SEMI-PARAMETRIC MODELING OF RISK EXPOSURE WITH MONOTONICITY CONSTRAINTS IN AUTOMOBILE INSURANCE

A PREPRINT

Roxane Turcotte
Département de mathématiques
Université du Québec à Montréal
turcotte.roxane@uqam.ca

September 13, 2024

ABSTRACT

Generalized additive models (GAM) allow the inclusion of smoothing functions in the modeling of model parameters. This makes it possible for a much more flexible structure between a regressor and an outcome variable since linearity is no longer imposed as in the case of a generalized linear model (GLM). This type of model makes it easier to analyze the relationship between variables. However, such flexibility is not always desirable when working with data associated with a practical context. In the case of car insurance data, certain form constraints are necessary to model risk exposure : the greater the risk exposure, the higher the probability of an accident. It thus has to be an increasing function, and this increase should be reflected in the premium. In this work, shape constraints to model risk exposure measured by mileage have been included in GAM and GAMLSS (generalized additive models of location, scale and shape) models. We use data from a Canadian company to illustrate the proposed approaches. We discuss the relevance of mileage as a measure of risk exposure and how the measure should be included in the modeling.

Keywords : Panel data, ratemaking, exposure, spline, monotonicity constraints, mileage

1 Introduction

Since telematics data collection has become widespread in the car insurance market, there has been interest in using mileage driven to model claims risk. Indeed, such a measure would have the advantage of offering the insured a degree of control over their insurance premium, as well as providing an incentive to reduce car use, which has numerous social benefits, notably in terms of pollution reduction. Classically, the exposure measure used by insurers is the duration of the contract on an annual basis. This means that the insured pays the same daily premium for each of their days, and would be reimbursed half the annual premium if they were to cancel their policy midterm. For coverage against vehicle theft, the use of this exposure measure seems appropriate. However, if we consider collision or third-party liability coverage, the use of the vehicle is the main cause of exposure to risk. However, few policyholders are able to provide a fairly accurate estimate of their annual mileage when the question is asked by the insurer at the start of the contract. By using duration as an exposure measure, policyholders sharing similar risk profiles may pay more than their actual risk because they compensate for those who use their vehicles more. In this light, the study of mileage as an exposure measure for coverages such as collision is relevant to actuarial science. Moreover, given that some research has concluded that including realized mileage may render variables such as the insured's gender insignificant

([3], [43]), a variable less socially acceptable than in the past, and even now prohibited in some jurisdictions ([1]), it is all the more important to take an interest in its modeling.

Several approaches have been proposed to improve pricing methods by including mileage. The idea of pricing based on vehicle use goes back to [44]. Telematics was not available at the time, so they proposed using a tax on gasoline or tires instead. Telematics data breathed new life into this idea with works such as [2], [4], [27], [43] and [18]. However, none of these have used shape-constrained splines to model distance traveled. In actuarial science, splines with monotonicity constraints have mainly been used in the context of mortality modeling, since the mortality rate is a monotonic non-decreasing function ([38],[40]).

Over the past decade, telematics data have become part of the new reality of automobile insurance underwriting. Both practitioners and researchers have taken an interest in the potential and value of this data. The use of GPS-gathered data on policyholders' driving habits could improve existing pricing models or create new insurance products. [43] had in fact concluded that the best pricing structure included both "traditional" variables and telematic data. Mileage as a measure of risk exposure was also considered. One of the challenges of using telematic data is the size of the data. A number of studies have focused on summarizing data, such as [45], [16] and [23], for example. Another approach was to study how much information needed to be collected before it started to become redundant ([12]). Annual mileage is one of the easiest pieces of GPS data to collect and analyze. The data being collected in real-time can come in handy for dynamic pricing. The impact of telematics on insurance products and insurability has been studied by [13], including ethical issues. We submit that the use of the annual mileage appears to respect privacy. It is not necessary to collect a detailed record of the journeys, and we don't consider the times when these kilometers were driven, which doesn't interfere with the person's lifestyle, which could be correlated with their socio-cultural group, for example.

In the random-effects model of [6], it was observed that the relationship between mileage and risk became decreasing when a certain threshold (in kilometers) was exceeded. This result was explained by the fact that users with mileage in the highest quantiles of the distribution are different risks from the typical insured. For example, they are more likely to use freeways, which are considered safer roads because there are no intersections and traffic is unidirectional. In another research, it was argued by [4] that the decreasing curve in the higher quantiles could be the result of a learning effect. By driving more, one would become a better risk, as one would have gained driving experience. Regardless of the explanation, this means that the marginal effect of mileage on accident risk is not observed, since the marginal risk of an additional kilometer is necessarily increasing. Indeed, driving an extra kilometer, rather than leaving your car parked, inevitably results in a greater exposure to collision risk, whatever might be the insured's risk profile. Instead, we observe the apparent effect, influenced by other uncontrolled factors. As the relationship between mileage and risk must necessarily be increasing, we wish to correct the irregularities observed. We therefore propose a semi-parametric longitudinal model built by random effects, but whose non-parametric terms estimating the relationship between distance driven and accident risk would be subject to an increasing shape constraint. Different splines exist and thus different ways of imposing spline monotonicity. Unlike [4] and [6], which used only P-splines, several splines were considered for this work. We choose to work with P-splines, cubic regression splines and thin-plate regression splines, and will explain the impact on modeling. The aim is to obtain a model that can be used in practice and that has a logical relationship between mileage and risk, and between duration and risk.

In this work, non-parametric terms subject to a monotonicity constraint will be used within a generalized additive model for location, scale and shape (GAMLSS). This type of modeling allows parametric or non-parametric terms to be included in the estimation of several parameters, see [35] or [39]. GAMLSSs generalize both generalized linear models ([31]) and generalized additive models ([19]), since any distribution can be used with this framework which is not limited to the exponential dispersion family. The idea of using GAMLSSs for insurance data modeling is not new ([21], [42], [41]). This is the framework used in [6]. However, the problem of a usable consistent model for pricing with mileage as the risk exposure measure has not been answered, which is tackled in this work.

The article is organized as follows : in the next section, basic notions about the splines used will be recalled, and procedures to introduce shape constraints will be discussed and compared. In Section 3, we present random-effects panel data models in which shape constraints have been included in the modeling of risk exposure measured in mileage and year. These models are estimated using a Canadian dataset provided by a private insurer presented in the following section. In Section 4, the estimation results with and without shape constraints are detailed, and we study the impact of these constraints on premium structure under different claims histories. Finally, Section 5 concludes.

2 Smoothing functions with monotonicity constraints

Several techniques are available to obtain a continuous and increasing f spline function to model the relationship between mileage and risk. The pool adjacent violators (PAVA) algorithm is considered one of the first techniques to produce a monotonic regression function. It was proposed by [7] and is based on isotonic regression, which involves projecting a non-parametric function into a set of increasing functions. [25] suggested this technique in a regression context. A major shortcoming of this approach was that it could not produce a smooth function. To correct this limitation, several two-step approaches based on the PAVA algorithm were proposed ([14], [28], [30] or [17]). These two-step procedures consisted of unconstrained smoothing in parameter space and monotonization. These techniques are the precursors of shape-constrained splines.

[46] proposes a penalized approach to fitting a monotonic cubic regression spline. The method is based on the piecewise polynomial representation and linear sufficient conditions to ensure monotonicity first used in [22]. Another approach is to use the non-negativity or non-positivity of the derivative of the smoothing function to obtain the monotonicity constraints (e.g. [49]). [34] proposed a way to estimate a strictly monotonic twice differentiable function by solving a homogeneous linear differential equation. However, the estimation algorithm was not optimal, partly due to the computational heaviness of the procedure. [49] integrated the method into a generalized regression model, while [29] generalized the work of [33] to add a convexity constraint, in addition to that of monotonicity. Approaches specifically adapted to the local support of B-Splines have also been developed. One notable fact is that there is a sufficient, but not necessary, condition to ensure spline monotonicity. A non-decreasing (non-increasing) sequence of parameters is sufficient to guarantee a monotonic increasing (decreasing) spline (see [37]). The first approach based on this condition is probably [24], where optimization is performed under this constraint. [20] and [36] have also proposed approaches involving linear constraints. [32] have proposed an approach that reparametrizes the B-spline coefficients to guarantee monotonicity, without performing constrained estimation. [26] examines the estimation of monotonic B-splines specifically in a counting model for longitudinal data. They use clinical trial data to illustrate their methods. For our work, we use a B-spline approach, a cubic spline regression approach and a numerical method based on constraints on the value of the derivative. The [32] method uses unconstrained estimation, while the [46] method is based on linear constraints.

This section presents three splines and techniques for for monotonizing them, meaning $f(x_i) > f(x_j), \forall x_i > x_j$. Splines are piecewise smoothing functions. By definition, smoothing functions that can be used in the context of GAMs or GAMLSSs can be written as a linear combination of basis functions ($b_j(x)$) and parameters (β_j):

$$f(x) = \sum_{j=1}^{q} b_j(x)\beta_j. \tag{2.1}$$

This feature preserves the linear form of the predictor, without requiring a log-linear link between the covariates and the mean parameter. For a detailed presentation of these splines, please consult [47]. The first approach is based on B-splines, the second on regression cubic splines and the third on thin-plate regression splines.

2.1 B-splines with constraints

P-splines are based on B-spline basis functions, estimated using a penalty to control smoothing[1]. B-splines are local basis functions for fitting a spline of order $(m+1)$. $m = 2$ corresponds to a cubic spline, which is considered optimal for obtaining a smooth spline, but well fitted to the data ([47]). It is defined by $(q + m + 2)$ k_j nodes. Local basis functions imply that each basis function is uniquely non-zero on the interval between the m + 3 adjacent nodes. It can

[1] Considering the method for ensuring monotonicity is related to the structure of the basis functions rather than the estimation procedure, B-splines are specifically referred to a few times in the following.

be represented in the additive form

$$s(x) = \sum_{j=1}^{q} B_j^m(x)\gamma_j.$$

The basics functions of a B-spline are usually defined in a recursive form

$$B_j^m(x) = \frac{x - k_j}{k_{j+m+1} - k_j} B_j^{m-1}(x) + \frac{k_{j+m+2} - x}{k_{j+m+2} - k_{j+1}} B_{j+1}^{m-1}(x), \quad \text{where} \tag{2.2}$$

$$B_j^{-1}(x) = \begin{cases} 1 & \text{si } k_j \leq x \leq k_{j+1} \\ 0 & \text{otherwise.} \end{cases}$$

According to [9], the first derivative of a B-spline with uniformly spaced nodes is

$$s'(x) = \frac{1}{h} \sum_{j=2}^{q} B_j^{m-1}(x)\Delta^1\gamma_j.$$

Since the basis functions of a B-spline are non-negative by definition, a sufficient condition for $s'(x) > 0$ is that $\Delta^1\gamma_j = \gamma_j - \gamma_{j-1} > 0$. This implies that an increasing sequence of parameters γ_j will produce a monotonically increasing function. Figure 2.1 illustrates how an increasing sequence of parameters ensures the monotonicity of the spline. Each dotted line represents one of the ($q = 10$) basis functions multiplied by an increasing sequence of parameters. The addition of each of the dotted lines forms the resulting spline shown in red. One of the basis functions has been represented by a wider line, in order to better observe the shape of a basis function.

To obtain an increasing sequence, we define the γ_j parameters as follows ([32]):

$$\gamma_j = \begin{cases} \beta_1 & \text{si } j = 1 \\ \beta_1 + \sum_{l=2}^{j} \exp(\beta_l) & \text{si } l = 2, ..., q \end{cases} \tag{2.3}$$

where β_l are unconstrained parameters. Defining $\tilde{\boldsymbol{\beta}} = (\beta_1, \exp(\beta_2), \exp(\beta_3), ..., \exp(\beta_q))^T$, the smoothing function can thus be represented by $g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}$, where $X_i = \{B_1^m(x_i), B_2^m(x_i), ..., B_q^m(x_i)\}$ is the i$^{\text{th}}$ row of the design matrix $\boldsymbol{X}$ and $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 1 & \dots & 0 \\ . & . & . & \dots & . \\ 1 & 1 & 1 & \dots & 1 \end{pmatrix}.$$

Using this approach, there's no need to perform a constrained optimization on the parameter space, since the structure of the model ensures an increasing spline.

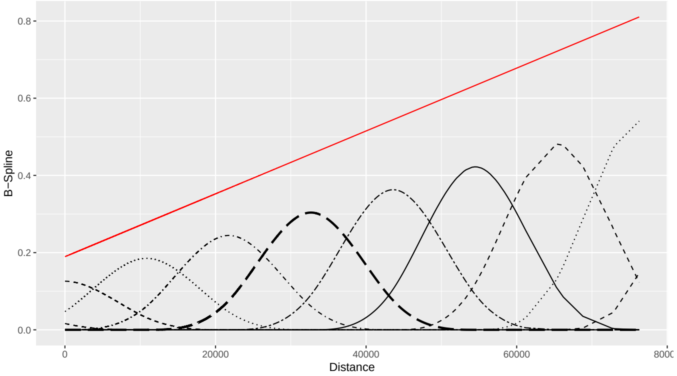

Figure 2.1: Illustration of a B-spline with increasing parameters

2.2 Regression cubic splines with constraints

A regression cubic spline is another method of defining cubic splines[2]. They are defined by q nodes k_j such that

$$s(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\boldsymbol{F}_j\boldsymbol{\beta} + c_j^+(x)\boldsymbol{F}_{j+1}\boldsymbol{\beta} = \sum_{j=1}^{q} b_j(x)\beta_j, \tag{2.4}$$

where

$$b_j(x) = \begin{cases} c_j^-(x)F_{j,i} + c_j^+(x)F_{j+1,i} + a_j^+(x) & \text{if } i = j+1, \\ c_j^-(x)F_{j,i} + c_j^+(x)F_{j+1,i} + a_j^-(x) & \text{if } i = j, \\ c_j^-(x)F_{j,i} + c_j^+(x)F_{j+1,i} & \text{otherwise.} \end{cases} \tag{2.5}$$

The parameter vector $\boldsymbol{\beta}$ is to be estimated. The terms $a_j^-(x)$, $a_j^+(x)$, $c_j^-(x)$, $c_j^+(x)$ are defined in Table 2.1. 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, the non-zero elements of the matrices $\boldsymbol{D}_{(q-2)\times(q)}$ and $\boldsymbol{B}_{(q-2)\times(q-2)}$ are defined in Table 2.1.

Basic functions for a cubic spline		
$a_j^-(x) = (k_{j+1} - x)/(k_{j+1} - k_j)$	$c_j^-(x) = \left((k_{j+1} - x)^3/(k_{j+1} - k_j) - (k_{j+1} - k_j)(k_{j+1} - x)\right)/6$	
$a_j^+(x) = (x - k_j)/(k_{j+1} - k_j)$	$c_j^+(x) = \left((x - k_j)^3/(k_{j+1} - k_j) - (k_{j+1} - k_j)(x - k_j)\right)/6$	
Definitions of non-zero elements of matrices D and B		
$D_{j,j} = 1/(k_{j+1} - k_j)$	$D_{j,j+1} = -(k_{j+1} - k_j)^{-1} - (k_{j+2} - k_{j+1})^{-1}$	$D_{j,j+2} = 1/(k_{j+2} - k_{j+1})$
$B_{j,j} = (k_{j+2} - k_j)/3$		$j \in \{1, ..., q-2\}$
$B_{j+1,j} = B_{j,j+1} = (k_{j+2} - k_{j+1})/6$		$j \in \{1, ..., q-3\}$

Table 2.1: Terms defining a cubic regression spline

[46] presents sufficient linear conditions to ensure the growth or decline of a cubic spline. Based on [48], sufficient necessary conditions to ensure monotonicity are that the values of $\kappa = s'(x_{i+1})(x_{i+1} - x_i)/(s(x_{i+1}) - s(x_i))$ and $\varphi = s'(x_i)(x_{i+1} - x_i)/(s(x_{i+1}) - s(x_i))$ lie in the space bounded by the curve shown in Figure 2.2. Let $\Delta_i = (s(x_{i+1}) - s(x_i))/(x_{i+1} - x_i)$ be the slope of the secant line of interval (x_i, x_{i+1}). κ and φ then represent the derivatives evaluated at each end of the interval divided by Δ_i. By ensuring monotonicity on each interval, the resulting spline will also be monotonic. The space delimited by the curve in Figure 2.2 is obtained by studying the behavior of $s(x)$, $s'(x)$ and $s''(x)$ in various scenarios and represents in fact the union of six spaces for the parameters that ensure

[2] In what follows, regression cubic splines will sometimes be abbreviated as "cubic splines"

monotonicity. For further details, please refer to [15]. Note that the constraint space of Figure 2.2 cannot be defined linearly. Based on [22], [46] proposes to restrict ourselves to zone A, in order to define linear constraints sufficient to ensure monotonicity. These constraints require that the values of κ and φ lie between 0 and 3. The approach proposed by [46] allows optimization with fewer constraints on the parameter space than the numerical approach presented in subsection 2.3.

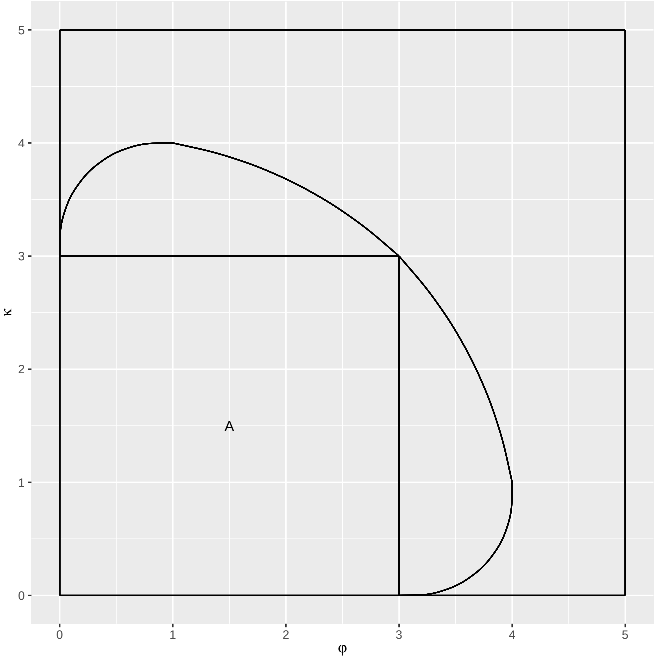

Figure 2.2: Constraint space

2.3 Numerical approach with TPRS

Thin Plate Regression Splines (TPRS) are a more general method than the ones previously discussed, see [11]. Among other things, they can be used to estimate multivariate splines ($d \geq 1$). One advantage of this method is that it is not necessary to specify node locations. For these splines, it is only required to specify a value m, called the rank of the spline. The spline can be written in the linear form

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

where $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$ are vectors of parameters to be estimated. The basis functions are defined by :

$$\eta_{md}(r) = \begin{cases} \dfrac{(-1)^{m+1+d/2}}{2^{2m-1}\Pi^{d/2}(m-1)!(m-d/2)!} r^{2m-d}\log(r) & d \text{ even,} \\ \dfrac{\Gamma(d/2-m)}{2^{2m}\Pi^{d/2}(m-1)!} r^{2m-d} & d \text{ odd} \end{cases} \tag{2.6}$$

where d is the dimension of the spline (here $d = 1$, as the spline is one-dimensional) and $M = \binom{m+d-1}{d}$. $\phi_j(\boldsymbol{x})$ are M linearly independent polynomials functions spanning the space of polynomials in $\mathbb{R}^d$. We have the constraint $\boldsymbol{T}^T\boldsymbol{\delta} = \boldsymbol{0}$, where $T_{i,j} = \phi_j(\boldsymbol{x}_i)$. The disadvantage of this method is that it requires a longer computation time than the other two methods, as the number of data increases. Thin-plate regression splines are used to reduce the space of $\boldsymbol{\delta}$ and thus the processing time, but the expression of the spline remains unchanged.

If there is little or no theory on monotonicity constraints for a spline, or if they are difficult to implement, it is possible to obtain a monotonically increasing spline by ensuring that the derivative is always positive (see Equation (2.7)). The values of the derivative evaluated at several points distributed over the entire domain are used as constraints. Optimization is therefore performed with a large number of constraints.

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} > 0 \tag{2.7}$$

This is an alternative method that can also be used with B-splines and cubic regression splines. In Section 4, the value of the derivative is evaluated at 100 points distributed over the entire domain. These 100 derivative evaluations were used to define the monotonicity constraints.

2.4 Comparison of approaches

The differences between the methods and some considerations for their intended use are discussed here. To this end, the basis functions of the three splines are illustrated in Figure 2.3 with unit parameters. The splines are expressed with ten basis functions each, fitted to the domain of the distance driven measured in kilometers for the illustration. The red line represents the spline resulting from these basis functions and coefficients equal to one. Once again, one of the basis functions has been represented by a wider line, in order to better observe its shape.

For the B-spline-based method, the vector of $\boldsymbol{\beta}$ coefficients is estimated, then reparameterized into $\boldsymbol{\gamma}$ coefficients that ensure an increasing sequence. In practice, we have observed that choosing a large number of nodes (which increases the number of parameters in the spline) can make it harder to achieve convergence. Indeed, the parameters associated with the nodes located on the highest quantiles of the distribution result from a sum of increasingly numerous exponential terms (see Equation (2.3)). Similarly, the greater the number of parameters, the more the first elements of the $\boldsymbol{\beta}$ vector interfere with the evaluation of several γ_j parameters. Another point to consider with regard to convergence is the choice of link function in the case where a large number of parameters are used to define the spline in a generalized model. By integrating the spline within an exponential link function, we obtain the exponential of a summation of exponential terms. However, constrained optimization is not required, since the structure of the model ensures an increasing spline, which can be an advantage.

The basis functions of cubic regression splines (second graph in Figure 2.3), unlike the basis functions of B-splines (first graph in Figure 2.3), are not locally defined. An increasing sequence of parameter vector coefficients is therefore not sufficient to ensure spline growth. If the spline is constrained to be increasing, and the basis functions are not local, the question arises as to whether it might be possible to correct the spline not locally, but over its entire domain, and thus "combat" the apparent effect observed in [4] and [6]. These two studies used only B-splines/P-splines. Considering that there are fewer observations in the tail of the distribution, it is not impossible that the shape irregularities observed are due to local overfitting caused by a lack of observations in this part of the domain. Furthermore, if we consider that the correlation between mileage and other driving habits is a phenomenon that affects the data over the whole domain, it may represent a wise choice to correct for shape irregularities by using non-zero basis functions over all or most of the domain.

Similarly, the basis functions of thin-plate regression splines are not locally defined either (third graph in Figure 2.3). In contrast to the other two approaches, TPRS has the notable advantage of not requiring the specification of node locations, whereas such specification is the result of subjective decisions. Besides, the absence of node locations avoids the need to define an extrapolation method beyond the value of the last node. For the contract duration, which is always between zero and one, the question of extrapolation is not a problem. On the other hand, the distance driven is not bounded. TPRS is recognized as a very good smoothing method, [47] even going so far as to describe it as an ideal smoother. For all these reasons, it was decided to include it in this comparative study.

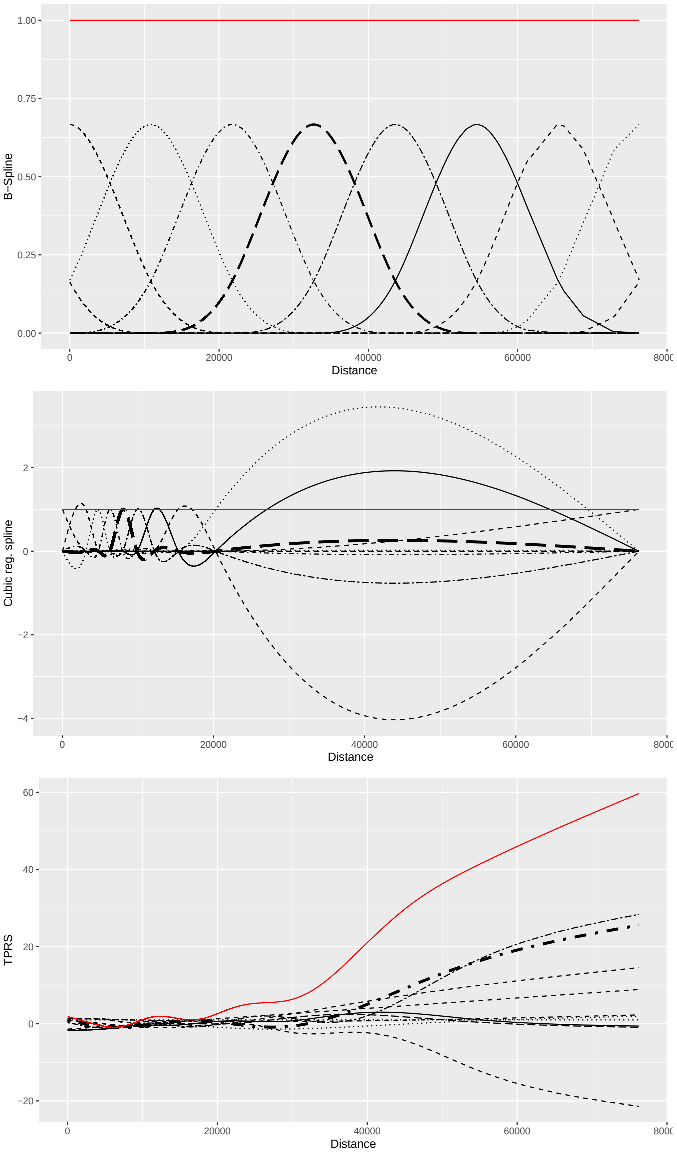

Figure 2.3: Illustration of the basis functions for each spline

3 Random effects model

Panel data is characterized by the observation of the same individual over time. An important advantage of panel data for motor insurance is that it enables longitudinal modeling of risk. Loss history is therefore used to capture some of the residual heterogeneity not explained by the segmentation variables. In other words, loss history would measure characteristics that cannot currently be collected, such as impatience at the wheel, propensity to be inattentive or lack of reflex ([10]). Studying the longitudinal modeling of risk of loss is therefore of interest to actuarial science.

In a research study by [6], two types of panel data models were considered : a random effect Poisson model and a fixed effect Poisson model. In both cases, the number of claims N of an insured i at time t is modeled using a Poisson distribution with mean $\mu_{i,t}$. The mean parameter integrates an effect α_i, random or fixed, which links contracts belonging to the same insured i:

$$N_{i,t} \sim \text{Poisson}(\mu_{i,t} = \alpha_i \lambda_{i,t}),$$

with $\lambda_{i,t} = \boldsymbol{X}_{i,t}\beta$ representing the a priori characteristics of the insured. In a random effects model, the α_i are independent and identically distributed variables. In a fixed effects model, the α_i are fixed parameters that must be estimated. [6] drew the conclusion that including a fixed individual factor made it possible to take into account the individual characteristics of policyholders that influence risk. As a result, they observed an almost linear relationship between mileage and accident risk. However, this fixed effects model cannot be used in practice since we would not know which individual factor to assign to a new policyholder. Nevertheless, this result justifies looking at a model whose exposure measure would be mileage, because such a proportional relationship between mileage and the risk of loss has a lot of potential. The random effects model can be used in practice since no factors need to be determined in advance for a new entrant. It would therefore be relevant to try to correct the shape irregularities that were observed.

Panel data models assume that all annual contracts belonging to the same driver i are dependent for $t = 1, \ldots, T_i$. If α always denotes the heterogeneity factor, the joint distribution of a random effects model for panel data can be expressed as:

$$\Pr[N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T}] = \int_0^\infty \left(\prod_{t=1}^{T_i} \Pr[N_{i,t} = n_{i,t} | \boldsymbol{x}_{i,1}, ..., \boldsymbol{x}_{i,T}, \alpha_i] \right) f(\alpha_i) d\alpha_i. \tag{3.1}$$

By selecting conjugate distributions for the count distribution and the random effect distribution, the likelihood is explicit and easier to use in practical situations.

The longitudinal models considered in this research work are all based on the multivariate negative binomial (MVNB) distribution. If we assume $N_{i,t}|\alpha \sim \text{Poisson}(\lambda_{i,t}\alpha)$, 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:

$$\Pr[N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T}] \quad = \quad \left(\prod_{t=1}^{T_i} \frac{\lambda_{i,t}^{n_{i,t}}}{n_{i,t}!} \right) \frac{\Gamma(n_{i,\bullet} + \nu)}{\Gamma(\nu)} \left(\frac{\nu}{\lambda_{i,\bullet} + \nu} \right)^{\nu} (\lambda_{i,\bullet} + \nu)^{-n_{i,\bullet}}, \tag{3.2}$$

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 and include exposition measures d, for the duration of the contract, and km, for the number of kilometers driven,

$$\lambda_{i,t} = \exp(d \cdot km \cdot \boldsymbol{x}_{i,t}\boldsymbol{\beta}) = g^{-1}(\eta_{i,t}). \tag{3.3}$$

This distribution is a generalization of the negative binomial distribution. It is a basic distribution for panel count data modelling with overdispersion ($\mathbb{E}[N_{i,t}] = \lambda_{i,t} < \mathbb{V}[N_{i,t}] = \lambda_{i,t} + (\lambda_{i,t})^2/\nu$). Because of $\lambda_{i,\bullet}$, it is not possible to simply view the longitudinal approach as a product of univariate distribution. If $\boldsymbol{n}_{i,(1:T)}$ denotes the claims history ($\boldsymbol{n}_{i,(1:T)} = [N_{i,1} = n_{i,1}, ..., N_{i,T} = n_{i,T}]$), the predictive distribution is a negative binomial distribution such that

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

whose predictive expectation is

$$E[N_{i,T+1} = n_{i,T+1} | \boldsymbol{n}_{i,(1:T)}] = \lambda_{i,T+1} \left(\frac{n_{i,\bullet} + \nu}{\lambda_{i,\bullet} + \nu} \right). \tag{3.4}$$

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 is the multivariate negative binomial. To work in a more general framework with any distribution, it is possible to consider generalized additive models for location, scale and shape (GAMLSS). It is possible to use this framework to include parametric or non-parametric terms in Equation (3.3). In the process of estimating the mean parameter, the parameters of each spline and those associated with the parametric terms are estimated separately. It is therefore possible to include either a reparameterization of the coefficient vector or to perform a constrained optimization on the parameters associated with the spline only. For further details on random effects modeling or MVNB, please refer to [5]. For more explanations on GAMLSS, please refer to [39]. [41] also specifically addressed the modeling of an MVNB regression with the GAMLSS framework.

4 Numerical applications

4.1 Data

The database comes from an important Canadian P&C insurance company and pertains to personal car insurance from the province of Ontario. Only policies that have been observed for a minimum of 100 days were retained in the analysis. In addition, only claims associated with road accidents (collision coverage) were included in this analysis, since we consider that automobile mileage has little association with the risk of theft or vandalism, for example.

The database information was partly collected using telematics devices placed on the vehicles. It was therefore possible to reliably determine the number of kilometers traveled by the vehicle during the insurance period. It should be noted that only 10 to 15 % of the insurer's overall portfolio decided to subscribe to the telematics program. According to the explanations received, mainly technophiles who appreciate new technologies and drivers at higher risk (such as inexperienced drivers or those with a bad driving record) subscribed to the telematics program since a discount was offered. The telematics portfolio therefore has a higher claim frequency than the overall portfolio, standing at 6 %. Considering that the models in practice are estimated with covariates, a small number of them were included to illustrate the segmentation. These variables are gender, marital status and vehicle use. The graphs in Figure 4.1 provide a visual representation of the distribution of observations across categories.

Figures 4.2 and 4.3 show the distribution of the classic exposure measure, i.e. contract duration, and mileage. Note that contracts that were observed for less than 100 days were withdrawn. The average contract lasted 0.645 year. As for mileage, note that the distribution is right-skewed. The average distance traveled is $10,398$ km, while the median is rather $8,561$ km. The maximum mileage observed is $76,272$ km.

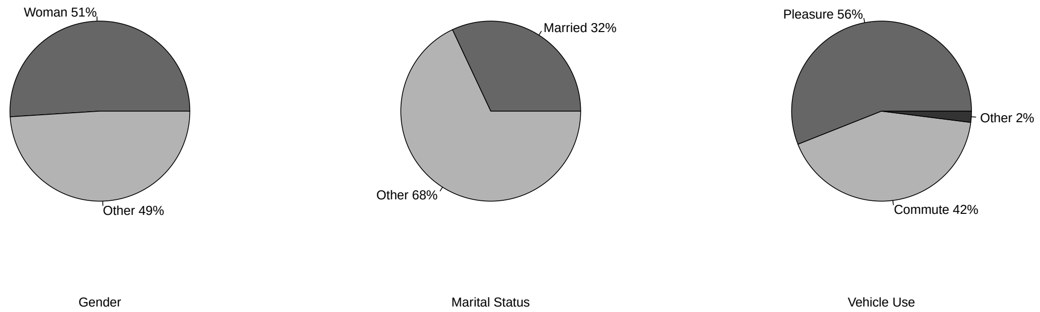

Figure 4.1: Répartition des variables explicatives par catégorie

The database has $59,685$ observations and Table 4.1 shows the attrition of the database over time. The database was split into an estimation database containing 80% of the policies and a test database containing the remaining 20%.

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 4.1: Distribution of the number of insurance periods for the database

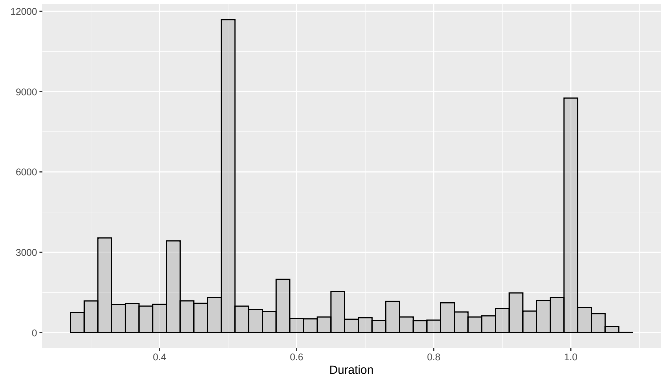

Figure 4.2: Histogram of risk exposure (in years). Each band has a length of 0.02 year.

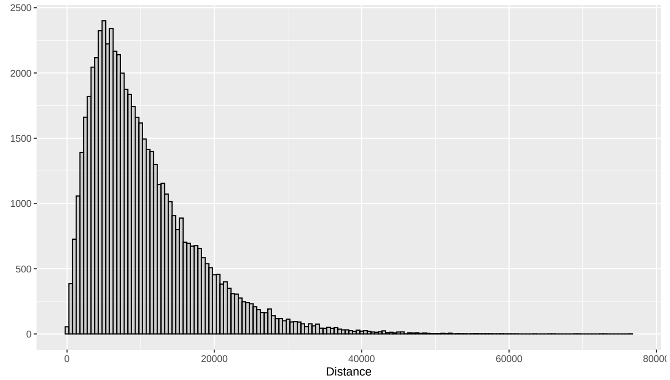

Figure 4.3: Histogram of distance driven (in km). Each band has a length of 500 km.

4.2 Results analysis

P-splines. Figures 4.4 and 4.5 present the results for the P-spline, a spline based on the B-spline basis functions estimated with a penalty. The decreasing shape of the mileage spline is again observed as it was in [4] and [6]. The spline associated with duration is also slightly decreasing on some segments of the domain. When an increasing constraint is imposed on the P-splines, we observe that the spline of duration flattens out. It is interesting to observe that if a monotonicity constraint is imposed on both splines, in a context where we model the risk of collision, it is the duration spline which flattens out while the mileage spline directs the level of the premium.

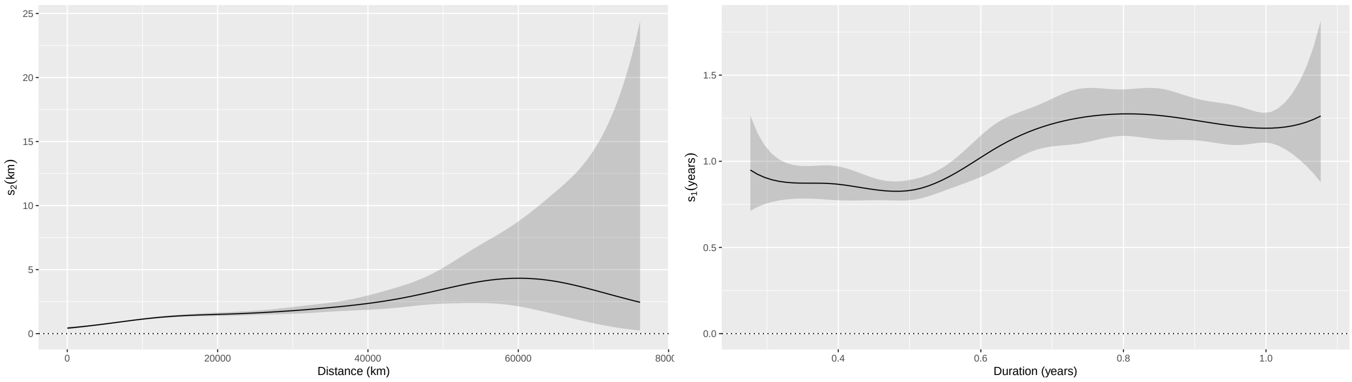

Figure 4.4: P-splines without monotonicity constraints

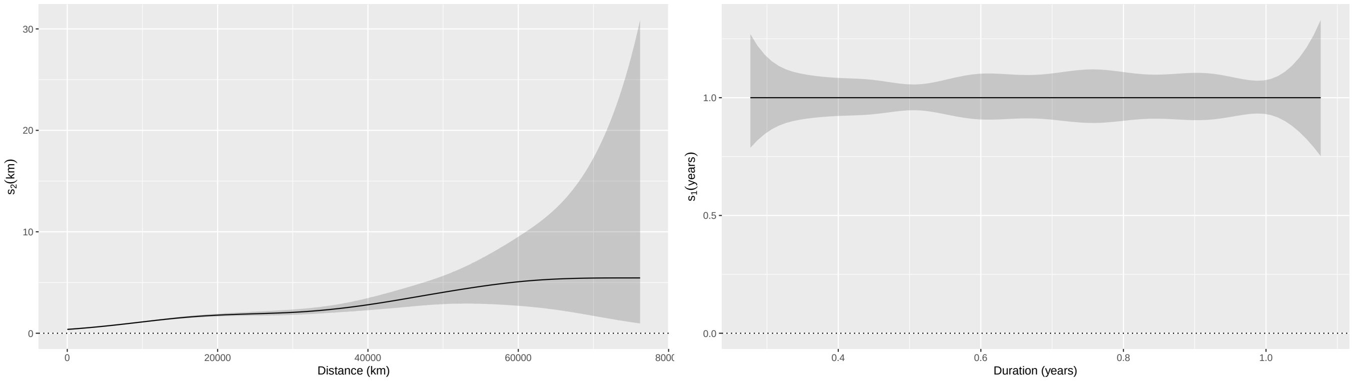

Figure 4.5: P-splines with monotonicity constraints

Cubic regression splines. Figures 4.6 and 4.7 present the results. It can be seen that by using a spline based on non-local basis functions (Figure 2.3), it is possible to obtain an increasing spline for the mileage without imposing monotonicity constraints. Local B-spline basis functions estimated this decrease with both Spanish data in [4], and Canadian data in [6]. The data becoming scarcer after 40,000 km (Figure 4.3) is a problem in this context. On the other hand, the spline associated with duration, whose observations are well distributed over its entire domain, presents several similarities with the spline in Figure 4.4. Yet it remains that the form obtained for the duration spline is not desirable to use it as an exposure measure in a pricing model, because the estimated spline is slightly decreasing over certain segments of the domain. Attempting to get a better fit for the duration by imposing monotonicity, we observe that the distance spline is virtually unchanged from the unconstrained cubic regression spline. For the duration spline, we find the same flattened shape as in Figure 4.5, probably meaning that the model requires the spline to decline to be optimal.

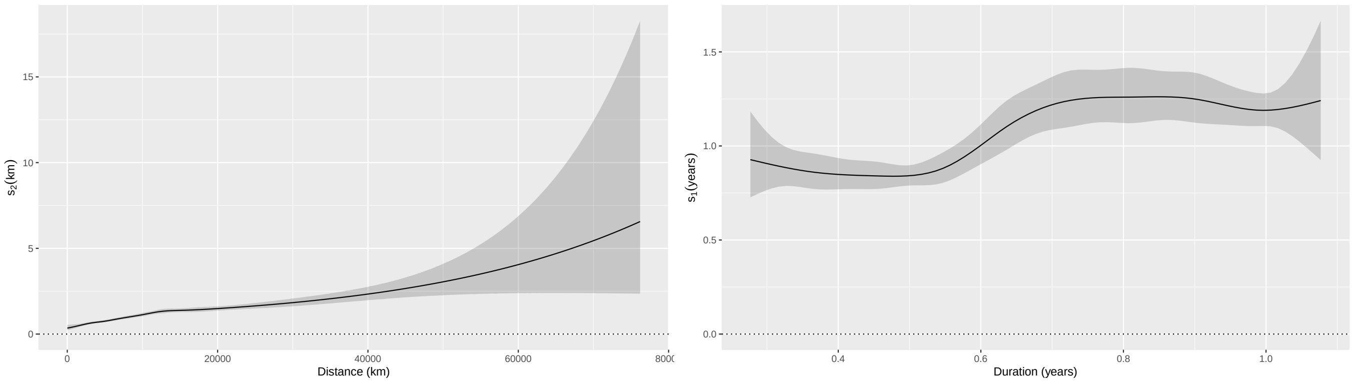

Figure 4.6: Cubic regression splines without monotonicity constraints

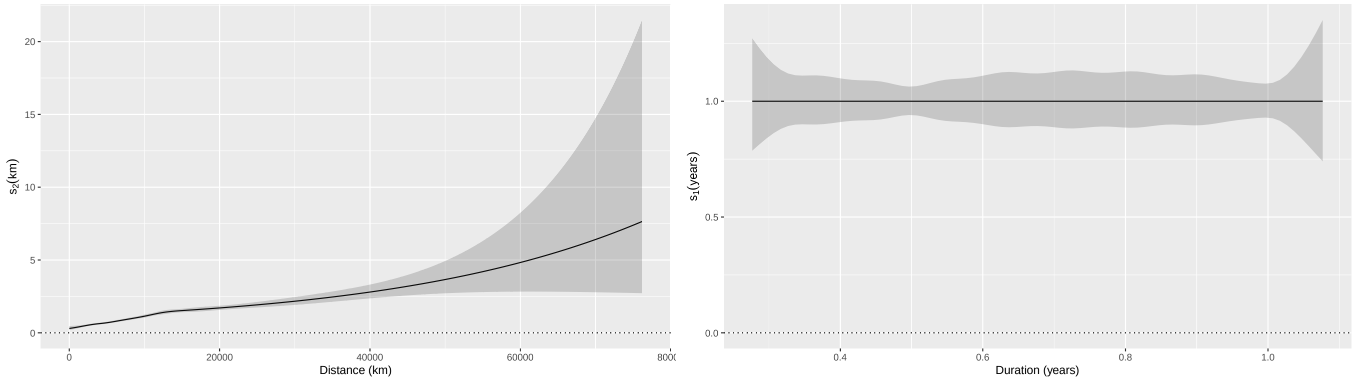

Figure 4.7: Cubic regression splines with monotonicity constraints

TPRS. Figures 4.8 and 4.9 present the estimation results. It can be noted that the unconstrained splines are essentially increasing, except for a slight decrease around the highest quantiles in the right-hand graph of Figure 4.8. There is a degree of stabilization in the unconstrained model for duration, which is corrected by the constrained model, without completely flattening the curve as in the two previous cases. Interestingly, rather than flattening the duration spline, the constraint makes it linear instead. Based on splines alone, this model could be justified for pricing. TPRS are not influenced by node location, since it is not necessary to define nodes as with the other two methods. TPRS basis are also not defined locally. Taken together, these features could explain why constrained TPRS models differ from the other two methods studied.

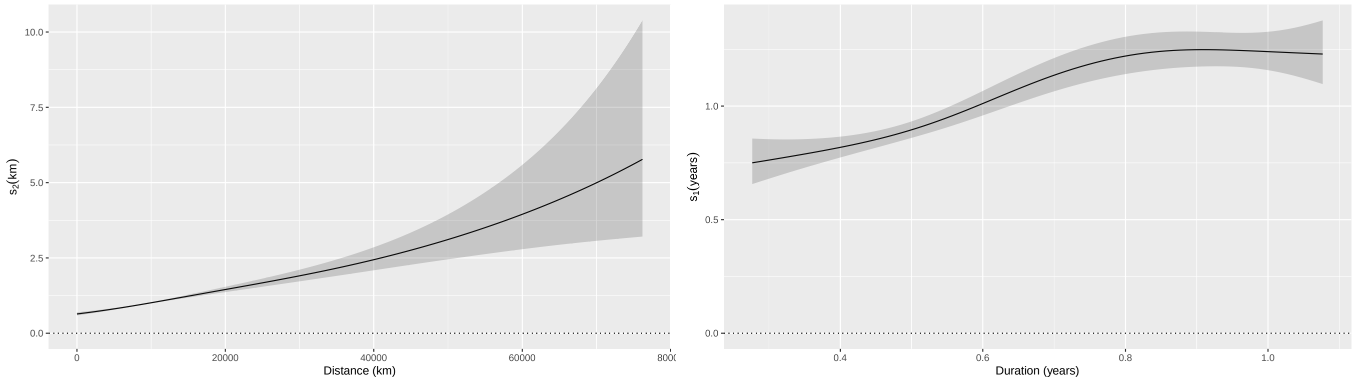

Figure 4.8: TPRS without monotonicity constraints

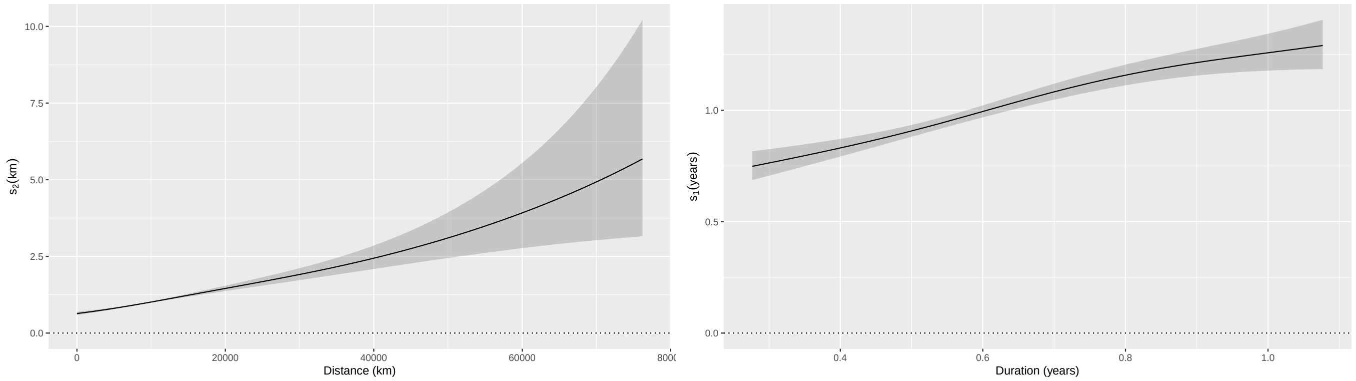

Figure 4.9: TPRS with monotonicity constraints

4.2.1 Additional remarks

Finally, one could mention some details on the different models estimation. First, the constrained duration curves are practically flat, but not completely. The right panel graph is enlarged in Figure 4.10 by removing the estimated confidence intervals. Looking at the values of the y-axis, the spline only marginally influences the mean of the estimate. Decreasing segments can also be obtained if the difference in y-axis values is below the error margin of the algorithm. In addition, the dimension of splines considerably affects the estimation time of models with monotonicity constraints. Up to nine knots, estimation times were reasonable, but could be counted in minutes or hours depending on whether the model included constraints.

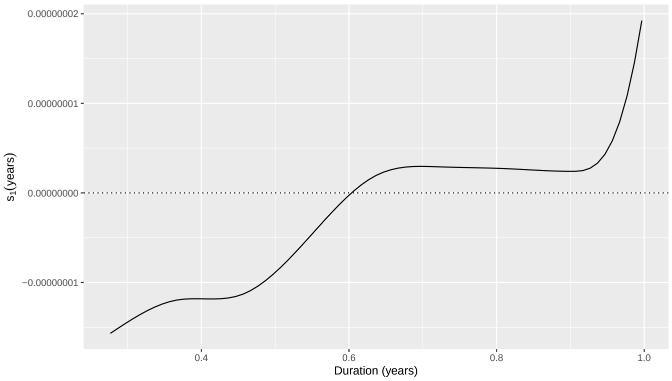

Figure 4.10: Cubic regression spline for duration with a monotonicity constraint

4.2.2 Goodness-of-fit Statistics

It remains to analyze the performance of these models on the data and to assess the cost of imposing shape constraints. Table 4.2 presents fit statistics on the in-sample dataset. We included the log-likelihood, AIC, and BIC to discuss the fit. The EDFs, or effective degrees of freedom, were included because they are used to determine the value of the AIC and BIC. The models all included five parametric terms, one for the intercept and four for the small set of covariates. The observed differences in the EDF values arise from the penalty imposed on the nonparametric terms (the splines) during estimation. We find that the unconstrained models are always superior to their constrained counterparts, which is understandable since the estimation procedure seeks to maximize the likelihood. If the optimal model were the one with monotone splines, it would be rendered by the unconstrained model. Despite the fact that the unconstrained TPRS model displays the worst log-likelihood among the unconstrained models, this spline has the best log-likelihood for the constrained models. Moreover, the gap between the model with and without constraints is less marked for the TPRS model. This result was intuitive considering Figures 4.4 to 4.9 where we observed that the curves with and without constraints were quite similar for the TPRS spline, unlike the other two splines. We also note that TPRS models penalize non-parametric terms more since their EDFs are lower than the other models. It follows form this that TPRS models obtain the best BIC scores, which penalize more for the number of parameters. Concerning strictly the models with constraints, the TPRS model obtains the best statistics whatever the criterion.

	Models with constraints			
Model	log-likelihood	AIC	BIC	EDF
P-spline	-10,663.59	21,367.50	21,544.46	20.17
Cubic reg. spline	-10,663.36	21,357.11	21,490.50	15.20
TPRS	-10,683.50	21,392.38	21,503.68	12.68
	Models with constraints			
Model	log-likelihood	AIC	BIC	EDF
P-spline	-10,694.21	21,422.04	21,569.53	16.81
Cubic reg. spline	-10,692.32	21,415.03	21,548.42	15.20
TPRS	-10,686.91	21,396.42	21,495.64	11.31

Table 4.2: In-sample statistics

Table 4.3 presents fit statistics on the test dataset $\boldsymbol{y}$ for all predictive distributions P estimated by the candidate models. Since these are discrete count data, not continuous data, appropriate statistics should be chosen for this type of data ([8]). Among the statistics selected are the logarithmic score and the Dawid-Sebastiani score (dss), as these two metrics make it possible to obtain significantly different scores between the MVNB models in order to separate them. The log score is equivalent to the log-likelihood calculated on the test data set. The Dawid-Sebastiani score represents the squared error normalized by the variance, to which a variance-based penalty is added so as not to systematically favor models estimating a higher variance. For an observation, we have $\text{dss}(P, y_i) = \left(\frac{y_i-\mu_p}{\sigma_p}\right)^2 + 2\log(\sigma_p)$ and the sum of the statistics for each of the observations in the test dataset is used to evaluate the final score. The table also includes the squared error since it is one of the most well-known and used error measures. Finally, the Poisson deviance is included, since the squared error corresponds to the normal deviance and a Poisson distribution is better suited to counting data. We can see that even on the test dataset, the statistics are better for the unconstrained models. We also observe that even among the unconstrained models, the P-spline model (Figure 4.4) obtains the best statistics, even though it shows a decrease in the mileage curve. These results therefore point in the direction that such models, to be optimal, should exhibit a decrease when considering duration and distance independently. The question then becomes, can we afford a slight deviation to allow for logical pricing? On the test dataset, the differences between the models with and without constraints are less significant than on the fitting dataset. Another thought to consider is the inclusion of splines associated with mileage and duration independently. This may not exactly reflect reality, but a multivariate version of exposure would be difficult to apply in practice ([4]). Other avenues in this direction could however be explored, since [4] were limited to a tensor product of P-splines.

		Models without constraints		
Model	Logarithmic	Dawid-Sebastiani	Poisson deviance	Squared error
B-spline	-2,645.45	-22,479.25	5,236.36	740.15
Cubic reg. spline	-2,645.51	-22,474.16	5,236.53	740.11
TPRS	-2,649.68	-22,320.39	5,245.18	740.57
		Models with constraints		
Model	Logarithmic	Dawid-Sebastiani	Poisson deviance	Squared error
B-spline	-2,647.52	-22,456.60	5,241.04	740.73
Cubic reg. spline	-2,647.04	-22,466.67	5,240.07	740.54
TPRS	-2,649.90	-22,321.90	5,245.67	740.68

Table 4.3: Fit statistics on test data set

4.3 Analysis of premium structure

This subsection presents premiums calculated with all six studied models. Table 4.4 contains the parametric term estimates for all models. Premiums, corresponding to the expectation of the distribution, were calculated on the basis of an average risk profile, i.e. a married woman using her car to go to work. Premiums were calculated for different durations (30%, 50%, 75% and 100% of the year) and different distances travelled (1000, 2000, 5000, 10000, 15000, 20000, 25000, 30000, 40000, 50000, 60000 and 70000 kilometers). We also study how the premium evolves according to different claims histories: without history (or a priori premium), a two-year claims-free history (or favorable predictive premium) and a history with two claims in two years (or unfavorable predictive premium).

Model	B-splines		Cubic reg. splines		TPRS	
	Unconstrained	Monotone	Unconstrained	Monotone	Unconstrained	Monotone
β_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)
β_{gender}	-0.030 (0.038)	-0.046 (0.038)	-0.030 (0.038)	-0.046 (0.038)	-0.030 (0.038)	-0.030 (0.038)
$\beta_{marital}$	0.141 (0.040)	0.147 (0.040)	0.141 (0.040)	0.147 (0.040)	0.142 (0.040)	0.143 (0.040)
$\beta_{use(Work)}$	0.715 (0.261)	0.783 (0.261)	0.718 (0.261)	0.781 (0.261)	0.766 (0.261)	0.772 (0.261)
$\beta_{use(Other)}$	0.693 (0.261)	0.768 (0.262)	0.696 (0.261)	0.768 (0.262)	0.704 (0.261)	0.709 (0.261)
ν	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 4.4: Parametric term estimates

4.3.1 A priori premiums

In the Appendix, Tables A.1, A.2 and A.3 present the premiums for models with and without constraints integrating respectively P-splines, cubic regression splines and TPRS. As could be observed in Figures 4.5 and 4.7, the premium does not vary according to contract duration in the monotone P-splines and cubic splines models. For these same splines, but without constraints, we observe irregularities, such as a decrease between kilometers $60,000$ and $70,000$, between durations 0.3 and 0.5, and between durations 0.75 and 1. As for the model with TPRS splines (Table A.3), we obtain a coherent premium structure, regardless of whether there are monotonicity constraints or not. The two structures are very similar as suggested by Figures 4.8 and 4.9.

Figure 4.11 illustrates the evolution of a priori premiums as a function of mileage for the different constrained splines studied. The duration influences the premium for the TPRS models. Thus, a curve for duration 0.5 and duration 1 are illustrated (broken lines). A unique curve is displayed for P-splines and cubic regression splines because duration does not impact the a priori premium. We observe that P-splines and cubic splines premium values are similar up to about 35,000 kilometers, which represents the majority of the portfolio (Figure 4.3). From there, the model with P-splines is more expensive, until around 58,000 kilometers, when the model with cubic splines increases more rapidly.

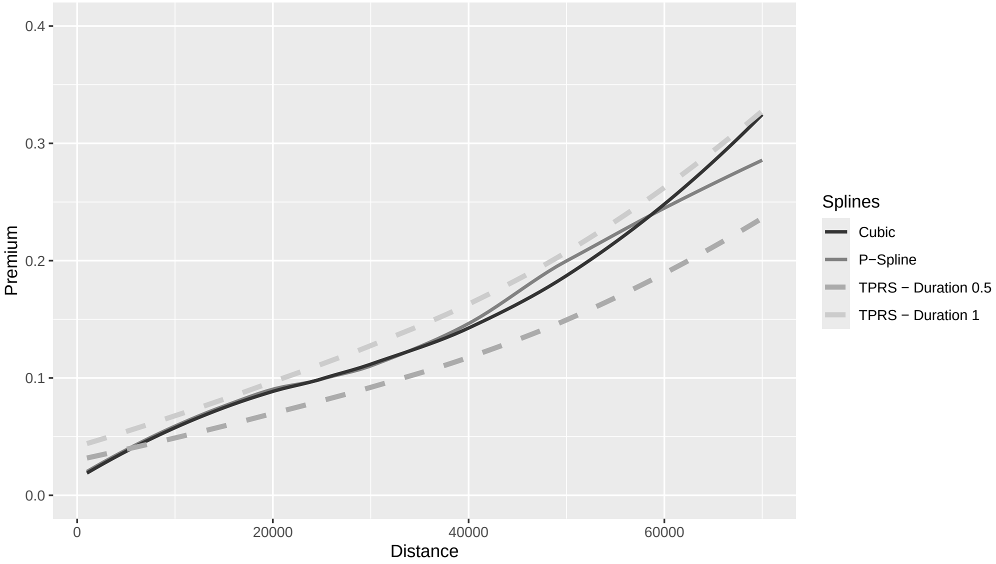

Figure 4.11: A priori premium based on mileage

Premiums for the TPRS model are generally higher than for the other two splines. In TPRS models, the premium varies with the duration of the contract, but not proportionally. Even considering that car use is the main source of risk exposure for collision coverage, it may make sense to increase the premium with the duration of the policy, since even a parked vehicle can be hit or hit-and-run.

Interestingly, the gap between the two duration curves of the TPRS splines is increasingly large as the distance increases. This implies that the model increases the premium more quickly when the kilometers are driven over the entire year rather than over a portion of the year. One would expect that driving a large number of kilometers over a shorter period of time would have a negative (upward) impact on the premium. However, if a large number of kilometers were driven in a short period of time, one can assume that freeways were used more. As mentioned earlier, these roads are considered less prone to collisions compared to city driving. If few kilometers were driven, for example 5,000 kilometers, the risk profiles between 6 months and one year are possibly more similar: a large proportion of kilometers driven in the city.

4.3.2 Predictive premiums

Tables A.4, A.5 and A.6 in the Appendix show how the premiums evolve after two years without claims for each of the models with and without constraints. The premium reduction is more tangible for the longer distances traveled. Indeed, there is little or no movement for the shortest distances. The a priori premium varies between 0.035 and 0.054 for 5,000 kilometers, while it varies from 0.035 to 0.053 for the same distance according to the favorable scenario after two years. If we considered 70,000 kilometers instead, the a priori premium would then vary from 0.237 to 0.328 according to the spline, but only from 0.223 to 0.302 according to a favorable scenario.

If we analyze by duration, there is not really a difference between the a priori and predictive premiums. Considering that the mileage is the main source of exposure to the risk of collision, it is reasonable not to reduce the premium just because the contract is active. For an insured driving only $1,000$ km per year, it is legitimate to wonder whether the risk has been sufficiently observed after two years to grant them a discount. Compared to the a priori premium, a reduction appears from approximately $10,000$ km.

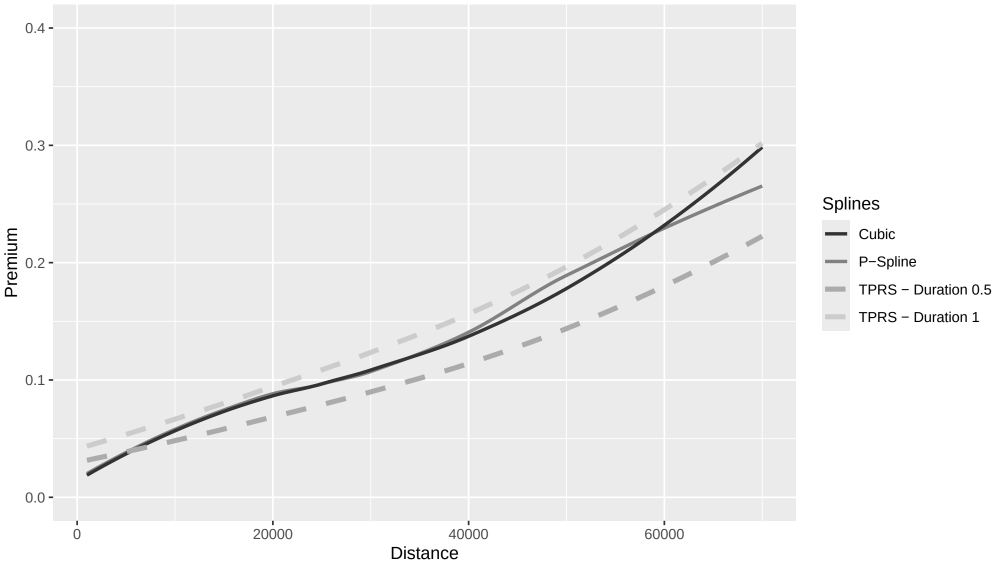

Figure 4.12: Evolution of the predictive premium according to mileage (favorable scenario)

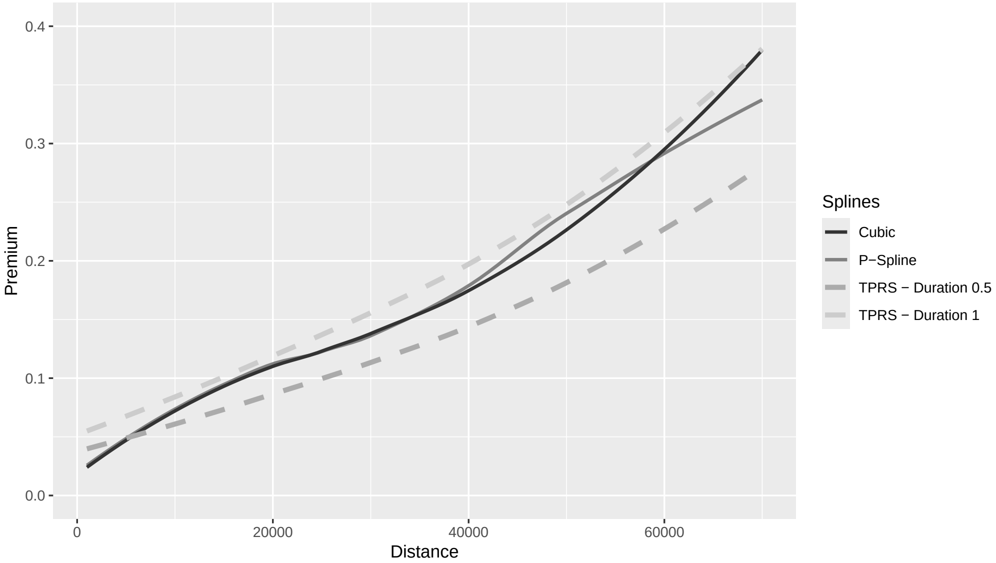

Figure 4.13: Evolution of the predictive premium according to mileage (unfavorable scenario)

Tables A.7 to A.9 present the predictive premiums for an unfavorable history (two claims after two years of observation). Compared with the a priori premiums, we see that the level of premiums is higher for both longer durations and longer distances. This characteristic may be desired to the extent that an insurer does not want to retain a bad risk. Figures 4.12 and 4.13 provide a visual summary of the predictive premiums according to the two scenarios considered. We observe that the shape of the splines is similar, but that the curves are translated upwards in the case of an unfavorable scenario.

We mentioned that the level of premiums is higher for both longer durations and longer distances in the unfavourable scenario, while the favourable scenario only grants a reduction for longer distances. These results are consistent since if a driver is insured for a short duration, the model predicts a low probability of claim. If the driver has, in fact, not claimed after two years, the model does not reduce the premium, as this was the expected result. Conversely, if the driver claims when the model estimated the probability to be low, the premium increases. By analyzing predictive premium Equation 3.4, we note that past claims ($n_{i,\bullet} = \sum_{t=1}^{T_i} n_{i,t}$) will affect the premium level more if the expectations of past years ($\lambda_{i,\bullet} = \sum_{t=1}^{T_i} \lambda_{i,t}$) are lower.

4.3.3 Impact of past claims

Finally, to illustrate the impact of past claims on the predictive premium, a priori premiums and premiums according to the favorable and unfavorable scenarios are shown in Figure 4.14 for the TPRS model. We note similarities between the a priori premiums and according to the favorable scenario in the first portion of the domain. Considering that the majority of the portfolio travels less than 30,000 kilometers, the reduction granted by the model is gradual and conservative. We also note that the unfavorable scenario curve moves further and further away from the a priori curve as distance increases. This is a desirable feature, since if the model considers a policyholder to be a bad risk, they will be penalized more and more the further they travels, thus encouraging the policyholder to reduce their risk exposure.

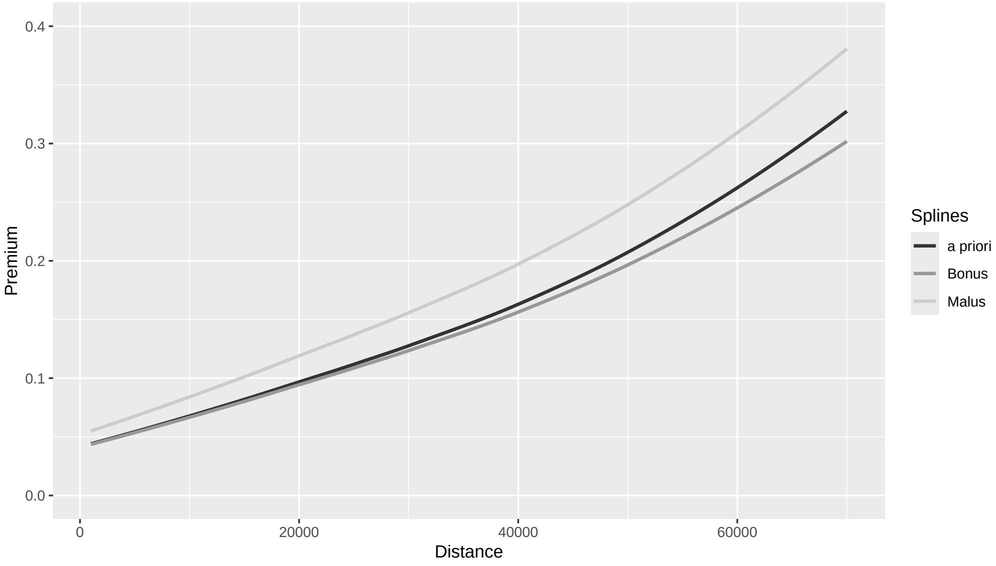

Figure 4.14: Evolution of the TPRS premium based on claims history

5 Conclusion

In this research work we explored the relationship between distance traveled and risk using different smoothing functions. The commonly used risk exposure measure, contract duration, was also included independently in the pricing models. Unlike earlier works in this line of study, a wider range of splines was considered to comparatively analyze the results. In addition to B-splines/P-splines, the smoothing functions that were used were cubic regression splines and thin plate regression splines with and without constraints to ensure monotonicity. We presented different methods to model a monotone function, in addition to exploring the characteristics of each spline. We concluded that using a monotone spline was done at a certain cost in terms of fit, but that the differences were less significant on the statistics from the test data set. One limitation addressed in this study concerns the data. We had mentioned that the portfolio was composed largely of young people, a clientele generally less loyal than their elders. It is therefore more difficult to collect a long history to fit a longitudinal model. However, the data obtained have the merit of being reliable since they were collected with a device installed in the vehicle rather than collected by a mobile application that can be forgotten to be activated. A next step would be to explore the dependence between duration and distance using, for example, multivariate splines. Indeed, one can reasonably assume that duration and distance may be correlated since policies with higher distances traveled are usually associated with longer policy durations. In conclusion, we can emphasize that when mileage measured by a GPS device is included in the pricing (rather than the insured's self-declaration), gender is not a significant covariate for any of the models (see Table 4.4). Marital status remains significant, however, so avenues other than mileage should be explored to replace this discriminating variable.

Acknowledgement

The author warmly thanks Jean-Philippe Boucher for his insightful comments, which significantly contributed to improving the quality of this manuscript. The author gratefully acknowledges the financial support provided by Fonds de recherche du Québec - Nature et technologies under grant 269533.

Statements and Declarations

No financial or non-financial interest to declare.

A A priori premiums

(a) Unconstrained

	Duration			
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
15000	0.074	0.059	0.089	0.084
20000	0.080	0.064	0.097	0.091
25000	0.086	0.068	0.104	0.098
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

	Duration			
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
15000	0.078	0.078	0.078	0.078
20000	0.091	0.091	0.091	0.091
25000	0.098	0.098	0.098	0.098
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 A.1: A priori premiums - P-splines

(a) Unconstrained

	Duration			
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
15000	0.064	0.058	0.086	0.080
20000	0.069	0.063	0.093	0.086
25000	0.076	0.069	0.102	0.093
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

	Duration			
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
15000	0.078	0.078	0.078	0.078
20000	0.087	0.087	0.087	0.087
25000	0.098	0.098	0.098	0.098
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 A.2: A priori premiums - Cubic regression splines

(a) Unconstrained

	Duration			
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
15000	0.050	0.058	0.077	0.081
20000	0.058	0.069	0.091	0.095
25000	0.067	0.079	0.105	0.110
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

	Duration			
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
15000	0.050	0.059	0.073	0.082
20000	0.059	0.070	0.086	0.097
25000	0.068	0.081	0.100	0.112
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 A.3: A priori premiums - TPRS

Predictive premiums (favorable scenario)

(a) Unconstrained

	Duration			
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
15000	0.073	0.058	0.087	0.083
20000	0.079	0.063	0.094	0.090
25000	0.085	0.067	0.101	0.096
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

	Duration			
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
15000	0.076	0.076	0.076	0.076
20000	0.089	0.089	0.089	0.089
25000	0.096	0.096	0.096	0.096
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 A.4: Predictive premiums (2 years without claims) - P-splines

(a) Unconstrained

	Duration			
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
15000	0.063	0.058	0.085	0.079
20000	0.068	0.062	0.091	0.084
25000	0.075	0.068	0.099	0.091
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

	Duration			
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
15000	0.076	0.076	0.076	0.076
20000	0.085	0.085	0.085	0.085
25000	0.095	0.095	0.095	0.095
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 A.5: Predictive premiums (2 years without claims) - Cubic regression splines

(a) Unconstrained

	Duration			
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
15000	0.049	0.057	0.076	0.079
20000	0.058	0.068	0.089	0.093
25000	0.066	0.078	0.102	0.107
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

	Duration			
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
15000	0.049	0.058	0.072	0.080
20000	0.058	0.069	0.085	0.095
25000	0.067	0.079	0.097	0.109
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 A.6: Predictive premiums (2 years without claims) - TPRS

Predictive premiums (Adverse scenario)

(a) Unconstrained

	Duration			
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
15000	0.090	0.071	0.108	0.102
20000	0.097	0.077	0.116	0.110
25000	0.104	0.083	0.125	0.118
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

	Duration			
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
15000	0.097	0.097	0.097	0.097
20000	0.113	0.113	0.113	0.113
25000	0.122	0.122	0.122	0.122
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 A.7: Predictive premiums (two claims) - P-splines

(a) Unconstrained

	Duration			
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
15000	0.078	0.071	0.105	0.097
20000	0.084	0.076	0.112	0.104
25000	0.092	0.084	0.123	0.113
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

	Duration			
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
15000	0.097	0.097	0.097	0.097
20000	0.108	0.108	0.108	0.108
25000	0.121	0.121	0.121	0.121
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 A.8: Predictive premiums (two claims) - Cubic regression splines

(a) Unconstrained

	Duration			
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
15000	0.062	0.072	0.095	0.099
20000	0.072	0.085	0.112	0.117
25000	0.083	0.098	0.128	0.134
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

	Duration			
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
15000	0.062	0.074	0.091	0.101
20000	0.073	0.087	0.107	0.119
25000	0.084	0.100	0.123	0.137
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 A.9: Predictive premiums (two claims) - TPRS

References

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

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

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

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

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

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

[7] H. Brunk, R. E. Barlow, D. J. Bartholomew, and J. M. Bremner. Statistical inference under order restrictions.(the theory and application of isotonic regression). Technical report, Missouri Univ Columbia Dept of Statistics, 1972.

[8] C. Czado, T. Gneiting, and L. Held. Predictive model assessment for count data. Biometrics, 65(4):1254–1261, 2009.

[9] C. De Boor. A practical guide to splines, volume 27. springer-verlag New York, 1978.

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

[11] J. Duchon. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. 1977.

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

[13] M. Eling and M. Kraft. The impact of telematics on the insurability of risks. The Journal of Risk Finance, 21(2):77–109, 2020.

[14] J. Friedman and R. Tibshirani. The monotone smoothing of scatterplots. Technometrics, 26(3):243–250, 1984.

[15] F. N. Fritsch and R. E. Carlson. Monotone piecewise cubic interpolation. SIAM Journal on Numerical Analysis, 17(2):238–246, 1980.

[16] G. Gao and M. V. Wüthrich. Feature extraction from telematics car driving heatmaps. European Actuarial Journal, 8(2):383–406, 2018.

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

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

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

[20] X. He and P. Shi. Monotone b-spline smoothing. Journal of the American statistical Association, 93(442):643–650, 1998.

[21] G. Heller, M. Stasinopoulos, B. Rigby, et al. 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, 2006.

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

[23] H. Jeong. Dimension reduction techniques for summarized telematics data. The Journal of Risk Management, Forthcoming, 2022.

[24] C. Kelly and J. Rice. Monotone smoothing with application to dose-response curves and the assessment of synergism. Biometrics, pages 1071–1085, 1990.

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

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

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

[28] E. Mammen. Estimating a smooth monotone regression function. The Annals of Statistics, pages 724–740, 1991.

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

[30] H. Mukerjee. Monotone nonparametric regression. The Annals of Statistics, pages 741–750, 1988.

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

[32] N. Pya and S. N. Wood. Shape constrained additive models. Statistics and computing, 25:543–559, 2015.

[33] J. O. Ramsay. Monotone regression splines in action. Statistical science, pages 425–441, 1988.

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

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

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

[37] L. Schumaker. Spline functions: basic theory. Cambridge university press, 2007.

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

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

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

[41] R. Turcotte and J.-P. Boucher. Gamlss for longitudinal multivariate claim count models. 2022.

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

[43] R. Verbelen, K. Antonio, and G. Claeskens. 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, 2018.

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

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

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

[47] S. N. Wood. Generalized additive models: an introduction with R. Chapman and Hall/CRC, 2017.

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

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