UNIVERSITÉ DU QUÉBEC À MONTRÉAL

APPROCHES STATISTIQUES SEMI-PARAMÉTRIQUES POUR LA MODÉLISATION LONGITUDINALE DU RISQUE EN ASSURANCE AUTOMOBILE

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN MATHÉMATIQUES

PAR

ROXANE TURCOTTE

AOÛT 2023

REMERCIEMENTS

Je tiens d'abord à remercier mon directeur de thèse, le professeur Jean-Philippe Boucher. Merci pour les commentaires et suggestions pour améliorer la qualité du contenu de cette thèse. Merci également pour le soutien et le temps pris pour répondre à mes questions sur divers sujets.

Je remercie la Chaire Co-operators en analyse des risques actuariels de l'UQAM (CARA) pour le partage des données qui ont servi à illustrer les modèles de cette thèse et pour le soutien financier tout au long du doctorat.

Je remercie également le Fonds québécois de la recherche sur la nature et les technologies et le Conseil de recherches en sciences naturelles et en génie du Canada pour leur soutien financier au cours de mes études doctorales.

Je remercie les membres de la Chaire CARA, spécialement le Professeur Mathieu Pigeon et mes collègues de bureau, Francis Duval et Juan-Sebastian Yanez, pour l'aide et les discussions diverses. J'espère vous recroiser dans le futur.

Je remercie mon ami Félix Locas pour son humour qui a su agrémenter ces années au doctorat. Je vais me souvenir de Floppa !

Un remerciement tout particulier à mes parents, Ginette et Richard, et à mon amoureux, Anthony, pour leur soutien inconditionnel. L'aide que vous m'avez apportée tout au long de mes études a fait une différence. Je vous aime tellement ! Merci infiniment.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

ACRONYMES

AIC Akaike information criterion.

BETA-NB Beta negative binomial.

BIC Bayesian information criterion.

DF Degrees of freedom.

EDF Effective degrees of freedom.

EM Expectation-maximisation.

FE Fixed effects.

GAM Generalised additive model.

GAMLSS Generalised additive model for location, scale and shape.

GLM Generalised linear model.

IRLS Iteratively re-weighted least square.

ML Maximum likelihood.

MLE Maximum likelihood estimator.

MVNB Multivariate negative binomial.

PIRLS Penalised iteratively re-weighted least square.

RE Ramdom effects.

TPRS Thin plate regression spline.

NOTATION

Nombres

x un scalaire ou une fonction.

X une variable aléatoire.

$\mathbf{A}$ un vecteur ou une matrice.

Unités

km des kilomètres.

h des heures.

RÉSUMÉ

Cette thèse de doctorat se concentre sur l'étude de modèles semi-paramétriques et longitudinaux pour la fréquence de réclamation en assurance automobile à des fins de tarification. Un des principaux objectifs de la thèse a été d'étudier le lien entre le kilométrage réellement parcouru par une automobile (collecté par un appareil télématique) et le risque d'une réclamation de type collision. Les modèles de régression additifs généralisés (GAM) et les modèles de régression additifs généralisés d'emplacement, d'échelle et de forme (GAMLSS) ont été utilisés pour permettre une modélisation plus flexible de la fréquence de réclamation. Ces modèles ont la particularité de permettre l'intégration de fonctions de lissage non paramétriques dans le prédicteur linéaire. Les fonctions de lissage qui ont été utilisées sont des splines, nommément les splines cubiques de régression, les P-splines et les splines de régression en plaque mince. La dépendance longitudinale entre les contrats a été considérée avec l'objectif de mieux intégrer l'historique de réclamation dans la tarification. Les distributions longitudinales utilisées pour la modélisation ont été construites par mélange de distributions conjuguées afin d'obtenir une expression explicite de la vraisemblance. Tous les modèles ont été illustrés à l'aide de données fournies par une grande compagnie d'assurance canadienne.

INTRODUCTION

La modélisation mathématique est au coeur de la gestion du risque dans le monde financier. Dans le domaine de l'assurance, le but même de l'activité commerciale est d'assumer le risque de perte inconnue et aléatoire en échange d'un montant fixé à l'émission du contrat. C'est donc dire que l'hypothèse de départ du principe d'assurance moderne est que l'expérience de perte puisse être résumée par une variable aléatoire, dont l'étude des caractéristiques est possible depuis l'apparition d'une théorie mathématisée des probabilités au 17e siècle. Le calcul différencié des primes d'assurance selon les caractéristiques inhérentes du risque est une responsabilité de base de l'actuaire. Entre l'apparition du principe de l'assurance environ 2000 avant notre ère et l'émergence des données massives provenant d'appareils connectés depuis une dizaine d'années, l'industrie de l'assurance a passé par différentes phases de transformation et les standards de pratique se sont graduellement formalisés. Le développement de la statistique inférentielle au 19e siècle a rendu possible le passage progressif de modèles déterministes à des modèles stochastiques permettant d'effectuer des prévisions probabilistes. Cette étape a été très importante pour estimer l'erreur associée à la prévision et mieux évaluer le risque associé à cette prévision. Ces techniques permettent de recueillir et d'analyser seulement un échantillon d'une population, pour ensuite inférer sur l'ensemble de la population. L'erreur est estimée, entre autres, en fonction de la taille de l'échantillon et du modèle choisi.

Une des tâches classiques du travail en actuariat est donc d'analyser les caractéristiques du risque, de récolter des informations sur le risque assuré, d'utiliser ces informations pour modéliser la probabilité de perte et de déterminer la prime à tarifer. L'industrie de l'assurance opère sur la base d'une tarification prospective : les réclamations passées servent à identifier les caractéristiques du risque associées à une plus grande probabilité de perte afin d'estimer le coût futur des indemnisations à verser. Pour modéliser la charge totale par unité d'exposition, la somme des indemnités à verser aux assurés par période d'assurance (généralement annuelle), on procède communément en deux étapes. On décompose la charge totale en sa composante de fréquence et de sévérité en introduisant le nombre de sinistres dans l'équation de la charge totale. Ainsi, la fréquence de réclamation correspond au nombre de sinistres par période d'assurance et la sévérité se définit comme la somme des indemnités versées divisée par le nombre de sinistres. Du produit de la fréquence et de la sévérité émane la charge totale. Cette charge est le risque financier que l'on souhaite modéliser par une variable aléatoire afin de déterminer la prime pure. La prime pure est la constante qui se rapproche le plus de . Elle représente donc la portion de la prime commerciale (prime déboursée par l'assuré) qui est impartie à compenser les pertes assurables, sans les divers frais d'exploitation de l'assureur. Pour dé-

terminer la valeur de la constante qui se rapproche le plus de , il convient de définir une mesure de distance. Bien qu'il pourrait être théoriquement possible de proposer plusieurs mesures, le contexte pratique limite les options. On utilise l'erreur quadratique pour mesurer la distance, car elle pénalise autant la surévaluation que la sous-évaluation du risque et que la constante résultante est l'espérance mathématique de . Cela permet d'obtenir une prime pure cohérente, ce qui ne serait pas nécessairement le cas avec d'autres mesures. Par exemple, l'erreur absolue ne serait pas une mesure de distance cohérente. En effet, la constante se rapprochant le plus de est la médiane. Or, la médiane de la distribution de perte est presque toujours nulle de par le fait que la plupart des risques assurés par les compagnies d'assurance non-vie pour les particuliers sont des risques de faible fréquence. Une prime nulle pour assurer un risque non nul est incohérent. Une hypothèse largement répandue étant de supposer l'indépendance entre les composantes de fréquence et de sévérité de la charge totale, il devient alors possible de séparer la modélisation de la charge totale, car l'espérance d'un produit de variables aléatoires indépendantes est égale au produit des espérances.

Dans cette thèse, il sera question de la modélisation de la composante de fréquence de la charge totale. Le modèle de base pour modéliser la fréquence de réclamation est un modèle linéaire généralisé (GLM) basé sur la distribution de Poisson. L'utilisation des GLMs s'est répandue dans l'industrie de l'assurance avec le développement au 21e siècle de la capacité computationnelle des ordinateurs et ces modèles se sont graduellement imposés comme le nouveau standard de l'industrie pour la modélisation de primes d'assurance. La théorie des GLMs repose sur la famille exponentielle linéaire. Cette famille de distributions est composée de toutes les distributions pouvant s'écrire sous une certaine forme générale. De cette forme générale, il est possible de construire une procédure d'estimation basée sur l'algorithme de Newton-Raphson. Un aspect important des GLMs est que des variables explicatives peuvent être incluses dans la modélisation du paramètre de moyenne , l'espérance de la distribution. Ceci permet d'identifier les meilleures variables de segmentation pour la tarification et d'établir un degré de certitude sur le pouvoir prédictif de ces termes paramétriques.

Malgré l'ensemble des avantages des GLMs, ce type de modèles possède plusieurs limites. D'abord, la théorie des GLMs ne permet pas l'inclusion de fonctions de lissage dans le prédicteur linéaire du modèle. Cela implique qu'un seul type de relation est possible entre les variables de segmentation et , une relation linéaire croissante ou décroissante. Les modèles additifs généralisés (GAM) sont une extension semi-paramétrique des GLMs. L'inclusion de termes non paramétriques comme des fonctions de lissage dans le prédicteur per-

met d'inclure de la flexibilité entre la variable explicative et le paramètre de moyenne de la distribution. Les GAMs peuvent donc être utilisés pour étudier le lien réel entre des variables continues et le risque.

En assurance, il est possible d'observer le même risque sur plus d'une période d'assurance, mais les modèles classiques ne permettent pas d'inclure une dépendance temporelle entre les contrats d'un même assuré. Pour un GAM, comme pour un GLM, la distribution sous-jacente choisie doit faire partie de la famille exponentielle linéaire. Ceci implique que plusieurs distributions utiles à l'actuariat ne peuvent pas être modélisées à l'aide de ces cadres. Afin d'utiliser des distributions permettant de modéliser la dépendance longitudinale entre les contrats d'un même assuré comme la binomiale négative multivariée, dont la distribution prédictive est une loi binomiale négative, nous nous tournons plutôt du côté des modèles de régression additifs généralisés d'emplacement, d'échelle et de forme (GAMLSS). Les modèles GAMLSSs sont à la fois une généralisation des GLMs et des GAMs, puisqu'il est possible d'inclure des fonctions de lissage dans la modélisation des paramètres.

La flexibilité additionnelle amenée par les modèles GAMs et GAMLSSs sera essentielle pour les modèles de fréquence abordés dans cette thèse. Avec le développement des technologies GPS, les compagnies d'assurance sont de plus en plus nombreuses à collecter des données télématiques. Ces systèmes permettent entre autres de collecter des informations comme la vitesse, l'accélération et le kilométrage qui concernent toutes directement le risque assuré. Cela permet d'approcher différemment la tarification et offre le potentiel de remplacer certaines variables de substitution (*proxy variables*, en anglais) qui ne font qu'exposer que certains comportements plus risqués sont plus courants parmi certains groupes, mais ne caractérisent pas directement le risque. Cela explique entre autres l'utilisation du genre, du statut marital ou du niveau d'étude dans la tarification en assurance automobile. Nous nous concentrerons principalement sur l'étude du kilométrage et sur la possibilité de l'utiliser comme mesure d'exposition au risque, en plus d'étudier plus largement la tarification au mérite à l'aide de modèles longitudinaux.

Spécifiquement, cette thèse est composée de trois chapitres, chacun correspondant à un article scientifique. Le premier chapitre porte sur l'étude de la relation entre le kilométrage réellement parcouru par une automobile et la fréquence de réclamation en utilisant des données télématiques. Cet article a été publié dans le journal *Risks* (Boucher & Turcotte, 2020). La relation a été étudiée en introduisant des P-splines dans les modèles de régression GAM et GAMLSS. Des effets fixes et aléatoires ont été inclus dans la modélisation pour introduire une dépendance entre les contrats d'un même assuré à travers le temps. Pour observer cor-

rectement la relation entre la distance parcourue et la fréquence des sinistres, il est montré qu'il convient d'utiliser une distribution de Poisson avec des effets fixes, car elle élimine l'hétérogénéité résiduelle qui n'a pas été correctement prise en compte par les modèles précédents basés sur les théories GAM et GAMLSS. Il est montré qu'il est possible de dériver une relation approximativement linéaire entre le kilométrage et la fréquence de réclamation.

Dans le deuxième chapitre, un modèle de type Bonus-Malus a été développé basé sur la modélisation semi-paramétrique d'une distribution longitudinale. Des variables explicatives continues ont été incluses dans la modélisation en utilisant des P-splines et des splines cubiques de régression pour améliorer le pouvoir prédictif. Une analyse des répercussions de différents scénarios de réclamations sur la prime selon les caractéristiques du modèle (paramétriques ou semi-paramétriques) a été réalisée. À titre comparatif, des modèles transversaux paramétriques et semi-paramétriques ont également été ajustés. Nous avons observé que la forme générale des splines associées à la même variable explicative était semblable, peu importe le type de modèle, ce qui peut s'avérer utile pour segmenter une variable continue en variable catégorielle. Finalement, la méthode d'estimation des modèles GAM et GAMLSS est expliquée afin de permettre l'utilisation des modèles GAMLSS pour la modélisation de n'importe quelle distribution utile à l'actuariat. En particulier, il est question du problème d'identification du modèle lorsque plusieurs splines sont incluses dans un modèle de régression comportant déjà un terme d'ordonnée à l'origine. Le deuxième volet de cette thèse a été publié par le *North American Actuarial Journal* (Turcotte & Boucher, 2023).

Le dernier chapitre reprend certains éléments du premier chapitre, mais pousse plus loin l'étude de la relation entre le kilométrage et la fréquence de réclamation. Cette fois, des contraintes de forme sur la spline sont imposées, en plus de considérer un plus grand éventail de splines, soit les P-splines, les splines cubiques de régression et les splines de régression en plaque mince. Les différences entre ces splines sont abordées. Pour qu'une métrique puisse être utilisée comme mesure d'exposition, elle doit, entre autres, être simple à calculer pour l'assureur, facile à comprendre pour les assurés et cohérente. Dans les premiers modèles, des irrégularités de forme pour une mesure d'exposition ont été observées. La spline associée au kilométrage devenait décroissante pour les quantiles élevés de la distribution, ce qui n'était pas cohérent. Ce résultat venait confirmer celui déjà observé sur des données espagnoles (Boucher *et al.*, 2017). Il est cependant nécessaire que la relation estimée entre le kilométrage et le risque soit toujours croissante pour respecter le principe que, plus on est exposé au risque, plus la probabilité d'accident est grande. L'objectif est donc de répondre à certaines incohérences de forme observées dans le premier projet, car le modèle

à effets aléatoires, contrairement au modèle à effets fixes, peut être utilisé en pratique pour la tarification. Cet article sera soumis pour publication.

CHAPITRE 1

A LONGITUDINAL ANALYSIS OF THE IMPACT OF DISTANCE DRIVEN ON THE PROBABILITY OF CAR ACCIDENTS

In the past decade, new technologies such as GPS-collected data have emerged, which offer new ways to approach car insurance pricing. Processing these data provides reliable information about drivers' behaviour. Before GPS and telematics devices, the insurance industry had to rely on proxy variables such as territory, gender and age of the drivers to measure risk. However, such covariates only describe the general behaviour of insured in those groups. For example, (Ayuso *et al.*, 2016b) shows that the differences observed in claims frequency between men and women are largely attributable to vehicle use ; (Verbelen *et al.*, 2018) reached a similar conclusion. In a social-political context where the use of gender in ratemaking is restricted or criticized, calculating premiums on more objective information is of interest.

One piece of GPS-collected information that is directly related to the risk insured is distance driven. The relevance of including this variable in ratemaking has been studied by (Ayuso *et al.*, 2014), (Ayuso *et al.*, 2016a), (Boucher *et al.*, 2013) and (Lemaire *et al.*, 2016), among others. (Boucher *et al.*, 2017) studied the effect of distance driven and policy duration time on claim frequency and challenged the usual ratemaking practice of using contract duration as the risk exposure measure. Mileage-based pricing can generate several benefits, notably on the environment, because it encourages policyholders to reduce their annual mileage. Establishing premiums on the basis of variables that the insured can control has the significant advantage of encouraging a positive change of habit in policyholders (see for example (Bolderdijk *et al.*, 2011) and (Tselentis *et al.*, 2016)). One can argue that distance driven is correlated with other driving habits resulting from driving experience (Ferreira & Minikel, 2010). Hence, if the model does not take this correlation into account, the resulting relationship between claim frequency and the distance driven would not give an appropriate representation of how the claim frequency could change when insureds change their driving habits. This is precisely what is tackled in this paper : we indeed focus on the "marginal" effect of distance driven. The objective of our paper is not to compute a premium, but mainly to understand how the distance impacts the claim frequency when all individual characteristics of policyholders have been considered.

We focus our analysis on the distance driven, yet other telematics variables could be of interest. In the study by (Verbelen *et al.*, 2018), driving time (daytime vs. nighttime) is studied along with the type of roads,

while (Ma *et al.*, 2018) find that speed and acceleration affect the expected claim frequency. (Ayuso *et al.*, 2014) analyse the effect of various covariates on the time before the first crash, and compare novice and experienced drivers. More recently, (Ayuso *et al.*, 2019) proposed to improve the traditional ratemaking methods by including information related to risk exposure and driving behaviour of insured. (Denuit *et al.*, 2019) use predictive rating with past telematics information in a credibility model. (Weidner *et al.*, 2016) study driving behaviour and vehicle use on different scales of analysis (maneuver, trip or insurance period) by means of form recognition and Fourier analysis methods. (Wüthrich, 2017) proposes to use speed and acceleration heat-maps to classify drivers into groups using K-means clustering. Each group is associated within a driving style and included as a categorical variable in a regression analysis. (Gao & Wüthrich, 2018) performed principal component analysis using singular value decomposition and bottleneck neural networks. The authors argue that a representation in two dimensions is sufficient to preserve most of the driving information, meaning that it is possible to obtain continuous representations with small-dimensional data. This representation could then be included in a Generalised Additive Model (GAM), as in the study by (Gao *et al.*, 2019). (Verbelen *et al.*, 2018) evaluate the predictive power and interpretability of telematics variables on claim frequency by comparing various types of models that include or exclude those telematics variables. The authors find that the best ratemaking structure includes both telematics and traditional covariates, while considering duration and mileage as exposure measures.

In Section 1.1, we present the dataset used for the numerical applications throughout this work, and we compare different exposure measures. In the following section, we used a GAM Poisson, as did (Boucher *et al.*, 2017), to link the distance driven with the number of claims. We observe the same relationship between distance and claims frequency ; however, we reject the "learning effect" explanation proposed by previous authors to explain the relationship, which we posit can be explained by the residual heterogeneity incorrectly captured by the underlying GAM model. Section 1.3 presents panel count data models that are better suited to explain individual heterogeneity. In Section 1.4, using Generalised Additive Models for Location, Scale and Shape (GAMLSS, see (Rigby & Stasinopoulos, 2005)) theory that generalises GAM, a multivariate count distribution for all the contracts of the same insured is developed, and a penalised log-likelihood is used to estimate the parameters. In Section 1.5, we use another approach based on a Poisson distribution with fixed effects to account for all individual characteristics, and show that an approximately linear relationship between the distance driven and claim frequency can be found. Section 1.6 concludes.

1.1 Summary of the Database

The dataset that has been used for our numerical analysis comes from an important Canadian P&C insurance company. We focus our analysis on personal car insurance from the province of Ontario.

In analysing telematics data, we must be careful before jumping to general conclusions about the driving behaviour of the whole portfolio. Indeed, policyholders who decided to place a telematics device on their car, or to download an application on their phone that tracks all their car trips, do not correspond to the general driver population. In our case, approximately 10% to 15% of the insurance company's portfolio chose to use the telematics option for their car insurance. Typically, these insureds correspond to one of the two following profiles :

1. Policyholders who are technophiles : they love new telematics technology and want detailed information about their driving habits. Summary driving data is indeed continuously available to policyholders via a website.
2. Young and/or bad drivers. To motive policyholders to buy the telematics option, insurance companies often offer an initial discount, and the renewal discounts range from 0% to 25% depending on driving experience [1]. Because auto insurance in Ontario is very expensive and often unaffordable for some drivers, all discounts are welcome for policyholders with high insurance premiums. As a result, an unusually high proportion of risky insureds uses telematics devices or telematics app.

In the dataset used, we observed the insureds for up to six insurance periods, with an average of 1.77 contracts per policyholder (see Table 1.1 for details). Only policyholders that have been observed at least 100 days were retained for the analysis. Since this is real data, it may contain some minor irregularities. The same table shows statistics for the number of claims, where we only kept claims related to road accidents. Indeed, we wanted to study accidents related to car usage and not, for example, those caused by floods, hail, theft or vandalism. The table shows statistics for a single insured period. We note that most policyholders do not claim, that the average claim frequency for the portfolio is 6.0%, and that the maximum number of claims observed is 3.

1. Please note that it was not legally possible for an Ontario insurance company to increase the insurance premium based on the telematics information when these data were collected.

1.1.1 Risk Exposure Measures

Table 1.2 summarizes the statistics of various risk exposure definitions :

1. Exposure time (the time between the start and the end of the insurance contract)
2. Distance driven
3. Number of trips
4. Hours driven.

Another candidate for risk exposure might be the self-reported approximation of the distance driven by the insured. However, as shown by many authors, such as (Lemaire *et al.*, 2016), the self-reported distance driven is not reliable and is often very different from the exact distance driven.

Exposure time, traditionally used by insurers, would be an appropriate measure of risk exposure if every driver had about the same car usage, which is not the case. Indeed, Table 1.2 shows that for an insured period, insureds drove between 7.1 and 76,272 km, with an average of 10,398 km. More specifically, the database also informed us about various types of car use by the insureds :

1. The maximum number of trips observed is 3,317 for a single insured period ; another insured used their car only 15 times.
2. A policyholder drove their car only for an hour during the whole insured period ; another driver used their car for more than 3000 h.

Consequently, there are important differences between driving uses and driving habits, which justifies the consideration of other measures than exposure time in the modeling.

Number of Insurance Periods	**1**	**2**	**3**	**4**	**5**	**6**
Number of policyholders	12 562	9 746	3 420	844	415	11
Proportion (%)	46.5	36.1	12.7	3.1	1.5	0.0

Table 1.1 – Distribution of the number of insurance periods for the database

	Average	Variance	Min.	Max.	25th pct	50th pct	75th pct
Exp. Time (in years)	0.645	0.060	0.277	1.079	0.463	0.540	0.912
Dist. Driven (in km)	10,398	55,138,376	7.1	76,272	5026	8561	13,836
Nb. of Trips	1083	383,165	15	3317	621	946	1434
Time Driven (in hours)	380	34,740	1	2159	248	356	483
Nb. of claims	0.060	0.061	0.000	3	0	0	0

Table 1.2 – Descriptive statistics for a single insured period

Figures 1.1 to 1.4 show histograms of different risk exposure measures under study. Except for exposure time, every other risk exposure distribution is right-skewed. Table 1.2 foreshadowed this result as the average was greater than the median for those risk exposures. This is another indication that some insureds make full use of their insurance time by making greater use of their car.

Figures 1.5 to 1.8 illustrate the links between claim frequency and risk exposure. A fairly clear linear trend seems to be emerging for the three non-traditional exposure measures for the first part of their respective curve, which contains most of the observations. However, we observe a strange relationship between the claims frequency and the risk exposures for higher quantiles of the distributions. We specify that each point on these graphs does not represent the same number of policyholders. Darker dots represent a larger number of policyholders.

Between the three usage-based exposure measures, our choice for a more detailed analysis is the distance driven. First, it seems to be the more objective risk measure out of the three. Indeed, the definition of a “trip” is not clear. For example, if the driver makes a quick stop to buy gas, does it count for one or two trips because the engine stopped ? For hours driven, does the time spent stopped at red lights and stuck in traffic count similarly to when the vehicle is moving ? Second, it would be hard to measure exposure only according to the number of trips from a marketing point of view because those who use their vehicle only to drive short distances would probably find it unfair.

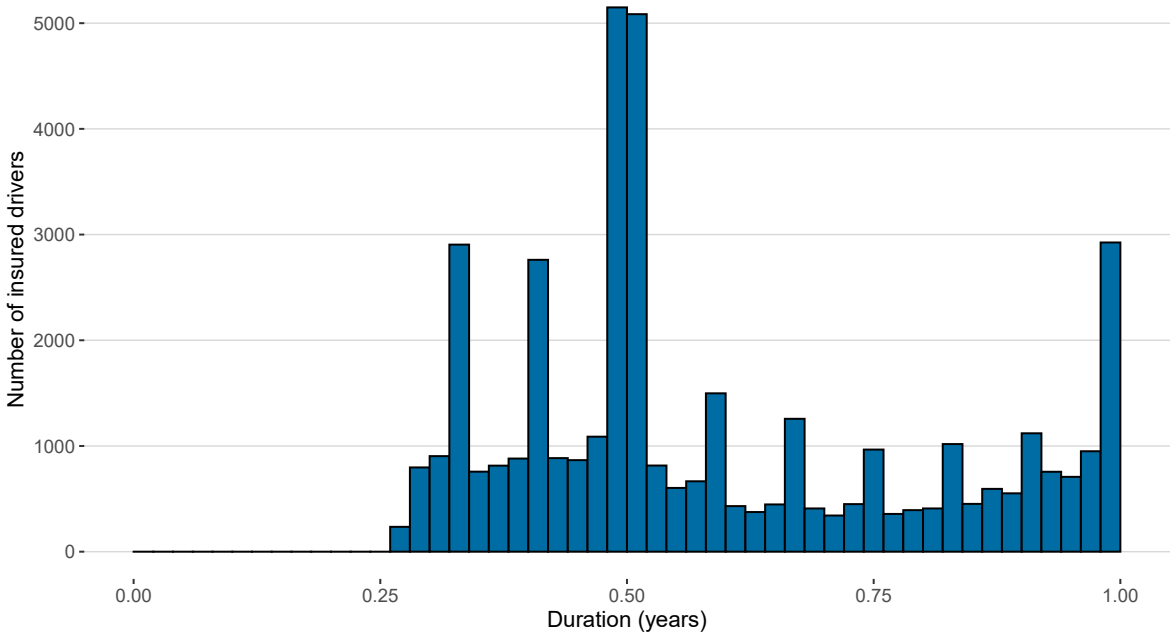

Figure 1.1 – Histogram of risk exposure (in years). Each band has a length of 0.02 year.

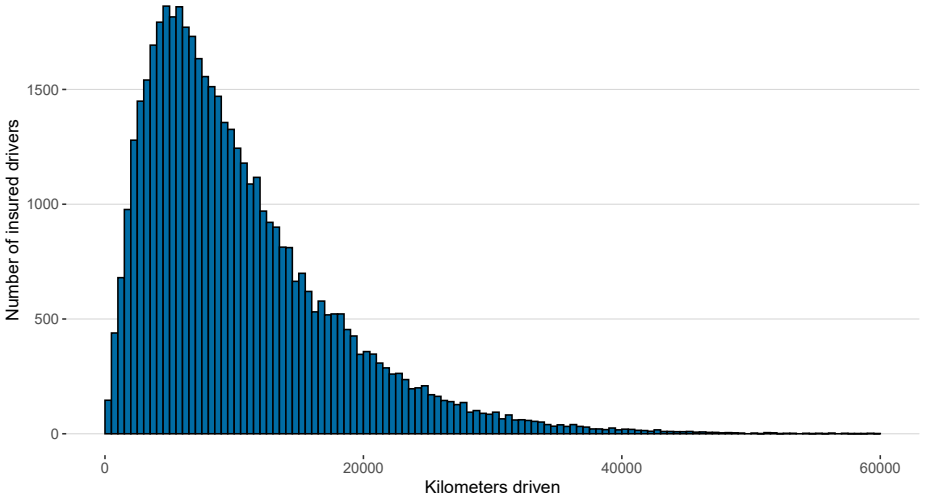

Figure 1.2 – Histogram of distance driven (in km). Each band has a length of 500 km.

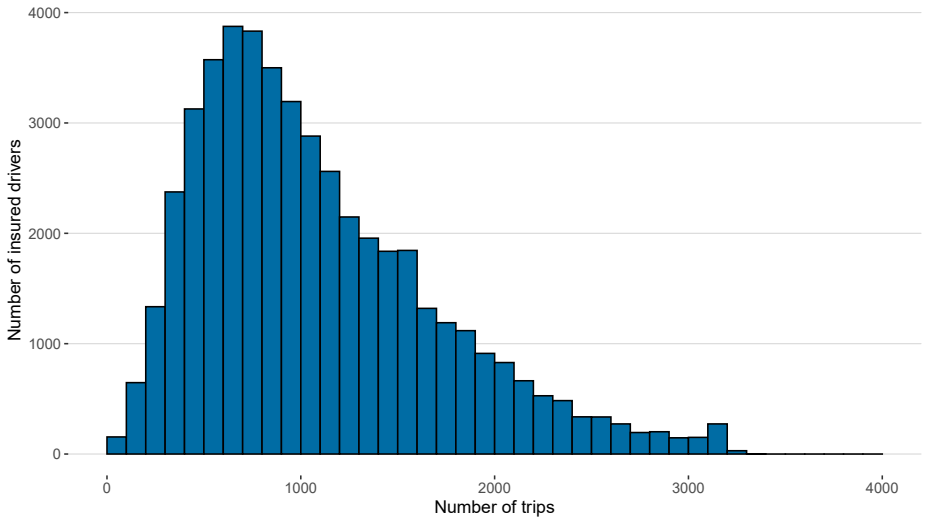

Figure 1.3 – Histogram of the number of trips. Each band has a length of 100 trips.

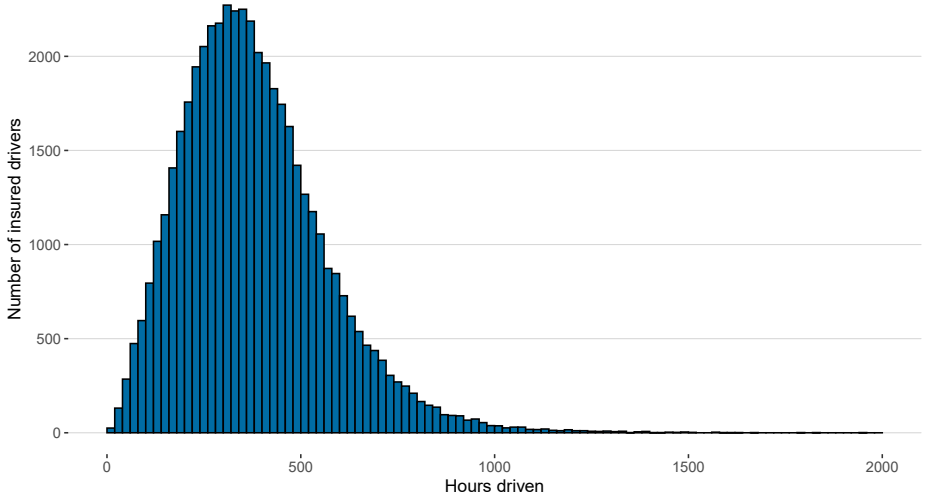

Figure 1.4 – Histogram of hours driven. Each band has a length of 2000 h.

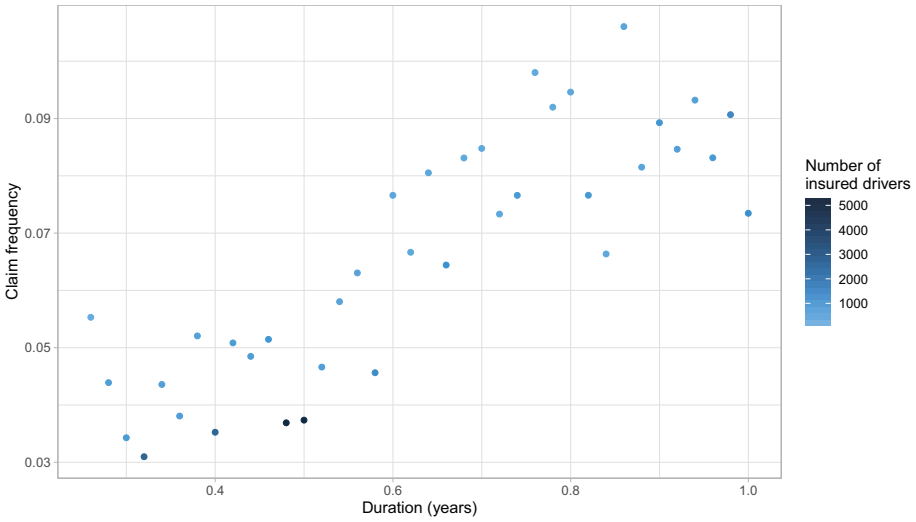

Figure 1.5 – Claims Frequency vs. Exposure Time. Data grouped by 0.02 year.

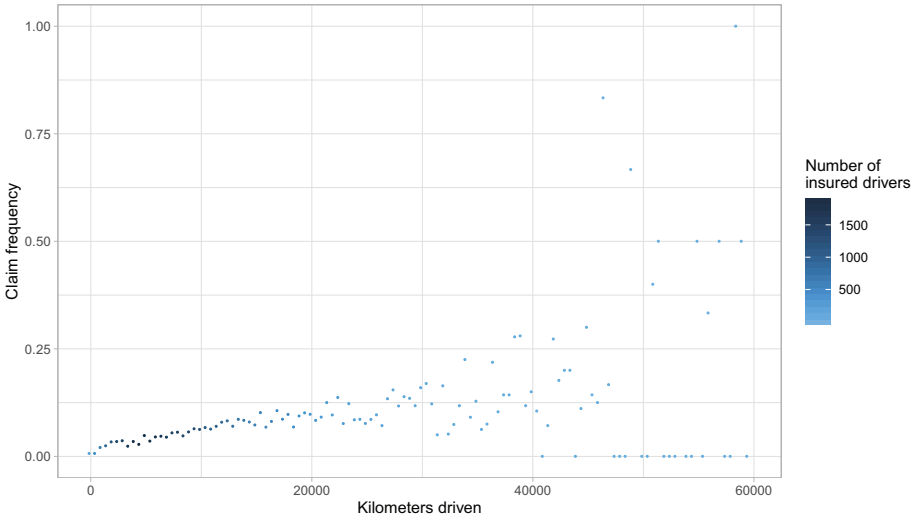

Figure 1.6 – Claims Frequency vs. Distance Driven. Data grouped by 500 km.

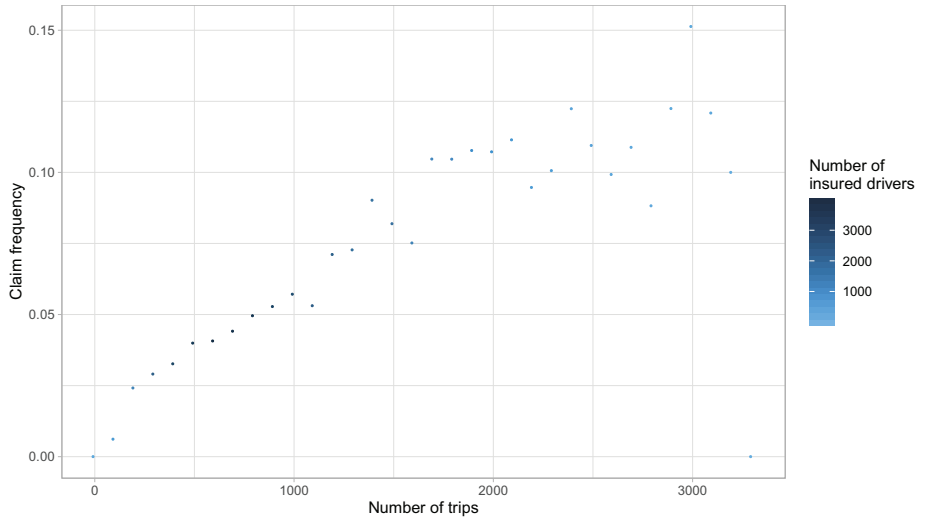

Figure 1.7 – Claims Frequency vs. Number of trips. Data grouped by 100 trips.

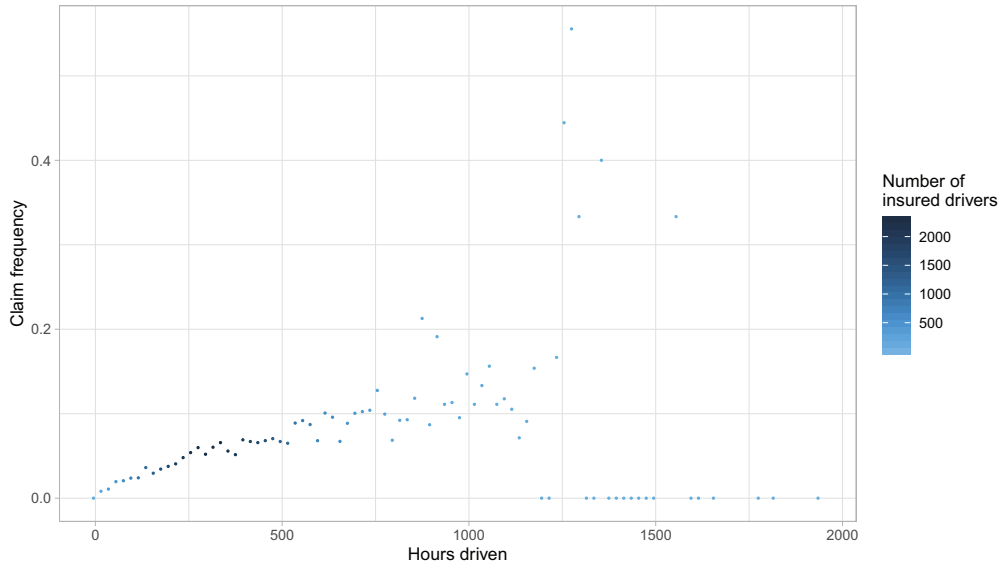

Figure 1.8 – Claims Frequency vs. Hours Driven. Data grouped by 20 hours.

1.2 Preliminary Risk Exposure Analysis

Traditionally, the starting point for the modeling of the number of claims N_i from a policyholder i exposed to the risk of a term t is modeled by a Poisson distribution of average $t\lambda_i$, where λ_i includes classic covariates used in pricing. For a Poisson regression, t is often referred to as an offset variable. Similarly, for an insured driving a car over a distance of d km, we are seeking a model with this form of proportionality between

distance driven and the expected claim frequency. Also, it could be interesting to combine and in the same model. To do this, one avenue is to use generalised additive models (GAM).

GAMs, introduced by (Hastie & Tibshirani, 1986), are an extension of the generalised linear models (GLM) theory. Consequently, as for the GLM, only distributions belonging to the linear exponential family could be used as the distribution of the response variable of a GAM. In a GLM, the linear predictor for an individual is given by $( ) =$, where $= [\ _1 \quad _2 \quad _3 \quad]$ is a vector of covariates and is a coefficient vector. For a GLM, the mean is given by a linear expression through a link function : GAMs relax the hypothesis of linearity, and smoothing functions of the covariates could be included in the predictor. For example, the mean for an individual could be given by $( ) = \ _0 + \ _1(\ _1) + \ _2(\ _2) + \ _3(\ _3)$, where $_0$ is an intercept, are smoothing functions and are covariates for $1 \ 2 \ 3$.

(Boucher *et al.*, 2017), by using a GAM Poisson model, analysed the influence of duration and distance driven on the number of claims with independent cubic splines and splines with a tensor product to introduce a dependence between those two risk exposure measures (see (Green & Silverman, 1993) for additional details on these smoothing functions). The model with independent cubic splines is the starting point of our analysis, and we evaluate the performance of this model on our data. The model $\log( ) = \ _0 + \ _1(\quad) + \ _2(\ )$ yields similar results to those obtained by (Boucher *et al.*, 2017), as it can be seen in Figure 1.9. Indeed, we observe a strongly increasing function for the first kilometers, then it stabilizes around 40,000 km. For the higher quantile of the distribution, there are very few observations, and the confidence interval is too wide to draw conclusions. As for $_2(\ )$, we observe a positive effect for the duration time, but no linear relationship because the function tends to stabilize. For the sake of completeness, the model with a tensor product has been fitted to our data. The tensor product includes dependency between the two exposure measures, and the fitted surface had a shape similar to that of (Boucher *et al.*, 2017). Considering the important differences between the European dataset used in (Boucher *et al.*, 2017) and our North American data, the similarity of the results is fairly interesting. First, climatic conditions are not the same given the significant accumulations of snow on the ground during Canadian winters. Second, the profile of policyholders between the two databases is not the same. The Spanish data focused exclusively on young drivers, while Canadian data's profiles are more diverse as explained in Section 1.1. It can also be noted that the data are collected over different years for the two databases, and the regulations differ from one country to another.

We investigate the smoothing functions on the scale of the response level ($\exp(\ _1(\quad))$ and $\exp(\ _2(\ ))$) to expose the multiplicative effect on , as illustrated in Figure 1.9. Specifically, with a log link function, we have

$$\begin{aligned} &= \exp(\qquad + \ _1(\quad) + \ _2(\ )) \\ &= \exp(\ _1(\quad))\exp(\ _2(\ ))\exp(\qquad) \\ &= \exp(\ _1(\quad))\exp(\ _2(\ )) \end{aligned} \tag{1.1}$$

In the study by (Boucher *et al.*, 2017), a "learning effect" is advanced to justify the look of $\ _1(\quad)$ (and $\exp(\ _1(\quad))$), where the expected number of claims seems to decrease as kilometers driven increases. We think that this effect cannot be used as an explanation. Indeed, most drivers in the insurance portfolio already have many years of driving experience. We do not think the extra 10,000–20,000 km adds enough experience to observe a learning effect.

Instead, we think that the shape of the smoothing function comes from the driver profiles : the lower quantiles of the distribution of the distance driven does not come from the same (type of) drivers as the higher quantiles. This means that models based on Figure 1.9 cannot be used to understand the relationship between the distance driven and the number of claims, and might not be used to set the premium for insured that suddenly change their driving habits, because it does not nearly tell us how their risk is changing.

As an example to illustrate the situation, we can suppose an insured who suddenly decides to drive 50,000 km instead of 40,000 km. Based on Figure 1.9, we would expect a decrease in the expected claims frequency. This is however impossible : the number of claims in the first 40,000 km cannot change, and the extra 10,000 km can only add other claims. In other words, if insureds choose to drive their cars rather than leaving it at home, the risk should always be greater. The slope could change as distance increases, but it should always be strictly positive since the risk is greater, meaning that the smoothing function (as the one observed in Figure 1.9) should always be increasing.

Our results, and those of (Boucher *et al.*, 2017), do not show a strictly positive relationship between claim number and distance driven. We think this can be explained by the residual individual heterogeneity of the model, which the basic Poisson GAM does not seem to capture correctly. One explanation comes from the fact that GAM supposes independence between all contracts of the same insured. We think that a more general model that relaxes this assumption should be used to correctly measure the impact of the distance

driven on the risk of accidents.

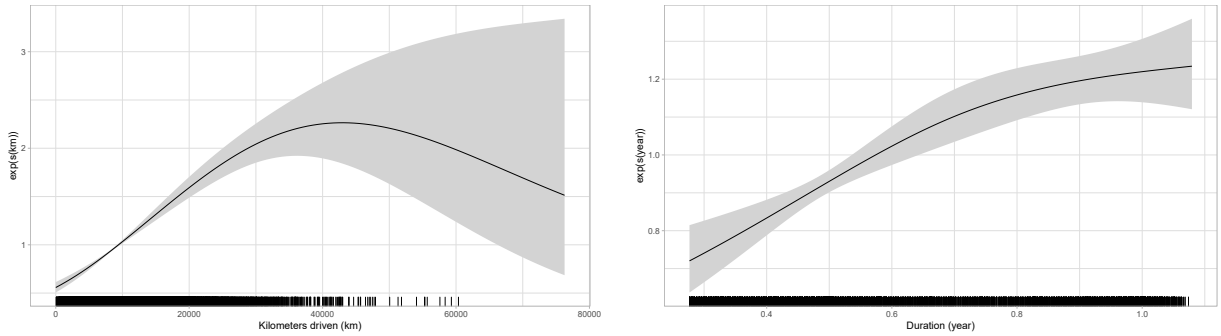

Figure 1.9 – $\exp(\ _1(\quad))$ and $\exp(\ _2(\quad))$ from the Poisson GAM estimated with Canadian data

1.3 Panel Data Modeling

When all observations are considered independent, we usually refer to it as cross-sectional data. Basic GLM and GAM are constructed under such an assumption. In non-life insurance, however, we can observe the same insured over many contracts. Consequently, we can generalise the approach by supposing a dependence between all those contracts. This dependence is usually justified by the fact that many important factors cannot be used as covariates in ratemaking.

Formally, suppose that we observe an insured for over contracts. To simplify the notation, the subscript is dropped and becomes . Instead of modeling the marginal distribution of each for $= 1$, we are now looking for the joint distribution (subscript is removed for convenience) :

$$\begin{aligned} & \Pr(\ _1 = \ _1 \quad _2 = \ _2 \qquad = \quad) \\ = \ & \Pr(\ _1 = \ _1) \quad \Pr(\ _2 = \ _2 \quad _1 = \ _1) \qquad \Pr(\quad = \qquad _1 = \ _1 \qquad _1 = \quad _1) \end{aligned}$$

where , $= 1$, is the number of reported accidents for insured period . There are many ways to construct multivariate count models (see (Inouye *et al.*, 2017) or (Molenberghs & Verbeke, 2006) for example). One popular way, which draws a parallel with the explanation of the unused covariates in the modeling, is to include an individual parameter in the mean parameter of the count distribution of each contract , which means that we have :

$$\text{Poisson}(\quad = \qquad) \tag{1.2}$$

where $= \exp(\quad)$ for 1 and 1 . For an insured , the key to the dependence then lies in the parameter which affects all the random variables for $= 1$. We can consider two different situations regarding this parameter :

1. All , $= 1$ are i.i.d. random variables that come from a selected prior distribution (we call this the random effects model, studied in detail in Section 1.4) ;
2. All , $= 1$ are unknown parameters that need to be estimated (we call this the fixed effects model, studied in detail in Section 1.5).

In both cases, random and fixed effects models give us the flexibility to create a joint distribution that allows for time dependence. However, even if they share some similarities, random and fixed effects models are different, and the differences between them are highlighted when we consider telematics data and the distance driven.

1.4 Random Effects

1.4.1 Model Specification

In random effects models, we suppose that , $= 1$, are random variables, with prior density $(\ )$. Conditionally on the random effects , all numbers of claims $_1 \quad _2$ from insured are independent. As shown in (Denuit *et al.*, 2007), the joint distribution of $_1$ can be expressed as :

$$\Pr[\quad _1 = \quad _1 \qquad = \quad] = \quad _0 \qquad \exp(\qquad)\frac{(\qquad)}{!} \qquad (\quad) \tag{1.3}$$

Many distributions can be used for , such as the gamma or the inverse Gaussian. If we suppose that follows a gamma distribution of mean 1 and variance $\frac{1}{}$, the joint distribution can be expressed as :

$$\Pr[\quad _1 = \quad _1 \qquad = \quad] = \quad \frac{(\quad)}{!} \quad \frac{(\quad + \ )}{(\ )} \quad \frac{}{+} \qquad +$$

where $= \quad _{=1}$ and $= \quad _{=1}$. This well-known distribution is the multivariate negative binomial distribution, or simply MVNB. This distribution is a generalisation of the negative binomial distribution. It is a basic distribution for panel count data modeling with overdispersion ($[\quad] =$ $[\quad] = \quad +(\quad)^2$). The correlation between and $_{+1}$ is $\text{ov}(\qquad _{+1}) = \qquad _{+1}$.

It can be shown that the first-order condition to obtain is :

$$\sum_{=1}^{} \sum_{=1}^{} \frac{+}{+} = 0 \tag{1.4}$$

This model is based on a distribution which is not a member of the linear exponential family. That means that GAM theory cannot be used to include smoothing functions. Instead, we use Generalised Additive Models for Location, Scale and Shape (GAMLSS) (see (Rigby & Stasinopoulos, 2005)) theory, that can be used for other distributions than the members of the linear exponential family of distribution. Moreover, a GAMLSS is more flexible because it can model a location parameter , a variance parameter (scale), a skewness parameter and a kurtosis parameter as additive functions of the covariates. The general form is given by

$$( ) = + \sum_{=1}^{} \tag{1.5}$$

where = . and are vectors with n elements. For each ( ), it is possible to add the desired number of additive terms. These terms could be, for example, smoothing functions or random effects.

A model does not need to specify each of the components of . For example, it is possible to use a GAMLSS that specify only the location parameter. In this case, would simply become = . For our telematics data, we choose to model the parameter with smoothing function by Equation (1.5), and is kept constant for all individuals.

Please note that in Equation (1.5), represents the parametric part that is present in GLMs and $\sum_{=1}^{}$ is the non-parametric part. is a known design matrix of dimensions and is a vector of parameters of length , which corresponds to the number of covariates in the parametric part of the model. As for and , they are respectively a known design matrix of dimensions and a vector of random variables of length . The shapes of and depend on the additive functions used. If a smooth function can be expressed in linear form, Equation (1.5) can be rewritten as

$$( ) = + \sum_{=1}^{} ( )$$

where is a smooth non-parametric function.

1.4.2 Numerical Illustration

To use GAMLSS, many distributions are available in the R package *gamlss*. Unfortunately, the MVNB distribution is not one of them (the distribution is however implemented by itself in the package *multinbmod*). Consequently, we have to write our own code for convenience. As shown in Equation (1.6), to fit the model, we maximise a penalised log-likelihood function , integrating a quadratic penalty , where is a penalty matrix for the vector of random effects parameters . The penalty matrix is very often define as , where different formulations are possible for . In our work, is a ( ) difference matrix of order r as defined in (Rigby & Stasinopoulos, 2005), taking $= 2$ (default). corresponds to the argument "order" in R function *pb.control* for P-splines. A hyper-parameter, noted here $^{+}$, controls the weight given to the penalty and, thus, the smoothness of the smoothing function. The greater its value, the smoother the resulting estimated function. $= 2$ penalties are added in the log-likelihood, one for each smoothing function included in the model. stands for the log-likelihood of the joint distribution associated with Equation (1.3).

$$= \quad 0\,5 \sum_{=1}^{2} \sum_{=1} \qquad (1.6)$$

The model is constructed as follows. As in Section 1.2, the mean parameter is represented by (1.1). On the other hand, unlike Section 1.2, we now work with cubic P-splines as our smoothing functions. Choosing the optimal number of nodes in a regression spline is not an obvious task. In Section 1.2, the number of nodes was determined by trial and error with different combinations for the two smooth functions. The number of nodes was chosen graphically using the representation of the smoothing function (Figure 1.9) to compromise between the accuracy of the data and smoothness. We now try a different approach that does not require us to select several nodes.

In a P-spline, we choose a relatively large number of knots, and wiggliness is controlled by a penalty parameter for each smoothing function. For instance, we used 20 knots for each spline, but "a relatively large number of knots" depends on your context and data. P-splines are smoothers based on B-splines with a difference penalty on coefficients of adjacent B-splines, which are strictly local polynomial functions (of a degree three, for our use). For further information on B-splines and P-splines, refer to (Eilers & Marx, 1996) and (Wood, 2017). P-splines have the advantage of offering flexibility without being cumbersome to implement.

To select the penalty parameters in ( ) associated with both P-splines of the contract duration and the

distance traveled, we test out multiple combinations of values of $= \ _1 \ _2$. We proceed in two steps : we first adjust the model for all the couples of a grid of parameters. Large steps are used to cover an interval ranging from small values to very large values for $_1$ and $_2$. Then, we examine the regions in which the parameter value models with the best AIC are obtained, and we restart a more specific search in these regions with smaller steps. Following multiple estimations of models, the best model was selected based on the AIC criteria that consider the number of *effective* degrees of freedom and the interpretability of the results.

Please note that this model was also fitted with a few covariates : gender (female or other), marital status (married or other) and vehicle usage (commute, pleasure or other). As shown in Equation (1.1), covariates can be added in the parameter. Adding covariates does not change the shape of the splines, but it does tend to increase the value of as more heterogeneity is explained.

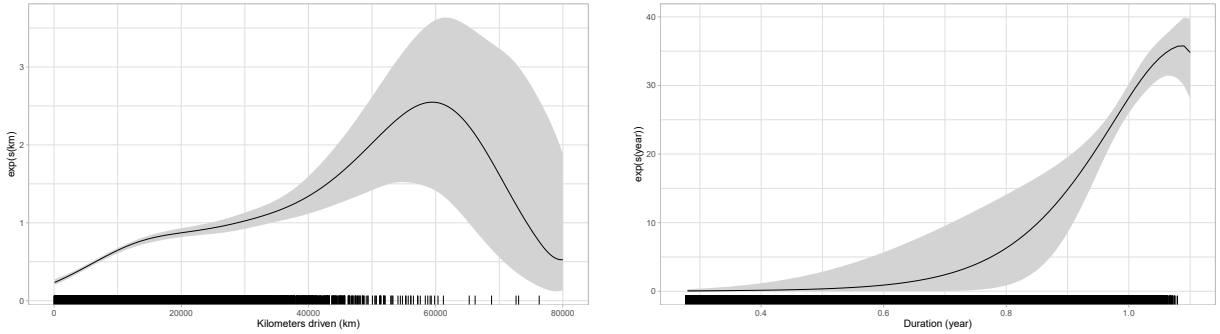

Figure 1.10 – $\exp(\ _1(\ \ ))$ and $\exp(\ _2(\ \ ))$ from the GAMLSS with random effects model estimated with Canadian data

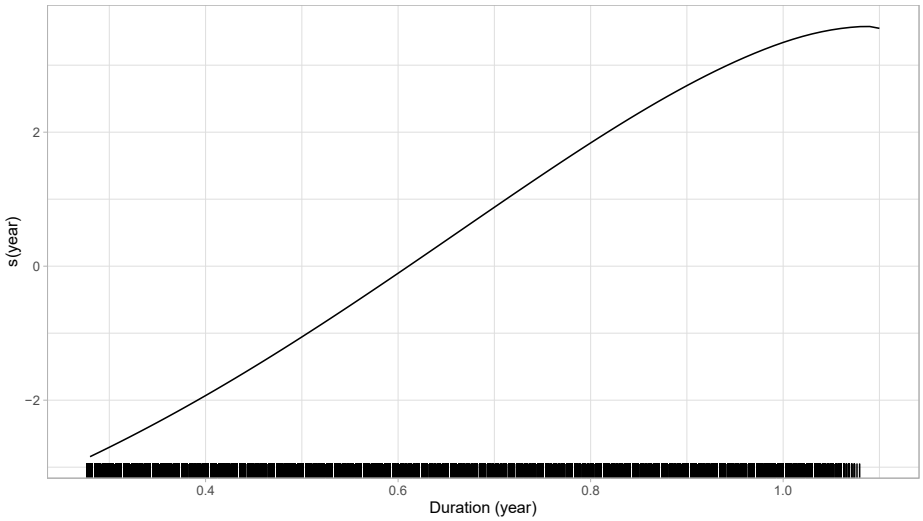

Figure 1.11 – $_2(\quad)$ from the GAMLSS with random effects model estimated with Canadian data

The fitted smooth functions are illustrated at Figure 1.10. We can see that the functions are very different from Figure 1.9. Indeed, even if we observed a similar decrease for the upper quantile of the distribution for the distance traveled, the highest point before the decrease is at a later point where the data are very scarce. This can be seen as an improvement over the previous model : we obtain a strictly positive relationship between claim number and distance driven until 60,000 km driven. For the contract duration, Figure 1.11 shows a nearly proportional relationship between the time exposure and the claims frequency, hence similarities with Figure 1.9.

As mentioned, the MVNB distribution is not available in the gamlss package and the confidence bands displayed in Figure 1.10 were generated by bootstrap. We turn to this alternative, because the penalty in P-splines does not allow calculating confidence intervals as easily as in the case of B-splines. Bootstrap is a common approach to deal with this problem. We resampled with replacement a sample of size . Given the estimated parameters from the maximum penalised log-likelihood procedure, except for the parameters of the spline for which we wish to construct the confidence bands, we estimate the parameters of the spline with this re-sample. We repeat these steps many times to construct an empirical distribution for each parameter of the spline. We construct the lower (upper) band of the spline by taking the 5 (95) percentile of each parameter distribution. Each repetition of this procedure could take some computational time, but it converges with a relatively small number of iterations. As is usually recommended, we used 1000 repetitions to find those values. We found that the values of the 5 and 95 percentile of a 100-repetition-

bootstrap are very similar to the percentiles of a 1000-repetitions-bootstrap. Considering the calculation time required for this procedure, this is an advantage that the bootstrap stabilizes quickly.

To conclude the MVNB, note that this distribution and other panel distribution for claim counts can be used for predictive rating, where it can be shown that the predictive distribution of depends on past values of and , for $= 1$ 1. To illustrate the situation, we obtained a value of $= 6\ 57$ for an MVNB with contract duration as an offset, but without driven distance, while the final GAMLSS model with 2 splines based on the MVNB generates $= 8\ 25$. Without going into details, given that our objective for this paper is to measure the impact of distance driven, it means that a rating structure based on MVNB with telematics information reduces the unexplained variance of the model, while offering smaller penalties/discounts for drivers who claim/do not claim. We refer to (Denuit *et al.*, 2019) for predictive rating models with telematics information.

1.5 Fixed Effects

1.5.1 Model Specification

In the fixed effects model, we consider each , 1 as an unknown parameter. One approach would then be to estimate all those parameters, as well as the parameters associated with the covariates, by maximum likelihood. That means that at least $+ \ +1$ parameters should be estimated, which is quite a high number of parameters given that is usually small for insurance datasets. The problem with this ML estimation is that it does not necessarily generate convergent estimates in the classical case of a fixed and . Moreover, the large number of parameters in the model causes what is called incidental problem, which means that an incorrect estimation of the fixed effects generates incorrect estimates of associated with covariates in the mean. In the case of a logistic regression, for example, it has been shown that the were indeed biased. However, hopefully, it has been shown that a fixed effects model based on a Poisson distribution does not have this problem (see (Cameron & Trivedi, 2013) for a detailed explanation).

Consequently, for a fixed-effect Poisson regression model of mean , it can be shown that the first-order condition to obtain each is simply

$$= \text{——} \tag{1.7}$$

where for each insured was directly estimated using MLE. For the parameters, the first condition by

MLE can be shown to be equal to :

$$\sum_{=1}^{} \sum_{=1} \frac{}{} = 0 \tag{1.8}$$

When looking closely at this equation, some details about estimation for fixed effects can be deduced.

1. When we compare Equations (1.4) (first-order condition equation of the random effects model) and (1.8), we see that when is large, or when 0, random and fixed effects models are equivalent. However, in our data, the number of contracts observed for each insured is small, while is significantly greater than zero. This results in different estimation equations between the two models.
2. Individuals observed for a single insured period, i.e., with $= 1$, are not considered in the estimation of the parameters;
3. Individuals who have not filed claims with the insurer do not contribute to the estimation either. Indeed, for an individual that does not have a claim, we have $\sum_{=1} 0 \frac{0}{} = 0$ which is constant, whatever the value of .
4. It is necessary to restrict the covariates included in to those that change over time. Consequently, this also rules out the inclusion of an intercept in the model.
5. If does not change over $= 1$ for an individual , this policyholder does not contribute to the estimation (even if they claimed). The ratio is the key element in the estimation of , where it is used to find the best "weight" to apply at each to approximate . In other words, to measure the specific effect of a covariate , the driving experience of an insured must be measured with and without the effect of . For the distance driven, this seems to be exactly what we are looking for. Indeed, as mentioned in Section 1.2, we are looking for the marginal impact of each extra kilometer driven when insureds decide to use their car rather than leaving it at home.

1.5.2 Poisson Fixed Effects and Smoothing Functions

We show that fixed effects modeling with smoothing functions is possible using the GAM theory. Indeed, as mentioned (Cameron & Trivedi, 2013), each can be seen as a simple covariate identifying the insured . Consequently, by

- removing insureds without claim,
- removing insured observed for only one insured period ,
- adding a factor covariate for insured identification,

the Poisson fixed effects model can be seen as a basic Poisson regression model without an intercept. Being part of the linear exponential family of distribution, GAM theory can then be used when smoothing functions are added to the mean parameter of the distribution.

1.5.3 Numerical Illustration

In practice, as mentioned, it is relatively easy to implement the fixed effects model with R; we simply used the *gam* function from the package *mgcv*. To include fixed effects in the model the intercept of the model is dropped. We include a unique identifier variable for each policyholder as a factor variable and we include the distance driven in the model using a cubic spline . As for the GAM of Section 1.2, we used a cubic spline for the modeling. The cubic spline yields to very similar results to those for a penalised spline, but for a fraction of the computation time. Unlike the two previous approaches, we decided to illustrate the usage of the Poisson fixed effects by not including a smoothing function for the duration because our objective in this research is to measure the marginal effect of the distance on the claim frequency. If we want to measure the risk of each additional kilometer the insured decides to drive, the duration of the contract is not important. Put another way, we want to construct a rating structure based solely on the distance driven as a risk measure. Figure 1.12 shows the results for the relationship between $\exp(\ (\ \ ))$ and claim frequency.

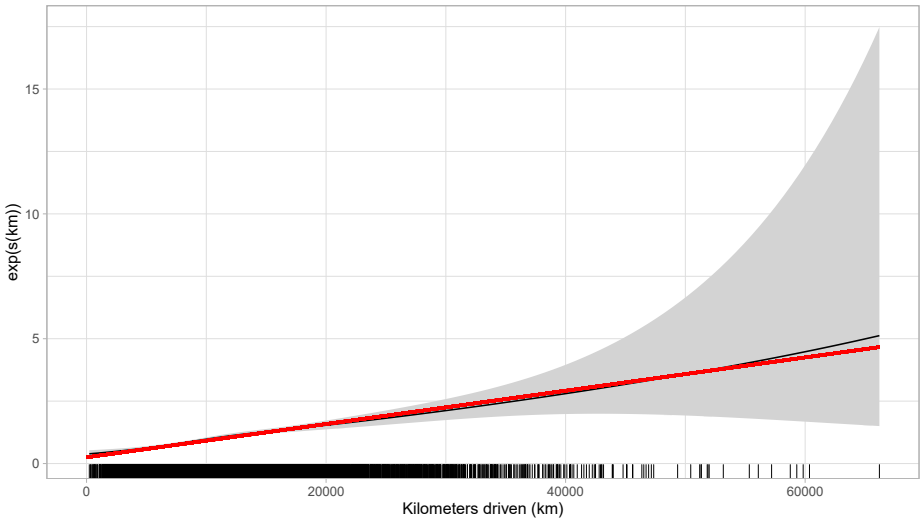

Figure 1.12 – GAM with fixed effects estimated with Canadian data

For the fixed effects model, we see that the relationship between the distance traveled on the claim frequency is always increasing, and is even almost linear. To highlight this linear effect, a line had been added to

the graph to show how close the relationship is to a linear relationship. Toward the end, there is a noticeable deviation from the linear relationship, but only 0.3% of the observations are beyond this point, which is not very significant. What has been called the "learning effect", observed in Section 1.2, has disappeared and we observe a much more logical and coherent relationship between distance traveled and frequency than before. The relationship between claim frequency and the distance driven should be understood as the marginal impact of each additional kilometer driven or not-driven. Explicitly, as we approximated $\exp(\ (\quad))$ by $0\ 25 + \frac{1}{15000}$ (the red line in Figure 1.12), we then have

$$
\begin{aligned}
&(\exp(\quad)\exp(\ (\qquad))) \\
&(\exp(\quad)(\quad + \qquad)) \\
&\quad 0\ 25\ \exp(\quad) + \frac{1}{15000}\exp(\quad)
\end{aligned}
$$

We see that the slope, i.e., the marginal impact of each additional kilometer driven or not-driven, is not the same for each insured because it depends on . To illustrate this difference, we use the estimated values of for several insureds. Figure 1.13 shows the relationship between claim frequency and distance driven for different individuals (the policyholder with the minimum, maximum, median, 25th and 75th percentile individual parameter value). With this model, we then reconcile the intuition that each kilometer should increase the risk for an individual, but that this increase could be different for each driver.

In summary, instead of referring to the "learning effect" to understand the left-hand graph of Figure 1.9, we should understand instead that typical insureds who drive more than 60,000 km per year are better risks *per kilometer* than insureds who drive approximately 40,000 km per year. That obviously does not mean that insureds that drive 40,000 km per year should drive 60,000 km to reduce their risk. The difference between insureds related to their risk *per kilometer* can be explained by many factors : more frequent use of the highway, higher proportion of driving outside rush hours, etc. However, for each driver, independently of their driving risk *per kilometer*, the risk of an accident will always **increase** for each additional kilometer driven (by approximately $\frac{1}{15\,000}$).

To conclude about the fixed effects model, note that the risk is still present even when the driving distance is zero. This is counter-intuitive because we can presume that someone who does not drive at all should have an expected claim frequency of zero. We agree. However, the real risk exposure is never completely null and the intercept could represent situations where an accident is possible even without driving a lot (e.g., it may occur very close to the insured's home). Moreover, even if the car would never actually be used,

hit and run situations are also possible.

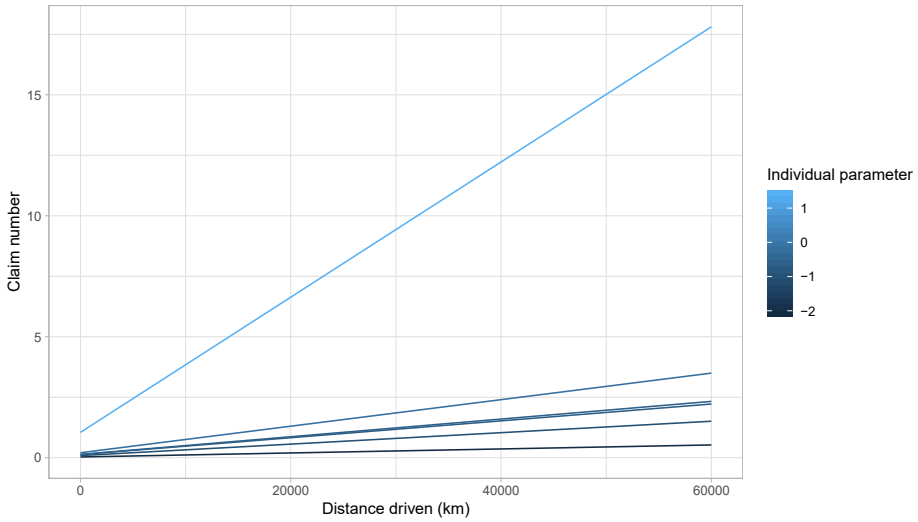

Figure 1.13 – Exposure measure for different individual parameters

1.5.4 Which Effect Should Be Used in Practice ?

Random and fixed effects seem to generate contradictory results, and we may wonder which model we should then use in practice, particularly for ratemaking. This has already been discussed in the actuarial literature by (Boucher & Denuit, 2006), but it is worth reexamining it in the context of telematics data, for distance driven in our case.

First, the fixed effects model is more general than the random effects model, which means that in case of contradictory results, fixed effects should always be preferred. Equation (1.3) can be derived as :

$$
\begin{aligned}
&\Pr[\ _1 = \ _1 \ = \ ] \\
&= \ _0 \Pr[\ _1 = \ _1 \ = \ _1 \ ] \ (\ _1 \ ) \\
&= \ _0 \ \Pr[\ = \ _1 \ ] \ (\ ) \\
&= \ _0 \ \exp(\ )\frac{(\ )}{!} \ (\ )
\end{aligned}
$$

We can see that we have to suppose an additional assumption : from the first to the second line of development, $(\ _1 \ )$ becomes $(\ )$. That means that we must suppose that random effects

are independent of observed covariates. Empirical analyses have shown that this is not the case. Indeed, as shown by (Boucher & Denuit, 2006), random effects do not have the same distribution for young drivers as for older ones, and depends on gender, for example. However, this is a typical assumption made in actuarial science, and (Boucher & Denuit, 2006) discusses the consequences of not satisfying this assumption. The authors concluded that the interpretation of random effects results are tricky.

On the other hand, fixed effects modeling, even if theoretically better, is not amenable to ratemaking :

- The model requires evaluating an individual parameter for each insured in the portfolio. This raises a problem for new policyholders.
- For a small value of , may be incorrectly estimated.
- As the model estimates each individual as —, policyholders without claims will have an expected number of claims of 0, meaning that the premium of these insureds should be zero.

As (Boucher & Denuit, 2006) conclude for basic ratemaking purposes, even if theoretically problematic, the random effects model should be preferred over a fixed effects model : random effects are flexible enough to compute premiums for new insureds, and do not generate a premium of 0 for insureds without claims. However, actuaries must understand that the parameters obtained by random effects models only indicate the apparent effect of the covariates, and not a causal effect (or what might be call the *real* impact).

To compare the fixed effects results with those of the random effects model, the approximate relationship for the median value of the individual parameter has been plotted over the smoothed function of distance traveled of the random effect approach (see Figure 1.14). Interestingly, the two curves are similar.

Regarding the use of the results of a fixed effects model, fixed effects should be used to understand the "true" relationship between covariates and claims experience. For ratemaking, fixed effects should be used to compute the premium surcharge for each additional kilometer the insureds drive. In our case, it represents an increase of $\frac{1}{15\,000}$ per km, for claim frequency. Using this approach, insurers will avoid the situation where an insured could see a premium reduction if, for example, he decides to drive 50,000 km instead of 40,000 km, as we saw with a basic GAM approach. Fixed effects can be used to construct PAYD insurance solely based on kilometers driven for self-service vehicles, where drivers' profile cannot be directly used for ratemaking. Research is required in this area.

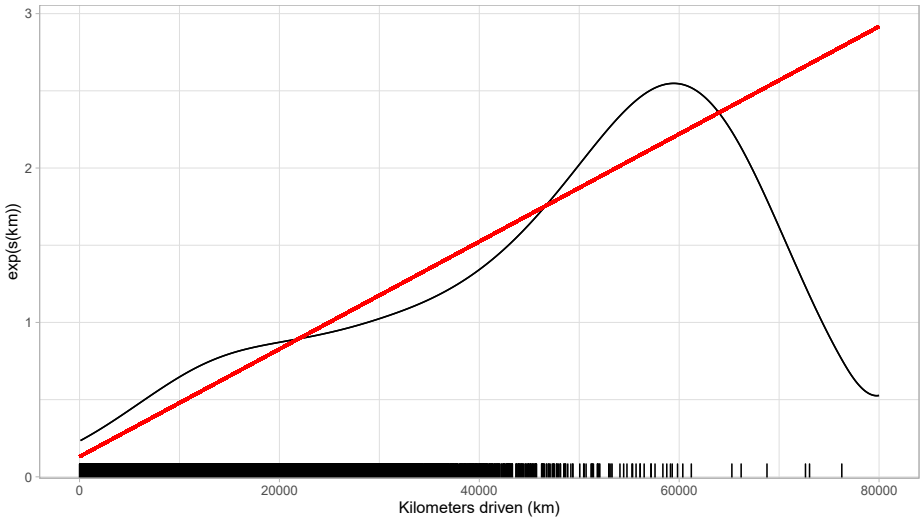

Figure 1.14 – Comparison between the random effect approach and the fixed-effect approach for the median value of the individual parameter

1.6 Conclusion

We have studied the relationship between claim frequency and the distance driven through different models by observing smooth functions. We first reproduced with our data the model proposed by (Boucher *et al.*, 2017) and observed what the authors called the "learning effect," where the expected number of claims seems to decrease as kilometers driven increase. Given that most drivers in the insurance portfolio already have many years of driving experience, we rejected the conclusion that an additional 10,000–20,000 km adds enough experience to observe a learning effect. Instead, we supposed that the residual heterogeneity was incorrectly captured by the underlying GAM model.

We then evaluated panel data models with fixed and random effects. Using GAMLSS theory, which generalises GAM, a multivariate count distribution for all the contracts of the same insured was developed. Smoothing functions were added to the mean parameter of the multivariate distribution, and a penalised log-likelihood was used to estimate the parameters. A grid of penalties, generating more than 1000 MVNB, was used to find the best distribution. However, again, the fitted smoothed function for the distance driven by the Poisson distribution with random effects did not seem to correctly describe the relationship between distance and claim frequency. Indeed, the expected number of claims still decreases disproportionately with kilometers driven.

We then used the Poisson with fixed effects to account for all individual characteristics. Because Poisson with fixed effects can be estimated by using covariates that identify each insured, we show that a simple GAM model without intercept can be used to include a smoothed function in the mean parameter. We then observed an approximately linear relationship between the distance driven and claim frequency when all individual characteristics have been accounted for in an individual parameter. This unravels the potential for the distance traveled as an exposure variable, even though this variable could not serve as a rating model. However, we think that the model proposed can be used to compute the premium surcharge for additional kilometers the insured wants to drive, or as the basis to construct PAYD insurance for self-service vehicle.

The new telematics data available in automobile insurance offers several new challenges. These data increase the possibility of identifying factors that make accidents more probable. Models like the fixed effects models proposed in the paper, make it possible to better capture the real effect of a covariate on risk. By using various models that do more than predict or calculate the insurance premium, research by insurers could shed light on risk in auto insurance. We therefore believe that many statistics compiled by telematics devices could be studied from such an angle in the future.

CHAPITRE 2

GAMLSS FOR LONGITUDINAL MULTIVARIATE CLAIM COUNT MODELS

Insurance policies cover multiple drivers and vehicles over several annual contracts. In predictive modeling and in ratemaking, this feature is not necessarily fully exploited to improve the model's predictive power. Consequently, this work centres around longitudinal driver-based claim frequency models because we believe the impact of using historical data will be most compelling on this pure premium component. Traditionally, pricing models break down the pure premium into two components : claim frequency and severity of insurable loss. The classic approach, widely accepted in the literature and in practice, is to assume independence between these two components and to model them separately, even if some recent papers have developed approaches to frequency-severity dependence (see (Delong *et al.*, 2021) or (Shi & Lee, 2022)). In point of fact, for the same reckless behaviour, the monetary consequences of a collision can change dramatically depending on whether the impact was with another vehicle, a truck or a cyclist, making it difficult to use severity in a time-dependent framework. These external considerations are unpredictable. However this at-fault accident is a strong indicator of future claim occurrence, as has been extensively studied in actuarial literature, in particular in the context of Bonus-Malus systems (BMS) (see, e.g., (Tremblay, 1992), (Denuit *et al.*, 2007) and (Lemaire, 2012)). Extending BMS by incorporating panel data has been proposed ((Boucher & Inoussa, 2014), (Verschuren, 2021)) as storage capacity had increased.

Time dependence has been extensively studied. (Frees & Wang, 2006) have notably included time dependence by using latent variables in the count regression and elliptical copulas in the severity model. In (Frees & Valdez, 2008), a three-component model for frequency, types of coverage, and severity of claims is introduced in a generalised linear model framework. This idea is taken up in (Yang & Shi, 2019) where this three-component model is incorporated into a longitudinal numerical application. They use loss data involving the building and content of government entities where censored subjects are rare over the six-year study period. They observed that the temporal correlation did not decrease significantly over the observation period. (Yan & Jeong, 2022) proposes an expectation-maximisation algorithm for longitudinal compound risk models. EM algorithm has also been used in (Tzougas, 2020) for the estimation of a Poisson-inverse gamma regression model with varying dispersion. The discrete nature of frequency data can become inconvenient as it limits the use of copulas. In (Shi & Valdez, 2014), the focus is only on claim frequency in a longitudinal context for a single insurance coverage. They use a "jitters" method to circumvent the limitations of copulas

and they use elliptical copulas to model subject dependence over time. This method allows the data to be continuous, while preserving the concordance-based measures of association as demonstrated in (Denuit & Lambert, 2005). (Pechon *et al.*, 2018) also focus on claim frequency analysis, but they consider the multiples individuals in the household, instead of looking at claim frequency at the policy level. They used a multivariate Poisson distribution with different random continuous effects modeled with Gaussian and Gamma distributions. The chosen mixtures are not rendering closed-form likelihoods, and the multiple integrations can make it difficult to implement the model in a practical context. On the other hand, they conclude that knowledge of the household's claims history can improve the individual prediction for each member, which can come in handy from a practical point of view. On the use of random effects for modeling the inherent heterogeneity of longitudinal P&C insurance data, (Jeong, 2020) offers a theoretical discussion of their use based on Bregman divergence. Random effects, telematics data or non-parametric terms are all different manners to account for residual heterogeneity. (Schelldorfer & Wuthrich, 2019) uses Combined Actuarial Neural Network (CANN) in the context of classical generalised linear model for cross-sectional frequency modeling. CANN were introduced in (Wüthrich & Merz, 2019) and a more detailed explanation can be found in (Wüthrich & Merz, 2023).

Relatedly, in this paper, we use the multivariate negative binomial distribution (MVNB) and the multivariate beta negative binomial distribution (Beta-NB) in a longitudinal framework that allows an *a posteriori* interpretation of longitudinal models for pricing as discussed in (Boucher *et al.*, 2008). These mixture models have a closed-form likelihood and long historical data does not complicate the implementation of the model, unlike the previously discussed mixture models. We follow a driver through time, not a policy or a car, as is usually the case, and we compare the results with models that lack a longitudinal component as a benchmark for predictive power improvement. In addition, we propose modeling the location parameter of each distribution using covariates and smoothing functions in a generalised additive model for location, scale and shape (GAMLSS) framework. The inclusion of nonlinear terms inside models based on distributions of the linear exponential family follows from the extension of generalised linear models ((Nelder & Wedderburn, 1972)) proposed by (Hastie & Tibshirani, 1986). For an exhaustive reference study on the subject of generalised additive models (GAM), see (Wood, 2017). Their use in actuarial science dates back to the early 2000s with research on spatial effects such as that done by (Fahrmeir *et al.*, 2003), (Denuit & Lang, 2004) and, more recently, (Henckaerts *et al.*, 2018). The use of GAMs to process telematic data has been widely studied with different objectives like classification ((Verbelen *et al.*, 2018), (Huang & Meng, 2019)), the study of risk exposure ((Boucher *et al.*, 2017), (Boucher & Turcotte, 2020)) or the risk factors of

near-miss events ((Guillen *et al.*, 2020)). The inclusion of linear regressors and nonlinear functions is limited to the location parameter for distributions of the linear exponential family in GAMs. A more general framework, generalised additive models for location, scale and shape (GAMLSS), has been proposed by (Rigby & Stasinopoulos, 2005), where any parametric distribution can be used. More details on this theory are given in (Stasinopoulos *et al.*, 2017). For insurance modeling, this step forward is worthwhile because it allows for more flexible modeling of more suitable distributions, such as zero-inflated and heavy-tailed distributions that are often not members of the linear exponential family (see (Heller *et al.*, 2006)), or the flexible modeling of scaling and shape parameters, as illustrated in (Heller *et al.*, 2007) and (Tzougas & Jeong, 2021). Random effects models are very common in actuarial science and have been studied a little in the context of GAMLSS ((Gilchrist *et al.*, 2009), or (Klein *et al.*, 2014)). However, to our knowledge, this had never been studied in a longitudinal setting before (Boucher & Turcotte, 2020), which is limited to the study of exposure measures. In that work, a pricing model had not been developed and, therefore, the contribution of the longitudinal and semi-parametric elements had not been evaluated. In (Tzougas & Frangos, 2014), they attempt a longitudinal model with the GAMLSS framework, but they use the assumption that the *a priori* explanatory variables remain constant over time, probably to circumvent the limitations associated with the use of the GAMLSS library in R. This assumption is not verified in practice. We did not use the GAMLSS library and our model does not assume any simplifying assumptions regarding the value of the explanatory variables. Our work clearly explains the process for implementing a parametric or non-parametric longitudinal model using the GAMLSS framework. In addition to implementing the MVNB approach for panel data, we show that the elements contained in this work make it possible to implement other multivariate distributions. For illustration, we use the Beta-NB, but the multivariate Sichel or Poisson Inverse Gaussian distribution could easily be used.

The current theory is not explicit enough to be able to use a GAMLSS with panel data. Panel data is important for modeling the dependence between the number of claims from the contracts of the same insured over time. It is, therefore, necessary to develop this theory to not be limited by simplifying assumptions. The objectives of this project are two-fold. First, we support the use of semi-parametric longitudinal models to improve predictive power and reduce the importance of past experience in the predictive rating. The inclusion of past experience is intended to introduce into pricing the unobservable elements of *a priori* pricing, such as impulsivity or less conservative behaviour. Semi-parametric models, by giving less importance to past experience, show that they are better suited to capturing *a priori* information. Secondarily, we discuss the use of GAMLSS in actuarial science. Applications have been demonstrated in the past using pre-coded

functions in an R package ((Heller *et al.*, 2006), (Heller *et al.*, 2007) and (Tzougas & Frangos, 2014)), but how to use this theory with any useful distribution for actuarial work, such as the MVNB and the Beta-NB, has not been explained. Clarifying the theory makes it possible to generalise the approach for all kinds of distributions such as the Gaussian inverse Poisson or the Sichel distribution.

Parametric distributions for claim counts are reviewed in Section 2.1, with the Poisson and negative binomial distributions for cross-section data, and the MVNB and Beta-NB distributions for longitudinal frameworks are introduced for the purpose of comparison and to establish notation. Section 2.2 explains the key elements of splines, penalised regression and GAMLSS in order to include non-parametric terms in the location parameter modeling. The database used for the numerical application is presented in Section 2.3.1 and Section 2.3 discusses the improvement found by moving from parametric models to semi-parametric longitudinal models and the advantages of our proposed approach. Section 2.4 concludes the paper.

2.1 Parametric modeling

We consider an insurance portfolio of policyholders observed over years. For each contract , $=1$, we define , a discrete random variable counting the number of claims for the policy period and a column-vector containing available explanatory factors at the beginning of period . In this vector, we may include , a scalar that measures the risk exposure. We suppose that there is independence between all insureds of the portfolio. The primary purpose of the ratemaking model is then to provide a prediction for :

$$\mathrm{E} \quad +1 \quad (1: \ ) \quad (1: \ +1)$$

where :

- $(1: \ )$ contains all past number of claims between time 1 and time for insured ;
- $(1: \ +1)$ contains all covariates used in the ratemaking, from contract 1 to $+1$. This usually corresponds to information about the age of the insured, the marital status of the insured, etc.

In this section, we look into parametric models for both :

- cross-section data models, for which independence is assumed between all annual contracts of the same policyholder (subsection 2.1.1) ;
- panel data models, for which we suppose dependence between all contracts written for the same driver (subsection 2.1.2).

Going through those models will be an opportunity to establish the notation. Additionally, all models pre-

sented in this section will be adjusted in their parametric and semi-parametric form for comparison in Section 2.3.

2.1.1 Cross-section Data Models

For cross-section data models, we suppose the independence between all drivers as well as between all contracts , for $= 1$ of the same insured . The probability function is then :

$$\Pr \quad = \quad _{(1: \quad 1)} \quad _{(1: \ )} = \Pr(\quad = \quad)$$

2.1.1.1 Poisson

To model the number of claims of insured , for contract , the Poisson distribution of mean is usually the starting point. It has a probability mass function defined as :

$$\Pr(\quad = \quad) = \frac{\exp(\quad)}{!}, \text{ with } \quad = \exp(\quad) = \quad ^{1}(\quad) \tag{2.1}$$

The mean parameter is expressed using a log link function and a linear combination of a vector of covariates and a vector of parameters . The Poisson distribution is part of the linear exponential family of distributions and it is therefore straightforward to estimate this model with GLM theory.

2.1.1.2 Negative binomial

The negative binomial distribution is very well suited for counting data with overdispersion. The probability mass function is :

$$\Pr(\quad = \quad) = \frac{(\quad + \quad)}{(\quad) \quad !} \quad (1 \quad) \quad , \text{ with } \quad = \exp(\quad) = \quad ^{1}(\quad) \tag{2.2}$$

The mean is $(\quad (1 \quad))$. The parameter to which the mean is proportional is expressed using a log link function and a linear combination of a vector of covariates and a vector of parameters . In general, the negative binomial is not part of the linear exponential family of distributions and GLM theory cannot be used to perform the estimation. We will explain in Section 2.2.3 that GAMLSS theory can be used instead.

2.1.2 Panel Data Models

Panel data models assume all annual contracts belonging to the same driver to be dependent for $= 1$. In a parametric approach, the joint distribution of the random vector $_{1}$ is needed.

The following presents the theory for a single insured. To simplify the notation, the subscript is dropped and becomes . Plenty of models allow for time dependence between random variables, e.g., conditional models, marginal models, and subject-specific models, but it has been shown that random effects models were the best suited for claim counts (see (Boucher *et al.*, 2008)). Indeed, predictive scoring for panel data models can be developed by introducing a heterogeneity factor. If denotes this heterogeneity factor, the joint distribution can be expressed as :

$$\Pr[\quad_1 = \quad_1 \qquad = \qquad] = \int_0 \prod_{=1} \Pr[\qquad = \qquad_1 \qquad\qquad] \quad (\quad) \tag{2.3}$$

By selecting a conjugate prior to the counting distribution, the likelihood is explicit and easier to work with in practical situations. The heterogeneity factor introduces the longitudinal dependency between all contracts , for $= 1$, of the same insured .

2.1.2.1 Experience Rating

The key element that follows from the development of the joint distribution is the ability to derive the predictive expectation of future claims at $+ 1$, conditional on past experience of contracts from time $= 1$. In a random effects model, as defined by Equation (2.3), the predictive expectation is

$$[\quad_{+1} = \quad_{+1} \quad_{(1:\ )}] = [\ [\quad_{+1} = \quad_{+1} \quad_{(1:\ )}\ ]] \tag{2.4}$$

This inclusion of past claims in the ratemaking is known as experience rating. Many approaches have been considered in order to achieve this goal, beginning with the individual credibility models of (Bühlmann, 1967) or (Albrecht, 1985), for example.

There are several justifications for experience-based pricing. Some policyholders exhibit riskier behaviour than others or live in more disaster-prone regions. The individual characteristics of each insured may partly explain this situation and some of them are often used as segmentation variables in regression models. However, many of these character-defining elements simply cannot be measured and used in pricing. For example, a negligent or reckless insured is more likely to suffer losses than a conscientious and attentive insured. Thus, past claims experience can be used to approximate the effect of these unmeasurable characteristics on the premium.

Insurers also justify experience-rating models because typical policyholders will often be reluctant to file a claim with their insurer because they are unfamiliar with the procedure. Consequently, policyholders who

have filed a claim in the past tend to be more willing to do so in the future, and to report accidents that they would not have before. Some might also become a bit less careful, knowing that the insurer will compensate them without issue and that the consequences of filing a claim are not so serious (moral hazard). From the insurer's point of view, it is therefore important to implement a rating structure that increases the premium for past claims, and rewards insureds with no claims. This structure assures the insurer that its insureds remain cautious and do not file claims for minor accidents.

2.1.2.2 Some Count Distributions for Longitudinal Data

2.1.2.2.1 Multivariate negative binomial

If we suppose Poisson( ), with a heterogenity factor that follows a gamma distribution of mean 1 and variance $\frac{1}{}$, the expected value of [ ] is unchanged and the joint distribution can be expressed as :

$$\Pr[_1 = _1 \quad = \quad] = \quad _{=1} \frac{}{!} \quad \frac{(\quad + \ )}{(\ )} \quad \frac{}{+} \quad (\quad + \ ) \tag{2.5}$$

where $= \ _{=1}$ and $= \ _{=1}$. The parameter could be modeled with regressors

$$= \exp(\quad) = \quad ^1(\quad) \tag{2.6}$$

This well-known distribution is the multivariate negative binomial distribution, or simply MVNB. This distribution is a generalisation of the negative binomial distribution. It is a basic distribution for panel count data modeling with overdispersion ([ ] = [ ] = $+(\quad)^2$). This distribution is interesting in the context of predictive ratemaking because the predictive distribution is a negative binomial distribution, as it can be shown that :

$$\Pr[\quad _{+1} = \quad _{+1} \quad _{(1: \ )}] = \frac{(\quad + \quad _{+1} + \ )}{(\quad + \ ) \quad _{+1}!} \quad \frac{ _{+1}}{+ \quad _{+1} +} \quad \quad \frac{+}{+ \quad _{+1} +}^{+}$$

It can also be shown that the predictive expected value can be expressed as :

$$[\quad _{+1} = \quad _{+1} \quad _{(1: \ )}] = \quad _{+1} \quad \frac{+}{+} \tag{2.7}$$

where $\frac{+}{+}$ acts as a correction factor between what actually happened and what the model predicted for the past contracts $= 1$. Because of , it is not possible to simply view the longitudinal

approach as a product of univariate distribution. In (Tzougas & Frangos, 2014), they got around this difficulty by assuming that was constant for all and thus, = . However, we do work with = =1 and the explanatory variables can evolve over time.

2.1.2.2.2 Beta negative binomial

Similarly, if the conditional distribution is a negative binomial distribution (the probability function expressed by equation (2.2)), with a heterogenity factor that follows a beta distribution such as ( ) = $\frac{(+)}{( ) ( )}$ $^{1}(1)$ ^{1}, the joint distribution can be expressed as

$$\Pr[\ _1 = \ _1 \ = \ ] = \ _{=1} \frac{(\ + \ )}{(\ ) \ !} \ \frac{(\ + \ )}{(\ ) \ (\ )} \frac{(\ + \ 1) \ (\ + \ 1)}{(\ + \ + \ + \ 2)}, \tag{2.8}$$

where = =1 and = =1 . The parameter could be modeled with regressors

$$= \exp(\ ) = \ ^{1}(\ ) \tag{2.9}$$

We can show that the predictive distribution can be expressed as :

$$\Pr[\ _{+1} = \ _{+1} \ _{(1: \ )}] = \frac{(\ _{+1} + \ _{+1})}{(\ _{+1}) \ _{+1}!} \frac{\mathrm{B}(\ _{+1} + \ + \ 1 \ _{+1} + \ + \ 1)}{\mathrm{B}(\ + \ 1 \ + \ 1)}$$

where $\mathrm{B}(\ )$ is the Beta function, with $\mathrm{B}(\ ) = \ _{0}^{1} \ ^{1}(1 \ ) \ ^{1}$. The predicted expected value is then given by :

$$[\ _{+1} = \ _{+1} \ _{(1: \ )}] = \frac{+ \ 1}{+ \ 2} \ _{+1} \tag{2.10}$$

In comparison with the *a priori* expected value, $\frac{+ \ 1}{+ \ 2}$ acts as a correction factor between what actually happened and what the model predicted for past contracts, up to a multiplicative factor $\frac{1}{}$.

2.2 Semi-parametric modeling

Generalised linear models have become the standard for pricing in the insurance industry ((Anderson *et al.*, 2007)). GAMs are an extension of the GLMs that were proposed by (Hastie & Tibshirani, 1986). GAMs introduce more flexibility in modeling, while keeping a somewhat similar framework to GLMs, which could be an advantage for a conservative industry like insurance. The main difference between a GLM and a GAM is the ability to include smoothing functions in the linear predictor of the location parameter. Although the

linear form of the predictor is kept, including smoothing functions implies that a log-linear relationship is no longer forced between and or (see Equations (2.1), (2.2), (2.6) and (2.9)). Of course, the implications of this change have several ramifications for the estimation or evaluation of the model, which will be covered in Section 2.2.2. The interpretation of the model is nonetheless very intuitive if one is familiar with GLM.

If GAMs allow for a less restrictive relationship between the location parameter and the covariates, GAMLSSs allow for a less restrictive relationship between the location parameter (or shape parameter or scale parameter) and the covariates, in addition to allowing the use of a wider range of distributions, including distributions that are more suited to the challenges of insurance data. In this work, we use the distributions presented in the previous section. In contrast, (Tzougas *et al.*, 2015) estimated several frequency and severity models for *a priori* ratemaking using distributions such as the Sichel distribution, the zero-inflated Poisson distribution, and the generalised gamma distribution that are not part of the linear exponential family of distributions. They then compared the models by analysing the differences in mean and variance parameters between the risk classes. Another useful application of GAMLSS for general insurance is spatial data analysis techniques, which are used with car theft data or to assess climate risk to houses, for example. (De Bastiani *et al.*, 2018) used GAMLSS to model spatial components in a Gaussian Markov random field model and they illustrated their proposed approach with Munich rent data. (Ramires *et al.*, 2021) proposed a clustering approach based on GAMLSS to consider latent variables in the modeling in order to minimise or correct anomalies such as unexplained bimodal distributions. They illustrated their approach with a practical example based on public health-insurance data. Approaches including GAMLSS in reserve models have also been proposed. For instance, the work of (Spedicato *et al.*, 2014) applies GAMLSS to development triangles to evaluate the distribution of the unpaid loss reserve in terms of best estimates and the shape of the distribution. The results are analysed in comparison with those of classical approaches, such as methods based on the Chain-Ladder technique.

As the variety of applications mentioned above shows, it would be interesting to examine the possibilities of this framework with time dependence and panel data.

2.2.1 Panel Data Models

This type of analysis using panel data would be practical in other applications as well. For example, the analysis of the impact of certain regressors on the frequency of claims should be carried out using an experience-

rating approach. Indeed, as mentioned by (Lemaire, 1998), several empirical studies have shown that the best predictor of the number of claims incurred by a driver in the future is past claims history. In other words, this means that the effect of certain regressors could be different when the model is also based on the number of past claims. In this sense, (Boucher & Inoussa, 2014) observed that the effect of the insured's age on the number of claims was different when an experience rating structure was implemented in automobile insurance. (Tzougas & Frangos, 2014) used GAMLSS with the Poisson-Gamma distribution, the Poisson Inverse Gaussian distribution, and the Sichel distribution to model the claim frequency distribution within a BMS with the objective of designing an optimal BMS for predictive ratemaking. Another notable example could be the use of GAMLSS by (Li & Tan, 2015) which used this framework to include covariates such as climate indices into the modeling of parameters in a nonstationarity time series to analyse the flood risk frequency. Considering that cross-sectional models have been successfully applied in actuarial work to efficiently measure the impact of continuous covariates, it is necessary to generalise the approach to longitudinal models. On this related subject, (Boucher & Turcotte, 2020) showed that the impact of mileage driven was not the same in a longitudinal model as in a model with cross-sectional data. In this work, we mainly use the flexibility of GAMLSS to model longitudinal data.

Thus, in this section, main points about including smoothing functions in the linear predictor of GAMs and GAMLSSs are covered. In order to correctly use distributions for longitudinal data, knowledge of smoothing functions, penalised regression and models generalising the theory of GAMs is required. These theories are often presented in detail in academic books, but we synthesise the important elements here. An introduction is given to help understand the key elements that are needed in order to fit a longitudinal model including regressors and smoothing functions in the modeling of the parameters. Many elements are taken from (Wood, 2017) and (Stasinopoulos *et al.*, 2017). Finally, note that several smoothing functions can be included in additive models but we focus our explanations on two one-dimensional smoothing functions, namely cubic regression splines and P-splines.

2.2.2 Theoretical Development of GAM

The objective is to include non-parametric terms in the linear predictor of parameters. First, we will explain what those so-called non-parametric terms that were considered in this work are and then we will discuss the difference between a GLM and a GAM to lead up to the introduction of GAMLSS in the next subsection.

2.2.2.1 Splines

Splines are smoothing functions defined piecewise by polynomials ; they could be expressed under the form of Equation (2.11). By definition, the smoothing functions used in the context of GAMs can be rewritten as a linear combination of basis functions (( )) and parameters ( ),

$$( ) = \sum_{=1} ( ) \tag{2.11}$$

This is the key element that allows us to preserve the linear form of the predictor, but to no longer require a log-linear relationship between and or . First, we illustrate what basis functions are using the cubic regression splines.

2.2.2.1.1 Cubic regression splines.

Cubic regression splines are common smoothing functions used with additive models. The cubic spline is a smoothing function built using pieces of a third-degree polynomial that are joined so that the function is continuous to the second derivative. One could have splines of different degrees. In particular, the cubic spline is used because natural cubic splines are the smoothest interpolators in the sense that if we define a spline using 1 polynomial pieces to fit a sample of data, the natural cubic spline minimises $( )^2$, where $_1$ and are the extremums of the data sample. The function $( )^2$ measures the oscillations of the function and the square prevents the positive values from cancelling out the negative values. The term "natural" cubic splines simply implies an additional hypothesis to define the function, which is $(_1) = ( ) = 0$.

In a smoothing context, rather than interpolation, we will choose knots, defining 1 intervals over the span of the data, where . Those knots are noted as for $= 1$. It is not required that the knots coincide with the location of a datum. However, the extreme values of the knots must cover the span of the data. One way to do this is to choose the knots at regular intervals, or quantiles, over all the data.

The cubic regression spline is defined using the following assumptions : the second derivative of the spline varies linearly in each interval (continuous curvature in each interval) and the first derivative of the spline is continuous on the knots (continuous curve at the joints). Following this, it can be shown that the cubic

spline $(\ )$, defined at point , is expressed as :

$$(\ ) = \quad (\ ) \quad + \ ^{+}(\ )\ _{+1} + \quad (\ ) \qquad + \ ^{+}(\ )\ _{+1} \ = \ \sum_{=1} \ (\ ) \tag{2.12}$$

where

$$(\ ) = \begin{array}{ll} (\ ) \quad + \ ^{+}(\ )\ _{+1} \ + \ ^{+}(\ ) & \text{if} \ = \ +1 \\ (\ ) \quad + \ ^{+}(\ )\ _{+1} \ + \quad (\ ) & \text{if} \ = \\ (\ ) \quad + \ ^{+}(\ )\ _{+1} & \text{otherwise.} \end{array} \tag{2.13}$$

The vector of parameters is to be estimated. The terms $(\ )$, $^{+}(\ )$, $(\ )$, $^{+}(\ )$ defining the basis functions are defined as :

$$(\ ) = \frac{\ +1}{\ +1} \qquad (\ ) = \frac{1}{6} \ \frac{(\ \ +1 \quad)^3}{\ +1} \quad (\ \ +1 \quad)(\ \ +1 \quad)$$

$$^{+}(\ ) = \frac{}{\ +1} \qquad ^{+}(\ ) = \frac{1}{6} \ \frac{(\ \quad)^3}{\ +1} \quad (\ \ +1 \quad)(\ \quad)$$

We also define , the $^{\text{th}}$ row of the matrix $= \begin{matrix} 0 \\ \\ 0 \end{matrix}$ where $= \ ^{1}$.

Finally, non-zero elements of matrices $(\ \ 2)\ (\ )$ and $(\ \ 2)\ (\ \ 2)$ are defined in Table 2.1.

$1 \qquad 2$	$1 \qquad 3$
$= \frac{\ +2}{3}$	$_{+1} \ = \quad _{+1} = \frac{\ +2 \quad \ +1}{6}$
$= (\ \ _{+1} \quad)\ ^{1}$	
$_{+1} = \quad (\ \ _{+1} \quad)\ ^{1} \quad (\ \ _{+2} \quad _{+1})\ ^{1}$	
$_{+2} = (\ \ _{+2} \quad _{+1})\ ^{1}$	

Table 2.1 – Non-zero elements of matrices and

2.2.2.1.2 B-splines.

Another way to express splines is to use B-splines basis, which are strictly local, unlike the basis functions given by Equation (2.13). This means that each basis function is non-zero only over the intervals between $+3$ adjoining knots, with $+1$ being the order of the basis. For example, $=2$ is for a cubic spline. B-splines with $=2$ are a different way of expressing a cubic spline, but with different basis functions that are on a local support. The initial purpose of B-splines was to offer a very stable basis for particular problems. However, for most uses, only poor statistical methods would start to show a noticeable stability enhancement. The value of B-splines basis comes from the method proposed by (Eilers & Marx, 1996), who have developed what has been known as P-splines, which will be discussed in more detail in the next subsection. In a nutshell, P-splines offer an interesting way of controlling wiggliness and they use B-splines basis. The difference between B-splines and P-splines is the inclusion of a penalty, which is what the "P" in P-splines stands for. Among other uses, P-splines provide an effective way to include constraints on the shape of the spline, such as monotonicity.

Similarly to smoothing functions used in generalised additive models, B-splines can be expressed in a linear form using basis functions $(\ )$:

$$(\ ) = \sum_{=1} (\ ) \tag{2.14}$$

where is the number of parameters defining the spline. However, it is necessary to define $+\ \ +2$ knots so that the data spanned between knots $_{+2}$ and $_{+1}$. In this manner, splines at the ends are well-defined over the intervals between $+3$ adjoining knots. Evenly spaced knots are usually chosen. There could be more than one way to express the B-spline basis function . The recursive definition is convenient and widespread in the literature (see (De Boor, 1978)). For 1 , $(\ )$ is defined as

$$(\ ) = \frac{}{+\ \ +1}\ \ {}^{1}(\ ) + \frac{+\ \ +2}{+\ \ +2\ \ \ +1}\ \ {}_{+1}{}^{1}(\ ) \quad \text{where} \tag{2.15}$$

$${}^{1}(\ ) = \begin{cases} 1 & \text{if} \quad {}_{+1} \\ 0 & \text{otherwise.} \end{cases}$$

2.2.2.1.3 Vector of covariates.

It is possible to use classical regression methods to assess the model that replaces continuous covariate with a spline $( )$ by replacing, inside the linear predictor, the vector $= (\ 1 \ \ 2 \ \ )$ with the matrix formed by the basis functions

$$= \begin{matrix} {}_1(\ 1) & {}_2(\ 1) & & (\ 1) \\ {}_1(\ 2) & {}_2(\ 2) & & (\ 2) \\ \vdots & \vdots & \ddots & \vdots \\ {}_1(\ \ ) & {}_2(\ \ ) & & (\ \ ) \end{matrix}$$

An identifiability problem might arise because an intercept ${}_0$ is nearly always included in the predictor. Consequently, could not be included directly like it is in the design matrix of a regression model that includes multiple covariates and an intercept. Let ${}_{(\ )} \ {}_{(\ )}$ be the dimension-reduced matrix, where $=$ $1.$ could be obtained by subtracting the column mean of from each column of the matrix . Put differently, we have

$$= \tag{2.16}$$

This equation follows from the constraint ${}_{=1} \ (\ \ ) = 0$, which changes neither the shape of the splines nor the value of the penalty hyperparameter (see Section 2.2.2.3).

It is worth pointing out that the basis functions (2.13) and (2.15) depend only on the locations of the knots and the continuous variable . This means that does not depend on the model and it could be included as-is in any predictor as if they were a set of covariates. However, one must be careful to differentiate between parametric (regular covariates) and non-parametric terms (smoothing functions) because the interpretability of some model statistics is lost for non-parametric terms.

If we were to link the aforementioned to the notation introduced in Section 2.1, let us suppose a count distribution where the mean parameter is computed with covariates , which could either be categorical or continuous. We include an intercept, categorical covariates and, for example, three continuous covariates. In a GAM, one or many continuous covariates could be replaced by a smoothing function, meaning that

can be represented as :

$$= 1 \underbrace{1}_{\text{categorical covariates}} \underbrace{\quad}_{} \underbrace{+1}_{\text{continuous covariates}} \underbrace{+2}_{} \underbrace{(\quad +3 \quad)}_{\text{smoothing function}} \tag{2.17}$$

Equation (2.17) can then be rewritten as :

$$= 1 \underbrace{1}_{\text{categorical covariates}} \underbrace{\quad}_{} \underbrace{+1}_{\text{continuous covariates}} \underbrace{+2}_{} \underbrace{1(\quad +3 \quad)}_{\text{smoothing function}} \underbrace{(\quad +3 \quad)}_{} \tag{2.18}$$

2.2.2.2 Controlling Wiggliness

It has been stated that the spline is defined by a certain number of knots, so we need to select their quantity. If their number is insufficient, the spline would be too smooth and if there are too many knots, the spline would be overfitted. The number of knots is a tuning parameter. To avoid selecting multiple knots and adjusting multiple regressions in order to find the right level of smoothness, it is common practice to use a penalised regression. One common way to proceed is to choose a number of knots slightly greater than what is considered adequate and the penalty prevents over-adjustment.

In a penalised regression, instead of minimising only the least squares like a classic regression (or maximising the likelihood), an additional penalty term is added for each smoothing function included in the linear predictor. For a linear regression with parametric covariates, the following equation is an example of a penalised linear regression model with two splines of $_1$ and $_2$ knots :

$$\overline{\quad} \quad \overline{\quad} \quad ^2 + \; _1 \quad _1 \quad + \; _2 \quad _2 \tag{2.19}$$

with

$$= (\; \underbrace{_0}_{} \quad \underbrace{_1}_{} \quad \underbrace{+1}_{} \quad \underbrace{+}_{} \quad \underbrace{+ \quad +1}_{} \quad \underbrace{+ \quad +}_{} \;) \tag{2.20}$$

where the coefficients associated with parametric terms are denoted ( ), the ones associated with spline 1 ( ) and the coefficients associated with spline 2 ( ) and so on if there are other smoothing functions. For a classic linear regression (Normal distribution), the matrix of weights $\overline{\quad}$ is the identity

matrix. An alternative representation of Equation (2.19) includes a scaling factor ( ) :

$$\overline{\quad} \qquad \overline{\quad} \qquad ^2 + (\ _1 \quad _1) \quad (\ _1 \quad _1) \quad + (\ _2 \quad _2) \quad (\ _2 \quad _2) \tag{2.21}$$

This is useful to point out because some statistical software uses this representation (like the "mgcv" package in **R**). The matrices , $1\ 2$ are penalties whose composition are different for each smoothing function. The penalties associated with cubic regression splines and P-splines are discussed at the end of this subsection. , $1\ 2$ are hyperparameters that control the wiggliness. The higher is, the more it penalises the adjusted spline and prevents excessive wiggliness.

Lastly, an important point is that, in a regression where the dimension of had been reduced for the sake of identifiability, the penalty matrix must also be adapted accordingly (see Equation 2.19). One way to proceed would be to find the orthogonal matrix of vector $_1$ using the QR decomposition and drop the first column. Let us denote this matrix . Then, $=$ and $=$. This step is obviously also necessary for any new data for which we want to make predictions : $=$.

2.2.2.2.1 Cubic regression splines.

The penalty of the cubic regression spline $(\ )$ is $(\ )^2 \quad = \quad ^1$, meaning that ^{1} . The intuition behind this penalty is that the second derivative describes the concavity of the function. If the function is very wiggly, there will be many inflection points. The square is used so that concavity and convexity do not cancel each other out.

Interestingly, any function that could be written as a linear combination of basis functions (i.e. $=$ $(\ )$), could be written as $(\ )^2 \quad =$ if the basis functions are twice differentiable. The penalty matrix can then be expressed solely as a function of the known basis functions. This comes in handy when defining penalties for various splines.

2.2.2.2.2 P-splines

As mentioned, P-splines use B-splines basis, but P-splines are used in the context of a penalised regression. A difference penalty is applied to the parameters to control the intensity of the smoothing. For a penalty order $= 2$ (not related to the order of the B-spline), the penalty is :

$$\mathcal{P} = \sum_{=1} (\ _{+1} \quad)^2 \tag{2.22}$$

or

$$\mathcal{P} = \tag{2.23}$$

where

$$= \begin{matrix} 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{matrix} \tag{2.24}$$

and thus . Any penalty order can be chosen. In comparison with cubic regression splines, this discrete penalty is harder to interpret in terms of the properties of the fitted smoothing. It penalises for difference in value from two consecutive parameters . If all the increments in were the same value, it would result in a linear function. If all were the same, it would result in a flat line.

2.2.2.3 Estimation Steps

We will revisit the estimation steps of a GLM before jumping into the estimation steps of a GAM. These two algorithms are closely related.

2.2.2.3.1 GLM.

For GLMs and GAMs, only distributions that are included in the linear exponential family could be used. Hence the following equation for the log-likelihood :

$$( ) = \qquad ( ) \quad ( ) + (\quad)$$
$$=1$$

where represents the observations of response variable , , and are functions, is a scale parameter and is the canonical parameter, which depends on the selected distribution. For the Poisson distribution $= \ , \ = \exp( ), \ = \ \log(\ !), \ = 1$ and $= \log(\ )$. In Equation (2.1), the parameter corresponds to and the definition of and are unchanged from Section 2.1.1. Lastly, we recall that the variance function is $( ) = \quad ( )$ and we define $( ) = 1 + (\quad) \quad ( ) \quad ( ) + \quad ( ) \quad ( )$. Taking $( )$ as 1 corresponds to Fisher weights and this definition of $(\ )$ has been used in this work.

The vector is generally estimated using the iteratively re-weighted least square algorithm (IRLS). It is first

required to initialise $= +$ and $= ( )$. is usually zero (it is meant to ensure that is finite). The next two steps are iterated to convergence.

1. Compute the vector $_1$ of the pseudodata $_1 = ( )( ) ( ) +$ and the nonzero elements of the diagonal of : $= ( ) ( )^2 ( )$,
2. The minimiser of the weighted least squares objective $_{=1} ( )^2$ is $^{[+1]} = ( ) ^1$.

The convergence of the deviance is often used as the stopping criterion of the algorithm.

2.2.2.3.2 GAM.

For known values of hyperparameter , the estimation of parameter vector by penalised iteratively re-weighted least square algorithm (PIRLS) is quite similar to the estimation of a GLM. Only a few adjustments are necessary to take into account the penalty. For a GAM, the minimiser of the weighted least squares objective is given by

$$^{[+1]} = ( ) ^1 \tag{2.25}$$

with

$$= \begin{matrix} 0 & _1 & ^1 & 0 \\ 0 & 0 & _2 & ^2 \end{matrix}$$

$$= _1 \; _1 \; _1$$

and

$$= \begin{matrix} ^1 & 0 \\ 0 & ^1 (+) \end{matrix}$$

and are defined as before in the GLM paragraph. For known values of hyperparameter , with this newly defined minimiser, the rest of the procedure is like the IRLS algorithm. In order to select the hyperparameter for a model with splines, it is not possible to use likelihood based statistics because a

non-penalised likelihood ($= 0$) would always, logically, maximise the likelihood. Cross-validation is a commonly used technique for assessing hyperparameters. Considering that leave-one-out cross-validation is computationally costly, a generalised cross validation score (GCV) is used more commonly in the context of additive models :

$$= \frac{_{=1}(\quad)^2}{[\quad(\quad)]^2} \tag{2.26}$$

where $(\quad)$ is the trace of the influence matrix which corresponds to the number of effective degrees of freedom (EDF). The influence matrix is

$$= \quad ^{1} \tag{2.27}$$

Briefly, EDF is used instead of degrees of freedom (DF) because of the penalty and its impacts on the parameters estimation procedure.

2.2.3 GAMLSS

The GAM framework only accommodates distributions that are included in the linear exponential family of distributions. The negative binomial distribution is not, in general, part of this family, and neither are the multivariate negative binomial or the beta negative binomial. To work in a more general framework with any distribution, it is possible to consider generalised additive models for location, scale and shape (GAMLSS). It is possible to use this framework to include parametric or non-parametric terms in Equations (2.2), (2.6) and (2.9). If only parametric terms are included, it is called a parametric GAMLSS and the model maintains some of its properties like on the asymptotic behaviour of the distribution. A parametric GAMLSS is like a generalisation of the GLM with any distribution. A GAMLSS that includes smoothing functions such as splines in Equations (2.2), (2.6) and (2.9) is a generalisation of GAM.

2.2.3.1 Estimation Steps

In (Rigby & Stasinopoulos, 2005), two algorithms are described to estimate the parameters that maximise the penalised likelihood given the hyperparameter . The RS algorithm (from (Rigby & Stasinopoulos, 1996)) and CG algorithm (from (Cole & Green, 1992)) are possible algorithms to be used to maximise the penalised likelihood. According to (Stasinopoulos *et al.*, 2017), the RS algorithm is generally more stable than the

CG algorithm. The RS algorithm maximises the penalised log-likelihood in an iterative way between the components of the model : the location ( ), the scale ( ) and the shape (and). Put another way, this means that there is an outer iteration step that successively estimates the parameters , , and by using the inner iteration steps, which includes a modified PIRLS algorithm, which (Rigby & Stasinopoulos, 1996) have called "a modified backfitting algorithm." The other parameters (for example , and) are assumed to be fixed while in the inner iteration of a parameter (for example). Those fixed values could either be the initialised value or the last estimated value of the parameter. The outer iteration stops when the global deviance (i.e., minus twice the fitted log-likelihood, see Equations (2.2), (2.5) and (2.8)) has converged.

The penalised log-likelihood is

$$= \sum_{=1}^{4} \sum_{=1} \qquad (2.28)$$

where is the likelihood and the second part is the summation of all penalties for every non-parametric term that could have been included in any of the k parameter(s) (, , and). is the number of non-parametric terms included in the parameter.

2.2.3.1.1 Inner iteration.

The inner iteration is inspired by the GLM/GAM estimation procedure. It is a local scoring algorithm that uses pseudo-data and penalised iteratively re-weighted least squares methods. The algorithm is presented in terms of the location parameter , but the same principle applies for scale and shape parameters (for a more general presentation, refer to (Stasinopoulos *et al.*, 2017)). So, given the current values of the parameter (or , and), the weight and the pseudo-observation are calculated. The weight is

$$= \quad — \quad — \qquad (2.29)$$

where the operator ' ' represents the Hadamard element-wise product (see Equation (2.31)), is defined as before (= ( )) and is either the second derivative of the log-likelihood with respect to for the standard Newton-Raphson scoring algorithm, or the expected value of this second derivative for the Fisher scoring algorithm or is (2.30) for a quasi Newton scoring algorithm.

$$= \quad — \quad — \qquad (2.30)$$

The pseudo-observation is then given by

$$= + \quad ^{1}$$

where is

$$= — = \quad — \qquad —$$

and

$$^{1} \quad = \quad _{1}{}^{1} \; _{1} \quad _{2}{}^{1} \; _{2} \qquad ^{1} \tag{2.31}$$

Once the weight and the pseudo-observation are calculated, an algorithm described as a modified backfitting algorithm is used to recalculate and until convergence of the global deviance is reached. This modified backfitting algorithm includes the penalised iteratively re-weighted least squares.

2.2.3.1.2 Modified backfitting algorithm.

In this procedure, the parameters of the vector are estimated iteratively by separating the estimation of the coefficients associated with parametric terms ( ), the ones associated with spline 1 ( ) and the coefficients associated with spline 2 ( ) and so on if there were other non-parametric terms (see Equation (2.20)).

1. The parameters for the parametric part are estimated minimising the weighted least square (see paragraph 'GLM' in Section 2.2.2.3) of $=$ using the calculated weights .
2. Afterwards, $=$ is calculated and is estimated using the penalised weighted least square with the appropriate penalty (see paragraph "GAM" in Section 2.2.2.3).
3. Finally, fit the penalised weighted least squares to $=$ to get the estimation of .

Those steps are repeated until the vector stops varying (convergence of the modified backfitting algorithm) and and are recalculated until the global deviance convergence (inner iteration convergence) is reached. If the model contains several parameters to be estimated, the procedure is repeated for an another parameter and so on until convergence of the global deviance (outer iteration convergence).

Lastly, because the model has been estimated for a given , the last step is to determine the value of .

(Stasinopoulos *et al.*, 2017) list several ways to achieve that. Among these, there is the GCV-based method presented above (see Equation (2.26)). The that minimises the GCV will be chosen.

2.3 Numerical Analysis

2.3.1 Dataset

To better understand the usefulness of the GAM and GAMLSS models presented in previous sections, we use an automobile insurance dataset as an example. The dataset contains information about the claim history from 2009 to the end of 2019. The data are provided by a major insurance company in Canada and concern the Canadian province of Ontario. The data are not balanced, meaning that some insureds are observed for one contract, while others may be observed for up to 11 contracts. For our example, we only work with passenger cars whose use is limited to pleasure, commuting or company cars. This excludes motorbikes, snowmobiles and passenger cars for specific purposes such as driver training. Finally, although we possess information on several types of coverages, we focus on collision coverage. Those decisions were taken to reduce data heterogeneity.

We have a very large database that covers over 11 years. For comparison, (Yang & Shi, 2019) had six years of data and (Frees & Valdez, 2008), although they had nine years of data, had an average subject observation length of only 2.08 years. We have 11 years of data and the average observation time for a policy is 5.15 years and 4.53 years per driver (drivers may jump in or out of an existing policy). Table 2.2 provides an overview of database attrition. Although it is normal that the largest values are found for the shortest durations, 38.13% of the drivers were still observed for more than five years. This quality data will allow us to propose a model for predictive ratemaking and illustrate it with a numerical application from which we can draw some conclusions.

Several covariates are available for our analysis. Table 2.3 provides an overview of the covariates that were selected for modeling. Theses covariates includes two continuous covariates and some descriptive statistics are provided in Table 2.4, from which several observations can be made. We observed that the dataset contains drivers who could be described as more experienced, while the 25th percentile corresponds to 18 years of driving experience. The typical insured lives outside Toronto, is married and uses their car for pleasure or commuting. We observed $42\ 377$ drivers and $90\ 86\%$ of them have no claims. Then, $8\ 22\%$, $0\ 84\%$, $0\ 07\%$ and $0\ 01\%$ of the drivers respectively have 1, 2, 3 and 4 claims.

Duration (in years)	1	2	3	4	5	6
# drivers	7 526	8 245	5 690	4 754	2 908	2 302
Duration (in years)	7	8	9	10	11	
# drivers	1 770	1 890	1 272	1 284	4 736	

Table 2.2 – Driver's duration of observation

	Variable	Description
1	Postal code	First letter of postal code : M (Toronto), P (North), K (East), L (Central), N (West)
2	Gender	Insured's gender : Female or Male/Other
3	Marital status	Insured's marital status : Married, Divorced, Separated, Single, Widow/Widower
4	Usage	Primary vehicle use : Pleasure, Business (company car), Commute, Farm (Farmer's car)
5	Number years driving	Number years driving as a continuous variable
6	Vehicle age	Vehicle age in years as a continuous variable

Table 2.3 – Description of covariates

Variable	Mean	25^{th} percentile	Median	75^{th} percentile
Number years driving	30.08	18.00	30.00	41.00
Vehicle age (in years)	6.31	2.00	6.00	9.00

Table 2.4 – Continuous covariate statistics

Finally, the database was separated into an in-sample and an out-of-sample dataset. Table 2.5 shows the distribution of observations between each of the subsamples and the total number of observations. The training set contains 75% of the policies chosen at random and their observations from the year 2009 to 2018. The validation set contains the remaining 25 % of the policies, from the year 2009 to 2018. Policies

written in 2019 were excluded because, as the contracts were annual, the majority expired after March 2020 and the beginning of the Covid-19 lockdown in Ontario. As an indication, the claim frequency of policies written in 2019 is half that of 2018. The expected value of the predictive distribution will contains all known past claim experience information available, given that a new insured's past claims experience is available to insurers from Ontario through the Autoplus database [1]. The database used was quite large, and this amount of information is not always available. If the intended usage would be a on a smaller dataset, simulation studies should be performed for evaluating the predictive sample performance.

	Number of observations	Years
Training sample	145 121	2009 - 2018
Validation sample	47 781	2009 - 2018
Remove due to Covid-19	25 874	2019
Total	218 776	

Table 2.5 – Number of observations in each subsample of the dataset

2.3.2 Cross-sectional data models

The models are fitted including all covariates described in Section 2.3.1 into the log-linear predictor from Equations (2.1) and (2.2). Table 2.6 shows the overdispersed negative binomial regression has an edge over the equidispersed Poisson model based on the widespread adjustment statistics AIC and BIC. A quasi-Poisson regression has also been performed on the data and the estimated dispersion parameter is 2.021. All this leads to the usual result of overdispersion of claim frequency data.

All estimates of the parametric terms for the fitted models in Section 2.3 can be found in Appendix 2.5. Table 2.7 shows how significant the signals from continuous variables are in both basic parametric models. From the p-value, it can be concluded that the log-linear relationship between these covariates and the mean parameter is statistically significant. It can be investigated whether a non-linear relationship would improve the fit of the model further. Table 2.6 shows better statistics for semi-parametric models for both the Poisson model and the negative binomial model. This implies that the additional parameters associated

1. Autoplus is a large database that all insurance companies in Ontario subscribe to. It contains information on all auto insurance histories in this Canadian province.

with the splines improve the log-likelihood of the model enough to be considered.

	AIC	BIC	Log-likelihood	EDF
Parametric models				
Poisson	30 977 96	31 126 24	15 473 98	15
Negative Binomial			—	16
Semi-parametric models				
Poisson cubic splines	30 871 27	31 062 03	15 416 34	19.2972
NB cubic splines	30 841 64		15 400 72	20.0960
Poisson P-splines	30 869 09	31 062 41	15 414 99	19.5564
NB P-splines		31 052 87	—	21.5134

Table 2.6 – In-sample goodness-of-fit statistics for cross-section data models

	Estimate	Std. Error	z value	Pr(z)
Poisson Model				
Number years driving	-0.0095	0.0014	-6.94	0.0000
Vehicle age (in years)	-0.0508	0.0044	-11.55	0.0000
NB Model				
Number years driving	-0.0097	0.0014	-6.94	0.0000
Vehicle age (in years)	-0.0509	0.0044	-11.45	0.0000

Table 2.7 – Estimated coefficients for continuous covariates in parametric cross-sectional data models

When analysing the shape of the splines in Figure 2.1 it can be seen that the splines are somewhat similar for the two splines considered. This was unsurprising because the number of EDF (Table 2.6) are similar and cubic regression splines and P-splines yield similar results in this situation. In addition, one may observe that the inclusion of splines really benefits the modeling of the "number of years driving" covariate as the shape of the function is far from linear when more flexibility is allowed. Only the splines of the Poisson models

are included, because the negative binomial models yield almost identical results. The similarity is striking for the spline associated with vehicle age, where each spline even crosses the zero dashed line at about the same place (around seven years). Regarding the number-of-years-driving spline, functions differ a bit more between the type of spline considered. Nonetheless, each spline shares interesting characteristics, which are where the functions are zero and a somewhat similar functions' minimum. The most noticeable difference between the cubic regression spline and the P-spline is the shape of the parabola under the zero line, but both of them indicate a shift at around 40 years of driving, even though it is not completely identical. This similarity between the splines is desirable in the sense that it seems to indicate that the models succeed in fetching the signal from the covariate and do not yield an over-fitted relationship that would depend on the model structure.

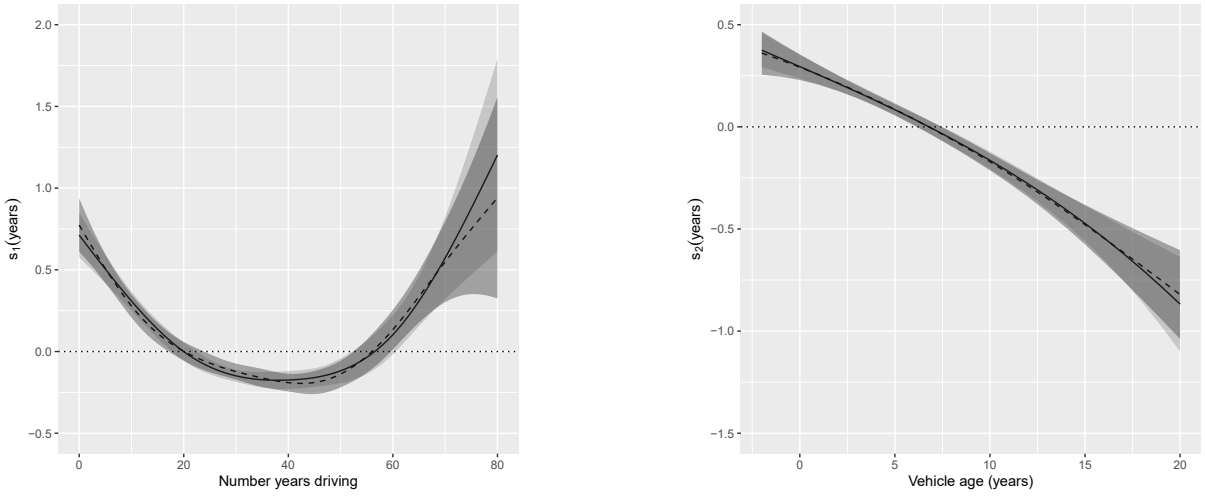

Figure 2.1 – Estimated cubic spline (solid) and P-spline (dashed) for the Poisson model

2.3.3 Panel data models

We fit panel data models on the data to check how adding the claims history to the modeling influences the results. Results are shown in Table 2.8. Compared to Table 2.6, we observe that the panel data models have the smallest value for the AIC. This is also the case for the BIC criterion, except for models including P-splines, as the BIC criterion penalises more depending on the number of parameters.

In theory, one would have to compare each of the possible combinations of splines, but in this section the primary objective is to illustrate the benefit of using longitudinal semi-parametric models. In any case, the

splines look about the same. Even though the P-splines shown in Figures 2.2 and 2.3 are different from Figure 2.1, they still share some common important features. First, for the number of years driving, we retrieve the parabolic shape again, the minimum is around 40 and the function crosses zero around 20 and 60. As for the vehicle age, the shape is bumpier, but is still zero around 8 and monotonically decreasing, except for the very first part of the function.

Traditionally, continuous variables are segmented and introduced as categorical variables in a model for ratemaking. As described in (Henckaerts *et al.*, 2018), continuous variables are not well-suited for insurance purposes, which explains the interest in investigating their segmentation. However, there is no consensus on how to proceed. In this work, by including continuous variables in different models by using smoothing functions, we have observed that the shifts in the splines are in about the same places, even if there are differences and the underlying model distribution is different. This can be an indication of how to segment the covariate. Furthermore, one can voluntarily fit a bumpier function to analyse what is going on with the smoothing function when more flexibility is allowed.

	AIC	BIC	Log-likelihood	EDF
Parametric models				
MVNB	30 908 99	31 067 16	15 438 50	16
Beta-NB			—	17
Semi-parametric models				
MVNB cubic splines	30 813 65		15 384 85	21.9784
Beta-NB cubic splines		31 039 97	—	22.9316
MVNB P-splines	30 857 56		15 401 94	26.8358
Beta-NB P-splines		31 125 43	—	27.5295

Table 2.8 – In-sample goodness-of-fit statistics for panel data models

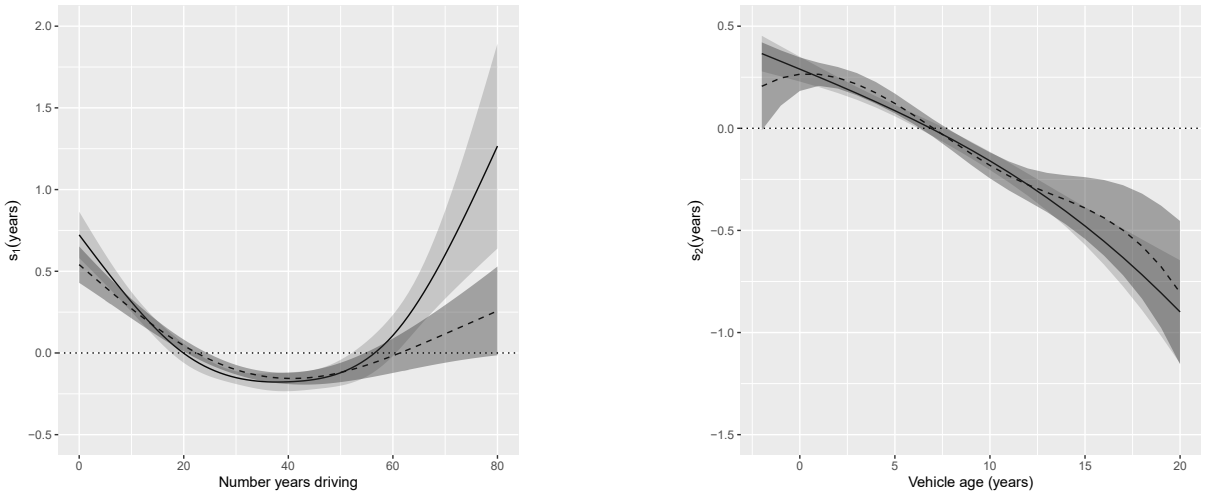

Figure 2.2 – Estimated cubic spline (solid) and P-spline (dashed) for the MVNB model

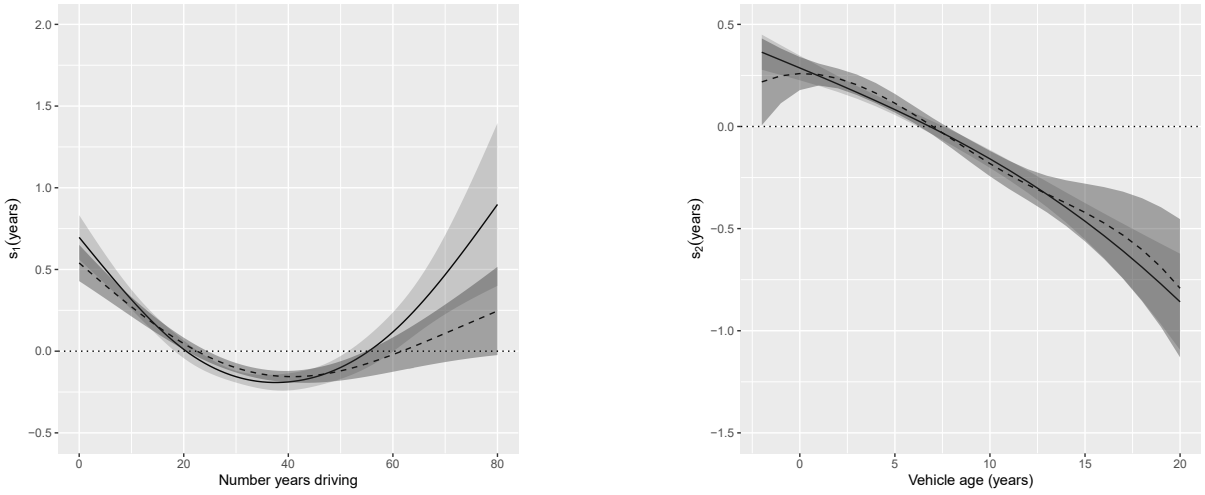

Figure 2.3 – Estimated cubic spline (solid) and P-spline (dashed) for the Beta-NB model

2.3.4 Analysing the Results

2.3.4.1 Scoring Rules

Model assessment is done using proper scoring rules for count data (see (Czado *et al.*, 2009)) on the test dataset. Those proper scoring rules are the logarithm score, the quadratic score, the spherical score, the ranked probability score, the Dawid-Sebastiani score and the squared error score. All six have been calcu-

lated, but only the logarithm score, the Dawid-Sebastiani score and the squared error score are presented in Table 2.9 because the other scores resulted in values that were too similar to help assess the models. Before discussing the results, let us present the three selected scores.

The logarithm score logs( ) is the probability of observing the test sample using the predictive distribution of the estimated model. It is calculated in the same way as the negative log-likelihood, but using the test sample, rather than the fitting sample. For one observation, we have :

$$\text{logs}( ) = \log( )$$

where is the probability of observing the data . The squared error ses( ) is a common distance measure used in statistics and it corresponds to the squared difference of the data and the estimated prediction :

$$\text{ses}( ) = ( )^2$$

Lastly, the Dawid-Sebastiani score dss( ) is a squared error that has been normalised by the variance, and a variance-based penalty has been added :

$$\text{dss}( ) = \frac{}{} ^2 + 2\log( )$$

To obtain the overall score, it suffices to sum the individual score for all the observations of the test set. The best model would minimise those scores. Considering the ses score is equivalent to the normal deviance, the deviance Poisson was added in Table 2.9 because the Poisson distribution is a more common distribution for modeling count data.

Table 2.9 shows the proper scoring rules and the Poisson deviance for the test set. The results clearly indicate that longitudinal models have better scores than cross-sectional models. The semi-parametric Beta-NB model minimised the most scores. For the squared-error score, the results are very similar and do not really help to assess the best model. The score was nonetheless included because it is widely used. The model that minimises all the three other scores is the cubic spline Beta-NB model. This tells us that the model that performed the best on predicting the out-of-sample dataset is a longitudinal semi-parametric model.

	Logarithmic	Dawid-Sebastiani	Poisson deviance	Squared error
Parametric models				
Poisson	5 132 46	89 317 24	10 240 54	1 023 73
NB	5 127 03	89 813 46	10 240 56	1 023 87
MVNB	5 122 86	90 050 49		
Beta-NB		—	10 235 00	1 023 81
Cubic splines				
Poisson	5 125 40	91 915 64	10 226 43	1 023 95
NB	5 119 42	92 774 36	10 225 63	1 024 10
MVNB	5 120 85	93 065 27	10 221 73	1 024 12
Beta-NB		—		
P-splines				
Poisson	5 123 85	92 091 24	10 223 33	1 023 87
NB		92 765 03	10 223 24	1 024 05
MVNB	5 133 30	91 709 53	10 223 24	1 023 86
Beta-NB	5 127 56	—		

Table 2.9 – Proper scoring rules and Poisson deviance on validation sample

2.3.4.2 Premium Comparison

We now compare the *a priori* premium for three different risk profiles, meaning the premium estimate for cross-sectional models or, equivalently, for longitudinal models without history. Those three profiles were selected because they represent low, medium and high risks. Covariate values for each profile are detailed in Table 2.10. Table 2.11 contains the mean and variance estimates for all models fitted in this work. First, let's discuss dispersion. From the results, we can see more or less equidispersion for low and medium risks but, for high risks, the difference between Poisson and the other models that allow overdispersion is noticeable. This is another indication that an equidispersed model is not suitable for claim frequency modeling purposes. Indeed, we would underestimate the real risk of our portfolio by using such a model.

For low and medium risks, longitudinal models tend to have higher premiums than cross-sectional models (with the exception of cubic spline models for medium risk). In longitudinal models, as can be seen in Table 2.12, the favourable history is going to reduce the premium. We can't have this correction in cross-sectional models, so the premium is lower right away. However, a favourable five-year history tends towards a lower predictive premium than the premium of cross-sectional models. Another observation from Table 2.12 is that semi-parametric models reduce the premium for a favourable history less rapidly in comparison to parametric models. Indeed, this was also observed in Table 2.11, where the parametric models estimate a higher premium for low and medium risk and the opposite for high risk. In a context where a two-year history could not be qualified as a mature and reliable experience, the prudence of semi-parametric models in premium reduction may be more welcomed for insurance usage.

	Postal Code	Gender	Marital status	Usage	Number years driving	Vehicle age
Low risk	K	Female	Married	Farm	45	15
Medium risk	L	Male/Other	Married	Commute	20	7
High risk	M	Male/Other	Single	Commute	2	2

Table 2.10 – Characteristics of risk profiles

Model	Low risk		Medium risk		High risk	
	Mean	Variance	Mean	Variance	Mean	Variance
Parametric models						
Poisson	0.0087	0.0087	0.0230	0.0230	0.0538	0.0538
NB	0.0087	0.0088	0.0232	0.0241	0.0547	0.0592
MVNB	0.0089	0.0089	0.0235	0.0238	0.0548	0.0569
Beta-NB	0.0089	0.0091	0.0235	0.0241	0.0548	0.0572
Cubic splines						
Poisson	0.0078	0.0078	0.0094	0.0094	0.0729	0.0729
NB	0.0079	0.0080	0.0101	0.0103	0.0745	0.0830
MVNB	0.0078	0.0078	0.0092	0.0093	0.0743	0.0779
Beta-NB	0.0080	0.0081	0.0097	0.0099	0.0728	0.0766
P-splines						
Poisson	0.0075	0.0075	0.0099	0.0099	0.0759	0.0759
NB	0.0075	0.0076	0.0101	0.0103	0.0780	0.0872
MVNB	0.0087	0.0088	0.0103	0.0104	0.0669	0.0698
Beta-NB	0.0086	0.0087	0.0106	0.0107	0.0665	0.0698

Table 2.11 – *A priori* premiums

Model	*A priori*	Number of years without claims				
		1	2	3	4	5
Parametric models						
MVNB	0.0235	0.9841	0.9686	0.9537	0.9392	0.9251
Beta-NB	0.0235	0.9862	0.9728	0.9598	0.9471	0.9347
Cubic splines						
MVNB	0.0092	0.9940	0.9881	0.9823	0.9765	0.9708
Beta-NB	0.0097	0.9946	0.9893	0.9840	0.9788	0.9737
P-splines						
MVNB	0.0103	0.9933	0.9867	0.9801	0.9737	0.9673
Beta-NB	0.0106	0.9942	0.9884	0.9827	0.9770	0.9714

Table 2.12 – Bonus depending on number of years without claims (medium risk) for mixture models

While differences have been observed in how premiums are reduced for an insured with a favourable history in a longitudinal model, it is interesting to analyse how premiums are penalised for claims in these same models. Table 2.13 presents the premium after one year of history according to claims experience. While the *a priori* premium of the MVNB model is lower than that of the Beta-NB model and the premium level of the MVNB model was lower than that of the Beta-NB model for policyholders with a favourable history, it can be observed that the MVNB model penalises claims more than the Beta-NB model. By analysing Equations (2.7) and (2.10), we find a clue that helps explain this result. For the MVNB model, we notice that the same parameter is found in the numerator and the denominator. In the fitted models, this parameter varied between 1.4495 and 1.5335. In the Beta-NB model, the parameters are different and the fitted models estimated significantly different values for the numerator and denominator. The value of the numerator (between 1.6803 and 1.7983) is significantly lower than the value of the denominator (between 232.3917 and 239.5266), which gives the Beta-NB model a more stable predictive premium, thus reducing the bonuses and maluses of the claims experience. Based on the results obtained for these data which are detailed in Table 2.9, this additional parameter of the Beta-NB model appears to provide a significant improvement to the modeling.

Model	*A priori*	1	2	3	4
Parametric models					
MVNB	0.0235	1.6630	2.3418	3.0207	3.6996
Beta-NB	0.0235	1.5732	2.1601	2.7471	3.3340
Cubic splines					
MVNB	0.0092	1.6422	2.2905	2.9387	3.5869
Beta-NB	0.0097	1.5477	2.1008	2.6539	3.2070
P-splines					
MVNB	0.0103	1.6415	2.2897	2.9379	3.5861
Beta-NB	0.0106	1.5480	2.1019	2.6557	3.2096

Table 2.13 – Malus depending on number of claims (medium risk) for mixture models

Lastly, there are mathematical reasons from the construction of the models that explain the results of Tables 2.12 and 2.13. Indeed, the heterogeneity distribution from Equation (2.3) represents the heterogeneity that the regressors of the *a priori* model could not capture and that affects all contracts from the same driver over time. Let's start with the MVNB model. It was mentioned that the parameter , which comes from the Gamma distribution, varied between 1.4495 and 1.5335. The higher values are associated with semi-parametric models and this is not insignificant. By model construction, the mean of the Gamma distribution is 1. The point estimate for the variance, 1 , is lower for semi-parametric models. Considering that the objective of including past experience is to capture the residual heterogeneity that is not explained by the *a priori* pricing variables, this means that semi-parametric models succeed in more accurately capturing the signal from the covariates, thus reducing residual heterogeneity variability and the importance of experience in predictive ratemaking. This results in the smaller bonuses and maluses for semi-parametric longitudinal models. In the case of the Beta-NB model, the mechanics operate somewhat differently. The structure of the model does not force the expectation to be unitary. The highest parameters and are still associated with semi-parametric models. The expectation of the beta distribution being $(+)$, the

expected value of is lower for the semi-parametric models.

2.4 Conclusion

In this work, the required steps to include smoothing functions within generalised model predictors were presented, using cubic regression splines and P-splines as examples. Because it does not use any simplifying assumptions as to the value of the *a priori* explanatory variables, the semi-parametric longitudinal count model approach described in this work could be used for practical situations. Smoothing functions were included in longitudinal count models to improve predictive power and better capture *a priori* information. Indeed, we observed the reduction of the random effect's variance in the MVNB model. Other approaches could be used to evaluate improvement in the predictive power of the *a priori* part. For instance, CANN (Wüthrich & Merz, 2019) tries to quantify the residual error to be captured : if the model improves substantially using such an approach, it is because an aspect of the modeled phenomenon has been neglected. We introduced the general ideas of GAMs and GAMLSSs, detailed their estimation methods and discussed possible smoothing functions to be included in the modeling of the parameter(s). To illustrate the usefulness of the models, they have been applied to insurance data. The results indicate that the inclusion of smoothing functions within longitudinal models is relevant to improve predictive power on the out-of-sample and reduces the weight of past experience in bonus-malus predictive premiums analysis in comparison with a parametric model. Two longitudinal models generalising the Poisson and negative binomial models were estimated and the results pointed to the relevance of the additional parameter of the Beta-NB distribution. Another interesting result was that the shifts in the estimated splines are in about the same places and share several important characteristics. This result was not necessarily expected at the beginning considering the significant differences between the cross-sectional and longitudinal models. This can be an indication of how to segment the covariate. In future work, it might be interesting to compare the vehicle approach and the driver approach or even to include dependency relationships between drivers in the same household. GLMs are already the industry standard for pricing, but it appears that GAMs/GAMLSSs would be a step forward in incorporating more flexibility in modeling.

	Poisson	NB	MVNB	Beta-NB
0	-2.9053 (0.0867)	-2.8910 (0.0883)	-2.8942 (0.0914)	2.0357 (0.4353)
1	-0.2510 (0.0717)	-0.2532 (0.0730)	-0.2481 (0.0760)	-0.2525 (0.0756)
	-0.1836 (0.0712)	-0.1844 (0.0725)	-0.1788 (0.0754)	-0.1824 (0.0750)
	-0.2158 (0.0722)	-0.2175 (0.0735)	-0.2112 (0.0764)	-0.2168 (0.0761)
	-0.2152 (0.0915)	-0.2183 (0.0930)	-0.2196 (0.0968)	-0.2254 (0.0965)
2	-0.1031 (0.0367)	-0.1041 (0.0373)	-0.0998 (0.0389)	-0.1018 (0.0388)
3	0.6007 (0.2683)	0.6131 (0.2735)	0.5442 (0.2766)	0.5152 (0.2834)
	0.3846 (0.2057)	0.3938 (0.2091)	0.3812 (0.2113)	0.3624 (0.2147)
	0.2400 (0.0398)	0.2429 (0.0405)	0.2451 (0.0421)	0.2416 (0.0420)
	0.3942 (0.2691)	0.4011 (0.2733)	0.3790 (0.2750)	0.3872 (0.2742)
4	0.0025 (0.0014)	0.0037 (0.0014)	-0.0010 (0.0014)	0.0063 (0.0014)
	-0.0342 (0.0044)	-0.0331 (0.0044)	-0.0362 (0.0045)	-0.0361 (0.0045)
	-0.4023 (0.0994)	-0.4037 (0.1009)	-0.3983 (0.1040)	-0.3919 (0.1035)
5	-0.0095 (0.0412)	-0.0097 (0.0419)	-0.0097 (0.0430)	-0.0098 (0.0430)
6	-0.0508 (0.1438)	-0.0509 (0.1451)	-0.0499 (0.1492)	-0.0496 (0.1486)
	-	-	1.4495 (0.2117)	-
	-	-	-	1.6803 (0.2889)
	-	-	-	232.3917 (92.0594)

Table 2.14 – Estimated parametric terms for parametric models

		Poisson	NB	MVNB	Beta-NB
	0	-3.5259 (0.0715)	-3.5165 (0.0728)	-3.5176 (0.0753)	1.3749 (0.4302)
1		-0.2492 (0.0718)	-0.2515 (0.0730)	-0.2500 (0.0757)	-0.2528 (0.0753)
		-0.1797 (0.0713)	-0.1808 (0.0726)	-0.1761 (0.0751)	-0.1803 (0.0747)
		-0.2314 (0.0722)	-0.2333 (0.0736)	-0.2297 (0.0762)	-0.2324 (0.0758)
		-0.2079 (0.0915)	-0.2102 (0.0931)	-0.2080 (0.0964)	-0.2169 (0.0962)
2		-0.1241 (0.0369)	-0.1260 (0.0374)	-0.1212 (0.0388)	-0.1235 (0.0387)
3		0.6370 (0.2684)	0.6516 (0.2735)	0.5883 (0.2757)	0.5655 (0.2820)
		0.4204 (0.2058)	0.4307 (0.2092)	0.4134 (0.2115)	0.3982 (0.2147)
		0.1592 (0.0415)	0.1612 (0.0421)	0.1611 (0.0436)	0.1603 (0.0435)
		0.2596 (0.2695)	0.2677 (0.2736)	0.2484 (0.2751)	0.2583 (0.2744)
4		0.1060 (0.1002)	0.1094 (0.1018)	0.1043 (0.1046)	0.1114 (0.1040)
		0.0303 (0.0429)	0.0333 (0.0435)	0.0285 (0.0446)	0.0301 (0.0445)
		-0.4520 (0.1439)	-0.4539 (0.1453)	-0.4455 (0.1489)	-0.4406 (0.1485)
		-	-	1.5335 (0.2318)	-
		-	-	-	1.7983 (0.3236)
		-	-	-	239.5266 (94.2778)

Table 2.15 – Estimated parametric terms for models including cubic splines

	Poisson	NB	MVNB	Beta-NB
0	-3.5218 (0.0715)	-3.5117 (0.0728)	-3.5271 (0.0759)	1.3719 (0.4298)
1	-0.2515 (0.0718)	-0.2542 (0.0731)	-0.2498 (0.0757)	-0.2548 (0.0754)
	-0.1818 (0.0713)	-0.1830 (0.0726)	-0.1786 (0.0751)	-0.1825 (0.0747)
	-0.2335 (0.0723)	-0.2361 (0.0736)	-0.2222 (0.0762)	-0.2278 (0.0758)
	-0.2101 (0.0916)	-0.2129 (0.0931)	-0.2173 (0.0965)	-0.2233 (0.0962)
2	-0.1246 (0.0369)	-0.1265 (0.0374)	-0.1115 (0.0388)	-0.1132 (0.0387)
3	0.6340 (0.2685)	0.6475 (0.2735)	0.5726 (0.2762)	0.5498 (0.2822)
	0.4166 (0.2058)	0.4262 (0.2093)	0.4060 (0.2111)	0.3869 (0.2146)
	0.1591 (0.0415)	0.1604 (0.0422)	0.1958 (0.0428)	0.1926 (0.0427)
	0.2569 (0.2696)	0.2650 (0.2736)	0.2956 (0.2749)	0.3051 (0.2741)
4	0.1036 (0.1002)	0.1068 (0.1018)	0.0646 (0.1042)	0.0718 (0.1037)
	0.0286 (0.0429)	0.0307 (0.0436)	0.0055 (0.0438)	0.0053 (0.0438)
	-0.4550 (0.1440)	-0.4569 (0.1453)	-0.4273 (0.1490)	-0.4202 (0.1484)
	-	-	1.5324 (0.2317)	-
	-	-	-	1.7950 (0.3229)
	-	-	-	238.9721 (93.9641)

Table 2.16 – Estimated parametric terms for models including P-splines

CHAPITRE 3

MODÉLISATION SEMI-PARAMÉTRIQUE DE L'EXPOSITION AU RISQUE AVEC CONTRAINTES DE MONOTONICITÉ EN ASSURANCE AUTOMOBILE

Depuis que la collecte de données télématiques s'est répandue dans le marché de l'assurance automobile, il y a un intérêt pour utiliser le kilométrage parcouru dans la modélisation du risque de réclamation. En effet, une telle mesure aurait l'avantage d'offrir à l'assuré un certain contrôle sur sa prime d'assurance, en plus d'inciter à la réduction de l'utilisation de la voiture, ce qui a de nombreux avantages sociaux, notamment concernant la réduction de la pollution. Classiquement, la mesure d'exposition en tarification chez les assureurs est la durée du contrat sur une base annuelle. Cela signifie que l'assuré paye la même prime quotidienne pour chacune de ses journées et qu'il serait remboursé la moitié de la prime annuelle s'il devait annuler sa police au milieu du terme. Pour une couverture protégeant contre le vol du véhicule, l'utilisation de cette mesure d'exposition semble appropriée. Toutefois, si on considère la couverture collision ou la responsabilité civile, l'utilisation du véhicule constitue la principale cause d'exposition au risque. En utilisant la durée comme mesure d'exposition, des assurés partageant des profils de risque similaires peuvent se retrouver à être solidaires de ceux qui utilisent davantage leur véhicule, étant donné que peu d'assurés sont capables de fournir une estimation assez juste de leur kilométrage annuel lorsque la question est posée par l'assureur en début de contrat. Dans cette optique, l'étude du kilométrage comme mesure d'exposition pour une couverture comme celle protégeant contre les collisions est pertinente pour la science actuarielle. D'ailleurs, considérant que certaines recherches sont venues à la conclusion que l'inclusion du kilométrage réellement parcouru pouvait rendre non significative des variables comme le genre de l'assuré ((Ayuso *et al.*, 2016b), (Verbelen *et al.*, 2018)), une variable socialement moins acceptable qu'autrefois, voire maintenant interdite dans certaines juridictions ((Agence des droits fondamentaux de l'Union européenne, Cour européenne des droits de l'homme et Conseil de l'Europe, 2018)), il est d'autant plus important de s'intéresser à sa modélisation.

Plusieurs modèles ont proposé d'inclure le kilométrage parcouru pour améliorer la tarification. L'idée d'une tarification modulée selon l'usage du véhicule remonte à (Vickrey, 1968). La télématique n'étant pas disponible à l'époque, ils ont plutôt proposé d'utiliser une taxe sur l'essence ou sur les pneus. Les données télématiques ont donné un second souffle à cette idée avec des travaux tels que (Ayuso *et al.*, 2019), (Boucher *et al.*, 2017), (Ma *et al.*, 2018), (Verbelen *et al.*, 2018) et (Guillen *et al.*, 2021). Toutefois, aucune de

ces recherches n'a utilisé des splines sous contraintes de forme pour modéliser la distance parcourue. En science actuarielle, des splines avec contraintes de monotonicité ont surtout été utilisées dans le contexte de la modélisation de la mortalité, puisque le taux de mortalité est une fonction monotone non décroissante ((Shang, 2019),(Tang *et al.*, 2022)).

Depuis une dizaine d'années, les données télématiques font partie de la nouvelle réalité de la tarification en assurance automobile. Autant les praticiens que les chercheurs se sont intéressés au potentiel et à la valeur de ces données. L'utilisation de données collectées par GPS sur les habitudes de conduite des assurés permettrait d'améliorer les modèles de tarification existants ou de créer de nouveaux produits d'assurance. (Verbelen *et al.*, 2018) avait d'ailleurs conclu que la meilleure structure de tarification incluait autant des variables dites «traditionnelles» que des données télématiques. Le kilométrage comme mesure d'exposition au risque avait également été considéré. Un des défis reliés à l'utilisation des données télématiques est la taille des données. Plusieurs recherches se sont intéressées à les résumer, tels que (Weidner *et al.*, 2016), (Gao & Wüthrich, 2018) et (Jeong, 2022), par exemple. Une autre approche a été d'étudier la quantité d'information qu'il était nécessaire de collecter avant que l'information commence à être redondante ((Duval *et al.*, 2022)). Le kilométrage annuel figure parmi les informations collectées par GPS les plus faciles à recueillir et analyser. Cela peut s'avérer pratique pour de la tarification dynamique, puisque les données sont collectées en temps réel. Les conséquences de la télématique sur les produits d'assurance et l'assurabilité ont été étudiées entre autres par (Eling & Kraft, 2020), notamment concernant des questions éthiques. L'utilisation du kilométrage annuel apparait selon nous comme une donnée qui respecte la vie privée. En effet, le détail des déplacements n'aurait pas besoin d'être conservé en mémoire et nous ne considérons pas les moments où ces kilomètres ont été conduits, ce qui n'interfère pas avec le mode de vie de la personne qui pourrait être corrélé à son groupe socioculturel, par exemple.

Dans le modèle à effets aléatoires de (Boucher & Turcotte, 2020), il était observé que la relation entre le kilométrage et le risque devenait décroissante lorsqu'un certain seuil (en kilomètres) était dépassé. Ce résultat était expliqué par le fait que les usagers dont le kilométrage se situait dans les quantiles les plus élevés de la distribution sont des risques différents de l'assuré typique. Par exemple, ils empruntent davantage les autoroutes qui sont considérées comme des routes plus sûres puisqu'il n'y a pas d'intersection et que la circulation y est unidirectionnelle. Dans une autre recherche, il était avancé par (Boucher *et al.*, 2017) qu'il pouvait s'agir d'un effet d'apprentissage. En conduisant davantage, on deviendrait un meilleur risque, car on aurait acquis de l'expérience de conduite. Peu importe l'explication, cela signifie que l'effet marginal du

kilométrage sur le risque d'accident n'est pas observé, puisque le risque marginal d'un kilomètre supplémentaire est forcément croissant. En effet, choisir de conduire un kilomètre supplémentaire, plutôt que de laisser sa voiture stationnée, expose forcément davantage le bien assuré au risque de collision, peu importe le profil de risque de l'assuré. On observe plutôt l'effet apparent, influencé par d'autres facteurs non contrôlés. La relation entre le kilométrage et le risque devant nécessairement être croissante, on souhaite corriger les irrégularités observées. Nous proposons donc un modèle longitudinal semi-paramétrique construit par effets aléatoires, mais dont les termes non paramétriques estimant la relation entre la distance parcourue et le risque d'accident seraient soumis à des contraintes de forme pour contraindre une relation croissante. Il existe différentes splines et différentes façons d'imposer la monotonicité de la spline. Contrairement à (Boucher *et al.*, 2017) et (Boucher & Turcotte, 2020) qui utilisaient uniquement les P-splines, plusieurs splines ont été considérées pour ce travail. Nous choisissons de travailler avec les P-Splines, les splines cubiques de régression et les splines de régression en plaque mince et nous expliquerons les impacts sur la modélisation. L'objectif est d'obtenir un modèle pouvant être utilisé en pratique qui a une relation logique entre le kilométrage et le risque ainsi qu'entre la durée et le risque.

Dans ce travail, les termes non paramétriques soumis à une contrainte de monotonicité seront utilisés à l'intérieur d'un modèle additif généralisé pour la moyenne, l'échelle et la forme (*generalised additive models for location, scale and shape* ou GAMLSS, en anglais). Ce type de modélisation permet d'inclure des termes paramétriques ou non paramétriques dans l'estimation de plusieurs paramètres, voir (Rigby & Stasinopoulos, 2005) ou (Stasinopoulos *et al.*, 2017). Les GAMLSSs généralisent à la fois les modèles linéaires généralisés ((Nelder & Wedderburn, 1972)) et les modèles additifs généralisés ((Hastie & Tibshirani, 1986)), puisque toute distribution peut être utilisée avec ce cadre et pas seulement les distributions membres de la famille exponentielle linéaire. L'idée d'utiliser les GAMLSSs pour la modélisation de données d'assurance n'est pas nouvelle ((Heller *et al.*, 2006), (Tzougas & Frangos, 2014), (Turcotte & Boucher, 2023)). C'est d'ailleurs le cadre qui a été utilisé dans (Boucher & Turcotte, 2020). Toutefois, le problème d'un modèle cohérent pouvant être utilisé en tarification qui utilise le kilométrage comme mesure d'exposition au risque n'a pas été répondu et c'est ce problème que nous ciblons avec ce travail.

L'article est organisé de la façon suivante : dans la prochaine section, les notions de base concernant les splines utilisées seront rappelées et la façon d'introduire des contraintes de forme sera détaillée. À la section 3.1.4, nous analyserons comparativement les différentes splines et contraintes de forme. À la section 3.2, nous présentons les modèles de données de panel à effets aléatoires dans lesquels des contraintes

de forme ont été incluses dans la modélisation de l'exposition au risque mesurée en kilométrage et en année. Ces modèles sont estimés en utilisant des données d'un assureur canadien présentées à la section suivante. À la section 3.3, nous présentons les modèles estimés avec et sans contraintes de forme et nous étudions l'impact de ces contraintes sur la structure des primes selon différents historiques de réclamations. Finalement, la section 3.4 conclut cet article.

3.1 Fonctions de lissage avec contraintes de monotonicité

Plusieurs techniques sont disponibles pour obtenir une fonction de spline continue et croissante pour modéliser le lien entre le kilométrage et le risque. L'algorithme des *pool adjacent violators* (PAVA) est considéré comme l'une des premières techniques pour produire une fonction de régression monotone. Il a été proposé par (Brunk *et al.*, 1972) et est basé sur la régression isotone qui consiste à projeter une fonction non paramétrique dans un ensemble de fonctions croissantes. (Kruskal, 1965) a suggéré cette technique dans un contexte de régression. Un défaut important de cette approche était qu'elle ne puisse pas produire une fonction lisse. Pour corriger cette limite, plusieurs approches en deux étapes basées sur l'algorithme PAVA ont été proposées ((Friedman & Tibshirani, 1984), (Mammen, 1991), (Mukerjee, 1988) or (Ghosh, 2007)). Ces procédures en deux étapes consistaient en un lissage sans contrainte sur l'espace des paramètres et une monotonisation. Ces techniques sont les précurseurs des splines avec contraintes de forme.

(Wood, 1994) propose une approche pénalisée pour ajuster une spline de régression cubique monotone. La méthode est basée sur la représentation polynomiale par morceaux ainsi que sur des conditions linéaires suffisantes pour assurer la monotonicité qui ont d'abord été utilisées dans (Hyman, 1983). Une autre approche consiste à utiliser la non-négativité ou la non-positivité de la dérivée de la fonction de lissage pour obtenir les contraintes de monotonicité (ex. (Zhang, 2004)). (Ramsay, 1998) a proposé une façon d'estimer une fonction strictement monotone deux fois différentiable en résolvant une équation différentielle linéaire homogène. L'algorithme d'estimation n'était toutefois pas optimal, en partie à cause de la lourdeur computationnelle de la procédure. D'autres se sont toutefois intéressés à cette approche. (Zhang, 2004) a intégré la méthode dans un modèle de régression généralisé, tandis que (Meyer, 2008) a généralisé le travail de (Ramsay, 1988) afin d'y ajouter une contrainte de convexité, en plus de celle de monotonicité. Des approches spécifiquement adaptées au support local des B-Splines ont également été développées. Un fait notable est qu'il existe une condition suffisante, non nécessaire pour assurer la monotonicité de la spline. Une séquence de paramètres non décroissante (non croissante) est suffisante pour garantir une spline mo-

notone croissante (décroissante) (voir (Schumaker, 2007)). La première approche basée sur cette condition est probablement celle de (Kelly & Rice, 1990) où on procède à une optimisation sous cette contrainte. (He & Shi, 1998) et (Rousson, 2008) ont également proposé des approches impliquant des contraintes linéaires. (Pya & Wood, 2015) ont proposé une approche qui reparamétrise les coefficients de la B-spline afin de garantir la monotonicité, sans procéder à une estimation sous contraintes. (Lu *et al.*, 2007) examine l'estimation de B-splines monotones spécifiquement dans un modèle de comptage pour des données longitudinales. Ils utilisent des données d'essai clinique pour illustrer leurs méthodes. Pour notre travail, on retient une approche basée sur les B-splines, une approche basée sur les splines cubiques de régression et une méthode numérique basée sur des contraintes concernant la valeur de la dérivée. La méthode de (Pya & Wood, 2015) procède à une estimation sans contrainte et celle de (Wood, 1994) est basée sur des contraintes linéaires.

Cette section fait la présentation de trois splines et de techniques permettant de les rendre monotones, signifiant que l'on a $(\ )$ $(\ )$. Les splines sont des fonctions de lissage définies par morceaux. Par définition, les fonctions de lissage qui peuvent être utilisées dans le contexte des GAMs ou des GAMLSSs peuvent être écrites sous la forme d'une combinaison linéaire entre des fonctions de base ($(\ )$) et des paramètres ( ) :

$$(\ ) = \sum_{=1} (\ ) \tag{3.1}$$

Cette caractéristique permet de préserver la forme linéaire du prédicteur, sans exiger un lien log-linéaire entre les variables explicatives et le paramètre de moyenne. Pour une présentation détaillée de ces splines, il est possible de consulter (Wood, 2017). La première approche repose sur les B-splines, la seconde se base sur les splines cubiques de régression et la troisième sur les splines de régression en plaque mince.

3.1.1 B-splines avec contraintes

Les P-splines sont fondées sur les fonctions de base B-splines, estimées en utilisant une pénalité pour contrôler le lissage[1]. Les fonctions de base B-splines sont des fonctions de base locales pour ajuster une spline de l'ordre $(\ + 1)$. Si $= 2$, cela correspond à une spline cubique qui est considérée comme optimale pour obtenir une spline lisse, mais bien ajustée aux données ((Wood, 2017)). Elle est définie par ($+$ $+2$) noeuds . Des fonctions de base locales impliquent que chaque fonction de base est uniquement non nulle

1. Considérant que la méthode pour assurer la monotonicité est reliée à la structure des fonctions de base plutôt qu'à la procédure d'estimation, on réfère quelques fois spécifiquement aux B-splines pour la suite.

sur l'intervalle compris entre les m + 3 noeuds adjacents. Elle peut être représentée sous la forme additive

$$( ) = \sum_{=1} ( )$$

Les fonctions de base d'une B-spline sont le plus souvent définies sous la forme récursive

$$( ) = \frac{}{+ +1} \; {}^{1}( ) + \frac{+ +2}{+ +2 \quad +1} \; {}_{+1}{}^{1}( ) \quad \text{où} \tag{3.2}$$

$${}^{1}( ) = \begin{cases} 1 & \text{si} \quad +1 \\ 0 & \text{sinon.} \end{cases}$$

D'après (De Boor, 1978), la première dérivée d'une B-spline avec des noeuds uniformément espacés est

$$( ) = \frac{1}{-} \sum_{=2} {}^{1}( ) \; {}^{1}$$

Les fonctions de base d'une B-splines étant non négative par définition, une condition suffisante pour que $( ) \quad 0$ est donc que ${}^{1} = {}_{1} \quad 0$. Cela implique qu'une séquence croissante des paramètres produira une fonction monotone croissante. La figure 3.1 illustre comment une séquence de paramètres croissants permet d'assurer la monotonicité de la spline. Chaque ligne pointillée représente une des ($= 10$) fonctions de base multipliées par une séquence croissante de paramètres. L'addition de chacune des lignes pointillées forme la spline résultante illustrée en rouge. Une des fonctions de base a été représentée par une ligne plus large afin de mieux observer la forme d'une fonction de base.

Afin d'obtenir cette séquence croissante, on définit les paramètres de cette façon (Pya & Wood, 2015) :

$$= \begin{cases} {}^{1} & \text{si} \quad = 1 \\ {}_{1} + \sum_{=2} \exp( ) & \text{si} \quad = 2 \end{cases} \tag{3.3}$$

où les paramètres sont des paramètres sans contrainte. Définissons $= (\; {}_{1} \; \exp(\; {}_{2}) \; \exp(\; {}_{3}) \quad \exp(\quad))$. La fonction de lissage peut ainsi être représentée par $(\quad) = \quad$, où $= \quad {}_{1}(\quad) \quad {}_{2}(\quad) \quad (\quad)$ est la i[e] rangée de la matrice de design et est

$$= \begin{matrix} 1 & 0 & 0 & & 0 \\ 1 & 1 & 0 & & 0 \\ 1 & 1 & 1 & & 0 \\ & & & & \\ 1 & 1 & 1 & & 1 \end{matrix}$$

En utilisant cette approche, il n'est pas nécessaire de procéder à une optimisation avec contraintes sur l'espace des paramètres puisque la structure du modèle assure une spline croissante.

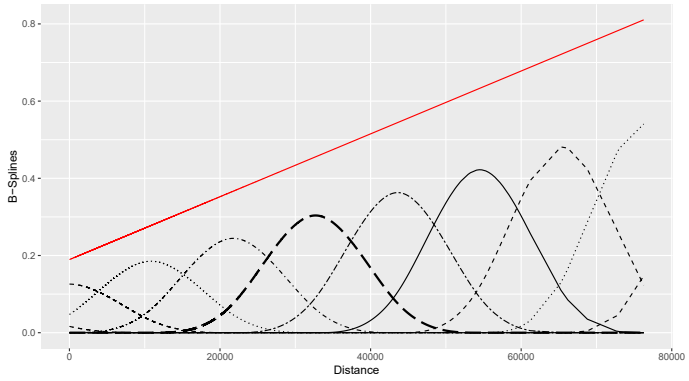

Figure 3.1 – Illustration d'une B-spline avec paramètres croissants

3.1.2 Splines cubiques de régression avec contraintes

Une spline cubique de régression est une autre méthode pour définir des splines cubiques [2]. Elles sont définies par noeuds telles que

$$( ) = \quad ( ) \quad + \quad ^{+}( ) \quad _{+1} + \quad ( ) \quad + \quad ^{+}( ) \quad _{+1} \quad = \quad \sum_{=1} \quad ( ) \tag{3.4}$$

où

$$( ) = \begin{cases} ( ) \quad + \ ^{+}( ) \quad _{+1} + \ ^{+}( ) & \text{si} \quad = \quad +1 \\ ( ) \quad + \ ^{+}( ) \quad _{+1} + \quad ( ) & \text{si} \quad = \\ ( ) \quad + \ ^{+}( ) \quad _{+1} & \text{sinon.} \end{cases} \tag{3.5}$$

2. Dans ce qui suit, les splines cubiques de régression seront parfois abrégées par «splines cubiques».

Le vecteur de paramètres $\quad$ doit être estimé. Les termes $\quad(\ )$, $\quad^{+}(\ )$, $\quad(\ )$, $\quad^{+}(\ )$ sont définis dans le tableau 3.1. On définit aussi $\quad$, la ^{e} rangée de la matrice $\quad=\begin{matrix}\\0\end{matrix}\quad$ où $\quad=\quad^{1}\quad$. Finalement, les éléments non nuls des matrices $\quad_{(\quad 2)\ (\ )}$ et $\quad_{(\quad 2)\ (\quad 2)}$ sont définis au tableau 3.1.

Fonctions de base pour une spline cubique			
$(\ )=(\quad_{+1}\quad)\ (\quad_{+1}\quad)$	$(\ )=(\quad_{+1}\quad)^{3}\ (\quad_{+1}\quad)\ (\quad_{+1}\quad)(\quad_{+1}\quad)\ \ 6$		
$^{+}(\ )=(\quad\quad)\ (\quad_{+1}\quad)$	$^{+}(\ )=(\quad\quad)^{3}\ (\quad_{+1}\quad)\ (\quad_{+1}\quad)(\quad\quad)\ \ 6$		
Définitions des éléments non nuls des matrices D et B			
$=1\ (\quad_{+1}\quad)$	$_{+1}=\ (\quad_{+1}\quad)^{\ 1}\ (\quad_{+2}\quad\ _{+1})^{\ 1}$	$_{+2}=1\ (\quad_{+2}\quad\ _{+1})$	
$=(\quad_{+2}\quad)\ 3$		$1\quad\ 2$	
$_{+1}\ =\quad_{+1}=(\quad_{+2}\quad\ _{+1})\ 6$		$1\quad\ 3$	

Table 3.1 – Termes définissant une spline de régression cubique

(Wood, 1994) présente des conditions linéaires suffisantes pour assurer la croissance ou la décroissance d'une spline cubique. Basées sur (Yan, 1987), des conditions nécessaires suffisantes pour garantir la monotonicité sont que les valeurs de $\quad=\quad(\quad_{+1})(\quad_{+1}\quad)\ (\ (\quad_{+1})\quad(\ ))$ et $\quad=\quad(\ )(\quad_{+1}\quad)\ (\ (\quad_{+1})\quad(\ ))$ se trouvent dans l'espace délimité par la courbe représentée à la figure 3.2. En posant $\quad=(\ (\quad_{+1})\quad(\ ))\ (\quad_{+1}\quad)$, $\quad$ et $\quad$ représentent les dérivées évaluées à chacune des extrémités de l'intervalle divisées par la pente de la droite sécante de cet intervalle. En assurant la monotonicité sur chaque intervalle, la spline résultante sera elle aussi monotone. L'espace délimité par la courbe de la figure 3.2 est obtenu par l'étude de $\quad(\ )$, $\quad(\ )$ et $\quad(\ )$ faite par (Fritsch & Carlson, 1980) et représente en réalité l'union de six espaces pour les paramètres qui respectent les conditions pour assurer la monotonicité. Notons que l'espace des contraintes ne peut pas se définir linéairement. Basé sur (Hyman, 1983), (Wood, 1994) propose de se restreindre à la zone A, afin de définir des contraintes linéaires suffisantes pour assurer la monotonicité. Ces contraintes imposent que les valeurs de $\quad$ et $\quad$ soient comprises entre 0 et 3. L'approche proposée par (Wood, 1994) permet de procéder à une optimisation comptant un nombre moins grand de contraintes sur l'espace des paramètres que l'approche numérique présentée à la sous-section 3.1.3.

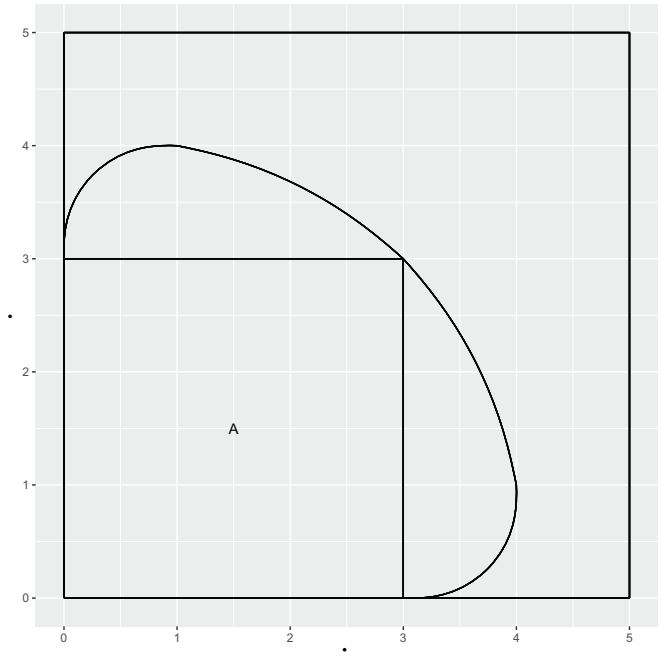

Figure 3.2 – Espace des contraintes

3.1.3 Approche numérique avec les TPRS

Les splines de régression en plaque mince (*Thin Plate Splines* ou TPRS, en anglais) sont une méthode plus générale que les deux splines précédemment abordées, voir (Duchon, 1977). Elles permettent entre autres d'estimer des splines multivariées ($d \geq 1$). Un avantage de cette méthode est qu'il n'est pas nécessaire de spécifier des emplacements pour les noeuds. Pour ces splines, il suffit de spécifier une valeur m, qu'on appelle le rang de la spline. La spline peut s'écrire sous la forme linéaire

$$s(\boldsymbol{x}) = \sum_{i=1}^{n} \delta_i \eta_{md}(||\boldsymbol{x} - \boldsymbol{x}_i||) + \sum_{j=1}^{M} \alpha_j \phi_j(\boldsymbol{x}),$$

où $\boldsymbol{\delta}$ et $\boldsymbol{\alpha}$ sont des vecteurs de paramètres à estimer. Les fonctions de base sont définies par :

$$\eta_{md}(r) = \begin{cases} \dfrac{(-1)^{m+1+d/2}}{2^{2m-1}\Pi^{d/2}(m-1)!(m-d/2)!} r^{2m-d}\log(r) & d \text{ pair}, \\ \dfrac{\Gamma(d/2-m)}{2^{2m}\Pi^{d/2}(m-1)!} r^{2m-d} & d \text{ impair} \end{cases} \tag{3.6}$$

où d est la dimension de la spline (ici $d = 1$, car la spline est unidimensionnelle) et $M = \binom{m+d-1}{d}$. Pour ce

qui est de ( ), ces fonctions sont des polynômes linéairement indépendants s'étendant sur l'espace des polynômes dans d'un degré inférieur à . On a la contrainte que = , où = ( ).

Cette méthode présente l'inconvénient qu'elle est plus lourde en termes de temps de calcul que les deux autres méthodes lorsque le nombre de données augmente. Les splines de régression en plaque mince servent à réduire le temps de calcul, mais l'expression de la spline reste inchangée.

S'il n'y a pas ou peu de théorie sur les contraintes de monotonicité pour une spline ou qu'elles sont difficiles à implémenter, il est possible d'obtenir une spline monotone croissante en s'assurant que la dérivée soit toujours positive (voir équation (3.7)). Une dérivée toujours positive assure la croissance de la spline. On utilise comme contraintes des valeurs de la dérivée évaluée à plusieurs endroits répartis sur l'ensemble du domaine. On procède donc à une optimisation avec un grand nombre de contraintes.

$$(\ _0) = \lim_{0} \frac{(\ _0 + \ ) \quad (\ _0)}{} \quad 0 \tag{3.7}$$

Il s'agit d'une méthode alternative qu'il est également possible d'utiliser avec les B-splines et splines cubiques de régression. À la section 3.3, la valeur de la dérivée est évaluée en 100 points répartis sur l'ensemble du domaine. Ces 100 évaluations de la dérivée ont servi à définir les contraintes de monotonicité.

3.1.4 Comparaison entre les approches

Nous expliquons ici les différences entre ces méthodes et certaines considérations pour l'usage souhaité. À cette fin, les fonctions de base des trois splines pour des paramètres unitaires sont illustrées à la figure 3.3. Les splines sont exprimées avec dix fonctions de base chacune, ajustées sur le domaine de la distance mesurée en kilomètres pour l'illustration. La ligne rouge représente la spline résultante de ces fonctions de base et de coefficients égaux à un. Encore une fois, une des fonctions de base a été représentée par une ligne plus large afin de mieux en observer la forme.

Pour la méthode basée sur les B-splines, le vecteur de coefficients est estimé, puis reparamétré en coefficients qui assurent une séquence croissante. En pratique, nous avons observé que choisir un grand nombre de noeuds (ce qui augmente le nombre de paramètres dans la spline) pouvait rendre la convergence plus difficile à atteindre. En effet, les paramètres associés aux noeuds situés sur les quantiles les plus

élevés de la distribution résultent d'une somme de termes exponentiels de plus en plus nombreux (voir l'équation (3.3)). Similairement, plus le nombre de paramètres est important, plus les premiers éléments du vecteur interfèrent dans l'évaluation de plusieurs paramètres . Un autre point à considérer concernant la convergence est le choix de la fonction de lien dans le cas où un nombre important de paramètres est utilisé pour définir la spline dans un modèle généralisé. En intégrant la spline à l'intérieur d'une fonction de lien exponentiel, on obtient l'exponentielle d'une sommation de termes exponentiels. Cependant, il n'est pas nécessaire de procéder à une optimisation sous contraintes puisque la structure du modèle assure une spline croissante, ce qui peut être un avantage.

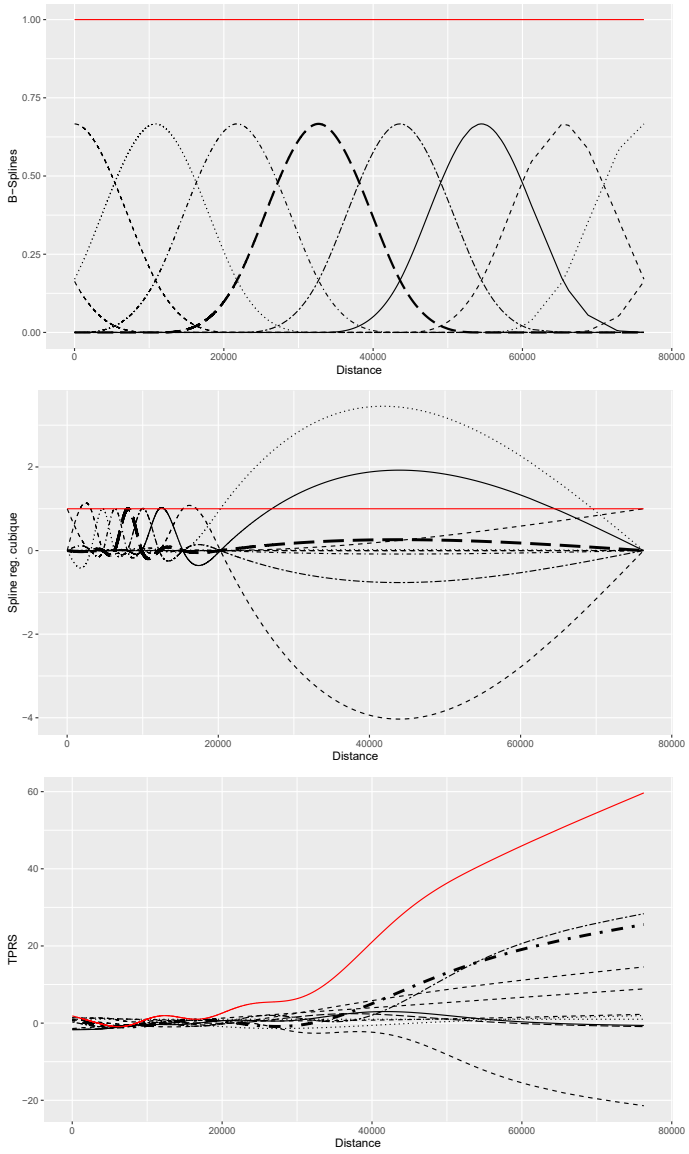

Figure 3.3 – Illustration des fonctions de base pour chacune des splines

Les fonctions de base des splines cubiques de régression (second graphique de la figure 3.3), contrairement aux fonctions de base des B-splines (premier graphique de la figure 3.3), ne sont pas définies localement. Une séquence croissante des coefficients du vecteur des paramètres n'est donc pas suffisante pour assurer la croissance de la spline. Si la spline est contrainte à être croissante et que les fonctions de base ne sont pas locales, on se demande s'il peut être possible de corriger la spline non pas localement, mais sur l'ensemble de son domaine et ainsi «combattre» l'effet apparent observé dans (Boucher *et al.*, 2017) et (Boucher & Turcotte, 2020). Ces deux recherches n'ont utilisé que les B-splines/P-splines. Considérant que les observations sont moins nombreuses dans la queue de la distribution, il n'est pas impossible que les irrégularités de forme observées soient dues à un surajustement local causé par un manque d'observations dans cette partie du domaine. De plus, si nous considérons que la corrélation entre le kilométrage et les autres habitudes de conduite est un phénomène qui affecte les données sur l'ensemble du domaine, cela peut représenter un choix judicieux de corriger les irrégularités de forme en utilisant des fonctions de base non nulles sur tout ou presque tout le domaine.

Dans le même ordre d'idées, les fonctions de base des splines de régression en plaque mince ne sont pas non plus locales (troisième graphique de la figure 3.3). Contrairement aux deux autres approches, les TPRS comportent l'avantage notable de ne pas nécessiter la spécification d'emplacement pour les noeuds, alors que cette spécification découle de décisions subjectives. De plus, l'absence d'emplacement pour les noeuds permet d'éviter d'avoir à définir une méthode d'extrapolation au-delà de la valeur du dernier noeud. Pour la durée du contrat, qui est toujours comprise entre zéro et un, la question de l'extrapolation n'est pas un problème. Par contre, la distance conduite n'est pas bornée. Les TPRS sont reconnues comme une très bonne méthode de lissage, (Wood, 2017) allant même jusqu'à les décrire comme étant un lisseur idéal. Pour toutes ces raisons, il a été décidé de l'inclure dans cette étude comparative.

3.2 Modèle à effets aléatoires

Les données de panel se caractérisent par l'observation à travers le temps d'un même individu. Les données de panel présentent l'avantage important pour l'assurance automobile de permettre la modélisation longitudinale du risque. L'historique de pertes est donc utilisé pour capturer une partie de l'hétérogénéité résiduelle qui n'a pas été expliquée par les variables de segmentation *a priori*. En d'autres mots, l'historique de perte viendrait mesurer des caractéristiques qui ne peuvent pas être actuellement recueillies comme l'impatience au volant, la propension à être inattentif ou le manque de réflexe (Denuit *et al.*, 2007). Étudier

la modélisation longitudinale du risque de perte est donc d'intérêt pour la science actuarielle.

Dans une première recherche de (Boucher & Turcotte, 2020), deux types de modèles pour données de panel avaient été considérés : un modèle Poisson à effets aléatoires et un modèle Poisson à effets fixes. Dans les deux cas, le nombre de réclamations d'un assuré au temps est modélisé par une loi de Poisson de moyenne en intégrant un effet , aléatoire ou fixe, qui lie les contrats appartenant à un même assuré :

$$\text{Poisson}(\quad = \quad)$$

avec $=$ représentant les caractéristiques *a priori* de l'assuré. Dans un modèle à effets aléatoires, les sont des variables indépendantes et identiquement distribuées. Dans un modèle à effets fixes, les sont des paramètres fixes qui doivent être estimés. (Boucher & Turcotte, 2020) tirait la conclusion qu'inclure un facteur individuel fixe permettait de tenir compte des caractéristiques individuelles des assurés qui influencent le risque et d'observer une relation presque linéaire entre le kilométrage et le risque d'accident. Cependant, ce modèle à effets fixes ne peut pas être utilisé en pratique puisqu'on ne saurait pas quel facteur individuel accorder à un nouvel assuré. Néanmoins, ce résultat justifie de s'intéresser à un modèle dont la mesure d'exposition serait le kilométrage, car une telle relation proportionnelle entre le kilométrage et le risque de sinistre a beaucoup de potentiel. Le modèle à effets aléatoires peut être utilisé en pratique puisqu'aucun facteur n'a besoin d'être déterminé à l'avance pour un nouvel entrant. Il serait donc pertinent d'essayer de corriger les irrégularités de forme qui ont été observées.

Les modèles de données de panel supposent que tous les contrats annuels appartenant au même conducteur sont dépendants pour $= 1$. Si désigne toujours le facteur d'hétérogénéité, la distribution jointe d'un modèle à effets aléatoires pour données de panel peut être exprimée comme suit :

$$\Pr[\quad _1 = \quad _1 \qquad = \quad] = \int_0 \prod_{=1} \Pr[\quad = \quad _1 \qquad] \quad (\quad) \tag{3.8}$$

En sélectionnant des distributions conjuguées pour la distribution de comptage et celle de l'effet aléatoire, la vraisemblance est explicite et plus facile à utiliser dans des situations pratiques.

Les modèles longitudinaux considérés dans ce travail de recherche sont tous basés sur la distribution binomiale négative multivariée (MVNB). Si nous supposons Poisson( ), dont le paramètre d'hétérogénéité est distribué selon une loi gamma de moyenne unitaire et de variance $\frac{1}{}$, la valeur espérée de

$[\quad]$ est inchangée et la distribution jointe peut s'exprimer par :

$$\Pr[\quad_1 = \quad_1 \qquad = \quad] = \quad_{=1} \frac{}{!} \frac{(\quad + \ )}{(\ )} \frac{}{+} \quad (\quad + \ ) \tag{3.9}$$

où $= \quad_{=1}$ et $= \quad_{=1}$. Le paramètre peut être modélisé avec des variables explicatives et inclure les mesures d'exposition , pour la durée du contrat, et , pour le nombre de kilomètres parcourus,

$$= \exp(\ _1(\quad) \quad _2(\qquad) \qquad) = \quad ^1(\quad) \tag{3.10}$$

où $_1$ et $_2$ sont des splines.

Cette distribution est une généralisation de la binomiale négative. Il s'agit d'une distribution classique pour les données de panel surdispersées ($[\quad] = \qquad [\quad] = \quad + (\quad)^2$) (Hausman *et al.*, 1984). Si $_{(1:\ )}$ désigne l'historique de réclamations ($_{(1:\ )} = [\quad_1 = \quad_1 \qquad = \quad]$), la distribution prédictive est une loi binomiale négative telle que

$$\Pr[\quad_{+1} = \quad_{+1} \quad_{(1:\ )}] = \frac{(\quad + \quad_{+1} + \ )}{(\quad + \ ) \quad_{+1}!} \frac{\quad_{+1}}{+ \quad_{+1} +} \qquad \frac{}{+ \quad_{+1} +}^{+}$$

dont l'espérance prédictive est

$$[\quad_{+1} = \quad_{+1} \quad_{(1:\ )}] = \quad_{+1} \frac{+}{+} \tag{3.11}$$

Les modèles GAMs peuvent seulement être utilisés avec les distributions faisant partie de la famille exponentielle linéaire. La binomiale négative ne fait pas partie de cette famille de distributions, pas plus que la binomiale négative multivariée. Afin de travailler dans un cadre plus général où le choix de distribution n'est pas restreint, il est possible de considérer les modèles additifs généralisés d'emplacement, d'échelle et de forme. Avec ce modèle-cadre, il est possible d'inclure des termes paramétriques ou non paramétriques dans l'équation (3.10). Dans le processus d'estimation du paramètre de moyenne, on estime séparément les paramètres de chaque spline et ceux associés aux termes paramétriques. Il est donc possible d'inclure soit une reparamétrisation du vecteur de coefficients ou de procéder à une optimisation sous contraintes sur les paramètres associés à la spline uniquement. Pour plus de détails sur la modélisation avec effets aléatoires ou sur la MVNB, il est possible de consulter (Boucher *et al.*, 2008). Pour davantage d'explications sur les GAMLSS, il est possible de consulter (Stasinopoulos *et al.*, 2017). (Turcotte & Boucher, 2023) a également abordé spécifiquement la modélisation d'une régression MVNB avec le cadre GAMLSS.

3.3 Applications numériques

3.3.1 Données

Cette base de données provient d'une grande compagnie en assurance de dommage et concerne les risques automobiles en ligne personnelle de la province de l'Ontario. Seulement les polices qui ont été observées un minimum de 100 jours ont été conservées dans l'analyse. De plus, seulement les réclamations associées à des accidents de la route (couverture collision) ont été recensées pour cette analyse puisque nous considérons que le kilométrage parcouru par l'automobile est peu associé au risque de vol ou de vandalisme, par exemple.

Une partie des informations de la base de données a été collectée à l'aide d'un appareil télématique installé à l'intérieur des véhicules. Il a donc été possible de déterminer de manière fiable le nombre de kilomètres parcourus par le véhicule pendant la période d'assurance. Il est à noter que seulement 10 à 15 % du portefeuille global de l'assureur a décidé de souscrire au programme de télématique. D'après les explications reçues, surtout les technophiles qui apprécient les nouvelles technologies et les conducteurs plus à risque comme les conducteurs inexpérimentés ou présentant un mauvais dossier de conduite ont souscrit au programme de télématique puisqu'un rabais était offert. Le portefeuille de télématique a donc une fréquence de réclamation plus élevée que le portefeuille global, s'établissant à 6%. Considérant que les modèles en pratique sont estimés avec des variables explicatives, un petit nombre d'entre elles ont été incluses pour illustrer la segmentation. Ces variables sont le genre, le statut marital et l'usage du véhicule. Les graphiques de la figure 3.4 offrent une représentation visuelle de la répartition des observations entre les catégories.

Les figures 3.5 et 3.6 présentent la distribution de la mesure d'exposition classique, soit la durée du contrat, et celle du kilométrage. On remarque que les contrats qui ont été observés moins de 100 jours ont été retirés. Le contrat moyen a duré $0\ 645$ an. Pour ce qui est du kilométrage, on remarque que la distribution est asymétrique à droite. La moyenne de la distance parcourue est de 10 398 km, alors que la médiane se situe plutôt à 8 561 km. Le kilométrage maximal observé est quant à lui de 76 272 km.

La base de données compte 59 685 observations et le tableau 3.2 présente l'attrition de la base de données au fil du temps. La base de données a été séparée en une base de données d'estimation comptant 80% des polices et une base de données test contenant le 20% restant.

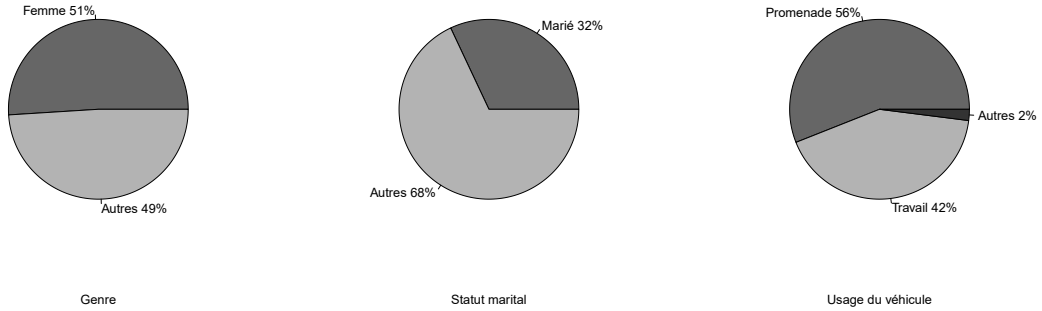

Figure 3.4 – Répartition des variables explicatives par catégorie

Nombre de périodes d'assurance	**1**	**2**	**3**	**4**	**5**	**6**
Nombre d'assurés	12 562	9 746	3 420	844	415	11
Proportion (%)	46.5	36.1	12.7	3.1	1.5	0.0

Table 3.2 – Répartition du nombre de périodes d'assurance observées

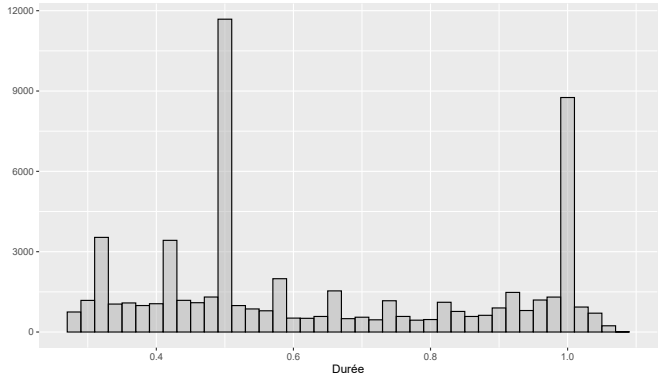

Figure 3.5 – Histogramme de la durée des polices (en année). Chaque bande a une largeur de 0.02 année.

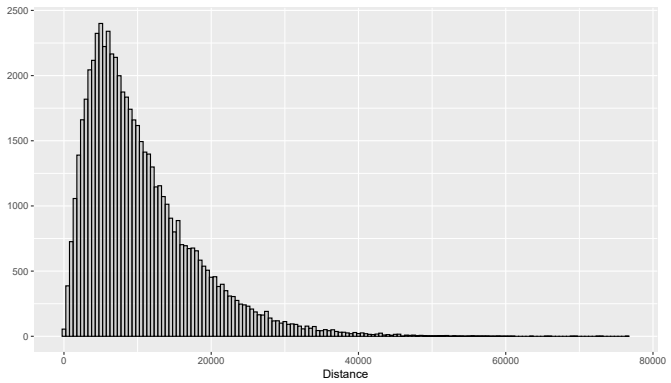

Figure 3.6 – Histogramme de la distance parcourue (en km). Chaque bande a une largeur de 500 km.

3.3.2 Analyse des résultats

3.3.2.1 P-splines.

Les figures 3.7 et 3.8 présentent les résultats pour la P-spline, une spline basée sur les fonctions de base B-splines estimées avec une pénalité. On retrouve encore une fois la forme décroissante observée dans (Boucher *et al.*, 2017) et (Boucher & Turcotte, 2020) pour la spline associée au kilométrage. La spline associée à la durée est également légèrement décroissante sur certains segments du domaine. Lorsqu'une contrainte de croissance est imposée sur les P-splines, on observe que la spline de la durée s'aplanit. Il est intéressant d'observer que si une contrainte de monotonicité est imposée sur les deux splines, dans un contexte où on modélise le risque de collision, c'est la spline de la durée qui s'aplanit alors que la spline du kilométrage dirige le niveau de la prime.

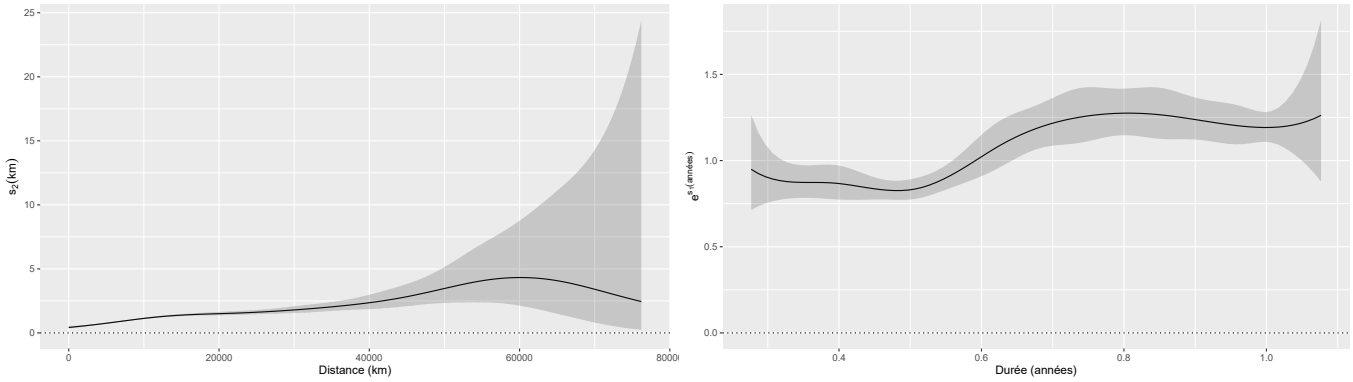

Figure 3.7 – P-splines sans contrainte de monotonicité

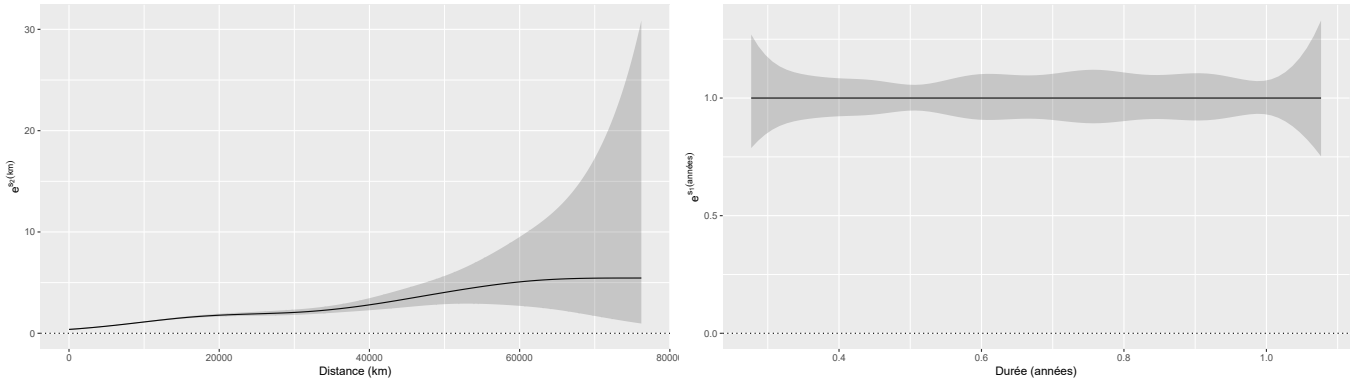

Figure 3.8 – P-splines avec contraintes de monotonicité

La contrainte de monotonicité se justifie d'après notre compréhension du risque : le risque de perte est marginalement croissant avec l'usage du véhicule tel que conclut dans (Boucher & Turcotte, 2020) avec le modèle à effets fixes. Toutefois, la monotonicité n'est pas en contradiction avec les résultats du modèle sans contrainte puisque le cas monotone est compris dans l'intervalle de confiance de ce modèle. Le rôle de la contrainte est donc d'établir une structure de prime sans possibilité d'arbitrage pour les assurés.

3.3.2.2 Splines cubiques de régression.

Les figures 3.9 et 3.10 présentent les résultats. On constate que pour obtenir une spline croissante pour le kilométrage, il n'était pas nécessaire d'imposer des contraintes de monotonicité, mais simplement d'utiliser une spline basée sur des fonctions de base qui ne soient pas locales (figure 3.3). Comme il est possible de l'observer à la figure 3.6, les données se font plus rares après 40 000 km. Les fonctions de base locales B-splines, autant avec des données espagnoles dans (Boucher *et al.*, 2017), que des données canadiennes dans (Boucher & Turcotte, 2020), estimaient cette décroissance. D'un autre côté, la spline associée à la durée, dont les observations sont bien réparties sur l'ensemble de son domaine, présente plusieurs similitudes avec la spline de la figure 3.7. Par contre, il demeure que la forme obtenue pour la spline de durée n'est pas souhaitable pour l'utiliser comme mesure d'exposition dans un modèle de tarification, car la spline estimée est légèrement décroissante sur certains segments du domaine. En essayant d'obtenir une meilleure forme pour la durée en imposant la monotonicité, on observe que la spline de la distance est pour ainsi dire inchangée par rapport à la spline de régression cubique sans contrainte. Pour la spline de la durée, on retrouve la même forme aplatie qu'à la figure 3.8, signifiant probablement que le modèle nécessite la décroissance de la spline pour être optimal.

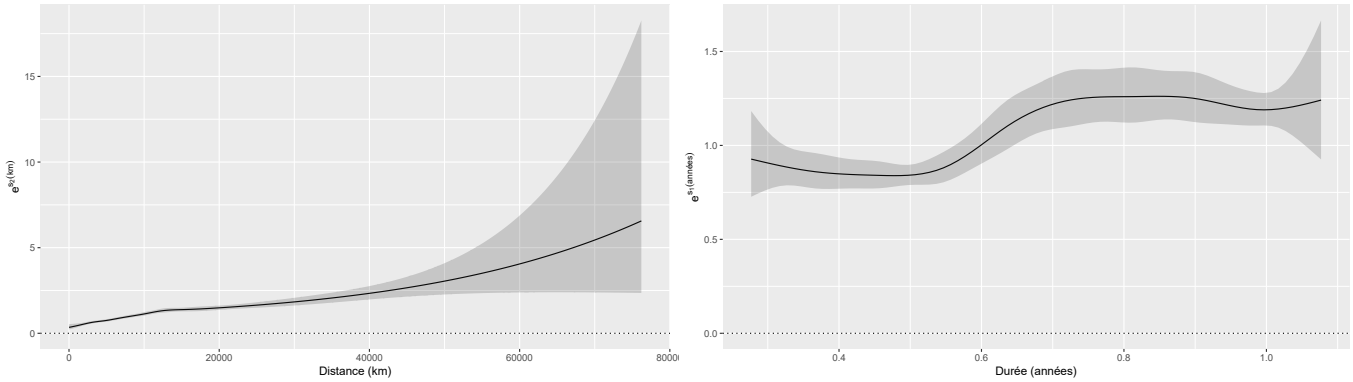

Figure 3.9 – Splines de régression cubiques sans contrainte de monotonicité

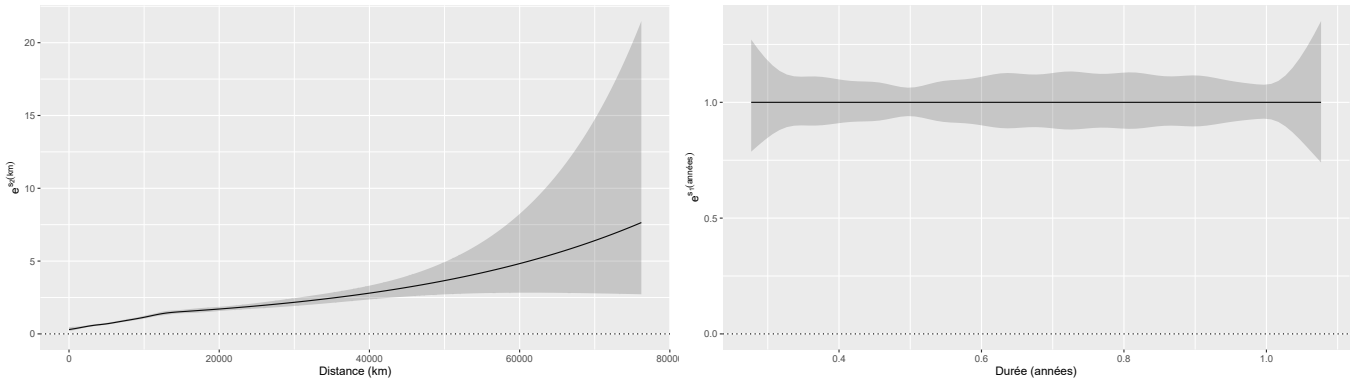

Figure 3.10 – Splines de régression cubiques avec contraintes de monotonicité

3.3.2.3 TPRS.

Les figures 3.11 et 3.12 présentent les résultats des estimations. On peut remarquer que les splines sans contrainte sont essentiellement croissantes, si on fait exception d'une légère décroissance autour des quantiles les plus élevés du graphique de droite de la figure 3.11. Il y a une certaine stabilisation dans le modèle sans contrainte pour la durée que vient corriger le modèle avec contraintes, sans pour autant aplanir complètement la courbe comme dans les deux cas précédents. Il est intéressant de constater que, plutôt que d'aplanir la spline de la durée, la contrainte la rend plutôt linéaire. Sur la base des splines seulement, ce modèle pourrait se justifier pour la tarification. Les TPRS ne sont pas influencées par la localisation des noeuds, puisqu'il n'est pas nécessaire d'en définir comme pour les deux autres méthodes. Les TPRS ne sont pas non plus définis localement. L'ensemble de ces caractéristiques pourrait expliquer que les modèles TPRS contraints se distinguent des deux autres méthodes étudiées.

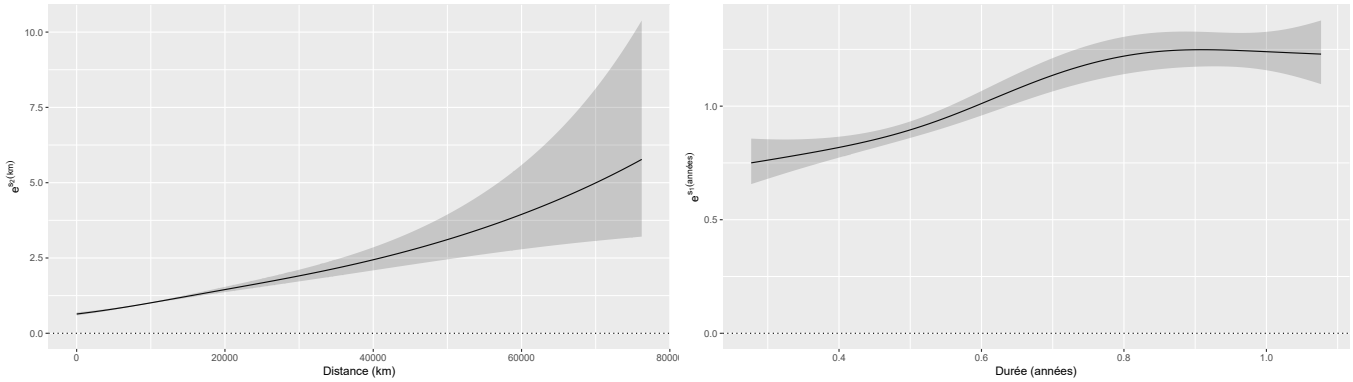

Figure 3.11 – TPRS sans contrainte de monotonicité

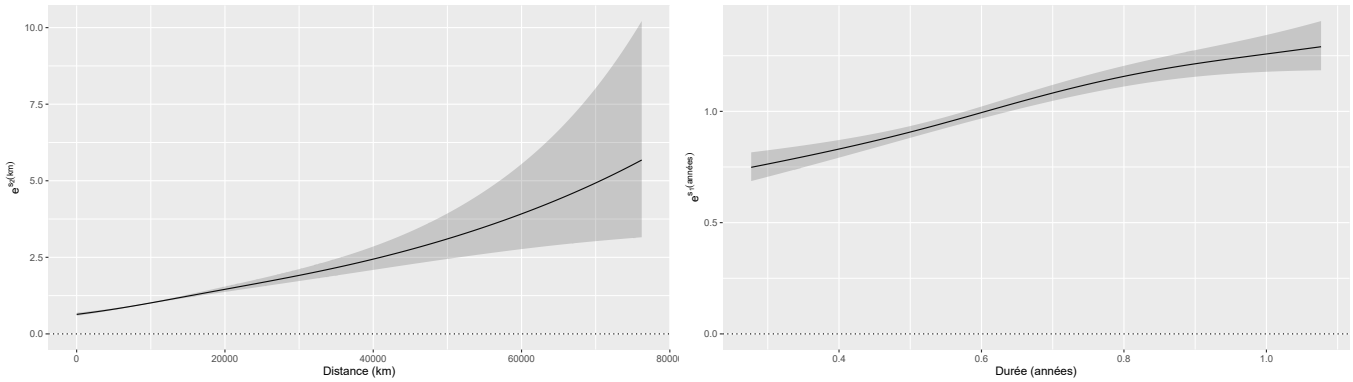

Figure 3.12 – TPRS avec contraintes de monotonicité

3.3.2.4 Remarques supplémentaires

En terminant, on pourrait mentionner quelques détails sur l'estimation des différents modèles. D'abord, les courbes contraintes de la durée sont pratiquement plates, mais pas complètement. On agrandit le graphique en enlevant les intervalles de confiance estimés à la figure 3.13. En observant les valeurs de l'axe des ordonnées, la spline n'influence que très marginalement la moyenne de l'estimé. On peut également obtenir des segments décroissants si la différence de valeurs en ordonnées se trouve sous la marge d'erreur de l'algorithme. De plus, la dimension des splines affecte considérablement le temps d'estimation des modèles avec contraintes de monotonicité. Jusqu'à 9 noeuds, les durées d'estimation étaient raisonnables, mais pouvait se compter en minutes ou en heures selon si le modèle comprenait des contraintes.

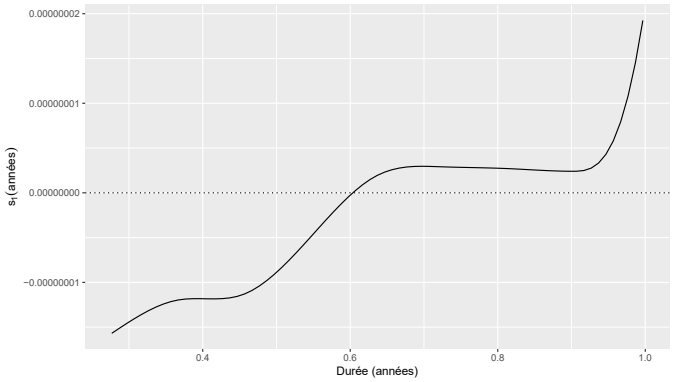

Figure 3.13 – Spline de régression cubique avec contrainte de monotonie pour la durée

3.3.2.5 Statistiques d'ajustement

Il reste à analyser la performance de ces modèles sur les données et d'évaluer le coût d'imposer des contraintes de forme. Le tableau 3.3 présente des statistiques d'ajustement sur l'ensemble de données d'estimation. Nous y avons inclus la log-vraisemblance, l'AIC et le BIC pour discuter de l'ajustement. Les EDF, ou degrés effectifs de liberté (*effective degrees of freedom*, en anglais), ont été inclus, car ils sont utilisés pour déterminer la valeur de l'AIC et du BIC. Les modèles comprenaient tous cinq termes paramétriques, soit un pour l'ordonnée à l'origine et quatre pour le petit ensemble de variables explicatives. Les différences observées entre les valeurs des EDF proviennent de la pénalité imposée aux termes non paramétriques (les splines) pendant l'estimation. On constate que les modèles sans contrainte sont toujours supérieurs à leur équivalent avec contraintes, ce qui est compréhensible puisque la procédure d'estimation cherche à maximiser la vraisemblance. Si le modèle optimal était le modèle avec la spline monotone, il serait rendu par le modèle sans contrainte. Malgré le fait que le modèle TPRS sans contrainte affiche la moins bonne log-vraisemblance parmi les modèles sans contrainte, c'est celui qui a la meilleure log-vraisemblance pour les modèles avec contraintes. D'ailleurs, l'écart entre le modèle avec et sans contraintes est moins marqué pour le modèle TPRS. Ce résultat était intuitif considérant les figures 3.7 à 3.12 où on observait que les courbes avec et sans contraintes étaient assez similaires pour la spline TPRS, contrairement aux deux autres splines. On remarque aussi que les modèles TPRS pénalisent davantage les termes non paramétriques puisque leurs EDFs sont inférieurs aux autres modèles. Cela fait en sorte que les modèles TPRS obtiennent le meilleur score BIC, qui pénalise davantage pour le nombre de paramètres. Concernant strictement les modèles avec contraintes, le modèle TPRS obtient les meilleures statistiques quel que soit le critère.

Modèles sans contrainte				
Modèle	log-vraisemblance	AIC	BIC	EDF
P-spline	-10,663.59	21,367.50	21,544.46	20.17
Spline reg. cubique	-10,663.36	21,357.11	21,490.50	15.20
TPRS	-10,683.50	21,392.38	21,503.68	12.68
Modèles avec contraintes				
Modèle	log-vraisemblance	AIC	BIC	EDF
P-spline	-10,694.21	21,422.04	21,569.53	16.81
Spline reg. cubique	-10,692.32	21,415.03	21,548.42	15.20
TPRS	-10,686.91	21,396.42	21,495.64	11.31

Table 3.3 – Statistiques d'ajustement

Le tableau 3.4 présente des statistiques d'ajustement sur l'ensemble de données test pour toutes les distributions prédictives estimées par les modèles candidats. Puisqu'il s'agit de données discrètes de comptage, et non des données continues, il faut choisir des statistiques appropriées pour ce type de données ((Czado *et al.*, 2009)). Parmi les statistiques retenues, on retrouve le score logarithmique et le score Dawid-Sebastiani (dss), car ces deux métriques permettent d'obtenir des scores sensiblement différents entre les modèles MVNB considérés afin de les départager. Le score logarithmique est l'équivalent de la log-vraisemblance calculée sur l'ensemble de données test. Le score Dawid-Sebastiani représente l'erreur quadratique qui a été normalisée par la variance, à laquelle on ajoute une pénalité basée sur celle-ci afin de ne pas avantager systématiquement les modèles estimant une plus grande variance. Pour une observation, on a dss( ) $= \frac{\ }{\ }^{2} + 2\log(\ )$ et la somme des statistiques pour chacune des observations de l'ensemble de données test permet d'évaluer le score final. Le tableau comprend également l'erreur quadratique puisqu'il s'agit d'une mesure d'erreur parmi les plus connues et utilisées. Finalement, la déviance Poisson est incluse, car l'erreur quadratique correspond à la déviance normale et qu'une distribution Poisson est mieux adaptée à des données de comptage. On peut constater que même sur l'ensemble de données test, les statistiques sont meilleures pour les modèles sans contrainte. On observe également que même parmi les modèles sans contrainte, le modèle P-spline (figure 3.7) obtient les meilleures statistiques, alors qu'il présente une décroissance dans la courbe du kilométrage. Ces résultats pointent donc dans la direction que ce type de modèles, pour être optimaux, devraient présenter une décroissance du kilométrage

lorsqu'on considère indépendamment la durée et la distance. La question devient donc, est-ce qu'on peut se permettre une légère déviation pour permettre une tarification logique. Sur l'ensemble de données test, les différences entre les modèles avec et sans contraintes sont moins importantes que sur l'ensemble de données d'ajustement. Une autre réflexion à considérer est sur l'inclusion des splines associées au kilométrage et la durée de manière indépendante. Cela ne reflète peut-être pas exactement la réalité, mais une version multivariée de l'exposition serait difficilement applicable en pratique (Boucher *et al.*, 2017). D'autres avenues en ce sens pourraient cependant être explorées, car (Boucher *et al.*, 2017) se sont limités à un produit tensoriel de P-splines.

Modèles sans contrainte				
Modèle	Logarithmique	Dawid-Sebastiani	Déviance Poisson	Erreur quadratique
P-spline	-2,645.45	-22,479.25	5,236.36	740.15
Spline reg. cubique	-2,645.51	-22,474.16	5,236.53	740.11
TPRS	-2,649.68	-22,320.39	5,245.18	740.57
Modèles avec contraintes				
Modèle	Logarithmique	Dawid-Sebastiani	Déviance Poisson	Erreur quadratique
P-spline	-2,647.52	-22,456.60	5,241.04	740.73
Spline reg. cubique	-2,647.04	-22,466.67	5,240.07	740.54
TPRS	-2,649.90	-22,321.90	5,245.67	740.68

Table 3.4 – Statistiques d'ajustement sur l'ensemble de données test

3.3.3 Analyse des primes

Cette sous-section présente des primes calculées avec les six modèles abordés précédemment. Le tableau 3.5 contient les estimations des termes paramétriques de l'ensemble des modèles. Les primes, correspondant à l'espérance de la distribution, ont été calculées en fonction d'un profil de risque moyen, soit une femme, mariée, qui utilise son véhicule pour aller travailler. Les primes ont été calculées pour différentes durées (30 %, 50%, 75% et 100% de l'année) et différentes distances parcourues (1 000, 2 000, 5 000, 10 000, 20 000, 30 000, 40 000, 50 000, 60 000 et 70 000 kilomètres). On étudie également comment évolue la prime selon différents historiques de réclamations : sans historique (ou prime *a priori*), un historique de deux ans sans réclamation (ou prime prédictive favorable) et un historique avec deux réclamations en deux ans (ou prime prédictive défavorable).

Modèle	P-spline		Spline reg. cubique		TPRS	
	Sans contrainte	Monotone	Sans contrainte	Monotone	Sans contrainte	Monotone
$_0$	-3.661 (0.261)	-3.712 (0.261)	-3.664 (0.261)	-3.713 (0.261)	-3.676 (0.261)	-3.680 (0.261)
	-0.030 (0.038)	-0.046 (0.038)	-0.030 (0.038)	-0.046 (0.038)	-0.030 (0.038)	-0.030 (0.038)
	0.141 (0.040)	0.147 (0.040)	0.141 (0.040)	0.147 (0.040)	0.142 (0.040)	0.143 (0.040)
( )	0.715 (0.261)	0.783 (0.261)	0.718 (0.261)	0.781 (0.261)	0.766 (0.261)	0.772 (0.261)
( )	0.693 (0.261)	0.768 (0.262)	0.696 (0.261)	0.768 (0.262)	0.704 (0.261)	0.709 (0.261)
	8.687 (0.669)	7.379 (0.582)	8.471 (0.653)	7.344 (0.579)	7.795 (0.608)	7.649 (0.598)

Table 3.5 – Estimations des termes paramétriques

3.3.3.1 Primes *a priori*

En annexe, les tableaux 3.6, 3.7 et 3.8 présentent les primes pour les modèles avec et sans contraintes intégrant respectivement des P-splines, des splines cubiques de régression et des TPRS. Comme il avait pu être observé aux figures 3.8 et 3.10, on remarque que la prime ne varie pas en fonction de la durée du contrat pour un même nombre de kilomètres parcourus dans les modèles avec P-splines et splines de régression cubiques monotones. Pour ces mêmes splines, mais sans contrainte, on constate des irrégularités, comme la décroissance entre les kilomètres 60 000 et 70 000, entre les durées 0.3 et 0.5 ainsi qu'entre 0.75 et 1. Pour ce qui est du modèle avec splines TPRS (figure 3.8), on obtient une structure de prime cohérente, peu importe s'il y a des contraintes de monotonicité ou non. Les deux structures sont très similaires comme le laissaient présager les figures 3.11 et 3.12.

La figure 3.14 illustre l'évolution des primes *a priori* en fonction du kilométrage pour les différentes splines sous contraintes étudiées. Pour les modèles avec P-splines et splines cubiques de régression, la durée d'exposition n'a pas d'impact sur la prime *a priori*. La durée influence toutefois la prime pour le modèle incluant des TPRS. Ainsi, une courbe pour une durée de 0.5 et une durée d'un an sont illustrées (lignes brisées). On observe que la valeur des primes des modèles avec P-splines et splines cubiques est semblable jusqu'à environ 35 000 kilomètres, ce qui représente la majorité du portefeuille (figure 3.6). Ensuite, le modèle avec P-splines est plus cher, jusqu'à environ 58 000 km où le modèle avec splines cubiques commence à augmenter plus rapidement.

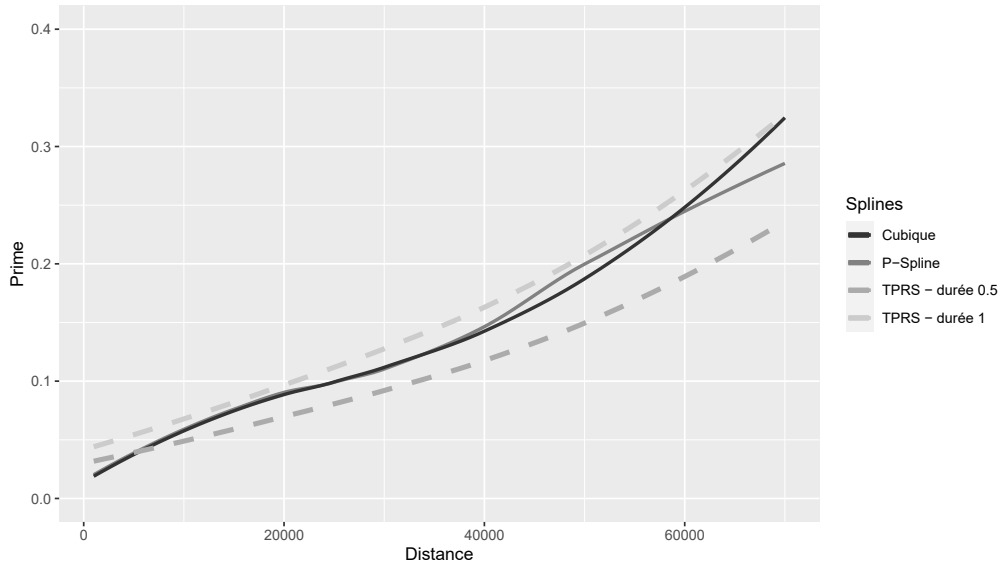

Figure 3.14 – Évolution de la prime *a priori* en fonction du kilométrage

Le niveau des primes du modèle TPRS est généralement plus élevé que pour les deux autres splines. Pour les modèles TPRS, la prime varie en fonction de la durée du contrat, mais pas de façon proportionnelle. Même en considérant que l'utilisation de la voiture constitue la principale source d'exposition au risque pour la couverture collision, il peut être logique d'augmenter la prime en fonction de la durée du contrat, car même un véhicule stationné peut être percuté ou victime d'un délit de fuite.

Il est intéressant de remarquer que l'écart entre les deux courbes de durée des splines TPRS est de plus en plus grand plus la distance augmente. Cela implique que le modèle augmente la prime plus rapidement lorsque les kilomètres sont conduits sur l'année entière plutôt que sur une portion d'année. On aurait pu s'attendre à ce que conduire un nombre important de kilomètres sur une durée moins longue ait un impact négatif (à la hausse) sur la prime. Toutefois, si un nombre important de kilomètres ont été conduits en peu de temps, on peut supposer que les autoroutes ont davantage été utilisées. Comme il a été mentionné précédemment, ces routes sont considérées comme moins propices aux collisions comparativement à la conduite en ville. Si peu de kilomètres ont été parcourus, par exemple 5 000 kilomètres, les profils de risque entre six mois et un an sont possiblement plus similaires : une proportion importante de kilomètres conduits en ville.

3.3.3.2 Primes prédictives

Les tableaux 3.9, 3.10 et 3.11 en annexe montrent comment évoluent les primes suite à deux années sans réclamation pour chacun des modèles avec et sans contraintes. La réduction de prime est plus tangible pour les distances parcourues plus élevées. En effet, il y a peu ou pas de mouvement pour les distances les plus basses. La prime *a priori* varie entre 0.035 et 0.054 pour 5 000 kilomètres, tandis qu'elle varie de 0.035 à 0.053 pour la même distance selon le scénario favorable après deux ans. Si on considérait plutôt 70 000 kilomètres, la prime *a priori* varierait alors de 0.237 à 0.328 selon la spline, mais seulement de 0.223 à 0.302 selon un scénario favorable. Si on analyse par durée, il n'y a pas vraiment de différence entre les primes *a priori* et prédictives. Considérant que le kilométrage est la principale source d'exposition au risque de collision, cela peut avoir un sens de ne pas réduire la prime seulement parce que le contrat est actif. Pour un assuré qui conduit seulement 1000 km par année, il est légitime de se demander si le risque a suffisamment été observé après deux ans pour lui accorder un rabais. Une réduction par rapport à la prime *a priori* commence à apparaître à partir d'environ 10 000 km.

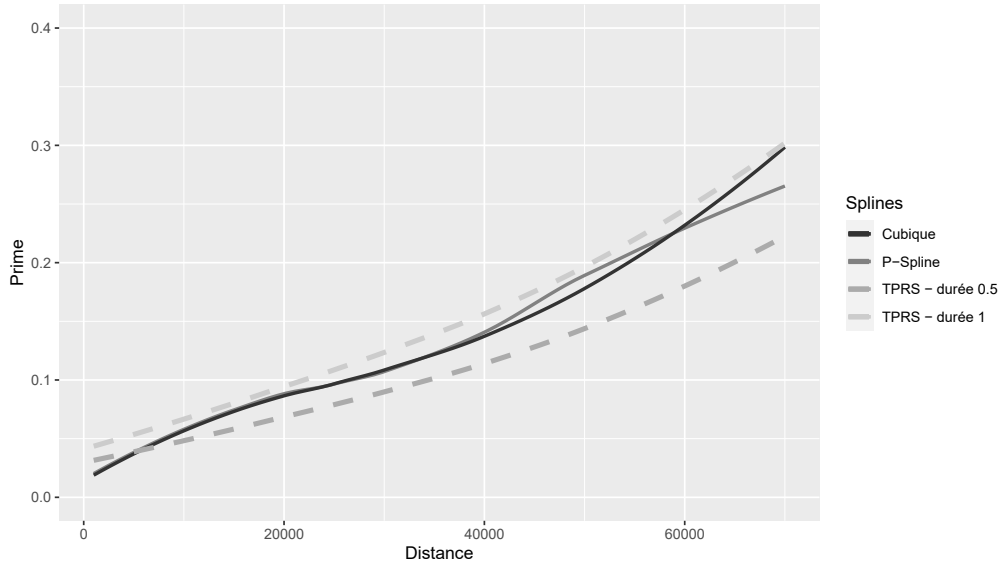

Figure 3.15 – Évolution de la prime prédictive en fonction du kilométrage (scénario favorable)

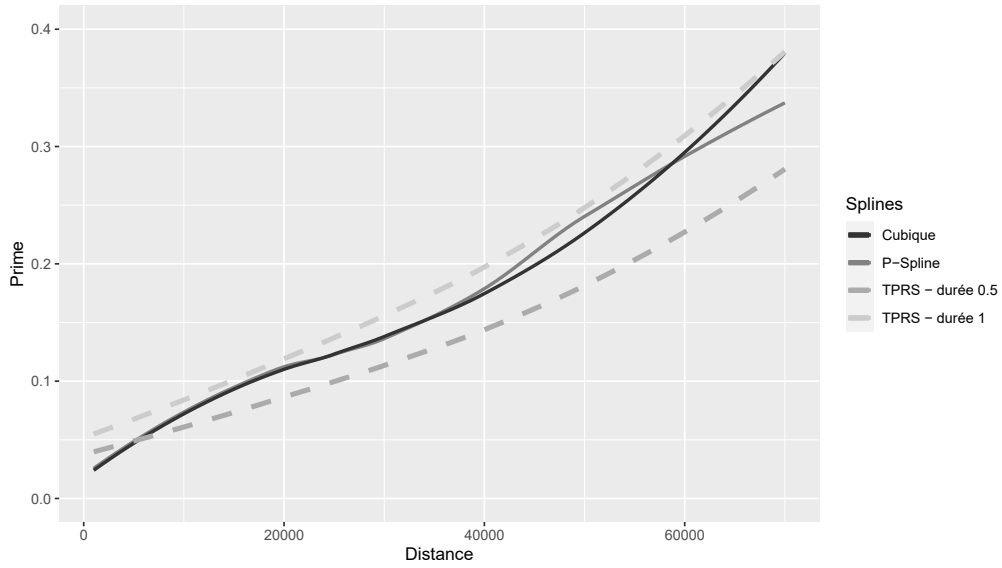

Figure 3.16 – Évolution de la prime prédictive en fonction du kilométrage (scénario défavorable)

Les tableaux 3.12 à 3.14 présentent les primes prédictives pour un historique défavorable (deux réclamations après deux années d'observation). En comparant avec les primes *a priori*, on constate que le niveau des primes est plus élevé autant pour les durées que les distances plus longues. Cette caractéristique peut être désirée dans la mesure où un assureur ne souhaite pas retenir un mauvais risque. Les figures 3.15 et 3.16 offrent un résumé visuel des primes prédictives selon les deux scénarios considérés. On observe que la forme des splines est similaire, mais que les courbes sont translatées vers le haut dans le cas d'un scénario défavorable.

Nous avons mentionné que le niveau des primes est plus élevé autant pour les durées que les distances plus longues dans le scénario défavorable, alors que le scénario favorable n'accorde une réduction que pour les distances plus longues. Ces résultats sont cohérents puisque si un conducteur est assuré pour une courte durée, le modèle prévoit une faible probabilité de réclamation. Si le conducteur n'a, en effet, pas réclamé après deux ans, le modèle ne réduit pas la prime, car c'était le résultat attendu. *A contrario*, si le conducteur réclame alors que le modèle estimait la probabilité faible, la prime augmente. En analysant l'équation 3.11 de la prime prédictive, on remarque que les réclamations passées ($=$ $_{=1}$) affecteront davantage le niveau de la prime si les espérances des années passées ($=$ $_{=1}$) sont moins élevées.

3.3.3.3 Impact des réclamations passées

Finalement, pour illustrer l'impact des réclamations passées sur la prime prédictive, les primes *a priori* et selon les scénarios favorable et défavorable sont illustrées à la figure 3.17 pour le modèle TPRS. On remarque la similitude des primes *a priori* et selon le scénario favorable dans la première portion du domaine. Considérant que la majorité du portefeuille parcourt moins de 30 000 kilomètres, la réduction accordée par le modèle est graduelle et conservatrice. On constate également que la courbe du scénario défavorable s'éloigne de plus en plus de la courbe *a priori* lorsque la distance augmente. Cette caractéristique est désirable dans la mesure où si le modèle considère un assuré comme un mauvais risque, cet assuré sera de plus en plus pénalisé au fur et à mesure qu'il parcourt plus de kilomètres, l'incitant ainsi à réduire son exposition au risque.

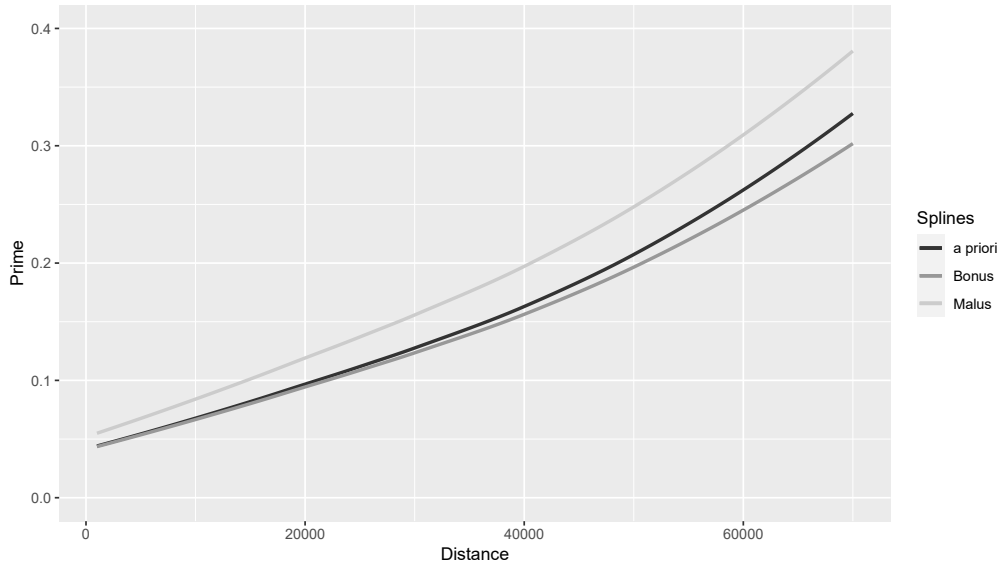

Figure 3.17 – Évolution de la prime TPRS en fonction de l'historique de réclamations

3.4 Conclusion

Dans ce travail de recherche, nous avons exploré la relation entre la distance parcourue et le risque en utilisant différentes fonctions de lissage. La mesure d'exposition au risque habituellement utilisée, la durée du contrat, a également été incluse de manière indépendante dans les modèles de tarification. Contrairement aux premiers travaux en ce sens, un plus large éventail de splines a été considéré pour analyser comparativement les résultats. En plus des B-splines/P-splines, les fonctions de lissage qui ont été utilisées sont les splines de régression cubiques et les splines de régression en plaque mince avec et sans contraintes pour

assurer la monotonicité. Nous avons présenté différentes méthodes pour modéliser une spline monotone, en plus d'explorer les caractéristiques de chacune. Nous avons conclu qu'utiliser une spline monotone s'effectuait à un certain coût en termes d'ajustement, mais que les différences étaient moins importantes sur les statistiques de l'ensemble de données test. Une limite abordée de cette étude concerne les données. Nous avions mentionné que le portefeuille était composé en plus grande partie de jeunes, une clientèle généralement moins fidèle que leurs aînés. Il est donc plus difficile de collecter un long historique pour ajuster un modèle longitudinal. Toutefois, les données obtenues ont le mérite d'être fiables puisqu'elles ont été collectées avec un appareil installé dans le véhicule plutôt que collectées par une application mobile qui peut être oubliée d'être activée. Une prochaine étape serait d'explorer la dépendance entre la durée et la distance en utilisant, par exemple, des splines multivariées. En effet, on peut raisonnablement supposer que la durée et la distance puissent être corrélées puisque la plupart des distances parcourues élevées sont associées à des durées de police plus grandes. Une autre piste pour des recherches futures serait de considérer d'autres modèles d'effets aléatoires. Dans ce travail, nous avons utilisé un effet aléatoire continu, mais il serait possible de considérer que les individus appartiennent à un groupe latent dont le nombre est fini ((Tseung *et al.*, 2023), (Chai Fung *et al.*, 2019)). Pour conclure, on peut souligner que lorsque le kilométrage mesuré par un appareil GPS est inclus dans la tarification (plutôt que l'auto-déclaration de l'assuré), le genre n'est pas une variable explicative significative pour aucun des modèles (voir tableau 3.5). Le statut marital reste cependant significatif, alors d'autres avenues que le kilométrage devraient être étudiées pour remplacer cette variable discriminante.

3.5 Annexe

3.5.1 Primes *a priori*

(a) Sans contrainte

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.026	0.021	0.031	0.029
2000	0.029	0.023	0.035	0.033
5000	0.040	0.032	0.049	0.046
10000	0.061	0.048	0.073	0.069
20000	0.080	0.064	0.097	0.091
30000	0.096	0.076	0.115	0.109
40000	0.126	0.100	0.151	0.143
50000	0.185	0.147	0.223	0.211
60000	0.231	0.183	0.278	0.263
70000	0.182	0.144	0.219	0.207

(b) Monotone

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.022	0.022	0.022	0.022
2000	0.025	0.025	0.025	0.025
5000	0.036	0.036	0.036	0.036
10000	0.057	0.057	0.057	0.057
20000	0.091	0.091	0.091	0.091
30000	0.105	0.105	0.105	0.105
40000	0.143	0.143	0.143	0.143
50000	0.206	0.206	0.206	0.206
60000	0.259	0.259	0.259	0.259
70000	0.278	0.278	0.278	0.278

Table 3.6 – Primes *a priori* - P-splines

(a) Sans contrainte

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.020	0.019	0.027	0.026
2000	0.025	0.023	0.034	0.032
5000	0.035	0.032	0.046	0.044
10000	0.052	0.048	0.069	0.065
20000	0.069	0.063	0.093	0.086
30000	0.085	0.076	0.112	0.102
40000	0.108	0.099	0.145	0.137
50000	0.140	0.129	0.189	0.177
60000	0.186	0.172	0.249	0.234
70000	0.250	0.230	0.330	0.310

(b) Monotone

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.019	0.019	0.019	0.019
2000	0.024	0.024	0.024	0.024
5000	0.035	0.035	0.035	0.035
10000	0.058	0.058	0.058	0.058
20000	0.087	0.087	0.087	0.087
30000	0.110	0.110	0.110	0.110
40000	0.142	0.142	0.142	0.142
50000	0.186	0.186	0.186	0.186
60000	0.246	0.246	0.246	0.246
70000	0.326	0.326	0.326	0.326

Table 3.7 – Primes *a priori* - Splines cubiques de régression

(a) Sans contrainte

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.027	0.032	0.042	0.044
2000	0.028	0.033	0.044	0.046
5000	0.033	0.038	0.051	0.053
10000	0.041	0.048	0.063	0.066
20000	0.058	0.069	0.091	0.095
30000	0.077	0.090	0.119	0.125
40000	0.098	0.116	0.153	0.160
50000	0.125	0.147	0.195	0.204
60000	0.159	0.187	0.247	0.259
70000	0.201	0.236	0.313	0.327

(b) Monotone

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.027	0.032	0.040	0.044
2000	0.028	0.034	0.042	0.047
5000	0.033	0.039	0.048	0.054
10000	0.041	0.049	0.060	0.067
20000	0.059	0.070	0.086	0.097
30000	0.077	0.092	0.113	0.127
40000	0.099	0.117	0.145	0.163
50000	0.125	0.149	0.184	0.207
60000	0.158	0.188	0.233	0.261
70000	0.199	0.237	0.293	0.328

Table 3.8 – Primes *a priori* - TPRS

3.5.2 Primes prédictives (scénario favorable)

(a) Sans contrainte

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.026	0.020	0.031	0.029
2000	0.029	0.023	0.035	0.033
5000	0.040	0.032	0.048	0.045
10000	0.060	0.047	0.072	0.068
20000	0.079	0.063	0.094	0.090
30000	0.094	0.075	0.112	0.106
40000	0.122	0.098	0.146	0.139
50000	0.178	0.142	0.212	0.201
60000	0.220	0.176	0.261	0.248
70000	0.175	0.140	0.208	0.198

(b) Monotone

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.022	0.022	0.022	0.022
2000	0.025	0.025	0.025	0.025
5000	0.036	0.036	0.036	0.036
10000	0.056	0.056	0.056	0.056
20000	0.089	0.089	0.089	0.089
30000	0.102	0.102	0.102	0.102
40000	0.138	0.138	0.138	0.138
50000	0.195	0.195	0.195	0.195
60000	0.242	0.242	0.242	0.242
70000	0.258	0.258	0.258	0.258

Table 3.9 – Primes prédictives (2 ans sans réclamation) - P-splines

(a) Sans contrainte

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.020	0.019	0.027	0.026
2000	0.025	0.023	0.033	0.031
5000	0.034	0.032	0.046	0.043
10000	0.052	0.047	0.068	0.064
20000	0.068	0.062	0.091	0.084
30000	0.083	0.075	0.109	0.099
40000	0.105	0.097	0.141	0.132
50000	0.136	0.126	0.181	0.170
60000	0.179	0.165	0.235	0.221
70000	0.236	0.218	0.307	0.289

(b) Monotone

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.019	0.019	0.019	0.019
2000	0.024	0.024	0.024	0.024
5000	0.035	0.035	0.035	0.035
10000	0.057	0.057	0.057	0.057
20000	0.085	0.085	0.085	0.085
30000	0.107	0.107	0.107	0.107
40000	0.137	0.137	0.137	0.137
50000	0.177	0.177	0.177	0.177
60000	0.230	0.230	0.230	0.230
70000	0.299	0.299	0.299	0.299

Table 3.10 – Primes prédictives (2 ans sans réclamation) - Splines cubiques de régression

(a) Sans contrainte

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.027	0.031	0.042	0.043
2000	0.028	0.033	0.044	0.046
5000	0.032	0.038	0.050	0.052
10000	0.040	0.047	0.062	0.065
20000	0.058	0.068	0.089	0.093
30000	0.075	0.088	0.116	0.121
40000	0.096	0.112	0.147	0.154
50000	0.121	0.142	0.186	0.194
60000	0.153	0.178	0.233	0.243
70000	0.191	0.223	0.290	0.302

(b) Monotone

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.027	0.032	0.039	0.044
2000	0.028	0.033	0.041	0.046
5000	0.032	0.038	0.047	0.053
10000	0.040	0.048	0.059	0.066
20000	0.058	0.069	0.085	0.095
30000	0.076	0.090	0.110	0.123
40000	0.096	0.114	0.140	0.156
50000	0.121	0.143	0.176	0.196
60000	0.152	0.179	0.219	0.244
70000	0.189	0.223	0.272	0.302

Table 3.11 – Primes prédictives (2 ans sans réclamation) - TPRS

3.5.3 Primes prédictives (scénario défavorable)

(a) Sans contrainte

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.032	0.025	0.038	0.036
2000	0.036	0.028	0.043	0.041
5000	0.049	0.039	0.059	0.056
10000	0.073	0.058	0.088	0.083
20000	0.097	0.077	0.116	0.110
30000	0.115	0.092	0.138	0.131
40000	0.151	0.120	0.180	0.171
50000	0.219	0.175	0.261	0.247
60000	0.270	0.216	0.321	0.305
70000	0.215	0.172	0.256	0.243

(b) Monotone

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.028	0.028	0.028	0.028
2000	0.032	0.032	0.032	0.032
5000	0.045	0.045	0.045	0.045
10000	0.071	0.071	0.071	0.071
20000	0.113	0.113	0.113	0.113
30000	0.130	0.130	0.130	0.130
40000	0.176	0.176	0.176	0.176
50000	0.248	0.248	0.248	0.248
60000	0.308	0.308	0.308	0.308
70000	0.329	0.329	0.329	0.329

Table 3.12 – Primes prédictives (deux réclamations) - P-splines

(a) Sans contrainte

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.025	0.023	0.034	0.032
2000	0.031	0.028	0.041	0.039
5000	0.043	0.039	0.057	0.053
10000	0.064	0.059	0.084	0.079
20000	0.084	0.076	0.112	0.104
30000	0.103	0.092	0.135	0.123
40000	0.130	0.120	0.174	0.164
50000	0.168	0.155	0.223	0.210
60000	0.221	0.204	0.291	0.274
70000	0.292	0.270	0.379	0.357

(b) Monotone

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.025	0.025	0.025	0.025
2000	0.031	0.031	0.031	0.031
5000	0.044	0.044	0.044	0.044
10000	0.073	0.073	0.073	0.073
20000	0.108	0.108	0.108	0.108
30000	0.136	0.136	0.136	0.136
40000	0.174	0.174	0.174	0.174
50000	0.225	0.225	0.225	0.225
60000	0.293	0.293	0.293	0.293
70000	0.381	0.381	0.381	0.381

Table 3.13 – Primes prédictives (deux réclamations) - Splines cubiques de régression

(a) Sans contrainte

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.034	0.040	0.052	0.055
2000	0.035	0.041	0.055	0.057
5000	0.041	0.048	0.063	0.066
10000	0.051	0.059	0.078	0.082
20000	0.072	0.085	0.112	0.117
30000	0.095	0.111	0.146	0.152
40000	0.121	0.141	0.185	0.193
50000	0.153	0.178	0.233	0.243
60000	0.192	0.224	0.292	0.305
70000	0.240	0.280	0.364	0.379

(b) Monotone

	Durée			
Distance	0.3	0.5	0.75	1
1000	0.034	0.040	0.050	0.055
2000	0.035	0.042	0.052	0.058
5000	0.041	0.048	0.060	0.067
10000	0.051	0.061	0.075	0.084
20000	0.073	0.087	0.107	0.119
30000	0.095	0.113	0.139	0.155
40000	0.121	0.144	0.176	0.197
50000	0.153	0.181	0.222	0.247
60000	0.192	0.226	0.277	0.308
70000	0.239	0.281	0.343	0.381

Table 3.14 – Primes prédictives (deux réclamations) - TPRS

CONCLUSION

Dans cette thèse de doctorat, différents termes non paramétriques ont été intégrés à des modèles longitudinaux de tarification en assurance automobile. Puisque les modèles transversaux ont été appliqués avec succès dans le domaine de l'assurance, il est pertinent d'étudier la modélisation de l'historique de réclamation. Il a été principalement question d'étudier la relation entre le kilométrage et le risque de perte ainsi que l'apport d'éléments longitudinaux et semi-paramétriques à la modélisation et la structure des primes. Dans le premier chapitre, des données télématiques ont été utilisées pour étudier la relation entre le kilométrage et le risque de réclamation en utilisant des modèles à effets aléatoires et à effets fixes. Les principales conclusions de ce projet ont été que le modèle à effets fixes, qui permettait de tenir compte des caractéristiques individuelles de l'assuré, ajustait une courbe pratiquement linéaire entre le kilométrage et le risque. Par contre, ce modèle ne pouvant pas être utilisé en pratique puisqu'il n'est pas possible d'évaluer le risque d'un nouvel assuré, un modèle à effets aléatoires a également été étudié. On observait toutefois une relation illogique entre le kilométrage et le risque. Puisque la spline associée au modèle à effets aléatoires ondule autour de celle du modèle à effets fixes ajustée en utilisant le paramètre individuel médian, on concluait que des caractéristiques individuelles non prises en charge venaient biaiser l'ajustement du modèle à effets aléatoires.

Dans le deuxième chapitre, un modèle de tarification a été développé afin d'évaluer l'utilité de modèles longitudinaux et semi-paramétriques pour améliorer le pouvoir prédictif. Ces modèles ont été utilisés pour mieux capturer l'information *a priori*. Par exemple, une réduction dans la valeur de l'estimateur de la variance de la distribution de l'effet aléatoire a été observée. Les espérances prédictives selon différents scénarios de réclamation ont été analysées. En réduisant l'importance de l'expérience passée dans la tarification prédictive, les modèles réussissent à mieux capturer des informations a priori. En effet, la prise en compte de l'expérience passée vise à introduire dans la tarification les éléments non mesurables dans une tarification a priori (impulsivité, imprudence, etc.) ainsi que ceux que le modèle n'a pas su capturer complètement (dû à un manque de flexibilité du modèle). On montre également les étapes pour inclure des termes non paramétriques dans la modélisation en utilisant les splines cubiques de régression et les P-splines comme exemples. Considérant que peu d'ouvrages succincts soient suffisamment détaillés pour permettre l'implémentation de ces techniques, cette section de la thèse contribue à la communauté scientifique en rendant ces notions plus accessibles.

Compte tenu des avantages d'un modèle de tarification longitudinal et semi-paramétrique, l'idée d'analyser la relation entre le kilométrage et le risque de réclamation pour le modèle à effets aléatoires a été reprise dans le troisième chapitre, mais en corrigeant les irrégularités de forme pour la tarification observées au premier chapitre. Pour corriger la décroissance observée au premier chapitre, trois approches possédant des caractéristiques différentes pour modéliser une spline monotone ont été présentées et analysées. Il a été observé que si la monotonicité était imposée à la fois sur les splines du kilométrage et de la durée, deux des trois types de splines aplanissaient la courbe associée à la durée du contrat pour la couverture collision. Considérant que les modèles intégrant ces deux splines obtenaient les meilleures statistiques d'ajustement, cela pourrait être une indication que si une seule mesure d'exposition devait être choisie pour la couverture collision, ce serait le kilométrage.

Dans la dernière décennie, l'intérêt grandissant pour les données collectées par GPS a permis de développer de nouvelles avenues pour la tarification en assurance automobile. Le développement de ces technologies permet de recueillir de nouvelles informations de manière de plus en plus fiable et de plus en plus rentable. Par contre, il est nécessaire d'en apprendre davantage sur les façons d'utiliser ces données pour la tarification dans une industrie aussi règlementée que l'assurance. Pour maintenir la stabilité financière de l'entreprise et lui assurer de remplir ses obligations envers ses assurés, il est nécessaire d'obtenir une évidence statistique convaincante et de satisfaire les régulateurs. Les changements sont donc souvent progressifs dans une industrie comme l'assurance. En ce sens, des modèles généralisant des techniques bien implémentées dans l'industrie consistent en une approche intéressante pour bonifier les modèles existants. Considérant que cette industrie utilise encore des critères discriminatoires comme le statut marital ou le niveau d'éducation, ce type de recherche est d'intérêt public pour une société qui cherche à réduire les écarts injustifiés entre les individus de différents groupes sociaux. Recueillir des informations sur ce qui se passe réellement derrière le volant et traiter mathématiquement ces informations est en ce sens fort pertinent.

BIBLIOGRAPHIE

Agence des droits fondamentaux de l'Union européenne, Cour européenne des droits de l'homme et Conseil de l'Europe. (2018). *Manuel de droit européen en matière de non-discrimination*. Office des publications de l'Union européenne.

Albrecht, P. (1985). An evolutionary credibility model for claim numbers. *ASTIN Bulletin : The Journal of the IAA*, *15*(1), 1–17.

Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E. & Thandi, N. (2007). A practitioner's guide to generalized linear models—a foundation for theory, interpretation and application.

Ayuso, M., Guillen, M. & Marín, A. M. P. (2016a). Using gps data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. *Transportation research part C : emerging technologies*, *68*, 160–167.

Ayuso, M., Guillen, M. & Nielsen, J. P. (2019). Improving automobile insurance ratemaking using telematics : incorporating mileage and driver behaviour data. *Transportation*, *46*(3), 735–752.

Ayuso, M., Guillén, M. & Pérez-Marín, A. M. (2014). Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. *Accident Analysis & Prevention*, *73*, 125–131.

Ayuso, M., Guillen, M. & Pérez-Marín, A. M. (2016b). Telematics and gender discrimination : some usage-based evidence on whether men's risk of accidents differs from women's. *Risks*, *4*(2), 10.

Bolderdijk, J. W., Knockaert, J., Steg, E. & Verhoef, E. T. (2011). Effects of pay-as-you-drive vehicle insurance on young drivers' speed choice : Results of a dutch field experiment. *Accident Analysis & Prevention*, *43*(3), 1181–1186.

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

Boucher, J.-P. & 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. & Guillen, 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. & Inoussa, R. (2014). A posteriori ratemaking with panel data. *ASTIN Bulletin : The Journal of the IAA*, *44*(3), 587–612.

Boucher, J.-P., Pérez-Marín, A. M. & 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.

Boucher, J.-P. & Turcotte, R. (2020). A longitudinal analysis of the impact of distance driven on the probability of car accidents. *Risks*, *8*(3), 91.

Brunk, H., Barlow, R. E., Bartholomew, D. J. & Bremner, J. M. (1972). *Statistical inference under order restrictions.(the theory and application of isotonic regression)*. Rapport technique, Missouri Univ Columbia Dept of Statistics.

Bühlmann, H. (1967). Experience rating and credibility. *ASTIN Bulletin : The Journal of the IAA*, *4*(3), 199–207.

Cameron, A. C. & Trivedi, P. K. (2013). *Regression analysis of count data*, volume 53. Cambridge university press.

Chai Fung, T., Badescu, A. L. & Sheldon Lin, X. (2019). A class of mixture of experts models for general insurance : Application to correlated claim frequencies. *ASTIN Bulletin : The Journal of the IAA*, *49*(3), 647–688. http://dx.doi.org/10.1017/asb.2019.25

Cole, T. J. & Green, P. J. (1992). Smoothing reference centile curves : the lms method and penalized likelihood. *Statistics in medicine*, *11*(10), 1305–1319.

Czado, C., Gneiting, T. & Held, L. (2009). Predictive model assessment for count data. *Biometrics*, *65*(4), 1254–1261.

De Bastiani, F., Rigby, R. A., Stasinopoulous, D. M., Cysneiros, A. H. & Uribe-Opazo, M. A. (2018). Gaussian markov random field spatial models in gamlss. *Journal of Applied Statistics*, *45*(1), 168–186.

De Boor, C. (1978). *A practical guide to splines*, volume 27. springer-verlag New York.

Delong, Ł., Lindholm, M. & Wüthrich, M. V. (2021). Making tweedie's compound poisson model more accessible. *European Actuarial Journal*, 1–42.

Denuit, M., Guillen, M. & Trufin, J. (2019). Multivariate credibility modelling for usage-based motor insurance pricing with behavioural data. *Annals of Actuarial Science*, *13*(2), 378–399.

Denuit, M. & Lambert, P. (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis*, *93*, 40–57.

Denuit, M. & Lang, S. (2004). Non-life rate-making with bayesian gams. *Insurance : Mathematics and Economics*, *35*(3), 627–647.

Denuit, M., Maréchal, X., Pitrebois, S. & Walhin, J.-F. (2007). *Actuarial modelling of claim counts : Risk classification, credibility and bonus-malus systems*. John Wiley & Sons.

Duchon, J. (1977). *Splines minimizing rotation-invariant semi-norms in Sobolev spaces*.

Duval, F., Boucher, J.-P. & Pigeon, M. (2022). How much telematics information do insurers need for claim classification ? *North American Actuarial Journal*, *26*(4), 570–590.

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

Eling, M. & Kraft, M. (2020). The impact of telematics on the insurability of risks. *The Journal of Risk Finance*, *21*(2), 77–109.

Fahrmeir, L., Lang, S. & Spies, F. (2003). Generalized geoadditive models for insurance claims data. *Blätter der DGVFM*, *26*(1), 7–23.

Ferreira, J. & Minikel, E. (2010). Pay-as-you-drive auto insurance in massachusetts : a risk assessment and report on consumer, industry and environmental benefits. *Conservation Law Foundation, Boston*.

Frees, E. W. & Valdez, E. A. (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association*, *103*(484), 1457–1469.

Frees, E. W. & Wang, P. (2006). Copula credibility for aggregate loss models. *Insurance : Mathematics and Economics*, *38*(2), 360–373.

Friedman, J. & Tibshirani, R. (1984). The monotone smoothing of scatterplots. *Technometrics*, *26*(3), 243–250.

Fritsch, F. N. & Carlson, R. E. (1980). Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, *17*(2), 238–246.

Gao, G., Meng, S. & Wüthrich, M. V. (2019). Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal*, *2019*(2), 143–162.

Gao, G. & Wüthrich, M. V. (2018). Feature extraction from telematics car driving heatmaps. *European Actuarial Journal*, *8*(2), 383–406.

Ghosh, D. (2007). Incorporating monotonicity into the evaluation of a biomarker. *Biostatistics*, *8*(2), 402–413.

Gilchrist, R., Kamara, A. & Rudge, J. (2009). An insurance type model for the health cost of cold housing : an application of gamlss. *REVSTAT–Statistical Journal*, *7*(1), 55–66.

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

Guillen, M., Nielsen, J. P. & Pérez-Marín, A. M. (2021). Near-miss telematics in motor insurance. *Journal of Risk and Insurance*, *88*(3), 569–589.

Guillen, M., Nielsen, J. P., Pérez-Marín, A. M. & Elpidorou, V. (2020). Can automobile insurance telematics predict the risk of near-miss events ? *North American Actuarial Journal*, *24*(1), 141–152.

Hastie, T. & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*(3), 297–310.

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

He, X. & Shi, P. (1998). Monotone b-spline smoothing. *Journal of the American statistical Association*, *93*(442), 643–650.

Heller, G., Stasinopoulos, M., Rigby, B. *et al*. (2006). The zero-adjusted inverse gaussian distribution as a model for insurance claims. Dans *Proceedings of the 21th International Workshop on Statistical Modelling*, volume 226233. Galway.

Heller, G. Z., Mikis Stasinopoulos, D., Rigby, R. A. & De Jong, P. (2007). Mean and dispersion modelling for policy claims costs. *Scandinavian Actuarial Journal*, *2007*(4), 281–292.

Henckaerts, R., Antonio, K., Clijsters, M. & Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal*, *2018*(8), 681–705.

Huang, Y. & Meng, S. (2019). Automobile insurance classification ratemaking based on telematics driving data. *Decision Support Systems*, *127*, 113156.

Hyman, J. M. (1983). Accurate monotonicity preserving cubic interpolation. *SIAM Journal on Scientific and Statistical Computing*, *4*(4), 645–654.

Inouye, D. I., Yang, E., Allen, G. I. & Ravikumar, P. (2017). A review of multivariate distributions for count data derived from the poisson distribution. *Wiley Interdisciplinary Reviews : Computational Statistics*, *9*(3), e1398.

Jeong, H. (2020). Testing for random effects in compound risk models via bregman divergence. *ASTIN Bulletin : The Journal of the IAA*, *50*(3), 777–798.

Jeong, H. (2022). Dimension reduction techniques for summarized telematics data. *The Journal of Risk Management*, *Forthcoming*.

Kelly, C. & Rice, J. (1990). Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics*, 1071–1085.

Klein, N., Denuit, M., Lang, S. & Kneib, T. (2014). Nonlife ratemaking and risk management with bayesian generalized additive models for location, scale, and shape. *Insurance : Mathematics and Economics*, *55*, 225–249.

Kruskal, J. B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society : Series B (Methodological)*, *27*(2), 251–263.

Lemaire, J. (1998). Bonus-malus systems : The european and asian approach to merit-rating. *North American Actuarial Journal*, *2*(1), 26–38.

Lemaire, J. (2012). *Bonus-malus systems in automobile insurance*, volume 19. Springer science & business media.

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

Li, J. & Tan, S. (2015). Nonstationary flood frequency analysis for annual flood peak series, adopting climate indices and check dam index as covariates. *Water Resources Management*, *29*(15), 5533–5550.

Lu, M., Zhang, Y. & Huang, J. (2007). Estimation of the mean function with panel count data using monotone polynomial splines. *Biometrika*, *94*(3), 705–718.

Ma, Y.-L., Zhu, X., Hu, X. & Chiu, Y.-C. (2018). The use of context-sensitive insurance telematics data in auto insurance rate making. *Transportation Research Part A : Policy and Practice*, *113*, 243–258.

Mammen, E. (1991). Estimating a smooth monotone regression function. *The Annals of Statistics*, 724–740.

Meyer, M. C. (2008). Inference using shape-restricted regression splines. *The Annals of Applied Statistics*, *2*(3), 1013–1033.

Molenberghs, G. & Verbeke, G. (2006). *Models for discrete longitudinal data*. Springer Science & Business Media.

Mukerjee, H. (1988). Monotone nonparametric regression. *The Annals of Statistics*, 741–750.

Nelder, J. A. & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society : Series A (General)*, *135*(3), 370–384.

Pechon, F., Trufin, J., Denuit, M. *et al.* (2018). Multivariate modelling of household claim frequencies in motor third-party liability insurance. *Astin Bulletin*, *48*(3), 969–993.

Pya, N. & Wood, S. N. (2015). Shape constrained additive models. *Statistics and computing*, *25*, 543–559.

Ramires, T. G., Nakamura, L. R., Righetto, A. J., Konrath, A. C. & Pereira, C. A. (2021). Incorporating clustering techniques into gamlss. *Stats*, *4*(4), 916–930.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical science*, 425–441.

Ramsay, J. O. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, *60*(2), 365–375.

Rigby, R. A. & Stasinopoulos, D. (1996). A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, *6*(1), 57–65.

Rigby, R. A. & 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.

Rousson, V. (2008). Monotone fitting for developmental variables. *Journal of Applied Statistics*, *35*(6), 659–670.

Schelldorfer, J. & Wuthrich, M. V. (2019). Nesting classical actuarial models into neural networks. *Available at SSRN 3320525*.

Schumaker, L. (2007). *Spline functions : basic theory*. Cambridge university press.

Shang, H. L. (2019). Dynamic principal component regression : Application to age-specific mortality forecasting. *ASTIN Bulletin : The Journal of the IAA*, *49*(3), 619–645.

Shi, P. & Lee, G. Y. (2022). Copula regression for compound distributions with endogenous covariates with applications in insurance deductible pricing. *Journal of the American Statistical Association*, *117*(539), 1094–1109.

Shi, P. & Valdez, E. A. (2014). Longitudinal modeling of insurance claim counts using jitters. *Scandinavian Actuarial Journal*, *2014*(2), 159–179.

Spedicato, G. A., Clemente, A. G. P. & Schewe, F. (2014). The use of gamlss in assessing the distribution of unpaid claims reserves. Dans *Casualty Actuarial Society E-Forum, Summer 2014-Volume 2*.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V. & De Bastiani, F. (2017). *Flexible regression and smoothing : using GAMLSS in R*. CRC Press.

Tang, K. H., Dodd, E. & Forster, J. J. (2022). Joint modelling of male and female mortality rates using adaptive p-splines. *Annals of Actuarial Science*, *16*(1), 119–135.

Tremblay, L. (1992). Using the poisson inverse gaussian in bonus-malus systems. *ASTIN Bulletin : The Journal of the IAA*, *22*(1), 97–106.

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

Tseung, S. C., Chan, I. W., Fung, T. C., Badescu, A. L. & Lin, X. S. (2023). Improving risk classification and ratemaking using mixture-of-experts models with random effects. *Journal of Risk and Insurance*.

Turcotte, R. & Boucher, J.-P. (2023). Gamlss for longitudinal multivariate claim count models. *North American Actuarial Journal*, 1–24.

Tzougas, G. (2020). Em estimation for the poisson-inverse gamma regression model with varying dispersion : An application to insurance ratemaking. *Risks*, *8*(3), 97.

Tzougas, G. & Frangos, N. (2014). The design of an optimal bonus-malus system based on the sichel distribution. In *Modern Problems in Insurance Mathematics* 239–260. Springer.

Tzougas, G. & Jeong, H. (2021). An expectation-maximization algorithm for the exponential-generalized inverse gaussian regression model with varying dispersion and shape for modelling the aggregate claim amount. *Risks*, *9*(1), 19.

Tzougas, G., Vrontos, S. D. & Frangos, N. E. (2015). Risk classification for claim counts and losses using regression models for location, scale and shape. *Variance*, *9*(1), 140–157.

Verbelen, R., Antonio, K. & Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, *67*(5), 1275–1304.

Verschuren, R. M. (2021). Predictive claim scores for dynamic multi-product risk classification in insurance. *ASTIN Bulletin : The Journal of the IAA*, *51*(1), 1–25.

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

Weidner, W., Transchel, F. W. & Weidner, R. (2016). Classification of scale-sensitive telematic observables for riskindividual pricing. *European Actuarial Journal*, *6*(1), 3–24.

Wood, S. N. (1994). Monotonic smoothing splines fitted by cross validation. *SIAM Journal on Scientific Computing*, *15*(5), 1126–1133.

Wood, S. N. (2017). *Generalized additive models : an introduction with R*. Chapman and Hall/CRC.

Wüthrich, M. V. (2017). Covariate selection from telematics car driving data. *European Actuarial Journal*, *7*(1), 89–108.

Wüthrich, M. V. & Merz, M. (2019). Yes, we cann ! *ASTIN Bulletin : The Journal of the IAA*, *49*(1), 1–3.

Wüthrich, M. V. & Merz, M. (2023). *Statistical foundations of actuarial learning and its applications*. Springer Nature.

Yan, T. & Jeong, H. (2022). Posterior ratemaking of compound loss using longitudinal data with em algorithm. *Available at SSRN 4000723*.

Yan, Z. (1987). Piecewise cubic curve-fitting algorithm. *mathematics of computation*, *49*(179), 203–213.

Yang, L. & Shi, P. (2019). Multiperil rate making for property insurance using longitudinal data. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, *182*(2), 647–668.

Zhang, J.-T. (2004). A simple and efficient monotone smoother using smoothing splines. *Journal of Nonparametric Statistics*, *16*(5), 779–796.