mathematics

MDPI

Article

Tariff Analysis in Automobile Insurance: Is It Time to Switch from Generalized Linear Models to Generalized Additive Models?

Zuleyka Díaz Martínez [1,†], José Fernández Menéndez [2,†] and Luis Javier García Villalba [3,*,†]

[1] Group of Analysis, Security and Systems (GASS), Department of Financial and Actuarial Economics & Statistics, Faculty of Economics and Business Administration, Universidad Complutense de Madrid (UCM), Campus Somosaguas, 28223 Madrid, Spain; zuleyka@ccee.ucm.es

[2] Department of Business Administration, Faculty of Economics and Business Administration, Universidad Complutense de Madrid (UCM), Campus Somosaguas, 28223 Madrid, Spain; jfernan@ccee.ucm.es

[3] Group of Analysis, Security and Systems (GASS), Department of Software Engineering and Artificial Intelligence (DISIA), Faculty of Computer Science and Engineering, Office 431, Universidad Complutense de Madrid (UCM), Calle Profesor José García Santesmases, 9, Ciudad Universitaria, 28040 Madrid, Spain

[*] Correspondence: javiergv@fdi.ucm.es; Tel.: +34-91-394-7638

[†] These authors contributed equally to this work.

Abstract: Generalized Linear Models (GLMs) are the standard tool used for pricing in the field of automobile insurance. Generalized Additive Models (GAMs) are more complex and computationally intensive but allow taking into account nonlinear effects without the need to discretize the explanatory variables. In addition, they fit perfectly into the mental framework shared by actuaries and are easier to use and interpret than machine learning models, such as trees or neural networks. This work compares both the GLM and GAM approaches, using a wide sample of policies to assess their differences in terms of quality of predictions, complexity of use, and time of execution. The results show that GAMs are a powerful alternative to GLMs, particularly when "big data" implementations of GAMs are used.

Keywords: automobile insurance; generalized additive models; splines; tariff analysis

MSC: 97M30

check for updates

Citation: Díaz Martínez, Z.; Fernández Menéndez, J.; García Villalba, L.J. Tariff Analysis in Automobile Insurance: Is It Time to Switch from Generalized Linear Models to Generalized Additive Models? *Mathematics* **2023**, *11*, 3906. https://doi.org/10.3390/math11183906

Academic Editors: Jing Yao, Xiang Hu and Jingchao Li

Received: 16 August 2023
Revised: 8 September 2023
Accepted: 11 September 2023
Published: 14 September 2023

1. Introduction: The Classical Tariff Analysis in the Actuarial Field

Since their original development by J. A. Nelder and R. W. M. Wedderburn [1], Generalized Linear Models (GLMs) have become a key methodology in applied statistics. GLMs burst onto the insurance scene in the early 1990s in works such as [2,3] and became part of the set of tools commonly used by actuaries, with a number of monographs devoted to showing their application, as well as those of their extensions, to the field of insurance [4–6].

The GLM provides a common framework that includes a broad range of regression models (ordinary least squares, logistic, Poisson, etc.) previously lacking a unified treatment. The essential requirement for a particular model that can be treated as a GLM is that the probability distribution of the variable analyzed belongs to the class of the exponential family [7,8]. In a GLM, the expectation of the dependent variable, $\mu = E(y)$, is modeled as a function of the linear predictor $\eta = \sum \beta_i x_i$, which includes the explanatory variables x_i. Therefore, we have that $\mu = l^{-1}(\eta)$ or, conversely, that $\eta = l(\mu)$. The function $l()$ is called a link function and can be selected with a certain degree of flexibility [9]. The variance $var(y)$ is customarily expressed in the GLM setting as a function of the mean in the following manner: $var(y) = \frac{\varphi}{A} V(\mu)$. In this expression, $V(\mu)$ is the so-called variance function, φ is the dispersion parameter or dispersion for short, and A is the known weight of each observation.

For instance, in the case of a Poisson regression model, the dependent variable, y, is the number of occurrences of an event in a time interval. By defining $\mu = E(y) = \lambda$ and taking, as is usually the case, a log link, the model becomes $\log(\lambda) = \sum \beta_i x_i$. Also, if the weights are $A = 1$, and because of the equality of mean and variance in a Poisson-distributed random variable, we obtain $var(y) = \lambda = \mu = \varphi V(\mu)$, which implies that the dispersion parameter is $\varphi = 1$ and the variance function is $V(\mu) = \mu$.

In any case, the explanatory variables enter the model through the linear predictor $\eta = \sum \beta_i x_i$, and therefore their effects are linear, making GLMs "linear" models. A classical ordinary least-squares model (OLS) is a GLM with $V(\mu) = 1$ as the variance function and the identity as the link function. This means that the parameter modeled is the mean of the dependent variable, μ, as a linear function of the covariates $x_i : \mu = \sum \beta_i x_i$. If x_i are continuous, their eventual nonlinear effects can be included in an OLS model, adding to the linear predictor of the corresponding quadratic, cubic, etc., terms, so that the linear predictor has the form $\eta = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \ldots$, or, likewise, in the case of a GLM, $l(\eta) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \ldots$.

Another, simpler way to introduce nonlinear effects into a GLM is by discretizing the continuous variables x_i.

The discretization implies an information loss regarding the original variable and requires the determination of the number of categories and the intervals that define the bins, which can be complex and subject to some degree of arbitrariness. Nevertheless, the discretization of an explanatory variable can help in applying and interpreting the model and provides a straightforward procedure to take into account an eventual nonlinear effect of the variable. Also, the discretization of the variable increases the number of parameters to be estimated. For a categorical variable with k levels, $k - 1$ parameters should be estimated. This provides enough flexibility for the model to be capable of including nonlinearities.

In the classical tariff analysis in the automobile insurance line of business, the customary practice consists of discretizing all the continuous explanatory variables used to estimate the risk associated with a policy (like the driver's age, horsepower of the vehicle, etc.) and generating, through the Cartesian product of the levels of these discretized variables, a set of "cells" or "policy groups". The cell a policy belongs to determines the premium that is charged. This simplifies the tariff application (the calculus of the premium of a new policy merely consists of identifying the cell that corresponds to it). The pure premium of a cell is usually estimated as the product of the claim frequency and the claim severity associated with the cell. Both the claim frequency and claim severity share the same structure, and they are quotients between a risk measure and a risk exposure [10]. The claim frequency is the ratio between the total number of claims in a cell and the total exposure of the cell, which is the total time that the policies in the cell have been in force. The claim severity is the mean claim size, i.e., the quotient between the total size of the claims in a cell and the number of claims originated by the policies in the cell.

The claim frequency is usually modeled as a rate model, i.e., a Poisson regression model where the number of claims in each cell is the dependent variable and the exposure (in this case, the total time the policies are in force) is taken as an offset [2,10], i.e., an explanatory variable whose coefficient is not estimated but taken as 1 [7].

With regard to the claim severity, as this is a continuous and positive variable, it is customarily modeled as the dependent variable in a Gamma regression in which the exposure (number of claims) is introduced as a weight for each observation, i.e., for each cell in this case [2,10]. Usually, both for the claim frequency and the claim severity, the link function of the regression model is the logarithm. This means that the risk and the associated premium corresponding to a cell can be expressed as the product of a factor specific to the cell multiplied by the risk of the cell taken as a baseline or reference. In the motor insurance field, this multiplicative effects model is generally considered more suitable for calculating premiums than an additive effects model [2,10].

These kinds of models are widely used for the claim frequency and the claim severity and can be considered standard in the automobile insurance field. A drawback is the need to determine how to discretize the explanatory variables, which can be somewhat arbitrary and may also require a huge amount of trial and error to find a discretization that groups the policies in sufficiently homogenous cells, but at the same time keeps low enough the number of cells so that the number of parameters in the model does not become unmanageable. Another problem is that the within-cell variability neglected by the model can be appreciable unless the cell size becomes very small and the number of cells becomes very high.

There exists, however, the possibility of working with the original non-discretized continuous variables. The main problem with this approach is how to treat nonlinear effects, for example, that of the driver's age, which does not show a linear relationship with the claim risk.

The most traditional way is to include polynomial terms (quadratic, cubic, etc.) as regressors to deal with the nonlinearities of the data, but this is far from being satisfactory because polynomials, in a regression context, tends to cause problems [11–14]. These problems are a consequence of the fact that polynomial regression is an ill-conditioned problem because its matrix model includes a Vandermonde matrix, and it is well-known that Vandermonde matrices have a very high condition number and pose serious problems of numerical instability [15–18]. Hence, polynomial regression should be used with caution, and the order of the polynomials should be kept as low as possible.

However, there are more elaborate alternatives that can solve these problems, which are based on the use of splines with Generalized Additive Models (GAM). The GAM [19] is a generalization of the GLM that introduces the explanatory variables in the linear predictor, not directly but through a spline of the variable, i.e., through a continuous piecewise polynomial of the variable that allows for the modeling of nonlinearities, thereby eliminating the problems associated with the simpler and more conventional polynomial terms. These models have been suggested as a reasonable alternative to the traditional pricing models based on discretized risk factors [10,20]. Instead of laboriously and somewhat arbitrarily calculating the values according to which each variable will be divided into a series of levels, GAMs allow this to be done automatically and maintain the traditional structure of regression models, which is easy to understand and interpret. For this reason, the aim of this work is to compare a classical model of tariff analysis in the automobile insurance field estimated using a GLM and discretized explanatory variables to a similar model estimated using a GAM, with its explanatory variables retaining their original, continuous, and non-discretized form. Our goal is to assess the improvement achieved by the greater amount of information supplied by the continuous variables and evaluate if this improvement exceeds the greater complexity, both conceptual and computational, of the GAM model.

Another kind of method widely used for tariff analysis in automobile insurance is de Bonus–Malus Systems (BMSs). A BMS includes a posteriori information about the track record of each policy or each policyholder in the portfolio. With this information, the policies are classified into different risk levels. Some a priori rating variables, like age, cubic capacity of the car, etc., can also be included in the model. Examples of this approach can be found in [21,22]. Nevertheless, we do not consider the Bonus–Malus Systems here.

The main finding of this work is that a Generalized Additive Model is a tool for calculating premiums in the field of automobile insurance that can take into account the nonlinear effect of some of the pricing variables in an automatic way and without the need to identify the intervals in which to divide the values of the rating factors, similar to a traditional GLM.

The rest of the work is organized as follows. Section 2 briefly reviews the scientific literature on GAM usage in several fields. Section 3 provides an overview of the theoretical foundations of Generalized Additive Models. Section 4 compares the results obtained using GAMs and GLMs for motor insurance ratings with a large sample from a Spanish insurance company. Finally, Section 5 presents the main results obtained in this work.

2. Literature Review

One possibility for addressing nonlinearities in data is to use modern methods from the field of machine learning and neural networks. For example, an early attempt by Mulquiney [23] to compare GLMs to MARS (Multivariate Adaptive Regression Splines) [24], MARTs (Multiple Additive Regression Trees) [25], and neural networks achieved mixed results but ultimately favored GLMs. Furthermore, a comparison between XGBoost and GLMs (logistic regression) for predicting motor insurance claims in [26] also favored GLMs. In contrast, Henckaerts and colleagues [27] found that carefully crafted Boosted Trees outperformed classical GLMs for insurance pricing. Ticconi [28] compared GLMs to neural networks and Support Vector Machines (SVMs) for credit insurance analysis, with the results demonstrating the superiority of SVMs. Baillargeon and colleagues [29] used a Hierarchical Attention Network (a type of neural network for document classification) to analyze risk factors from a textual dataset describing car accidents, using a GLM as a baseline for comparison. Delong and Wüthrich [30] trained a neural network to develop models for the process of claim payments and claims incurred for property or bodily injury. They used GLMs, Generalized Additive Models, and Regression Trees as starting models to feed into the neural network. A good overview of other applications of machine learning tools to actuarial science can be found in [31].

As GAMs are particularly suitable for modeling nonlinear relationships between variables, they have been widely and routinely used in several research fields since their inception. For example, in the field of biology, they have been used to analyze the spatial distribution of fishes and vegetal species [32–40]. Comparisons of the relative performance of models estimated using GLMs or GAMs can be found in [41–45]. GAMs have also been frequently combined or compared to GLMs in environmental sciences [46–49], climatology [50], meteorology [51], neuroscience [52,53], and genetics [54]. In the actuarial literature, some works have applied the GAM methodology to various problems in the insurance field [55–60].

3. Generalized Additive Models

A Generalized Additive Model (GAM) is very similar to a GLM but with the key difference that the explanatory variables are introduced in the model through an "additive" predictor of the form $\eta = \beta_0 + f_1(x_1) + \ldots + f_p(x_p)$. This predictor replaces the linear predictor of the form $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$, typical of a GLM. The result is that some smooth functions $f_k(x_k)$ of the independent variables replace the simpler linear terms of the form $\beta_k x_k$, typical of a GLM. The functions $f_k()$ can be selected from different function spaces, provided that they are flexible enough to approximate any functional shape; however, usually, they are some type of spline or kernel smoother [13,14,19]. Also, the fact that each explanatory variable enters the model through a specific additive term in the predictor, enables individualizing and clearly analyzing its effect on the response, independently of the other variables. Efficient algorithms for estimating these models can be found in [19].

In each model, the specific peculiarities of the functions $f_k()$ are evaluated as part of the fitting process. The model loses the "parametric" character of the GLM and becomes what is usually termed a "semiparametric" model [61]. These semiparametric models are a good compromise between a fully parametric model (simpler to fit and interpret but less flexible in incorporating nonlinear effects) and a totally nonparametric one, where the functions $f_k()$ are a priori entirely arbitrary. A nonparametric model has enormous flexibility but is very difficult to fit and interpret [62]. In the semiparametric case, the functions $f_k()$ are usually constructed, as mentioned previously, using some kind of kernel smoother or, more frequently, splines.

A kernel smoother enables the implementation of "local" regression: the estimated value in a point x_0 is calculated using the observed values at points close to it. To do this, a weight is assigned to every point x. The weight is determined by a kernel $K(x_0, x)$, and the closer x is to x_0, the greater the weight value. Different kinds of kernels are used in practice. Some good revisions of this kind of technique can be found in [63,64].

A spline is a linear combination of polynomials, each defined over a specific bounded interval and null outside of this interval. In this way, it is possible to address one of the key problems of polynomials when used in interpolation and estimation: their non-local character, i.e., the fact that they are defined over the entire real line, so small changes in their values in one region can result in big changes in other, very distant regions. The most frequently used splines are cubic ones, composed of third-order polynomials. Splines are widely used in statistics and numerical analysis, and there is a huge amount of literature devoted to them. A classical reference on the subject is [65].

The boundaries of the intervals in which the splines are defined are a set of points termed knots. Given a set of knots $\xi_1 \leq \xi_2 \ldots \leq \xi_n$ belonging to an interval $[a,b]$, a cubic spline is a function based on a set of cubic polynomials, each one defined in one of the intervals $(a,\xi_1),(\xi_1,\xi_2),\ldots,(\xi_n,b)$ and null in the rest of them. The spline is defined so that it is continuous, and with first- and second-order derivatives, it is also continuous at the knots [13]. With these conditions, it can be easily seen that given n knots, a cubic spline is defined by $n+4$ parameters. A natural cubic spline is a cubic spline that satisfies the four additional conditions that its second and third derivatives at the points a and b are zero. A natural cubic spline is defined by n parameters. In fact, given n knots in an interval $[a,b]$, the set of splines in this interval and with these knots forms a vector space of dimension $n+4$ in the case of cubic splines, or n in the case of natural cubic ones.

There are many different ways to express the basis of these vector spaces. One of them is the basis of truncated powers of the space of cubic splines [14]. This basis includes the four monomial terms 1, x, x^2, and x^3 plus a set of n (one for each knot) third powers of the positive part functions of the form $\left(X-\xi_j\right)_+^3$. The positive part function $\left(X-\xi_j\right)_+$ is defined as:

$$\begin{cases} 0, & \text{if } X-\xi_j < 0 \\ X-\xi_j, & \text{if } X-\xi_j > 0 \end{cases} \tag{1}$$

Independently of the basis selected, if we denote $s_j()$ as the functions of the basis of the cubic splines defined over an interval (and with a specified set of knots), we find that every smooth function defined over such an interval can be approximated with a linear combination of s_j. In this way, the functions $f_k()$ of the predictor of a GAM can be expressed as linear combinations of the functions of the basis $f_k(x_k) \approx \sum_j \beta_j s_j(x_k)$, and the estimation of $f_k()$ reduces to the estimation of the coefficients β_j of these linear combinations. If the GAM reduces to an OLS model (because the dependent variable is normally distributed, and the link function is the identity), β_j can be estimated using least squares by minimizing the quantity $\parallel Y-\sum \beta_j s_j(x) \parallel^2$ in the usual way. This is completely analogous to the minimization of $\parallel Y-\sum \beta_j x^j \parallel^2$ in the case of a polynomial regression, where the monomials x^j play the role of the $s_k(x)$ basis functions.

This approach generalizes straightforwardly when the GAM corresponds to a Generalized Linear Model (i.e., when the variable is not normal or the link function is not the identity). Thus, the estimation of a GAM would be similar to that of a GLM, with the only change being the replacement of x^j with the splines $s_k(x)$. These types of splines, with the knots fixed in advance, are often called regression splines [14]. When used, the complexity or roughness of the model is controlled by adding or removing knots. The use of regression splines is very simple because it is merely a slight generalization of a conventional regression model. However, they pose the problem that one has to select the number and location of the knots, and the results can be quite sensitive to this choice. Moreover, the models constructed by adding and removing knots are not nested, which makes it difficult to select the most appropriate model.

It is possible to eliminate the problem of selecting knots through regularization, i.e., by adding a quadratic additional term that penalizes the curvature of the estimated functions

f_k to the quantity to be minimized, $\parallel Y - \sum \beta_j s_j(x) \parallel^2$. Usually, this term is of the form $\int (f''(x))^2 dx$. In this case, a penalized regression spline is obtained, which controls the complexity or roughness of the model through the weight assigned to the quadratic penalization term: the more the weight, the lesser the roughness. In the case of penalized regression splines, it is still necessary to choose the knots, but this selection has little impact on the final results, provided that the number of knots selected is high enough to lead to a good fitting of the model.

An additional generalization is provided by smoothing splines. In this case, the set of knots is maximal: every available value of the independent variable for which there is an observation becomes a knot. Such a model has an exceedingly high number of parameters and perfectly interpolates all the observations. This model lacks any interest, so it is necessary to reduce its complexity through regularization. This is done by selecting a function f from a determined function space $\mathcal{H}$ that minimizes the sum of a term penalizing curvature, along with another term that, for example, in the case of an OLS model, represents the sum of squared errors:

$$\min_{f \in \mathcal{H}} \left\{ \sum (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx \right\} \tag{2}$$

The value of the smoothing parameter λ is chosen a priori and controls the amount of penalization of the curvature of f, so that the greater its value, the smoother the estimated model. If the function space $\mathcal{H}$ consists of differentiable functions with an absolutely continuous first derivative in an interval that contains the observations, it can be proven that the solution to the minimization problem above is a natural cubic spline with knots at the observations [13]. The function $f()$, the solution to the minimization problem, can thus be represented as a linear combination of the basis elements of the vector space of natural cubic splines, $f = \sum \beta_i s_i$. Therefore, the minimization problem becomes the following: $\min_\beta (y - S\beta)'(y - S\beta) + \lambda \beta' \Omega \beta$.

Here, we have that S is a matrix whose elements are $S_{ij} = s_j(x_i)$; Ω is a matrix of the elements $\Omega_{ij} = \int s_i'' s_j'' dx$; β is a vector whose elements are the parameters β_i that we want to estimate; and y is the vector with the observed values of the dependent variable. Formulated in this way, the problem is easily solved, and its solution is $\hat{\beta} = (S'S + \lambda\Omega)^{-1} S'y$ [13,14].

The conclusion is that the smoothing spline that best fits the data is the natural cubic spline with knots at the observations, which can be expressed as $\hat{f} = \sum \hat{\beta}_i s_i$. The values predicted by the model are $\hat{y} = S\hat{\beta} = S(S'S + \lambda\Omega)^{-1} S'y = Hy$. They are obtained as a linear transformation, Hy, of the observations. The matrix $H = S(S'S + \lambda\Omega)^{-1} S'$ plays a similar role to that of the hat matrix in an OLS regression. In OLS, the trace of the hat matrix provides the dimension of the linear subspace over which it projects the vector y of the observations. This dimension indicates the number of parameters, or degrees of freedom, of the model. Similarly, in the case of smoothing splines, the trace $tr(H)$ also indicates the "effective" degrees of freedom of the model. In general, they are not integer numbers, depend on the value of the smoothing parameter λ, and provide a measure of the complexity of the model. A high value of λ strongly penalizes the curvature of the spline f, which, therefore, tends to approach a straight line. Conversely, a low value of λ implies that the effective degrees of freedom of the model are high, and f becomes rougher, more complex, and "wrinkled". In fact, the complexity of the model can be controlled by selecting both the value of λ and the number of effective degrees of freedom.

The smoothing splines tend to be computationally expensive, but other than this, they have a good number of desirable properties and also easily generalize to higher dimensions. For example, in two dimensions, the term that penalizes the curvature of the spline becomes $\int \int \left[\left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x_1 \partial x_2} \right) + \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 \right] dx_1 dx_2$ (instead of $\int (f''(x))^2 dx$). These smoothing splines in two or more dimensions are usually termed thin-plate splines [66].

4. Comparing the Two Approaches

As mentioned, the aim of this work is to compare the traditional approach to tariff analysis in the automobile insurance field, i.e., the one based on GLMs with discretized explanatory variables—the rating factors—with the one based on GAMs, which allows for the introduction of nonlinearities in a very flexible way and without discretizing the rating factors. To do this, we used data from the automobile policy portfolio of a Spanish insurance firm for the year 2005. The data were slightly cleaned, discarding some extremely atypical or clearly erroneous values; otherwise, they were used as found in the insurer's database. From the full portfolio, five random training samples, each containing 500,000 policies, were extracted. With each of these samples, the GLMs and GAMs were estimated and compared. Also, five additional random samples of 200,000 policies were extracted and used as test samples. Each one of these test samples was paired with one of the training samples to check the predictive accuracy of the estimated models. Initially, two kinds of models were estimated: one for the number of claims using Poisson regressions, and one for the claim size using Gamma regressions. The rating factors used as the independent variables in the models are described in Table 1.

Table 1. Rating Factors.

Name	Variable	Characteristics
TYPE_VE	Type or category of vehicle	Categorical variable with 6 levels.
USAGE	Usage of the vehicle	Categorical variable with 20 levels.
NATURE	Nature of the vehicle	Categorical variable with 2 levels.
PLACES	Number of seats	Count variable; treated as categorical with 8 levels.
AMBIT	Circulation area of the vehicle	Categorical variable with 8 levels.
VEH_AGE	Age of the vehicle	Continuous variable. Discretized with 16 levels.
DRI_AGE	Age of the driver	Continuous variable. Discretized with 11 levels.
LIC_YEARS	Number of years of the driving license	Continuous variable. Discretized with 9 levels.
WEIGHT_POW	Power-to-weight ratio	Continuous variable. Discretized with 3 levels.
GENDER	Gender of the driver	Categorical variable with 2 levels (male/female).
ZONA	The different Spanish regions and some big cities	Categorical variable with 65 levels.
DIESEL	Does the vehicle have a diesel engine?	Categorical variable with 2 levels (diesel/gasoline).

For the Generalized Linear Models, the continuous rating factors were discretized in the same way as the insurance firm did. The levels of the rating factors defined a set of cells, and within these cells, the values of the dependent variable (number of claims and claim size) were aggregated. In this manner, all the explanatory variables used in the GLMs were categorical. Additionally, a set of "naïve" GLMs was also estimated. In these models, the rating factors were discretized, but not according to the levels used by the insurance firm (levels that reflect its ample experience in policy pricing and require a long process of fine-tuning to find the most appropriate discretization of the variables), but using a set of categories obtained simply from the quantiles of each variable. Nevertheless, the number of levels chosen for the discretization was the same as that used by the insurer, so the discretization was not fully "naïve", as it included some of the insurer's ratings.

For the GAMs, the continuous explanatory variables were introduced into the models using splines.

For each of the five training samples of 500,000 policies, three models (a GLM, a naive GLM, and a GAM) were estimated. These models were used to predict the values of the dependent variables for the test sets, calculating the sum of their respective absolute prediction errors as a measure of predictive power. The absolute prediction errors obtained for the test samples of 200,000 observations are shown in Table 2.

Table 2. Absolute prediction errors.

Poisson								
Sample	**Naïve GLM**	**GLM**	**GAM**					**BAM**
			$k = 10$	$k = 15$	$k = 20$	$k = 30$	$k = 50$	$k = 30$
1	15,785.96	15,774.31	15,772.9	15,770.87	15,771.02	15,770.24	15,770.84	15,772.34
2	16,065.91	16,050.05	16,050.54	16,048.58	16,048.20	16,047.72	16,046.31	16,048.67
3	15,792.93	15,779.81	15,780.58	15,778.65	15,778.43	15,777.71	15,777.37	15,778.17
4	15,765.74	15,750.54	15,750.68	15,748.98	15,749.11	15,748.74	15,748.12	15,749.06
5	15,883.12	15,872.55	15,868.71	15,867.85	15,867.50	15,867.49	15,868.31	15,868.63

Gamma									
Sample	**Naïve GLM**	**GLM**	**GAM**						**BAM**
			$k = 4$	$k = 10$	$k = 15$	$k = 20$	$k = 30$	$k = 50$	$k = 4$
1	3,085,250	3,069,263	3,056,845	3,076,293	3,077,411	3,091,096	3,106,039	3,118,692 *	3,056,219
2	3,053,085	3,045,601	3,037,337	3,043,671	3,052,295	3,062,141	3,072,514	3,083,956 *	3,038,045
3	3,175,236	3,167,669	3,155,742	3,166,577	3,177,928	3,182,010	3,189,077	3,206,048 *	3,153,131
4	2,887,690	2,884,199	2,880,680	2,882,199	2,885,353	2,890,385	2,897,359	2,911,412 *	2,880,103
5	3,225,151	3,216,130	3,203,817	3,209,767	3,215,150	3,219,094	3,234,659	3,239,390 *	3,203,398

(*) A spline with $k = 30$ has been used for VEH_AGE.

The GLMs were estimated using the standard R function, glm. For the GAMs, the gam function of the R mgcv package [67] was used with thin-plate regression splines and different values of the parameter k. This value indicated the maximum allowable dimension of the spline space and hence its degrees of freedom. Also, Table 2 shows the sums of the absolute prediction errors for these different values of k. For the Poisson models, in general, the smallest prediction error was reached for $k = 50$. Note that the huge differences between the values for the Poisson and Gamma models are due to the nature of the dependent variables (number of claims in the Poisson case, claim size in the Gamma case).

The mgcv package includes a “big data” function (bam) used to estimate the GAMs. This is very similar to the gam function but is designed to work with very big datasets and use multiple CPU cores in parallel, resulting in very efficient memory usage and shorter execution times. With the bam function, it is not advisable to use thin-plate splines because of their high computational complexity [68], so other types of splines, like the cubic regression ones that we used here, are preferable. Although these splines may lead to slightly worse results, they are much quicker to evaluate. We can see in Table 2 that, in fact, for $k = 4$, the prediction errors of the GAMs evaluated with the bam function are very close to those of the gam function, but the execution times are, as discussed below, a lot shorter and comparable to those of the GLM models.

In Table 2, the values corresponding to the Gamma regressions with $k = 50$ are marked with an asterisk because, in this case, the continuous variable VEH_AGE does not have enough different values to use a spline with $k = 50$ (the covariate has fewer unique values than the specified maximum degrees of freedom). Hence, we used a spline with $k = 30$ for this variable, and splines with $k = 50$ for the rest of the continuous variables (DRI_AGE, LIC_YEARS, and WEIGHT_POW).

To check if the improvement in the absolute prediction error obtained when using a GAM, instead of a GLM or naïve GLM, was statistically significant, we followed [69], which recommended conducting a Kolmogorov–Smirnov test. Although this test was used in [69] in a time-series prediction context, nothing prevents its usage in the more general case of comparing the prediction errors of two different models. Table 3 summarizes the p-values of the Kolmogorov–Smirnov tests for the null hypothesis that the absolute prediction errors obtained with the GAM are not stochastically lesser than the errors obtained with the corresponding GLM or naïve GLM. Therefore, rejecting the null hypothesis means that the cumulative distribution function (cdf) of the absolute error (AE) of the prediction of the GAM lies above and to the left of the corresponding cdf of the AE of the GLM or naïve GLM. This would mean that the bulk of the distribution of the GAM prediction errors concentrates on values significantly smaller than those of the GLM or naïve GLM prediction errors.

Table 3 shows the p-values obtained by comparing the prediction errors of the GAMs to those of the corresponding GLMs and checking the hypothesis that the errors of the GAMs were not (stochastically) less than those of the GLMs. The lower the p-values, the greater the confidence that the GAMs outperformed the GLMs in terms of the absolute prediction error. In general, the absolute prediction errors of the GAMs were significantly better than those of the GLMs (at a 95% or 90% confidence level), except in the case of the Poisson regression models, where the absolute errors of the GAMs were not significantly better than those of the GLMs (although they were significantly better than those of naïve GLMs).

Table 3. p-values of the Kolmogorov–Smirnov tests.

	Poisson		**Gamma**	
Sample	**GAM vs. Naïve GLM**	**GAM vs. GLM**	**GAM vs. Naïve GLM**	**GAM vs. GLM**
1	0.0256	0.0966	0.0092	0.0033
2	0.0561	0.1624	0.0351	0.0651
3	0.0246	0.2012	0.0971	0.0499
4	0.0804	0.3357	0.1945	0.0202
5	0.0335	0.1716	0.0131	0.09861

A problem that was found in the Poisson GAMs was that of overdispersion, but this can be treated in a similar manner to that in the case of the more conventional GLMs. Among these solutions, the mgcv package enables estimating, for example, a negative binomial, a quasi-likelihood model, etc.

A relatively simple way of estimating the dispersion parameter and checking if it is close to 1, as it should be, is by using the quotient between the sum of the squared Pearson residuals and the degrees of freedom of the model [70]. In our case, for all the samples and models estimated, the values of this quotient ranged between 1.17 and 1.21, and therefore the overdispersion did not cause a serious problem in any case.

As for the Gamma regressions, it appears that there is a certain overfitting, both for the default value of k ($k = 10$) and the rest of the values used ($k = 15$, 20, 30, and 50). This can be seen in Figure 1, that depicts the values taken by the splines of the continuous rating factors for the Gamma regressions with a value of $k = 30$ (the red lines, with a 95% confidence band in grey around them). In this figure, there is a series of oscillations in the splines that do not reflect any foreseeable effect of the explanatory variables, but are a mere consequence of a too-high value for the degrees of freedom allowed in the splines.

As the absolute prediction errors shown in Table 2 grew with the value of k, different values of this parameter, smaller than the default value $k = 10$, were tested. The value of k with the lesser value of the absolute prediction error was $k = 4$. Figure 2 shows the splines for the continuous independent variables of the Gamma regression models with $k = 4$. One can see that the apparent oscillations visible in the models with higher values of k, like the ones displayed in Figure 1, have disappeared.

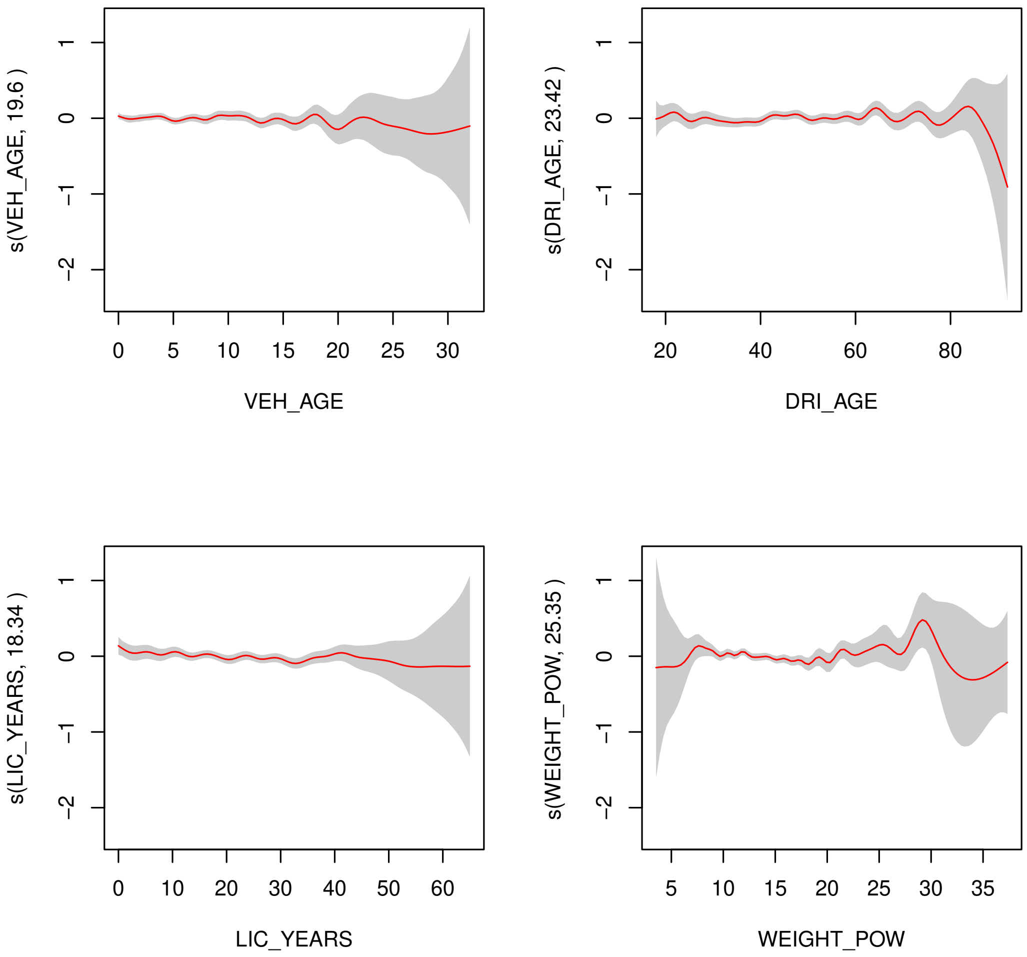

Figure 1. Overfitting in Gamma regression with $k = 30$.

In general, the results were very similar for the GAMs and the more classical GLMs with the "sophisticated" discretization of the continuous variables implemented by the insurance firm. Both kinds of models yielded results that outperformed those obtained with the GLMs with "naive" discretization, so they can be considered alternative solutions to the problem of coping with nonlinear effects in the automobile insurance field. The GLMs with discretized rating factors require fine-tuning, which can be painstaking and tedious, to determine how to discretize the continuous explanatory variables, i.e., to determine the number of levels and the intervals that define them. As for the GAMs, they are more complex models, but the fine-tuning process is more "automated" and quick, and in practice, it reduces to check that the complexity of the model, as measured by the degrees of freedom of the splines used, is high enough to properly reflect the nonlinear effect of every independent variable.

A problem with the GAMs was that their higher complexity entailed greater requirements in terms of RAM and significantly longer CPU execution times. Table 4 shows the execution times (in seconds) for a set of Poisson and Gamma regressions, with different sample sizes, both for GLMs and GAMs.

The aforementioned bam function of the mgcv package not only makes more efficient use of the RAM and is faster than the gam function but is also designed to be executed in parallel using multiple threads or multiple cores through the parallel package [71]. In our case, it was executed using three cores in a single computer. It can be seen in Table 3 that the execution times of the bam function are comparable to those of the GLMs and one order of magnitude lower than those of the gam function. Moreover, their predictive performance, as measured with the absolute prediction error, is only slightly worse than that of the GAMs with thin-plate splines. This clearly shows that the bam function provides a neat and quick alternative for estimating GAMs since it is as fast as the more conventional glm function for GLMs with discretized variables. It should be noted that the sample sizes in Table 4 are far smaller for the Gamma than for the Poisson models because the former were estimated using the observations with at least one claim, and thus the vast majority of policies, that is, the ones with zero claims, were discarded. In practice, only one out of twenty-five policies has one or more claims per year. We chose sample sizes according to this proportion.

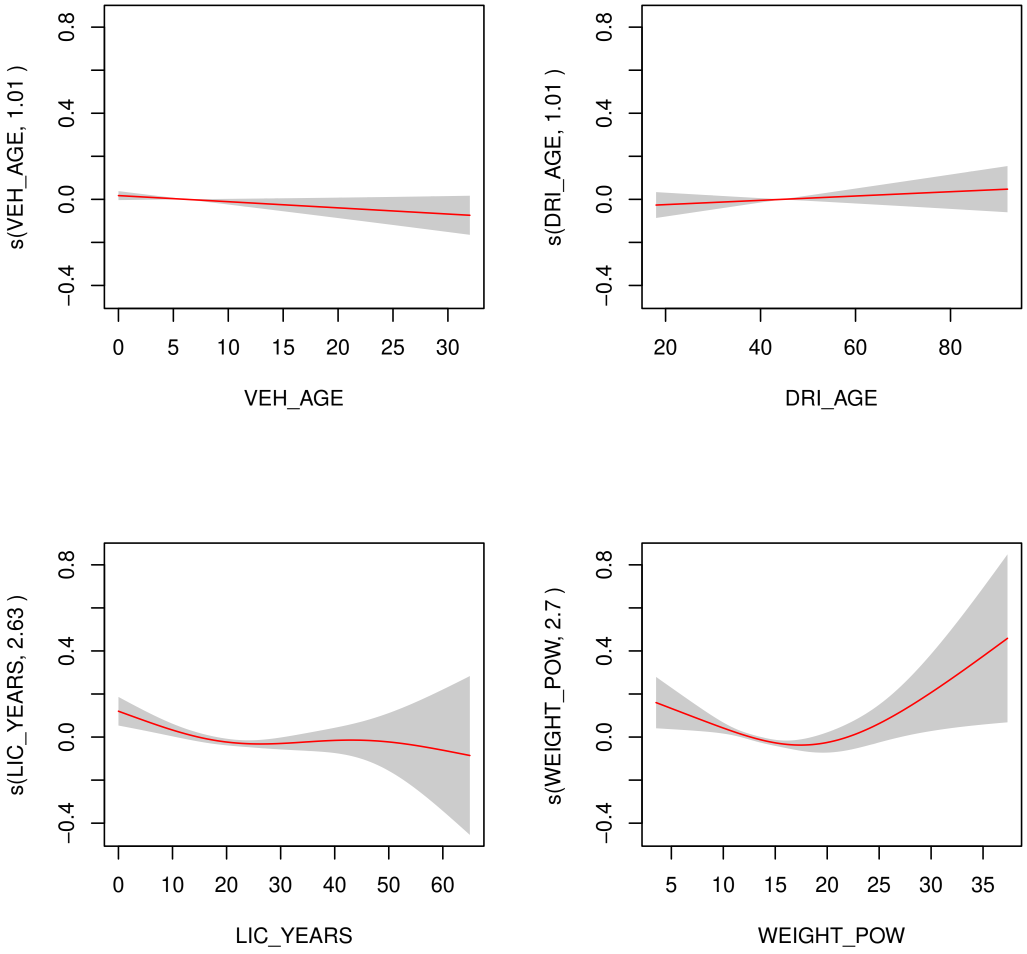

Figure 2. Gamma regression with $k = 4$.

Table 4. Execution times (in seconds).

Poisson				Gamma			
Sample Size	**GLM**	**GAM**	**BAM**	**Sample Size**	**GLM**	**GAM**	**BAM**
125,000	136.19	1591.18	201.77	5000	3.59	183.75	36.12
250,000	259.34	2326.81	348.30	10,000	6.39	289.92	36.08
375,000	370.55	3505.14	457.88	15,000	9.61	440.50	37.47
500,000	506.30	4793.21	616.93	20,000	12.48	763.18	40.33
625,000	640.68	6128.69	780.73	25,000	13.10	1014.32	54.19
750,000	757.25	8707.59	865.03	30,000	17.87	1570.21	56.51

Figure 3 shows a graphical image of the runtimes. Although these times increased exponentially with the sample size for the GAMs, with the bam function, they were almost identical to the times for estimating a GLM using the glm function.

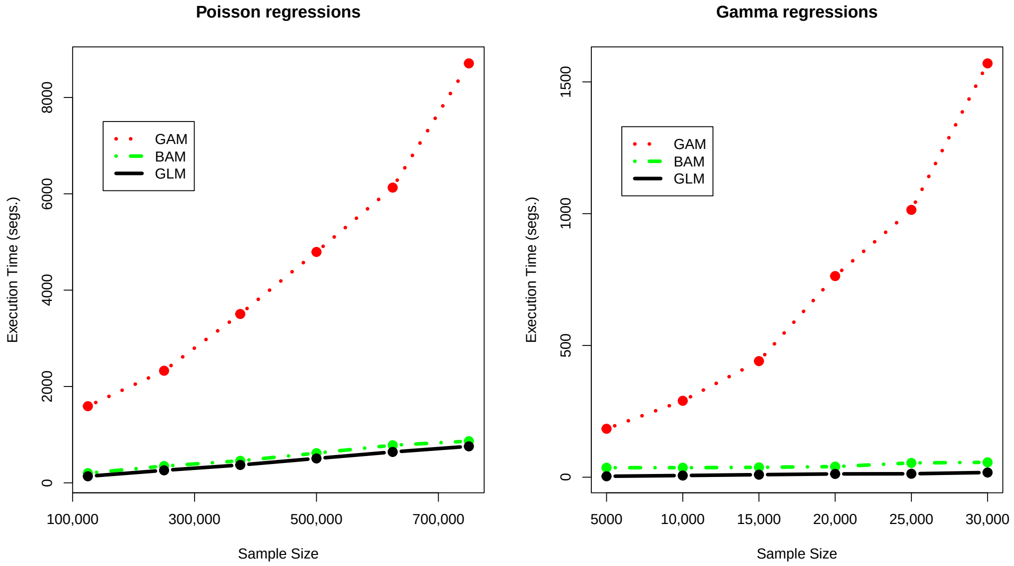

Figure 3. Execution times.

5. Conclusions

GLMs are the basic pricing tool used in the field of automobile insurance. Usually, they are used by taking the continuous rating factors and discretizing them. This implies the definition of a series of "cells", so that each policy belongs to one of them, and its premium is the one corresponding to that cell (all the policies in the same cell have the same premium). This facilitates the computation of the premium of a new policy because it reduces, in practice, to determine the cell the new policy belongs to. It also enables taking into account in a very simple manner the eventual nonlinear effects of the rating factors. As a drawback, the discretization of the rating factors implies a certain amount of information loss and also entails a certain amount of arbitrariness in the choice of the number of levels and intervals that define the discretized variables. A careful analysis and choice of these levels are, therefore, absolutely necessary to obtain an acceptable tariff model.

As a more sophisticated alternative to discretization, it is possible to replace the GLMs with their generalization, the GAMs. The latter are more computationally intensive and with them, it is no longer possible to compute the premium by merely looking at the cell the policy belongs to (since there are no cells at all), but they enable taking into account the nonlinear effects of the rating factors in a very straightforward fashion and without discretizing them. Consequently, there is no information loss, as happens in the case of GLMs with discretized explanatory variables, and it is possible to model the relationship between the rating factors and the key ratios (claim frequency, claim severity, etc.) the insurer is interested in with greater detail. The downsides of the GAMs are the greater complexity in the models, the greater computational load, and, therefore, the longer CPU time required for their estimation. Furthermore, it is not possible to determine the premium for a new policy simply by using a table that shows the set of premium values according to the levels of the rating factors. In contrast, in the case of GAMs, the calculation of the premium for a new policy requires determining the value predicted by the model for the specific values of the rating factors of that policy.

5.1. *Discussion and Future Lines of Work*

In this work, we compared both approaches—traditional GLMs with discretized variables and more sophisticated GAMs without discretization—with the aim of determining whether the use of GAMs, which are more complex and computationally intensive, is advantageous enough in comparison to GLMs to justify their utilization. To do this, we used a huge sample of car policies from a Spanish insurer and estimated and compared a series of GLMs and GAMs for the two key ratios, claim frequency and claim severity, commonly used for tariff analysis in the automobile insurance field. As is conventional, Poisson regressions were used to model the claim frequency, and Gamma regressions were used to model the claim severity.

Regarding the quality of the predictions of both types of models, the most relevant issue is that the prediction errors for the GAMs were, in general, slightly better (i.e., minor) than those for the GLMs (see Table 2). Also, in the case of the Poisson regressions, the reduction in the prediction error when passing from GLMs to GAMs was notably smaller (approximately 25% on average) than the reduction when passing from naive GLMs to GLMs (we used this last reduction as a reference). In the case of the Gamma regressions, the reduction in the prediction error when passing from GLMs to GAMs was slightly greater (16.6% on average) than the reduction when passing from naive GLMs to GLMs, although we needed to manually adjust the degrees of freedom of the models (through the value of k) to avoid overfitting.

Table 5 shows the improvements in the prediction error when moving from GLMs to GAMs, relative to the improvement in the error when moving from naïve GLMs to GLMs (i.e., the values recorded in the table are of the form $\frac{error\ GAM - error\ GLM}{error\ GLM - error\ naïve}$). We can see in this table that the improvement in the predictions' quality (measured by the absolute prediction error) is not very appealing when considered in absolute terms, but compared to the improvement experienced when moving from naive GLMs to GLMs, one can see that the error reduction is very noticeable, especially in the case of the Gamma regressions.

Table 5. Improvement in the prediction error.

	Sample					Mean	Standard Deviation
	1	**2**	**3**	**4**	**5**		
Poisson	0.3494	0.1469	0.1601	0.1184	0.4787	0.2507	0.1567
Gamma	0.7768	1.1042	1.5762	1.0080	1.3649	1.1660	0.3115

With regard to execution times, the greater complexity of the GAMs implies longer execution times compared to the GLMs. According to Table 3, in the case of the Poisson regressions, these times were, on average, approximately ten times higher for the GAMs than for the GLMs and about 61 times higher in the case of the Gamma regressions. These differences in the execution times are large enough to pose a serious problem when working with big samples and tilt the balance in favor of the less demanding GLMs. Nevertheless, these differences almost vanish if the bam function of the same R package (mgcv) is used instead of the gam (especially for the Poisson regressions). The bam function was designed to deal with big datasets, and its results were very similar to those obtained with the gam function but with a dramatic reduction in execution times. Also, this function can be executed in parallel in a cluster with multiple CPUs, speeding up the calculations and providing additional reductions in execution times. Table 3 shows that the execution times of the bam function were very similar and of the same order of magnitude as those of the gam function in the case of the Poisson regressions, and only about five times greater (instead of 61 times if the bam function was not used) for the Gamma regressions. Thus, GAMs estimated using the bam function are a competitive alternative to GLMs.

The rise of modern machine learning and deep learning methods offers a novel and powerful alternative to traditional, regression-based methods for rate analysis in insurance. Many of these new tools, such as those based on trees (random forest, for example) and those based on neural networks (deep learning), lack the interpretability of classical methods and are also very computationally intensive, but their success in areas such as image analysis and natural language processing clearly shows that they will be increasingly important in the future in the actuarial field. A natural way of extending this work is to compare GAMs and their ability to deal with nonlinearities to methods such as deep learning. Some work has been carried out in this direction, for example in [72,73].

5.2. Managerial Implications

As for the managerial implications, we can conclude that GAMs should be considered a powerful and realistic choice for tariff analysis in the automobile insurance field. They improve the quality of the models by reducing the prediction errors but at the expense of a higher complexity of the models and longer execution times. When using GAMs, it is no longer necessary to discretize the continuous risk factors, which saves a significant amount of work in determining the levels of the discretized variables and fine-tuning the models. Nevertheless, in exchange for this capability of working directly with the continuous variables, we lose the possibility of assigning a premium to a new policy by simply searching in a table that collects the premiums corresponding to each level of the rating factors.

Moreover, the use of GAMs is not completely automatic (in the sense that one could trust the results obtained using the default values of the parameters), but sometimes it is necessary for some manual fine-tuning of the models, as seen in the case of the Gamma regressions, where there was certain overfitting that forced us to manually choose the models' degrees of freedom.

Broadly speaking, the nonlinearities observed in the rating factors' effects on the key ratios (claim frequency and claim severity) used in car insurance do not seem very pronounced, as can be seen, for example, in Figure 2. This means that GAMs will not be as useful for modeling these nonlinearities as they are in certain fields of natural sciences where they are commonly used, as previously mentioned, and where the nonlinear effects of the explanatory variables are noticeably more pronounced. Nonetheless, nonlinearities exist, and GAMs prove to be a very useful tool and more sophisticated than traditional GLMs with discretized variables, which, nowadays, is standard practice in automobile insurance pricing.

Author Contributions: All authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the THEIA (Techniques for Integrity and Authentication of Multimedia Files of Mobile Devices) UCM project (FEI-EU-19-04). This work was also funded by the THEIA I (Techniques for Integrity, Authentication, and Scene Recognition in Multimedia Files of Mobile Devices—Part I) UCM project (FEI-EU-21-01). This work was also funded by the Ministerio de Ciencia e Innovación under project number PID2021-125133NB-I00.

Data Availability Statement: Data has been obtained under a confidentiality agreement and cannot be publicly disclosed.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nelder, J.A.; Wedderburn, W.M. Generalized Linear Models. *J. R. Stat. Soc. Ser. A* **1972**, *135*, 370–384. [CrossRef]
2. Brockman, M.J.; Wright, T.S. Statistical motor rating: Making effective use of your data. *J. Inst. Actuar.* **1992**, *119*, 457–543. [CrossRef]
3. Haberman, S.; Renshaw, A.E. Generalized Linear Models and Actuarial Science. *J. R. Stat. Soc. Ser. D* **1996**, *45*, 407–436. [CrossRef]
4. Denuit, M.; Hainaut, D.; Trufin, J. *Effective Statistical Learning Methods for Actuaries I. GLMs and Extensions*; Springer: Cham, Switzerland, 2019.
5. Goldburd, M.; Khare, A.; Tevet, D.; Guller, D. *Generalized Linear Models for Insurance Rating*, 2nd ed.; Casualty Actuarial Society: Arlington, VA, USA, 2020.
6. Wüthrich, M.V.; Merz, M. *Statistical Foundations of Actuarial Learning and Its Applications*; Springer International Publishing: Berlin/Heidelberg, Germany, 2023. [CrossRef]
7. Agresti, A. *Foundations of Linear and Generalized Linear Models*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
8. Dobson, A.J. *An Introduction to Generalized Linear Models*, 2nd ed.; Chapman & Hall/CRC: Boca Raton, FL, USA, 2002.
9. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd ed.; Chapman & Hall/CRC: Boca Raton, FL, USA, 1989.
10. Ohlsson, E.; Johansson, B. *Non-Life Insurance Pricing with Generalized Linear Models*; Springer: Heidelberg, Germany, 2010.
11. Fox, J. *Applied Regression Analysis and Generalized Linear Models*, 3rd ed.; Sage Publications: Thousand Oaks, CA, USA, 2008.
12. Gentle, J.E. *Computational Statistics*; Springer: New York, NY, USA, 2009.
13. Green, P.J.; Silverman, B.W. *Nonparametric Regression and Generalized Linear Models*; Chapman & Hall: London, UK, 1994.

14. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, NY, USA, 2009.
15. Demmel, J.W. *Applied Numerical Linear Algebra*; SIAM: Philadelphia, PA, USA, 1997.
16. Gentle, J.E. *Matrix Algebra. Theory, Computations, and Applications in Statistics*; Springer: New York, NY, USA, 2007.
17. Seber, G.A.F.; Lee, A.J. *Linear Regression Analysis*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2003.
18. Seber, G.A.F. *A Matrix Handbook for Statisticians*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
19. Hastie, T.; Tibshirani, R. *Generalized Additive Models*; Chapman & Hall/CRC: London, UK, 1990.
20. de Jong, P.; Heller, G.Z. *Generalized Linear Models for Insurance Data*; Cambridge University Press: New York, NY, USA, 2008.
21. Mahmoudvand, R.; Hassani, H. Generalized Bonus-Malus Systems with a Frequency and a Severity Component on an Individual Basis in Automobile Insurance. *ASTIN Bull. J. IAA* **2009**, *39*, 307–315. [CrossRef]
22. Si, J.; He, H.; Zhang, J.; Cao, X. Automobile insurance claim occurrence prediction model based on ensemble learning. *Appl. Stoch. Model. Bus. Ind.* **2022**, *38*, 1099–1112. [CrossRef]
23. Mulquiney, P. *Application of Soft-Computing Techniques in Accident Compensation*; Institute of Actuaries of Australia's (IAAust) Accident Compensation Seminar, 2004; Institute of Actuaries of Australia: Sydney, Australia, 2004.
24. Friedman, J.H. Multivariate adaptive regression splines. *Ann. Stat.* **1991**, *19*, 1–67. [CrossRef]
25. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29* , 1189–1232. [CrossRef]
26. Pesantez-Narvaez, J.; Guillen, M.; Alcañiz, M. Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression. *Risks* **2019**, *7*, 70. [CrossRef]
27. Henckaerts, R.; Côté, M.P.; Antonio, K.; Verbelen, R. Boosting insights in insurance tariff plans with tree-based machine learning methods. *N. Am. Actuar. J.* **2021**, *25*, 255–285. [CrossRef]
28. Ticconi, D. *Individual Claims Reserving in Credit Insurance Using GLM and Machine Learning*; Dipartimento di Scienze Statistiche, Sapienza Università di Roma: Rome, Italy, 2018.
29. Baillargeon, J.T.; Lamontagne, L.; Marceau, E. Mining actuarial risk predictors in accident descriptions using recurrent neural networks. *Risks* **2020**, *9*, 7. [CrossRef]
30. Delong, Ł.; Wüthrich, M.V. Neural Networks for the Joint Development of Individual Payments and Claim Incurred. *Risks* **2020**, *8*, 33. [CrossRef]
31. Blier-Wong, C.; Cossette, H.; Lamontagne, L.; Marceau, E. Machine learning in P&C insurance: A review for pricing and reserving. *Risks* **2020**, *9*, 4. [CrossRef]
32. Bailey, D.; Collins, M.; Gordon, J.; Zuur, A.; Priede, I. Long-term changes in deep-water fish populations in the northeast Atlantic: A deeper reaching effect of fisheries? *Proc. R. Soc. Lond. B Biol. Sci.* **2009**, *276*, 1965–1969. [CrossRef] [PubMed]
33. Drexler, M.; Ainsworth, C.H. Generalized Additive Models Used to Predict Species Abundance in the Gulf of Mexico: An Ecosystem Modeling Tool. *PLoS ONE* **2013**, *8*, e64458 . [CrossRef] [PubMed]
34. Grüss, A.; Drexler, M.; Ainsworth, C.H. Using delta generalized additive models to produce distribution maps for spatially explicit ecosystem models. *Fish. Res.* **2014**, *159*, 11–24. [CrossRef]
35. Heger, A.; Ieno, E.; King, N.; Morris, K.; Bagley, P.; Priede, I. Deep-sea pelagic bioluminescence over the Mid-Atlantic Ridge. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **2008**, *55*, 126–136. [CrossRef]
36. Mondal, S.; Vayghan, A.H.; Lee, M.A.; Wang, Y.C.; Semedi, B. Habitat Suitability Modeling for the Feeding Ground of Immature Albacore in the Southern Indian Ocean Using Satellite-Derived Sea Surface Temperature and Chlorophyll Data. *Remote Sens.* **2021**, *13*, 2669. [CrossRef]
37. Murase, H.; Nagashima, H.; Yonezaki, S.; Matsukura, R.; Kitakado, T. Application of a generalized additive model (GAM) to reveal relationships between environmental factors and distributions of pelagic fish and krill: A case study in Sendai Bay, Japan. *ICES J. Mar. Sci. J. Du Cons.* **2009**, *66*, 1417–1424. [CrossRef]
38. Potts, S.E.; Rose, K.A. Evaluation of GLM and GAM for estimating population indices from fishery independent surveys. *Fish. Res.* **2018**, *208*, 167–178. [CrossRef]
39. Sagarese, S.R.; Frisk, M.G.; Cerrato, R.M.; Sosebee, K.A.; Musick, J.A.; Rago, P.J. Application of generalized additive models to examine ontogenetic and seasonal distributions of spiny dogfish (Squalus acanthias) in the Northeast (US) shelf large marine ecosystem. *Can. J. Fish. Aquat. Sci.* **2014**, *71*, 847–877. [CrossRef]
40. Valavi, R.; Guillera-Arroita, G.; Lahoz-Monfort, J.J.; Elith, J. Predictive performance of presence-only species distribution models: A benchmark study with reproducible code. *Ecol. Monogr.* **2022**, *92* . [CrossRef]
41. Hua, C.; Zhu, Q.; Shi, Y.; Liu, Y. Comparative analysis of CPUE standardization of Chinese Pacific saury (Cololabis saira) fishery based on GLM and GAM. *Acta Oceanol. Sin.* **2019**, *38*, 100–110. [CrossRef]
42. Thuiller, W.; Araújo, M.B.; Lavorel, S. Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales. *J. Veg. Sci.* **2003**, *14*, 669–680. [CrossRef]
43. Momberg, M.; Ryan, P.G.; Hedding, D.W.; Schoombie, J.; Goddard, K.A.; Craig, K.J.; Le Roux, P.C. Factors determining nest-site selection of surface-nesting seabirds: A case study on the world's largest pelagic bird, the Wandering Albatross (Diomedea exulans). *IBIS* **2023**, *165*, 190–203. [CrossRef]
44. Yu, H.; Jiao, Y.; Carstensen, L.W. Performance comparison between spatial interpolation and GLM/GAM in estimating relative abundance indices through a simulation study. *Fish. Res.* **2013**, *147*, 186–195. [CrossRef]
45. Rocca, F.D.; Milanesi, P. The Spread of the Japanese Beetle in a European Human-Dominated Landscape: High Anthropization Favors Colonization of Popillia japonica. *Diversity* **2022**, *14*, 658. [CrossRef]

46. Gujral, H.; Sinha, A. Association between exposure to airborne pollutants and COVID-19 in Los Angeles, United States with ensemble-based dynamic emission model. *Environ. Res.* **2021**, *194*, 110704. [CrossRef]
47. Lee, W.; Lim, Y.H.; Ha, E.; Kim, Y.; Lee, W.K. Forecasting of non-accidental, cardiovascular, and respiratory mortality with environmental exposures adopting machine learning approaches. *Environ. Sci. Pollut. Res.* **2022**, *29*, 88318–88329. [CrossRef]
48. Li, L.; Blomberg, A.J.; Stern, R.A.; Kang, C.M.; Papatheodorou, S.; Wei, Y.; Liu, M.; Peralta, A.A.; Vieira, C.L.; Koutrakis, P. Predicting Monthly Community-Level Domestic Radon Concentrations in the Greater Boston Area with an Ensemble Learning Model. *Environ. Sci. Technol.* **2021**, *55*, 7157–7166. [CrossRef]
49. Tan, Y.; Zeng, Z.; Liang, H.; Weng, X.; Yao, H.; Fu, Y.; Li, Y.; Chen, J.; Wei, X.; Jing, C. Association between Perfluoroalkyl and Polyfluoroalkyl Substances and Women's Infertility, NHANES 2013–2016. *Int. J. Environ. Res. Public Health* **2022**, *19*, 15348. [CrossRef]
50. Pourghasemi, H.R.; Rossi, M. Landslide susceptibility modeling in a landslide prone area in Mazandarn Province, north of Iran: A comparison between GLM, GAM, MARS, and M-AHP methods. *Theor. Appl. Climatol.* **2016**, *130*, 609–633. [CrossRef]
51. Osah, S.; Acheampong, A.A.; Fosu, C.; Dadzie, I. Regression models for predicting daily IGS zenith tropospheric delays in West Africa: Implication for GNSS meteorology and positioning applications. *Meteorol. Appl.* **2021**, *28*, e2030. [CrossRef]
52. Egger, S.T.; Bobes, J.; Seifritz, E.; Vetter, S.; Schuepbach, D. Functional transcranial Doppler: Selection of methods for statistical analysis and representation of changes in flow velocity. *Health Sci. Rep.* **2021**, *4*, e400. [CrossRef] [PubMed]
53. Thompson, P.A.; Watkins, K.E.; Woodhead, Z.V.J.; Bishop, D.V.M. Generalized models for quantifying laterality using functional transcranial Doppler ultrasound. *Hum. Brain Mapp.* **2023**, *44*, 35–48. [CrossRef]
54. Cui, E.H.; Song, D.; Wong, W.K.; Li, J.J. Single-cell generalized trend model (scGTM): A flexible and interpretable model of gene expression trend along cell pseudotime. *Bioinformatics* **2022**, *38*, 3927–3934. [CrossRef]
55. Antonio, K.; Beirlant, J. Issues in claims reserving and credibility: A semiparametric approach with mixed models. *J. Risk Insur.* **2008**, *75*, 643–676. [CrossRef]
56. Breuer, A.; Staudt, Y. Equalization Reserves for Reinsurance and Non-Life Undertakings in Switzerland. *Risks* **2022**, *10*, 55. [CrossRef]
57. Denuit, M.; Lang, S. Non-life rate-making with Bayesian GAMs. *Insur. Math. Econ.* **2004**, *35*, 627–647. [CrossRef]
58. England, P.D.; Verrall, R.J. Stochastic claims reserving in general insurance. *Br. Actuar. J.* **2002**, *8*, 443–518. [CrossRef]
59. Staudt, Y.; Wagner, J. Assessing the Performance of Random Forests for Modeling Claim Severity in Collision Car Insurance. *Risks* **2021**, *9*, 53. [CrossRef]
60. Verschuren, R.M. Predictive claim scores for dinamic multi-product risk clasiffication in insurance. *ASTIN Bull. J. IAA* **2021**, *51*, 1–25. [CrossRef]
61. Wang, Y. *Smoothing Splines. Methods and Applications*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2011.
62. Faraway, J.J. *Extending the Linear Model with R*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2006.
63. Bowman, A.W.; Azzalini, A. *Applied Smoothing Techniques for Data Analysis*; Oxford University Press: New York, NY, USA, 1997.
64. Loader, C. *Local Regression and Likelihood*; Springer: New York, NY, USA, 1999.
65. Wahba, G. *Spline Models for Observational Data*; SIAM: Philadelphia, PA, USA, 1990.
66. Wood, S.N. Thin Plate Regression Splines. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2003**, *65*, 95–114. [CrossRef]
67. Wood, S. mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation; R Package Version 1.9.0. 2023. Available online: https://cran.r-project.org/web/packages/mgcv/index.html (accessed on 13 July 2023).
68. Wood, S.N. *Generalized Additive Models: An Introduction with R*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2017.
69. Hassani, H.; Silva, E.S. A Kolmogorov-Smirnov Based Test for Comparing the Predictive Accuracy of Two Sets of Forecasts. *Econometrics* **2015**, *3*, 590–609. [CrossRef]
70. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*, 4th ed.; Springer: New York, NY, USA, 2003.
71. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2023.
72. Denuit, M.; Hainaut, D.; Trufin, J. *Effective Statistical Learning Methods for Actuaries I. Neural Networks and Extensions*; Springer: Cham, Switzerland, 2019.
73. Denuit, M.; Hainaut, D.; Trufin, J. *Effective Statistical Learning Methods for Actuaries I. Tree-Based Methods and Extensions*; Springer: Cham, Switzerland, 2020.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.