence de vols pour chacune des combinaisons de modalités :

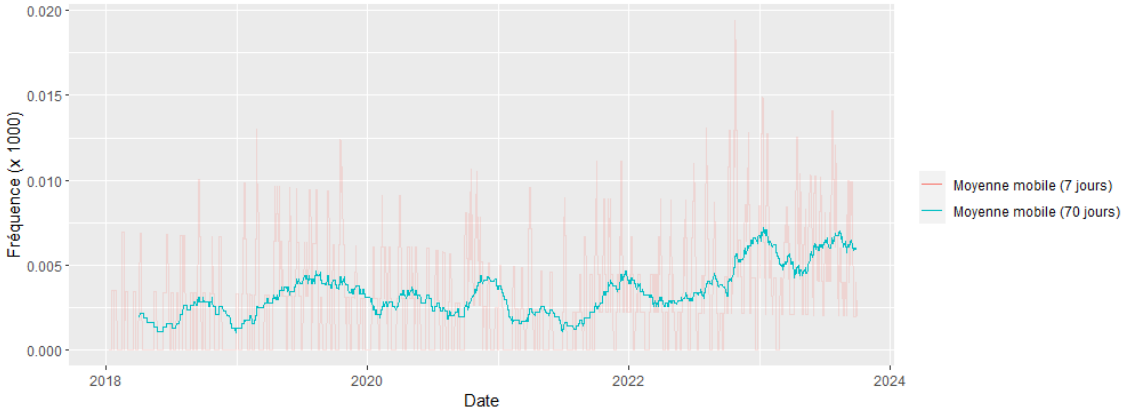
- À la figure 5.5a, la moyenne mobile 70 jours semble se maintenir dans une fréquence d'à peu près 0.0025 jusqu'à la moitié de 2021. Ensuite, un double saut survient jusqu'à atteindre un sommet à la moitié de 2022, pour ensuite redescendre drastiquement jusqu'au début de 2024 pour se stabiliser ;
- À la figure 5.5b, on observe que la moyenne mobile 70 jours de la fréquence ne dépasse jamais la valeur de 0.005 et va même jusqu'à atteindre une valeur très près de zéro au début de 2021 ;
- À la figure 5.5c, on observe une fluctuation de la moyenne mobile atteignant les mêmes valeurs basses et les mêmes valeurs hautes jusqu'à la fin de 2022. Ensuite, elle atteint un sommet au début de 2023 pour redescendre et remonter presque aussitôt à peu près à la même valeur que le sommet précédent ;
- À la figure 5.5d, la fréquence est plus élevée que pour les autres modalités pendant presque toute la période, et elle est assez stable jusqu'au début de l'année 2022. Par ailleurs, une forte tendance à la hausse de la fréquence est observée à partir de 2022, atteignant son maximum au début de 2023 avant de redescendre.



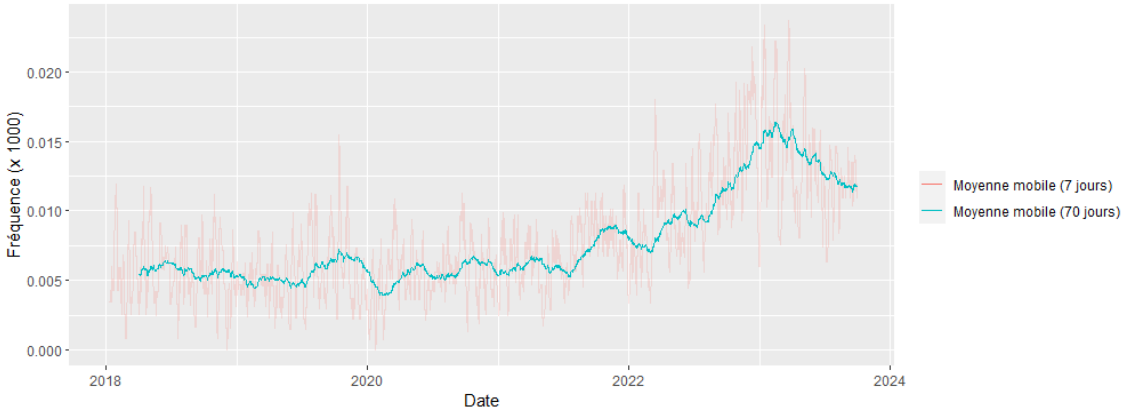
(a) Fréquence quotidienne de vols avec Chevaux $\leq$ 190 et Empat $\leq$ 2745



(b) Fréquence quotidienne de vols avec Chevaux $\leq$ 190 et Empat > 2745



(c) Fréquence quotidienne de vols avec Chevaux > 190 et Empat $\leq$ 2745



(d) Fréquence quotidienne de vols avec Chevaux > 190 et Empat > 2745

Figure 5.5 – Moyenne mobile quotidienne de la sinistralité avec les covariables Chevaux et Empat

Similairement à ce que nous avons fait pour la variable Chevaux, la figure 5.6 illustre la distribution de la fréquence de vols quotidienne pour chacune des combinaisons de modalités :

— On constate que les données avec $(X_1 = 1, X_2 = 1)$, ont une concentration importante à zéro, surpassant presque toutes les autres valeurs avec une distribution très restreinte ;
— Pour $(X_1 = 0, X_2 = 1)$ et $(X_1 = 1, X_2 = 0)$, on remarque une concentration importante à zéro, mais cette fois-ci les distributions peuvent atteindre des valeurs beaucoup plus grandes ;
— Avec $(X_1 = 0, X_2 = 0)$, une proportion de valeurs nulles réduite est observée, comparativement aux trois autres, et la répartition des données est plus équilibrée.

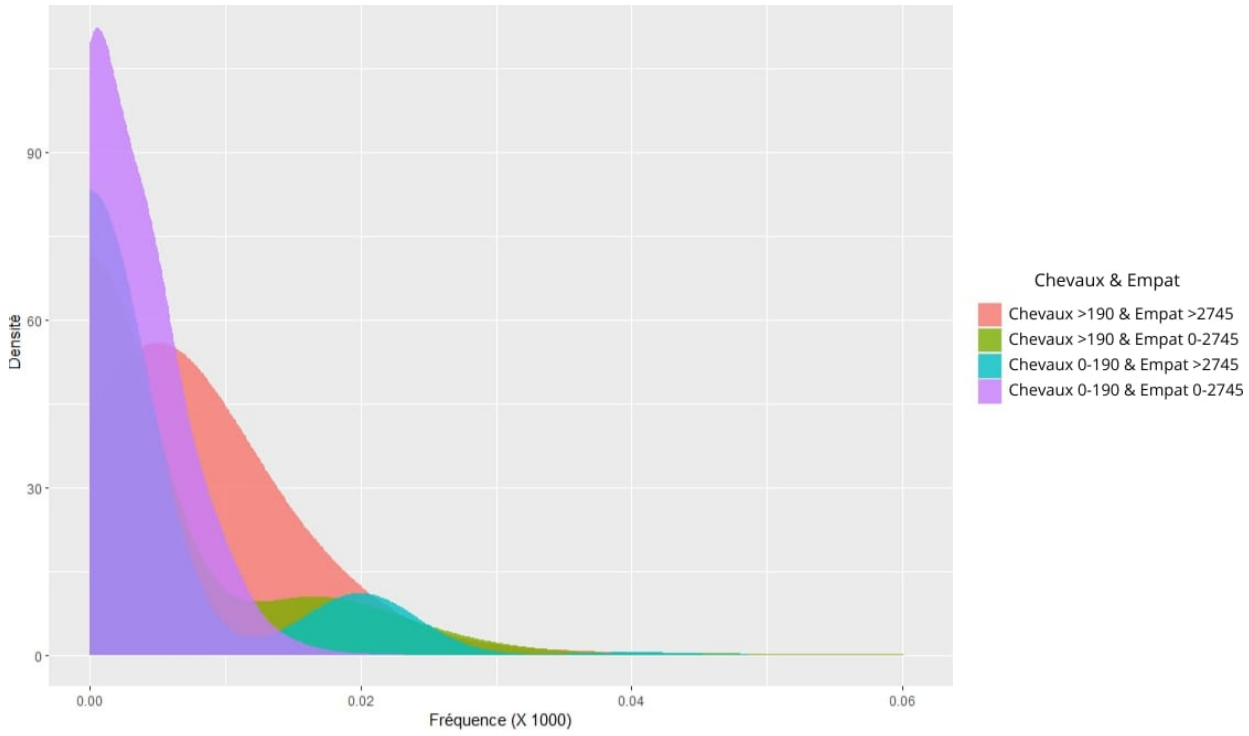
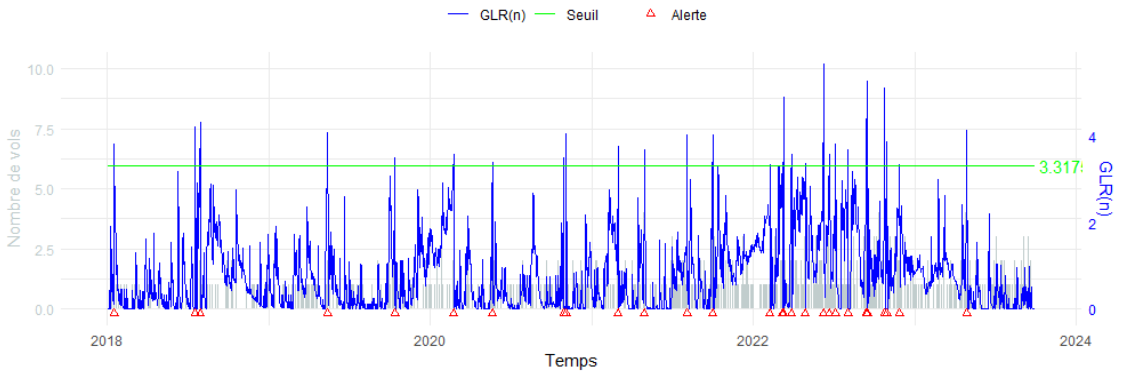


Figure 5.6 – L'estimation par noyau de la fréquence par jour pour chacune des modalités des covariables Chevaux et Empat

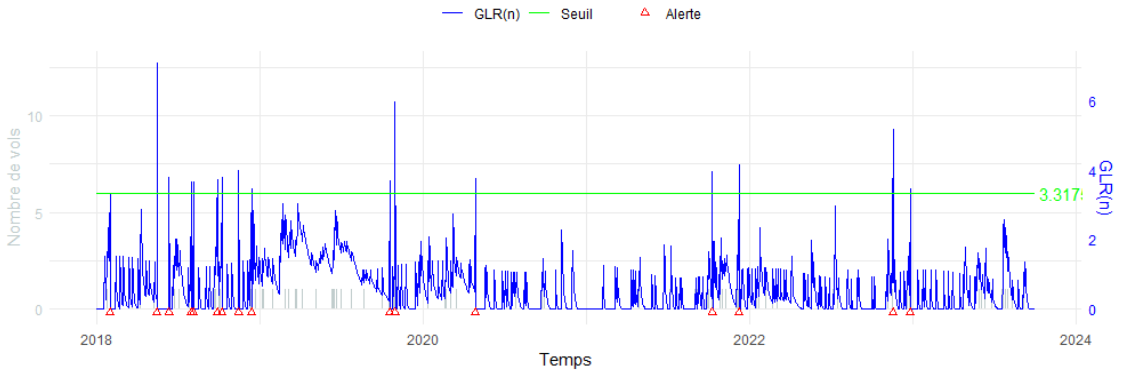
5.3.2.1 Différents GLR indépendants

Les 4 approches GLR générées sont illustrées à la figure 5.7. Quelques observations pertinentes peuvent être faites :

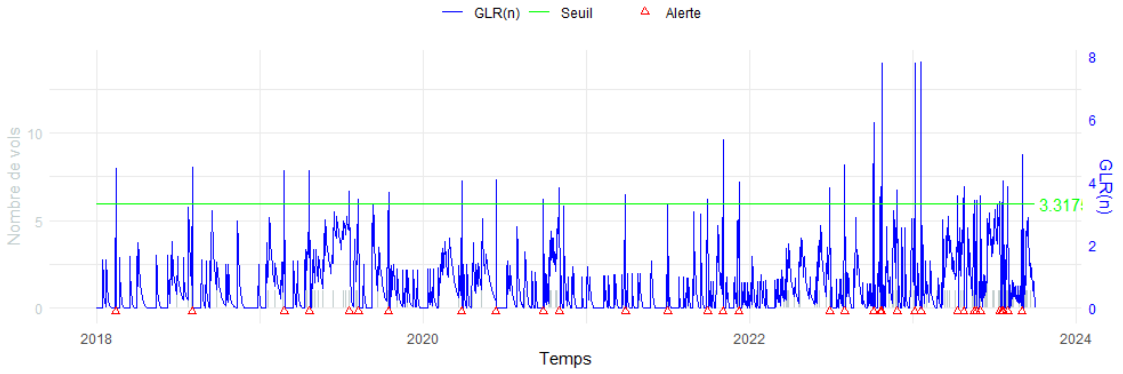
- Une abondance d'alertes est constatée dans toutes les figures, sauf dans la figure 5.7d, et ce, de manière constante tout au long des observations sous surveillance ;
- La figure 5.7d est similaire à la figure 5.4, avec moins d'alertes et celles-ci apparaissent sensiblement aux mêmes périodes ;
- La figure 5.7b a beaucoup d'alertes en 2018, ce qui diffère des autres figures ;
- Pour les figures 5.7b et 5.7c, on observe un nombre de vols très faible et souvent même de zéro ;
- Pour la figure 5.7a, on remarque un nombre de vols assez faible, mais beaucoup moins de zéro que les figures 5.7b et 5.7c, mais beaucoup plus que la figure 5.5d.



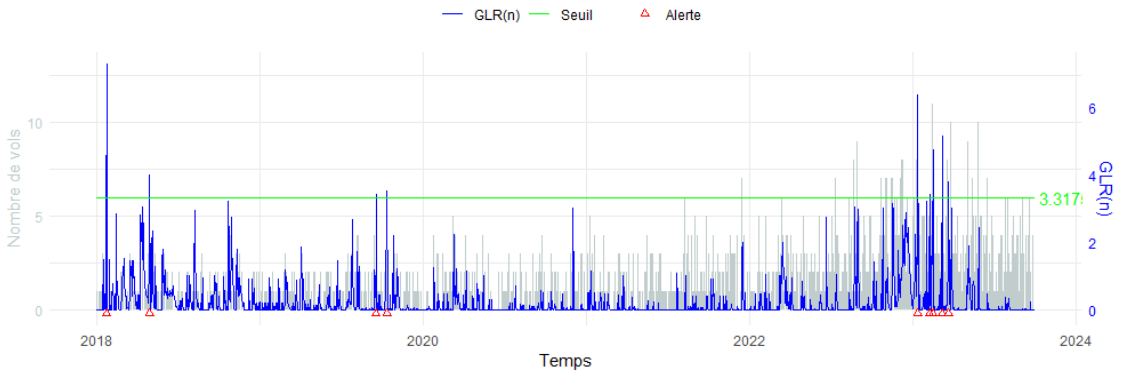
(a) GLR avec Chevaux $\leq$ 190 et Empat $\leq$ 2745



(b) GLR avec Chevaux $\leq$ 190 et Empat > 2745



(c) GLR avec Chevaux > 190 et Empat $\leq$ 2745



(d) GLR avec Chevaux > 190 et Empat > 2745

Figure 5.7 – Application d'approches GLR indépendantes pour les quatre modalités des variables Chevaux et Empat

5.3.2.2 Approche par GLR-étendu

Tout comme dans la sous-section 5.3.1.2, nous appliquerons maintenant le GLR-étendu en ajoutant la covariable Empat. Nous utiliserons le même procédé et les mêmes paramètres que pour le modèle avec la covariable Chevaux, mais la moyenne sera la suivante :

$$\mu_{0,t}^{j} = d_t \exp\left(\beta_0 + \beta_1 t + \sum_{s=1}^{S} [\beta_{2s}\cos(\omega s t) + \beta_{2s+1}\sin(\omega s t)] + \sum_{j=1}^{2} \gamma_j X_j\right), \tag{5.6}$$

et nous utiliserons ces moyennes pour calculer l'équation de la log-vraisemblance (5.2) ainsi que les équations des dérivées (5.3) et (5.4). Une fois ces calculs effectués, nous obtiendrons le GLR-étendu présenté dans la figure 5.8.

On observe que la figure 5.8 présente des alertes en 2018 et 2019, aucune en 2020, et une concentration importante d'alertes à la fin de 2022. En comparant cette figure avec la figure 5.4, soit le GLR-étendu avec la variable Chevaux, nous notons une alerte supplémentaire au début de 2018, une alerte supplémentaire au milieu de 2019, une autre en 2020, et trois alertes supplémentaires en 2023, soit un total de six alertes supplémentaires. En ce qui concerne les valeurs GLR(n), elles sont généralement plus élevées dans la figure 5.8 que dans la figure 5.4. L'inclusion de Empat permet de réduire la densité des pics, tout comme l'a fait l'inclusion de la covariable Chevaux lors de son inclusion dans le GLR de base. Les valeurs GLR(n) ont contri-

bué à lisser les valeurs du GLR et à diminuer les fluctuations extrêmes. La présence de Empat semble rendre le modèle plus stable.

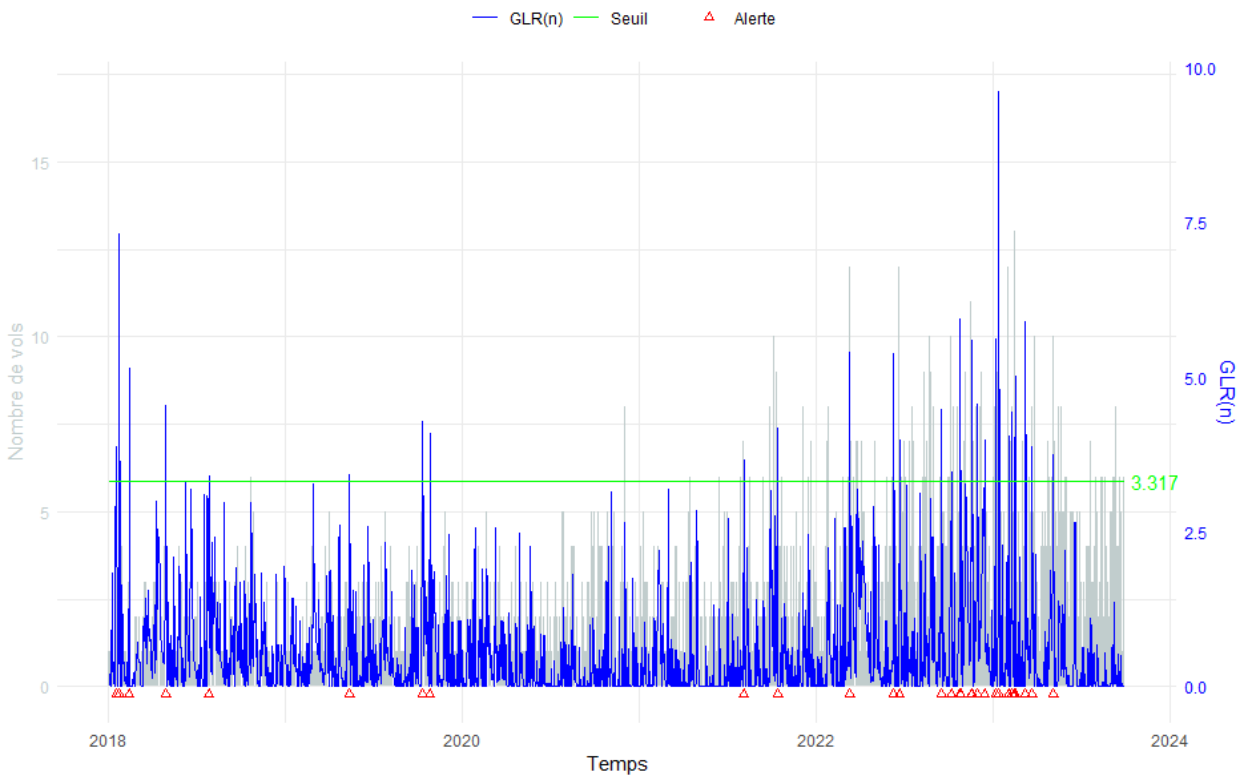


Figure 5.8 – Application du modèle GLR sur nos données en prenant compte de l'exposition, les variables Chevaux et Empat

En comparant la figure 5.8 avec la figure 5.7d de la sous-section précédente 5.3.2.1, on remarque que les figures se ressemblent beaucoup, à la différence que la figure 5.8 contient beaucoup plus d'alertes. La figure 5.8 compte 30 alertes en tout. Il y a 28 alertes à la figure 5.7a, 16 alertes à la figure 5.7b, 34 alertes pour la figure 5.7c et 9 alertes pour la figure 5.7d. Le GLR-étendu, encore une fois, semble faire un bon compromis entre les quatre GLR par modalité.

5.3.3 Ajout du régresseur Type

La troisième variable à ajouter sera le type de véhicule, soit les camions ou tous les autres véhicules qui ne sont pas des camions (répertorié comme Type dans la base de données). Nous intégrons ainsi la nouvelle

variable suivante dans le modèle :

$$X_3 = \begin{cases} 1, & \text{si le véhicule n'est pas un camion} \\ 0, & \text{sinon.} \end{cases}$$

Étant donné que Chevaux, Empat et Type ont tous deux modalités, cela implique l'estimation de huit approches GLR indépendantes. Pour analyser les effets des variables rajoutées dans le GLR, nous avons ainsi divisé le portefeuille d'assurés en huit groupes. Les distributions et les fréquences de chaque classe de risque sont présentées à l'annexe A.

Comme nous l'avons fait pour la variable Empat, la figure A.1 illustre la distribution de la fréquence de vols quotidienne pour chacune des combinaisons de modalités :

- Toutes les classes de risque ont une concentration importante en zéro, surpassant presque toutes les autres valeurs sauf ($X_1 = 0, X_2 = 0, X_3 = 0$) ;
- La classe de risque ($X_1 = 0, X_2 = 0, X_3 = 0$) possède une proportion de valeurs nulles réduite et la répartition des données est plus équilibrée.

L'évolution de la fréquence de vols des moyennes mobiles 70 jours pour chacune des combinaisons de modalités qui sont des camions :

- La figure A.2, semble se maintenir dans une fréquence d'à peu près 0.0025 jusqu'à la moitié de 2021. Ensuite, elle atteint un sommet à environ la moitié de 2022 pour ensuite redescendre en dessous de 0.005 ;
- La figure A.3, représente une fréquence très faible allant même parfois à zéro. Elle oscille entre 0 et 0.005 jusqu'au début de 2022 pour augmenter continuement par la suite jusqu'à se stabiliser à environ 0.01 en 2024 ;
- La figure A.4, représente une fréquence qui oscille entre 0 et 0.005 sur toute la période ;
- La figure A.5, semble se maintenir à 0.005 jusqu'à la moitié de 2021 pour ensuite atteindre un maximum au début de 2024.

L'évolution de la fréquence de vols des moyennes mobiles 70 jours pour chacune des combinaisons de modalités qui ne sont pas des camions :

— La figure A.6, représente une fréquence oscillante entre 0.002 et 0.003 ;
— La figure A.7, représente une fréquence très faible qui dépassera deux fois la valeur de 0.005 et atteint zéro au début 2024 pour revenir aux alentours de 0.005 ;
— La figure A.8, oscillera entre 0 et 0.005 ;
— La figure A.9, semble se maintenir aux alentours de de 0.005 jusqu'à la moitié de 2022 pour ensuite atteindre un sommet au début de 2024.

Globalement, la figure A.5 représente la fréquence de vols la plus élevée. On constate que cette figure ressemble beaucoup à la figure 2.1a de la sous-section 2.1. De plus, on remarque que la fréquence de la figure A.5 est supérieure à la fréquence de la figure 2.1a.

5.3.3.1 Différents GLR indépendants

Les résultats pour les GLR indépendants avec l'ajout de Type sont présentés à l'annexe B. Quelques observations pertinentes peuvent être faites :

— À la figure B.1, on obtient 41 alertes. Une grosse masse d'alertes est observée à partir de la fin de 2021. On remarque un nombre de vols très faible et parfois de zéro ;
— À la figure B.2, on obtient 47 alertes. On observe une grosse masse d'alertes à partir de la fin de 2023. On remarque un nombre de vols très faible et souvent de zéro ;
— À la figure B.3, on obtient 14 alertes. Les alertes sont assez disparates sur toute la durée. On observe que le nombre de vols est presque toujours de zéro ;
— À la figure B.4, on obtient 22 alertes. Une grosse masse d'alertes est observée à la fin de 2022. Le nombre de vols est élevé et la présence de zéro est moindre ;
— À la figure B.5, on obtient 13 alertes. Les alertes sont disparates sur toute la durée. On observe un nombre de vols faible et parfois de zéro ;
— À la figure B.6, on obtient 10 alertes. Une seule masse d'alertes est observable en 2019. On remarque que le nombre de vols est presque toujours de zéro ;
— À la figure B.7, on obtient 7 alertes. Il y a 3 alertes en 2018, 1 alerte en 2019 et 3 autres en 2021. On remarque que le nombre de vols est presque toujours de zéro ;
— À la figure B.8, on obtient 10 alertes. Les alertes sont concentrées en 2018, suivi d'une alerte à la fin de 2019 et une autre à la fin de 2020. On observe un nombre de vols très faible et souvent de zéro.

Pour résumé, le GLR indépendant représentant le plus d'alertes est la figure B.2. À l'inverse, le GLR indépendant contenant le moins d'alertes est la figure B.7. Pour ce qui est de la figure B.4, elle ressemble énormément aux GLR-étendu obtenus des figures 5.4 et 5.8 avec des alertes survenant sensiblement aux mêmes périodes.

5.3.3.2 Approche par GLR-étendu

Tout comme dans la sous-section 5.3.2.2, nous appliquerons maintenant le GLR-étendu en ajoutant la covariable Type. Nous suivrons le même procédé que pour le modèle avec la covariable Empat, mais la moyenne sera la suivante :

$$\mu_{0,t}^{j} = d_t \exp\left(\beta_0 + \beta_1 t + \sum_{s=1}^{S} [\beta_{2s}\cos(\omega st) + \beta_{2s+1}\sin(\omega st)] + \sum_{j=1}^{3} \gamma_j X_j\right), \tag{5.7}$$

et nous utiliserons les équations nécessaires pour obtenir le GLR-étendu, à savoir les équations (5.2), (5.3), et (5.4).

La figure 5.9, montre des alertes en 2018 et 2019, une seule alerte en 2020, et une concentration importante d'alertes à la fin de 2022. Elle contient 44 alertes, soit 14 alertes de plus que le GLR-étendu précédent de la figure 5.8 de la sous-section 5.3.2.2. L'ajout de la variable Type au GLR augmente une fois de plus le nombre d'alertes, dépassant celles observées dans les figures 5.4 et 5.8. Les valeurs GLR(n) sont plus lisses et les fluctuations extrêmes sont moindres. La présence de Type semble rendre le modèle plus stable. En comparant la figure 5.9 avec les GLR indépendants de l'annexe B, on remarque que les figures 5.9 et B.4 se ressemblent beaucoup, à la différence que la figure 5.9 contient plus d'alertes.

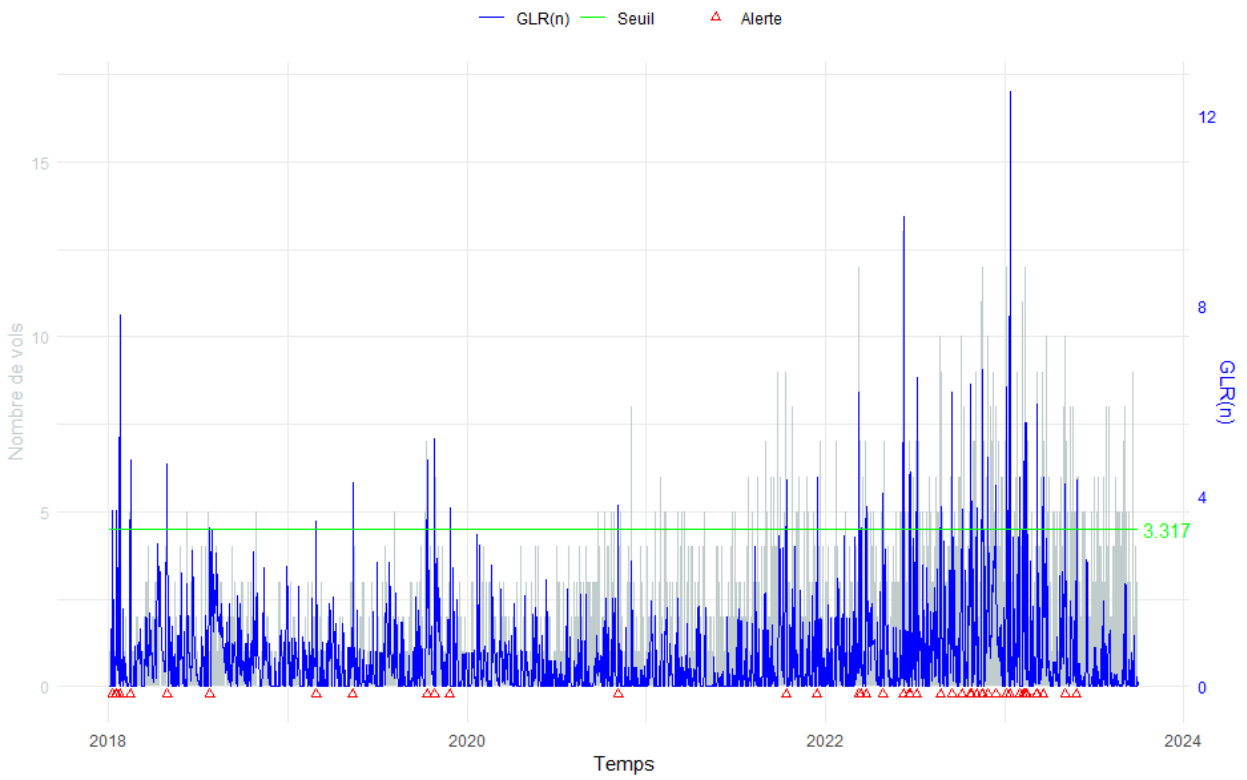


Figure 5.9 – Application du modèle GLR sur les données en prenant compte de l'exposition et les variables Chevaux, Empat et Type

5.3.3.3 Analyse GLR-étendu globale

Il est clair que l'ajout de covariables est essentiel pour améliorer la détection des anomalies et l'identification des tendances dans les données. En intégrant progressivement des covariables dans les graphiques du GLR-étendu, nous augmentons la sensibilité du modèle, ce qui permet de mieux saisir les variations sous-jacentes. Nous renommerons les modèles GLR-étendu pour simplifier les futures références comme suit :

- GLR(0) : Application du modèle GLR de base, sans aucune variable explicative (voir la figure 4.6 pour le résultat appliqué sur les données) ;
- GLR(1) : Application du modèle GLR sur les données en prenant en compte l'exposition et la variable Chevaux (voir la figure 5.4 pour le résultat appliqué sur les données) ;
- GLR(2) : Application du modèle GLR sur les données en prenant en compte l'exposition ainsi que les

variables Chevaux et Empat (voir la figure 5.8 pour le résultat appliqué sur les données) ;

— GLR(3) : Application du modèle GLR sur les données en prenant en compte l'exposition ainsi que les variables Chevaux, Empat et Type (voir la figure 5.9 pour le résultat appliqué sur les données).

Dans le graphique GLR(1), où une seule covariable est utilisée, on observe une sensibilité limitée pour la détection des anomalies. En revanche, les graphiques GLR(2) et GLR(3), qui ajoutent successivement des covariables, montrent une augmentation notable du nombre d'alertes détectées. Cela illustre comment l'ajout de covariables améliore la sensibilité du modèle à identifier des anomalies ou des événements significatifs. Chaque nouvelle covariable intégrée dans le GLR-étendu renforce la capacité du modèle à détecter des changements qui pourraient ne pas être visibles avec un modèle plus simple, mettant en évidence l'importance de considérer des facteurs additionnels qui influencent les données.

Cependant, l'augmentation de la complexité du modèle avec des covariables supplémentaires peut également accroître le risque de fausses alertes. Il est donc essentiel de calibrer le modèle avec soin pour maintenir un équilibre entre sensibilité et spécificité. Si d'autres covariables avaient été ajoutées, il aurait été probablement nécessaire de réajuster le seuil pour répondre aux besoins de l'analyse. L'application du GLR-étendu, enrichie par des covariables, améliore l'analyse en la rendant plus robuste face aux séries temporelles influencées par divers facteurs. De manière générale, toutes les alertes doivent faire l'objet d'une investigation. Lorsqu'une seule alerte isolée se produit, il est possible qu'il s'agisse d'une valeur aberrante ou d'une fausse alerte.

5.3.3.4 Diagnostics avec les données utilisées

Concrètement, avec les données de vols que nous avons utilisées pour l'application des approches, il est intéressant de faire quelques observations sur les résultats obtenus. Ainsi, si la compagnie d'assurance qui nous a partagé les données avait utilisé quotidiennement le modèle GLR(3) avec ses données de vols, grâce aux alertes que nous voyons dans la figure 5.9, elle aurait été avertie régulièrement au début de l'année 2018, et à de nombreuses reprises à partir du 11 mars 2022.

Lorsqu'on analyse les alertes, il est essentiel de comprendre les causes de ces valeurs aberrantes. Il est donc nécessaire d'analyser les résultats obtenus afin de déterminer les causes de ces anomalies. Les GLR indépendants montrent comment les classes de risque réagissent et fournissent des informations sur les alertes ainsi que sur leurs dates de survenance. Le GLR(3) est un bon compromis entre les différents GLR

indépendants.

Les GLR indépendants ne sont pas un bon indicateur hiérarchique des risques pour les différentes classes de risque. Par contre, lorsque le GLR indépendant ressemble au GLR(3), donc possède moins d'alertes tout en survenant sensiblement aux mêmes périodes, on peut alors supposer que la classe de risque utilisée dans l'obtention du GLR indépendant a fortement contribué au GLR(3). Par conséquent, cette classe de risque contribuera énormément aux valeurs aberrantes obtenues une fois appliquées dans le GLR(3). Dans notre cas, le GLR indépendant de la figure B.4, correspondant aux modalités Chevaux > 190, Empat > 2745 et Type de type camion, est similaire au GLR(3). De plus, si la compagnie d'assurance veut surveiller une classe de risque en particulier avec des paramètres ajustés aux besoins de l'analyse, le GLR indépendant pourrait être un outil très intéressant.

Les fréquences des différentes classes de risque présentées à l'annexe A peuvent aider à identifier les variations susceptibles de contribuer aux alertes du GLR(3). En somme, en analysant les fréquences des classes de risque des figures A.2 à A.9, on constate que les modalités Chevaux > 190, Empat > 2745 et Type de type camion possèdent généralement les fréquences les plus élevées. La figure A.5, présente la fréquence la plus élevée et possède ces modalités. Cette constatation est corroborée par la comparaison de la figure A.5 avec la figure 2.1a. La figure A.5 possède une fréquence supérieure à la figure 2.1a lorsque cette classe de risque est utilisée. On peut aussi remarquer que les deux figures ont des courbes très similaires. Ce qui suggère une réduction de l'exposition et un nombre de vols constant ou légèrement réduit. Cela implique donc que $(X_1 = 0, X_2 = 0, X_3 = 0)$ capture bien les tendances réelles et que cette classe de risque est généralement responsable de nombres élevés de vols. De plus, l'analyse des covariables de la section 2.2 vient appuyer les constatations précédentes indiquant que ces modalités sont les plus risquées.

Les mêmes analyses précédentes peuvent être appliquées sur les GLR-étendu des sous-sections 5.3.1 et 5.3.2. Pour le GLR(1), la modalité Chevaux > 190 sera la modalité la plus risquée. Pour ce qui est du GLR(2), les modalités Chevaux > 190 et Empat > 2745 seront les modalités les plus risquées.

Si la compagnie d'assurance avait utilisé le modèle GLR(3), elle aurait investigué les alertes subséquentes apparaissant à partir du 11 mars 2022. Elle aurait constaté que les véhicules de type camion avec une puissance moteur supérieure à 190 et un empattement supérieur à 2745 sont les véhicules les plus risqués. Plusieurs mesures auraient alors pu être mises en place. En voici quelques-unes :

— L'ajustement des polices d'assurance : Établir des primes plus élevées pour les véhicules les plus à risque. Proposer des options de couverture adaptées ;
— L'encouragement de technologies anti-vols : Offrir des réductions de primes pour les assurés qui installent des dispositifs de sécurité avancés, tels que des systèmes de localisation GPS, des systèmes d'alarmes ou des dispositifs de coupe-courant ;
— La collaboration avec les fabriquants : Travailler avec les fabricants de véhicules pour développer des véhicules avec des caractéristiques de sécurité améliorées pour dissuader les voleurs.

Pour conclure, contrairement au modèle GLR de base sans covariable ajoutée de la sous-section 4.3.5, qui ne permet pas de recueillir d'informations supplémentaires sur les covariables susceptibles de contribuer aux valeurs aberrantes, le modèle GLR-étendu le permet.

5.3.4 Ajout d'autres régresseurs

Nous avons montré le résultat du modèle avec une approche GLR(3) intégrant ainsi 3 covariables dans le modèle GLR. Il serait possible d'aller encore plus loin avec un GLR(4), un GLR(5) ou, de manière plus générale, un GLR(n) pour n'importe quelle valeur entière de n représentant le nombre de covariables à ajouter dans l'approche. Pour appliquer un tel modèle, nous devons toutefois créer des données en regroupant les assurés par jour selon toutes les classes de risque possibles, avec le nombre de vols dans chaque classe et les expositions correspondantes. Cela conduit à une base de données avec au minimum 2^n classes de risque pour chaque instant t, augmentant ainsi considérablement la taille des données à traiter. Ceci est un problème majeur de l'approche, car cette augmentation de données augmente significativement le temps de calcul de la statistique GLR(n). Pour illustration, pour générer la figure 5.9 qui se base sur un GLR(3), le temps de calcul informatique en R a été d'environ 24 heures. En plus de limiter le nombre de covariables, cette contrainte du modèle nous empêche aussi d'ajouter des covariables ayant de nombreuses modalités, comme le modèle ou la marque du véhicule assuré. Il faudrait probablement penser à une autre manière de travailler avec un tel modèle dans le futur si nous voulons améliorer la précision de l'analyse.

CONCLUSION

Ce mémoire met en évidence l'importance de l'utilisation de modèles statistiques avancés pour surveiller les vols de voitures, apportant des perspectives nouvelles et des outils pratiques pour les compagnies d'assurance. Les modèles Farrington et GLR, initialement conçus pour la surveillance des maladies infectieuses, démontrent une adaptabilité prometteuse dans le contexte des vols de véhicules. Ils permettent de détecter des anomalies dans les données et fournissent des alertes précoces, aidant ainsi les compagnies à mieux gérer les risques.

Le modèle Farrington, basé sur la comparaison d'un sous-ensemble de données historiques, est efficace pour identifier la saisonnalité et les variations des données. Cependant, il présente des limitations en raison de son utilisation partielle des données et de sa sensibilité aux changements récents qui pourraient ne pas être capturés par l'historique. Les résultats montrent que le modèle peut détecter des anomalies, mais son efficacité dépend de la qualité et de la pertinence des données utilisées pour la comparaison. Une analyse plus approfondie des seuils et une révision continue des critères de comparaison sont nécessaires pour améliorer la précision du modèle.

D'autre part, le modèle GLR offre une approche plus globale en exploitant toutes les données disponibles et en capturant les tendances saisonnières de manière paramétrique. Sa capacité à maximiser le rapport de vraisemblance pour identifier les points de changement permet de détecter avec précision les variations significatives. Le GLR est robuste dans divers scénarios de détection de valeurs aberrantes et s'adapte bien aux fluctuations des données de vols de voitures. Son utilisation dans le cadre de la surveillance des contrôles statistiques des processus (CSP) offre des opportunités intéressantes pour améliorer la détection et la gestion des risques dans le secteur des assurances.

Le choix entre le modèle Farrington et le modèle GLR dépend du contexte : le modèle Farrington est idéal pour les analyses épidémiologiques avec des tendances saisonnières, tandis que le GLR convient mieux aux analyses nécessitant une plus grande flexibilité statistique. Le modèle GLR est avantageux pour sa réactivité et sa flexibilité à intégrer plusieurs variables explicatives, ce qui le rend efficace pour capter des changements rapides et irréguliers dans les données. Cependant, cette complexité peut rendre son interprétation plus difficile. Le modèle Farrington est plus simple et est idéal pour détecter des motifs saisonniers réguliers ainsi que des augmentations progressives, étant ainsi plus adapté aux périodes prolongées. Le choix

entre ces deux modèles dépend des besoins spécifiques en termes de réactivité et de complexité de la saisonnalité des données analysées.

Les résultats obtenus avec ces modèles indiquent que les compagnies d'assurance peuvent tirer profit de leur mise en oeuvre pour suivre les tendances et les anomalies dans les statistiques de vols de voitures. En intégrant ces outils dans leurs systèmes de gestion des risques, elles peuvent améliorer leur réactivité face aux événements imprévus et élaborer des stratégies proactives pour minimiser les pertes. Les modèles offrent également une base solide pour l'analyse prédictive, permettant aux compagnies d'anticiper les évolutions futures et de mieux se préparer à des fluctuations potentielles dans les tendances de vols.

Pour une mise en oeuvre efficace, il est crucial de continuer à affiner les modèles en tenant compte des spécificités des données utilisées et des évolutions dans les tendances de vols. Une collaboration étroite entre les statisticiens, les analystes de données et les experts en assurance est nécessaire pour adapter et améliorer les modèles en fonction des besoins réels du marché. De plus, la formation continue des utilisateurs sur l'interprétation des résultats et l'application des modèles dans la prise de décision est essentielle pour maximiser les avantages de ces outils.

En résumé, ce mémoire démontre que l'application de modèles statistiques tels que le modèle Farrington et le modèle GLR dans le domaine des vols de voitures offre des perspectives novatrices pour la gestion des risques en assurance. Ces modèles, grâce à leur capacité à détecter des anomalies et à fournir des alertes précoces, constituent des outils précieux pour renforcer la surveillance et la résilience des compagnies face aux défis croissants de la criminalité liée aux véhicules. Leur utilisation représente une avancée significative vers une approche plus axée sur les données dans la gestion des assurances, avec des implications positives pour la réduction des pertes et l'amélioration de la sécurité des véhicules assurés.

ANNEXE A

STATISTIQUES DESCRIPTIVES DES COVARIABLES CHEVAUX, EMPAT ET TYPE

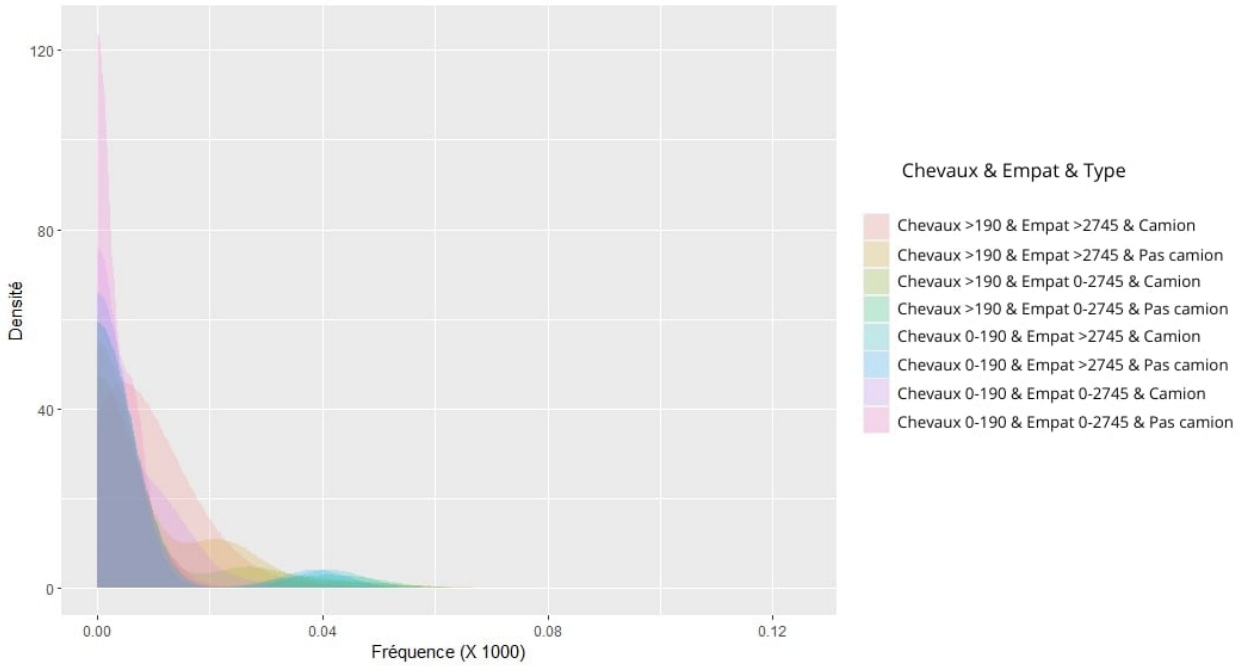


Figure A.1 – L'estimation par noyau de la fréquence par jour pour chacune des modalités des covariables Chevaux, Empat et Type

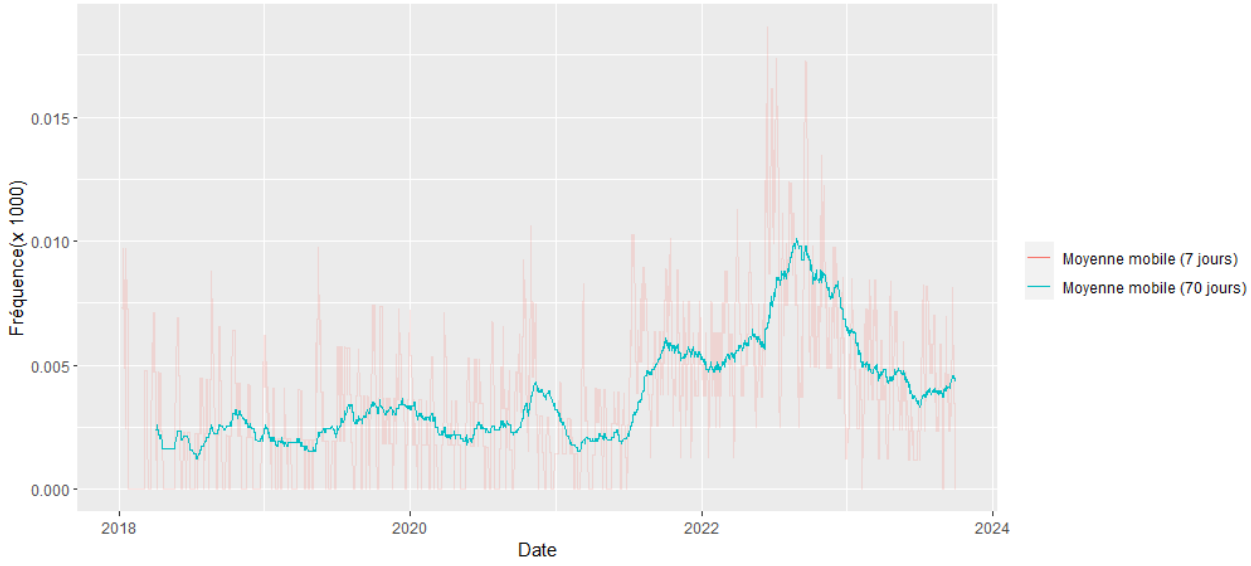


Figure A.2 – Fréquence de vols pour les assurés avec Chevaux ≤ 190, Empat ≤ 2745 et Type de type camion

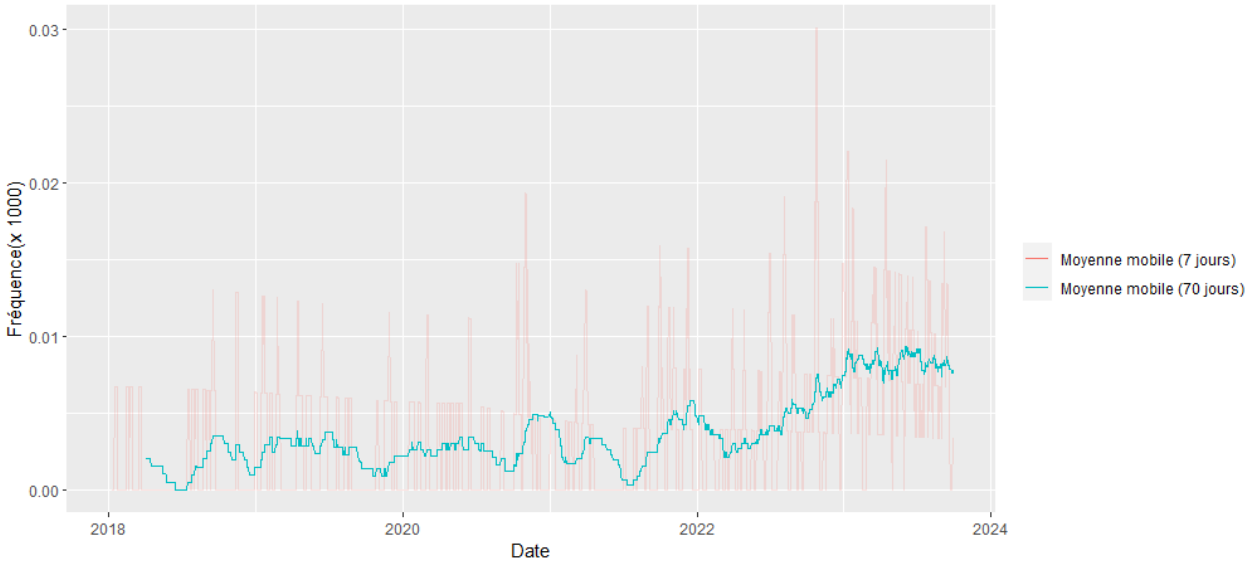


Figure A.3 – Fréquence de vols pour les assurés avec Chevaux > 190, Empat ≤ 2745 et Type de type camion

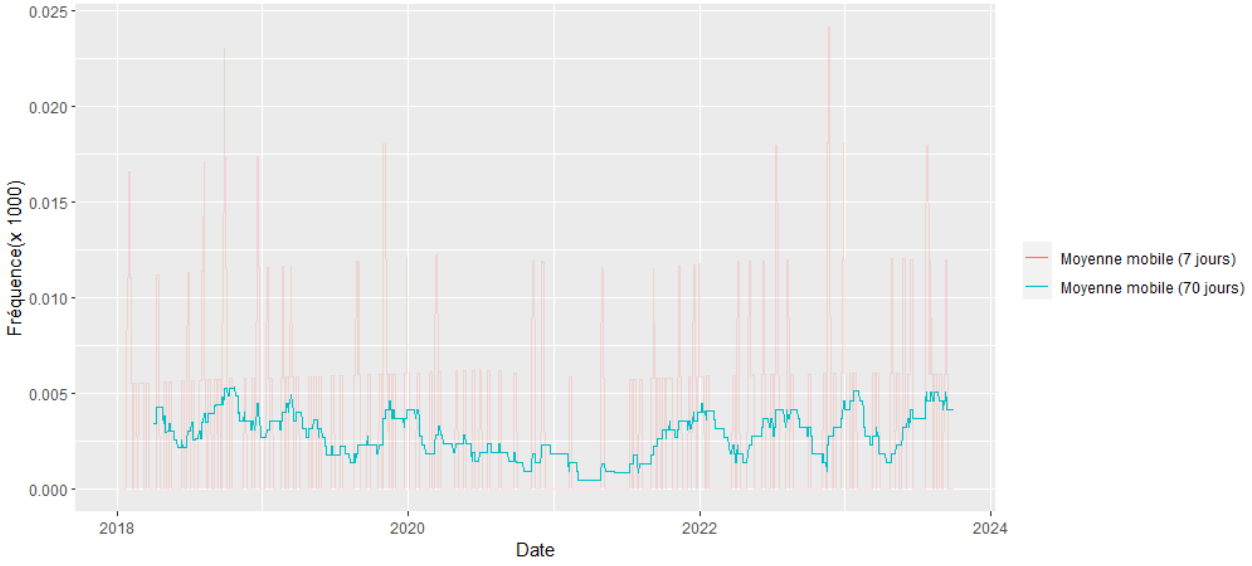


Figure A.4 – Fréquence de vols pour les assurés avec Chevaux ≤ 190, Empat > 2745 et Type de type camion

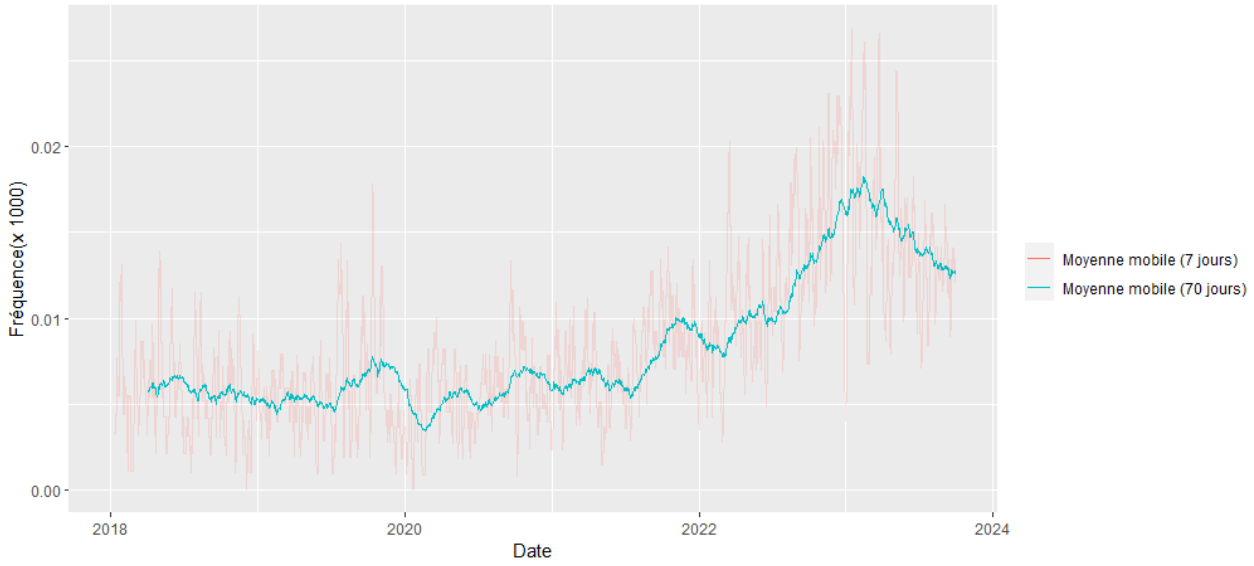


Figure A.5 – Fréquence de vols pour les assurés avec Chevaux > 190, Empat > 2745 et Type de type camion

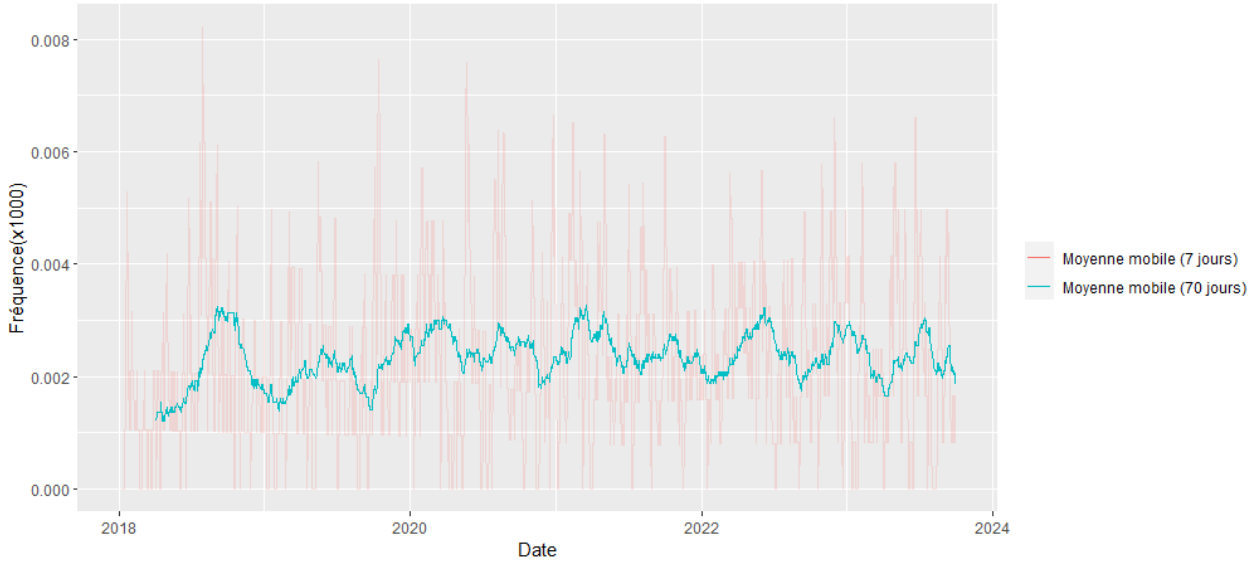


Figure A.6 – Fréquence de vols pour les assurés avec Chevaux ≤ 190, Empat ≤ 2745 et Type avec tous les véhicules autres que les camions

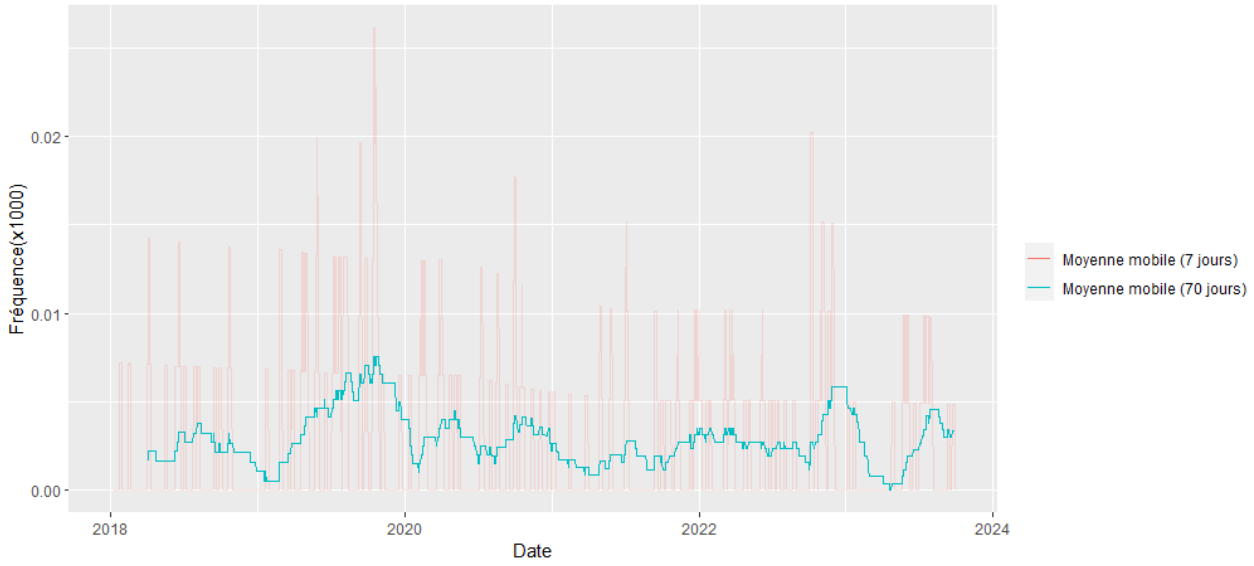


Figure A.7 – Fréquence de vols pour les assurés avec Chevaux > 190, Empat ≤ 2745 et Type avec tous les véhicules autres que les camions

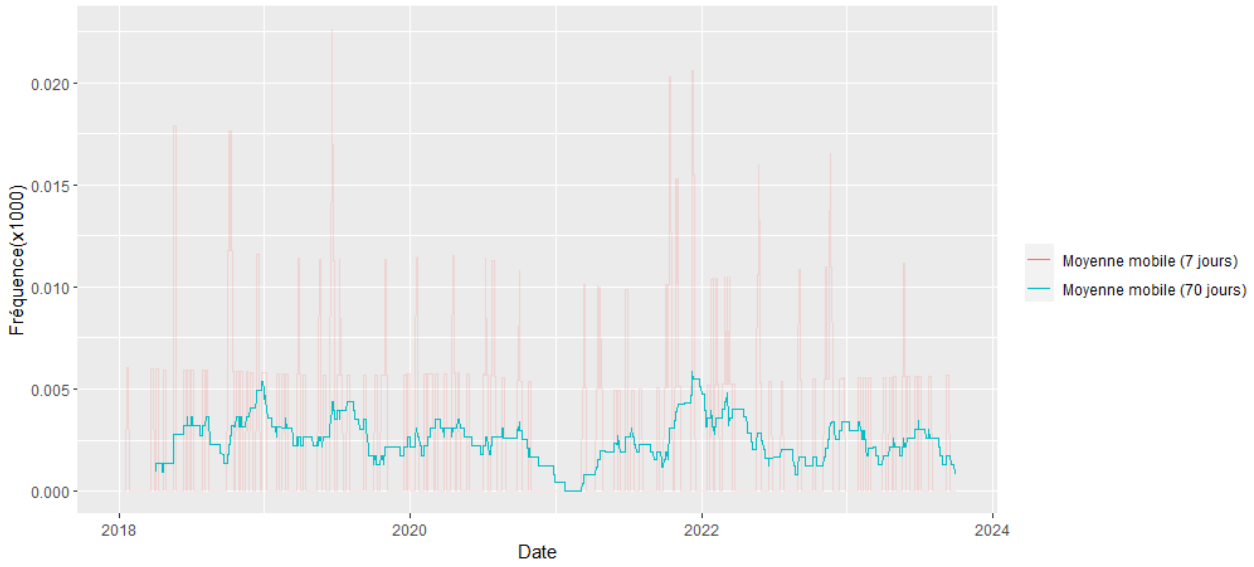


Figure A.8 – Fréquence de vols pour les assurés avec Chevaux ≤ 190, Empat > 2745 et Type avec tous les véhicules autres que les camions

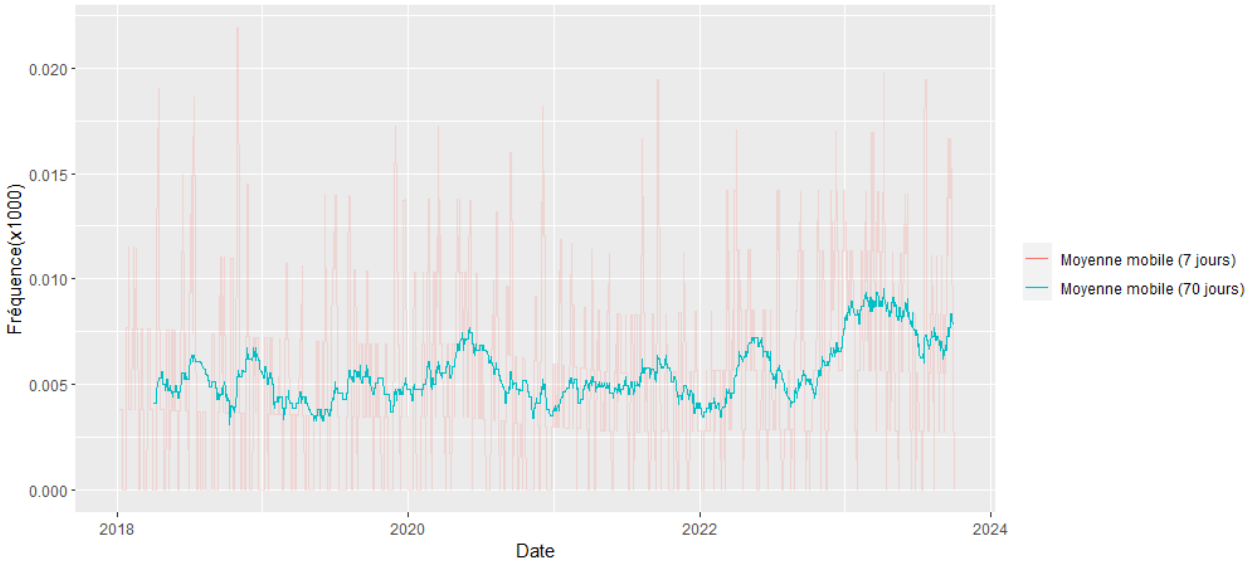


Figure A.9 – Fréquence de vols pour les assurés avec Chevaux > 190, Empat > 2745 et Type avec tous les véhicules autres que les camions

ANNEXE B

GLR INDÉPENDANTS POUR LES COVARIABLES CHEVAUX, EMPAT ET TYPE

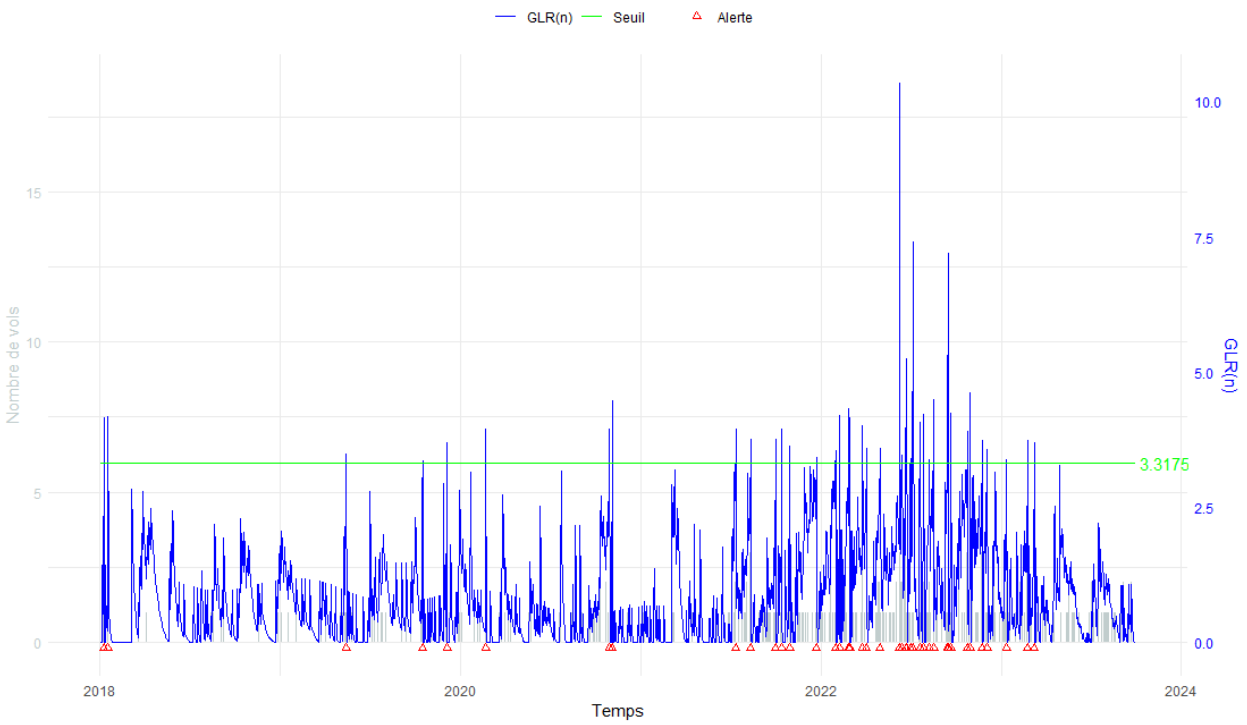


Figure B.1 – GLR pour les assurés avec Chevaux ≤ 190, Empat ≤ 2745 et Type de type camion

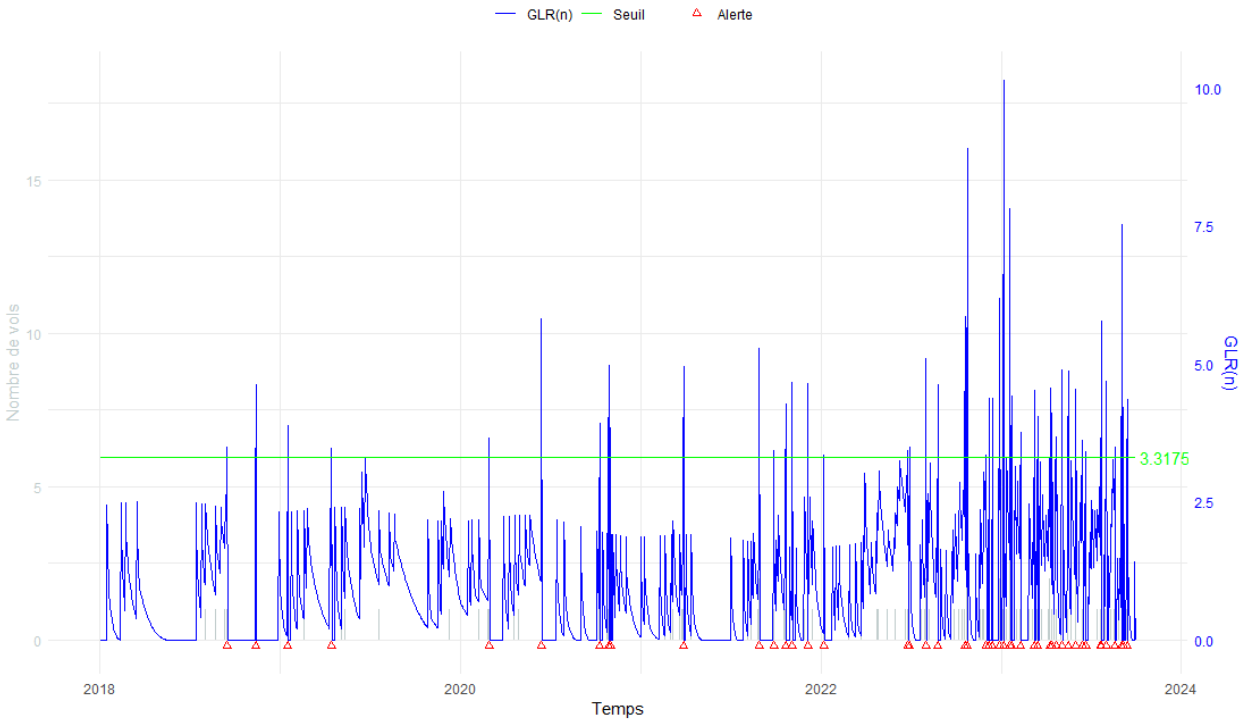


Figure B.2 – GLR pour les assurés avec Chevaux > 190, Empat ≤ 2745 et Type de type camion

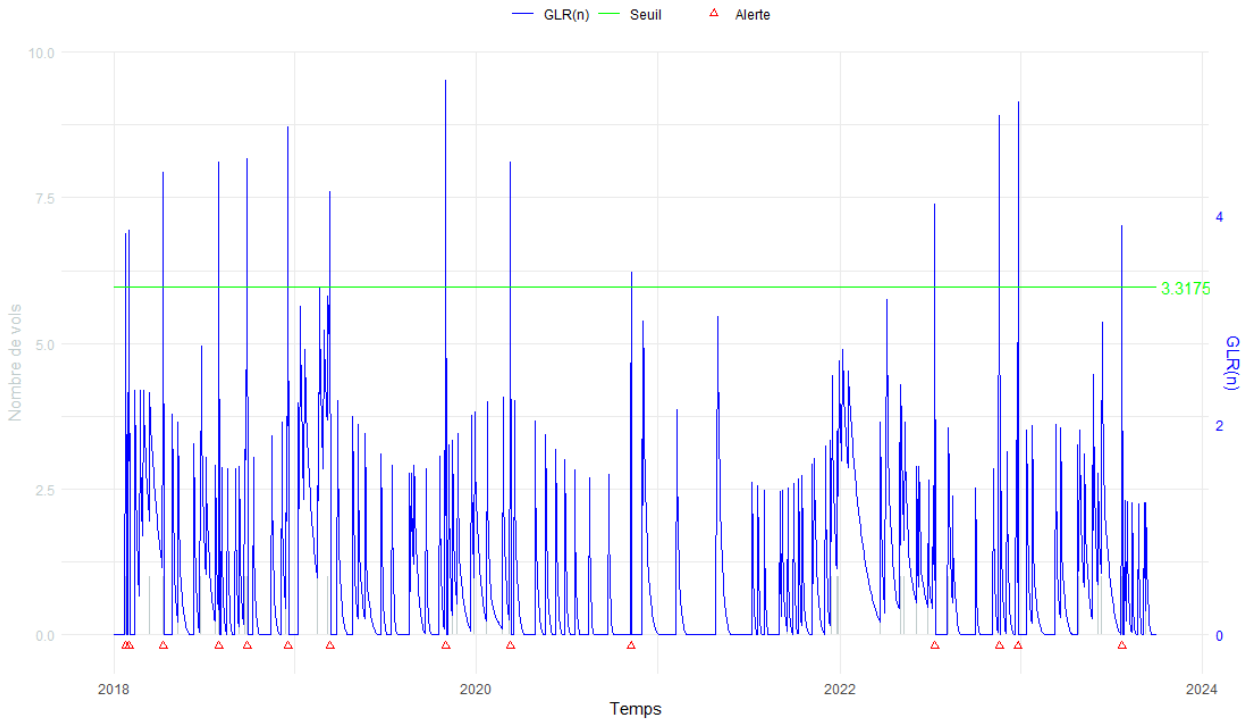


Figure B.3 – GLR pour les assurés avec Chevaux ≤ 190, Empat > 2745 et Type de type camion

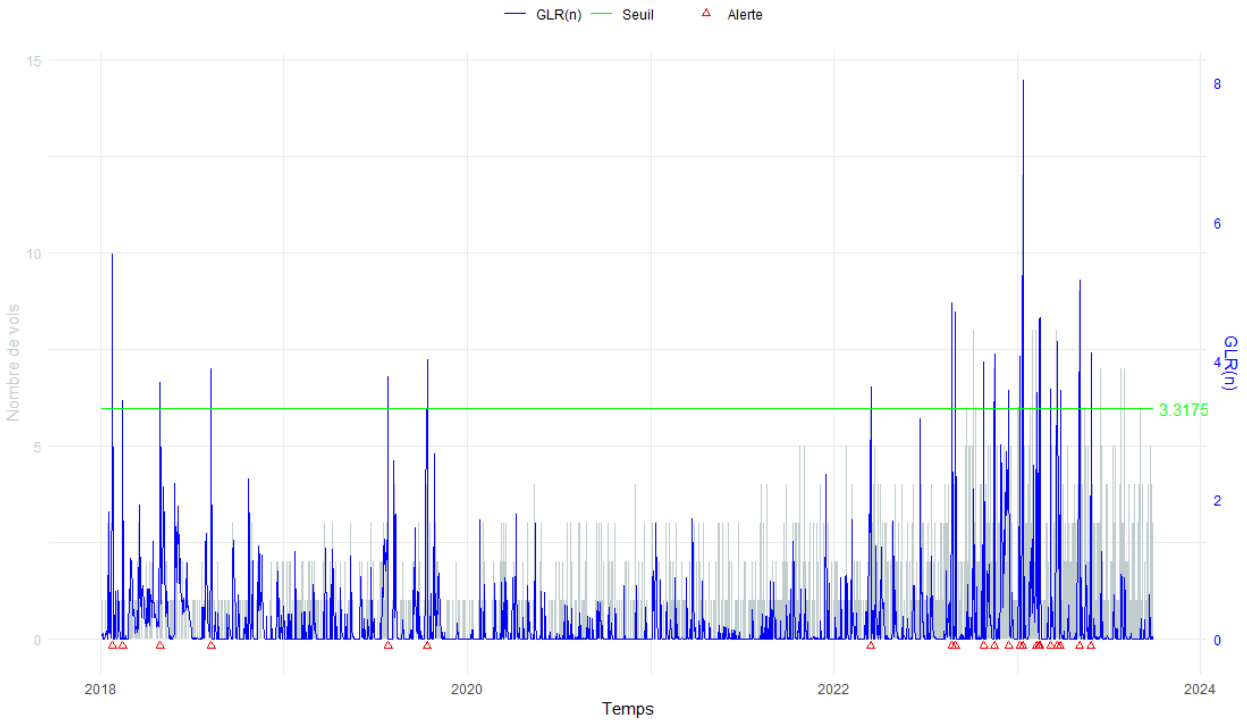


Figure B.4 – GLR pour les assurés avec Chevaux > 190, Empat > 2745 et Type de type camion

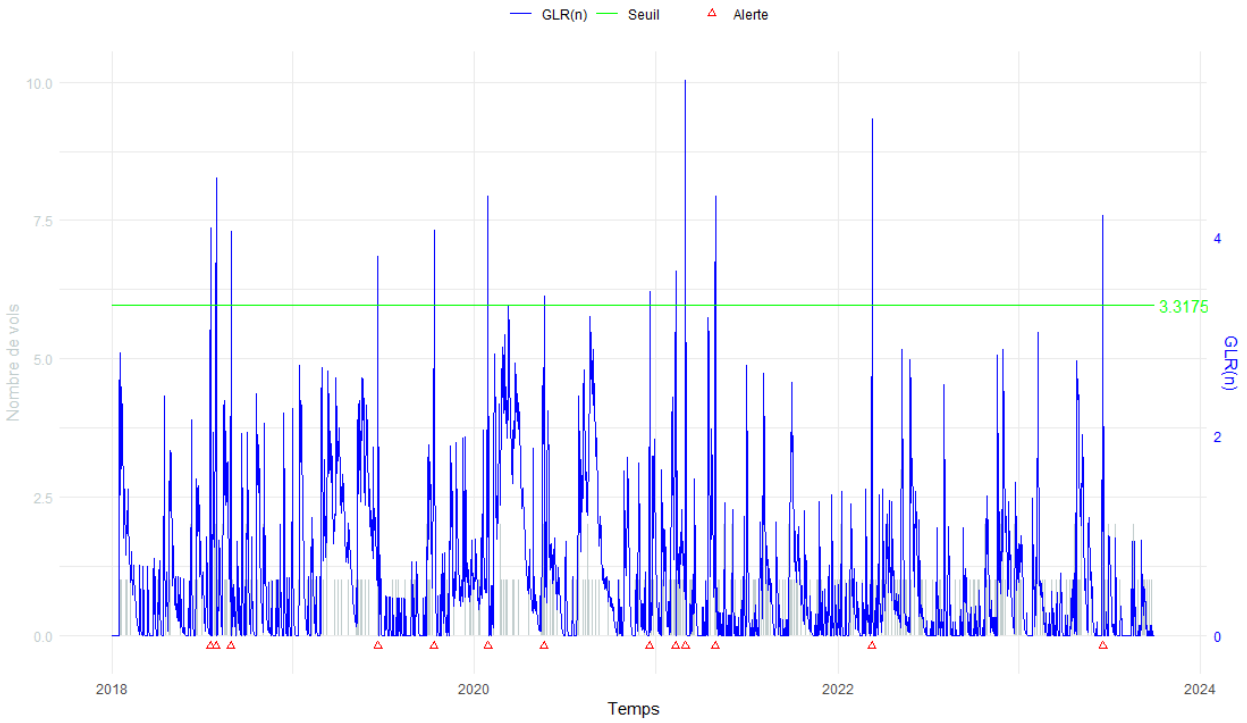


Figure B.5 – GLR pour les assurés avec Chevaux ≤ 190, Empat ≤ 2745 et Type de type autre que camion

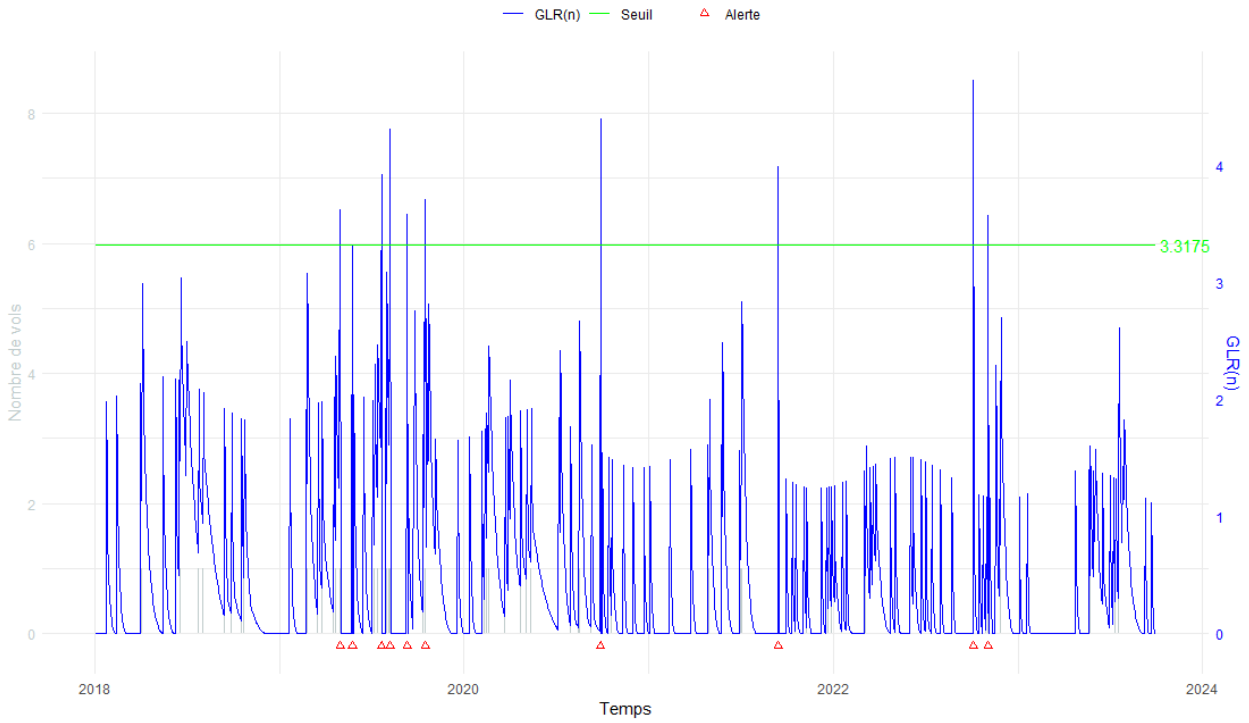


Figure B.6 – GLR pour les assurés avec Chevaux > 190, Empat ≤ 2745 et Type de type autre que camion

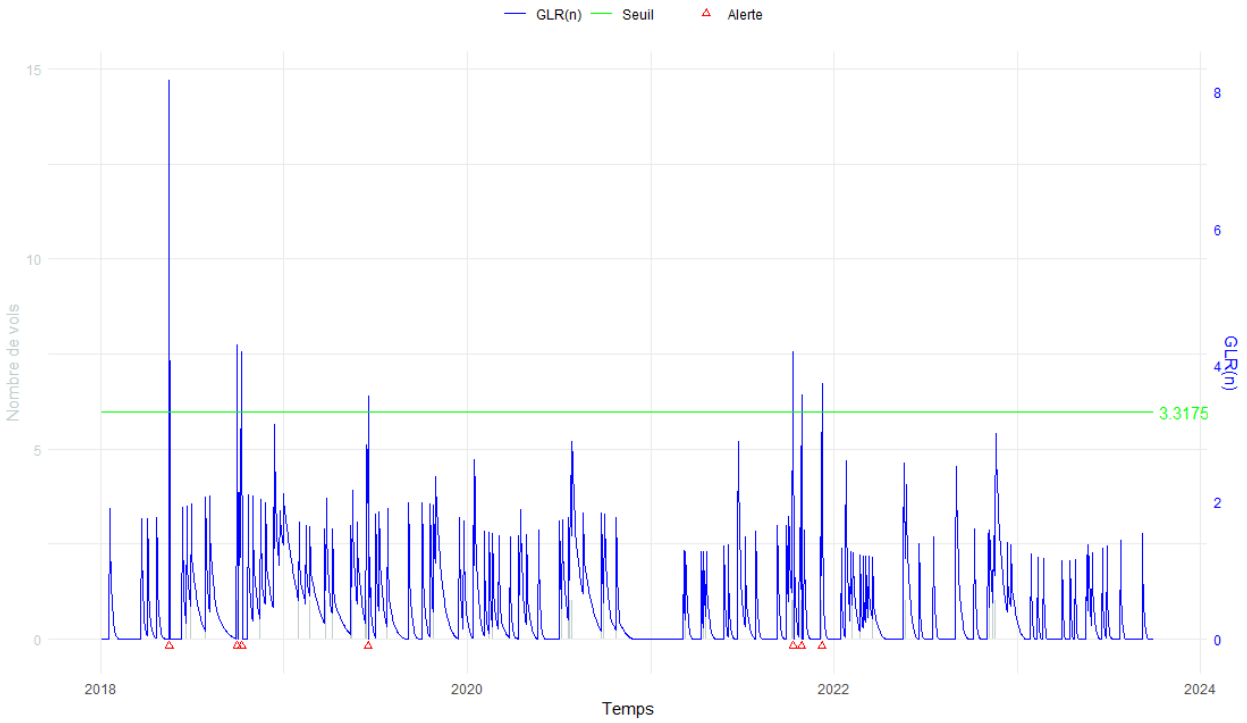


Figure B.7 – GLR pour les assurés avec Chevaux ≤ 190, Empat > 2745 et Type de type autre que camion

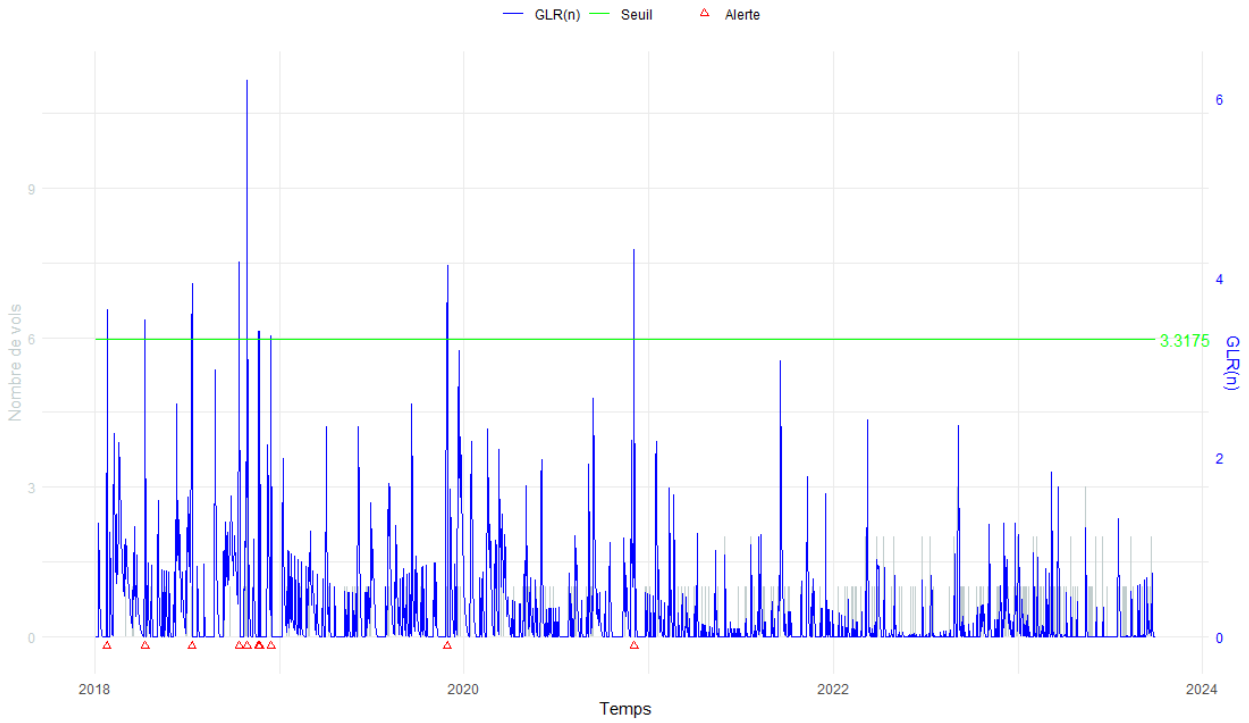


Figure B.8 – GLR pour les assurés avec Chevaux > 190, Empat > 2745 et Type de type autre que camion

BIBLIOGRAPHIE

Arcand, F. (2024). Ottawa « serre la vis ». *La Presse*. Récupéré le 2024-05-08 de https://www.lapresse.ca/actualites/politique/2024-05-20/plan-d-action-contre-le-vol-de-vehicules/ottawa-serre-la-vis.php

Ayres, I. et Levitt, S. D. (1998). Measuring Positive Externalities from Unobservable Victim Precaution : An Empirical Analysis of Lojack. *The Quarterly Journal of Economics*, *113*(1), 43–77. Récupéré de https://ideas.repec.org/a/oup/qjecon/v113y1998i1p43-77..html

Basseville, M. et Nikiforov, I. V. (1993). Detection of abrupt changes : theory and application. *Technometrics*, *36*, 550. Récupéré de https://api.semanticscholar.org/CorpusID:121394317

Braga, A. et Clarke, R. (2014). Explaining high-risk concentrations of crime in the city : Social disorganization, crime opportunities, and important next steps. *Journal of Research in Crime and Delinquency*, *51*, 480–498. http://dx.doi.org/10.1177/0022427814521217

Brown, R. L., Durbin, J. E. et Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the royal statistical society series b-methodological*, *37*, 149–163. Récupéré de https://api.semanticscholar.org/CorpusID:117391322

Casella, G. et Berger, R. (2024). *Statistical inference*. CRC Press.

Clarke, R. V. et Harris, P. M. (1992). Auto theft and its prevention. *Crime and Justice*, *16*, 1 – 54. Récupéré de https://api.semanticscholar.org/CorpusID:144083901

Davison, A. C. et Tsai, C.-L. (1992). Regression model diagnostics. *International Statistical Review / Revue Internationale de Statistique*, *60*(3), 337–353. Récupéré le 2024-04-16 de http://www.jstor.org/stable/1403682

de Jong, P. et Heller, G. (2008). *Generalized Linear Models for Insurance Data*. International Series on Actuarial Science. Cambridge University Press. Récupéré de https://books.google.ca/books?id=DW0syblRyHUC

Deheuvels, P. (1977). Estimation non paramétrique de la densité par histogrammes généralisés. *Revue de statistique appliquée*, *25*(3), 5–42.

Ekblom, P. (1997). Gearing up against crime : A dynamic framework to help designers keep up with the adaptive criminal in a changing world. *2*, 249–265.

Farrell, G., Tseloni, A. et Tilley, N. (2011). The effectiveness of car security devices and their role in the crime drop. *Criminology & Criminal Justice - CRIMINOL CRIM JUSTICE*, *11*, 21–35. http://dx.doi.org/10.1177/1748895810392190

Farrington, C. P., Andrews, N. J., Beale, A. D. et Catchpole, M. A. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *159*(3), 547–563. Récupéré le 2023-11-07 de

http://www.jstor.org/stable/2983331

Fonti, V. et Belitser, E. (2017). Feature selection using lasso. *VU Amsterdam research paper in business analytics*, *30*, 1–25.

Frisén, M. et Wessman, P. (1999). Evaluations of likelihood ratio methods for surveillance. differences and robustness. *Communications in Statistics - Simulation and Computation*, *28*, 597–622. Récupéré de https://api.semanticscholar.org/CorpusID:123156571

Genuer, R. et Poggi, J.-M. (2017). Arbres CART et Forêts aléatoires,Importance et sélection de variables. working paper or preprint

Gérard, J.-F. (2023). Voici le parcours emprunté par les véhicules volés au canada. *Radio-Canada*. Récupéré le 2024-07-30 de https://ici.radio-canada.ca/nouvelle/2008954/vol-voitures-volees-itineraire-parcours#:~:text=Environ%209500%20v%C3%A9hicules%20ont%20%C3%A9t%C3%A9%20vol%C3%A9s%20%C3%A0%20Toronto%20l'an%20dernier.&text=Les%20vols%20de%20voiture%20sont,Service%20de%20police%20de%20Toronto.

Höhle, M. et Paul, M. (2008). Count data regression charts for the monitoring of surveillance time series. *Computational Statistics & Data Analysis*, *52*(9), 4357–4368. http://dx.doi.org/https://doi.org/10.1016/j.csda.2008.02.015. Récupéré de https://www.sciencedirect.com/science/article/pii/S0167947308000716

Kim, H.-J. et Siegmund, D. (1989). The likelihood ratio test for a change-point in simple linear regression. *Biometrika*, *76*(3), 409–423. Récupéré le 2024-04-05 de http://www.jstor.org/stable/2336108

Kondofersky, I. et Theis, F. J. (2018). Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical Learning with Sparsity : The Lasso and Generalizations. Boca Raton : CRC Press. *Biometrics*, *74*(2), 769–769. http://dx.doi.org/10.1111/biom.12895. Récupéré de https://doi.org/10.1111/biom.12895

Lai, T. L. (1995). Sequential changepoint detection in quality control and dynamical systems. *Journal of the royal statistical society series b-methodological*, *57*, 613–644. Récupéré de https://api.semanticscholar.org/CorpusID:125569699

Lawson, A. B. et Kleinman, K. P. (2005). Introduction : Spatial and syndromic surveillance for public health. Récupéré de https://api.semanticscholar.org/CorpusID:58380475

Levi, M. et Maguire, M. (2004). Reducing and preventing organised crime : An evidence-based critique : Special issue : Evidence-based approaches to regulating "organized crime" (guest editor : Michael levi). *Crime*, *41*. http://dx.doi.org/10.1023/B:CRIS.0000039600.88691.af

Marquardt, D. W. et Snee, R. D. (1975). Ridge regression in practice. *The American Statistician*, *29*(1), 3–20.

Noufaily, A., Enki, D. G., Farrington, P., Garthwaite, P., Andrews, N. et Charlett, A. (2013). An improved algorithm for outbreak detection in multiple surveillance systems. *Statistics in medicine*, *32*(7), 1206—1222. http://dx.doi.org/10.1002/sim.5595. Récupéré de https://europepmc.org/articles/pmc3692796?pdf=render

Rogerson, P. A. et Yamada, I. (2004). Approaches to syndromic surveillance when data consist of small regional counts. *MMWR supplements*, *53*, 79–85. Récupéré de https://api.semanticscholar.org/CorpusID:24477337

Salmon, M., Schumacher, D. et Höhle, M. (2016). Monitoring count time series in r : Aberration detection in public health surveillance. *Journal of Statistical Software*, *70*(10), 1–35. http://dx.doi.org/10.18637/jss.v070.i10. Récupéré de https://www.jstatsoft.org/index.php/jss/article/view/v070i10

Skinner, K. R., Montgomery, D. C. et Runger, G. C. (2003). Process monitoring for multiple count data using generalized linear model-based control charts. *International Journal of Production Research*, *41*, 1167 – 1180. Récupéré de https://api.semanticscholar.org/CorpusID:121443074

Stroup, D. F., Williamson, G. D., Herndon, J. L. et Karon, J. M. (1989). Detection of aberrations in the occurrence of notifiable diseases surveillance data. *Statistics in Medicine*, *8*(3), 323–329. http://dx.doi.org/https://doi.org/10.1002/sim.4780080312. Récupéré de https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780080312

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. Récupéré le 2024-04-09 de http://www.jstor.org/stable/2346178

Tilley, N. et Laycock, G. (2018). Developing a knowledge base for crime prevention : Lessons learned from the british experience. *Crime Prevention and Community Safety*, *20*. http://dx.doi.org/10.1057/s41300-018-0053-8

Wagner, J. M. et Shimshak, D. G. (2007). Stepwise selection of variables in data envelopment analysis : Procedures and managerial perspectives. *European Journal of Operational Research*, *180*(1), 57–67. http://dx.doi.org/https://doi.org/10.1016/j.ejor.2006.02.048. Récupéré de https://www.sciencedirect.com/science/article/pii/S0377221706002839

Webb, B. et Laycock, G. (1992). Tackling car crime : The nature and extent of the problem.

Welsh, B. et Farrington, D. (2003). Effects of improved street lighting on crime protocol for a systematic review. *Campbell Systematic Reviews*, *1*. http://dx.doi.org/10.1002/CL2.14

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, *9*(1), 60–62. Récupéré le 2024-06-09 de http://www.jstor.org/stable/2957648

Wood, S. (2006). *Generalized Additive Models : An Introduction With R*, volume 66. http://dx.doi.org/10.1201/9781315370279