UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MODÉLISATIONS NON-PARAMÉTRIQUES DE LA FRÉQUENCE EN ASSURANCE AUTOMOBILE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

ACTUARIELLES ET FINANCIÈRES

PAR

GABRIEL ALEPIN

NOVEMBRE 2018

REMERCIEMENTS

J'ai découvert le monde de la recherche par l'entremise d'un stage avec le groupe de recherche Quantact après ma première année universitaire. Sous la supervision de Jean-Philippe Boucher, j'ai alors eu un avant-goût de ce qu'impliquerait la poursuite de mes études à la maîtrise. J'aimerais le remercier de m'avoir transmis cet intérêt pour la recherche, mais aussi pour ses conseils, son aide et son temps. À la suite de ce stage, la décision de réaliser une maîtrise est devenue très facile à prendre, tout comme celle du choix de mon directeur de recherche. Merci Jean-Philippe !

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

L'arrivée d'appareils télématiques permet de mettre en pratique de vieilles idées à propos d'une assurance automobile basée sur l'utilisation réelle du véhicule assuré, soit une assurance de type UBI (*Usage-based insurance*). À l'aide de données possédant le kilométrage exact parcouru, nous allons nous intéresser à la modélisation de la fréquence de réclamations par des modèles additifs généralisés avec paramètres de position, d'échelle et de forme (GAMLSS), poursuivant le travail de Boucher *et al.* (2017). Pour cela, nous allons utiliser des fonctions de lissage par P-splines pour le kilométrage et la durée d'exposition. Diverses lois discrètes de probabilité seront étudiées, dont la distribution binomiale négative multivariée (MVNB) à données longitudinales. Les impacts potentiels au niveau de la tarification seront également présentés par l'étude de différents profils d'assurés.

Mots-clés : modèle additif généralisé avec paramètres de position, d'échelle et de forme (GAMLSS), modèle additif généralisé (GAM), modèle linéaire généralisé (GLM), données longitudinales, données de panel, segmentation, modèle de fréquence, actuariat, tarification en assurance automobile, *Pay-as-you-drive* (PAYD), *Usage-based insurance* (UBI).

INTRODUCTION

Le marché de l'assurance automobile est énorme. Au Québec seulement, plus de 5 millions de véhicules ont été assurés en 2016, pour un total des primes d'assurance de plus de 3,3 milliards de dollars [1].

Traditionnellement, les assureurs établissent la prime d'un assuré en fonction du risque qu'il représente. Il est toutefois difficile de mesurer précisément le degré de ce risque. Les assureurs s'en remettent alors à des caractéristiques de risque, soit des variables qu'ils considèrent corrélées avec le vrai niveau de risque de l'assuré selon leurs analyses, telles que le sexe, l'âge, la cote de crédit, le kilométrage annuel, le lieu de résidence et le modèle de véhicule pour n'en nommer que quelques-unes. De plus, la durée du contrat d'assurance sert typiquement d'unité d'exposition et est considérée proportionnelle au risque. Cela signifie que, pour un même profil de risque, une assurance de 6 mois est 2 fois moins coûteuse qu'une même assurance de 1 an.

Cette façon de procéder possède comme défaut qu'elle n'est pas basée sur l'usage réel du véhicule. L'assuré doit fournir une estimation de son kilométrage annuel au début du contrat, mais l'assureur ne peut que difficilement valider cette information, si bien que les estimations sont souvent fausses. De plus, les assurés ont avantage à sous-estimer leur kilométrage, car, de façon générale, plus un assuré parcourt de kilomètres, plus sa prime augmente.

---

1. Rapport sommaire du Groupement des assureurs automobiles (GAA) : https ://gaa.qc.ca/Pdf/fr/Rapport-Sommaire-FR-2017.pdf

La facilité actuelle à installer des appareils télématiques sur les véhicules assurés ou, plus simplement, à utiliser les données fournies directement par le téléphone portable de l'assuré permet de remettre d'actualité de vieilles idées à propos d'une assurance automobile basée sur l'usage du véhicule assuré, nommée UBI pour *Usage-based insurance*, et non uniquement sur la durée du contrat.

On retrouve d'ailleurs avec Vickrey (1968) les balbutiements d'une théorie relative à une assurance automobile de type UBI. Vickrey suggère une prime basée sur le kilométrage pour mieux refléter l'utilisation des voitures assurées. Une façon de mesurer le kilométrage passerait par une vérification annuelle de l'odomètre du véhicule assuré, mais la facilité de l'époque à « reculer » les odomètres posait problème. Il était donc difficile de mettre en pratique une telle assurance avec les technologies du moment. Comme alternatives, il suggère alors d'inclure des frais d'assurance sur l'achat de pneus ou d'essence, mais ces méthodes ne sont pas sans défaut.

Aujourd'hui, les produits d'assurance automobile de type UBI se multiplient. Leur popularité grandissante ouvre des possibilités en recherche concernant le traitement et l'utilisation des nouvelles données que les appareils télématiques permettent de collecter. On discerne 2 grandes catégories de produits UBI, soit PAYD[2] et PHYD[3] (Tselentis *et al.*, 2016). La catégorie PAYD se concentre sur le kilométrage parcouru, le type de routes emprunté, comme une autoroute, une route de campagne ou un boulevard, et sur les heures de conduites notamment, soit sur les trajets de l'assuré. L'autre catégorie, soit PHYD, est davantage orientée sur la façon de conduire de l'assuré et peut mesurer les excès de vitesse ainsi que les accélérations et les freinages brusques. Évidemment, un produit d'assurance

2. *Pay-as-you-drive.*

3. *Pay-how-you-drive.*

peut également être basé sur un mélange entre les catégories PAYD et PHYD.

Les produits de type PAYD sont davantage étudiés en recherche. Ils incluent notamment les produits de type *Per-Mile-Premium* (PMP) et utilisent le kilométrage parcouru comme unité d'exposition plutôt que la durée d'exposition(Litman, 2011, 2005, 2001). À ce sujet, de nombreux articles notent une relation non proportionnelle entre la fréquence de réclamations et le kilométrage (Boucher *et al.*, 2017, 2013; Litman, 2011; Bordoff et Noel, 2008).

Une tarification PAYD basée sur le kilométrage mesuré par un appareil télématique possède de nombreux avantages (Lemaire *et al.*, 2016; Tselentis *et al.*, 2016; Bordoff et Noel, 2008; Litman, 2005).

Parmi ceux-ci, on note une amélioration de la justesse actuarielle de la prime, puisque la tarification classique ne peut pas tenir compte du kilométrage de façon adéquate faute de données fiables. Ainsi, la prime chargée reflète mieux le risque représenté par chaque assuré. Une autre conséquence est que les assurés ont davantage de contrôle sur leur prime puisqu'elle est directement liée à la distance qu'ils parcourent.

De plus, on constate une réduction de la fraude sur le kilométrage déclaré, car le kilométrage est désormais mesuré précisément plutôt qu'estimé par l'assuré.

En procédant de cette façon, on rend également l'assurance plus abordable pour les personnes parcourant peu de kilomètres en voiture, ce qui pourrait réduire la part de conducteurs non assurés.

Une telle tarification ajoute aussi un incitatif à moins conduire, ce qui mène à une réduction potentielle du nombre de voitures sur les routes et, possiblement, du nombre d'accidents. D'ailleurs, s'il y a moins de voitures sur les routes, on peut aussi s'attendre à une diminution de la congestion routière et de la pollution

générée par l'automobile. Avec un produit de type PHYD, l'incitatif va même plus loin en encourageant les assurés à mieux conduire, ce qui devrait aussi diminuer le nombre d'accidents et améliorer le bilan routier.

Les appareils télématiques pourraient même servir à communiquer automatiquement avec les services d'urgence en cas d'accident ou à retrouver des véhicules volés.

Évidemment, une tarification PAYD comprend également des désavantages (Lemaire *et al.*, 2016). Tout d'abord, il y a le coût et l'installation de l'appareil télématique à considérer. De plus, les primes sont davantage imprévisibles puisque le kilométrage annuel est souvent variable d'une année à l'autre. L'aspect légal est aussi à considérer, puisque certaines juridictions n'ont toujours pas adapté leurs lois sur l'assurance aux produits PAYD. Par exemple, une surcharge à l'assuré pour kilométrage excessif ou une tarification rétrospective peuvent être illégales dans certaines juridictions.

Il faut aussi considérer le côté intrusif de ces appareils qui collectent des données sur le comportement des assurés. Ces données doivent être sécurisées par respect pour la protection de la vie privée. Des questions éthiques sont d'ailleurs à être considérées, notamment à savoir si ces données peuvent être partagées ou non en cas d'enquête policière.

Un autre point dont il faut tenir compte est celui du transfert des coûts (Litman, 2005). Avec une tarification classique où il est impossible de correctement tenir compte du kilométrage, les assurés parcourant de courtes distances se retrouvent à payer une prime plus élevée que le risque qu'ils représentent et, donc, à « financer » ceux qui parcourent de longues distances. Ainsi, les conducteurs à faible kilométrage devraient opter pour une assurance PAYD basée sur le kilométrage parcouru pour économiser. S'ils le font, la prime provenant d'une tarification

classique devrait augmenter pour éviter la sélection adverse, puisque ceux qui la finançaient ont changé de produit d'assurance. Par conséquent, si la tarification classique devient plus dispendieuse, la tarification PAYD devrait gagner encore plus en popularité, rendant la prime issue d'une tarification classique toujours plus chère.

Dans ce mémoire, nous allons nous intéresser à la modélisation de la fréquence de réclamations en assurance automobile avec un produit de type PAYD. Nous allons d'abord, au chapitre 1, étudier les modèles connus et appliqués en assurance, soit les modèles linéaires généralisés (GLM) et les modèles additifs généralisés (GAM). Puis, au chapitre 2, nous allons introduire les modèles additifs généralisés avec paramètres de position, d'échelle et de forme (GAMLSS). Finalement, au chapitre 3, nous appliquerons des modèles GAMLSS sur des données d'assurance de type PAYD en prenant le soin de les comparer avec des modèles plus classiques en assurance automobile.

CHAPITRE I

INTRODUCTION AUX MODÈLES

1.1 Quelques définitions préalables

En assurance, la prime pure n'est que le produit de la fréquence et de la sévérité. En considérant la prime pure comme une variable aléatoire, on peut alors utiliser des modèles mathématiques et tenir compte d'une certaine part de hasard dans les primes pures futures.

Pour évaluer le montant à charger à un assuré, l'objectif est de trouver une constante c qui s'approche le plus possible de la variable aléatoire de la prime pure S. Dans ce mémoire, la distance $d_2(S, c) = E\left[(S - c)^2\right]$ a été retenue. La valeur de la constante c qui minimise cette distance correspond à l'espérance de la prime pure, c'est-à-dire $c = E[S]$. Si l'on considère la fréquence et la sévérité comme des variables aléatoires indépendantes, on obtient alors :

$$E[S] = E[\text{Prime Pure}] = E[\text{Fréquence}] \times E[\text{Sévérité}].$$

Avec le résultat ci-dessus, on constate que l'évaluation de la prime à charger à un assuré peut se faire en traitant les composantes de la fréquence et de la sévérité séparément.

Dans ce mémoire, on se concentrera uniquement sur la modélisation de la fréquence. Ceci revient à considérer une sévérité fixe de 1$ pour toutes les réclama-

tions dans l'évaluation de la prime.

1.2 Modélisation de la fréquence de réclamations

Il y a plusieurs possibilités pour la modélisation de la fréquence de réclamations. Les modèles linéaires généralisés, ou GLM, sont considérés comme la norme dans l'industrie, mais récemment, les modèles additifs généralisés, ou GAM, gagnent en popularité. Toutefois, avant de les présenter en profondeur, revenons à la base avec le modèle de régression linéaire.

1.2.1 Régression linéaire

Supposons que nous désirons modéliser la fréquence de réclamations Y et que nous avons n observations de cette variable aléatoire, que nous noterons y_i où $i = 1, ..., n$. Sous le modèle de régression linéaire, on pose des hypothèses d'indépendance et d'homoscédasticité pour la variable réponse. Dans notre contexte, on peut également supposer que nous connaissons k variables explicatives X, c'est-à-dire des caractéristiques des assurés qui peuvent contribuer à « expliquer » leur fréquence de réclamation. Un tel modèle peut s'exprimer comme ceci :

$$Y_i = \beta_0 + x_{i1} \times \beta_1 + x_{i2} \times \beta_2 + ... + x_{ik} \times \beta_k + \epsilon_i, \tag{1.1}$$

où β_0 représente l'ordonnée à l'origine, x_{ij} correspond à la j^{e} caractéristique de risque connue de l'assuré i et β_j est le coefficient relié à la j^{e} caractéristique de risque. Dans ce modèle, ϵ_i correspond à l'erreur d'estimation et suit une loi Normale$(0, \sigma^2)$.

L'équation (1.1) peut être réécrite sous la forme matricielle

$$Y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_i,$$

où $\boldsymbol{X}$ est une matrice de design de dimension $n \times (k+1)$ contenant les variables

explicatives, dont une variable constante valant 1 pour tenir compte de l'ordonnée à l'origine, $\boldsymbol{X}_i$ est la ligne de cette matrice correspondant à l'assuré i et $\boldsymbol{\beta}$ est un vecteur de paramètres $(k+1) \times 1$.

Ceci revient à dire que la variable réponse suit une distribution normale, soit $Y_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2)$ pour $i = 1, ..., n$.

Pour modéliser le nombre de réclamations, qui est une variable discrète positive, la régression linéaire présente donc un inconvénient majeur, c'est-à-dire que la distribution normale est une distribution continue pouvant prendre des valeurs négatives.

1.2.2 Modèles linéaires généralisés

Les modèles linéaires généralisés, ou GLM [1], permettent de contrer cet inconvénient. En effet, avec ces modèles, développés par Nelder et Wedderburn (1972), la distribution de la variable réponse s'élargit à la famille exponentielle linéaire.

Cette famille regroupe l'ensemble des distributions, discrètes ou continues, dont les lois de probabilité peuvent s'écrire sous la forme :

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

où θ et ϕ sont respectivement appelés les paramètres canonique et de dispersion. De plus, $c(\cdot)$ est une fonction quelconque qui dépend de y et ϕ et $a(\cdot)$ est une autre fonction quelconque, mais dépendant uniquement de θ.

L'espérance et la variance des membres de cette famille de distributions peuvent

1. *Generalized linear models.*

être exprimées sous les formes générales

$$E[Y] = a'(\theta),$$
$$Var[Y] = \phi a''(\theta),$$

où $a'(\theta)$ et $a''(\theta)$ sont les dérivées première et seconde de $a(\theta)$ par rapport à θ. Selon le choix de la distribution de la variable réponse, il est donc possible de ne pas rencontrer de l'hétéroscédasticité, contrairement aux modèles de régression linéaire.

Parmi les distributions faisant partie de cette famille, il y a notamment les lois normale, Poisson, binomiale, gamma et inverse-gaussienne. On peut se référer à McCullagh et Nelder (1989) pour plus de détails.

Les GLM permettent également de spécifier une fonction de lien $g(\cdot)$ entre la moyenne de la distribution de la variable réponse μ et les variables explicatives, soit

$$g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta}, \tag{1.2}$$

où $\boldsymbol{X}_i$ est la ligne correspondant à l'assuré i de la matrice des variables explicatives et $\boldsymbol{\beta}$ est le vecteur des coefficients.

La fonction de lien $g(\cdot)$ est habituellement sélectionnée de façon à avoir un domaine correspondant à celui du paramètre de moyenne μ. Par exemple, pour une distribution Bernoulli, le paramètre de moyenne est compris entre 0 et 1. Ainsi, une fonction de lien probit ou logit pourrait être utilisée, car le domaine correspond à celui du paramètre de moyenne. D'autres facteurs peuvent aussi être considérés dans la sélection de la fonction de lien. Dans un contexte d'assurance, la fonction de lien logarithmique permet d'exprimer le paramètre de moyenne par

des facteurs multiplicatifs, qu'on appelle aussi relativités (McCullagh et Nelder, 1989).

Il est à noter que, lorsque qu'on sélectionne la distribution normale pour la variable réponse avec la fonction de lien identité, on retrouve exactement le modèle de régression linéaire.

Pour plus de détails sur la famille exponentielle linéaire, les fonctions de lien, les GLM et leur application en assurance, McCullagh et Nelder (1989) et De Jong *et al.* (2008) constituent d'excellentes références.

1.2.3 Modèles additifs généralisés

Les modèles additifs généralisés, ou GAM[2], constituent une extension des GLM. Ces modèles, développés par Hastie et Tibshirani (1986), se distinguent des GLM uniquement par le prédicteur linéaire qui peut désormais accueillir des fonctions de lissage, soit des termes non paramétriques. Par exemple, la structure d'un GAM en notation vectorielle pourrait s'écrire comme ceci :

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + f_1(\boldsymbol{x}_1) + f_2(\boldsymbol{x}_2) + f_3(\boldsymbol{x}_3, \boldsymbol{x}_4), \tag{1.3}$$

où les fonctions $f_1(\cdot)$, $f_2(\cdot)$ et $f_3(\cdot)$ sont des fonctions de lissage qui dépendent d'une ou plusieurs variables explicatives $\boldsymbol{x}_i$. Mis à part ces fonctions, l'équation (1.3) est en tout point identique à l'équation (1.2), c'est-à-dire que $\boldsymbol{X}$ est une matrice de design, $\boldsymbol{\beta}$ est un vecteur de paramètres et le prédicteur linéaire est lié à la moyenne $\boldsymbol{\mu}$ à travers une fonction de lien $g(\cdot)$. Pour davantage d'informations sur les GAM et leurs fonctions de lissage, on peut se référer à Boucher *et al.* (2017), Côté (2016) et Wood (2006).

---

2. *Generalized additive models.*

1.3 Application des modèles additifs généralisés

Boucher *et al.* (2017) ont présenté une application des GAM pour la modélisation de la fréquence de réclamations en assurance automobile. Ils avaient à leur disposition une base de données d'un assureur espagnol [3] provenant d'un produit d'assurance automobile de type PAYD. Ils avaient ainsi accès au kilométrage exact parcouru par chaque assuré.

De façon traditionnelle, les assureurs ont recours à des modèles GLM et le kilométrage fait partie des caractéristiques de risque contenues dans la matrice $\boldsymbol{X}$ de l'équation (1.2). Habituellement, le kilométrage est transformé en variables binaires, ce qui crée différentes classes. Il y a 2 raisons principales justifiant cela. Tout d'abord, sans appareil télématique, il est impossible pour les assureurs de connaître le véritable kilométrage des assurés, ce qui rend les données à ce sujet peu fiables. En second lieu, plusieurs recherches parviennent à la conclusion que la fréquence de réclamations est non proportionnelle au kilométrage (Boucher *et al.*, 2017, 2013; Litman, 2011; Bordoff et Noel, 2008). Ainsi, utiliser le kilométrage sans transformation revient à présumer qu'il suit une relation linéaire avec la fréquence, ce qui est faux.

L'utilisation d'appareils télématiques permet d'obtenir le kilométrage exact des assurés. Avec un modèle GLM où le kilométrage est transformé en variables binaires, il faudrait utiliser une multitude de variables binaires afin de mieux décrire la relation entre le kilométrage et le nombre de réclamations. Une autre option est l'utilisation d'un modèle GAM comprenant une fonction de lissage pour le kilométrage pour tenir compte de sa non-linéarité avec la fréquence.

C'est sous ce contexte que Boucher *et al.* (2017) ont choisi d'utiliser des modèles

3. Il s'agit de la même base de données utilisée au chapitre 3.

GAM avec une fonction de lissage pour cette variable.

Par ailleurs, l'approche traditionnelle en assurance prétend que la fréquence de réclamations est proportionnelle à la durée d'exposition. En d'autres mots, cela signifie que, pour un même risque, une assurance d'un an est 2 fois plus coûteuse que la même assurance de 6 mois.

Afin de confirmer cette hypothèse, ils ont également décidé d'appliquer une seconde fonction de lissage sur la durée d'exposition.

D'abord, ils se sont intéressés à l'utilisation de splines cubiques individuelles pour le kilométrage et la durée d'exposition pour un GAM suivant une distribution Poisson avec un lien logarithmique et une ordonnée à l'origine :

$$\log(\mu_i) = \beta_0 + f_1(km_i) + f_2(d_i), \qquad i = 1, ..., n, \tag{1.4}$$

où β_0 correspond à l'ordonnée à l'origine, n représente le nombre total d'assurés et $f_1(\cdot)$ et $f_2(\cdot)$ correspondent respectivement aux splines cubiques pour le kilométrage km et la durée d'exposition d.

Ils ont constaté que les fonctions de lissage, autant pour le kilométrage que pour la durée, ne sont clairement pas linéaires. Leur résultat pour la durée contredit l'approche traditionnelle en assurance voulant que la relation entre le nombre de réclamations et la durée d'exposition soit proportionnelle.

Puis, ils ont fait un second modèle en utilisant plutôt un produit tensoriel, un lissage à plusieurs dimensions, pour tenir compte de l'interaction entre le kilométrage et la durée d'exposition :

$$\log(\mu_i) = \beta_0 + f(km_i, d_i), \qquad i = 1, ..., n, \tag{1.5}$$

où $f(.,.)$ correspond au produit tensoriel prenant comme arguments le kilométrage

et la durée d'exposition.

Le modèle avec produit tensoriel avait pour objectif de visualiser l'effet combiné du kilométrage et de la durée, à savoir si l'effet de la durée varie en fonction du kilométrage et vice versa. Il s'agit d'une façon de vérifier si le premier modèle tient compte adéquatement de la dépendance entre le kilométrage et la durée.

Finalement, ils ont comparé ces 2 modèles avec un GLM Poisson avec une fonction de lien logarithmique ayant 5 variables binaires pour le kilométrage et un ajustement de proportionnalité pour la durée d'exposition :

$$\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \log(d_i), \quad i = 1, ..., n, \tag{1.6}$$

où β_0 à β_5 sont des paramètres. Les valeurs des variables x_1 à x_5 sont présentées dans le tableau 1.1. Il s'agit d'un modèle plus classique en assurance permettant d'évaluer l'amélioration que procurent les modèles (1.4) et (1.5).

Tableau 1.1: Représentation du kilométrage en variables binaires

Variable	Valeur
x_1	Vaut 1 si $km \geq 1000$
x_2	Vaut 1 si $5\,000 < km \leq 10\,000$
x_3	Vaut 1 si $10\,000 < km \leq 15\,000$
x_4	Vaut 1 si $15\,000 < km \leq 20\,000$
x_5	Vaut 1 si $km > 20\,000$

Concernant la modélisation des 2 GAM, Boucher *et al.* ont fixé le nombre de noeuds pour le kilométrage et la durée, puis ont estimé les paramètres de lissage par la minimisation du score de validation croisée généralisé[4] (GCV).

4. *Generalized cross validation score*, qui sera défini formellement à la section 2.3.2.

Le GCV a également servi de critère pour comparer la qualité des 3 modèles entre eux. Ainsi, selon le GCV, le GAM avec produit tensoriel possède un très léger avantage sur le GAM à splines cubiques individuelles alors que le GLM est en retrait de ces 2 modèles.

Tableau 1.2: Comparaison du critère GCV pour les 3 modèles de Boucher *et al.* (2017)

Modèle	GCV
GAM à splines cubiques	0,38412
GAM avec produit tensoriel	0,38403
GLM	0,38640

De plus, Boucher *et al.* ont réalisé une analyse des résidus de prédiction en se servant de données de validation. Ils ont constaté que les résultats des 3 modèles étaient très semblables.

Par la suite, ils ont ajouté des variables explicatives à leurs modèles. Pour les profils d'individus étudiés, ils ont remarqué que les modèles GAM (1.4) et (1.5) génèrent des primes pures semblables, mais que ces primes sont différentes du modèle GLM (1.6).

Finalement, Boucher *et al.* ont présenté un exemple de structure tarifaire pouvant être réalisée à partir d'un GAM.

Ce mémoire va poursuivre le travail fait par Boucher *et al.* en s'intéressant aux modèles additifs généralisés avec paramètres de position, d'échelle et de forme (GAMLSS) avec le même ensemble de données. Pour un autre exemple d'application de modèles GAM sur des données d'assurance automobile, voir Verbelen *et al.* (2016).

CHAPITRE II

INTRODUCTION AUX MODÈLES ADDITIFS GÉNÉRALISÉS AVEC PARAMÈTRES DE POSITION, D'ÉCHELLE ET DE FORME

Les modèles additifs généralisés avec paramètres de position, d'échelle et de forme, ou GAMLSS [1], constituent une généralisation des GAM. Proposés par Rigby et Stasinopoulos (2005), les GAMLSS assouplissent l'hypothèse de la distribution de la variable réponse en ne la limitant plus à la famille exponentielle linéaire. De plus, les GAMLSS permettent la modélisation des paramètres de variance σ, d'asymétrie ν et d'aplatissement τ en plus de celle du paramètre de la moyenne μ par l'entremise des fonctions de lien $g_1(\cdot)$ à $g_4(\cdot)$:

$$g_1(\boldsymbol{\mu}) = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \boldsymbol{Z}_{j1}\boldsymbol{\gamma}_{j1}, \tag{2.1}$$

$$g_2(\boldsymbol{\sigma}) = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \boldsymbol{Z}_{j2}\boldsymbol{\gamma}_{j2}, \tag{2.2}$$

$$g_3(\boldsymbol{\nu}) = \boldsymbol{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \boldsymbol{Z}_{j3}\boldsymbol{\gamma}_{j3}, \tag{2.3}$$

$$g_4(\boldsymbol{\tau}) = \boldsymbol{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \boldsymbol{Z}_{j4}\boldsymbol{\gamma}_{j4}, \tag{2.4}$$

où les paramètres μ, σ, ν et τ sont représentés sous leur forme vectorielle de dimen-

1. *Generalized additive models for location, scale and shape.*

sion $n \times 1$ correspondant aux n données. De plus, les termes $\boldsymbol{X\beta}$ représentent les parties paramétriques du modèle et les termes $\boldsymbol{Z\gamma}$, les parties non paramétriques. Pour la partie paramétrique, $\boldsymbol{X}_k$, pour k allant de 1 jusqu'à 4, est une matrice connue de design de dimension $n \times J'_k$, où J'_k correspond au nombre de variables explicatives composant $\boldsymbol{X}_k$, et $\boldsymbol{\beta}_k$ est un vecteur de coefficients de longueur J'_k. Du côté non paramétrique, $\boldsymbol{Z}_{jk}$ est une matrice de design connue de dimension $n \times q_{jk}$ et $\boldsymbol{\gamma}_{jk}$ est un vecteur de variables aléatoires de longueur q_{jk}.

Le modèle permet d'ajouter le nombre de termes additifs désirés J_k pour chaque fonction de lien $g_k(\cdot)$. Évidemment, dans l'éventualité où J_k serait égal à 0, il n'y a pas de terme additif pour la k^{e} fonction de lien et le terme de sommation disparaît de l'expression de $g_k(\cdot)$.

Il est à noter que les composantes $\boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}$ peuvent servir à modéliser différents types de termes additifs, comme des fonctions de lissage ou des effets aléatoires, selon la structure qu'on leur attribue dans le modèle.

2.1 Les splines cubiques

Les splines constituent l'un des choix possibles comme fonctions de lissage. Il s'agit principalement de fonctions polynomiales définies par morceaux. Dans cette section, un résumé des résultats présentés et détaillés par Boucher *et al.* (2017) sera réalisé. Cette section sera donc fortement inspirée par cet article et par le mémoire de Steven Côté sur lequel cet article est basé (Côté, 2016). Pour davantage d'information sur les splines, Green et Silverman (1993) et Wood (2006) constituent de bonnes références.

Les splines cubiques sont des splines dont le polynôme de degré maximal en est un de degré 3. Elles possèdent certaines propriétés intéressantes et, pour cette raison, nous allons nous concentrer sur celles-ci.

Supposons que nous avons comme données $\{x_j, y_j\}$ pour $j = 1, ..., n$ avec $x_j < x_{j+1}$. Les splines cubiques pourraient alors être utilisées soit dans un contexte d'interpolation, soit dans un contexte d'ajustement ou de lissage.

2.1.1. Spline cubique d'interpolation

Dans un contexte d'interpolation, l'objectif est de trouver une fonction passant par tous les points de notre jeu de données. Si l'on désire trouver une fonction lisse, les splines cubiques s'avèrent alors très utiles.

En effet, afin d'obtenir une courbe continue, l'interpolation par spline cubique[2] impose que les dérivées premières et secondes de la fonction, qui sera par morceaux, soient continues aux noeuds d'interpolation (les données dans ce cas-ci) et entre les noeuds.

En posant $h_j = x_{j+1} - x_j$, nous obtennons alors la fonction $s(\cdot)$ de lissage suivante :

$$s(x) = a_j^-(x)y_j + a_j^+(x)y_{j+1} + c_j^-(x)s''(x_j) + c_j^+(x)s''(x_{j+1}), \quad x_j \leq x \leq x_{j+1},$$

où

$$a_j^-(x) = (x_{j+1} - x)/h_j, \qquad c_j^-(x) = \left[(x_{j+1} - x)^3/h_j - h_j(x_{j+1} - x)\right]/6,$$
$$a_j^+(x) = (x - x_j)/h_j, \qquad c_j^+(x) = \left[(x - x_j)^3/h_j - h_j(x - x_j)\right]/6.$$

Pour obtenir les valeurs des dérivées secondes, il suffit de résoudre le système d'équations basé sur la condition de continuité des dérivées premières aux noeuds

$$\frac{h_j}{6}s''(x_j) + \frac{(x_{j+2} - x_j)}{3}s''(x_{j+1}) + \frac{h_{j+1}}{6}s''(x_{j+2}) = \frac{y_{j+2} - y_{j+1}}{h_{j+1}} - \frac{y_{j+1} - y_j}{h_j},$$

où $j = 1, ..., n - 2$.

---

2. Traduction libre de *cubic interpolating spline.*

Nous avons alors n dérivées secondes inconnues avec $n-2$ équations. Pour pouvoir résoudre ce système d'équations, il faut donc poser 2 conditions supplémentaires. Habituellement, ces conditions portent sur les bornes x_1 et x_n. Lorsqu'on pose $s''(x_1) = 0$ et $s''(x_n) = 0$, la fonction $s(\cdot)$ obtenue se nomme spline cubique naturelle.

Il a d'ailleurs été prouvé [3] que la spline cubique naturelle est la fonction d'interpolation continue sur $[x_1, x_n]$ possédant des dérivées premières continues la plus lisse possible selon la minimisation du critère

$$\int_{x_1}^{x_n} g''(x)^2 dx.$$

Ce critère constitue une façon de mesurer l'oscillation d'une fonction. En effet, plus une fonction sera ondulée, plus sa dérivée première variera et plus le carré de sa dérivée seconde sera grand.

2.1.2 Spline cubique d'ajustement

L'interpolation n'est toutefois pas idéale dans toutes les situations. En effet, par sa construction, elle va généralement accorder trop d'importance aux variations locales au détriment de la tendance générale qu'expriment les données.

L'ajustement, ou le lissage, est alors une alternative fort intéressante. Sous ce contexte, on cherche à respecter 2 objectifs opposés, soit de trouver une courbe s'approchant le plus possible des données tout en demeurant la plus lisse possible. Ainsi, la tendance générale est favorisée aux variations locales.

Posons $g(\cdot)$, la fonction de lissage recherchée. Précédemment, dans un contexte d'interpolation, nous avions que $g(x_j) = y_j$. Pour un contexte d'ajustement, cette

3. Côté (2016) et Wood (2006) détaillent cette preuve.

hypothèse est assouplie et les $g(x_j)$, pour $j = 1, ..., n$, sont désormais considérés comme des paramètres qu'il faudra estimer.

Le critère d'ajustement s'exprime alors ainsi :

$$\sum_{j=1}^{n}(y_j - g(x_j))^2 + \lambda \int_{x_1}^{x_n} g''(x)^2 dx. \tag{2.5}$$

Le premier terme de la fonction (2.5) correspond à la somme des carrés des différences entre les observations et la fonction $g(\cdot)$ alors que le deuxième terme constitue une pénalité visant à réduire les oscillations de la fonction. Le paramètre de lissage λ peut être estimé par différentes méthodes (que nous verrons à la section 2.3), mais il sera fixé arbitrairement pour l'instant.

La minimisation du critère (2.5) mène alors au respect des 2 objectifs de l'ajustement, soit que la fonction $g(\cdot)$ approche le plus possible les données sans toutefois subir trop de variations.

De plus, il peut être prouvé que la spline cubique naturelle est la fonction $g(\cdot)$ qui minimise le critère (2.5) [4]. Afin de respecter le contexte d'ajustement dans lequel nous nous trouvons, il sera alors question de la spline cubique d'ajustement [5].

2.2 La régression par splines cubiques pénalisées

Les splines cubiques présentées à la section 2.1 présentent toutefois un inconvénient qui peut être important selon leur application, soit le fait qu'il y ait autant de paramètres à estimer que de données. En effet, en travaillant avec des bases de données regroupant des centaines de milliers d'entrées, comme c'est souvent le cas en assurance, nous pouvons nous retrouver face à de très longs délais de calculs

4. Côté (2016) et Wood (2006) détaillent cette preuve.

5. Traduction libre de *cubic smoothing spline.*

numériques.

La régression par spline cubique pénalisée[6] constitue une solution à ce problème. Il s'agit d'une spline cubique dont le nombre de paramètres est limité, ce qui réduit le temps requis pour les calculs tout en conservant les propriétés importantes des splines.

Il existe différentes options pour la régression par spline cubique pénalisée, mais nous nous concentrerons sur les P-splines. Ce choix est fait de façon à faciliter l'utilisation de l'extension *gamlss* dans le logiciel *R* qui offre la possibilité d'utiliser ce type de spline pénalisée.

Les P-splines, développés par Eilers et Marx (1996), sont basées sur les B-splines. Dans cette section, les B-splines seront d'abord étudiées, puis les P-splines suivront.

2.2.1 Les B-splines

Les B-splines sont essentiellement des fonctions polynomiales définies par morceaux et connectées par des noeuds. Plusieurs ouvrages, tels que De Boor *et al.* (1978), Dierckx (1995) et Schumaker (2007), traitent en profondeur de ces splines.

Contrairement à ce que nous avons vu à la section 2.1 avec les splines cubiques d'interpolation et d'ajustement, il n'y a pas nécessairement autant de noeuds que de données pour les B-splines. En fait, le choix du nombre de noeuds dépend de l'utilisateur. Dans ce mémoire, nous travaillerons avec des noeuds équidistants, puisque cela simplifie les calculs.

Supposons que nous avons désormais comme données $\{x_i, y_i\}$ pour $i = 1, ..., n$.

---

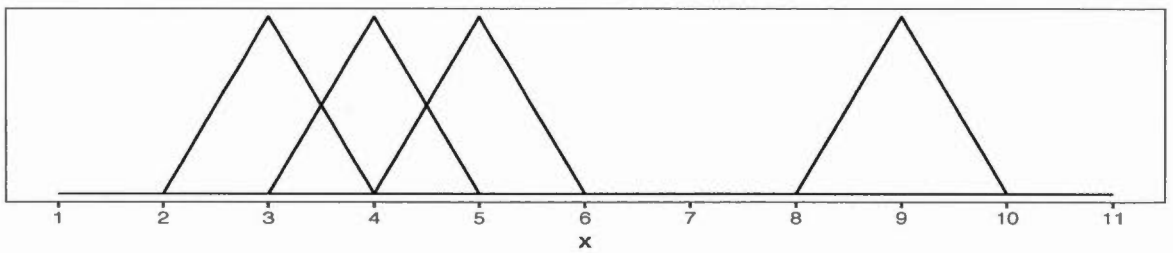
6. Traduction libre de *penalized cubic regression spline.*

Le domaine des données est donc $[x_{min}, x_{max}]$. Il s'agira également du domaine de la fonction de lissage par B-splines. Pour positionner les noeuds, il faut ensuite sélectionner arbitrairement un nombre d'intervalles m pour séparer notre domaine en parties égales. De plus, le degré d des B-splines doit être choisi. Comme nous nous intéressons aux splines cubiques, nous travaillerons avec $d = 3$. Au total, il y aura $q = m + d$ B-splines pour composer la fonction de lissage.

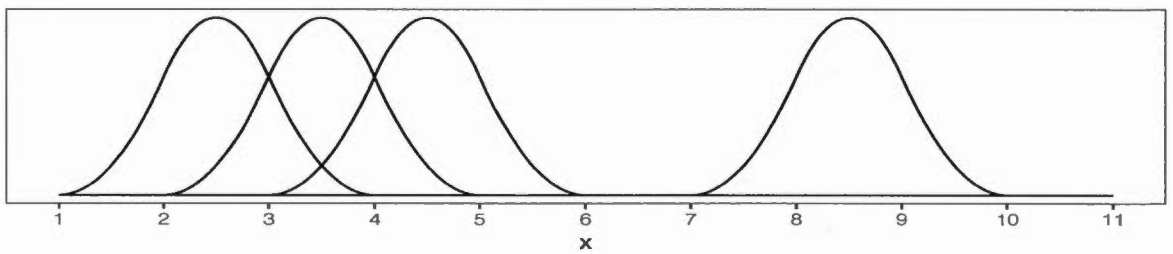
Eilers et Marx (1996) dressent un résumé intéressant accompagné d'une représentation graphique claire des B-splines. Chacune des q B-splines sera constituée de $d + 1$ morceaux polynomiaux de degré d attachés ensemble à d noeuds intérieurs. À ces noeuds intérieurs, les dérivées d'une B-spline sont continues jusqu'à l'ordre $d - 1$. Pour avoir d noeuds intérieurs, une B-spline est donc comprise à l'intérieur d'un domaine de $d + 2$ noeuds voisins. Sur ce domaine, elle sera positive et à l'extérieur de celui-ci, elle sera nulle. En d'autres mots, les B-splines ont comme particularité d'être des fonctions très locales. De plus, comme les B-splines s'étendent sur plusieurs noeuds, certaines B-splines vont se chevaucher. À ce sujet, une B-spline peut chevaucher jusqu'à $2d$ morceaux polynomiaux des B-splines voisines.

La figure 2.1, inspirée de la représentation de Eilers et Marx, illustre bien ces propriétés. Elle a été obtenue en calculant des B-splines sur un domaine allant de 1 à 11 et en divisant ce domaine par $m = 10$ intervalles (les noeuds intérieurs correspondent à chaque entier entre 1 et 11). De plus, les résultats pour les degrés 1 jusqu'à 3 sont présentés. Toutefois, seulement certaines des fonctions B-splines sont affichées afin d'alléger les figures résultantes. Il s'agit des 3^e, 4^e, 5^e et 9^e B-splines pour les degrés 1 et 2 et des 4^e, 5^e, 6^e et 10^e B-splines pour le degré 3.

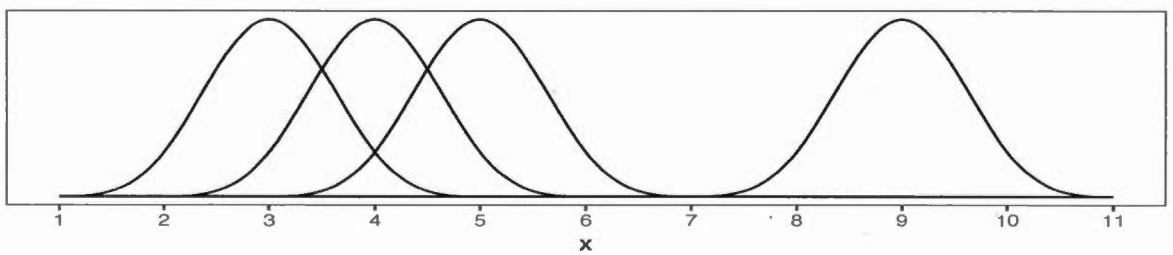
Pour que les B-splines aux extrémités du domaine $[x_{min}, x_{max}]$ aient l'espace nécessaire pour déployer leurs morceaux polynomiaux, il faut étirer notre domaine



(a) Degré 1



(b) Degré 2



(c) Degré 3

Figure 2.1: Comparaison de fonctions B-splines se chevauchant et individuelles de degré 1 (a), 2 (b) et 3 (c)

et ajouter $2d$ noeuds à notre séquence de noeuds équidistants, soit d noeuds au début et à la fin. Il y aura donc $m + 2d + 1$ noeuds au total. De plus, comme ces noeuds supplémentaires se retrouvent à l'extérieur du domaine de la fonction de lissage, ils ont un impact négligeable sur la fonction de lissage finale.

Le tableau 2.1 dresse un résumé des paramètres et des caractéristiques des B-splines.

Tableau 2.1: Résumé des paramètres et des caractéristiques des B-splines

Domaine	$[x_{min}, x_{max}]$
Nombre d'intervalles sur le domaine	m
Degré	d (3 pour des splines cubiques)
Nombre de B-splines	$q = m + d$
Nombre total de noeuds	$m + 2d + 1$

Les B-splines sont généralement définies par l'entremise de leur propriété de récursion, qui est largement utilisée pour faciliter les calculs numériques de ces fonctions (De Boor, 1972; Cox, 1972) :

$$B_{j,d+1}(x) = \frac{x - x_j}{x_{j+d} - x_j} B_{j,d}(x) + \frac{x_{j+d+1} - x}{x_{j+d+1} - x_{j+1}} B_{j+1,d}(x),$$
$$B_{j,1}(x) = \begin{cases} 1, & \text{si } x_j \leq x < x_{j+1}, \\ 0, & \text{sinon}, \end{cases} \tag{2.6}$$

où $B_{j,d+1}(x)$ représente la j^{e} fonction B-spline de degré d évaluée au point x et $j = 1, ..., m + d$. Les x_j représentent les noeuds avec x_1 comme premier noeud et x_{m+2d+1} comme dernier noeud. Il s'agit de la définition des B-splines telle que retrouvée dans Dierckx (1995). C'est donc dire que x_1 jusqu'à x_d ainsi que x_{m+d+1} jusqu'à x_{m+2d+1} sont des noeuds situés à l'extérieur du domaine.

Avec l'équation (2.6), il sera question de B-spline normalisée, car la somme de toutes les fonctions de base pour n'importe quel x compris sur le domaine est égale à 1, soit

$$\sum_{j=1}^{q} B_{j,d+1}(x) = 1,$$

où, rappelons-nous, $q = m + d$ correspond au nombre de fonctions de base composant la fonction de lissage.

La fonction de lissage résultante des B-splines $s(\cdot)$ est constituée de la somme des q B-splines multipliées par un coefficient a_j telle que

$$s(x) = \sum_{j=1}^{q} a_j B_{j,d+1}(x). \tag{2.7}$$

L'estimation des coefficients a_j peut s'effectuer par la minimisation de la somme de la différence entre les données y_i et la fonction $s(\cdot)$ au carré, c'est-à-dire la minimisation de S de l'équation ci-dessous :

$$\begin{aligned} S &= \sum_{i=1}^{n} \Big(y_i - s(x_i) \Big)^2, \\ &= \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{q} a_j B_{j,d+1}(x_i) \right)^2. \end{aligned} \tag{2.8}$$

Toutefois, le choix du nombre de noeuds influence grandement la forme finale de la fonction de lissage. Sélectionner un nombre trop élevé de noeuds mène à du surajustement alors qu'un nombre insuffisant de noeuds peut mener à du sous-ajustement.

Pour corriger cette lacune, plusieurs options s'offrent à nous, dont celle des P-splines de Eilers et Marx (1996).

Exemple sur les B-splines

Pour illustrer l'application de B-splines, prenons le jeu de données *cars*, disponible gratuitement sur le logiciel *R*. Il s'agit de 50 données datant des années 1920 mesurant la distance requise en pieds pour immobiliser des voitures selon leur vitesse en miles à l'heure.

Tout d'abord, il faut choisir le degré des B-splines. Posons $d = 3$. Il faut également déterminer le nombre d'intervalles m. En prenant $m = 5$, on se retrouve avec $q = m + d = 8$ fonctions B-splines, pour un total de $m + 2d + 1 = 12$ noeuds. Les coefficients estimés par la minimisation de l'équation (2.8) ainsi que la courbe obtenue sont présentés au tableau 2.2 et à la figure 2.2 respectivement.

Tableau 2.2: Coefficients estimés pour l'exemple de B-splines pour la distance de freinage

$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	$\hat{a}_5$	$\hat{a}_6$	$\hat{a}_7$	$\hat{a}_8$
-18,26	10,23	13,33	28,62	53,67	46,19	102,89	113,84

En répétant cet exemple avec $m = 1$ et $m = 10$ à la figure 2.3, on voit clairement que la forme de la courbe ajustée dépend grandement du nombre d'intervalles m choisis et, par conséquent, du nombre de noeuds sélectionnés. Avec $m = 10$, le surajustement est flagrant alors qu'avec $m = 1$, l'ajustement semble être bon. Toutefois, on ne peut pas prendre une valeur inférieure à 1 pour m, donc on ne peut pas se rendre à apercevoir de façon nette du sous-ajustement.

Figure 2.2: Courbe ajustée de la distance de freinage selon la vitesse

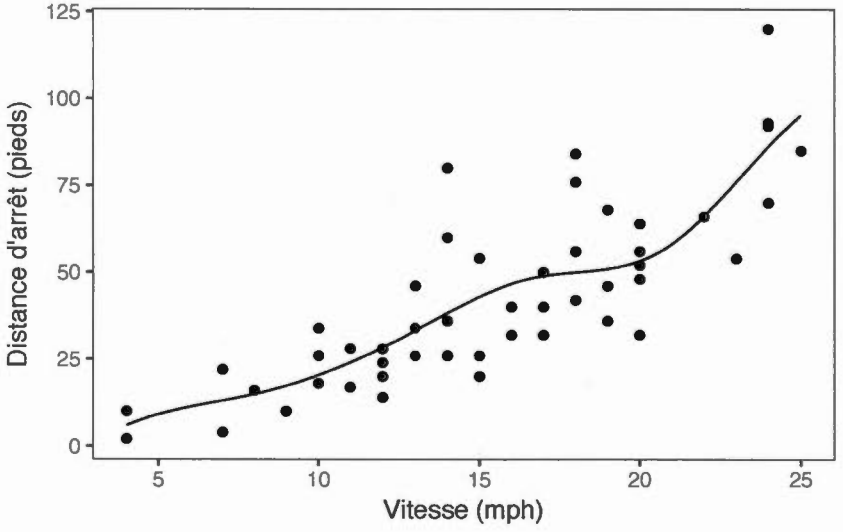
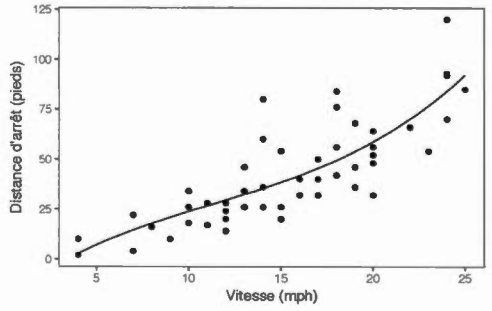
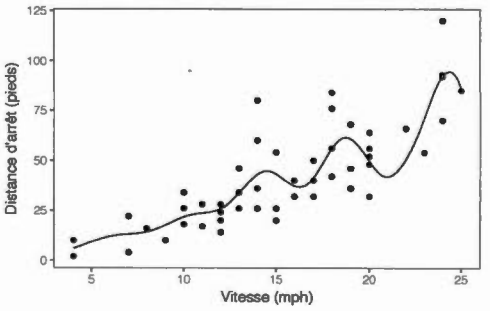


Figure 2.3: Courbe ajustée de la distance de freinage selon la vitesse pour différentes valeurs de m



(a) $m = 1$



(b) $m = 10$

2.2.2 Les P-splines

Eilers et Marx (1996) ont trouvé une solution concernant le choix du nombre de noeuds pour les fonctions de lissage par B-splines. À ce sujet, Wood (2006) considère même que le véritable intérêt statistique lié aux B-splines résulte du travail de Eilers et Marx.

Ces derniers ont proposé d'ajouter un terme de pénalité à l'équation (2.8) basé sur la différence d'ordre r des coefficients au carré, soit

$$S = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{q} a_j B_{j,d+1}(x_i) \right)^2 + \lambda \sum_{j=r+1}^{q} \left(\Delta^r a_j \right)^2, \tag{2.9}$$

où $\Delta a_j = a_j - a_{j-1}$. Notons également que, pour $r > 1$, $\Delta^r a_j = \Delta^{r-1}(\Delta a_j)$.

Cette pénalité est contrôlée par le paramètre de lissage λ. En procédant ainsi, plutôt que de limiter le nombre de noeuds, on limite l'influence du nombre de noeuds sur la fonction de lissage quitte à avoir un peu plus de noeuds que nécessaire.

La pénalité repose sur la variation des coefficients des B-splines, ce qui constitue une façon de mesurer les oscillations d'une fonction de lissage par B-splines. En effet, plus cette fonction va osciller, plus ses coefficients vont varier. De plus, ce modèle donne le loisir de choisir l'ordre de la différence des coefficients.

La forme de l'équation (2.9) est très semblable à celle de l'équation (2.5). En minimisant S pour estimer les coefficients a_j, le premier terme de l'équation (2.9) a pour objectif que la fonction de lissage soit le plus près des données alors que le deuxième terme cherche plutôt à obtenir la courbe la plus lisse possible. Le résultat est donc très similaire à ce que nous avons vu avec les splines cubiques d'ajustement à la section 2.1.2.

L'équation (2.9) peut être réécrite sous la forme matricielle

$$S = (\boldsymbol{y} - \boldsymbol{Ba})^T(\boldsymbol{y} - \boldsymbol{Ba}) + \lambda(\boldsymbol{D}_r\boldsymbol{a})^T(\boldsymbol{D}_r\boldsymbol{a}), \tag{2.10}$$

où $\boldsymbol{a}$ est le vecteur $q \times 1$ des coefficients a_j, $\boldsymbol{D}_r$ est une matrice de dimension $(q-r)\times q$ représentant Δ^r, et $\boldsymbol{B}$ est une matrice $n \times q$ dont l'élément $b_{i,j}$ correspond à $B_{j,d+1}(x_i)$.

Il peut être montré que la minimisation de l'équation (2.9) pour un λ fixé s'obtient lorsque les coefficients a_j valent :

$$\hat{\boldsymbol{a}} = (\boldsymbol{B}^T\boldsymbol{B} + \lambda\boldsymbol{D}_r^T\boldsymbol{D}_r)^{-1}\boldsymbol{B}^T\boldsymbol{y}. \tag{2.11}$$

Le choix du paramètre λ permet d'augmenter ou de réduire le poids de la pénalité. D'ailleurs, avec $\lambda = 0$, on retrouve sans surprise la fonction de lissage par B-splines, puisque le terme de pénalité disparaît. Pour un λ très élevé, la pénalité devient beaucoup trop coûteuse pour ajouter la moindre ondulation à la courbe de lissage. Dans cette situation, nous obtenons alors une droite comme fonction de lissage.

Toutefois, il est toujours nécessaire de déterminer le nombre de noeuds et la valeur du paramètre de lissage λ. À ce sujet, Ruppert (2002), Ruppert *et al.* (2003) et Wood (2006) parviennent tous à la même conclusion, soit que le paramètre de lissage λ est celui qui a la plus grande importance sur la forme finale de la courbe de lissage. Ce paramètre peut être estimé par la minimisation du critère GCV ou du critère AIC que nous verrons dans la prochaine section. De plus, le choix du nombre de noeuds n'est pas critique. Comme l'explique Wood (2006), le nombre de noeuds vient simplement imposer une limite supérieure sur la flexibilité d'un terme. C'est plutôt le paramètre de lissage λ qui contrôle les degrés de liberté effectifs (EDF). Ainsi, l'ajustement du modèle est plutôt insensible au nombre de noeuds pourvu qu'il n'y ait pas un nombre de noeuds trop faible.

Établir le nombre de noeuds demeure donc au choix de l'utilisateur, mais tant que ce choix procure au modèle suffisamment de flexibilité, il n'y aura pas vraiment de différence sur la courbe de lissage obtenue parce que celle-ci est contrôlée par le paramètre de lissage λ.

Exemple sur les P-splines

Reprenons l'exemple de la fin de la section 2.2.1 avec la base de données *cars*, mais pour des fonctions de lissage par P-splines.

Réutilisons $m = 10$ intervalles, où nous avions un problème de surajustement dans l'exemple sur les B-splines, et posons $d = 3$ comme degré pour les B-splines et une pénalité de différence d'ordre $r = 2$. Nous allons donc avoir $q = m + d = 13$ fonctions B-splines, pour un total de $m + 2d + 1 = 17$ noeuds.

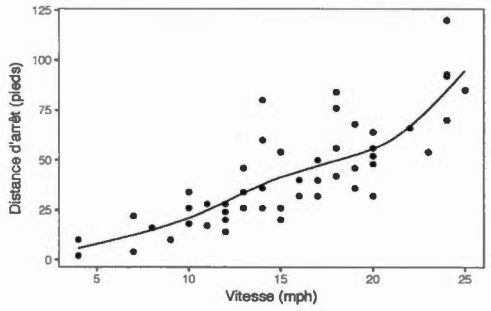
Pour $\lambda = 1$, avec l'équation (2.11), nous trouvons les coefficients présentés au tableau 2.3.

Tableau 2.3: Coefficients estimés pour l'exemple de P-splines pour la distance de freinage

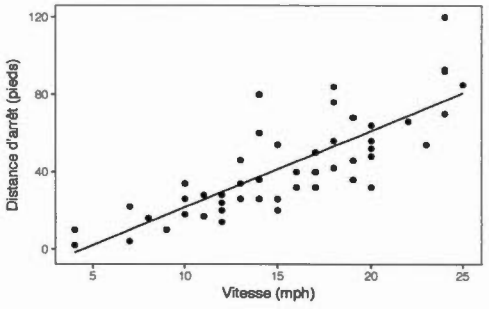
$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	$\hat{a}_5$	$\hat{a}_6$	$\hat{a}_7$	$\hat{a}_8$	$\hat{a}_9$	$\hat{a}_{10}$	$\hat{a}_{11}$	$\hat{a}_{12}$	$\hat{a}_{13}$
1,4	5,8	10,3	15,2	21,6	30,8	40,4	45,9	51,8	57,4	73,5	95,0	115,7

La figure 2.4 présente les courbes de lissage par P-splines obtenues pour $\lambda = 1$ et $\lambda = 100\,000$. On peut remarquer que le surajustement avec $m = 10$ pour le lissage par B-splines est corrigé en grande partie par l'ajout d'une pénalité avec $\lambda = 1$. Si le terme de pénalité est très grand, comme avec $\lambda = 100\,000$, on obtient alors une droite.

Figure 2.4: Courbe ajustée de la distance de freinage selon la vitesse pour différents λ



(a) $m = 10$ et $\lambda = 1$



(b) $m = 10$ et $\lambda = 100\,000$

2.3 Choisir le paramètre de lissage

Jusqu'ici, nous avons fixé le paramètre de lissage de façon arbitraire. Il est toutefois possible de l'estimer, notamment par le critère d'information de Akaike (AIC[7]) ou par le critère du GCV.

2.3.1 Établir le paramètre de lissage par le AIC

De façon générale, plus un modèle compte de paramètres, mieux il s'ajuste aux données. Toutefois, avoir trop de paramètres peut entraîner d'autres problèmes pour un modèle, comme de présenter du surajustement ou une complexité supplémentaire peu justifiable par rapport au gain d'ajustement.

Une façon de mesurer la significativité du gain d'ajustement passe par le critère AIC (Akaike, 1974). Ce critère repose entièrement sur le principe de la log-vraisemblance pénalisée et revient à maximiser la log-vraisemblance d'un modèle,

7. *Akaike information criterion.*

mais en tenant compte d'une pénalité sur les degrés de liberté effectifs du modèle :

$$AIC = -2\ell + 2 \times EDF, \tag{2.12}$$

où ℓ est la log-vraisemblance du modèle et EDF, les degrés de liberté effectifs du modèle.

Généralement, c'est le nombre de paramètres qui est utilisé plutôt que les EDF. Toutefois, comme l'expliquent James *et al.* (2013), les paramètres associés aux noeuds des P-splines subissent des contraintes importantes, c'est-à-dire qu'ils sont fortement restreints. Recourir aux EDF devient alors une façon plus appropriée de mesurer la flexibilité des P-splines.

Nous reviendrons à la section 2.5.2 sur la façon de calculer les EDF.

Selon ce critère, le modèle minimisant le AIC est celui à favoriser. Ainsi, trouver le paramètre de lissage λ optimal selon ce critère revient à trouver celui qui minimise le AIC.

Le critère BIC[8],

$$BIC = -2\ell + log(n) \times EDF, \tag{2.13}$$

où ℓ est la log-vraisemblance du modèle et n correspond au nombre de données, peut également être utilisé. Similaire au critère AIC, le critère BIC revient à maximiser la log-vraisemblance d'un modèle sujette à une pénalité. Ainsi, le modèle minimisant le BIC est celui à favoriser.

---

8. *Bayesian information criterion.*

2.3.2 Établir le paramètre de lissage par le critère GCV[9]

Le score de validation croisée généralisée (GCV[10]) repose, comme son nom l'indique, sur le principe de validation croisée. Il s'agit de mesurer la qualité d'ajustement d'un modèle avec des données n'ayant pas servi à estimer ce modèle.

Le GCV est une généralisation du score de validation croisée ordinaire (OCV[11]). L'OCV est défini comme ceci :

$$\text{OCV} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_i^{[-i]} - y_i\right)^2, \tag{2.14}$$

où $\hat{f}_i^{[-i]}$ représente le modèle estimé en excluant la i^{e} donnée et évalué pour cette donnée, n correspond au nombre total de données et y_i est la i^{e} donnée de la variable réponse qu'on modélise.

Pour éviter d'avoir à estimer n modèles, il peut être montré que

$$\text{OCV} = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{f}_i)^2}{(1 - H_{ii})^2}, \tag{2.15}$$

où $\hat{f}$ est le modèle ajusté avec l'ensemble des données et $\boldsymbol{H}$, la matrice chapeau.

Pour les P-splines, la matrice chapeau peut s'écrire

$$\boldsymbol{H} = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B} + \lambda\boldsymbol{D}_r^T\boldsymbol{D}_r)^{-1}\boldsymbol{B}^T, \tag{2.16}$$

où λ est le paramètre de lissage et les matrices $\boldsymbol{B}$ et $\boldsymbol{D}_r$ sont telles que définies à la section 2.2.2.

---

9. Cette section sera un résumé de Wood (2006), Côté (2016) et Boucher *et al.* (2017) concernant le critère GCV.

10. *Generalized cross validation score.*

11. *Ordinary cross validation score.*

Les poids $(1 - H_{ii})$ de l'équation (2.15) peuvent être remplacés par le poids moyen $tr(\boldsymbol{I} - \boldsymbol{H})/n$, où $\boldsymbol{I}$ est la matrice identité de dimension $n \times n$, ce qui donne le score GCV :

$$\text{GCV} = \frac{n \sum_{i=1}^{n} (y_i - \hat{f}_i)^2}{[tr(\boldsymbol{I} - \boldsymbol{H})]^2}. \tag{2.17}$$

Le GCV possède comme avantages sur l'OCV la propriété d'invariance (voir section 4.5.2 et 4.5.3 de Wood (2006)) et une simplicité accrue pour les calculs numériques, ce qui explique son utilisation plus répandue.

Selon le critère GCV, la meilleure valeur pour le paramètre de lissage λ est celle qui minimise l'équation (2.17).

Exemple de sélection de λ par le critère GCV Reprenons l'exemple des sections précédentes sur la distance de freinage selon la vitesse. Précédemment, avec l'exemple de la section 2.2.2 sur les P-splines, nous avions fixé 2 valeurs pour λ, soit 1 et 100 000. Nous pourrions vouloir connaître le paramètre de lissage optimal selon le critère GCV.

Pour cela, il suffit de calculer le score GCV obtenu pour différentes valeurs de λ et sélectionner le λ qui minimise ce score. La figure 2.5 illustre le score GCV selon le paramètre λ. Pour $\lambda = 109{,}2$, le score GCV est minimisé avec une valeur de 244,09.

La figure 2.6 montre la courbe de lissage correspondante au paramètre de pénalité optimal selon le GCV.

Figure 2.5: Score GCV selon le paramètre de lissage pour l'exemple de la distance de freinage

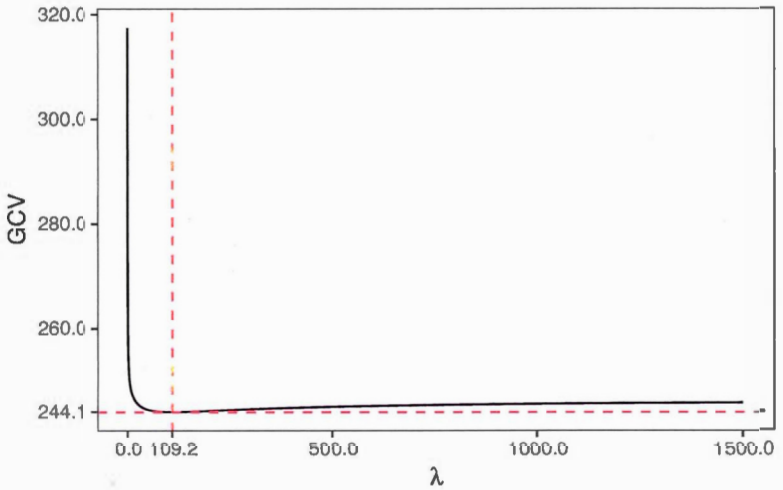
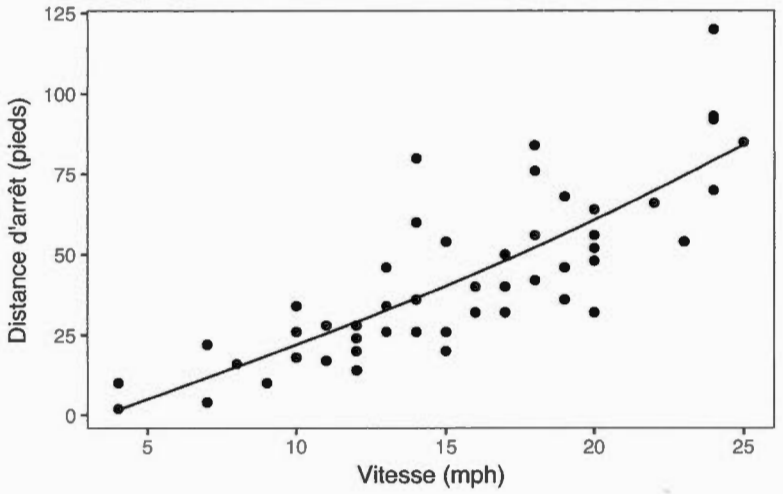


Figure 2.6: Courbe ajustée de la distance de freinage selon la vitesse pour le λ optimal selon le critère GCV



2.4 Quelques distributions discrètes de probabilité

Contrairement aux modèles GLM et GAM, les distributions de probabilité des modèles GAMLSS ne se limitent pas à la famille exponentielle linéaire. Ainsi, une plus grande variété de distributions peut être utilisée avec les modèles GAMLSS. Comme nous nous intéressons particulièrement à la modélisation du nombre de réclamations, qui est une variable aléatoire discrète, nous nous concentrerons exclusivement sur des distributions discrètes.

Des informations supplémentaires sur les distributions qui seront présentées dans cette section ainsi que sur leur application en assurance se trouvent notamment en consultant Boucher *et al.* (2008, 2007) et Boucher et Denuit (2006).

Dans les sous-sections suivantes, Y_i représentera donc le nombre de réclamations de l'observation i, où i va de 1 jusqu'au nombre total d'observations n. De plus, à l'exception de la distribution binomiale négative multivariée (MVNB), toutes les observations sont considérées indépendantes, y compris celles concernant le même assuré.

2.4.1 Poisson

La distribution Poisson, introduite comme son nom l'indique par Siméon Denis Poisson (1837), est une loi de probabilité discrète dont la fonction de masse s'écrit :

$$\Pr[Y_i = k] = \frac{\exp(-\lambda_i)\lambda_i^k}{k!}, \qquad k = 0, 1, 2, ...,$$

où λ_i représente le paramètre de la distribution Poisson pour l'assuré i et est strictement positif.

De plus, la distribution Poisson fait partie de la famille exponentielle linéaire, ce qui permet de l'utiliser pour des modèles GLM et GAM.

Une propriété importante de la Poisson est l'équidispersion, c'est-à-dire que son espérance et sa variance sont égales :

$$E[Y_i] = \lambda_i,$$
$$Var[Y_i] = \lambda_i.$$

Cette propriété peut toutefois causer des ennuis pour la modélisation de la fréquence de réclamations en assurance, puisque les données d'assurance présentent habituellement de la surdispersion. Les prochaines distributions qui seront présentées corrigent ce problème avec l'ajout d'un paramètre d'hétérogénéité qui élimine l'exigence d'équidispersion.

2.4.2 Binomiale négative de type 2

Une première approche pour permettre la surdispersion des données repose sur l'ajout d'un paramètre d'hétérogénéité Θ à la distribution Poisson.

Supposons désormais que $Y_i|\Theta_i \sim \text{Poisson}(\lambda_i\Theta_i)$ où $\Theta_i \sim \text{Gamma}(\alpha_i^{-1},\alpha_i^{-1})$.

Au final, Y_i suit une distribution binomiale négative de type 2 dont la fonction de probabilité s'écrit :

$$\Pr[Y_i = k] = \frac{\Gamma(\alpha_i^{-1}+k)}{\Gamma(\alpha_i^{-1})k!}\left(\frac{\alpha_i^{-1}}{\alpha_i^{-1}+\lambda_i}\right)^{\alpha_i^{-1}}\left(\frac{\lambda_i}{\alpha_i^{-1}+\lambda_i}\right)^{k}, \qquad k = 0, 1, 2, ...,$$

où α_i est strictement positif et représente le paramètre de surdipersion de l'assuré i et λ_i est un autre paramètre strictement positif lié à l'assuré i.

À ce sujet, λ_i correspond également à l'espérance du nombre de réclamations de l'assuré i :

$$E[Y_i] = \lambda_i.$$

Puisque le paramètre α_i est strictement positif, on se retrouve avec un terme de variance plus grand que l'espérance et, donc, dans une situation de surdispersion :

$$Var[Y_i] = \lambda_i + \alpha_i \lambda_i^2.$$

Il est important de noter que la binomiale négative de type 2 est une généralisation de la distribution Poisson. En effet, pour $\alpha_i \to 0$, on trouve que $Y_i \sim \text{Poisson}(\lambda_i)$.

2.4.3 Binomiale négative de type 1

La distribution du paramètre d'hétérogénéité Θ_i peut être modifiée afin de trouver d'autres lois de probabilité pour Y_i. En effet, nous pouvons supposer que $Y_i|\Theta_i \sim \text{Poisson}(\lambda_i\Theta_i)$ où $\Theta_i \sim \text{Gamma}(\lambda_i\alpha_i^{-1}, \lambda_i\alpha_i^{-1})$.

Au final, Y_i suit alors une binomiale négative de type 1 dont la distribution s'écrit :

$$\Pr[Y_i = k] = \frac{\Gamma(\alpha_i^{-1}\lambda_i + k)}{\Gamma(\alpha_i^{-1}\lambda_i)k!}(1+\alpha_i)^{-\lambda_i/\alpha_i}(1+\alpha_i^{-1})^{-k}, \qquad k = 0, 1, 2, ...,$$

où α_i et λ_i sont les paramètres de surdispersion et de moyenne respectivement de l'assuré i et sont tous les deux strictement positifs.

Si l'espérance de Y_i sous une binomiale négative de type 1 est équivalente à celle d'une binomiale négative de type 2, ce n'est pas le cas de la variance :

$$E[Y_i] = \lambda_i,$$
$$Var[Y_i] = \lambda_i + \alpha_i\lambda_i.$$

En effet, il y a une différence sur le deuxième terme de la variance, soit que λ_i est à la puissance 1 et non 2.

Tout comme pour la binomiale négative de type 2, pour $\alpha_i \to 0$, on trouve que $Y_i \sim \text{Poisson}(\lambda_i)$. La binomiale négative de type 1 est donc elle aussi une généralisation de la distribution Poisson.

2.4.4 Poisson inverse-gaussienne

Le paramètre d'hétérogénéité peut également suivre une distribution autre que la gamma. La Poisson inverse-gaussienne de type $(r+1)$ s'obtient d'ailleurs en supposant que $Y_i|\Theta_i \sim \text{Poisson}(\lambda_i\Theta_i)$ et que Θ_i suit une loi inverse-gaussienne de moyenne 1 et de variance $\tau_i\lambda_i^{r-1}$.

La fonction de masse de probabilité de la Poisson inverse-gaussienne de type $(r+1)$ s'écrit :

$$\Pr[Y_i = k] = \frac{\lambda_i^k}{k!}\left(\frac{2}{\pi\tau_i\lambda_i^{r-1}}\right)^{0.5} \exp(-\tau_i\lambda_i^{r-1})(1+2\tau_i\lambda_i^r)^{-s_i/2}K_{s_i}(z_i), \quad k = 0,1,2,...,$$

où s_i et z_i sont définis ainsi :

$$s_i = k - 0.5,$$
$$z_i = \frac{(1+2\tau_i\lambda_i^r)^{0.5}}{\tau_i\lambda_i^{r-1}},$$

et où $K_j(\cdot)$ est une fonction Bessel modifiée de seconde espèce respectant :

$$K_{-0.5}(a) = \left(\frac{\pi}{2a}\right)^{0.5}\exp(-a),$$
$$K_{0.5}(a) = K_{-0.5}(a),$$
$$K_{s+1}(a) = K_{s-1}(a) + \frac{2s}{a}K_s(a).$$

Les paramètres τ_i et λ_i sont strictement positifs et représentent respectivement les paramètres de surdispersion et de moyenne de l'assuré i :

$$E[Y_i] = \lambda_i,$$
$$Var[Y_i] = \lambda_i + \tau_i\lambda_i^{r+1}.$$

On remarque que la forme de la variance dépend du type de Poisson inverse-gaussienne sélectionné. Pour $r = 1$, on retrouve une forme de variance très sem-

blable à la binomiale négative de type 2 alors qu'elle s'apparente davantage à la binomiale négative de type 1 pour $r = 0$.

Pour $\tau_i \to 0$, on trouve que la Poisson inverse-gaussienne revient à une Poisson(λ_i).

Dans ce mémoire, afin de limiter le nombre de modèles utilisés, nous allons uniquement considérer le cas $r = 1$ pour la Poisson inverse-gaussienne.

2.4.5 Poisson gonflée à zéro

En assurance, pour modéliser le nombre de réclamations, il y a typiquement un grand pourcentage des données pour lesquelles aucune réclamation n'est réalisée. Ceci peut occasionner des problèmes d'ajustement lorsqu'une distribution Poisson est utilisée, puisque celle-ci a tendance à sous-estimer le nombre d'assurés qui ne réclameront pas.

Une façon de corriger ce défaut est de recourir à la distribution Poisson gonflée à 0. Il s'agit d'une distribution Poisson dont la masse de probabilité est augmentée à 0 telle que :

$$\Pr[Y_i = k] = \begin{cases} \phi_i + (1 - \phi_i)\exp(-\lambda_i), & k = 0, \\ (1 - \phi_i)\dfrac{\exp(-\lambda_i)\lambda_i^k}{k!}, & k = 1, 2, ..., \end{cases} \tag{2.18}$$

où ϕ_i est le paramètre pour gonfler la probabilité d'avoir 0 réclamation de l'assuré i et λ_i est le paramètre strictement positif d'une distribution Poisson concernant l'assuré i. Évidemment, ϕ_i doit être compris en 0 et 1 afin que l'équation (2.18) soit bel et bien une distribution Poisson dont la fonction de masse de probabilité est gonflée à 0.

En procédant ainsi, nous avons alors que l'espérance de Y_i dépend de 2 paramètres,

soit ϕ_i et λ_i :

$$E[Y_i] = (1 - \phi_i)\lambda_i. \tag{2.19}$$

Par ailleurs, la variance de Y_i est supérieure à son espérance pour $\phi_i > 0$:

$$Var[Y_i] = (1 - \phi_i)(\lambda_i + \phi_i\lambda_i^2).$$

Cette distribution admet donc de la surdispersion.

Sans surprise, pour $\phi_i = 0$, on retrouve la distribution Poisson de moyenne λ_i.

2.4.6 Binomiale négative multivariée (MVNB)

Jusqu'à maintenant, avec les différentes distributions présentées à la section 2.4, toutes les observations ont été considérées indépendantes. Toutefois, il s'agit d'une hypothèse qui est fort probablement fausse avec des données d'assurance. En effet, comme le même assuré est observé plusieurs fois à travers différentes années, on peut s'attendre à ce qu'il y ait une dépendance entre les observations d'un même assuré.

La distribution binomiale négative multivariée (MVNB), développée par Hausman *et al.* (1984), introduit justement une dépendance entre les observations d'un même assuré tout en conservant l'hypothèse d'indépendance entre les assurés. En d'autres mots, elle permet l'utilisation de données longitudinales contrairement aux distributions précédentes qui se limitent à des données transversales.

Précédemment, nous avons modélisé Y_i, soit le nombre de réclamations de l'observation i, où i va de 1 jusqu'au nombre total n d'observations. Avec la MVNB, on considère désormais le nombre de réclamations de l'assuré i au temps t, représenté par $Y_{i,t}$, où i va de 1 jusqu'à m, le nombre total d'assurés, et t va de 1 jusqu'à T_i,

le nombre total de fois où l'assuré i est observé. Les données restent les mêmes que l'on travaille avec Y_i ou $Y_{i,t}$, mais la différence réside dans la façon de les traiter.

La MVNB suppose que $Y_{i,t}|\Theta_i \sim \text{Poisson}(\lambda_{i,t}\Theta_i)$ et que Θ_i suit une loi gamma de moyenne 1 et de variance ν. De plus, sachant Θ_i, les variables $Y_{i,1}$ jusqu'à Y_{i,T_i} sont considérées indépendantes. Avec ces hypothèses, on trouve que la distribution jointe de $Y_{i,1}, ..., Y_{i,T_i}$ est :

$$\Pr[Y_{i,1} = k_{i,1}, ..., Y_{i,T_i} = k_{i,T_i}] = \left(\prod_{t=1}^{T_i} \frac{\lambda_{i,t}^{k_{i,t}}}{k_{i,t}!}\right) \frac{\Gamma(k_{i,\bullet} + \nu)}{\Gamma(\nu)} \left(\frac{\nu}{\lambda_{i,\bullet} + \nu}\right)^{\nu} \left(\lambda_{i,\bullet} + \nu\right)^{-k_{i,\bullet}},$$

où $\lambda_{i,\bullet} = \sum_{t=1}^{T_i} \lambda_{i,t}$, $k_{i,\bullet} = \sum_{t=1}^{T_i} k_{i,t}$ et $k_{i,t}$ est un entier positif pour tout couple (i,t).

Comme son nom le suggère, la MVNB a des liens avec la binomiale négative. À ce sujet, la distribution de $Y_{i,1}$ peut être réécrite sous la forme d'une loi binomiale négative de type 2 de paramètres ν et $\tau = \nu$:

$$\Pr[Y_{i,1} = k_{i,1}] = \frac{\Gamma(\nu + k_{i,1})}{\Gamma(\nu) k_{i,1}!} \left(\frac{\tau}{\tau + \lambda_i}\right)^{\nu} \left(\frac{\lambda_i}{\tau + \lambda_i}\right)_{i,1}^{k}, \qquad k_{i,1} = 0, 1, 2, ...,$$

pour i allant de 1 jusqu'au nombre total d'assurés m.

Pour $t > 1$, la distribution de $Y_{i,t}|Y_{i,t-1}, ..., Y_{i,1}$ revient également à une loi binomiale négative de type 2, mais de paramètres $\nu^* = \nu + \sum_{j=1}^{t-1} k_{i,j}$ et $\tau^* = \nu + \sum_{j=1}^{t-1} \lambda_{i,j}$, telle que montrée par l'équation (2.20) :

$$\Pr[Y_{i,t} = k|Y_{i,t-1} = k_{i,t-1}, ..., Y_{i,1} = k_{i,1}] = \frac{\Gamma(\nu^* + k)}{\Gamma(\nu^*) k!} \left(\frac{\tau^*}{\tau^* + \lambda_i}\right)^{\nu^*} \left(\frac{\lambda_i}{\tau^* + \lambda_i}\right)^{k}, \tag{2.20}$$

où k est un entier positif et où $k_{i,t-1}, ..., k_{i,1}$ correspondent aux réalisations connues de $Y_{i,t-1}, ..., Y_{i,1}$.

En utilisant la méthode du conditionnement [12], on peut trouver facilement l'espérance et la variance de $Y_{i,t}$:

$$E[Y_{i,t}] = \lambda_{i,t},$$
$$Var[Y_{i,t}] = \lambda_{i,t} + \frac{1}{\nu}\lambda_{i,t}^2.$$

Comme le montre la forme de la variance de $Y_{i,t}$, la distribution MVNB admet de la surdispersion.

Lorsque vient le temps d'établir une prévision du nombre de réclamations pour un nouvel assuré, l'une des possibilités est d'utiliser $E[Y_{i,1}]$. Toutefois, lorsqu'il s'agit d'un assuré déjà observé, c'est-à-dire un assuré pour lequel $T_i \geq 1$, la dépendance entre les observations d'un même assuré avec la distribution MVNB permet de se servir des observations passées de cet assuré afin d'améliorer la prévision de son nombre de réclamations :

$$E[Y_{i,T_i+1}|y_{i,1}, ..., y_{i,T_i}] = \lambda_{i,T_i+1}\frac{\nu + \sum_{j=1}^{T_i} y_{i,j}}{\nu + \sum_{j=1}^{T_i} \lambda_{i,j}}. \tag{2.21}$$

Dans l'équation (2.21), la fraction sur le côté droit agit comme un facteur d'ajustement sur la prévision λ_{i,T_i+1} qui serait réalisée pour un assuré n'ayant jamais été observé. Ce facteur vient augmenter ou réduire cette prévision selon l'expérience passée de l'assuré.

2.5 GAMLSS avec P-splines

Revenons aux équations (2.1) à (2.4) du début du chapitre 2. Celles-ci peuvent être réécrites sous une forme générale avec $k = 1, 2, 3, 4$:

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}, \tag{2.22}$$

---

12. $E[A] = E[E[A|B]]$ et $Var[A] = E[Var[A|B]] + Var[E[A|B]]$.

où $\boldsymbol{\theta}_1 = \boldsymbol{\mu}$, $\boldsymbol{\theta}_2 = \boldsymbol{\sigma}$, $\boldsymbol{\theta}_3 = \boldsymbol{\nu}$ et $\boldsymbol{\theta}_4 = \boldsymbol{\tau}$, soit les paramètres de moyenne, de variance, d'asymétrie et d'aplatissement respectivement.

2.5.1 Estimation d'un GAMLSS avec des paramètres de lissage fixés

L'estimation d'un modèle GAMLSS pour des paramètres de lissage fixés peut se faire par la maximisation de la log-vraisemblance pénalisée :

$$\ell_p = \ell - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\boldsymbol{\gamma}_{jk}^T\boldsymbol{G}_{jk}\boldsymbol{\gamma}_{jk}, \tag{2.23}$$

où ℓ_p représente la log-vraisemblance pénalisée et ℓ, la log-vraisemblance du modèle. Le terme de droite constitue la pénalité. Le paramètre p représente le nombre de fonctions de lien utilisées dans le modèle. Par exemple, pour un GAM, p serait égal à 1, puisqu'il n'y a que la fonction de lien relié à $\boldsymbol{\mu}$.

Les structures des matrices $\boldsymbol{Z}$ de l'équation (2.22) et $\boldsymbol{G}$ de l'équation (2.23) dépendent du type de fonction additive utilisée.

Pour un terme additif (j,k) de P-splines, la matrice $\boldsymbol{Z}_{jk}$, de dimension $n \times q_{jk}$, contient les q_{jk} fonctions B-splines comme vues à la section 2.2.1 et représentées sous la forme matricielle par $\boldsymbol{B}$ à la section 2.2.2. La matrice $\boldsymbol{G}_{jk}$ est définie comme ceci :

$$\boldsymbol{G}_{jk} = \lambda_{jk}\boldsymbol{D}_r^T\boldsymbol{D}_r, \tag{2.24}$$

où $\boldsymbol{D}_r$ représente une matrice $(q_{jk} - r) \times q_{jk}$ des différences d'ordre r et est définie comme à la section 2.2.2. De plus, λ_{jk} correspond au paramètre de lissage pour le terme (j,k) de l'équation (2.22).

2.5.2 Mesurer les degrés de liberté effectifs (EDF)

Pour comparer différents modèles GAMLSS entre eux, le critère AIC, présenté à la section 2.3.1, s'avère une option intéressante en raison de sa simplicité. Pour pouvoir l'utiliser, il faut mesurer les degrés de liberté effectifs (EDF) des modèles à comparer.

La méthode qui sera présentée dans cette section est équivalente à celle utilisée dans l'extension *gamlss* du logiciel *R*. Elle est exacte lorsque le modèle contient au maximum 1 fonction de lissage, mais ne constitue qu'une approximation autrement (Rigby et Stasinopoulos, 2005).

Pour calculer les EDF d'un modèle, il faut d'abord définir la matrice diagonale des poids itératifs[13] $\boldsymbol{W}_{ks}$:

$$\boldsymbol{W}_{ks} = \frac{-\partial^2 \boldsymbol{\ell}}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\eta}_s^T} = -\text{diag}\left(\frac{\partial^2 \ell_i}{\partial \eta_{ik} \partial \eta_{is}}\right),$$

où k et s vont de 1 jusqu'au nombre de fonctions de lien $g(\cdot)$ employées dans le modèle et définies à l'équation (2.22). À ce propos, le vecteur $\boldsymbol{\eta}_k$ est aussi défini comme à cette équation et est composé des éléments η_{ik}, où i va de 1 jusqu'à n, le nombre de données. De plus, $\boldsymbol{\ell}$ correspond à un vecteur composé de la log-vraisemblance associée à chaque donnée, soit ℓ_i.

Ensuite, nous pouvons définir la matrice de lissage[14] $\boldsymbol{S}_{jk}$ correspondant au j^{e} terme de lissage de la fonction de lien $g_k(\cdot)$ de l'équation (2.22) :

$$\boldsymbol{S}_{jk} = \boldsymbol{Z}_{jk}(\boldsymbol{Z}_{jk}^T \boldsymbol{W}_{kk} \boldsymbol{Z}_{jk} + \boldsymbol{G}_{jk})^{-1} \boldsymbol{Z}_{jk}^T \boldsymbol{W}_{kk},$$

---

13. Traduction libre de *diagonal matrix of iterative weights.*

14. *Smoothing matrix.*

où $\boldsymbol{Z}_{jk}$ et $\boldsymbol{G}_{jk}$ sont définies pour des fonctions de lissage par P-splines comme à la section 2.5.1.

Pour trouver les EDF d'un terme de lissage (j, k), il suffit de calculer la trace de la matrice $\boldsymbol{S}_{jk}$ associée.

Les EDF d'un modèle sans fonction de lissage correspondent aux degrés de liberté traditionnels. Pour les calculer, il suffit d'additionner le nombre total de paramètres dans le modèle.

Lorsqu'un modèle compte une ou plusieurs fonctions de lissage, il est plus facile de calculer les EDF du modèle comme étant la somme des EDF associés à chaque fonction de lien $g_k(\cdot)$. Pour la portion paramétrique $\boldsymbol{X}_k\boldsymbol{\beta}_k$ de la fonction de lien $g_k(\cdot)$, les EDF correspondent à la longueur J'_k du vecteur $\boldsymbol{\beta}_k$. À ce nombre, il faut ajouter les EDF de chaque terme de lissage de la fonction de lien concernée et soustraire $(m_k - 1)$, où m_k représente le nombre de fonctions de lissage de la fonction de lien $g_k(\cdot)$.

Le fait de soustraire $(m_k - 1)$ aux EDF s'explique mieux avec un exemple. Supposons que la fonction de lien $g_k(\cdot)$ contient 2 fonctions de lissage. Autrement dit, $m_k = 2$. Lorsqu'on considère ces 2 fonctions ensemble, elles ont une certaine forme. Cependant, lorsque nous les prenons séparément, la première fonction pourrait être plus élevée et la seconde plus faible ou vice versa de telle sorte qu'elles ont toujours la même forme qu'initialement lorsqu'on considère leur effet commun. Soustraire $(m_k - 1)$ permet de retirer les degrés de liberté en trop qui ont été fournis au modèle. Dans cet exemple, il y a donc un degré de liberté qui doit être retiré.

Une fonction de lien $g_k(\cdot)$ comptant au moins une fonction de lissage n'a pas besoin d'une constante dans sa partie paramétrique $\boldsymbol{X}_k\boldsymbol{\beta}_k$ puisque la constante

se retrouve incluse (« avalée ») par la fonction de lissage.

L'extension *gamlss* fonctionne un peu autrement pour le calcul des EDF. Tout d'abord, il faut laisser une constante dans la partie paramétrique de la fonction de lien $g_k(\cdot)$ même lorsqu'il y a au moins une fonction de lissage. De plus, pour chaque fonction de lissage, un coefficient multiplicatif est estimé avec la fonction. Le calcul des EDF du modèle se fait également par l'addition des EDF associés à chaque fonction de lien $g_k(\cdot)$. Pour la portion paramétrique des fonctions de lien, les changements concernent l'ajout des coefficients multiplicatifs associés à chaque fonction de lissage ainsi que la constante. Concernant la portion non paramétrique, les EDF de chaque terme de lissage de la fonction de lien concernée sont ajoutés, puis on retire 2 degrés pour chaque terme de lissage de la fonction de lien, soit 1 degré pour la constante et un second pour le coefficient multiplicatif ajouté précédemment. Lorsqu'il y a m_k termes de lissage dans une fonction de lien, on se retrouve à enlever 1 degré pour la constante m_k fois, soit $m_k - 1$ fois de trop, d'où l'équivalence avec l'autre méthode énoncée plus tôt.

2.5.3 Estimation d'un GAMLSS avec des paramètres de lissage non fixés

Lorsque les paramètres de lissage ne sont pas connus, différentes méthodes existent pour les estimer.

Dans ce mémoire, nous allons nous concentrer sur la méthode par le AIC. Plus concrètement, il s'agit de trouver et sélectionner les paramètres de lissage qui minimisent le AIC d'un modèle.

Une façon sous-optimale d'y arriver est de créer une grille de valeurs pour les paramètres de lissage, puis d'estimer le modèle comme à la section 2.5.1, c'est-à-dire en maximisant la log-vraisemblance pénalisée pour estimer les paramètres. En calculant la log-vraisemblance du modèle obtenu ainsi que ses degrés de liberté

effectifs, nous avons alors son AIC. Il suffit ensuite de sélectionner le modèle avec le AIC minimal.

Cette méthode peut certainement être améliorée. L'utilisation d'algorithmes d'ajustement rétroactif[15] accélère la maximisation de la log-vraisemblance pénalisée, ce qui améliore nettement le temps total d'estimation (Rigby et Stasinopoulos, 2005).

15. Traduction libre de *backfitting algorithm.*

CHAPITRE III

APPLICATION À L'ASSURANCE AUTOMOBILE

Avec la popularité grandissante des appareils télématiques à bord des véhicules assurés, on peut recueillir davantage d'informations sur les assurés, notamment leur kilométrage exact parcouru. Une tarification basée conjointement sur la durée du contrat et sur le kilométrage devient alors possible à appliquer grâce à ces données additionnelles. Pour cela, les modèles GAMLSS deviennent très intéressants puisque, comme discuté précédemment, ils constituent une généralisation des modèles GAM qui eux-mêmes constituent une extension des modèles GLM, soit les 2 types de modèles principalement utilisés par les assureurs. Les modèles GAMLSS permettent donc de facilement comparer des modèles plus traditionnels en assurance, comme un modèle GLM avec la distribution Poisson, avec des modèles basés sur des distributions ne faisant pas partie de la famille exponentielle linéaire. De plus, avec les modèles GAMLSS, il est possible d'ajouter des fonctions de lissage, ce qui peut être intéressant à tester avec le kilométrage et la durée.

L'application de ces modèles a été réalisée avec le logiciel de statistiques *R*. À ce sujet, il est bon de savoir que Stasinopoulos *et al.* (2007) ont développé une extension nommée *gamlss* permettant d'ajuster rapidement ce type de modèles à l'aide d'un algorithme d'ajustement rétroactif[1]. D'ailleurs, ils ont publié un livre, Stasi-

1. Traduction libre de *backfitting algorithm*.

nopoulos *et al.* (2017), détaillant davantage leur extension. D'anciennes versions de ce livre, comme Rigby et Stasinopoulos (2009), sont disponibles gratuitement et fournissent tous les détails concernant les distributions incluses dans leur extension.

3.1 Données

Les données utilisées dans ce mémoire pour comparer différents modèles proviennent d'un assureur espagnol. Au total, il y a 129 116 observations et elles sont réparties à travers les années 2009 à 2011. Le tableau 3.1 présente la répartition de ces observations à travers les années, de même que le nombre de numéros distincts de contrat.

Tableau 3.1: Nombre d'observations et de numéros distincts de contrat selon les années considérées pous les données

	Année			Années regroupées
	2009	2010	2011	
Nombre d'observations	17 419	40 208	71 489	129 116
Nombre de contrats distincts	11 974	24 143	39 240	43 936

Il est intéressant de noter qu'en moyenne, un numéro de contrat est observé plus d'une fois par année. Ceci peut notamment s'expliquer par un renouvellement de l'assurance ou par la mise à jour d'informations liées au contrat. Sur le total de 129 116 données, on remarque 43 936 numéros distincts de contrat, soit environ 3 observations par numéro de contrat.

3.1.1 Types de réclamations

De plus, avec cette base de données, on remarque que les réclamations des assurés sont regroupées et classées selon leur type, c'est-à-dire selon qu'il s'agisse de dommages corporels ou matériels et si la responsabilité de l'accident est attribuée à l'assuré ou non. Le tableau 3.2 présente ces 4 types.

Tableau 3.2: Types de réclamations de la base de données

Type de réclamations	Description
$nb1$	Dommages matériels avec responsabilité
$nb2$	Dommages matériels sans responsabilité
$nb3$	Dommages corporels avec responsabilité
$nb4$	Dommages corporels sans responsabilité

Dans ce mémoire, nous allons nous concentrer sur les dommages matériels non-responsables $nb2$. Cette décision a été prise par souci de continuité avec le travail de Boucher *et al.* (2017), mais il est à noter que nous aurions également pu utiliser $nb1$ [2].

3.1.2 Informations sur les assurés

En plus des types de réclamations, la base de données contient des informations sur le kilométrage exact parcouru, la durée du contrat d'assurance, l'âge de l'assuré, l'âge du véhicule, le sexe de l'assuré et le type de stationnement utilisé par l'assuré. Les tableaux 3.3 et 3.4 présentent respectivement les statistiques des variables quantitatives et la répartition des assurés à travers les variables qualitatives de la base de données.

2. Les modèles utilisés plus loin dans cette section pour la modélisation de $nb2$ ont aussi été appliqués pour $nb1$ à des fins comparatives et des résultats similaires ont été observés.

Tableau 3.3: Statistiques descriptives des variables quantitatives de la base de données

Variable	Moy.	Écart-type	Min.	Max.	Quantiles		
					25%	50%	75%
nb2	0,07	0,28	0,00	4,00	0,00	0,00	0,00
Durée	0,66	0,32	0,003	1,00	0,40	0,70	1,00
Kilométrage	6155,18	5723,44	0,01	75 014,42	2065,70	4645,99	8465,08
Âge	25,48	3,16	18,00	37,00	23,00	25,00	28,00
Âge du véhicule	7,42	4,54	0,00	34,00	4,00	7,00	11,00

Tableau 3.4: Répartition des assurés selon leur sexe et leur type de stationnement

	Sexe		Type de stationnement	
	Homme	Femme	Garage privé	Voie publique
Nombre d'observations	69 829	59 287	97 188	31 928
Pourcentage (%)	54,08	45,92	75,27	24,73

On remarque tout d'abord qu'il y a un nombre maximal de 4 réclamations pour *nb2* et un nombre moyen de réclamations de 0,07. Concernant le kilométrage, 75% des assurés parcourent moins que 8500 km et le kilométrage moyen est de 6155 km. Toutefois, il faut demeurer prudent avec ces nombres, puisqu'ils englobent des assurés couverts pour moins d'une année. À ce propos, la durée est calculée comme étant la fraction du nombre de jours du contrat d'assurance sur une année complète, basée sur 365 jours. La durée minimale correspond à 1 jour assuré et la durée moyenne est de 0,66 année, ce qui correspond à environ 8 mois.

L'analyse des statistiques descriptives est particulièrement importante pour la variable de l'âge des assurés. En effet, on constate que l'intervalle d'âge est compris

entre 18 et 37 ans et que 75% des assurés ont 28 ans ou moins. À ce sujet, on peut présumer que le produit d'assurance relatif à cette base de données visait une clientèle plus jeune, soit une clientèle qui paye typiquement une prime d'assurance élevée et pour laquelle un rabais potentiel lié aux appareils télématiques pourrait être plus attrayant. De plus, l'âge du véhicule, relativement élevé avec une moyenne de 7,42 ans, concorde avec le fait d'avoir une clientèle plus jeune.

Du côté des variables qualitatives, on remarque qu'il y a un peu plus d'hommes que de femmes dans la base de données, à 54% contre 46%, et qu'environ 3 assurés sur 4 stationnent leur véhicule dans un garage privé plutôt que sur les voies publiques.

Une étude plus précise du nombre de réclamations de type $nb2$, noté Y_{nb2}, permet de constater que cette variable présente de la surdispersion, puisque sa variance est plus élevée que son espérance :

$$E[Y_{nb2}] = 0{,}07418136,$$
$$Var[Y_{nb2}] = 0{,}07805051,$$
$$Var[Y_{nb2}] = 1{,}052158 \times E[Y_{nb2}].$$

3.1.3 Influence du kilométrage et de la durée sur la fréquence de réclamations

Comme nous nous intéressons particulièrement à la modélisation de la fréquence de réclamations par le kilométrage et la durée, une analyse plus approfondie de ces variables est de mise. Le tableau 3.5 détaille la répartition du nombre de réclamations pour des dommages matériels de type non-responsable, soit $nb2$.

Concernant le kilométrage et la durée, les figures 3.1 et 3.2 présentent la répartition du nombre d'assurés par classe de kilométrage et de durée. Les classes sont d'une longueur de 250 km et de 0,02 année respectivement. Pour le kilométrage, on peut constater qu'une grande proportion des assurés parcourent moins de 30 000 km sur une année de contrat et que, de façon générale, plus le kilométrage est important,

Tableau 3.5: Répartition du nombre de réclamations pour des dommages matériels non-responsables

Nombre de réclamations	Type $nb2$	
	Fréquence	Pourcentage (%)
0	120 114	93,0280
1	8454	6,5476
2	521	0,4035
3	26	0,0201
4	1	0,0008
5 et plus	0	0
Total	129 116	100

moins il y a d'assurés dans cette classe. Quant à la durée, le tiers des assurés se retrouvent dans la classe la plus élevée, c'est-à-dire avec une durée allant de 0,98 à 1 an. Les deux tiers restants des assurés ont une répartition relativement uniforme à travers les autres classes.

Afin d'avoir une idée plus précise de l'impact du kilométrage sur la fréquence de réclamations, la figure 3.3 présente le nombre moyen de réclamations de type $nb2$ selon le kilométrage. Pour obtenir cette figure, les assurés ont été regroupés dans des classes construites par intervalle de 500 km. En raison du peu de données disponibles pour les kilométrages élevés, la figure ne va pas au-delà de 30 000 km. Il est toutefois important de garder en tête que le nombre d'assurés varie d'une classe à l'autre. La figure 3.3 permet de constater que le kilométrage ne semble pas avoir une relation linéaire avec la fréquence de réclamations. Pour des petites valeurs, une hausse du kilométrage fait augmenter rapidement la fréquence de réclamations, mais cet effet semble s'estomper plus le kilométrage augmente.

Figure 3.1: Répartition du nombre d'assurés par tranche de kilométrage

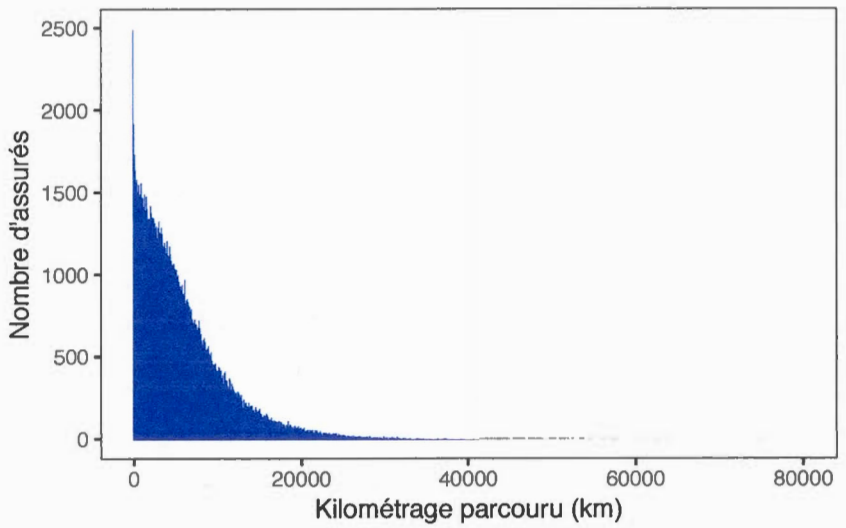


Figure 3.2: Répartition du nombre d'assurés par tranche de durée de contrat

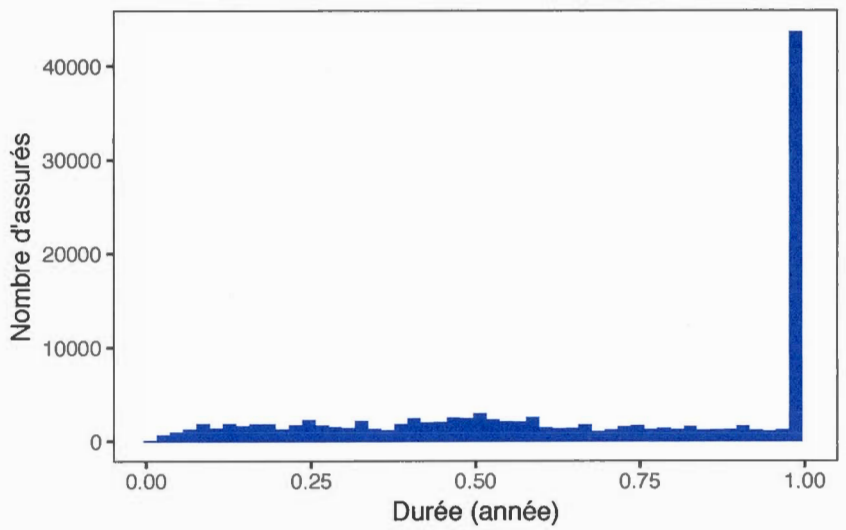
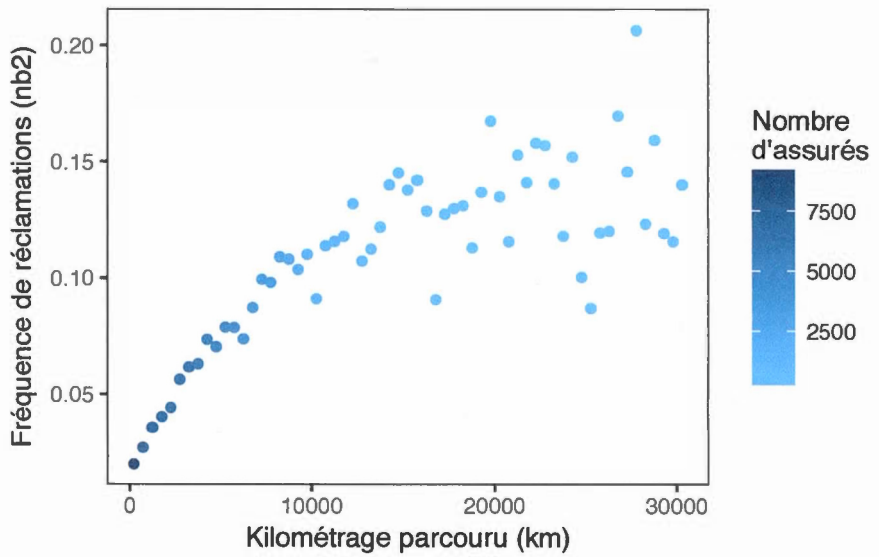


Figure 3.3: Fréquence de réclamations pour $nb2$ selon la classe de kilométrage parcouru

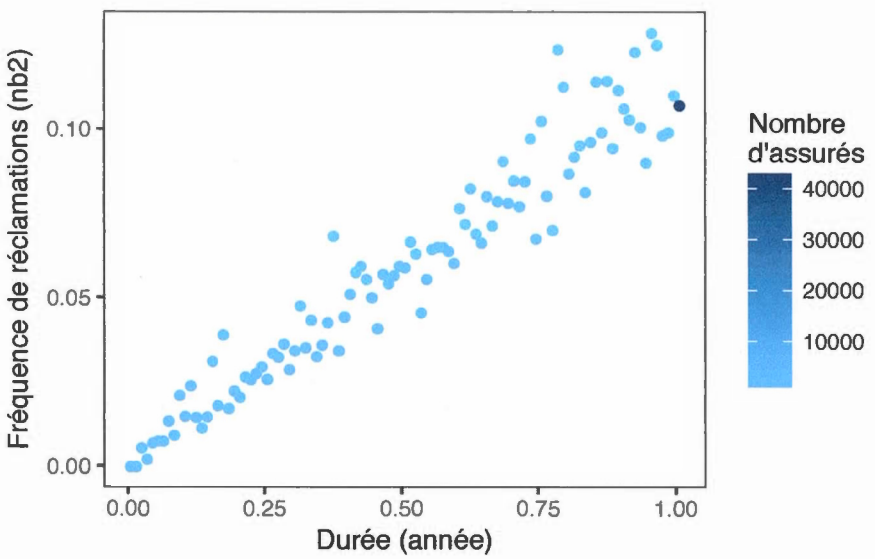


Un exercice similaire a également été réalisé pour visualiser l'impact de la durée sur la fréquence de réclamations à la figure 3.4. Les assurés ont cette fois été regroupés par des classes d'une longueur de 0,01 année, puis le nombre moyen de réclamations par classe a été calculé. À l'exception de la dernière classe, couvrant la période de 0,99 à 1 an, le nombre d'assurés est relativement uniforme d'une classe à l'autre. On constate en analysant la figure 3.4 qu'il semble y avoir une relation proportionnelle entre la durée et la fréquence de réclamations.

3.1.4 Traitement des données

Afin d'incorporer les informations fournies par la base de données sur les assurés à différents modèles, il faut choisir comment les traiter. Dans ce mémoire, il a été décidé que l'âge des assurés ainsi que l'âge des véhicules soient transformés en variables binaires par la façon présentée au tableau 3.6.

Figure 3.4: Fréquence de réclamations pour *nb2* selon la classe de durée de contrat



De plus, 3 données ont été jugées aberrantes et retirées. Il s'agit de données où il y avait davantage de réclamations de type *nb2* que de kilométrage parcouru avec au moins 250 jours de durée de contrat.

Par ailleurs, puisqu'il n'y a que très peu de données avec un kilométrage supérieur à 30 000 km (voir figure 3.1), il aurait été difficile d'obtenir des résultats significatifs pour des kilométrages aussi élevés. Il a donc été décidé que les 736 données concernées soient retirées de la modélisation.

Au total, 739 données ont ainsi été rejetées, ce qui correspond à 0,57% des données totales. Il reste donc 128 377 données.

Également, afin de pouvoir valider les modèles qui seront générés, il est préférable de mettre de côté une partie des données. De cette façon, nous pourrons valider la qualité de ces modèles avec des observations n'ayant pas servi à leur estimation. Pour cela, nous avons décidé d'utiliser 2 tiers des données pour l'estimation et le tiers restant pour la validation. Les données de validation ont été obtenues en

Tableau 3.6: Représentation des variables explicatives en variables binaires

Variable	Valeur
x_1	Vaut 1 (constante)
x_2	Vaut 1 si l'assuré a 25 ans ou moins
x_3	Vaut 1 si l'assuré a plus de 25 ans et au plus 30 ans
x_4	Vaut 1 si le véhicule a 2 ans ou moins
x_5	Vaut 1 si le véhicule a plus de 2 ans et au plus 5 ans
x_6	Vaut 1 si le véhicule a plus de 5 ans et au plus 10 ans
x_7	Vaut 1 si l'assuré est un homme
x_8	Vaut 1 si l'assuré a un garage privé

pigeant aléatoirement 14 634 numéros de contrat, correspondant exactement au tiers du nombre de numéros de contrat. En procédant ainsi, nous nous retrouvons avec 42 799 observations pour la validation. Les 85 578 données restantes serviront donc à l'estimation.

3.2 Modèles classiques

En assurance automobile, la façon classique de modéliser la fréquence de réclamations se traduit par l'utilisation d'un modèle GLM avec la distribution Poisson. Des variables explicatives sont utilisées et liées au paramètre de moyenne λ par l'entremise de la fonction logarithmique afin de mieux ajuster la fréquence au risque représenté par chaque assuré. De plus, la durée du contrat d'assurance est considérée proportionnelle à la fréquence. Cette hypothèse est en lien avec la théorie du processus de Poisson, soit un cas particulier de processus de comptage, où l'espérance est directement proportionnelle au temps d'exposition. Ainsi, les assureurs ajustent typiquement la fréquence au prorata de la durée, calculée habituellement en fraction d'année.

La fonction de lien $g(\cdot)$ d'un tel modèle s'exprime alors comme ceci :

$$g(\lambda_i) = \log(\lambda_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \log(d_i), \tag{3.1}$$

où λ_i est le paramètre de moyenne lié à l'assuré i, $\boldsymbol{X}_i$ est la i^{e} ligne de la matrice $\boldsymbol{X}$ regroupant les variables explicatives connues des assurés, $\boldsymbol{\beta}$ est un vecteur de coefficients et d_i est la durée du contrat du i^{e} assuré.

Afin de servir de référence, ce modèle a été appliqué sur les données. Les variables explicatives utilisées pour la matrice $\boldsymbol{X}$ sont celles présentées au tableau 3.6. À ce sujet, il a été déterminé que toutes ces variables sont significatives à l'intérieur du modèle à un seuil de 5%.

En variant la distribution de probabilités utilisée, des modèles GAMLSS similaires peuvent également être appliqués. Pour cela, les distributions binomiales négatives de type 1 et 2, Poisson inverse-gaussienne ainsi que Poisson gonflée à zéro présentées à la section 2.4 constituent de bonnes options à considérer. Afin qu'elles soient plus facilement comparables à la Poisson, il suffit de reprendre la fonction de lien $g(\lambda_i)$ du paramètre de la moyenne présentée à l'équation (3.1), soit une fonction de lien logarithmique, et de ne pas appliquer de fonction de lien sur le second paramètre de ces distributions. La distribution MVNB peut aussi être considérée, mais avec une légère modification à la fonction de lien de l'équation (3.1) pour tenir compte des données longitudinales qu'elle admet :

$$g(\lambda_{i,t}) = \log(\lambda_{i,t}) = \boldsymbol{X}_{i,t}\boldsymbol{\beta} + \log(d_{i,t}), \tag{3.2}$$

où les indices (i,t) correspondent à la t^{e} observation de l'assuré i.

L'intérêt pour la distribution MVNB repose sur l'ajout d'une dépendance entre les observations d'un même assuré qui n'est pas présente avec les autres distributions.

Comme il a été vu à la section 3.1.2, le nombre de réclamations de type $nb2$ qu'on tente de modéliser est surdispersé, comme c'est le cas fréquemment en assurance. Utiliser une distribution autre que la Poisson parmi celles mentionnées plus tôt présente comme principal avantage sur la Poisson le fait d'admettre de la surdispersion plutôt que de l'équidispersion.

Dans un souci de clarté, les modèles de cette section seront référés comme étant les modèles classiques.

Le tableau 3.7 compare la qualité de ces modèles selon le critère AIC. Tous les modèles à données transversales performent de manière similaire, à l'exception du modèle Poisson, largement en retrait avec un AIC de 44 356,83. De ce groupe, le modèle avec la distribution binomiale négative de type 2 est le meilleur selon ce critère avec un AIC de 44 317,57. Toutefois, si la distribution MVNB à données longitudinales est aussi considérée, c'est alors le modèle avec cette distribution qui surpasse les autres avec un AIC de 44 271,10.

Tableau 3.7: Comparaison de la qualité des modèles classiques selon le critère AIC

Modèle	AIC
Poisson	44 356,83
Binomiale négative de type 2	**44 317,57**
Binomiale négative de type 1	44 324,04
Poisson inverse-gaussienne	44 317,95
Poisson gonflé à zéro	44 317,72
MVNB	**44 271,10**

Une autre façon de mesurer la qualité globale des modèles repose sur la comparaison de la répartition du nombre de réclamations de chaque modèle avec le nombre observé de réclamations.

En utilisant les données d'estimations, on remarque au tableau 3.8 que le modèle Poisson est celui qui performe le moins bien. En effet, ce modèle est celui dont les prédictions s'éloignent le plus de ce qui est observé. Tous les autres modèles ont des prédictions qui s'approchent beaucoup des observations. Du lot, le modèle Poisson inverse-gaussienne est celui qui se démarque le plus par sa précision.

Tableau 3.8: Comparaison de la prédiction du nombre de réclamations des modèles classiques avec les données d'estimation

Modèle	Nombre de réclamations					
	0	1	2	3	4	≥ 5
Poisson	79 797,18	5727,05	261,60	8,91	0,24	0,01
Bin. Nég. type 2	79 899,56	5532,10	341,81	20,29	1,18	0,07
Bin. Nég. type 1	79 888,02	5555,45	332,98	17,65	0,86	0,04
Poisson inv.-gaus.	79 897,99	5536,32	338,31	20,93	1,36	0,10
Poisson gonflé à 0	79 900,75	5526,62	350,38	16,59	0,63	0,02
MVNB	79 875,59	5568,17	331,67	18,51	1,00	0,06
Observé	79 893	5545	339	17	1	0

En répétant le même exercice avec les données de validation cette fois, les conclusions changent quelque peu. Le modèle MVNB est celui dont les prédictions sont les plus près de la réalité, et ce, de façon nette sur les autres modèles. Comme plus tôt, le modèle Poisson est celui dont les prédictions s'éloignent le plus des observations.

Tableau 3.9: Comparaison de la prédiction du nombre de réclamations des modèles classiques avec les données de validation

Modèle	Nombre de réclamations					
	0	1	2	3	4	≥ 5
Poisson	39 371,54	3053,90	150,82	5,57	0,17	0
Bin. Nég. type 2	39 434,09	2939,17	195,40	12,51	0,79	0,04
Bin. Nég. type 1	39 423,24	2959,65	188,06	10,49	0,54	0,02
Poisson inv.-gaus.	39 441,53	2934,21	192,48	12,81	0,90	0,07
Poisson gonflé à 0	39 443,50	2927,77	200,04	10,25	0,42	0,02
MVNB	39 645,39	2762,78	164,18	9,13	0,49	0,03
Observé	39 585	2820	172	5	0	0

3.3 Modèles classiques avec kilométrage

Une manière d'améliorer les modèles réalisés à la section 3.2 passe par l'ajout du kilométrage comme variable explicative. De façon classique, les assureurs s'en remettent à l'estimation fournie en début de contrat par les assurés. Toutefois, comme il a été expliqué plus tôt, cette estimation est souvent fausse et il n'existe pas vraiment de mécanisme de validation de cette information chez les assureurs. Ceux-ci doivent donc être prudents en traitant le kilométrage, puisque cette variable n'est pas mesurée de façon fiable.

Dans ce contexte, l'intégration du kilométrage aux modèles classiques peut se faire en réutilisant sensiblement les mêmes fonctions de lien que celles de la section 3.2, soit pour les distributions à données transversales

$$g(\lambda_i) = \log(\lambda_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \log(d_i),$$

et pour la distribution MVNB

$$g(\lambda_{i,t}) = \log(\lambda_{i,t}) = \boldsymbol{X}_{i,t}\boldsymbol{\beta} + \log(d_{i,t}),$$

mais en modifiant la matrice $\boldsymbol{X}$ des variables explicatives pour ajouter le kilométrage.

Dans ce mémoire, les données utilisées contiennent le kilométrage exact parcouru par les assurés. La situation est quelque peu différente de celle des assureurs puisque l'information disponible sur le kilométrage est fiable dans notre cas. Néanmoins, ajouter le kilométrage aux modèles précédents en le traitant par la façon traditionnelle en assurance permet d'avoir une bonne idée des pratiques actuelles de l'industrie.

Comme les données des assureurs sur le kilométrage ne sont ni fiables, ni précises habituellement, le kilométrage est la plupart du temps transformé en variables

binaires. Le tableau 3.10 présente une façon de procéder. Les intervalles sélectionnés sont plutôt grands à 5000 km justement en raison du manque de précision des données. À ce propos, il est difficile pour les assureurs de savoir si la différence de kilométrage entre un assuré qui parcourt 15 000 km et un autre qui en parcourt 18 000 est significative, puisque ces nombres proviennent d'estimations.

Ainsi, les variables explicatives x_9 à x_{13} sont ajoutées à la matrice $\boldsymbol{X}$ contenant déjà les variables explicatives x_1 à x_8 du tableau 3.6.

Tableau 3.10: Représentation en variables explicatives binaires du kilométrage

Variable	Valeur
x_9	Vaut 1 si $km < 5000$
x_{10}	Vaut 1 si $5000 \leq km < 10\,000$
x_{11}	Vaut 1 si $10\,000 \leq km < 15\,000$
x_{12}	Vaut 1 si $15\,000 \leq km < 20\,000$
x_{13}	Vaut 1 si $20\,000 \leq km < 25\,000$

Implicitement, l'utilisation des variables binaires x_9 à x_{13} fait en sorte que la classe de référence pour le kilométrage est constituée des assurés parcourant au moins 25 000 km.

Sur l'ensemble des variables, seules x_1 à x_{10} s'avèrent significatives dans un modèle GLM à distribution Poisson avec fonction de lien logarithmique. Les autres modèles de cette section vont également recourir à ces 10 variables explicatives afin qu'ils soient plus facilement comparables entre eux, ce qui signifie que les variables x_{11} à x_{13} sont abandonnées. Par conséquent, la classe de référence pour le kilométrage est désormais constituée des assurés parcourant au moins 10 000 km.

Le tableau 3.11 compare par le critère AIC la qualité d'ajustement des modèles

de la section 3.2 en leur ajoutant les variables explicatives x_9 et x_{10} basées sur le kilométrage. Tous ces nouveaux modèles, auxquels nous référerons comme étant les modèles classiques avec kilométrage, présentent un AIC plus faible (et donc meilleur) que ceux de la section précédente présentés au tableau 3.7.

Parmi les modèles classiques avec kilométrage, le modèle Poisson est celui dont le AIC est le plus élevé. Les autres modèles à données transversales présentent une meilleure qualité d'ajustement selon ce critère. De ce groupe, le modèle avec la distribution binomiale négative de type 2 a un très léger avantage sur les modèles avec la distribution Poisson inverse-gaussienne et Poisson gonflée à zéro. Toutefois, ces modèles sont éclipsés par celui avec la distribution MVNB qui a un AIC nettement inférieur.

Tableau 3.11: Comparaison de la qualité des modèles classiques selon le critère AIC en ajoutant le kilométrage

Modèle	AIC
Poisson	44 257,02
Binomiale négative de type 2	**44 222,23**
Binomiale négative de type 1	44 227,20
Poisson inverse-gaussienne	44 222,49
Poisson gonflé à zéro	44 222,65
MVNB	**44 175,90**

Comme précédemment, on peut aussi comparer les prévisions du nombre de réclamations de chaque modèle. Au tableau 3.12, on constate avec les données d'estimation que le modèle ayant la distribution Poisson est celui dont les prévisions s'éloignent le plus des nombres observés pour chaque niveau de réclamation. Pour les autres modèles, les prévisions sont très semblables. Du lot, on note que les modèles avec les distributions binomiale négative de type 1 et MVNB sont légèrement

plus près des nombres observés.

Ces résultats changent lorsque les données de validation sont utilisées. Au tableau 3.13, on s'aperçoit que le modèle avec la distribution MVNB est désormais plus précis que les autres modèles concernant les nombres d'assurés pour chaque niveau de réclamation. Sans surprise, c'est encore le modèle avec la distribution Poisson qui s'avère le plus imprécis.

Tableau 3.12: Comparaison de la prédiction du nombre de réclamations des modèles classiques avec kilométrage avec les données d'estimation

Modèle	Nombre de réclamations					
	0	1	2	3	4	≥ 5
Poisson	79 809,99	5702,54	272,27	9,90	0,29	0,01
Bin. Nég. type 2	79 906,61	5516,55	349,05	21,41	1,29	0,09
Bin. Nég. type 1	79 895,02	5542,14	338,63	18,26	0,90	0,05
Poisson inv.-gaus.	79 905,31	5520,16	345,90	22,04	1,47	0,12
Poisson gonflé à 0	79 907,38	5509,24	359,71	17,92	0,72	0,03
MVNB	79 885,99	5545,97	341,79	20,05	1,14	0,06
Observé	79 893	5545	339	17	1	0

Ces modèles semblent mieux que ceux de la section 3.2. Toutefois, la problématique avec les modèles classiques avec kilométrage repose sur les assurés qui se trouvent près des limites d'intervalles de kilométrage. Comme les assureurs se basent habituellement sur les estimations de kilométrage des assurés, ceux-ci peuvent être surchargés ou, à l'inverse, obtenir un rabais parce qu'ils se retrouvent dans un certain intervalle de kilométrage. Toutefois, à la fin de l'année, ils peuvent avoir parcouru la même distance totale que des assurés d'une classe voisine. Les assurés ont donc un plus grand incitatif à mentir pour économiser.

Tableau 3.13: Comparaison de la prédiction du nombre de réclamations des modèles classiques avec kilométrage avec les données de validation

Modèle	Nombre de réclamations					
	0	1	2	3	4	≥ 5
Poisson	39 327,57	3094,15	154,34	5,75	0,17	0,02
Bin. Nég. type 2	39 389,97	2982,21	196,67	12,34	0,76	0,05
Bin. Nég. type 1	39 395,48	2987,78	187,91	10,29	0,51	0,03
Poisson inv.-gaus.	39 419,53	2957,61	191,56	12,40	0,84	0,06
Poisson gonflé à 0	39 421,27	2950,90	199,31	10,09	0,41	0,02
MVNB	39 651,68	2750,84	169,02	9,88	0,56	0,02
Observé	39 585	2820	172	5	0	0

3.4 Modèles avancés

Lorsque le kilométrage parcouru par les assurés est mesuré par l'entremise d'un appareil télématique, l'utilisation de cette variable dans la modélisation de la fréquence de réclamations devient plus intéressante puisque les données sur les distances parcourues sont dorénavant précises et fiables. Avec la qualité accrue des données concernant cette variable, des modèles plus avancés, comme des GAM ou des GAMLSS avec une fonction de lissage pour le kilométrage, s'avèrent alors plus attrayants. En effet, une fonction de lissage permet à une variable de « s'exprimer » plus librement qu'une relation linéaire ou qu'une fonction à paliers (voir le tableau 3.10 de la section 3.3).

Dans cette section, les modèles considérés sont des GAMLSS avec une fonction de lissage par P-splines pour le kilométrage. De plus, l'hypothèse de proportionnalité entre la fréquence et la durée est assouplie au profit d'une seconde fonction de lissage par P-splines pour la durée. Cette seconde fonction permet davantage de flexibilité pour l'effet de la durée sur la fréquence qui, selon les résultats de Boucher *et al.* (2017), n'auraient pas une relation proportionnelle. Par souci de clarté, les modèles de cette section seront appelés les modèles avancés.

Puisque la spline cubique d'ajustement minimise le critère (2.5), il a été décidé de travailler avec des splines cubiques pour les 2 fonctions de lissage. Ainsi, pour reprendre la notation résumée au tableau 2.1, nous avons $d = 3$. De plus, les domaines du kilométrage et de la durée sont divisés en $m = 35$ intervalles. Ce nombre a été choisi de façon à fournir une borne supérieure pour les degrés de liberté suffisamment grande (voir la section 2.2.2), mais il aurait pu être différent sans vraiment impacter les résultats. Au total, il y a donc 38 fonctions B-splines qui vont composer la fonction de lissage par P-splines du kilométrage et 38 autres pour celle de la durée.

Concernant la pénalité d'ordre r relative aux P-splines, le choix a été fait de tester 2 possibilités, soit des pénalités d'ordre 2 ou d'ordre 3. À ce propos, il ne semble pas y avoir une raison de préférer un ordre à un autre dans la littérature scientifique.

Finalement, les paramètres de lissage sont estimés par la minimisation du critère AIC présenté à la section 2.3.1.

3.4.1 Distributions à données transversales

Pour ces modèles avancés, nous allons d'abord nous concentrer sur les distributions à données transversales, c'est-à-dire en excluant la distribution MVNB. Nous y reviendrons à la prochaine sous-section.

Pour les modèles avancés, la fonction de lien du paramètre de la moyenne λ_i est semblable à celle des modèles classiques présentée par l'équation (3.1). Il s'agit encore d'une fonction logarithmique. Toutefois, comme mentionné précédemment, une fonction de lissage par P-splines, $f_1(\cdot)$, est ajoutée pour le kilométrage. De plus, la durée se retrouve désormais traitée par une seconde fonction de lissage par P-splines, $f_2(\cdot)$.

Nous avons ainsi comme fonction de lien :

$$g(\lambda_i) = \log(\lambda_i) = \boldsymbol{X}_i\boldsymbol{\beta} + f_1(km_i) + f_2(d_i), \tag{3.3}$$

où km_i et d_i représentent respectivement le kilométrage et la durée d'exposition de l'assuré i, $\boldsymbol{\beta}$ est un vecteur de coefficients de longueur 8 et $\boldsymbol{X}_i$ correspond à la ligne associée au i^{e} assuré de la matrice des variables explicatives. Cette matrice, de dimension $n \times 8$, regroupe les valeurs des 8 variables binaires présentées au tableau 3.6, soit les mêmes que pour les modèles classiques, pour l'ensemble des n assurés.

Le tableau 3.14 présente le AIC des modèles avancés pour les distributions à données transversales selon des pénalités d'ordre 2 ou 3 pour les fonctions de lissage par P-splines. On y remarque que de passer d'un ordre 2 de pénalité à un ordre 3 fait augmenter légèrement le AIC pour les modèles à distribution Poisson et Poisson inverse-gaussienne. Pour les autres modèles, on note une petite diminution du AIC. De plus, autant pour un ordre 2 de pénalité que pour un ordre 3, on constate que tous les modèles ayant une distribution admettant de la surdispersion surpassent nettement le modèle à distribution Poisson selon le AIC. De ce groupe, c'est le modèle à distribution binomiale négative de type 2 qui possède le plus faible AIC pour les 2 ordres de pénalité considérés.

Tableau 3.14: Comparaison selon le critère AIC de la qualité des modèles avancés avec une pénalité d'ordre k pour les fonctions de lissage par P-splines

Modèle	AIC pour $k = 2$	AIC pour $k = 3$
Poisson	44 129,40	44 131,82
Binomiale négative de type 2	**44 097,94**	**44 095,33**
Binomiale négative de type 1	44 099,48	44 096,50
Poisson inverse-gaussienne	44 098,06	44 099,42
Poisson gonflé à zéro	44 098,21	44 096,39

En analysant les fonctions de lissage obtenues, présentées par les figures 3.5 à 3.14, on remarque qu'elles ont des tendances très semblables, et ce, peu importe la distribution et l'ordre de pénalité retenus.

Pour le kilométrage, on remarque que les fonctions augmentent rapidement et presque linéairement pour les 7500 premiers kilomètres, puis croissent plus lentement jusqu'à 10 000 km pour se stabiliser ensuite. Pour des kilométrages très élevés de plus de 25 000 km, la zone ombragée, correspondant à l'écart-type associé aux fonctions, devient très grande. Il faut donc demeurer prudent avec la

tendance des fonctions pour ces valeurs élevées de kilométrage.

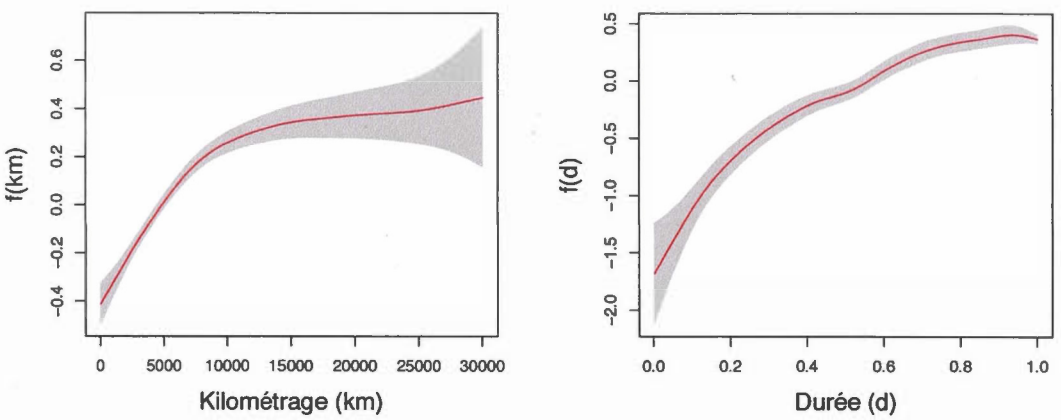
De façon générale, les fonctions de lissage obtenues pour la durée croissent rapidement et de façon plutôt linéaire entre 0 et 0,2 année, puis plus lentement de 0,2 à 0,8 année, mais toujours avec une tendance plutôt linéaire, sauf vers 0,5 année où il y a un petit creux. Finalement, de 0,8 à 1 an, les fonctions semblent atteindre un plateau.

Pour obtenir l'effet direct de la durée sur la fréquence, il faut prendre l'exponentielle des fonctions de lissage de la durée en raison du lien logarithmique. En faisant cela, le résultat global est que nous n'avons pas des fonctions qui indiquent une proportionnalité entre la fréquence et la durée d'exposition à cause du petit creux à 0,5 année et du plateau entre 0,8 et 1 an. Par conséquent, comme l'ont constaté Boucher *et al.* (2017), l'industrie se base sur une fausse hypothèse en présumant cette relation proportionnelle.

Les fonctions de lissage obtenues semblent plus stables en utilisant une pénalité d'ordre 2. En effet, celles du modèle avancé Poisson illustrées à la figure 3.5 deviennent plus ondulées avec une pénalité d'ordre 3. Un phénomène semblable se produit aussi pour les fonctions de lissage de la durée pour les modèles avancés avec les distributions Poisson inverse-gaussienne et Poisson gonflée à 0.

Néanmoins, en combinant l'effet des fonctions de lissage du kilométrage et de la durée aux figures 3.15 à 3.19, on réalise que ces ondulations supplémentaires n'ont pas un effet marqué sur les surfaces engendrées. En effet, celles-ci ont toutes une forme plutôt similaire. Ces figures sont obtenues en présumant des valeurs de 0 pour les variables explicatives x_2 à x_8 (se référer au tableau 3.6 ainsi qu'à l'équation 3.3), c'est-à-dire en utilisant uniquement la constante, le kilométrage et la durée pour calculer la fréquence.

Figure 3.5: Fonctions de lissage du kilométrage et de la durée avec la pénalité d'ordre 2 pour le modèle Poisson

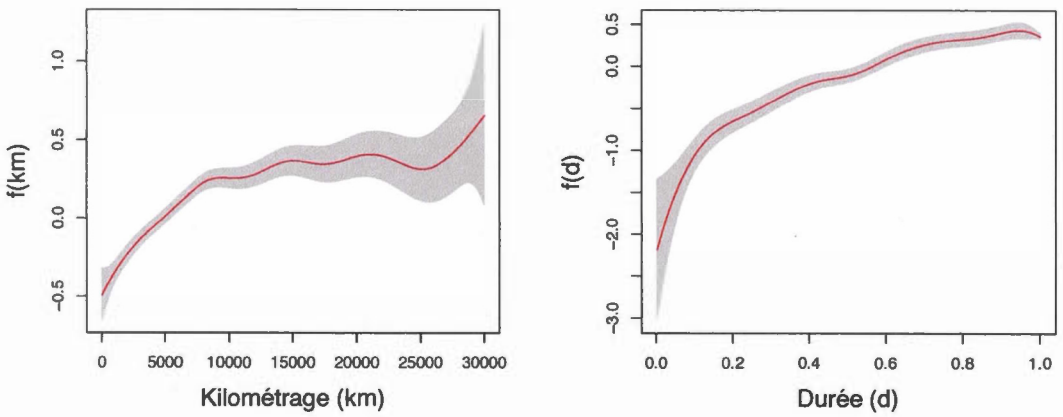


(a) Kilométrage

(b) Durée

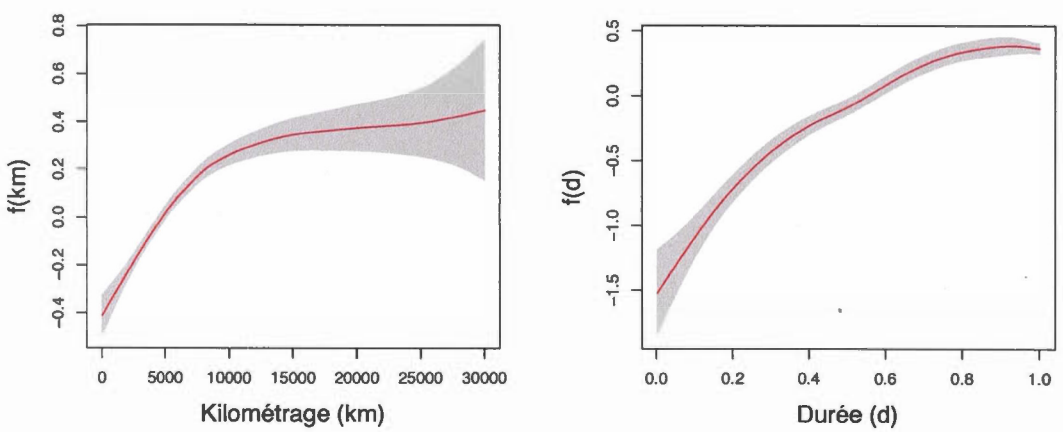
Le creux vers 0,5 année des fonctions de lissage de la durée est bien visible sur la plupart des figures 3.15 à 3.19. Il pourrait s'agir d'un léger surajustement puisque le modèle avancé Poisson avec pénalité d'ordre 2 et les modèles avancés avec la distribution binomiale négative de type 2 ne semblent pas en être particulièrement affecté.

Figure 3.6: Fonctions de lissage du kilométrage et de la durée avec la pénalité d'ordre 3 pour le modèle Poisson



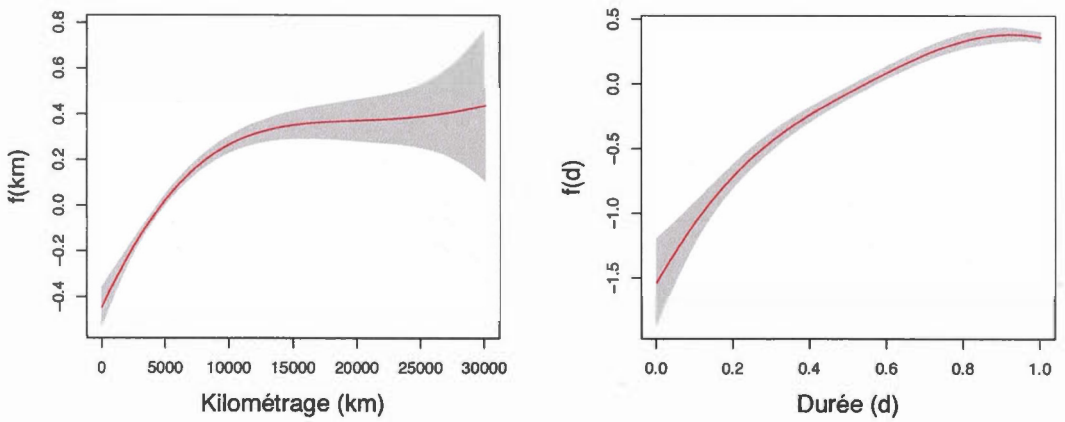
(a) Kilométrage (b) Durée

Figure 3.7: Fonctions de lissage du kilométrage et de la durée avec la pénalité d'ordre 2 pour le modèle avec la distribution binomiale négative de type 2



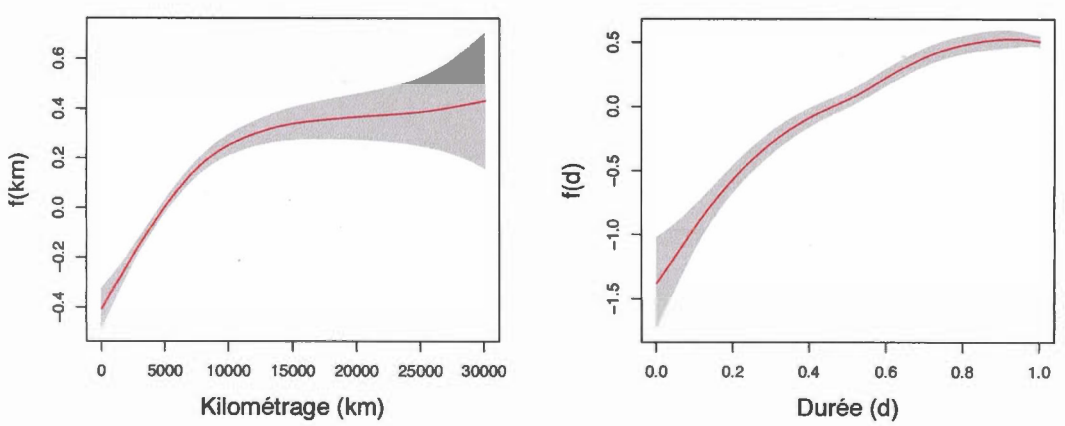
(a) Kilométrage (b) Durée

Figure 3.8: Fonctions de lissage du kilométrage et de la durée avec la pénalité d'ordre 3 pour le modèle avec la distribution binomiale négative de type 2



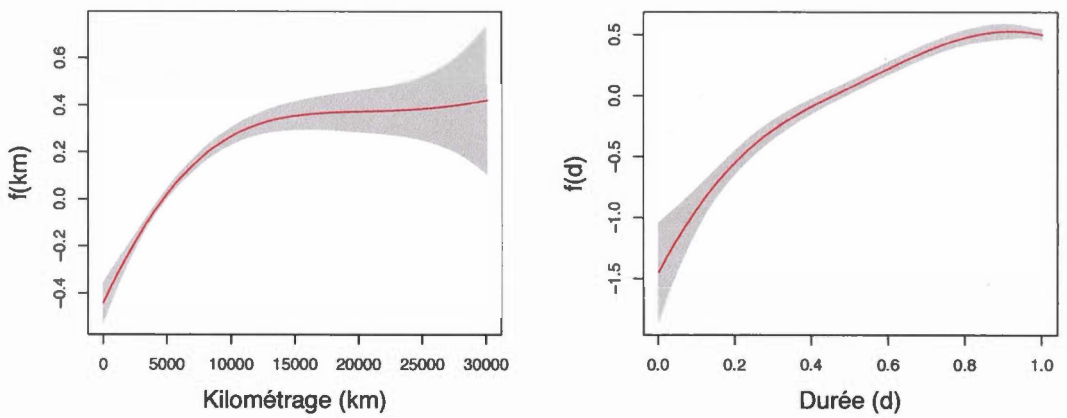
(a) Kilométrage (b) Durée

Figure 3.9: Fonctions de lissage du kilométrage et de la durée avec la pénalité d'ordre 2 pour le modèle avec la distribution binomiale négative de type 1



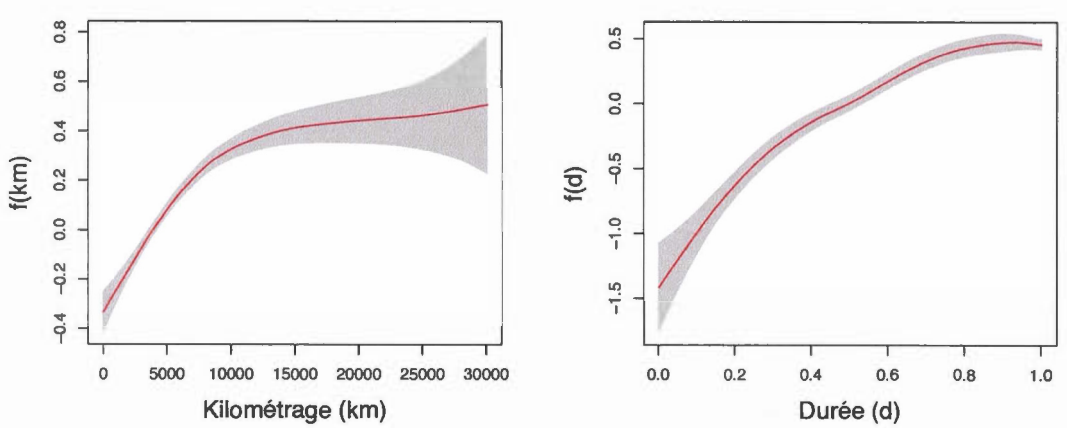
(a) Kilométrage (b) Durée

Figure 3.10: Fonctions de lissage du kilométrage et de la durée avec la pénalité d'ordre 3 pour le modèle avec la distribution binomiale négative de type 1



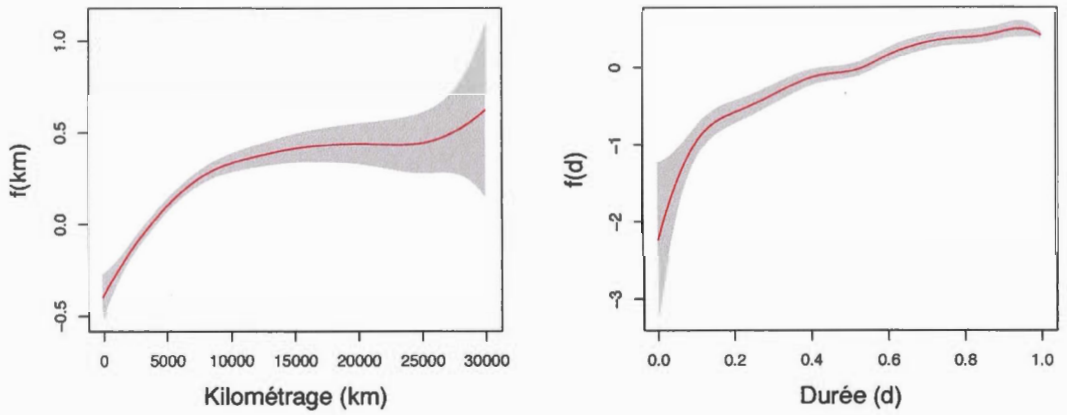
(a) Kilométrage (b) Durée

Figure 3.11: Fonctions de lissage du kilométrage et de la durée avec la pénalité d'ordre 2 pour le modèle avec la distribution Poisson inverse-gaussienne



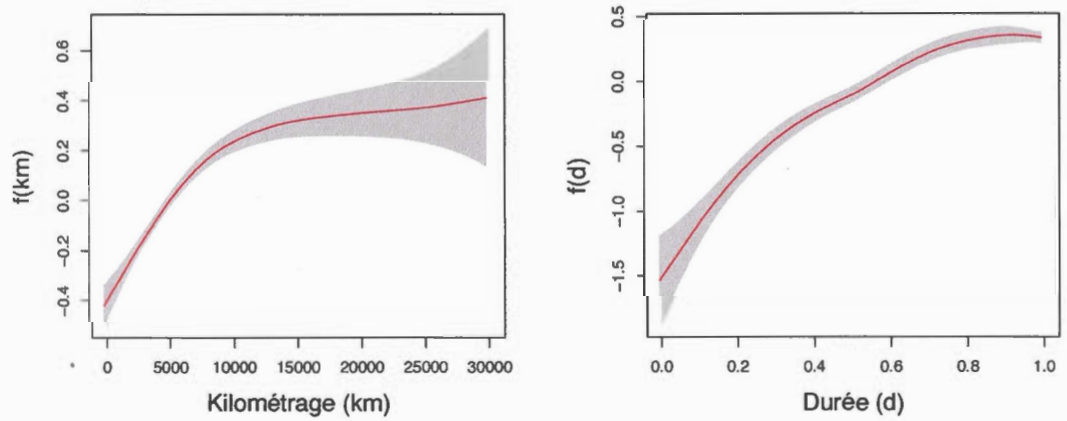
(a) Kilométrage (b) Durée

Figure 3.12: Fonctions de lissage du kilométrage et de la durée avec la pénalité d'ordre 3 pour le modèle avec la distribution Poisson inverse-gaussienne



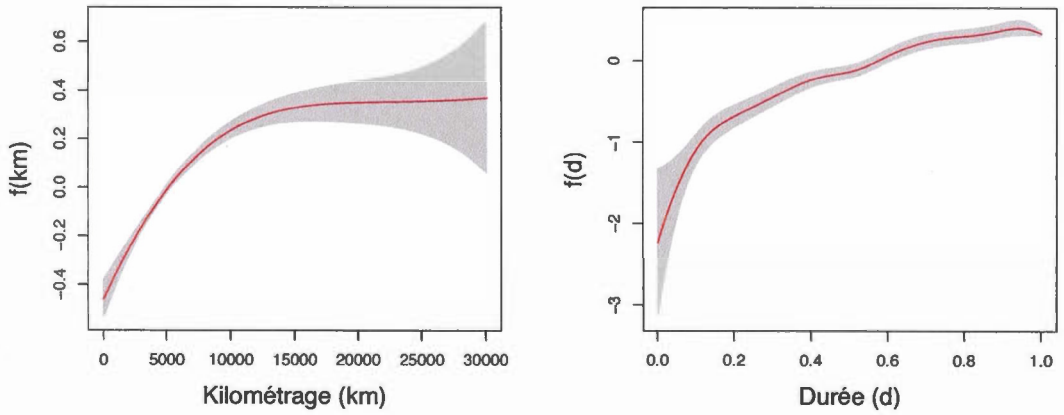
(a) Kilométrage (b) Durée

Figure 3.13: Fonctions de lissage du kilométrage et de la durée avec la pénalité d'ordre 2 pour le modèle Poisson gonflé à 0



(a) Kilométrage (b) Durée

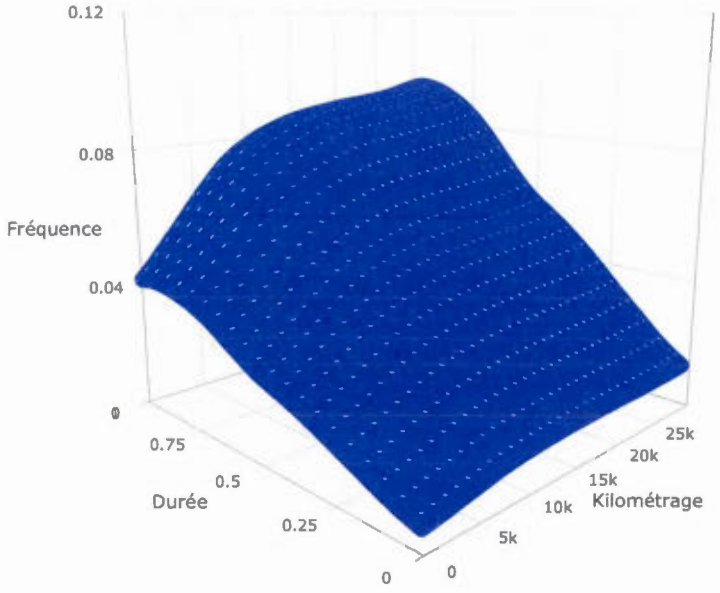
Figure 3.14: Fonctions de lissage du kilométrage et de la durée avec la pénalité d'ordre 3 pour le modèle Poisson gonflé à 0



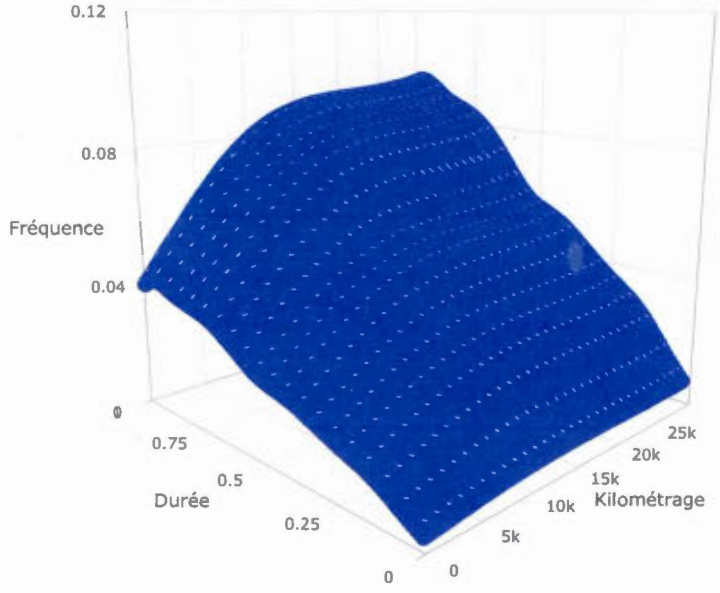
(a) Kilométrage

(b) Durée

Figure 3.15: Effet combiné du kilométrage et de la durée sur la fréquence pour les modèles avec la distribution Poisson

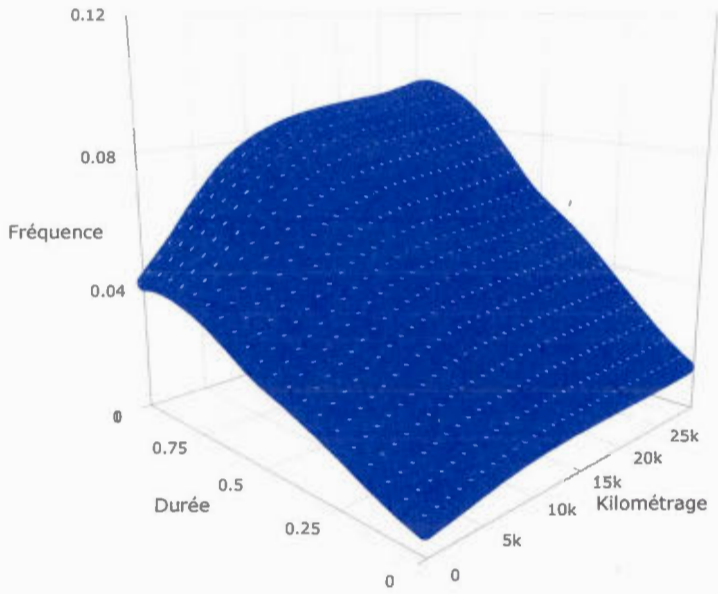


(a) Pénalité d'ordre 2

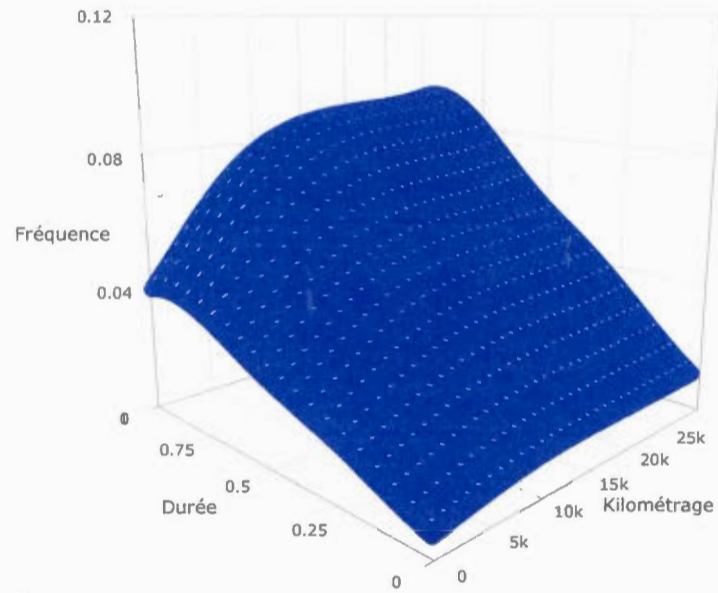


(b) Pénalité d'ordre 3

Figure 3.16: Effet combiné du kilométrage et de la durée sur la fréquence pour les modèles avec la distribution binomiale négative de type 2

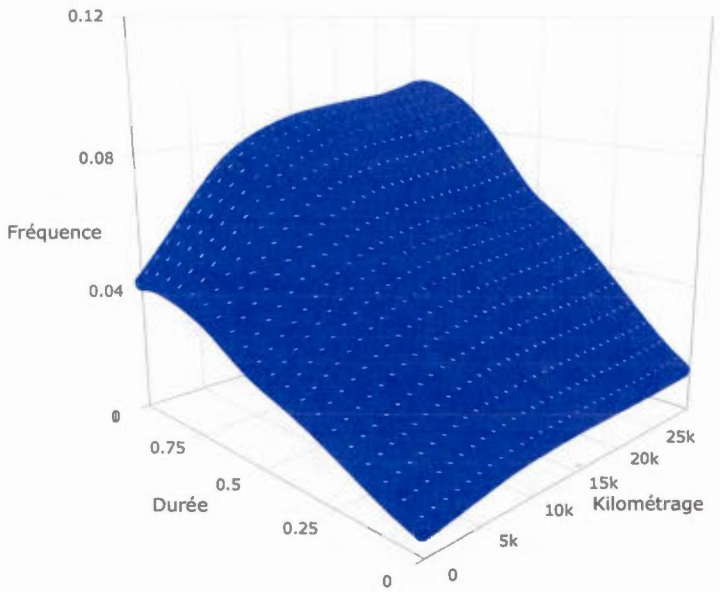


(a) Pénalité d'ordre 2

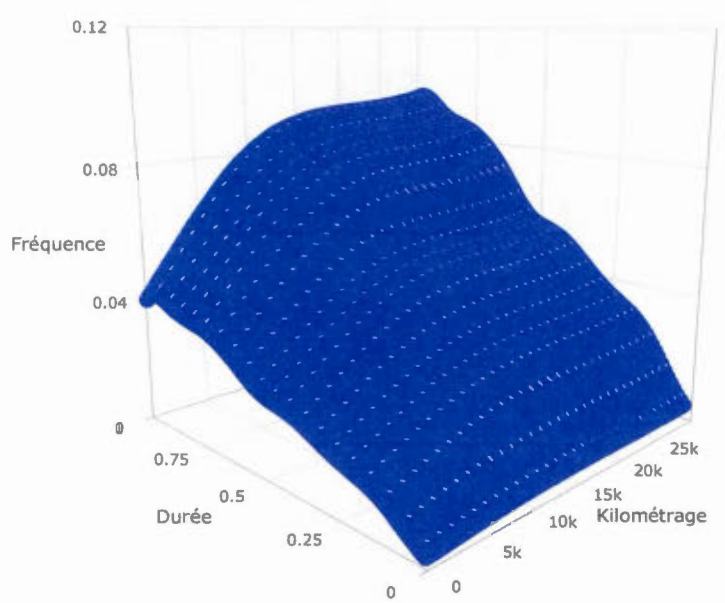


(b) Pénalité d'ordre 3

Figure 3.17: Effet combiné du kilométrage et de la durée sur la fréquence pour les modèles avec la distribution binomiale négative de type 1

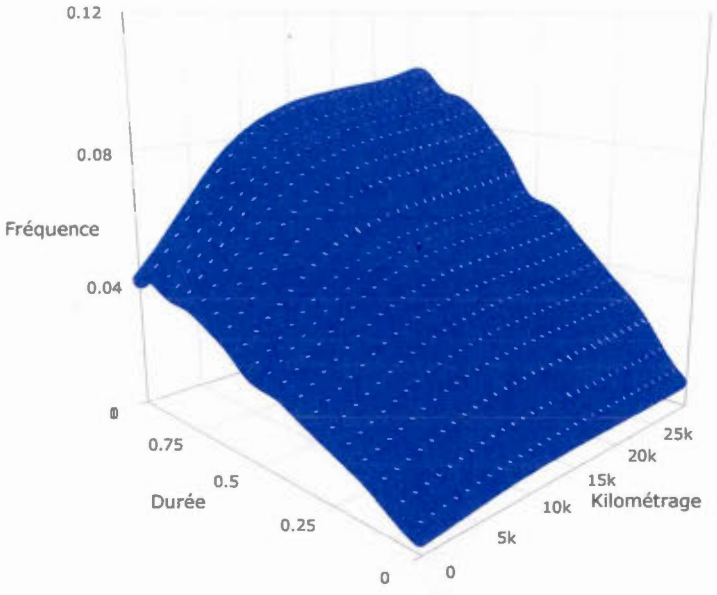


(a) Pénalité d'ordre 2

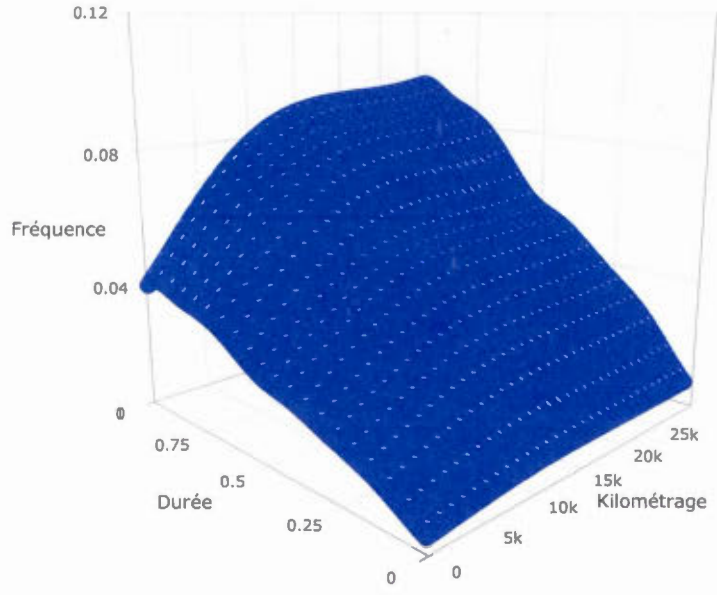


(b) Pénalité d'ordre 3

Figure 3.18: Effet combiné du kilométrage et de la durée sur la fréquence pour les modèles avec la distribution Poisson inverse-gaussienne

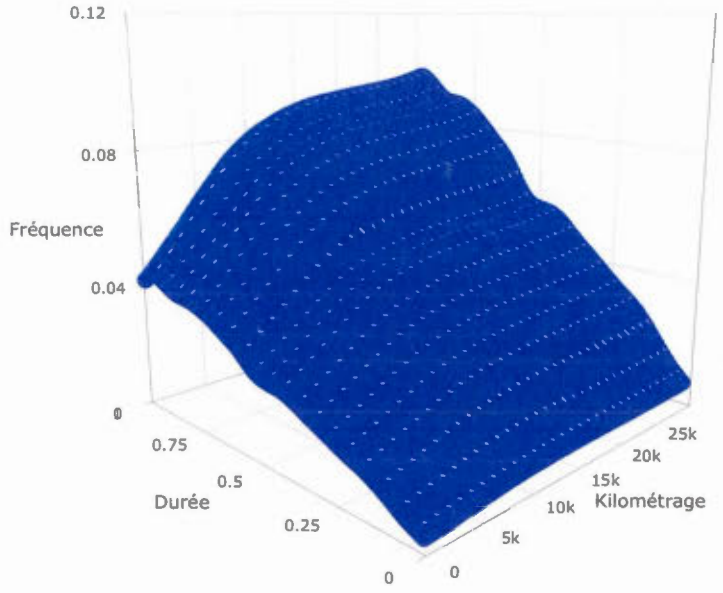


(a) Pénalité d'ordre 2

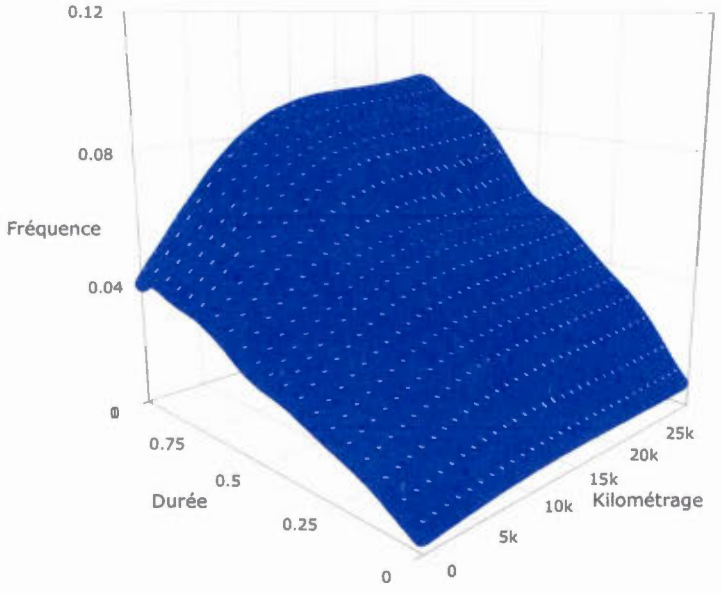


(b) Pénalité d'ordre 3

Figure 3.19: Effet combiné du kilométrage et de la durée sur la fréquence pour les modèles Poisson gonflés à 0



(a) Pénalité d'ordre 2



(b) Pénalité d'ordre 3

3.4.2 Distribution MVNB

La distribution MVNB possède comme principal avantage sur les autres distributions présentées dans ce mémoire l'ajout d'une dépendance entre les observations d'un même assuré. Ainsi, la prime prédictive associée à un assuré est ajustée par un facteur multiplicatif en fonction de son expérience passée, ce qui permet de « récompenser » les bons assurés, c'est-à-dire ceux qui ne réclament pas ou peu, avec un rabais et de « punir » les moins bons à l'aide d'une surcharge.

Comme la MVNB considère l'aspect longitudinal des données, les mêmes données sont utilisées, mais les indices changent. Pour cette distribution, l'indice i représente un assuré et non une observation comme c'est le cas pour les distributions à données transversales. De plus, un indice temporel t est ajouté pour distinguer les observations d'un même assuré i.

Pour les modèles avancés avec la distribution MVNB, il faut alors revoir légèrement l'équation 3.3 de la fonction de lien du paramètre de moyenne pour incorporer les nouveaux indices :

$$g(\lambda_{i,t}) = \log(\lambda_{i,t}) = \boldsymbol{X}_{i,t}\boldsymbol{\beta} + f_1(km_{i,t}) + f_2(d_{i,t}),$$

où les indices (i,t) signifient la t^{e} observation de l'assuré i. Mis à part ce changement, tout le reste est identique, c'est-à-dire que $\boldsymbol{\beta}$ est toujours un vecteur de coefficients de longueur 8, que $f_1(\cdot)$ et $f_2(\cdot)$ correspondent aux fonctions de lissage du kilométrage et de la durée respectivement et que $\boldsymbol{X}$ est une matrice de dimension $n \times 8$ contenant les variables explicatives x_1 à x_8 présentées au tableau 3.6. Dans ce cas-ci, $\boldsymbol{X}_{i,t}$ correspond à la ligne de la matrice $\boldsymbol{X}$ associée à la t^{e} observation de l'assuré i.

En raison de la complexité accrue de la distribution MVNB, il a été difficile de minimiser le critère AIC afin d'estimer les paramètres de lissage des fonctions $f_1(\cdot)$

et $f_2(\cdot)$. Idéalement, il aurait fallu appliquer un algorithme de minimisation, mais, dans le but de vérifier plus rapidement si la distribution MVNB pouvait offrir un meilleur AIC que les autres modèles avancés, une autre méthode a été appliquée.

Celle-ci consiste tout d'abord à tester quelques valeurs pour les paramètres de lissage du kilométrage et de la durée afin de trouver ceux qui, à première vue, semblent minimiser le AIC. Ensuite, on établit une grille de valeurs pour les paramètres de lissage autour de celles trouvées précédemment afin de vérifier si, près de celles-ci, il n'y a aurait pas un autre couple de paramètres de lissage fournissant un AIC encore plus faible. Cette méthode a pour avantages d'être rapide et facile à appliquer, mais au détriment d'une plus grande précision. De plus, il est possible que le couple de paramètres de lissage trouvé ne fournisse qu'un minimum relatif pour le AIC et non le minimum absolu.

Comme à la section 3.4.1 l'ordre des pénalités ne semblait pas avoir un impact important, il a été décidé de se concentrer sur la pénalité d'ordre 2 uniquement et d'appliquer la méthode décrite ci-dessus. Le tableau 3.15 présente la grille des AIC obtenus selon les valeurs de paramètres de lissage utilisées. Avec celle-ci, le paramètre de lissage associé au kilométrage va de 5500 à 8750 par bonds de 250 et celui associé à la durée va de 4000 à 6250 par bonds de 250. Le AIC minimal est de 44 054,52 est obtenu avec des paramètres de lissage de 8000 pour le kilométrage et de 6000 pour la durée. Ce AIC est largement inférieur à celui des autres modèles avancés avec une pénalité d'ordre 2 (voir tableau 3.14).

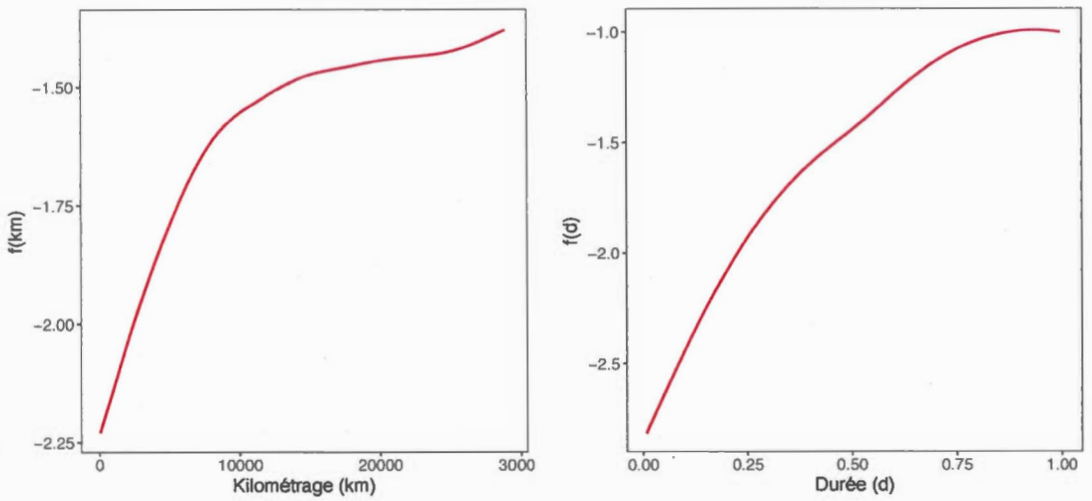
Pour le modèle avancé MVNB, les fonctions de lissage estimées, présentées à la figure 3.20, ont la même forme que celles de la section 3.4.1. L'effet combiné du kilométrage et de la durée sur le paramètre de moyenne $\lambda_{i,t}$ est lui aussi très semblable aux autres modèles avancés.

Par ailleurs, il est intéressant de noter que l'écart entre le plus petit AIC et le plus

Tableau 3.15: Valeur du AIC pour un modèle avec la distribution MVNB selon les paramètres de lissage du kilométrage et de la durée

	λ_d									
λ_{km}	4000	4250	4500	4750	5000	5250	5500	5750	6000	6250
5500	44 055,46	44 055,41	44 055,52	44 055,29	44 055,47	44 055,38	44 055,33	44 055,25	44 055,20	44 055,12
5750	44 055,36	44 055,28	44 055,27	44 055,27	44 055,28	44 055,36	44 055,34	44 055,25	44 055,16	44 055,08
6000	44 055,44	44 055,44	44 055,64	44 055,24	44 055,24	44 055,29	44 055,25	44 055,16	44 055,07	44 054,99
6250	44 055,42	44 055,34	44 055,56	44 055,27	44 055,28	44 055,23	44 055,16	44 055,07	44 054,98	44 054,90
6500	44 055,28	44 055,33	44 055,40	44 055,21	44 055,26	44 055,18	44 055,10	44 055,03	44 054,97	44 054,95
6750	44 055,36	44 055,24	44 055,37	44 055,29	44 055,11	44 055,06	44 055,05	44 054,98	44 054,91	44 054,91
7000	44 055,30	44 055,21	44 055,34	44 055,23	44 055,10	44 055,05	44 054,99	44 054,93	44 054,89	44 054,80
7250	44 055,27	44 055,39	44 055,29	44 055,11	44 055,09	44 055,02	44 054,95	44 054,88	44 054,87	44 054,74
7500	44 055,45	44 055,18	44 055,10	44 055,06	44 055,02	44 054,97	44 054,96	44 054,86	44 054,78	44 054,76
7750	44 055,21	44 055,15	44 055,06	44 055,02	44 054,97	44 054,92	44 054,74	44 054,82	44 054,82	44 054,72
8000	44 055,19	44 055,08	44 055,05	44 054,99	44 054,96	44 054,90	44 054,93	44 054,80	**44 054,52**	44 054,64
8250	44 055,16	44 055,07	44 055,02	44 054,98	44 054,94	44 054,93	44 054,79	44 054,82	44 054,82	44 054,75
8500	44 055,14	44 055,06	44 054,99	44 054,97	44 054,89	44 054,89	44 054,77	44 054,71	44 054,74	44 054,53
8750	44 055,13	44 055,05	44 054,97	44 054,91	44 054,87	44 054,81	44 054,77	44 054,74	44 054,75	44 054,73

Figure 3.20: Fonctions de lissage du kilométrage et de la durée avec la pénalité d'ordre 2 pour le modèle avec la distribution MVNB



(a) Kilométrage

(b) Durée

grand dans le tableau 3.15 n'est d'à peine 1,12, ce qui est très peu. De plus, les paramètres estimés de chaque modèle de la grille sont très semblables entre eux. On peut se questionner à savoir si une meilleure technique d'estimation pour les paramètres de lissage apporterait une amélioration significative au modèle.

À ce sujet, la figure 3.22 présente la relation entre l'inverse du paramètre ν de la distribution MVNB et le AIC des modèles engendrés par la grille. Bien qu'il n'y ait pas de relation évidente, on y constate que la variation de l'inverse du paramètre ν est très faible, se déroulant à la cinquième et à la sixième décimale. L'intérêt de s'intéresser à cette transformation du paramètre ν repose sur le fait qu'elle représente l'hétérogénéité résiduelle qui n'est pas capturée par les régresseurs. Ainsi, plus cette valeur est faible, mieux le modèle capture l'hétérogénéité. Également, plus cette valeur est faible et plus la variance est faible pour une même

Figure 3.21: Effet combiné du kilométrage et de la durée sur la fréquence pour le modèle avec la distribution MVNB et des pénalités d'ordre 2

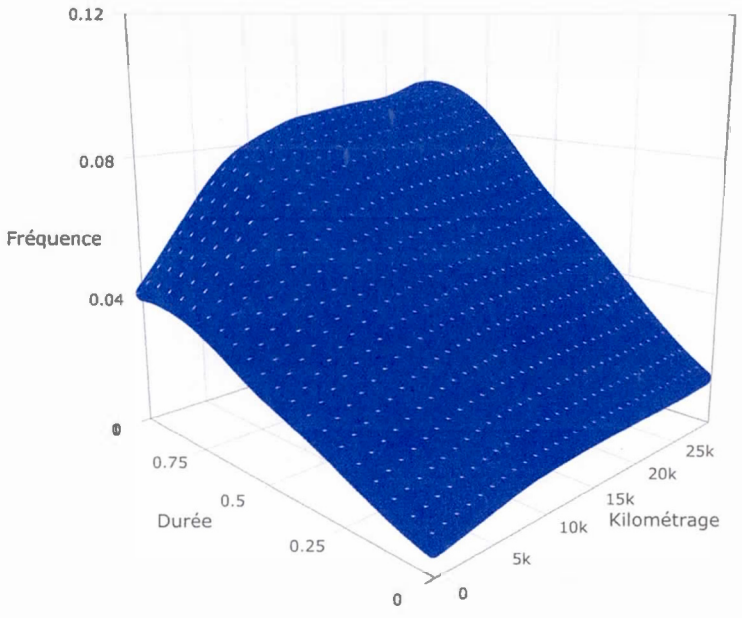
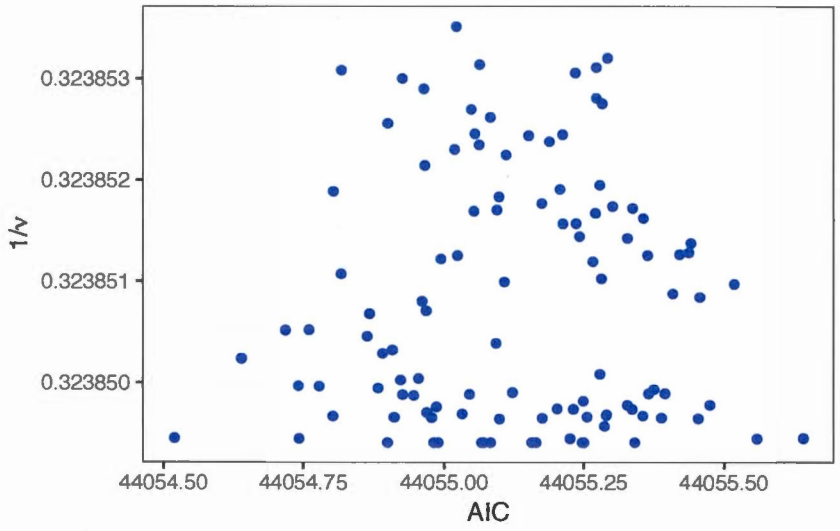


Figure 3.22: Relation entre l'inverse du paramètre ν et le AIC pour des modèles avancés à distribution MVNB variant les valeurs de paramètres de lissage



valeur du paramètre de moyenne $\lambda_{i,t}$, car, sous la MVNB :

$$Var[Y_{i,t}] = \lambda_{i,t} + \frac{1}{\nu}\lambda_{i,t}^2.$$

Le tableau 3.16 présente l'hétérogénéité des 3 modèles à distribution MVNB utilisés, soit les modèles MVNB classique, MVNB classique avec kilométrage et MVNB avancé. On y constate que le modèle MVNB avancé est celui avec l'hétérogénéité résiduelle la plus faible, ce qui correspond aux attentes. En utilisant des fonctions de lissage pour le kilométrage exact parcouru et la durée, les assureurs désirent améliorer la qualité de leurs modèles et c'est ce qui est observé ici.

La valeur du paramètre d'hétérogénéité résiduelle affecte la prime prédictive. Celle-ci, présentée à l'équation 2.21, est réécrite ici :

$$E[Y_{i,T_i+1}|y_{i,1},...,y_{i,T_i}] = \lambda_{i,T_i+1}\frac{\nu + \sum_{j=1}^{T_i} y_{i,j}}{\nu + \sum_{j=1}^{T_i} \lambda_{i,j}}.$$

Pour rappel, la prime prédictive est constituée d'une prime de base λ_{i,T_i+1} à la-

Tableau 3.16: Comparaison de l'hétérogénéité résiduelle des différents modèles avec la distribution MVNB

Modèle	$1/\nu$
MVNB classique	0,353210
MVNB classique avec km	0,3414675
MVNB avancé	0,3238495

quelle est appliquée un facteur d'ajustement pour tenir compte de l'expérience passée de l'assuré.

En supposant une prime de base constante de 0,1000, soit $\lambda_{i,t} = 0{,}10000$ pour tout t, nous pouvons aisément vérifier l'impact de la diminution du paramètre d'hétérogénéité résiduelle sur la prime prédictive d'un assuré de 10 ans selon son nombre de réclamations. Cet exemple est présenté au tableau 3.17.

Tableau 3.17: Impact de la diminution du paramètre d'hétérogénéité sur la prime prédictive d'un assuré de 10 ans selon son nombre de réclamations

		Nombre de réclamations				
Modèle	A priori	0	1	2	3	4
MVNB classique	0,1000	0,0261	0,1000	0,1739	0,2478	0,3217
MVNB classique avec km	0,1000	0,0254	0,1000	0,1745	0,2491	0,3236
MVNB avancé	0,1000	0,0245	0,1000	0,1755	0,2511	0,3266

On y constate que plus le paramètre d'hétérogénéité résiduelle $(1/\nu)$ est petit, plus le facteur d'ajustement de la prime devient important, c'est-à-dire plus les rabais ou surcharges qu'il octroie sont grands. Ainsi, les assurés avec un bon historique sont davantage favorisés et ceux avec un mauvais historique, davantage pénalisés.

3.4.3 Comparaison de la qualité d'ajustement des modèles avancés

Jusqu'à maintenant, nous nous sommes surtout attardés aux fonctions de lissage ainsi qu'au AIC des modèles avancés. Comme précédemment, nous pouvons aussi nous intéresser à la prédiction de la répartition du nombre de réclamations afin de voir l'ajustement que présente chaque modèle avancé comparativement à la répartition observée.

Avec les données d'estimation, on constate au tableau 3.18 que le modèle avancé Poisson est celui qui s'ajuste le moins bien aux observations. Tous les autres modèles avancés s'ajustent très bien à ce qui est observé. C'est particulièrement le cas avec le modèle avancé Poisson gonflé à 0 dont les prédictions sont quasiment exactes.

Tableau 3.18: Comparaison de la prédiction du nombre de réclamations des modèles avancés avec pénalité d'ordre 2 avec les données d'estimation

Modèle	Nombre de réclamations					
	0	1	2	3	4	≥ 5
Poisson	79 798,34	5724,99	262,27	9,14	0,26	0
Bin. Nég. type 2	79 891,09	5551,37	332,35	19,06	1,07	0,06
Bin. Nég. type 1	79 884,68	5561,85	330,19	17,39	0,84	0,14
Poisson inv.-gaus.	79 889,80	5554,78	329,55	19,58	1,21	0,08
Poisson gonflé à 0	79 891,89	5544,62	341,81	16,05	0,61	0,02
MVNB	79 881,17	5569,13	325,74	17,95	0,96	0,05
Observé	79 893	5545	339	17	1	0

En passant aux données de validation, les résultats changent à la faveur du modèle avancé MVNB. Au tableau 3.19, on s'aperçoit que ce modèle s'ajuste beaucoup mieux à ces nouvelles données que les autres modèles avancés. Complètement en

retrait, on retrouve encore une fois le modèle avancé Poisson. Tous les autres modèles avancés présentent un ajustement semblable entre eux par rapport aux observations.

Tableau 3.19: Comparaison de la prédiction du nombre de réclamations des modèles avancés avec pénalité d'ordre 2 avec les données de validation

Modèle	Nombre de réclamations					
	0	1	2	3	4	≥ 5
Poisson	39 400,41	3045,53	131,81	4,14	0,10	0,01
Bin. Nég. type 2	39 457,00	2948,56	167,29	8,69	0,44	0,02
Bin. Nég. type 1	39 431,92	2970,79	170,35	8,53	0,39	0,02
Poisson inv.-gaus.	39 446,04	2958,75	167,58	9,10	0,51	0,02
Poisson gonflé à 0	39 446,77	2955,92	171,82	7,23	0,24	0,02
MVNB	39 644,97	2765,97	161,66	8,90	0,47	0,03
Observé	39 585	2820	172	5	0	0

3.5 Améliorer la modélisation de la distribution Poisson gonflée à zéro

Les GAMLSS procurent une option intéressante par rapport aux GLM et aux GAM qui n'a pas été appliquée dans ce mémoire jusqu'à maintenant, soit d'appliquer des régresseurs à travers une fonction de lien pour un second paramètre.

L'espérance de la distribution Poisson gonflée à zéro, présentée à l'équation (2.19), compte 2 paramètres, λ et ϕ, et il est intéressant de voir si l'utilisation d'une seconde fonction de lien peut améliorer le AIC du modèle avancé Poisson gonflé à 0.

Comme précédemment, une fonction de lien logarithmique, $g_1(\cdot)$ est retenue pour le paramètre λ. Pour ϕ, un paramètre prenant une valeur comprise entre 0 et 1 pour « gonfler » à zéro une distribution Poisson, la fonction de lien logit est retenue.

Choisir les régresseurs pour chaque fonction de lien s'avère une tâche qui est simple en apparence, mais qui demande beaucoup de travail. De nombreuses combinaisons de régresseurs ont été testées et comparées par l'entremise du critère AIC. Parmi celles-ci, il y a notamment l'utilisation de tous les régresseurs pour les 2 fonctions de lien et les multiples possibilités de répartition des régresseurs entre les 2 fonctions de lien.

Après plusieurs tentatives avec les régresseurs, on retient le modèle avec les fonctions ci-dessous :

$$g_1(\lambda_i) = \log(\lambda_i) = \boldsymbol{X}_i\boldsymbol{\beta} + f_2(d_i),$$
$$g_2(\phi_i) = \text{logit}(\phi_i) = \log\left(\frac{\phi_i}{1-\phi_i}\right) = \alpha_0 + f_1(km_i),$$

où $g_1(\cdot)$ est pratiquement identique à la fonction de lien des modèles avancés de l'équation (3.3). La différence réside dans le retrait de la fonction de lissage $f_1(\cdot)$

du kilométrage. En effet, cette fonction est plutôt utilisée avec la fonction de lien $g_2(\cdot)$ et est accompagnée d'une constante, soit α_0. La matrice $\boldsymbol{X}$ des variables explicatives est elle aussi identique, c'est-à-dire qu'elle regroupe les variables x_1 à x_8 du tableau 3.6 pour les n assurés.

Afin d'éviter toute confusion avec les modèles précédents, celui-ci sera nommé modèle avancé modifié.

Comme avec les modèles avancés avec des distributions à données transversales, des pénalités d'ordre 2 et 3 ont été testées et, encore une fois, il n'y a pas une grande différence sur le AIC entre les 2 ordres de pénalité. Avec le tableau 3.20, on constate que le AIC du modèle avancé modifié est quelque peu meilleur que le modèle avancé Poisson gonflé à 0.

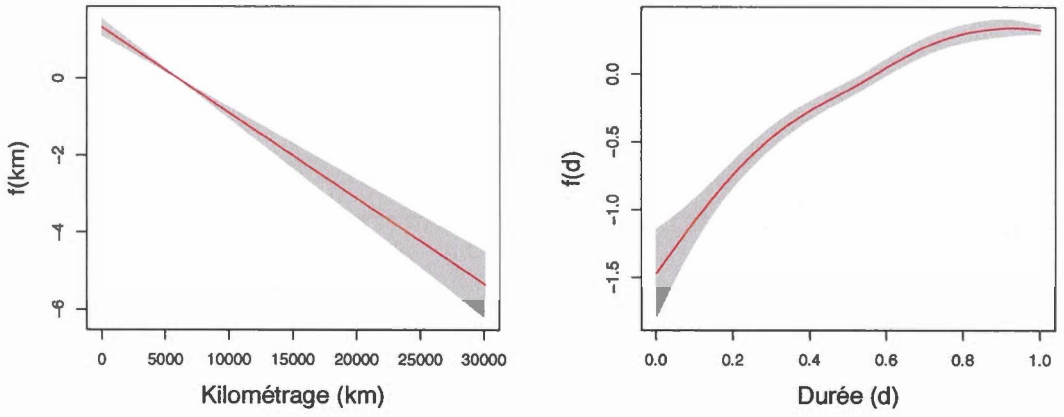
Tableau 3.20: Comparaison selon le critère AIC de la qualité des modèles avancés et avancés modifiés avec la distribution Poisson gonflée à 0 avec une pénalité d'ordre k pour les fonctions de lissage par P-splines

Modèle	AIC pour $k = 2$	AIC pour $k = 3$
Poisson gonflé à zéro avancé	44 098,21	44 096,39
Poisson gonflé à zéro avancé modifié	**44 093,22**	**44 094,23**

Aux figures 3.23 et 3.24, les fonctions de lissages du kilométrage et de la durée sont présentées pour des pénalités d'ordre 2 et 3 respectivement. Pour la durée, les fonctions obtenues sont très similaires à celles des modèles avancés. Ceci n'est guère surprenant puisque la durée est associée à une fonction de lien logarithmique dans tous les cas. Pour le kilométrage, les fonctions de lissage sont complètement différentes, ce qui est normal en raison de la fonction de lien logit utilisée pour le paramètre ϕ.

Pour la pénalité d'ordre 2, $f_1(\cdot)$ décroît de manière linéaire. En passant à la

Figure 3.23: Fonctions de lissage du kilométrage et de la durée avec la pénalité d'ordre 2 pour le modèle modifié avec la distribution Poisson gonflée à 0



(a) Kilométrage

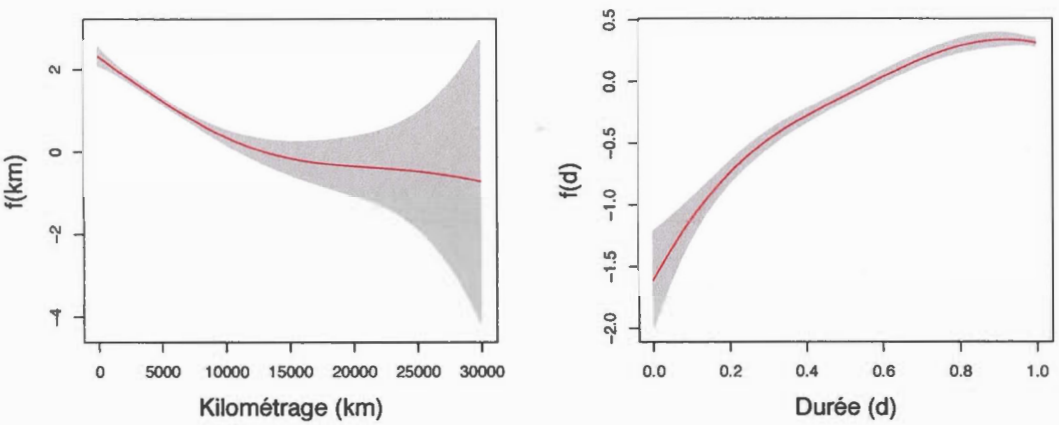
(b) Durée

pénalité d'ordre 3, nous observons également que $f_1(\cdot)$ décroît de manière linéaire, mais seulement pour les 12 500 premiers kilomètres. Ensuite, la fonction de lissage semble atteindre un plateau, mais l'écart-type de la fonction, représenté par la zone ombragée, augmente avec le kilométrage et devient très grand, ce qui complique l'interprétation.

Une fonction de lissage décroissante pour le kilométrage implique que, sous la fonction de lien logit, le paramètre ϕ diminue avec le kilométrage. Par conséquent, nous obtenons que, pour un assuré parcourant de plus courtes distances, le paramètre ϕ est plus élevé, ce qui signifie que sa probabilité de ne pas réclamer est davantage « gonflée » que pour un assuré parcourant plus de kilomètres.

La figure 3.25 présente l'effet combiné de la durée et du kilométrage sur l'espérance du modèle avancé modifié. Pour l'obtenir, les variables explicatives x_2 à x_8 ont

Figure 3.24: Fonctions de lissage du kilométrage et de la durée avec la pénalité d'ordre 3 pour le modèle modifié avec la distribution Poisson gonflée à 0



(a) Kilométrage

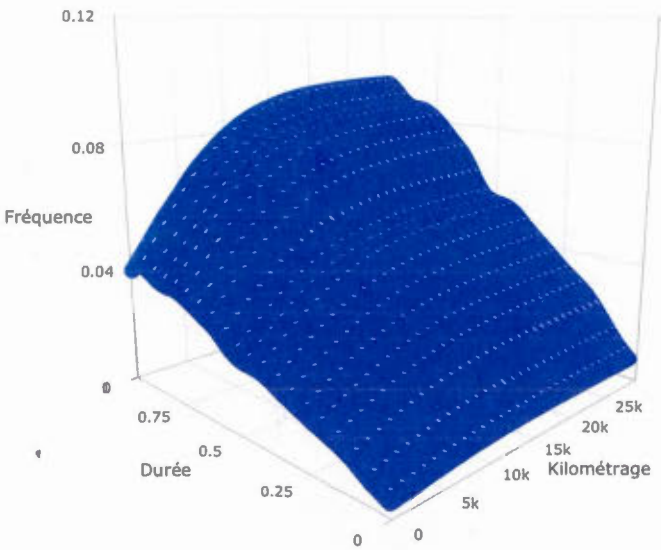
(b) Durée

été présumées égales à 0. Malgré la modification des fonctions de lien, les surfaces obtenues pour des pénalités d'ordre 2 et 3 sont très similaires à celles des modèles avancés à distribution Poisson gonflée à 0. Le creux à 0,5 année est bien visible pour les 2 ordres de pénalité. De plus, ces figures semblent manquer un peu de lissage par leurs ondulations, c'est-à-dire de présenter un peu trop de surajustement.

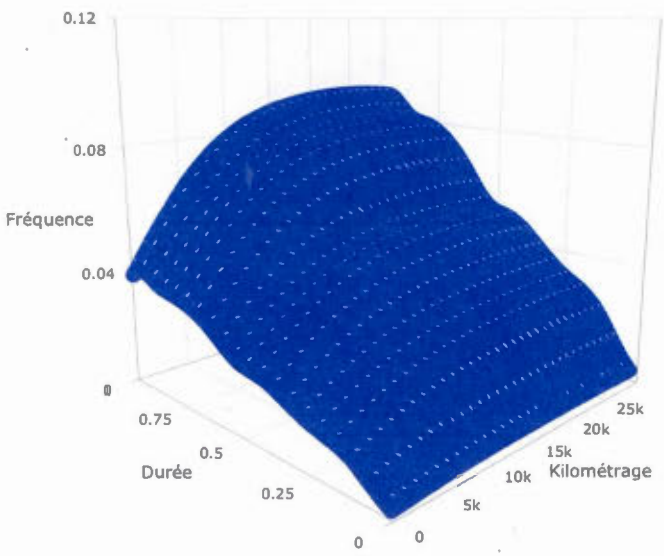
Une façon de vérifier si nous nous retrouvons effectivement en présence de surajustement consiste à tester ces modèles sur les données d'estimation et de validation. Au tableau 3.21, avec les données d'estimation, on constate que le modèle avancé Poisson gonflé à 0 est mieux ajusté aux nombres observés de réclamations que le modèle avancé modifié. Malgré tout, ce dernier fait un travail honnête comme ajustement.

En passant aux données de validation, la situation s'inverse. Au tableau 3.22,

Figure 3.25: Effet combiné du kilométrage et de la durée sur la fréquence pour les modèles modifiés avec la distribution Poisson gonflée à 0



(a) Pénalité d'ordre 2



(b) Pénalité d'ordre 3

Tableau 3.21: Comparaison de la prédiction du nombre de réclamations pour les modèles avec la distribution Poisson gonflée à 0 avec pénalité d'ordre 2 pour les données d'estimation

Modèle	Nombre de réclamations					
	0	1	2	3	4	≥ 5
Poisson gonflé à 0	79 891,89	5544,62	341,81	16,05	0,61	0,02
Poisson gonflé à 0 modifié	79 872,53	5579,23	328,95	13,82	0,45	0,02
Observé	79 893	5545	339	17	1	0

nous constatons que le modèle avancé modifié s'approche énormément plus des observations. Ce modèle semble donc plus robuste à un changement de données.

Tableau 3.22: Comparaison de la prédiction du nombre de réclamations pour les modèles avec la distribution Poisson gonflée à 0 avec pénalité d'ordre 2 pour les données de validation

Modèle	Nombre de réclamations					
	0	1	2	3	4	≥ 5
Poisson gonflé à 0	39 446,77	2955,92	171,82	7,23	0,24	0,02
Poisson gonflé à 0 modifié	39 646,13	2765,84	162,97	6,84	0,22	0
Observé	39 585	2820	172	5	0	0

3.6 Impact sur les primes

Jusqu'à maintenant, les modèles présentés ont été comparés selon des critères reposant sur l'ensemble des données d'estimation ou de validation. Or, 2 modèles peuvent performer similairement en considérant l'ensemble des assurés, mais très différemment d'un profil d'assuré à un autre.

Pour prendre en compte cette réalité, cette section va étudier les primes générées pour quelques profils d'assurés selon les modèles utilisés. En procédant ainsi, l'impact sur la tarification des différents modèles est également analysé.

Dans ce mémoire, nous nous concentrons sur la modélisation de la fréquence de réclamations. Ainsi, les primes dont il sera question représentent l'espérance du nombre de réclamations d'un assuré.

Le tableau 3.23 présente les 3 profils d'assurés étudiés, soit respectivement un profil à bas risque, un second à risque modéré et un troisième à haut risque. Le profil de bas risque correspond à une femme âgée de plus de 30 ans et ayant un véhicule de plus de 10 ans. Un assuré dont l'âge est compris entre 25 et 30 ans et conduisant un véhicule âgé de 2 à 5 ans constitue le profil de risque modéré. Finalement, le profil à risque élevé représente un homme de 25 ans ou moins avec un véhicule de 2 ans ou moins.

Tableau 3.23: Caractéristiques du risque pour les 3 profils d'assurés étudiés

Profil	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Bas risque	1	0	0	0	0	0	0	1
Risque modéré	1	0	1	0	1	0	1	1
Haut risque	1	1	0	0	0	1	1	1

Ces profils de risque seront analysés selon 3 scénarios d'exposition, présentés au tableau 3.24. On y note que le kilométrage considéré va de 7500 à 15 000 km et que la durée d'exposition va de 0,5 à 1 an.

Le tableau 3.25 présente les primes prédictives pour les modèles classiques. Comme ces modèles n'utilisent pas le kilométrage, l'exposition est uniquement basée sur la durée, fixée à 1 an.

Tableau 3.24: Niveaux d'exposition au risque considérés pour chaque profil d'assurés étudié

Type d'exposition	Kilométrage (km)	Durée (année)
a	7500	0,5
b	12 000	0,8
c	15 000	1

Pour la distribution MVNB, les primes présentées dans les tableaux 3.25, 3.26 et 3.27 sont des primes a priori, c'est-à-dire des primes chargées à des nouveaux assurés pour lesquels l'assureur n'a pas d'historique.

Il est intéressant de noter que les primes présentées au tableau 3.25 sont très semblables d'une distribution à l'autre pour chaque combinaison du profil de risque et du type d'exposition. Sans surprise, on constate que la variance des primes est plus élevée pour les distributions autres que la Poisson, soit celles admettant de la surdispersion.

Tableau 3.25: Primes prédictives pour les modèles classiques selon les 3 profils de risque étudiés et une exposition de 1 an

	Bas risque		Risque moyen		Haut risque	
Distribution	Moy.	Var.	Moy.	Var.	Moy.	Var.
Poisson	0,0686	0,0686	0,1001	0,1001	0,1378	0,1378
Bin. Nég. type 2	0,0686	0,0705	0,1002	0,1043	0,1379	0,1458
Bin. Nég. type 1	0,0690	0,0713	0,1004	0,1036	0,1373	0,1418
Poisson inv.-gau.	0,0686	0,0705	0,1002	0,1043	0,1379	0,1458
Poisson gonflée à 0	0,0685	0,0711	0,1001	0,1056	0,1379	0,1483
MVNB	0,0689	0,0706	0,1005	0,1041	0,1378	0,1445

Les primes prédictives pour les modèles classiques avec kilométrage sont présentées au tableau 3.26. Encore une fois, les primes sont pratiquement identiques entre les distributions et la variance des primes avec la distribution Poisson est plus faible qu'avec les autres distributions.

Sous l'exposition c, soit avec un kilométrage de 15 000 km et une durée d'un an, ces primes peuvent être comparées avec celles du tableau 3.25, puisque la durée est identique. On constate que ces dernières sont inférieures à celles du tableau 3.26 avec l'exposition c. Ceci s'explique évidemment par l'ajout du kilométrage qui permet d'améliorer la segmentation.

Pour les modèles avancés avec la pénalité d'ordre 2, les primes prédictives se trouvent au tableau 3.27. Encore une fois, les distributions produisent sensiblement les mêmes primes et la distribution Poisson procure une variance presque toujours plus faible en raison de sa contrainte d'équidispersion. À ce propos, les exceptions proviennent des distributions Poisson inverse-gaussienne et Poisson gonflée à 0 qui semblent agir différemment sous l'exposition de type a. En effet, leurs primes prédictives pour cette exposition sont légèrement plus faibles que les autres, ce qui fait que leurs variances, bien que surdispersées, demeurent sous les variances de la distribution Poisson.

Par ailleurs, en comparant les tableaux 3.26 et 3.27 entre eux, on constate que les primes sont plus élevées pour les modèles avancés que pour les modèles classiques avec kilométrage, sauf pour les profils de risque moyen et élevé avec une exposition de type c (1 an et 15 000 km) où les primes sont similaires. Il pourrait s'agir d'une conséquence de la proportionnalité de la durée sur la fréquence pour les modèles classiques avec kilométrage. Cette approche pourrait sous-estimer la fréquence de ces modèles pour des durées inférieures à 1.

Au tableau 3.28, les primes prédictives pour le modèle avancé à distribution Pois-

son gonflée à 0 et le modèle avancé modifié sont présentées. Pour ces 2 modèles, des pénalités d'ordre 2 pour les fonctions de lissage ont été retenues. Bien que la gestion des variables explicatives à travers les fonctions de lien soient différentes entre ces 2 modèles, il est intéressant de constater que les primes sont quasiment identiques. Toutefois, la variance des primes est plus faible avec le modèle avancé modifié pour les expositions de type b et c. Ces expositions correspondent à des kilométrages élevés, soit respectivement de 12 000 km sur 0,8 année et de 15 000 km sur un an. Ceci peut s'expliquer par le paramètre de gonflement ϕ qui diminue lorsque le kilométrage augmente avec le modèle avancé modifié plutôt que d'être fixe comme sous le modèle avancé.

Jusqu'à maintenant, les primes pour les modèles à distribution MVNB ont été calculées pour des nouveaux assurés, c'est-à-dire des assurés pour lesquels un assureur n'a pas d'historique. Toutefois, avec la MVNB, une dépendance entre les observations d'un même assuré est introduite. Ainsi, les primes s'ajustent à chaque année selon l'historique des assurés et sont dites a posteriori.

Les tableaux 3.29, 3.30 et 3.31 présentent les primes a priori et a posteriori pour les 3 profils d'assuré selon une exposition d'un an et 15 000 km pour la distribution MVNB sous les modèles classique, classique avec kilométrage et avancé respectivement. Pour cet exemple de tarification, les caractéristiques du risque ainsi que les primes a priori sont considérées constantes dans le temps. Il est également présumé que l'assureur possède un historique de 10 ans sur les assurés.

On constate dans ces tableaux qu'un assuré est surchargé s'il présente plus de réclamations que ce que l'assureur prévoyait sur l'horizon de temps analysé. Si c'est l'inverse, alors il reçoit un rabais. L'équation (2.21) présente la formule utilisée pour calculer les primes a posteriori.

Tableau 3.26: Primes prédictives pour les modèles classiques avec kilométrage selon les 3 profils de risque et les types d'exposition étudiés

Expo.	Distribution	Bas risque		Risque moyen		Haut risque	
		Moy.	Var.	Moy.	Var.	Moy.	Var.
a	Poisson	0,0387	0,0387	0,0529	0,0529	0,0723	0,0723
	Bin. Nég. type 2	0,0387	0,0393	0,0529	0,0540	0,0724	0,0744
	Bin. Nég. type 1	0,0389	0,0401	0,0530	0,0546	0,0720	0,0742
	Poisson inv.-gau.	0,0387	0,0393	0,0529	0,0540	0,0723	0,0744
	Poisson gonflée à 0	0,0387	0,0394	0,0529	0,0543	0,0724	0,0751
	MVNB	0,0389	0,0394	0,0531	0,0540	0,0722	0,0740
b	Poisson	0,0677	0,0677	0,0924	0,0924	0,1264	0,1264
	Bin. Nég. type 2	0,0676	0,0694	0,0924	0,0957	0,1264	0,1326
	Bin. Nég. type 1	0,0681	0,0701	0,0926	0,0954	0,1259	0,1298
	Poisson inv.-gau.	0,0676	0,0694	0,0924	0,0957	0,1264	0,1326
	Poisson gonflée à 0	0,0676	0,0699	0,0924	0,0968	0,1265	0,1347
	MVNB	0,0681	0,0697	0,0930	0,0959	0,1265	0,1320
c	Poisson	0,0846	0,0846	0,1155	0,1155	0,1580	0,1580
	Bin. Nég. type 2	0,0845	0,0873	0,1155	0,1207	0,1580	0,1677
	Bin. Nég. type 1	0,0851	0,0877	0,1158	0,1193	0,1574	0,1622
	Poisson inv.-gau.	0,0845	0,0873	0,1156	0,1207	0,1580	0,1677
	Poisson gonflée à 0	0,0845	0,0881	0,1155	0,1223	0,1581	0,1709
	MVNB	0,0851	0,0876	0,1162	0,1208	0,1582	0,1667

Tableau 3.27: Primes prédictives pour les modèles avancés avec pénalité d'ordre 2 selon les 3 profils de risque et les types d'exposition étudiés

Expo.	Distribution	Bas risque		Risque moyen		Haut risque	
		Moy.	Var.	Moy.	Var.	Moy.	Var.
a	Poisson	0,0477	0,0477	0,0620	0,0620	0,0817	0,0817
	Bin. Nég. type 2	0,0477	0,0486	0,0621	0,0635	0,0818	0,0842
	Bin. Nég. type 1	0,0479	0,0494	0,0621	0,0640	0,0812	0,0837
	Poisson inv.-gau.	0,0459	0,0467	0,0598	0,0611	0,0787	0,0809
	Poisson gonflée à 0	0,0459	0,0469	0,0597	0,0614	0,0787	0,0816
	MVNB	0,0488	0,0496	0,0633	0,0646	0,0830	0,0852
b	Poisson	0,0836	0,0836	0,1088	0,1088	0,1432	0,1432
	Bin. Nég. type 2	0,0836	0,0861	0,1087	0,1130	0,1431	0,1506
	Bin. Nég. type 1	0,0842	0,0868	0,1091	0,1125	0,1428	0,1472
	Poisson inv.-gau.	0,0831	0,0856	0,1081	0,1123	0,1423	0,1497
	Poisson gonflée à 0	0,0829	0,0862	0,1080	0,1136	0,1423	0,1520
	MVNB	0,0830	0,0852	0,1075	0,1113	0,1410	0,1474
c	Poisson	0,0894	0,0894	0,1164	0,1164	0,1532	0,1532
	Bin. Nég. type 2	0,0894	0,0923	0,1163	0,1212	0,1531	0,1616
	Bin. Nég. type 1	0,0900	0,0927	0,1166	0,1202	0,1526	0,1574
	Poisson inv.-gau.	0,0890	0,0919	0,1159	0,1207	0,1526	0,1610
	Poisson gonflée à 0	0,0889	0,0927	0,1158	0,1222	0,1526	0,1637
	MVNB	0,0901	0,0927	0,1167	0,1211	0,1531	0,1606

Tableau 3.28: Primes prédictives pour le modèle avancé à distribution Poisson gonflée à 0 et le modèle avancé modifié selon les 3 profils de risque et les types d'exposition étudiés

		Bas risque		Risque moyen		Haut risque	
Expo.	Modèle	Moy.	Var.	Moy.	Var.	Moy.	Var.
a	Avancé	0,0459	0,0469	0,0597	0,0614	0,0787	0,0816
	Avancé modifié	0.0459	0.0466	0.0596	0.0608	0.0786	0.0806
b	Avancé	0,0829	0,0862	0,1080	0,1136	0,1423	0,1520
	Avancé modifié	0.0830	0.0837	0.1079	0.1091	0.1422	0.1443
c	Avancé	0,0889	0,0927	0,1158	0,1222	0,1526	0,1637
	Avancé modifié	0.0892	0.0896	0.1159	0.1166	0.1528	0.1540

Tableau 3.29: Primes prédictives d'un assuré observé 10 ans avec le modèle classique à distribution MVNB selon les 3 profils de risque étudiés et une exposition d'un an et de 15 000 km (type c)

		Nombre de réclamations				
Profil	A priori	0	1	2	3	4
Bas risque	0,0689	0,0554	0,0750	0,0946	0,1142	0,1338
Risque moyen	0,1005	0,0742	0,1004	0,1266	0,1528	0,1790
Haut risque	0,1378	0,0927	0,1254	0,1581	0,1909	0,2236

Tableau 3.30: Primes prédictives d'un assuré observé 10 ans avec le modèle classique avec kilométrage à distribution MVNB selon les 3 profils de risque étudiés et une exposition d'un an et de 15 000 km (type c)

		Nombre de réclamations				
Profil	A priori	0	1	2	3	4
Bas risque	0,0851	0,0659	0,0885	0,1110	0,1335	0,1560
Risque moyen	0,1162	0,0832	0,1116	0,1400	0,1684	0,1968
Haut risque	0,1582	0,1027	0,1378	0,1728	0,2079	0,2430

Tableau 3.31: Primes prédictives d'un assuré observé 10 ans avec le modèle avancé à distribution MVNB selon les 3 profils de risque étudiés et une exposition d'un an et de 15 000 km (type c)

		Nombre de réclamations				
Profil	A priori	0	1	2	3	4
Bas risque	0,0901	0,0697	0,0923	0,1149	0,1375	0,1600
Risque moyen	0,1167	0,0847	0,1121	0,1396	0,1670	0,1944
Haut risque	0,1531	0,1023	0,1355	0,1686	0,2018	0,2349

CONCLUSION

L'arrivée constante de nouvelles technologies poussent l'ensemble des secteurs à s'adapter continuellement et l'assurance automobile n'y échappe pas. C'est le cas en ce moment même avec les appareils télématiques. Malgré leur apparition sur le marché québécois il y a déjà quelques années, les assureurs sont encore à valider et à améliorer l'utilisation qu'ils font des données recueillies par ces appareils.

Dans cette optique, Boucher *et al.* (2017) se sont intéressés aux GAM afin d'incorporer, sous une fonction de lissage, le kilométrage à la modélisation de la fréquence en assurance auto.

Afin de poursuivre leur travail, dans ce mémoire, un plus large éventail de distributions a été appliqué sur les données. Ceci a été rendu possible par l'utilisation de GAMLSS. Présentés au chapitre 2, ces modèles retirent l'hypothèse des GAM stipulant que la distribution de la variable réponse doit provenir de la famille exponentielle linéaire. Les fonctions de lissage par P-splines sont également présentées à ce chapitre.

Au chapitre 3, plusieurs types de modèles ont été proposés et appliqués. Tout d'abord, il y a eu les modèles classiques, soit des modèles qui s'approchent de ce que les assureurs font de manière plus traditionnelle. Puis, les modèles classiques avec kilométrage ont été obtenus en ajoutant des variables représentant des classes de kilométrage aux modèles précédents. Finalement, il y a eu les modèles avancés, soit des modèles intégrant des fonctions de lissage par P-splines pour le kilométrage et la durée.

Pour ces 3 types de modèles, différentes distributions discrètes de probabilité ont été utilisées. De manière générale, la distribution Poisson, la seule qui est membre de la famille exponentielle linéaire parmi les distributions analysées, est celle qui présente la moins bonne qualité générale d'ajustement. À l'opposé se trouve la distribution MVNB, la seule distribution parmi celles analysées à données longitudinales. Bien qu'elle présente un défi supplémentaire d'estimation lorsqu'on lui ajoute des fonctions de lissage, elle demeure celle qui, globalement, présente la meilleure qualité d'ajustement.

Parmi tous ces modèles, les modèles avancés sont ceux qui présentent les meilleurs AIC. L'utilisation d'une fonction de lissage pour le kilométrage a permis de constater la non-linéarité de cette variable sur la fréquence. De plus, retirer l'hypothèse de proportionnalité entre la durée et la fréquence au profit d'une seconde fonction de lissage s'est avérée justifié, puisque la courbe obtenue n'est pas linéaire.

Par ailleurs, les GAMLSS permettent aussi de modéliser des paramètres autres que celui de la moyenne avec des fonctions de lien. À cet effet, le chapitre 3 présente également un exemple d'application avec la distribution Poisson gonflée à 0. En déplaçant la fonction de lissage du kilométrage d'un modèle avancé de la fonction de lien du paramètre de moyenne à celle du paramètre de gonflement ϕ_i, on obtient un ajustement légèrement meilleur selon le critère AIC.

Afin d'améliorer les modèles présentés dans ce mémoire, plusieurs avenues devraient être étudiées. Tout d'abord, il faudrait trouver une meilleure technique d'estimation pour les modèles avancés à distribution MVNB. De plus, tester un plus large éventail de distributions à données longitudinales permettrait de comparer plus facilement la MVNB à d'autres distributions. Finalement, il serait intéressant d'avoir accès à des données télématiques beaucoup plus complètes, de façon à pouvoir intégrer davantage d'informations sur la façon de conduire des assurés

dans les modèles. Par exemple, le nombre de freinages et d'accélérations brusques, les excès de vitesse, le type de routes empruntées ou les heures de conduite pourraient être introduits dans les modèles.

RÉFÉRENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Bordoff, J. E. et Noel, P. J. (2008). Pay-as-you-drive auto insurance : A simple way to reduce driving-related harms and increase equity. the brookings institution.

Boucher, J.-P., Côté, S. et Guillen, M. (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models. *Risks*, *5*(4), 54.

Boucher, J.-P. et Denuit, M. (2006). Fixed versus random effects in poisson regression models for claim counts : A case study with motor insurance. *ASTIN Bulletin : The Journal of the IAA*, *36*(1), 285–301.

Boucher, J.-P., Denuit, M. et Guillén, M. (2007). Risk classification for claim counts : A comparative analysis of various zeroinflated mixed poisson and hurdle models. *North American Actuarial Journal*, *11*(4), 110–131.

Boucher, J.-P., Denuit, M. et Guillén, M. (2008). Models of insurance claim counts with time dependence based on generalization of poisson and negative binomial distributions. *Variance*, *2*(1), 135–162.

Boucher, J.-P., Pérez-Marín, A. M. et Santolino, M. (2013). Pay-as-you-drive insurance : the effect of the kilometers on the risk of accident. Dans *Anales del Instituto de Actuarios Españoles*, volume 19, 135–154. Instituto de Actuarios Españoles.

Côté, S. (2016). Modèles additifs généralisés dans la modélisation de l'impact du kilométrage et de l'exposition au risque en assurance automobile.

Cox, M. G. (1972). The numerical evaluation of b-splines. *IMA Journal of Applied Mathematics*, *10*(2), 134–149.

De Boor, C. (1972). On calculating with b-splines. *Journal of Approximation theory*, *6*(1), 50–62.

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C. et De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-Verlag New York.

De Jong, P., Heller, G. Z. *et al.* (2008). *Generalized linear models for insurance data*, volume 10. Cambridge University Press Cambridge.

Dierckx, P. (1995). *Curve and surface fitting with splines*. Oxford University Press.

Eilers, P. H. et Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, 89–102.

Green, P. J. et Silverman, B. W. (1993). *Nonparametric regression and generalized linear models : a roughness penalty approach*. CRC Press.

Hastie, T. et Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*, 297–318.

Hausman, J. A., Hall, B. H. et Griliches, Z. (1984). Econometric models for count data with an application to the patents-r&d relationship.

James, G., Witten, D., Hastie, T. et Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Lemaire, J., Park, S. C. et Wang, K. C. (2016). The use of annual mileage as a rating variable. *ASTIN Bulletin : The Journal of the IAA*, *46*(1), 39–69.

Litman, T. (2001). Distance-based vehicle insurance feasibility, benefits and costs : Comprehensive technical report. *Victoria Transport Policy Institute, Victoria*.

Litman, T. (2005). Pay-as-you-drive pricing and insurance regulatory objectives. *Journal of Insurance Regulation*, *23*(3), 35.

Litman, T. (2011). Distance-based vehicle insurance feasibility, benefits and costs : Comprehensive technical report. *Victoria Transport Policy Institute, Victoria*.

McCullagh, P. et Nelder, J. A. (1989). Generalized linear models, no. 37 in monograph on statistics and applied probability.

Nelder, J. A. et Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A, General*, *135*, 370–384.

Poisson, S. D. (1837). Probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilitiés. *Paris, France : Bachelier*, *1*, 1837.

Rigby, R. et Stasinopoulos, D. (2009). A flexible regression approach using gamlss in r. *London Metropolitan University, London*.

Rigby, R. A. et Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, *54*(3), 507–554.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of computational and graphical statistics*, *11*(4), 735–757.

Ruppert, D., Wand, M. P. et Carroll, R. J. (2003). *Semiparametric regression.* Numéro 12. Cambridge university press.

Schumaker, L. (2007). *Spline functions : basic theory.* Cambridge University Press.

Stasinopoulos, D. M., Rigby, R. A. *et al.* (2007). Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, *23*(7), 1–46.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V. et De Bastiani, F. (2017). *Flexible Regression and Smoothing : Using GAMLSS in R.* CRC Press.

Tselentis, D. I., Yannis, G. et Vlahogianni, E. I. (2016). Innovative insurance schemes : pay as/how you drive. *Transportation Research Procedia*, *14*, 362–371.

Verbelen, R., Antonio, K. et Claeskens, G. (2016). Unraveling the predictive power of telematics data in car insurance pricing.

Vickrey, W. (1968). Automobile accidents, tort law, externalities, and insurance : an economist's critique. *Law and Contemporary Problems*, *33*(3), 464–487.

Wood, S. (2006). *Generalized Additive Models : An Introduction with R.* CRC Press.