GAMLSS FOR LONGITUDINAL MULTIVARIATE CLAIM COUNT MODELS

A PREPRINT

Roxane Turcotte,
Chaire Co-operators en analyse des risques actuariels
Departement de mathematiques
Universite du Quebec a Montreal
turcotte.roxane@courrier.uqam.ca

Jean-Philippe Boucher
Chaire Co-operators en analyse des risques actuariels
Departement de mathematiques
Universite du Quebec a Montreal
boucher.jean-philippe@uqam.ca

November 21, 2022

ABSTRACT

By generalising GAMs, GAMLSSs allow parametric or semi-parametric modelling of one or more parameters of distributions which are not members of the linear exponential family. Consequently, these GAMLSS approaches offer an interesting theoretical framework to allow the use of several potentially helpful distributions in actuarial science. GAMLSS theory is coupled with longitudinal approaches for counting data because these approaches are essential to predictive pricing models. Indeed, they are mainly known for modelling the dependence between the number of claims from the contracts of the same insured over time. Considering that the models' cross-sectional counterparts have been successfully applied in actuarial work and the importance of longitudinal models, we show that the proposed approach allows to quickly implement multivariate longitudinal models with non-parametric terms. This semi-parametric modelling is illustrated using a dataset from a major insurance company in Canada. An analysis is then conducted on the improvement of predictive power that the use of historical data and non-parametric terms in the modelling allows. Our approach differs from previous studies by the fact that it does not use any simplifying assumptions as to the value of the a priori explanatory variables and that we have carried out a predictive pricing integrating non-parametric terms within the framework of the GAMLSS in an explicit way, which makes it possible to reproduce the same type of study using other distributions.

Keywords Predictive Ratemaking, Longitudinal Data, Claim Count, Generalised Additive Models (GAM), Generalised Additive Models for Location, Scale and Shape (GAMLSS)

1 Introduction

Insurance policies cover multiple drivers and vehicles over several annual contracts. In predictive modelling 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)). 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 and Inoussa (2014), Verschuren (2021)) as storage capacity had increased.

Time-dependence has been extensively studied. Frees and Wang (2006) have notably included time dependence by using latent variables in the count regression and elliptical copulas in the severity model. In Frees and 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 and 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 does not decrease significantly over the observation period. The discrete nature of frequency data can become inconvenient as it limits the use of copulas. In Shi and 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 and 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 modelled 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.

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 modelling the location parameter of each distribution using covariates and smoothing functions in a Generalised additive models 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 and Wedderburn (1972)) proposed by Hastie and 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 and 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 and Meng (2019)), the study of risk exposure (Boucher et al. (2017), Boucher and 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 and Stasinopoulos (2005), where any parametric distribution can be used. More details on this theory are given in Stasinopoulos et al. (2017). For insurance modelling, this step forward is worthwhile because it allows for more flexible modelling 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 modelling of scaling and shape parameters, as illustrated in Heller et al. (2007) and Tzougas and 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 and Turcotte (2020), which is limited to the study of exposure measures. In this 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 and 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 NB-Beta, 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 modelling 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 behavior. 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 and 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 generalize 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, 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 3 explains the key elements of splines, penalised regression and GAMLSS in order to include non-parametric terms in the location parameter modelling. The database used for the numerical application is presented in Section 4.1 and Section 4 discusses the improvement found by moving from parametric models to semi-parametric longitudinal models and the advantages of our proposed approach. Section 5 concludes the paper.

2 Parametric modelling

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

$$\mathrm{E}\left[N_{i,T+1}|\boldsymbol{n}_{i,(1:T)}, \boldsymbol{x}_{i,(1:T+1)}\right],$$

where :

- $\boldsymbol{n}_{i,(1:T)}$ contains all past number of claims between time 1 and time T for insured i;
- $\boldsymbol{x}_{i,(1:T+1)}$ contains all covariates used in the ratemaking, from contract 1 to $T+1$. 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 T annual contracts of the same policyholder i (subsection 2.1);

- panel data models, for which we suppose dependence between all contracts written for the same driver i (subsection 2.2).

Going through those models will be an opportunity to establish the notation. Additionally, all models presented in this section will be adjusted in their parametric and semi-parametric form for comparison in Section 4.

2.1 Cross-section Data Models

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

$$\Pr\big(N_{i,T} = n | \boldsymbol{n}_{i,(1:T-1)}, \boldsymbol{x}_{i,(1:T)}\big) = \Pr(N_{i,T} = n | \boldsymbol{x}_{i,T}) \, .$$

Poisson. To model the number of claims of insured i, for contract t, the Poisson distribution of mean $\lambda_{i,t}$ is usually the starting point. It has a probability mass function defined as:

$$\Pr(N_{i,t} = n_{i,t} | \boldsymbol{x}_{i,t}) = \frac{\exp(-\lambda_{i,t})\lambda_{i,t}^{n_{i,t}}}{n_{i,t}!}, \text{ with } \lambda_{i,t} = \exp(\boldsymbol{x}_{i,t}\boldsymbol{\beta}) = g^{-1}(\eta_{i,t}). \tag{2.1}$$

The mean parameter is expressed using a log link function g and a linear combination of a vector of covariates $\boldsymbol{x}_{i,t}$ and a vector of parameters $\boldsymbol{\beta}$. The Poisson distribution is part of the linear exponential family of distributions and it is therefore straightforward to estimate this model with GLM theory.

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

$$\Pr(N_{i,t} = n_{i,t} | \boldsymbol{x}_{i,t}, \theta) = \frac{\Gamma(n_{i,t} + r_{i,t})}{\Gamma(r_{i,t})\Gamma(n_{i,t})}\theta^{n_{i,t}}(1-\theta)^{r_{i,t}}, \text{ with } r_{i,t} = \exp(\boldsymbol{x}_{i,t}\boldsymbol{\beta}) = g^{-1}(\eta_{i,t}). \tag{2.2}$$

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 3.3 that GAMLSS theory can be used instead.

2.2 Panel Data Models

Panel data models assume all annual contracts belonging to the same driver i to be dependent for $t = 1, \dots, T$. In a parametric approach, the joint distribution of the random vector $N_{i,1}, \dots, N_{i,T}$ is needed. 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 α denotes this heterogeneity factor, the joint distribution can be expressed as:

$$\Pr[N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T}] = \int_0^\infty \left(\prod_{t=1}^{T} \Pr[N_{i,t} = n_{i,t} | \boldsymbol{x}_{i,1}, ..., \boldsymbol{x}_{i,T}, \alpha_i] \right) f(\alpha_i) d\alpha_i. \tag{2.3}$$

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

2.2.1 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 $T+1$, conditional on past experience of contracts from time $t = 1, \ldots, T$. In a random effects model as defined by Equation (2.3), the predictive expectation is

$$E[N_{i,T+1} = n_{i,T+1}|\boldsymbol{n}_{i,(1:T)}] = E[E[N_{i,T+1} = n_{i,T+1}|\boldsymbol{n}_{i,(1:T)}, \alpha]]. \tag{2.4}$$

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, the 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.

2.2.2 Some Count Distributions for Longitudinal Data

Multivariate negative binomial. If we suppose $N_{i,t}|\alpha \sim \text{Poisson}(\lambda_{i,t}\alpha)$, with a heterogenity factor α that follows a gamma distribution of mean 1 and variance $\frac{1}{\nu}$, the expected value of $E[N_{i,t}]$ is unchanged and the joint distribution can be expressed as:

$$\Pr[N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T}] \quad = \quad \left(\prod_{t=1}^{T} \frac{\lambda_{i,t}^{n_{i,t}}}{n_{i,t}!}\right) \frac{\Gamma(n_{i,\bullet} + \nu)}{\Gamma(\nu)} \left(\frac{\nu}{\lambda_{i,\bullet} + \nu}\right)^{\nu} (\lambda_{i,\bullet} + \nu)^{-n_{i,\bullet}}, \tag{2.5}$$

where $n_{i,\bullet} = \sum_{t=1}^{T} n_{i,t}$ and $\lambda_{i,\bullet} = \sum_{t=1}^{T} \lambda_{i,t}$. The parameter $\lambda_{i,t}$ could be modelled with regressors

$$\lambda_{i,t} = \exp(\boldsymbol{x}_{i,t}\boldsymbol{\beta}) = g^{-1}(\eta_{i,t}). \tag{2.6}$$

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 modelling with overdispersion ($\mathbb{E}[N_{i,t}] = \lambda_{i,t} < \mathbb{V}[N_{i,t}] = \lambda_{i,t} + (\lambda_{i,t})^2/\nu$). 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:

$$\Pr[N_{i,T+1} = n_{i,T+1}|\boldsymbol{n}_{i,(1:T)}] = \frac{\Gamma(n_{i,\bullet} + n_{i,T+1} + \nu)}{\Gamma(n_{i,\bullet} + \nu) n_{i,T+1}!} \left(\frac{\lambda_{i,T+1}}{\lambda_{i,\bullet} + \lambda_{i,T+1} + \nu}\right)^{n_{i,T+1}} \left(\frac{\lambda_{i,\bullet} + \nu}{\lambda_{i,\bullet} + \lambda_{i,T+1} + \nu}\right)^{n_{i,\bullet}+\nu}.$$

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

$$E[N_{i,T+1} = n_{i,T+1}|\boldsymbol{n}_{i,(1:T)}] = \lambda_{i,T+1}\left(\frac{n_{i,\bullet}+\nu}{\lambda_{i,\bullet}+\nu}\right), \tag{2.7}$$

where $\left(\frac{n_{i,\bullet}+\nu}{\lambda_{i,\bullet}+\nu}\right)$ acts as a correction factor between what actually happened and what the model predicted for the past contracts $t = 1, \ldots, T$. Because of $\lambda_{i,\bullet}$, it is not possible to simply view the longitudinal approach as a product of univariate distribution. In Tzougas and Frangos (2014), they got around this difficulty by assuming that $\lambda_{i,t}$ was constant for all t and thus, $\lambda_{i,\bullet} = T_i \times \lambda_i$. However, we do work with $\lambda_{i,\bullet} = \sum_{t=1}^{T} \lambda_{i,t}$ and the explanatory variables can evolve over time.

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 θ that follows a beta distribution such as $g(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$, the joint distribution can be expressed as

$$\Pr[N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T}] = \left(\prod_{t=1}^{T}\frac{\Gamma(n_{i,t}+r_{i,t})}{\Gamma(r_{i,t})n_{i,t}!}\right)\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(r_{i,\bullet}+\beta-1)\Gamma(n_{i,\bullet}+\alpha-1)}{\Gamma(r_{i,\bullet}+\beta+n_{i,\bullet}+\alpha-2)}, \tag{2.8}$$

where $n_{i,\bullet} = \sum_{t=1}^{T} n_{i,t}$ and $r_{i,\bullet} = \sum_{t=1}^{T} r_{i,t}$. The parameter $r_{i,t}$ could be modelled with regressors

$$r_{i,t} = \exp(\boldsymbol{x}_{i,t}\boldsymbol{\beta}) = g^{-1}(\eta_{i,t}). \tag{2.9}$$

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

$$\Pr[N_{i,T+1} = n_{i,T+1}|\boldsymbol{n}_{i,(1:T)}] = \frac{\Gamma(n_{i,T+1}+r_{i,T+1})}{\Gamma(r_{i,T+1})n_{i,T+1}!}\frac{\mathrm{B}(r_{i,T+1}+r_{i,\bullet}+\beta-1, n_{i,T+1}+n_{i,\bullet}+\alpha-1)}{\mathrm{B}(r_{i,\bullet}+\beta-1, n_{i,\bullet}+\alpha-1)},$$

where $\mathrm{B}(\cdot,\cdot)$ is the Beta function, with $\mathrm{B}(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\partial t$. The predicted expected value is then given by:

$$E[N_{i,T+1} = n_{i,T+1}|\boldsymbol{n}_{i,(1:T)}] = \left(\frac{n_{i,\bullet}+\alpha-1}{r_{i,\bullet}+\beta-2}\right)r_{i,T+1}. \tag{2.10}$$

In comparison with the *a priori* expected value, $\left(\frac{n_{i,\bullet}+\alpha-1}{r_{i,\bullet}+\beta-2}\right)$ acts as a correction factor between what actually happened and what the model predicted for past contracts, up to a multiplicative factor $\left(\frac{\beta-1}{\alpha}\right)$.

3 Semi-parametric modelling

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 and Tibshirani (1986). GAMs introduce more flexibility in modelling, 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 $\boldsymbol{x}_{i,t}$ and $\lambda_{i,t}$ or $r_{i,t}$ (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 3.2. 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 modelling 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.

3.1 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 and 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 and Frangos (2014) used GAMLSS with the Poisson-Gamma distribution, the Poisson Inverse Gausian 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 and Tan (2015) that used this framework to include covariates such as climate indices into the modelling 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 and 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 modelling 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.

3.2 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.

3.2.1 Splines

Splines are smoothing functions defined piecewise by polynomials; they could be expressed under the form of Equation (3.1). By definition, the smoothing functions used in the context of GAMs can be rewritten as a linear combination of basis functions ($b_i(x)$) and parameters (β_i),

$$f(x) = \sum_{i=1}^{q} b_i(x)\beta_i. \tag{3.1}$$

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 $\boldsymbol{x}_{i,t}$ and $\lambda_{i,t}$ or $r_{i,t}$. First, we illustrate what basis functions are using the cubic regression splines.

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 together 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 $n-1$ polynomial pieces to fit a sample of n data, the natural cubic spline minimises $\int_{x_1}^{x_n} f''(x)^2$, where x_1 and x_n are the extremums of the data sample. The function $f''(x)^2$ 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 $f''(x_1) = f''(x_n) = 0$.

In a smoothing context, rather than interpolation, we will choose q knots, defining $q-1$ intervals over the span of the data, where $q < n$. Those knots are noted as k_j for $j = 1, \ldots, q$. 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 $f(x)$, defined at point x, is expressed as:

$$f(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\boldsymbol{F}_j\boldsymbol{\beta} + c_j^+(x)\boldsymbol{F}_{j+1}\boldsymbol{\beta} = \sum_{i=1}^{q} b_i(x)\beta_i, \tag{3.2}$$

where

$$b_i(x) = \begin{cases} c_j^-(x)F_{j,i} + c_j^+(x)F_{j+1,i} + a_j^+(x) & \text{if } i = j+1, \\ c_j^-(x)F_{j,i} + c_j^+(x)F_{j+1,i} + a_j^-(x) & \text{if } i = j, \\ c_j^-(x)F_{j,i} + c_j^+(x)F_{j+1,i} & \text{otherwise.} \end{cases} \tag{3.3}$$

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

$$a_j^-(x) = \frac{k_{j+1} - x}{k_{j+1} - k_j}, \qquad c_j^-(x) = \frac{1}{6}\left[\frac{(k_{j+1} - x)^3}{k_{j+1} - k_j} - (k_{j+1} - k_j)(k_{j+1} - x)\right],$$
$$a_j^+(x) = \frac{x - k_j}{k_{j+1} - k_j}, \qquad c_j^+(x) = \frac{1}{6}\left[\frac{(x - k_j)^3}{k_{j+1} - k_j} - (k_{j+1} - k_j)(x - k_j)\right].$$

We also define $\boldsymbol{F}_j$, the j^{th} row of the matrix $\boldsymbol{F} = \begin{bmatrix} 0 \\ F^- \\ 0 \end{bmatrix}$, where $\boldsymbol{F^-} = \boldsymbol{B}^{-1}\boldsymbol{D}$. Finally, non-zero elements of matrices $\boldsymbol{D}_{(q-2)\times(q)}$ and $\boldsymbol{B}_{(q-2)\times(q-2)}$ are defined in Table 3.1.

$j \in \{1, ..., q-2\}$	$j \in \{1, ..., q-3\}$
$B_{j,j} = \frac{1}{3}\left(\frac{1}{k_{j+1}-k_j} + \frac{1}{k_{j+2}-k_{j+1}}\right)$	$B_{j+1,j} = B_{j,j+1} = \frac{1}{6(k_{j+2}-k_{j+1})}$
$D_{j,j} = \frac{1}{k_{j+1}-k_j}$	
$D_{j,j+1} = \frac{-1}{k_{j+1}-k_j} - \frac{-1}{k_{j+2}-k_{j+1}}$	
$D_{j,j+2} = \frac{1}{k_{j+2}-k_{j+1}}$	

Table 3.1: Non-zero elements of matrices $\boldsymbol{B}$ and $\boldsymbol{D}$

B-splines. Another way to express splines is to use B-splines basis, which are strictly local unlike the basis functions given by Equation (3.3). This means that each basis function is non-zero only over the intervals between $m+3$ adjoining knots, with $m+1$ being the order of the basis. For example, $m=2$ is for a cubic spline. B-splines with $m=2$ 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 and 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 offers 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 $B_i^m(x)$:

$$f(x) = \sum_{i=1}^{q} B_i^m(x)\beta_i, \tag{3.4}$$

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

$$B_i^m(x) = \frac{x-k_i}{k_{i+m+1}-k_i}B_i^{m-1}(x) + \frac{k_{i+m+2}-x}{k_{i+m+2}-k_{i+1}}B_{i+1}^{m-1}(x), \quad \text{where} \tag{3.5}$$

$$B_i^{-1}(x) = \begin{cases} 1 & \text{if } k_i \le x \le k_{i+1} \\ 0 & \text{otherwise.} \end{cases}$$

Vector of covariates. It is possible to use classical regression methods to assess the model that replaces continuous covariate x with a spline $f(x)$ by replacing, inside the linear predictor, the vector $\boldsymbol{X} = (x_1, x_2, \dots x_n)$ with the matrix formed by the basis functions

$$\boldsymbol{X}^S = \begin{pmatrix} b_1(x_1) & b_2(x_1) & \dots & b_q(x_1) \\ b_1(x_2) & b_2(x_2) & \dots & b_q(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ b_1(x_n) & b_2(x_n) & \dots & b_q(x_n) \end{pmatrix}.$$

An identifiability problem might arises because an intercept β_0 is nearly always included in the predictor. Consequently, $\boldsymbol{X}^S$ could not be included directly like it is in the design matrix of a regression model that includes multiple covariates and an intercept. Let $\tilde{X}^S_{(n)\times(\tilde{q})}$ be the dimension-reduced matrix, where $\tilde{q} = q - 1$. $\tilde{X}^S$ could be obtained by substracting the column mean of X^S from each column of the matrix X^S. Put differently, we have

$$\tilde{X}^S = X^S - \mathbf{1}\mathbf{1}^T X^S/n. \tag{3.6}$$

This equation follows from the constraint $\sum_{i=1}^{n} f(x_i) = 0$, which changes neither the shape of the splines nor the value of the penalty hyperparameter (see Section 3.2.3).

It is worth pointing out that the basis functions (3.3) and (3.5) depend only on the locations of the knots and the continuous variable x. This means that $\tilde{X}^S$ does not depend on the model and it could be included as-is in any predictor η 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, let us suppose a count distribution where the mean parameter is computed with covariates $\boldsymbol{x}_{i,t}$ which could either be categorical or continuous. We include an intercept, r 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 $\boldsymbol{x}_{i,t}$ can be represented as:

$$\boldsymbol{x}_{i,t} = \left(1, \underbrace{x_{1_{i,t}}, ..., x_{r_{i,t}}}_{\text{categorical covariates}}, \underbrace{x_{r+1_{i,t}}, x_{r+2_{i,t}}}_{\text{continuous covariates}}, \underbrace{f(x_{r+3_{i,t}})}_{\text{smoothing function}}\right). \tag{3.7}$$

Equation (3.7) can then be rewritten as:

$$\boldsymbol{x}_{i,t} = \left(1, \underbrace{x_{1_{i,t}}, ..., x_{r_{i,t}}}_{\text{categorical covariates}}, \underbrace{x_{r+1_{i,t}}, x_{r+2_{i,t}}}_{\text{continuous covariates}}, \underbrace{b_1(x_{r+3_{i,t}}), ..., b_{\tilde{q}}(x_{r+3_{i,t}})}_{\text{smoothing function}}\right). \tag{3.8}$$

3.2.2 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 p parametric covariates, the following equation is an example of a penalised linear regression model with two splines of $\tilde{q}_1$ and $\tilde{q}_2$ knots:

$$||\sqrt{\boldsymbol{W}}\boldsymbol{y} - \sqrt{\boldsymbol{W}}\boldsymbol{X}\boldsymbol{\beta}||^2 + \Lambda_1\boldsymbol{\beta}_{S_1}^T\tilde{\boldsymbol{S}}_1\boldsymbol{\beta}_{S_1} + \Lambda_2\boldsymbol{\beta}_{S_2}^T)\tilde{\boldsymbol{S}}_2\boldsymbol{\beta}_{S_2}, \tag{3.9}$$

with

$$\boldsymbol{\beta} = (\underbrace{\beta_0, ..., \beta_{p-1}}_{\boldsymbol{\beta}^{PAR}}, \underbrace{\beta_{p+1}, ..., \beta_{p+\tilde{q}_1}}_{\boldsymbol{\beta}_{S_1}}, \underbrace{\beta_{p+\tilde{q}_1+1}, ..., \beta_{p+\tilde{q}_1+\tilde{q}_2}}_{\boldsymbol{\beta}_{S_2}}), \tag{3.10}$$

where the coefficients associated with parametric terms are denoted ($\boldsymbol{\beta}_{PAR}$), the ones associated with spline 1 ($\boldsymbol{\beta}_{S_1}$) and the coefficients associated with spline 2 ($\boldsymbol{\beta}_{S_2}$) and so on if there are other smoothing functions. For a classic linear regression (Normal distribution), the matrix of weights $\sqrt{\boldsymbol{W}}$ is the identity matrix. An alternative representation of Equation (3.9) includes a scaling factor (σ_k) :

$$||\sqrt{\boldsymbol{W}}\boldsymbol{y} - \sqrt{\boldsymbol{W}}\boldsymbol{X}\boldsymbol{\beta}||^2 + (\Lambda_1 \cdot \sigma_1)\boldsymbol{\beta}_{S_1}^T(\tilde{\boldsymbol{S}}_1/\sigma_1)\boldsymbol{\beta}_{S_1} + (\Lambda_2 \cdot \sigma_2)\boldsymbol{\beta}_{S_2}^T)(\tilde{\boldsymbol{S}}_2/\sigma_2)\boldsymbol{\beta}_{S_2}. \tag{3.11}$$

This is useful to point out because some statistical software uses this representation (like the "mgcv" package in **R**). The matrices $\tilde{\boldsymbol{S}}_k$, $k \in \{1, 2\}$ 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. Λ_k, $k \in \{1, 2\}$ are hyperparameters that control the wiggliness. The higher Λ_k 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 X^S had been reduced for the sake of identifiability, the penalty matrix $\boldsymbol{S}$ must also be adapted accordingly (see Equation 3.9). One way to proceed would be to find the orthogonal matrix of vector $\mathbf{1}_{1\times n}X^S_{n\times q}$ using the QR decomposition and drop the first column. Let us denote this matrix $\boldsymbol{Z}$. Then, $\tilde{\boldsymbol{X}}^S = \boldsymbol{X}^S\boldsymbol{Z}$ and $\tilde{\boldsymbol{S}}^S = \boldsymbol{Z}^T\boldsymbol{S}^S\boldsymbol{Z}$. This step is obviously also necessary for any new data $\boldsymbol{X}'^S$ for which we want to make predictions : $\tilde{\boldsymbol{X}}'^S = \boldsymbol{X}'^S\boldsymbol{Z}$.

Cubic regression splines. The penalty of the cubic regression spline $f(x)$ is $\int_{k_1}^{k_{q+1}} f''(x)^2\partial x = \boldsymbol{\beta}^T D^T B^{-1} D\boldsymbol{\beta}$, meaning that $S \equiv D^T B^{-1} D$. 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 f that could be written as a linear combination of basis functions (i.e. $f = \sum_j \beta_j b_j(x)$), could be written as $\int f''(x)^2\partial x = \boldsymbol{\beta}^\intercal\boldsymbol{S}\boldsymbol{\beta}$ if the basis functions are twice differentiable. The penalty matrix $\boldsymbol{S}$ can then be expressed solely as a function of the known basis functions. This comes in handy when defining penalties for various splines.

P-splines As mentioned, P-splines uses 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 $m = 2$ (not related to the order of the B-spline), the penalty is:

$$\mathcal{P} = \sum_{i=1}^{q}(\beta_{i+1} - \beta_i)^2 \tag{3.12}$$

or

$$\mathcal{P} = \boldsymbol{\beta}^T\boldsymbol{P}^T\boldsymbol{P}\boldsymbol{\beta}, \tag{3.13}$$

where

$$\boldsymbol{P} = \begin{pmatrix} -1 & 1 & 0 & \dots & \dots \\ 0 & -1 & 1 & 0 & \dots \\ . & . & . & . & . \\ . & . & . & . & . \end{pmatrix} \tag{3.14}$$

and thus $S \equiv \boldsymbol{P}^T\boldsymbol{P}$. 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 β_i. If all the increments in $\boldsymbol{\beta}$ were the same value, it would result in a linear function. If all β_i were the same, it would result in a flat line.

3.2.3 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.

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-likelihoood:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i\theta_i - b(\theta_i)\}/a(\phi) + c(\phi, y_i),$$

where $\boldsymbol{y}$ represents the observations of response variable $\boldsymbol{Y}$, b, a and c are functions, ϕ is a scale parameter and θ is the canonical parameter, which depends on the selected distribution. For the Poisson distribution $a = \phi$, $b = \exp(\theta)$, $c = -\log(y!)$, $\phi = 1$ and $\theta = \log(\mu)$. In Equation (2.1), the parameter λ corresponds to μ and the definition of η and g are unchanged from Section 2.1. Lastly, we recall that the variance function is $V(\mu) = Var(Y)/\phi$ and we define $\alpha(\mu_i) = 1 + (y_i - \mu_i)\{V'(\mu_i)/V(\mu_i) + g''(\mu_i)/g'(\mu_i)\}$. Taking $\alpha(\mu_i)$ as 1 corresponds to Fisher weights and this definition of $\alpha(\mu_i)$ has been used in this work.

The vector $\boldsymbol{\beta}$ is generally estimated using the iteratively re-weighted least square algorithm (IRLS). It is first required to initialise $\hat{\mu}_i = y_i + \delta_i$ and $\hat{\eta}_i = g(\hat{\mu}_i)$. δ_i is usually zero (it is meant to ensure that $\hat{\eta}_i$ is finite). The next two steps are iterated to convergence.

1. Compute the vector $\boldsymbol{z}_{n\times 1}$ of the pseudodata $\boldsymbol{z}_{n\times 1} = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i)/\alpha(\mu_i) + \hat{\eta}_i$ and the nonzero elements of the diagonal of $\boldsymbol{W}$: $w_i = \alpha(\hat{\mu}_i)/\{g'(\hat{\mu}_i)^2 V(\hat{\mu}_i)\}$,
2. The minimiser of the weighted least squares objective $\sum_{i=1}^{n} w_i(z_i - \boldsymbol{X}_i\boldsymbol{\beta})^2$ is $\boldsymbol{\beta}^{[k+1]} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{z}$.

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

GAM. For known values of hyperparameter $\boldsymbol{\Lambda}$, the estimation of parameter vector $\hat{\boldsymbol{\beta}}$ 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

$$\hat{\boldsymbol{\beta}}^{[k+1]} = (\boldsymbol{X}_{plus}^T\boldsymbol{w}_{plus}\boldsymbol{X}_{plus})^{-1}\boldsymbol{X}_{plus}^T\boldsymbol{w}_{plus}\boldsymbol{z}_{plus}, \tag{3.15}$$

with

$$\boldsymbol{X}_{plus} = \begin{pmatrix} X_{n\times p} & \tilde{X}^{S_1}_{n\times q_1} & \tilde{X}^{S_2}_{n\times q_2} \\ 0 & \Lambda_1 \cdot \tilde{S}^1_{q_1\times q_1} & 0 \\ 0 & 0 & \Lambda_2 \cdot \tilde{S}^2_{q_2\times q_2} \end{pmatrix},$$

$$\boldsymbol{z}_{plus}^T = \begin{pmatrix} z_{n\times 1}^T & 0_{q_1\times 1}^T & 0_{q_2\times 1}^T \end{pmatrix},$$

and

$$\boldsymbol{w}_{plus} = \begin{pmatrix} w_{1\times n} & 0 \\ 0 & 1_{1\times(q_1+q_2)} \end{pmatrix}$$

$\boldsymbol{z}$ and $\boldsymbol{w}$ are defined as before in the GLM paragraph. For known values of hyperparameter $\boldsymbol{\Lambda}$, with this newly defined minimiser, the rest of the procedure is like the IRLS algorithm. In order to select the hyperparameter Λ for a model with splines, it is not possible to use likelihood based statistics because a non-penalised likelihood ($\Lambda = 0$) 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:

$$GCV = \frac{n\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{[n - tr(\boldsymbol{A})]^2}, \tag{3.16}$$

where $tr(\boldsymbol{A})$ is the trace of the influence matrix which corresponds to the number of effective degrees of freedom (EDF). The influence matrix is

$$\boldsymbol{A} = \left(\boldsymbol{X}_{plus}^T \boldsymbol{w}_{plus} \boldsymbol{X}_{plus}\right)^{-1} \boldsymbol{X}^T \boldsymbol{w} \boldsymbol{X}. \tag{3.17}$$

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

3.3 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.

3.3.1 Estimation Steps

In Rigby and Stasinopoulos (2005), two algorithms are described to estimate the parameters $\hat{\boldsymbol{\beta}}$ that maximise the penalised likelihood given the hyperparameter $\boldsymbol{\Lambda}$. The RS algorithm (from Rigby and Stasinopoulos (1996)) and CG algorithm (from Cole and 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 (μ), the scale (σ) and the shape (ν and τ). Put another way, this means that there is an outer iteration step that successively estimates the parameters μ, σ, ν and τ by using the inner iteration steps, which includes a modified PIRLS algorithm, which Rigby and Stasinopoulos (1996) have called "a modified backfitting algorithm." The other parameters (for example σ, ν and τ) are assumed to be fixed while in the inner iteration of a parameter (for example μ). 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

$$l_p = l - \sum_{k=1}^{4}\sum_{j=1}^{J_k} \Lambda_{k,j} \boldsymbol{\beta}_{S_{k,j}}^T \boldsymbol{S}_{k,j} \boldsymbol{\beta}_{S_{k,j}} \tag{3.18}$$

where l 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) (μ, σ, ν and τ). J_k is the number of non-parametric terms included in the k^{th} parameter.

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 μ, 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 $\hat{\mu}$ (or $\hat{\sigma}$, $\hat{\nu}$ and $\hat{\tau}$), the weight ω and the pseudo-observation z are calculated. The weight ω is

$$\omega = -\boldsymbol{f} \circ \left(\frac{\partial \mu}{\partial \eta}\right) \circ \left(\frac{\partial \mu}{\partial \eta}\right), \tag{3.19}$$

where the operator '$\circ$' represents the Hadamard element-wise product (see Equation (3.21)), η is defined as before ($\eta = g(\mu)$) and $\boldsymbol{f}$ is either the second derivative of the log-likelihood with respect to μ for the standard Newton-Raphson scoring algorithm, or the expected value of this second derivative for the Fisher scoring algorithm or $\boldsymbol{f}$ is (3.20) for a quasi Newton scoring algorithm.

$$\boldsymbol{f} = -\left(\frac{\partial l}{\partial \mu}\right) \circ \left(\frac{\partial l}{\partial \mu}\right). \tag{3.20}$$

The pseudo-observation is then given by

$$z = \eta + \omega^{-1} \circ u,$$

where u is

$$u = \frac{\partial l}{\partial \eta} = \left(\frac{\partial l}{\partial \mu}\right) \circ \left(\frac{\partial \mu}{\partial \eta}\right)$$

and

$$\omega^{-1} \circ u = \left(\omega_1^{-1} u_1, \omega_2^{-1} u_2, ..., \omega_n^{-1} u_n\right). \tag{3.21}$$

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

Modified backfitting algorithm. In this procedure, the parameters of the vector $\boldsymbol{\beta}$ are estimated iteratively by separating the estimation of the coefficients associated with parametric terms ($\boldsymbol{\beta}_{PAR}$), the ones associated with spline 1 ($\boldsymbol{\beta}_{S_1}$) and the coefficients associated with spline 2 ($\boldsymbol{\beta}_{S_2}$) and so on if there were other non-parametric terms (see Equation (3.10)).

1. Next, the parameters $\boldsymbol{\beta}_{PAR}$ for the parametric part are estimated minimising the weighted least square (see paragraph 'GLM' in Section 3.2.3) of $\varepsilon = z - \tilde{X}^{S_1}_{n\times\tilde{q}_1}\hat{\boldsymbol{\beta}}_{S_1} - \tilde{X}^{S_2}_{n\times\tilde{q}_2}\hat{\boldsymbol{\beta}}_{S_2}$ using the calculated weights ω .
2. Afterwards, $\varepsilon = z - X_{n\times p}\hat{\boldsymbol{\beta}}_{PAR} - \tilde{X}^{S_2}_{n\times\tilde{q}_2}\hat{\boldsymbol{\beta}}_{S_2}$ is calculated and $\boldsymbol{\beta}_{S_1}$ is estimated using the penalised weighted least square with the appropriate penalty (see paragraph "GAM" in Section 3.2.3).
3. Finally, fit the penalised weighted least squares to $\varepsilon = z - X_{n\times p}\hat{\boldsymbol{\beta}}_{PAR} - \tilde{X}^{S_1}_{n\times\tilde{q}_1}\hat{\boldsymbol{\beta}}_{S_1}$ to get $\boldsymbol{\beta}_{S_2}$.

Those steps are repeated until the vector $\boldsymbol{\beta}$ stops varying (convergence of the modified backfitting algorithm) and z and ω 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 Λ_k, the last step is to determine the value of Λ_k. Stasinopoulos et al. (2017) list several ways to achieve that. Among these, there is the GCV-based method presented above (see Equation (3.16)). The Λ that minimises the GCV will be chosen.

4 Numerical Analysis

4.1 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 concerns 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 and Shi (2019) had six years of data and Frees and 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 4.1 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 4.2 provides an overview of the covariates that were selected for modelling. Theses covariates includes two continuous covariates and some descriptive statistics are provided in Table 4.3, 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 4.1: Driver's duration of observation

	Variable	Description
x_1	Postal code	First letter of postal code: M (Toronto), P (North), K (East), L (Central) et N (West)
x_2	Gender	Insured's gender: Female or Male/Other
x_3	Marital status	Insured's marital status: Married, Divorced, Separated, Single, Widow/Widower
x_4	Usage	Primary vehicle use: Pleasure, Business (company car), Commute and Farm (Farmer's car)
x_5	Number years driving	Number years driving as a continuous variable
x_6	Vehicle age	Vehicle age in years as a continuous variable

Table 4.2: 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 4.3: Continuous covariate statistics

Finally, the database was separated into an in-sample and an out-of-sample dataset. Table 4.4 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].

	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 4.4: Number of observations in each subsample of the dataset

4.2 Cross-sectional data models

The models are fitted including all covariates described in Section 4.1 into the log-linear predictor from Equations (2.1) and (2.2). Table 4.5 shows the overdispersed Negative binomial regression having 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 4 can be found in Appendix A. Table 4.6 shows how significant the signals from continuous variables are in both parametric basic 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 4.5 shows better statistics for semi-parametric models for both the Poisson model and the Negative binomial model. This implies that the additional parameters associated 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	**30,945.99**	**31,104.15**	**−15,456.99**	16
Semi-parametric models				
Poisson cubic splines	30,871.27	31,062.03	−15,416.34	19.2972
NB cubic splines	30,841.64	**31,040.29**	−15,400.72	20.0960
Poisson p-splines	30,869.09	31,062.41	−15,414.99	19.5564
NB p-splines	**30,840.20**	31,052.87	**−15,398.59**	21.5134

Table 4.5: In-sample goodness-of-fit statistics for cross-section data models

[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.

	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 4.6: Estimated coefficients for continuous covariates in parametric cross-sectional data models

When analysing the shape of the splines in Figure 4.1 it can be seen that the splines are somewhat similar for the two splines considered. This was unsurprising because the number of EDF (Table 4.5) are similar and cubic regression splines and p-splines yields similar results in this situation. In addition, one may observe that the inclusion of splines really benefits the modelling 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 indicates a shift at around 40 years 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.

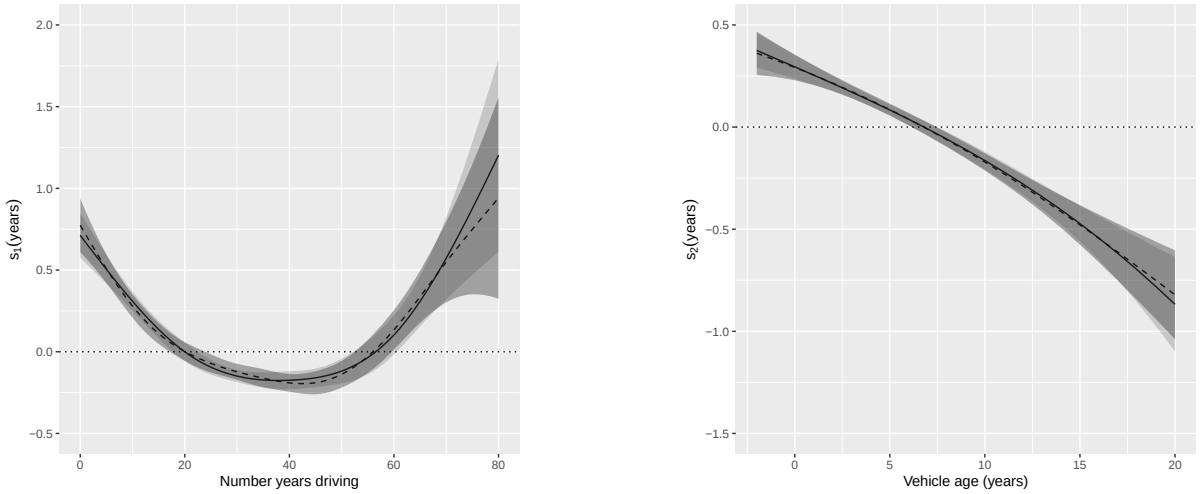

Figure 4.1: Estimated cubic spline (solid) and P-spline (dashed) for the Poisson model

4.3 Panel data models

We fit panel data model on the data to check how adding the claims history to the modelling influences the results. Results are shown in Table 4.7. Compared to Table 4.5, 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 combination 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 4.2 and 4.3 are different from Figure 4.1, 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	**30,902.53**	**31,070.58**	**−15,434.26**	17
Semi-parametric models				
MVNB cubic splines	30,813.65	**31,030.92**	−15,384.85	21.9784
Beta-NB cubic splines	**30,813.28**	31,039.97	**−15,383.71**	22.9316
MVNB p-splines	30,857.56	**31,122.84**	−15,401.94	26.8358
Beta-NB p-splines	**30,853.29**	31,125.43	**−15,399.12**	27.5295

Table 4.7: In-sample goodness-of-fit statistics for panel data models

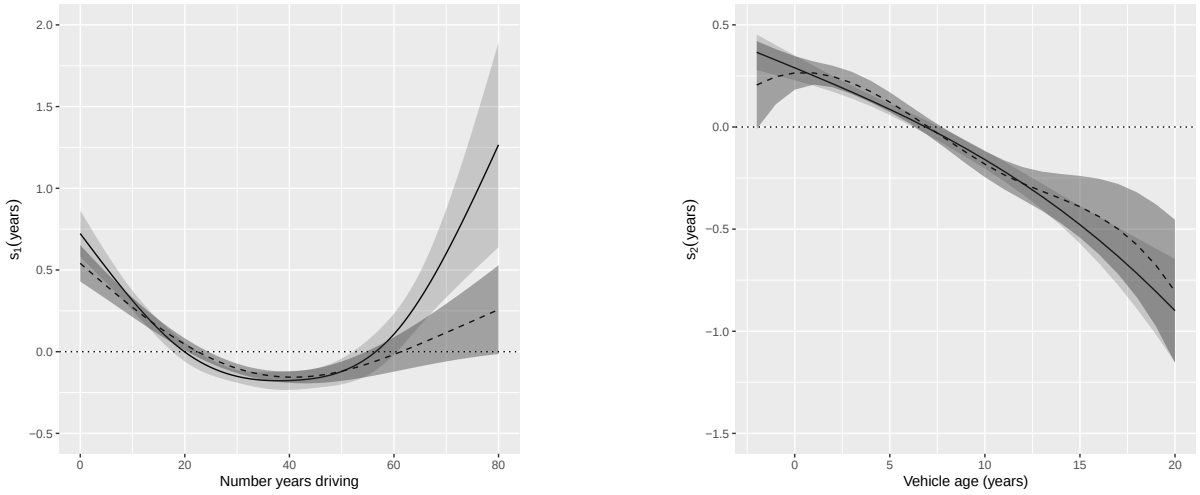

Figure 4.2: Estimated cubic spline (solid) and P-spline (dashed) for the MVNB model

4.4 Analysing the Results

4.4.1 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 calculated, but only the logarithm score, the Dawid-Sebastiani score and the squared error score are presented in Table 4.8 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.

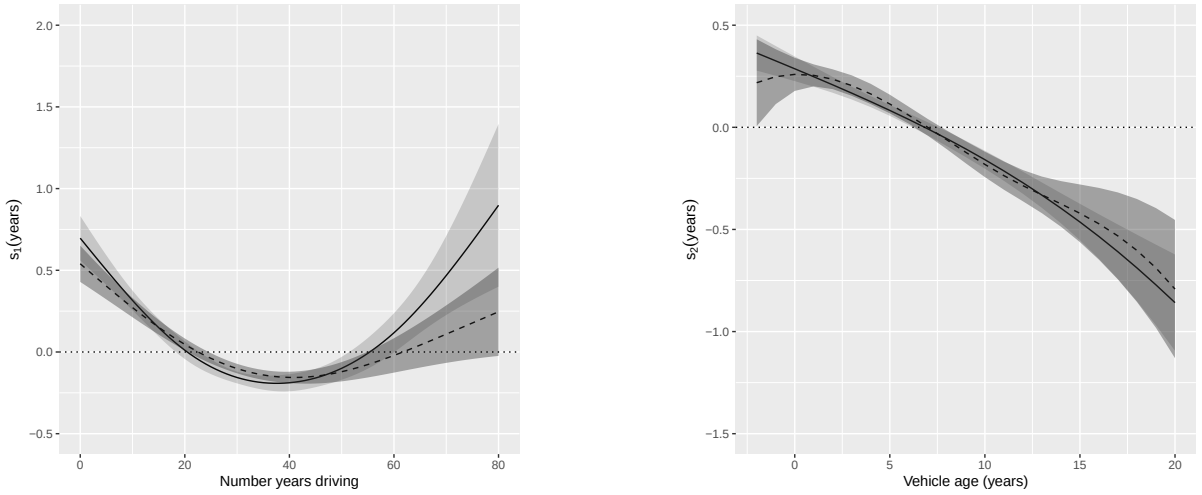

Figure 4.3: Estimated cubic spline (solid) and P-spline (dashed) for the Beta-NB model

The logarithm score $\text{logs}(P, y)$ is the probability of observing the test sample $\boldsymbol{y}$ using the predictive distribution P 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:

$$\text{logs}(P, y) = -\log(p_y),$$

where p_y is the probability of observing the data y. The squared error $\text{ses}(P, y)$ is a common distance measure used in statistics and it corresponds to the squared difference of the data y and the estimated prediction μ_p :

$$\text{ses}(P, y) = (y - \mu_p)^2.$$

Lastly, the Dawid-Sebastiani score $\text{dss}(P, n)$ is a squared error that has been normalised by the variance, and a variance-based penalty has been added:

$$\text{dss}(P, y) = \left(\frac{y - \mu_p}{\sigma_p}\right)^2 + 2\log(\sigma_p).$$

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 4.8 because the Poisson distribution is a more common distribution for modelling count data.

Table 4.8 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.

4.4.2 Premiums 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 risk. Covariate values for each profile are detailed in Table 4.9. Table 4.10

	Logarithmic	Dawid-Sebastiani	Poisson deviance	Squared error
Parametric models				
Poisson	5,132.46	−89,317.24	10,240.54	1,023.73
NegBin	5,127.03	−89,813.46	10,240.56	1,023.87
MVNB	5,122.86	−90,050.49	**10,228.81**	**1,023.71**
Beta-NB	**5,121.31**	**−91,302.40**	10,235.00	1,023.81
Cubic splines				
Poisson	5,125.40	−91,915.64	10,226.43	1,023.95
Negbin	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	**5,116.77**	**−93,835.96**	**10,219.90**	**1,023.96**
P-splines				
Poisson	5,123.85	−92,091.24	10,223.33	1,023.87
Negbin	**5,118.26**	−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	**−92,927.10**	**10,220.95**	**1,023.76**

Table 4.8: Proper scoring rules and Poisson deviance on validation sample

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 modelling 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 4.11, 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 4.11 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 4.10, 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 4.9: Characteristics of risk profiles

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 4.12 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 demoninator. 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 NB-Beta model a more stable predictive premium,

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
Negbin	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
Negbin	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
Negbin	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 4.10: *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 4.11: Bonus depending on number of years without claims (medium risk) for mixture models

thus reducing the bonuses and maluses of the claims experience. Based on the results obtained for these data which are detailed in Table 4.8, this additional parameter of the NB-Beta model appears to provide a significant improvement to the modelling.

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 4.12: 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 4.11 and 4.12. 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 ν, 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, $1/\nu$, 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 α and β are still associated with semi-parametric models. The expectation of the Beta distribution being $\alpha/(\alpha+\beta)$, the expected value of θ is lower for the semi-parametric models.

5 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. Smoothing functions were included in longitudinal count models to improve predictive power and better capture *a priori* information. The general ideas of GAMs and GAMLSSs were introduced, their estimation methods detailed and possible smoothing functions to be included in the modelling of the parameter(s) were discussed. 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 and reduces the weight of past experience in predictive premiums 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 estimated splines, even if there might be some differences, still shared several important characteristics. This result was not necessarily expected at the beginning considering the significant differences between the cross-sectional and longitudinal models. 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 modelling.

A Estimated paramters

		Poisson	NB	MVNB	Beta-NB
	β_0	-2.9053 (0.0867)	-2.8910 (0.0883)	-2.8942 (0.0914)	2.0357 (0.4353)
x_1	β_K	-0.2510 (0.0717)	-0.2532 (0.0730)	-0.2481 (0.0760)	-0.2525 (0.0756)
	β_L	-0.1836 (0.0712)	-0.1844 (0.0725)	-0.1788 (0.0754)	-0.1824 (0.0750)
	β_N	-0.2158 (0.0722)	-0.2175 (0.0735)	-0.2112 (0.0764)	-0.2168 (0.0761)
	β_P	-0.2152 (0.0915)	-0.2183 (0.0930)	-0.2196 (0.0968)	-0.2254 (0.0965)
x_2	$\beta_{M/O}$	-0.1031 (0.0367)	-0.1041 (0.0373)	-0.0998 (0.0389)	-0.1018 (0.0388)
x_3	β_{Div}	0.6007 (0.2683)	0.6131 (0.2735)	0.5442 (0.2766)	0.5152 (0.2834)
	β_{Sep}	0.3846 (0.2057)	0.3938 (0.2091)	0.3812 (0.2113)	0.3624 (0.2147)
	β_{Sin}	0.2400 (0.0398)	0.2429 (0.0405)	0.2451 (0.0421)	0.2416 (0.0420)
	β_{Wid}	0.3942 (0.2691)	0.4011 (0.2733)	0.3790 (0.2750)	0.3872 (0.2742)
x_4	β_{Bus}	0.0025 (0.0014)	0.0037 (0.0014)	-0.0010 (0.0014)	0.0063 (0.0014)
	β_{Com}	-0.0342 (0.0044)	-0.0331 (0.0044)	-0.0362 (0.0045)	-0.0361 (0.0045)
	β_{Far}	-0.4023 (0.0994)	-0.4037 (0.1009)	-0.3983 (0.1040)	-0.3919 (0.1035)
x_5	β_{yrslic}	-0.0095 (0.0412)	-0.0097 (0.0419)	-0.0097 (0.0430)	-0.0098 (0.0430)
x_6	β_{vehage}	-0.0508 (0.1438)	-0.0509 (0.1451)	-0.0499 (0.1492)	-0.0496 (0.1486)
	ν	-	-	1.4495 (0.2117)	-
	α	-	-	-	1.6803 (0.2889)
	β	-	-	-	232.3917 (92.0594)

Table A.1: Estimated parametric terms for parametric models

		Poisson	NB	MVNB	Beta-NB
	β_0	-3.5259 (0.0715)	-3.5165 (0.0728)	-3.5176 (0.0753)	1.3749 (0.4302)
x_1	β_K	-0.2492 (0.0718)	-0.2515 (0.0730)	-0.2500 (0.0757)	-0.2528 (0.0753)
	β_L	-0.1797 (0.0713)	-0.1808 (0.0726)	-0.1761 (0.0751)	-0.1803 (0.0747)
	β_N	-0.2314 (0.0722)	-0.2333 (0.0736)	-0.2297 (0.0762)	-0.2324 (0.0758)
	β_P	-0.2079 (0.0915)	-0.2102 (0.0931)	-0.2080 (0.0964)	-0.2169 (0.0962)
x_2	$\beta_{M/O}$	-0.1241 (0.0369)	-0.1260 (0.0374)	-0.1212 (0.0388)	-0.1235 (0.0387)
x_3	β_{Div}	0.6370 (0.2684)	0.6516 (0.2735)	0.5883 (0.2757)	0.5655 (0.2820)
	β_{Sep}	0.4204 (0.2058)	0.4307 (0.2092)	0.4134 (0.2115)	0.3982 (0.2147)
	β_{Sin}	0.1592 (0.0415)	0.1612 (0.0421)	0.1611 (0.0436)	0.1603 (0.0435)
	β_{Wid}	0.2596 (0.2695)	0.2677 (0.2736)	0.2484 (0.2751)	0.2583 (0.2744)
x_4	β_{Bus}	0.1060 (0.1002)	0.1094 (0.1018)	0.1043 (0.1046)	0.1114 (0.1040)
	β_{Com}	0.0303 (0.0429)	0.0333 (0.0435)	0.0285 (0.0446)	0.0301 (0.0445)
	β_{Far}	-0.4520 (0.1439)	-0.4539 (0.1453)	-0.4455 (0.1489)	-0.4406 (0.1485)
	ν	-	-	1.5335 (0.2318)	-
	α	-	-	-	1.7983 (0.3236)
	β	-	-	-	239.5266 (94.2778)

Table A.2: Estimated parametric terms for models including cubic splines

		Poisson	NB	MVNB	Beta-NB
	β_0	-3.5218 (0.0715)	-3.5117 (0.0728)	-3.5271 (0.0759)	1.3719 (0.4298)
x_1	β_K	-0.2515 (0.0718)	-0.2542 (0.0731)	-0.2498 (0.0757)	-0.2548 (0.0754)
	β_L	-0.1818 (0.0713)	-0.1830 (0.0726)	-0.1786 (0.0751)	-0.1825 (0.0747)
	β_N	-0.2335 (0.0723)	-0.2361 (0.0736)	-0.2222 (0.0762)	-0.2278 (0.0758)
	β_P	-0.2101 (0.0916)	-0.2129 (0.0931)	-0.2173 (0.0965)	-0.2233 (0.0962)
x_2	$\beta_{M/O}$	-0.1246 (0.0369)	-0.1265 (0.0374)	-0.1115 (0.0388)	-0.1132 (0.0387)
x_3	β_{Div}	0.6340 (0.2685)	0.6475 (0.2735)	0.5726 (0.2762)	0.5498 (0.2822)
	β_{Sep}	0.4166 (0.2058)	0.4262 (0.2093)	0.4060 (0.2111)	0.3869 (0.2146)
	β_{Sin}	0.1591 (0.0415)	0.1604 (0.0422)	0.1958 (0.0428)	0.1926 (0.0427)
	β_{Wid}	0.2569 (0.2696)	0.2650 (0.2736)	0.2956 (0.2749)	0.3051 (0.2741)
x_4	β_{Bus}	0.1036 (0.1002)	0.1068 (0.1018)	0.0646 (0.1042)	0.0718 (0.1037)
	β_{Com}	0.0286 (0.0429)	0.0307 (0.0436)	0.0055 (0.0438)	0.0053 (0.0438)
	β_{Far}	-0.4550 (0.1440)	-0.4569 (0.1453)	-0.4273 (0.1490)	-0.4202 (0.1484)
	ν	-	-	1.5324 (0.2317)	-
	α	-	-	-	1.7950 (0.3229)
	β	-	-	-	238.9721 (93.9641)

Table A.3: Estimated parametric terms for models including p-splines

References

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., and Thandi, N. (2007). A practitioner's guide to generalized linear models—a foundation for theory, interpretation and application.

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

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

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

Cole, T. J. and 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., and 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., and 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., and Wüthrich, M. V. (2021). Making tweedie's compound poisson model more accessible. European Actuarial Journal, pages 1–42.

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

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

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

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

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

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

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

Gilchrist, R., Kamara, A., and 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.

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

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

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

Heller, G. Z., Mikis Stasinopoulos, D., Rigby, R. A., and 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., and Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes. Scandinavian Actuarial Journal, 2018(8):681–705.

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

Klein, N., Denuit, M., Lang, S., and 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.

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.

Li, J. and 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.

Nelder, J. A. and 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.

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

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

Rigby, R. A. and 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.

Shi, P. and 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., and Schewe, F. (2014). The use of gamlss in assessing the distribution of unpaid claims reserves. In Casualty Actuarial Society E-Forum, Summer 2014-Volume 2.

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

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

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

Tzougas, G. and 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., and 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., and 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.

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

Yang, L. and 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.