mathematics

MDPI

Article

Generalised Additive Modelling of Auto Insurance Data with Territory Design: A Rate Regulation Perspective

Shengkun Xie * and Kun Shi

Global Management Studies, Ted Rogers School of Management, Toronto Metropolitan University, Toronto, ON M5B 2K3, Canada

* Correspondence: shengkun.xie@torontomu.ca; Tel.: +1-416-979-5000 (ext. 543474)

Abstract: Pricing using a Generalised Linear Model is the gold standard in the auto insurance industry and rate regulation. Generalised Additive Model applications in insurance pricing are receiving increasing attention from academic researchers and actuarial pricing professionals. The actuarial practice has constantly shown evidence of significantly different premium rates among the different rating territories. In this work, we build predictive models for claim frequency and severity using the synthetic Usage Based Insurance (UBI) dataset variables. First, we conduct territorial clustering based on each location's claim counts and amounts by grouping those locations into a smaller set, defined as a cluster for rating purposes. After clustering, we incorporate these clusters into our predictive model to determine the risk relativity for each factor level. Through predictive modelling, we have successfully identified key factors that may be helpful for the rate regulation of UBI. Our work aims to fill the gap between individual-level pricing and rate regulation using the UBI database and provides insights on consistency in using traditional rating variables for UBI pricing. Our main contribution is to outline how GAM can address a more complicated functionality of risk factors and the interactions among them. We also contribute to demonstrating the territory clustering problem in UBI to construct the rating territories for pricing and rate regulation. We find that relativity for high annual mileage driven is almost three times that associated with low annual mileage level, which implies its importance in premium calculation. Overall, we provide insights into how UBI can be regulated through traditional pricing factors, additional factors from UBI datasets and rating territories derived from basic rating units and the driver's location.

Keywords: auto insurance regulation; generalised linear models; generalised additive models; rate making; predictive modeling; usage based insurance

MSC: 62H20; 62J12; 62P05

check for updates

Citation: Xie, S.; Shi, K. Generalised Additive Modelling of Auto Insurance Data with Territory Design: A Rate Regulation Perspective. *Mathematics* **2023**, *11*, 334. https://doi.org/10.3390/math11020334

Academic Editor: Tahir Choulli

Received: 7 December 2022
Revised: 3 January 2023
Accepted: 6 January 2023
Published: 9 January 2023

1. Introduction

Usage-based insurance (UBI) and its pricing and regulation problems have received much attention in actuarial science and insurance technology [1–4]. UBI aims to use in-vehicle devices or smartphones to collect drivers' driving behaviour and driving trajectory, which may lead to a better classification of auto insurance risk. In [5], machine learning techniques, including a logistic regression approach, were used to extensively investigate how driving behaviour variables affect the prediction of the risk probability and claim frequency. The study in [5] illustrates how interpretability and high predictive accuracy can be achieved in the machine learning modelling framework of auto insurance pricing. Additionally, research in [6] shows improved underwriting performance at the early stage of adoption and a higher return on equity for insurance companies. However, research on exploring UBI data for pricing and how rate regulation can be made with UBI is still in the premature stage. In [7], a study of predictive analytics using UBI shows how they contradict the aggregate nature of insurance. The pricing based on the individual prediction may deeply shake the homogeneity hypothesis behind the insurance pooling. This motivates us to explore how traditional risk factors may potentially be helpful and how they can be used to address rate regulation problems.

Research on connecting UBI with a traditional auto insurance policy is fundamental to a better understanding of the difference between individual-level and collective-based pricing. Traditional rate-making and rate regulation are based on conventional variables related to insured drivers and vehicles. The major risk factors associated with drivers may include the driver's accident history and a variable that often combines gender, age and car use. For instance, this type of major risk factor used in auto insurance rate regulation in Canada consists of Driving Record (DR) and Class (i.e., the type of use). Through these two major factors, regulators aim to identify the key information from policyholders so that the insurance risks can be further discriminated. The relativities of these major risk factors provide benchmark values for the auto insurance industry. For regulating UBI, it is crucial to explore continuous variables such as Insured Age, Car age, Credit Scores, and Annual Mileage Driven from UBI data to address their impact on the predictive modelling of insurance risk. The contribution to the predictive models from these variables may be significantly different when it comes to the UBI, although they are proven to be important risk factors in auto insurance pricing. Therefore, further investigation of the significance associated with these rating variables is needed to better understand their impact on UBI pricing. This motivated us to conduct this study, which mainly focuses on the conventional pricing factors from the UBI database to address the significance of pricing factors, the interaction between crucial variables, and relativity estimates of rating factors. Studying these aspects will help us regulate auto insurance and provide insights into the connections and differences between individual-level and collective-based pricing.

Pricing using a Generalised Linear Model (GLM) is the gold standard in the auto insurance industry, and rate regulation [8–12]. In particular, it plays a crucial role in rate regulation due to its model interpretability. In [13], two classification techniques, including GLM, are used to investigate the claim frequency using UBI. In addition, the modelling tools facilitate the use of telemetric data to improve risk management in insurance. In [14], GLM was used to model UBI, and the study shows the potential applications of UBI by insurance companies for setting up auto insurance premium rates. The extension of modelling using GLM to UBI data is natural, and much of the UBI predictive modelling is based on GLM. The linearity assumption enables an easy understanding and interpretation of the effect of a risk factor on the model responses: either loss severity or loss frequency or loss cost. To facilitate rate regulation, variable selections or variable importance measures associated with linear models are much more accessible to conduct than non-linear models such as neural networks or decision-tree-based methods. The machine learning models often outperform GLM or other linear models [15–17], but they are considered to be black box models and may create obstacles in communication among the fields associated with the regulatory practice. This helps to answer why, in rate regulation, regulators are reluctant to use machine learning models to estimate relativities of major risk factors, which are served as benchmark values for auto insurance companies. Because of this, this work also illustrates how we can calculate risk relativities for major risk factors we considered using GLM. However, GLM cannot identify a more complicated relationship between risk factors and the response, such as claim frequency. Therefore, the interpretability of the model in this regard needs to be improved. Additionally, balancing the interpretability of the models used and the predictive power and accuracy of the models [18,19] is required; to achieve this, we may have to seek alternative solutions. This motivates us to focus on statistical models that play a role between linear models and complicated non-linear machine learning models. On the one hand, the model has desired interpretability due to the requirement of rate regulation practice of auto insurance. On the other hand, we expect some non-linear functional patterns to be addressed by specific components of proposed modelling techniques, which maintain the interpretability of models.

Applications of Generalised Additive Models (GAM) in insurance pricing are receiving increasing attention from academic researchers and actuarial pricing professionals [20–22]. In [21], GAM was applied to near-miss event data from a sample of drivers to identify the risk factors associated with a higher risk of near-miss occurrence. The study reveals that certain factors are associated with a higher expected number of near-miss events, which may be useful when implementing dynamic risk monitoring through telematics. In [9], GAM was used to model loss frequency and severity that involved both categorical and continuous numerical variables, but the study shows that GLM can approximate the GAM model closely, which leads to a similar premium structure when they are used for pricing. The additivity nature of GAM maintains the desired model interpretability in terms of the significance of contribution by each component. In addition, the ease of producing variable importance measures associated with each component makes GAM even more interpretable in explaining which risk factors contribute more or less. On the other hand, the flexibility of specifying functionality using parametric and non-parametric approaches improves the ability to explain the impact on the response using non-linear functions. From an actuarial pricing perspective, some monotonic functions, such as the relativity function of Driving Records, may be required and must be imposed on pricing. GAM provides us flexibility in specifying the functionality of the given predictor based on the need, resulting in more practically applying predictive modelling techniques for rate regulation purposes. Another reason for using GAM to predict insurance losses or claim counts is to reduce the dimension of output model parameters. This is particularly important and meaningful when adding additional variables from UBI datasets to the traditional risk factors for regulation purposes. Because of this, the total number of estimated model parameters becomes more manageable. The estimation of functionality using a set of basis function within each additive component provide an easy and intuitive explanation of the relationship between risk factors and response. This motivates us to investigate the functionality of risk factors and their impact to claim frequency and severity modelling using the UBI database. This work will fill the gap in improving connections between risk factors and the response variable using a more complicated function, such as spline [23].

Territory risk clustering [24–26] and its use in predictive models have been critical aspects of auto insurance pricing [27,28]. Rating territory has been considered a key factor in calculating auto insurance premiums. Depending on how the basic rating unit is set for pricing, these rating units may consist of postal codes, zip codes, or a larger residential area. The risk relativities associated with the rating territory are discriminative from one region to the other. The actuarial practice has constantly shown evidence of significantly different premium rates among rating territories. The main reason for this is if drivers rely heavily on personal vehicles, then the loss cost will be higher than in other areas where people have more choices in terms of local transportation, therefore, a high premium in that area. Because of this, clustering the territory risk is essential in auto insurance pricing. However, to the best of our knowledge, literature has yet to be found in the research on clustering territory risk for UBI. This may call for a study of how clustering of the territory risk can be incorporated into the predictive modelling of loss frequency or loss severity or both. In this work, we build predictive models for claim frequency and severity using the synthetic UBI dataset variables. To achieve this goal, first, we conduct territorial clustering based on each location's claim counts and amounts by grouping those locations into a smaller set and define them as a cluster for rating purposes. Grouping the neighbouring areas or similar loss severity or frequency helps improve the credibility of the risk measures; therefore, it helps improve the stability of modelling results. After clustering, we incorporate these clusters into our predictive model to determine the risk relativity for each factor level. Within the modelling using GAM, we consider models both with and without interactions between variables of interest. We are interested in obtaining the numerical risk factor functionality conditional on different categorical risk factor levels. This allows us to obtain insights into how risk factors are related and whether or not some interactions between them can be excluded due to the less significant impact.

In auto insurance rate regulation, rate making and risk classification are typical tasks. Therefore, determining the risk-level relativity and territory risk clustering are two primary focuses. When working with UBI, insurance pricing is determined at the individual level based on personal risk characteristics, and an investigation into how we can predict the claim frequency and claim amounts using the UBI database for the risk factors available from such a database is needed. However, from a rate regulation perspective, each driver's driving patterns or behaviours are of no interest, but pricing factors from UBI may help predict claim frequency and severity. Our work does not aim to build a predictive model for individual-level pricing for UBI. Instead, we fill the gap between individual-level pricing and rate regulation using the UBI database and provide insights on potential consistency in using traditional rating variables for UBI pricing. Although an insurance company may use many other factors linked to specific drivers, where the type of risk is more detailed, regulators will only aim to determine the actuarial fairness based on major risk factors. Because of this, the approaches used to evaluate the level of risk are all group based. Since how rate regulation on traditional group-based insurance pricing can be extended to UBI is still an open question, regulation of UBI from an actuarial perspective needs to be well addressed. Therefore, predictive modelling problems using traditional variables from the UBI database must be explored. Our main contribution is to outline how GAM can address a more complicated functionality of risk factors and their interaction among them. We also contribute to demonstrating the territory clustering problem in UBI to construct the rating territories for pricing and rate regulation. The rest of this paper is organised as follows. In Section 2, GAM modelling techniques, territory clustering, and risk relativity estimates are discussed. Next, in Section 3, the application to UBI synthetic data is presented and analyzed. Finally, we conclude our findings, provide further remarks and outline future work in Section 4.

2. Materials and Methods

2.1. Data

The dataset used in this research is a synthetic dataset that consists of a portfolio of 100,000 auto UBI policies provided by [29]. The dataset can be found in http://www2.math.uconn.edu/~valdez/data.html (accessed on 1 November 2021). This synthetic dataset contains three types of variables: traditional policy variables, driving patterns-related variables, and response variables, which are claim amounts and claim counts. There are 52 variables, including categorical variables such as marital status and car use and numerical variables such as annual mileage driven and credit scores. In addition, some variables are used to describe driving patterns, such as the speeds of acceleration and braking. Some variables capture the intensity of left or right turns. Due to the imbalance of the dataset, the information related to claim counts and claim amounts is limited. To enrich the information and become more useful for the illustration of predictive modelling for insurance loss using UBI, the dataset was generated using SMOTE technique [30] from a relatively smaller real-world telematics dataset. The application of oversampling via SMOTE is due to the limited available observations and the imbalanced nature of insurance data. Furthermore, this dataset is of transactional type, and each record reports if there is a claim. If so, then there is an associated loss amount of claim. For an additional introduction of telematics variables and descriptive data analysis of variables in this dataset, readers can refer to [29]; however, this is not required, and more detailed information will not affect the self-contained property of this paper.

In the present study, we focus on traditional policy variables and territory clustering problems to address statistical modelling and clustering with applications in UBI. Since we aim to provide the general guideline for rate regulation, we focus on the prediction of claim amounts and claim frequency based on the major risk factors only. We specifically focus on modelling using Insured Age, Gender, Car Age, Marital Status, Car Use, Credit Scores, Annual Mileage Driven, and Years of no Claims. Within this predictive modelling, we also incorporate clustering techniques to define new territories to reduce the dimensionality of the current level of territory variable. As we can see from Table 1, there are many levels of territory codes that indicate different locations of losses, but claim counts and total exposures associated with each territory level are low for many levels. This may cause a credibility issue regarding the estimate of loss cost and average loss. This implies that regrouping of the territory is needed to improve credibility. We observe a significant positive correlation between the Claim Probability and average loss, which is typical in auto insurance. We also observe a negative correlation between the Claim Probability and Credit Scores, as well as the claim probability and Annual Mileage Driven. To get an idea of the distribution of variables used in our predictive modelling using GAM, we present the five-number summary statistics for numerical variables, including Insured Age, Car Age, Credit Scores, Annual Mileages Driven, Year of no Claims, and Claim Amounts. The results are summarized in Table 2. We observe that Annual Mileages Driven, Years of no Claims, and Claim Amounts are quite right-skewed. The frequency distributions of those categorical variables, including Insured Sex, Marital Status, Car Use, and Claim Counts, are shown in Figure 1. From the presented results, we see that the frequency distributions of these variables are quite imbalanced. These data characteristics demonstrate a certain level of complexity for the given data, implying more advanced statistical techniques may be considered.

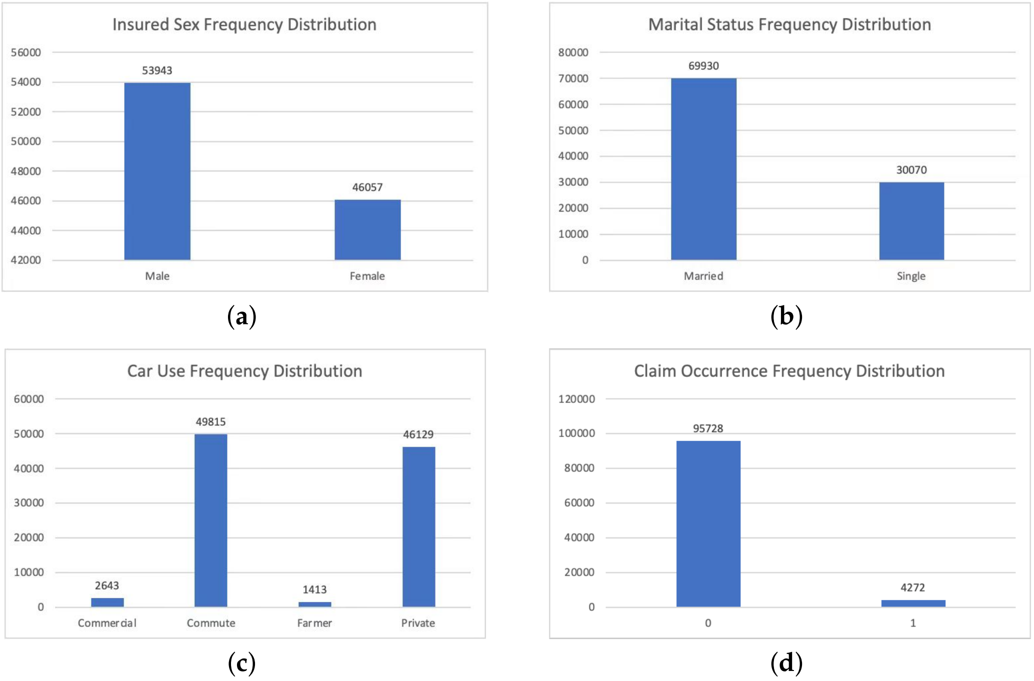
Table 1. Key summary statistics by original territory code (i.e., without clustering). Note, Loss Cost = Sum of Claim Amounts/Total Exposures; Average Loss = Sum of Claim Amounts/Sum of Claim Counts.

Territory Code	Sum of Claim Counts	Total Exposures	Sum of Claim Amounts	Average of Credit Score	Average of Annual Miles Drive	Claim Probability	Loss Cost	Average Loss
11	0	52	0	826	8675	0	0	NA
12	42	1296	146,074	813	9244	3.24%	113	3478
13	57	1277	170,183	808	9122	4.46%	133	2986
14	50	1245	131,196	801	8789	4.02%	105	2624
15	83	1885	245,435	797	8728	4.40%	130	2957
18	109	2710	468,572	803	9167	4.02%	173	4299
23	58	1533	202,581	802	9156	3.78%	132	3493
24	67	1865	127,610	799	8170	3.59%	68	1905
26	75	1504	243,652	798	8343	4.99%	162	3249
30	105	2716	309,225	793	8573	3.87%	114	2945
31	103	2612	325,271	800	9103	3.94%	125	3158
32	113	2307	308,753	801	9292	4.90%	134	2732
33	89	1893	316,630	796	9401	4.70%	167	3558
35	122	3279	338,824	797	8971	3.72%	103	2777
36	85	1962	208,506	800	8670	4.33%	106	2453
37	75	1938	215,604	803	9500	3.87%	111	2875
38	102	2839	357,571	813	9902	3.59%	126	3506
39	166	3686	385,668	805	9772	4.50%	105	2323
43	162	3842	430,690	806	9730	4.22%	112	2659
52	93	2366	240,624	799	9204	3.93%	102	2587
54	62	1175	151,053	801	8770	5.28%	129	2436
57	85	1530	310,372	803	8781	5.56%	203	3651
59	69	1562	332,517	804	8860	4.42%	213	4819
60	52	987	204,317	798	9116	5.27%	207	3929
61	43	944	227,861	799	8842	4.56%	241	5299
62	45	1034	113,990	804	8780	4.35%	110	2533
63	61	1365	272,002	801	8982	4.47%	199	4459
64	57	1304	221,465	799	8817	4.37%	170	3885
65	51	1320	187,664	804	8914	3.86%	142	3680
66	80	1770	220,030	801	9057	4.52%	124	2750
67	58	1379	180,420	796	9169	4.21%	131	3111
68	81	1499	221,893	798	8996	5.40%	148	2739
69	98	1731	286,512	800	8943	5.66%	166	2924
70	70	1406	272,776	793	8849	4.98%	194	3897
71	97	1532	415,526	795	8807	6.33%	271	4284
72	104	1891	433,089	798	8786	5.50%	229	4164
73	103	2035	428,190	800	8770	5.06%	210	4157
74	101	2064	600,002	798	9143	4.89%	291	5941
75	86	1733	371,354	799	8674	4.96%	214	4318
76	91	1585	331,613	798	8962	5.74%	209	3644
77	90	1733	388,935	795	8961	5.19%	224	4321
78	52	1287	150,199	800	9450	4.04%	117	2888
79	52	1357	124,776	791	9389	3.83%	92	2400
80	59	1505	165,601	794	9396	3.92%	110	2807
81	49	1320	79,334	797	9458	3.71%	60	1619
82	51	1234	106,663	795	9316	4.13%	86	2091
83	74	1569	292,613	794	9476	4.72%	186	3954
84	90	2552	253,480	803	9379	3.53%	99	2816
85	102	3288	314,241	805	9203	3.10%	96	3081
86	81	2209	184,604	795	9571	3.67%	84	2279
87	83	2372	184,865	802	9326	3.50%	78	2227
88	98	2511	214,715	804	8597	3.90%	86	2191
89	68	2404	204,761	810	9007	2.83%	85	3011
90	30	972	83,123	815	9619	3.09%	86	2771
91	43	1034	57,003	814	10356	4.16%	55	1326

Using this database, we compare the results obtained from the empirical study (no models involved) and modelling with and without rating territories that we design from clustering. We also address how these major risk factors from the UBI database can be used for rate regulation purposes and how risk relativity can be obtained from GLM.

Table 2. Five number summary statistics (i.e., Minimum, 1st Quartile, Median, 3rd Quartile, and Maximum) and the mean value for numerical variables used in this work including the response variable Claim amounts where zero losses are excluded. Note that the minimum value −2 for Car Age is due to the fact that buying a newer model can occur up to two years in advance.

	Insured Age	**Car Age**	**Credit Score**	**Annual Miles Driven**	**Years of No Claims**	**Claim Amounts**
Min.	16.00	−2.00	422.0	0	0.00	0.77
1st Qu.	39.00	2.00	766.0	6214	15.00	786.27
Median	51.00	5.00	825.0	7456	29.00	1988.60
Mean	51.38	5.64	800.9	9124	28.84	3561.13
3rd Qu.	63.00	8.00	856.0	12,427	41.00	4037.89
Max.	103.00	20.00	900.0	56,731	79.00	104,074.89



(**a**) (**b**)

(**c**) (**d**)

Figure 1. The frequency distributions for categorical variables used in this work, including Insured Sex (**a**), Marital Status (**b**), Car Use (**c**) and Claim Counts (**d**), which is the response variable for model that predicts claim probability.

2.2. Generalised Additive Models

Generalised Additive Models are an extension of Generalised Linear Models [31,32], which have been used as an important modelling tool in risk analysis, particularly for auto insurance pricing and rate regulation problems. The beauty of using GAM instead of GLM is that GAM facilitates the estimate of linear, additive model components that can be used to reflect how those factors affect the response through a clear functional relationship in terms of those predictor variables, independently, for a more interpretable model. Additionally, an easy way of incorporating non-linear components through a specification of an interaction term between different variables makes the modelling using GAM even more powerful. Investigation of dependency or interaction between pricing variables using GAM is new in auto insurance pricing and rate regulation, and we aim to discover the interesting patterns behind the pricing variables and insights into how they interact using UBI data.

Mathematically speaking, modelling using GAM with K predictors with an assumption of no interaction among them can be represented as follows:

$$f(X_1, X_2, \ldots, X_K) = \beta_0 + \sum_{k=1}^{K} f_k(X_k), \tag{1}$$

where $f_k(X_k)$ is the kth additive component of the model that represents a non-linear smoothing function of X_k. The idea behind GAM is to model functionality between a response Y and K different predictors through an additive, linear in parameters regression problem where the following squared loss function is minimised

$$L(\beta_0, \ldots, \beta_k) = \left(Y - \beta_0 - \sum_{k=1}^{K} f_k(X_k)\right)^2. \tag{2}$$

Of course, the functionality of $f(X_1, X_2, \ldots, X_K)$ in the GAM model is only an approximation, as we do not have a ground truth model to reflect the actual relationship. However, such linear and additive approximation often performs better than other regression models such as polynomial regression [33], piecewise linear regression [34]. Additionally, the loss function can be extended by including an extended term allowing control of the smoothness of the curve estimate. In this work, we use a set of spline basis functions $\{b_{kl}(X_k)\}_{l=1,2,\ldots,L_k}$ for the kth predictor to model each additive component $f_k(X_k)$ in the model. We control the sparsity of $\{b_{kl}(X_k)\}$ by adding a penalty term to the loss function. Therefore, the penalised version of loss function in (2) becomes:

$$L^p(\beta_0, \ldots, \beta_{kl}) = \left(Y - \beta_0 - \sum_{k=1}^{K} \sum_{l=1}^{L_k} \beta_{kl} b_{kl}(X_k)\right)^2 + \sum_{k=1}^{K} \sum_{l=1}^{L_k} \lambda_k \left(\beta_{k,l+1} - \beta_{kl}\right)^2. \tag{3}$$

Note that the parameters were produced using the mgcv R package, which automatically selects optimal wiggliness-penalty, λ_k to produce the functionality between the numerical risk factor X_k and the response variables Y. Here, λ_k controls the wiggliness of the smoothed terms or the B-spline basis by penalising the squared difference of two consecutive coefficients of basis functions with each factor, and it is cross-validated in terms of its choice. This work uses REML (restricted maximum likelihood) to produce smooth terms. Other methods, such as GCV (generalised cross-validation) and Mallow's C_p are available in the mgcv package, and they can also be used to achieve integrated smoothness estimation of functionality [35]. A more complicated smoothing function for $f_k(X_k)$ other than splines [36] is possible when it is needed, but this is outside the scope of this paper.

In this work, both claim frequency and claim amount are modelled by GAM. The GAM model of the claim amount can be described as follows:

$$\begin{aligned} Y_i^l &= \beta_0 + \beta_1 InsuredAge + \beta_2 InsuredSex + \beta_3 Marital + \beta_4 CarUse + s(CarAge_i) \\ &+ s(CreditScore_i) + s(AnnualMiles_i) + s(YearsNoClaim_i) + \epsilon_i, \end{aligned} \tag{4}$$

where $s(\cdot)$ is a spline function used to capture the smooth functionality between the response variable and the given predictor variable. Y_i^l represents the ith claim amount. The error term ϵ_i, is assumed to independently, identically follow a Gamma distribution, a popular loss distribution selection. As we mentioned earlier, the advantage of using GAM is its flexibility in constructing functions of numerical variables controlled by other categorical variables at a certain level. This allows us to see how the functional behaviour of a variable of interest will depend on other factors to discover their potential interaction. For example, when we aim to study how Annual Miles Driven depend on the type of Car use, the model becomes the following for the case in which CarUse is not one of the predictors.

$$\begin{aligned} Y_i^l &= \beta_0 + \beta_1 InsuredAge + \beta_2 InsuredSex + \beta_3 Marital + s(CreditScore_i) \\ &+ \ s(CarAge_i) + \ s(AnnualMiles_i) + s(YearsNoClaim_i | InsuredSex) + \epsilon_i, \end{aligned} \quad (5)$$

Since GAM is an extension of GLM, the setting of an error function, link function, and variance function shares the commonality of GLM. Although there are many other choices, we use log link and Gamma distribution as error distribution due to their popularity in practice. The model form of GAM for claim probability is identical to the claim amount except for the different response variable and error function, which is assumed to be binomial.

2.3. Incorporating Designed Territories to Predictive Modelling

In auto insurance pricing, the territory risk is critical and plays an important role, since the risk levels have been well discriminated against, and those basic rating units are formed into clusters. Thanks to the clustering effect, the overall variation of loss cost is minimised. Using Model (4), we further use *Location.Cluster* as a variable to indicate the location or rating territory, and we have to design such territory from the UBI dataset. In this data, there are 57 territories. These territories belong to either an Urban or a Rural region. In order to take the region into consideration, we re-organised this location information by appending region to the territories to create $57 \times 2 = 114$ different locations. To conduct the clustering, we applied the K-mean clustering algorithm to the predicted claim frequency from a logistic regression model and claim amounts from the GAM model. This selection is based on our intensive study of interpretable clustering using low-dimensional feature vectors (i.e., two-dimensional). Our study shows that clustering based on the claim frequency and claim amounts leads to the most discriminative clusters. This two-dimensional feature vector in clustering enables a visualisation of clustering results, which is critical in rate regulation, where interpretability is highly desirable. This method outperforms the one that uses principal component analysis (PCA) to extract low-dimensional feature vectors from all risk characteristics available to us. Unfortunately, the PCA approach does not help improve the discriminative power in clustering territory risk.

We should realize that this work has two layers of predictive modelling when rating territory is incorporated into the model. In auto insurance, rating territory has been proven to have the most discriminative power in risk analysis, therefore, we will investigate how our design of new rating territories affect the predicted outcomes of the risk measures, including claim probability and claim amounts. After incorporating the designed rating territory to the GAM for claim amounts, the new model becomes:

$$\begin{aligned} Y_i^l &= \beta_0 + \beta_1 InsuredAge + \beta_2 InsuredSex + \beta_3 Marital + \beta_4 CarUse + \beta_5 Location.Cluster \\ &+ s(CarAge_i) + s(CreditScore_i) + s(AnnualMiles_i) + s(YearsNoClaim_i) + \epsilon_i. \end{aligned} \quad (6)$$

The model for claim frequency prediction is similar to (6), but some variables that do not have significant impact to the prediction are removed from the model. The model can be described as follows:

$$\begin{aligned} Y_i^c &= \alpha_0 + \alpha_1 CarUse + \alpha_2 Location.Cluster + s(CarAge_i) + s(CreditScore_i) \\ &+ s(AnnualMiles_i) + s(YearsNoClaim_i) + \epsilon_i, \end{aligned} \quad (7)$$

where Y_i^c represents the claim counts.

2.4. Deriving Relativities for Risk Factors

In GAM modelling, functional patterns are obtained for those numerical variables and interaction between different risk factors can also be investigated. In rate regulation, an essential aspect of evaluating the risk factors is estimating their risk relativity, and we rely on a generalised linear model to complete this task. The main reason for using GLM is that the risk relativity needs to be derived at the factor level, including numerical and categorical variables. Of course, in theory, one may derive the risk relativity based on the estimated functional component for a given continuous risk factor. However, this may be only practically achievable and useful for a monotonic functional pattern, as a more detailed local pattern may not lead to a meaningful implication, and one needs to explain why the local fluctuation of risk relativity should be considered. However, this is outside the scope of this work.

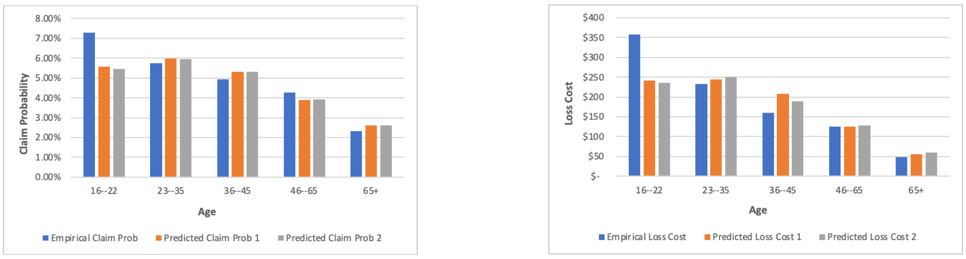
To obtain a set of relativity for each factor at each level, we regroup each numerical variable to produce different levels. For instance, we regroup the Insured Ages into the following classes: 16 to 22, 23 to 35, 36 to 45, 46 to 65, and 65 and above. For Car Ages, we group them into less than 0, 1 to 5, 6 to 10, 11 to 15, and greater than 16. In the regrouping of Insured Ages and Car Ages, we maintain consistency with common practices in Canadian rate regulation. In addition, variables such as Annual Mileage Driven and Years of no Claims, appeared in rate regulation less often, as far as we know. The groupings were made by considering the total number of risk exposures, particularly for the first and the last group. The same principle is applied to variables of Credit Scores and Years of no Claims. For the detail of the grouping level of these two variables, readers can refer to the relativity output shown in the result section.This regrouping helps to improve the stability of risk relativity for each factor level as later, the claim counts and claim amounts are aggregated and will be taking a log-scale transformation via a link function within GLM to handle the heterogeneity of data.

There are two steps in deriving risk relativity. First, we fit the data after regrouping for each individual observation to GLM (i.e., logistic regression) using claim frequency as a response for the model to obtain the coefficient of the ith factor at the jth level, denoted by α_{ij}. We then repeat the same procedure by changing the response variable to claim severity, which is the claim amount, to obtain β_{ij}, which is the coefficient of the ith factor and jth level. If a level has been selected as a basis in GLM, the coefficient of such a factor level is zero. Since the log link is used for both GLMs that are used to model severity and frequency, we have to transform the coefficients back to the original scale; therefore, the risk relativity of the ith factor and jth factor level becomes $exp\{\alpha_{ij} + \beta_{ij}\}$. If we further select the first level in each factor as the basis, the final relativity estimate is $\frac{exp\{\alpha_{ij}+\beta_{ij}\}}{exp\{\alpha_{i1}+\beta_{i1}\}}$. This is to ensure that the selected factor level will have relativity to be one. To estimate relativity, we use the same set of predictors for the models we considered; although they may exist, some of them are not statistically significant in one of the models. However, this does not affect our estimate of risk relativity, as their relativities will be close to one when such a factor does not significantly impact the prediction outcome. Note that the intercept coefficients of both models will form a basis rate, which can be treated as an overall average premium for all drivers. It is equal to $exp\{\alpha_0 + \beta_0\}$. So, the risk relativity at a level of a given risk factor is a modifier that is used to reflect the relative risk level when a driver falls into that category.

3. Results

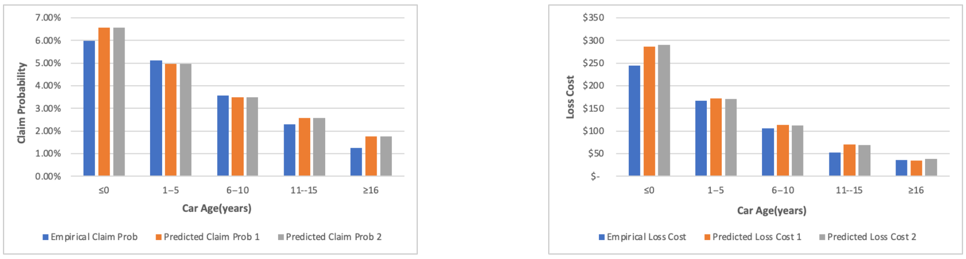
This section will present and discuss the results from various statistical models, including GLM, GAM, GAM with the designed territory, and GAM with the interaction of variables. We will present the comparative results to demonstrate the strength of using GAM. When producing functionality of numerical variables in GAM, the penalised B-spline was used, and the optimal selection of a number of basis functions was cross-validated through an internal mechanism of the mgcv R package. This is how we achieve the smoothness estimation for the GAM model. Finally, the results on risk relativities for risk factors at each level, obtained from GLM mode, are analyzed and discussed.

Since we focus on the rate regulation aspect, we first investigate the group-based distribution of numerical risk factors for the claim probability and loss cost. Then, we compute the predicted probability and loss cost by different groups after predicting the claim frequency and severity using GAMs. Finally, the results are separated by the prediction with and without designed clusters as an additional rating factor and the empirical estimates, where the claim probability and average claim amount are computed based on the actual observations of claim occurrences. These grouped distributions provide an overall pattern of claim frequency and loss cost depending on the rating factors we consider. The results are reported in Figures 2–6, respectively, for Driver Age, Credit Scores, Car Age, Annual Mileage Driven and Years of no Claims. We observe that the major pattern of distributions of these rating variables from the predictive models (with and without designed clusters) is in line with the empirical distributions. However, there is still a discrepancy for some factor levels; for instance, the group of youngest drivers. The main reason for this discrepancy may be due to a smaller number of exposures for the young driver class. From Figures 2 and 3, we do not see a significant impact from the inclusion of clusters as an additional rating factor. The distributional pattern for Credit Scores is similar to Driver Age or Car Age, with more distinguishable results from the predictive models. This may tell us that predictive modelling can balance overall loss or claim frequency patterns; therefore, it is more useful and more practical than empirical results. In particular, from the results presented in Figure 5, we notice the power of GAM in capturing the desired distributional pattern for Annual Mileage Driven. For example, we expect that risk will increase with the increase in Annual Mileage driven, but due to insufficient observations, we may not be able to determine this from the observed data. However, the GAM modelling can pick up this expected pattern, even though the information presented in the dataset is limited, especially for the GAM predictive modelling with designed clusters. This may suggest that incorporating clusters to the GAM for rate making is necessary to better discriminate the underlying risk. The significant increasing or decreasing patterns of distributions may imply that the rating factors are helpful in clearly discriminating claim probability of loss cost by different factor levels. Therefore, the risk relativities computed later become more distinguishable among different levels with a risk factor.



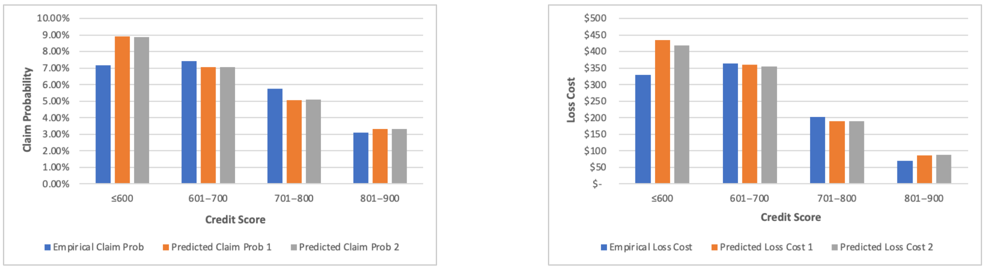
(**a**) (**b**)

Figure 2. Comparison of empirical and the predicted values of claim probability and loss cost by different age groups. (**a**) presents the results of claim probability. (**b**) corresponds to the loss cost. The orange bar shows results for predictions without using clusters. The grey bar presents results for predictions with designed clusters.



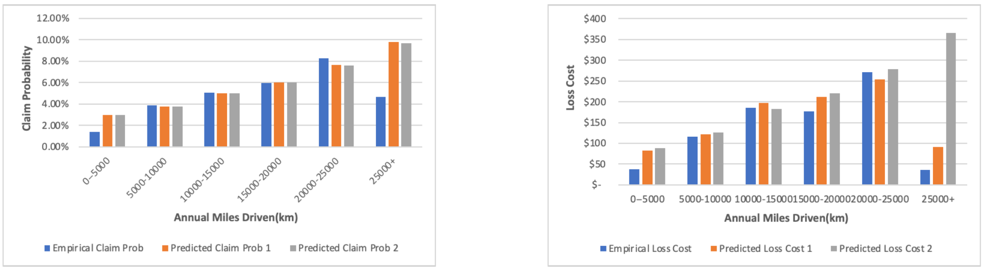
(**a**) (**b**)

Figure 3. Comparison of empirical and predicted values of claim probability and loss cost by car age groups. (**a**) presents the results of claim probability. (**b**) corresponds to the loss cost. The orange bar shows results for predictions without using clusters. The grey bar presents results for predictions with designed clusters.



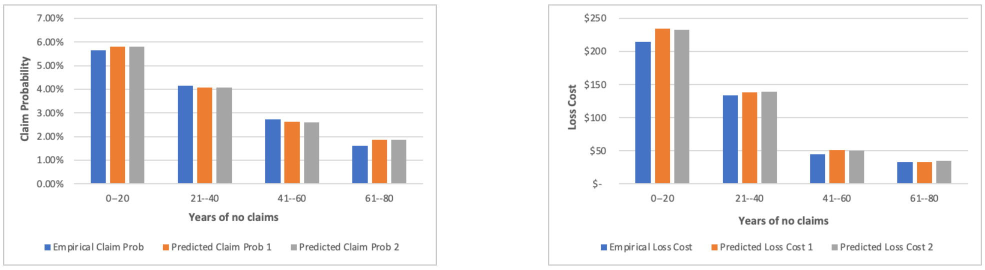
(**a**) (**b**)

Figure 4. Comparison of empirical and predicted values of claim probability and loss cost by Credit Scores. (**a**) presents the results of claim probability. (**b**) corresponds to the loss cost. The orange bar shows results for predictions without using clusters. The grey bar presents results for predictions with designed clusters.



(**a**) (**b**)

Figure 5. Comparison of empirical and the predicted values of claim probability and loss cost by Annual Mileage Driven (**a**) presents the results of claim probability. (**b**) corresponds to the loss cost. The orange bar shows results for predictions without using clusters. The grey bar presents results for predictions with designed clusters.



(**a**) (**b**)

Figure 6. Comparison of empirical and predicted values of claim probability and loss cost by Years of no Claims (**a**) presents the results of claim probability. (**b**) corresponds to the loss cost. The orange bar shows results for predictions without using clusters. The grey bar presents results for predictions with designed clusters.

Based on our results, we notice some merits for the inclusion of designed clusters in GAM, as it seems to outperform other cases in connecting the responses and the Annual Mileage Driven. The Annual Mileage Driven is particularly important for UBI pricing as it measures important drivers' driving characteristics. We present the parameter output for GAM models, including both prediction models, to understand the driving force affecting the claim frequency and amounts. The results are shown in Table 3 for frequency modelling and in Table 4 for claim amounts modelling. First, the set of variables deemed essential have significantly different results for modelling frequency and severity, but they share a commonality. In modelling frequency, Car Age, Car Use, Credit Scores, Clusters, Annual Mileage Driven, and Year of no Claims are statistically significant. In contrast, Car Use, Clusters, Insured Age, Car Age, Credit Scores, and Year of no Claims are substantial for the GAM model in predicting claim amounts. The exciting thing related to UBI is the Annual Mileage Driven. This variable is only significant in predicting frequency, but not the claim amounts. However, given that the pattern of loss cost distribution of Annual Mileage Driven presented in Figure 5 is linearly strongly increasing, we can infer that this tendency is mainly due to the significant impact of the frequency distribution of such a variable. Given that the Annual Mileage Driven variable represents the usage of the vehicle and is a driving force of high claim probability, which further leads to increased loss cost, it may become a major risk factor for rate regulation of UBI on a collective basis. This means that the distribution of Annual Mileage Driven can be estimated based on industry-level data to obtain a set of benchmark values for different levels of such risk factors. In this case, rate regulation of UBI can be started by extending the traditional focus to include an additional key factor that can describe UBI's key characteristics, such as the Annual Mileage Driven variable.

Table 3. The model outputs from Logistic Regression (with designed clusters) used to predict claim frequency.

	Estimate	**Std. Error**	**z Value**	**Pr (>|z|)**
(Intercept)	0.075	0.172	0.434	0.664
Car.age	−0.075	0.004	−17.813	0.000 ***
Car.useCommute	−0.192	0.082	−2.335	0.020 *
Car.useFarmer	−0.859	0.236	−3.633	0.0003 ***
Car.usePrivate	−0.216	0.088	−2.463	0.014 *
Credit.score	−0.003	0.0002	−18.908	0.000 ***
Location.Cluster2	0.221	0.095	2.317	0.020 *
Location.Cluster3	−1.070	0.416	−2.573	0.010 *
Location.Cluster4	0.199	0.070	2.837	0.005 **
Location.Cluster5	0.445	0.320	1.393	0.164
Location.Cluster6	−0.221	0.084	−2.638	0.008 **
Location.Cluster7	0.102	0.112	0.915	0.360
Location.Cluster8	−0.678	0.457	−1.484	0.138
Location.Cluster9	0.010	0.069	0.138	0.890
Location.Cluster10	0.187	0.075	2.492	0.013 *
Location.Cluster11	0.080	0.517	0.155	0.877
Location.Cluster12	−0.512	0.119	−4.304	0.00002 ***
Location.Cluster13	−0.058	0.078	−0.740	0.459
Location.Cluster14	0.104	0.067	1.546	0.122
Annual.miles.drive	0.00003	0.00000	7.069	0.000 ***
Years.noclaims	−0.012	0.001	−10.511	0.000 ***

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

One of the key characteristics in GAM modelling is the additive component, which can be used to reflect if the impact from such a variable is significant and if the functional relationship is significantly non-linear. We display the relevant results in Table 5 to explore this. In Table 5, the effective degree of freedom for all additive components is greater than one, and they are all statistically significant, indicating a non-linearity of each additive component. So, we further investigate the functional patterns for each numerical risk factor. Figure 7 displays the functional pattern of Car Age, Credit Scores, Insured Age, and Year of no Claims, respectively. They are the four smooth terms included in the GAM for predicting claims. The results show that the sampling errors become much larger for the groups with fewer risk exposures. This leads to the confidence intervals being much wider for those groups. With the increase in the Car Age, the effect on the claim amounts decreases. From the functional pattern of the Car Age, we see that the new car has a much greater impact on the claim amounts. The general functional pattern of Credit Scores decreases with the increase in scores, implying that a driver with higher Credit Scores seems to represent a greater risk. Additionally, it is interesting that the pattern is periodic, with a frequency of approximately 20. After 80 years old, the claim amounts' impact continues to be negative. As for Years of no Claims, the trend appears to decrease with the increase in the number of Years of no Claims. After the number of Years of no claim reaches 60, the impact on claim amounts becomes positive.

Table 4. The model outputs from GAM (with clusters) used to predict claim amounts.

	Estimate	**Std. Error**	**t Value**	**Pr (>|t|)**
(Intercept)	7.831	0.151	51.920	0.000 ***
Insured.sexMale	0.024	0.043	0.558	0.577
MaritalSingle	0.007	0.049	0.151	0.880
Car.useCommute	0.100	0.111	0.896	0.370
Car.useFarmer	−0.852	0.311	−2.738	0.006 **
Car.usePrivate	0.018	0.119	0.149	0.881
Location.Cluster2	−0.002	0.128	−0.018	0.986
Location.Cluster3	−0.514	0.588	−0.874	0.382
Location.Cluster4	0.309	0.095	3.266	0.001 **
Location.Cluster5	0.027	0.418	0.064	0.949
Location.Cluster6	−0.097	0.113	−0.861	0.389
Location.Cluster7	0.531	0.148	3.578	0.0004 ***
Location.Cluster8	1.130	0.585	1.931	0.054
Location.Cluster9	0.018	0.092	0.191	0.849
Location.Cluster10	0.252	0.100	2.514	0.012 *
Location.Cluster11	1.543	0.653	2.362	0.018 *
Location.Cluster12	0.152	0.158	0.964	0.335
Location.Cluster13	0.059	0.106	0.557	0.577
Location.Cluster14	0.229	0.090	2.541	0.011 *
Annual.miles.drive	0.00000	0.00001	0.362	0.718

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table 5. The statistical significance of smooth terms for Insured Age, Car Age, Credit Scores, and Years of no Claims, respectively.

	edf	**Ref.df**	F	***p*-Value**
s (Insured.age)	7.560	8.351	1.973	0.044 *
s (Car.age)	6.748	7.709	4.325	0.000 ***
s (Credit.score)	7.460	8.339	14.568	0.000 ***
s (Years.noclaims)	8.012	8.617	3.276	0.001 ***

Note: * $p < 0.05$; *** $p < 0.001$.

In rate regulation, the rating variables are separated by vehicles and drivers. The Car age is often used to determine the vehicle's rate group and is considered a factor that affects the vehicle's rate group. The other three variables are related to drivers and are continuous in nature. Their impact on claim frequency and severity can be estimated by using risk relativity at each factor level. Therefore, the rate regulation of UBI can be extended to include the estimate of these risk factors, which may be used by traditional rate regulation one way or another, but obtaining the risk relativity for each level of a given factor can better reflect the pricing at an individual level. We should notice that rate regulation is still a collective basis rate making, but with some details that can be used to drive individual-level pricing toward the correct direction. So, Car Age, Credit Scores, Insured Age, and Years of no Claims can become additional rating variables besides the Annual Mileage Driven.

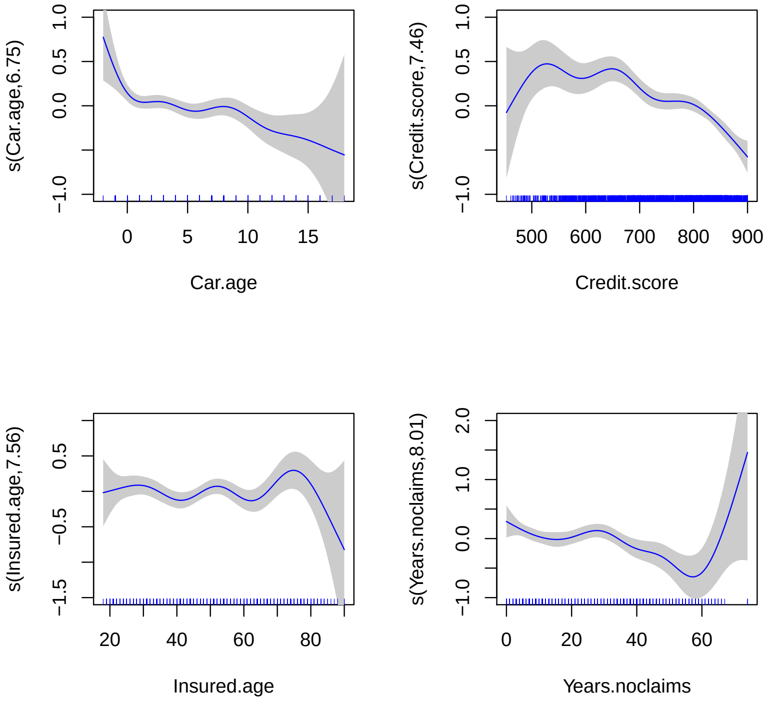
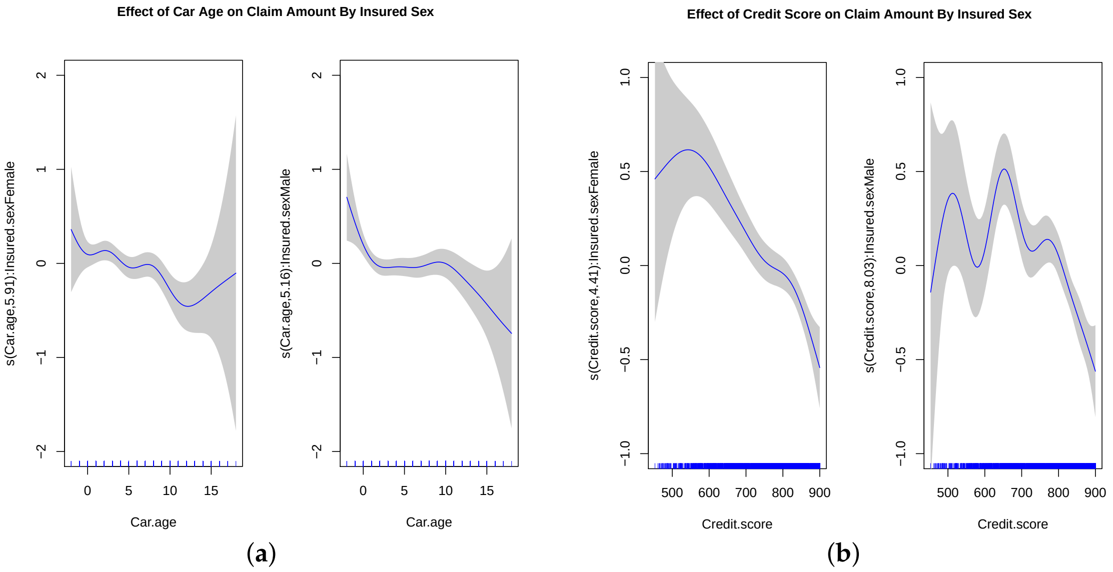


Figure 7. The plots of fitted spline functions, respectively, for Car age, Credit Scores, Annual miles drive, and Years of no claims.

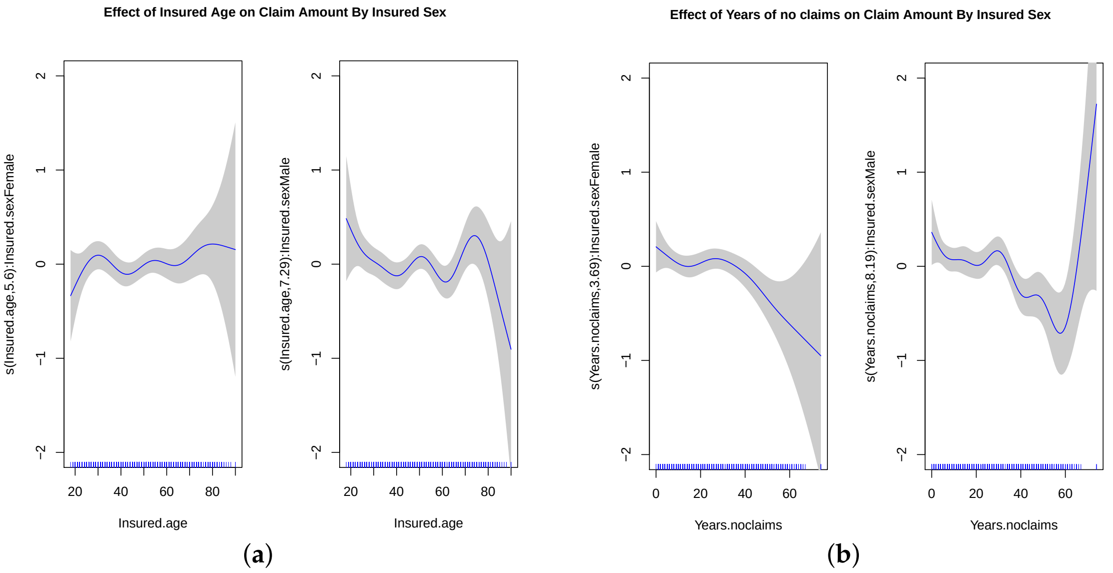
We observe that most of the categorical risk factors are traditional rating variables. In UBI, the new rating variables are mostly continuous, as the data were the outcome of continuously monitoring drivers' driving behaviours or location. Although we have yet to reach the stage of analysing such data, we should prepare for how one can explore this type of data to enhance the regulation of UBI. Because of this, our next focus is on investigating how these functional patterns of numerical variables interact with other variables, particularly the categorical variable. For instance, we may ask if the impact on the claim amounts from Car Age differs from female drivers to male drivers. Figures 8 and 9 show the conditional functional pattern for the numerical variables on gender. Overall, these functional patterns do not deviate significantly from the normal range of variables we consider. The difference is mainly located at the values either near the lower bound or upper bound of the interval, where we have higher sampling errors. This may suggest that functional patterns of Car Age, Credit Scores, Insured Age, or Years of no Claims do not depend on gender. Therefore, in rate regulation, gender should not be one of the rating variables to discriminate the potential risk. This result coincides with the requirement by the European Union in rate regulation of insurance.

Figures 10 and 11 show the results of how Car Age, Credit Scores, Insured Age and Years of no Claims interact with Marital status. First, we observe that the Car Age does not significantly interact with Marital Status, which makes sense to us. However, it seems to behave differently for the functional pattern of Credit Scores, Insured Age, and Years of no Claims between married and single drivers. This may imply that consideration needs to be given when pricing auto insurance policies for the different marital statuses of drivers. As for the interaction with Car Use, we also observe material differences among different levels of Car Use for the variables we consider. This may imply that pricing needs to be determined by different types of Car Use to further capture different risk levels among those with insurance. The obtained results are displayed in Figures 12–15. Note that Marital status, Car Use, and Insured Age are often combined to make another variable to avoid the issue prevented by regulation rule. For instance, in Canada, these three variables are combined to become a new variable called Type of Use, one of the major factors considered in rate regulation. Because of the potential interactions illustrated in Figures 12–15, a regulator may have to separate the data by these variables or the new variable that combines all these together to improve the performance of rate regulation of UBI.



(**a**) (**b**)

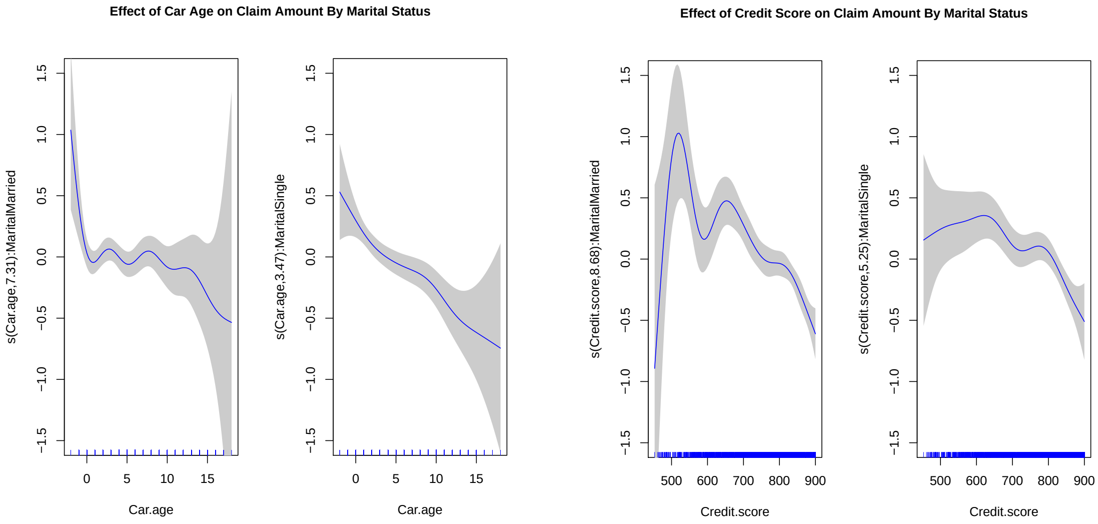
Figure 8. The plots of fitted spline functions for Car Age and Credit Scores, separated by Insured Sex. (**a**) Car Age by Insured Sex; (**b**) Credit Scores by Insured Sex.



(**a**) (**b**)

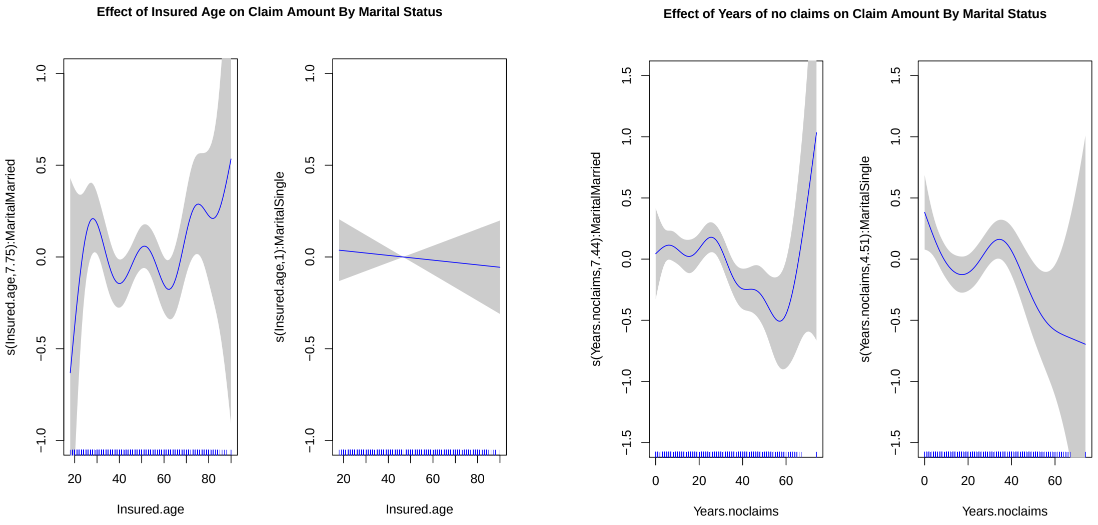
Figure 9. The plots of fitted spline functions for Insured Age and Years of no Claims, separated by Insured Sex. (**a**) Insured Age by Insured Sex; (**b**) Year of no Claims by Insured Sex.

To show the statistical significance of the designed rating territory, we conducted the ANOVA F test to compare two GAM models, one with rating territory and the other without rating territory. From the obtained results shown in Tables 6 and 7, we observe that the p-values for testing the reduction in model variation are equal to zero, implying that contributions from rating territory are all significant for the two models, i.e., the model for predicting claim probability and the model for predicting claim amounts. Furthermore, rating territory as a model predictor influences claims probability more than claim amounts because of small deviance, leading to a much higher value of test statistics. This is also verified by model performance measures GCV (i.e., Generalised Cross-Validation Statistic) and AIC (i.e., Akaike Information Criterion) shown in Table 8. Both cross-validated model errors and AIC values are smaller for GAM models with rating territory as an additional model predictor. Although we realise that BIC (i.e., Bayesian Information Criterion) is higher for the model with Territory, this may be due to many factor-level estimations that need to be carried out to increase BIC's values. The dataset also contains a fair amount of extreme losses, leading to high values of GCV for the GAM model that is used to predict claim amounts. This may be another reason for the inconsistency in performance evaluation.



(**a**) (**b**)

Figure 10. The plots of fitted spline functions for Car Age and Credit Scores, separated by Marital Status. (**a**) Car Age by Marital Status; (**b**) Credit Scores by Marital Status.



(**a**) (**b**)

Figure 11. The plots of fitted spline functions for Insured Age and Years of no Claims, separated by Marital Status. (**a**) Insured Age by Marital Status; (**b**) Years of no Claims by Marital Status.

Finally, rate relativities for different risk factors are estimated through GLM modelling. The reason for using GLM rather than GAM modelling is that we have to calculate the risk relativity for each level of a given risk factor, including both numerical and categorical variables. For numerical values, we must first group them into different categories. The modelling assumes that the risk relativity is the same for each level of a given risk factor. These estimated risk relativities are summarised in Table 9. Using GLM, we derive the risk relativity for claim probability, claim amounts, and the loss cost. The loss cost relativity for Insured Age appears to be decreasing with the increase in age. Gender and Marital have no discriminative power, and the relativity decreases with the increase in Car Age. Particularly, new cars have much higher risk relativity than older cars. Commercial and commuter Cars are riskier than other types of Car Use. The relativity of high Credit Scores is much lower than that of low Credit Scores. This implies that drivers who have high Credit Scores are a greater risk. The relativity of Years of no Claims is also significantly decreased with the increase in the number of Years of no Claims. Different rating territories have different relativities, some of which have either extremely low or very high relativity, such as Cluster 3 and 11. The Annual Mileage Driven variable has significantly high relativities caused by the high claim probability. To further study the functionality of relativity for important risk factors, including Annual Mileage Driven, Credit Scores, and Years of no Claims, we summarised the results obtained in Figures 16–18. From the displayed results, we observe that Annual Mileage Driven as a rating variable is dominated by claim frequency, and the relativities estimated by using claim amounts are close for most of the levels of such factors. The sudden decrease in relativity for the bracket with the largest amount of driving may be due to insufficient observation of large losses, as we still observe high relativity for claim probability. Unlike the Annual Mileage Driven variable, the risk relativities for Credit Scores and Years of no Claims are significantly affected by claim frequency and amounts.

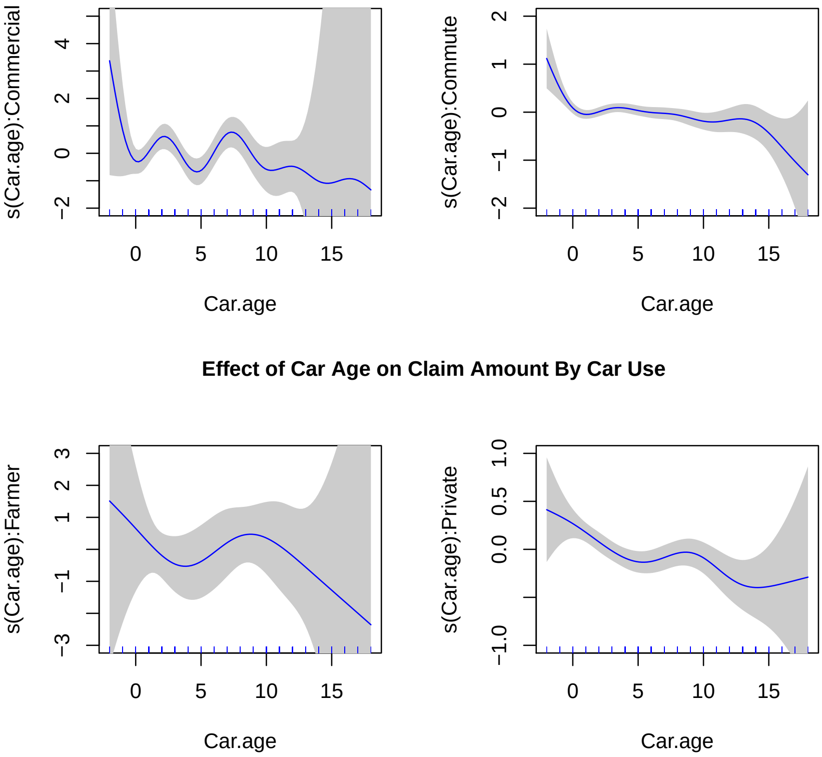


Figure 12. The plots of fitted spline functions for Car Age, separated by Car Use.

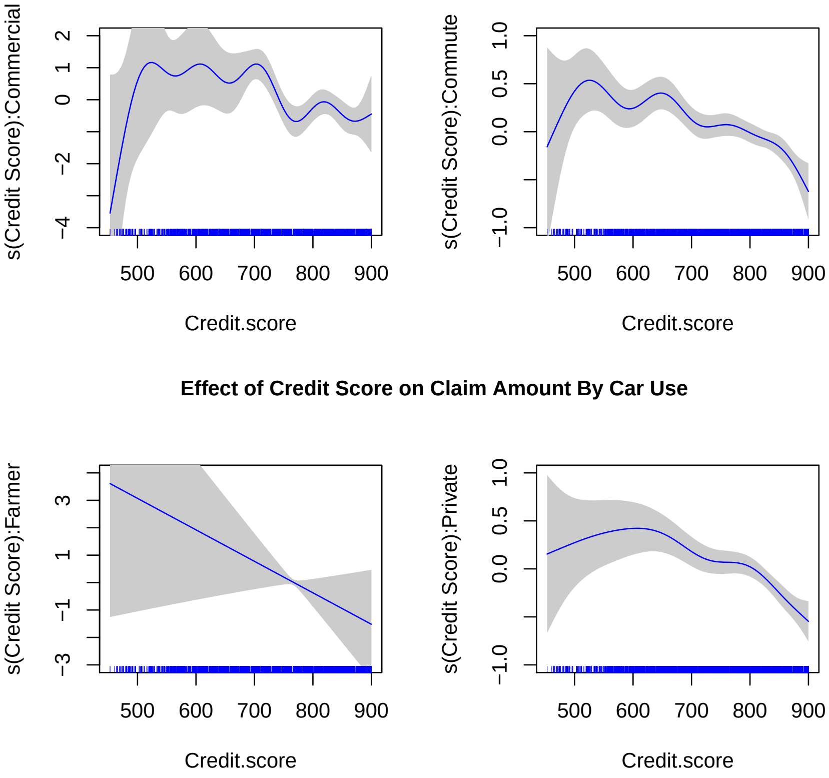


Figure 13. The plots of fitted spline functions for Credit Scores, separated by Car Use.

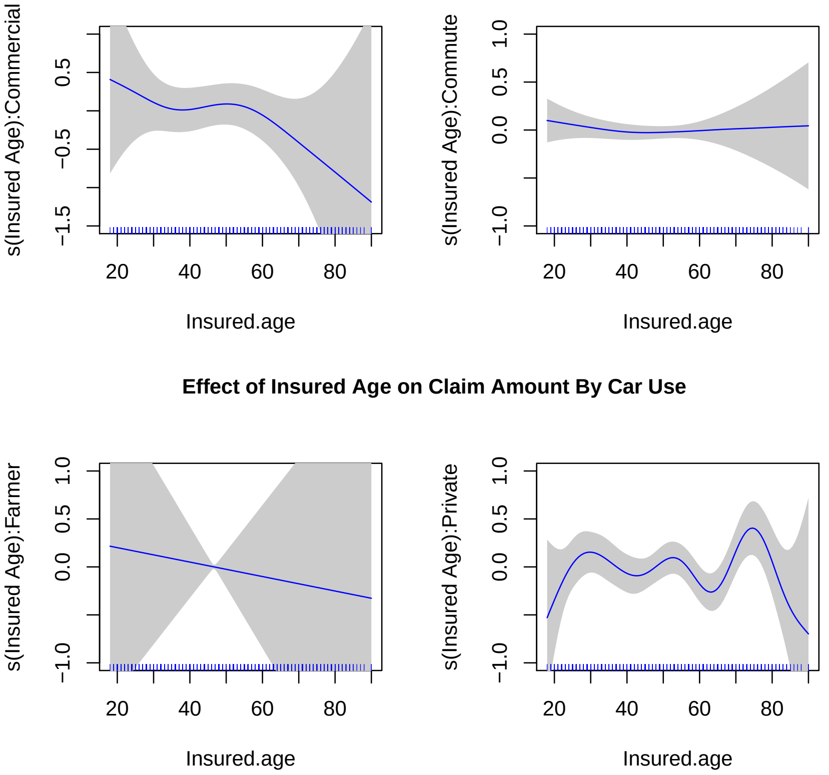


Figure 14. The plots of fitted spline functions for Insured Age, separated by Car Use.

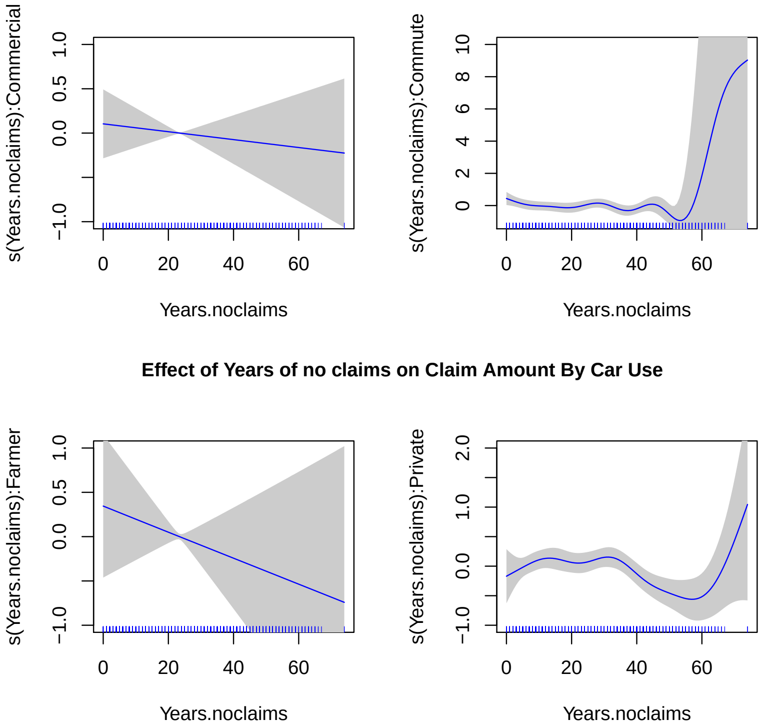


Figure 15. The plots of fitted spline functions for Years of no Claims, separated by Car Use.

Table 6. Statistical test results and summary statistics for the GAM modelling of claim probability, with and without designed rating territories. Note that the test is based on Deviance, which is a goodness-of-fit statistic for the model, calculated based on the log likelihood-ratio; Pr (>F) is the *p*-value of the test. "Resid. Df" means Residual Degree of freedom; "Resid. Dev" means Residual Standard Deviation; "Df" means Degree of freedom.

Statistic	N	Mean	St. Dev.	Min	Max
Resid. Df	2	99,955.380	9.308	99, 948.800	99,961.960
Resid. Dev	2	25,491.370	75.184	25,438.200	25,544.530
Df	1	−13.164		−13.164	−13.164
Deviance	1	−106.327		−106.327	−106.327
F	1	8.077		8.077	8.077
Pr (>F)	1	0.000		0	0

Table 7. Statistical test result and summary statistics for the GAM modelling of claim amounts, with and without designed rating territories. Note that the test is based on Deviance, which is a goodness-of-fit statistic for the model, calculated based on the log likelihood-ratio; Pr (>F) is the *p*-value of the test. "Resid. Df" means Residual Degree of freedom; "Resid. Dev" means Residual Standard Deviation; "Df" means Degree of freedom.

Statistic	N	Mean	St. Dev.	Min	Max
Resid. Df	2	3812.783	9.292	3806.213	3819.354
Resid. Dev	2	4873.449	68.969	4824.681	4922.217
Df	1	−13.141		−13.141	−13.141
Deviance	1	−97.536		−97.536	−97.536
F	1	4.553		4.553	4.553
Pr (>F)	1	0.00000		0.00000	0.00000

Table 8. Performance comparison between GAM models for predicting claim probability and claim amounts, respectively. Note: GCV: Generalised Cross-Validation Statistic; AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion.

Claim Probability Model			
	GCV	AIC	BIC
GAM with clusters	0.04021174	34,076.88	34,527.22
GAM without clusters	0.04026616	34,156.86	34,481.85
Claim Amounts Model			
GAM with clusters	26,227,331	70,317.51	70,658.48
GAM without clusters	27,176,211	70,383.03	70,641.44

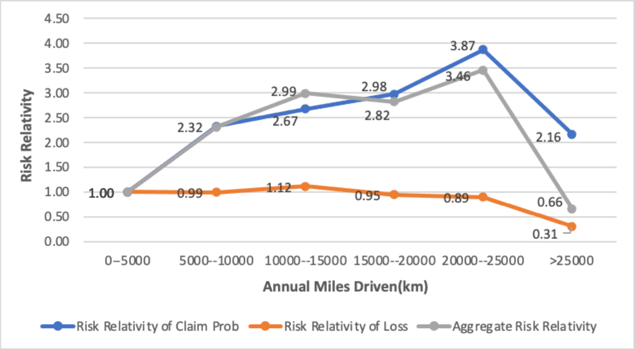


Figure 16. Risk relativities by grouped Annual Mileage Driven based on loss severity, frequency, and loss cost (aggregate risk), respectively.

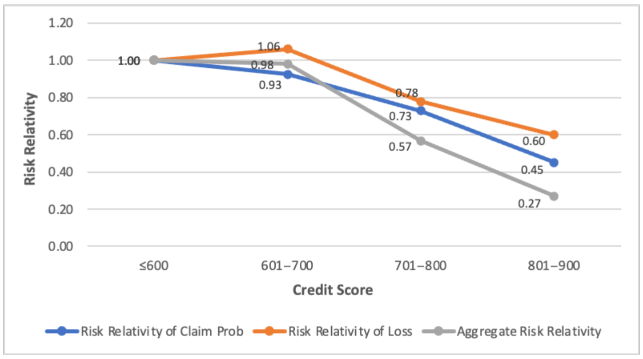


Figure 17. Risk relativities by grouped Credit Scores based on loss severity, frequency, and loss cost (aggregate risk), respectively.

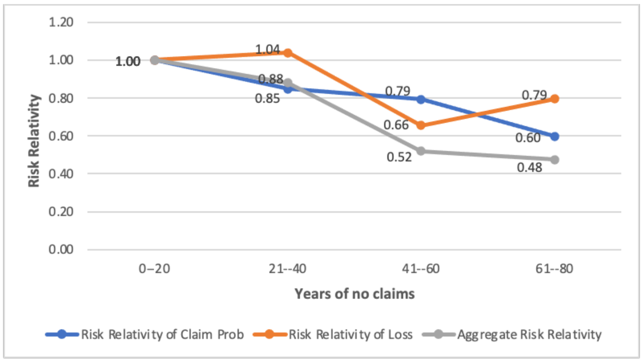


Figure 18. Risk relativities by grouped Years of no Claims based on loss severity, frequency, and loss cost (aggregate risk), respectively.

Table 9. Risk relativity estimate for categorical factors and numerical factors with designed levels.

Variable	Risk Relativity of Claim Prob (With Poisson)	Risk Relativity of Claim Amount (With Gamma)	Aggregate Risk Relativity (Poisson + Gamma)
Insured.Age. 16–22	1.00	1.00	1.00
Insured.Age. 23–35	0.69	0.91	0.63
Insured.Age. 36–45	0.65	0.75	0.49
Insured.Age. 46–65	0.71	0.79	0.57
Insured.Age. 65+	0.53	0.89	0.47
Female	1.00	1.00	1.00
Male	0.99	1.03	1.03
Car.Age ≤ 0	1.00	1.00	1.00
Car.Age. 1–5	0.91	0.91	0.83
Car.Age. 6–10	0.64	0.83	0.53
Car.Age. 11–15	0.41	0.64	0.26
Car.Age $\geq$16	0.24	0.66	0.16
Married	1.00	1.00	1.00
Single	1.03	0.99	1.03
Car.Use.Commercial	1.00	1.00	1.00
Car.Use.Commute	0.81	1.05	0.85
Car.Use.Farmer	0.42	0.44	0.18
Car.Use.Private	0.81	0.96	0.77
Credit.Score $\leq$600	1.00	1.00	1.00
Credit.Score. 601–700	0.93	1.06	0.98
Credit.Score. 701–800	0.73	0.78	0.57
Credit.Score. 801–900	0.45	0.60	0.27
Location.Cluster1	1.00	1.00	1.00
Location.Cluster2	1.18	0.95	1.12
Location.Cluster3	**0.35**	**0.45**	**0.16**
Location.Cluster4	1.23	1.36	1.67
Location.Cluster5	1.48	0.95	1.40
Location.Cluster6	0.79	0.89	0.70
Location.Cluster7	1.07	1.57	1.69
Location.Cluster8	0.49	3.55	1.74
Location.Cluster9	1.00	0.99	1.00
Location.Cluster10	1.20	1.26	1.51
Location.Cluster11	1.04	5.04	5.23
Location.Cluster12	0.60	1.07	0.65
Location.Cluster13	0.93	1.03	0.96
Location.Cluster14	1.11	1.21	1.34
Annual.Miles. 0–5000	1.00	1.00	1.00
Annual.Miles. 5000–10,000	**2.32**	0.99	**2.31**
Annual.Miles. 10,000–15,000	**2.67**	1.12	**2.99**
Annual.Miles. 15,000–20,000	**2.98**	0.95	**2.82**
Annual.Miles. 20,000–25,000	**3.87**	0.89	**3.46**
Annual.Miles. 25,000+	2.16	0.31	**0.66**
Years.noclaims. 0–20	1.00	1.00	1.00
Years.noclaims. 21–40	0.85	1.04	0.88
Years.noclaims. 41–60	0.79	0.66	0.52
Years.noclaims. 61–80	0.60	0.79	0.48

4. Conclusions

Automobile insurance has shifted from collective to individual-level pricing as usage-based insurance has become more prevalent. As a result of such a shift in focus from traditional pricing, it is intended to further discriminate drivers to attract more drivers with reasonable risks. By better distinguishing between insurance risk classes, the insurance company can enhance the implementation of actuarial pricing principles, as each driver should be responsible for his/her own risk. From a rate regulation standpoint, the rate-making and rate classification techniques associated with UBI are still under development, and rate regulation for UBI is premature. In this work, we have devoted ourselves to exploring what rating variables can potentially be used for rate regulation of UBI, where the variables are highly discriminative on an aggregated basis and can be truly reflective when it comes to the measure of individual-level risk. Additionally, as regulation is collective, these major risk factors could be easily aggregated, while at the same time retaining some representative risks. We conducted predictive modelling based on GAM to forecast claim probability and amounts. Our study showed that variables of Insured Age, Car Age, Credit Scores, Annual Mileage Driven, and Years of no Claims might serve as major risk factors for regulating UBI. The estimated risk relativity for these variables at the group level may be used as a benchmark for regulation purposes. Particularly, we find that relativity for high annual mileage driven is almost three times greater than that associated with low annual mileage level, which implies its importance in premium calculation. Additionally, the risk relativity estimate for Year of no claims is in line with the estimates of relativity for Driving Record (DR) that were previously studied in [37]. It is considered an extension of the study of DR, as the year of no claims variable covers almost a lifetime of driving history, but DR only focuses on the first seven years with right truncation at year seven. In addition, the GAM modelling technique is beneficial for capturing the functional pattern of a given risk factor. The exploration of such functionality over major risk factors reveals the important non-linear relationship between major risk factors and claim probability or amounts. Moreover, the additive nature of its components in the GAM model helps improve its interpretability.

From a rate regulation perspective, rating territory is critical, as it has been proven that territory risk is highly discriminative and contributes significantly to insurance pricing. We have also extended this study to territory clustering using a low-dimensional feature subset of insurance risk. Our study has demonstrated that a combination of claim frequency and severity becomes an optimal feature vector, leading to a suboptimal design of rating territory. The sub-optimality applies in the sense that it is optimal among all other choices. The zero p-values for the significance test of extending the GAM model by including rating territory show strong evidence of the importance of such a variable in modelling the claim probability and amounts. Furthermore, relativities associated with territory level are significantly discriminative, meaning that premium level will be affected considerably when the territory is used as a factor in pricing. This finding extends the similar results from traditional auto insurance to UBI and confirms that rating territory is key for both types of insurance. Since this study aims to illustrate how rating territory can be retained to regulate UBI, their relativities associated with different levels of rating territory were produced. The estimated risk relativity suggested that the designed rating territory has a discriminative power to differentiate the level of territory risk. Combining Urban or Rural with the driver's location area to design rating territory allows us to create more basic rating units for a better clustering result. Our future work will further develop statistical techniques suitable for regulating UBI and continue identifying more major risk factors from the UBI database. In addition, we will extend our predictive modelling by including telematics variables. Due to the high dimensionality and complexity of telematics variables, we plan to find a way of transforming telematics variables into a low-dimensional feature subspace. Therefore, it is necessary to have explainable and interpretable statistical models for auto insurance rate regulation. However, we will not derive the risk relativity for telematics variables; instead, we will consider how they can be used to improve predictive modelling performance. All of these factors contribute to a better understanding of the complex UBI system.

Author Contributions: Conceptualization, S. Xie; methodology, S. Xie; software, K.Shi; validation, S. Xie and K.Shi; formal analysis, S. Xie and K.Shi; investigation, S. Xie and K.Shi; resources, S. Xie and K.Shi; data curation, S. Xie and K.Shi; writing—original draft preparation, S. Xie; writing—review and editing, S. Xie; visualization, K.Shi; supervision,S. Xie; project administration, S. Xie. All authors have read and agreed to the published version of the manuscript.

Funding: There is no funding support for this research project.

Data Availability Statement: The data belongs to the regulator and they are available upon request. Code is available upon request.

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

References

1. Śliwiński, A.; Kuryłowicz, Ł. Usage-based insurance and its acceptance: An empirical approach. *Risk Manag. Insur. Rev.* **2021**, *24*, 71–91. [CrossRef]
2. Arumugam, S.; Bhargavi, R. A survey on driving behavior analysis in usage based insurance using big data. *J. Big Data* **2019**, *6*, 86. [CrossRef]

3. Hu, X.; Zhu, X.; Ma, Y.L.; Chiu, Y.C.; Tang, Q. Advancing usage-based insurance—A contextual driving risk modelling and analysis approach. *IET Intell. Transp. Syst.* **2019**, *13*, 453–460. [CrossRef]
4. Händel, P.; Ohlsson, J.; Ohlsson, M.; Skog, I.; Nygren, E. Smartphone-based measurement systems for road vehicle traffic monitoring and usage-based insurance. *IEEE Syst. J.* **2013**, *8*, 1238–1248. [CrossRef]
5. Huang, Y.; Meng, S. Automobile insurance classification ratemaking based on telematics driving data. *Decis. Support Syst.* **2019**, *127*, 113156. [CrossRef]
6. Che, X.; Liebenberg, A.; Xu, J. Usage-Based Insurance—Impact on Insurers and Potential Implications for InsurTech. *N. Am. Actuar. J.* **2021**, 1–28. [CrossRef]
7. Barry, L.; Charpentier, A. Personalization as a promise: Can Big Data change the practice of insurance? *Big Data Soc.* **2020**, *7*, 2053951720935143. [CrossRef]
8. Zhang, J.; Miljkovic, T. *Ratemaking for a New Territory: Enhancing glm Pricing Model with a Bayesian Analysis*; Casualty Actuarial Society: Arlington County, VA, USA, 2019.
9. Henckaerts, R.; Antonio, K.; Clijsters, M.; Verbelen, R. A data driven binning strategy for the construction of insurance tariff classes. *Scand. Actuar. J.* **2018**, *2018*, 681–705. [CrossRef]
10. Bian, Y.; Yang, C.; Zhao, J.L.; Liang, L. Good drivers pay less: A study of usage-based vehicle insurance models. *Transp. Res. Part Policy Pract.* **2018**, *107*, 20–34. [CrossRef]
11. Ohlsson, E.; Johansson, B. *Non-Life Insurance Pricing with Generalized Linear Models*; Springer: Berlin, Germany, 2010; Volume 2.
12. Francis, R.A.; Geedipally, S.R.; Guikema, S.D.; Dhavala, S.S.; Lord, D.; LaRocca, S. Characterizing the performance of the conway-maxwell poisson generalized linear model. *Risk Anal. Int. J.* **2012**, *32*, 167–183. [CrossRef]
13. Cunha, L.; Bravo, J.M. Automobile Usage-Based-Insurance: Improving Risk Management using Telematics Data. In Proceedings of the 2022 17th Iberian Conference on Information Systems and Technologies (CISTI), Madrid, Spain, 22–25 June 2022; pp. 1–6.
14. Ma, Y.L.; Zhu, X.; Hu, X.; Chiu, Y.C. The use of context-sensitive insurance telematics data in auto insurance rate making. *Transp. Res. Part Policy Pract.* **2018**, *113*, 243–258. [CrossRef]
15. Kuo, K.; Lupton, D. Towards explainability of machine learning models in insurance pricing. *arXiv* **2020**, arXiv:2003.10674.
16. Denuit, M.; Charpentier, A.; Trufin, J. Autocalibration and Tweedie-dominance for insurance pricing with machine learning. *Insur. Math. Econ.* **2021**, *101*, 485–497. [CrossRef]
17. 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.
18. Pena-Reyes, C.A.; Sipper, M. Fuzzy CoCo: Balancing accuracy and interpretability of fuzzy models by means of coevolution. In *Accuracy Improvements in Linguistic Fuzzy Modeling*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 119–146.
19. Casillas, J.; Cordón, O.; Herrera, F.; Magdalena, L. Interpretability improvements to find the balance interpretability-accuracy in fuzzy modeling: An overview. In *Interpretability Issues in Fuzzy Modeling*; Casillas, J., Cordón, O., Herrera, F., Magdalena, L., Eds.; Springer: Berlin/Heidelberg, Germany, 2003; pp. 3–22.
20. Boucher, J.P.; Turcotte, R. A longitudinal analysis of the impact of distance driven on the probability of car accidents. *Risks* **2020**, *8*, 91. [CrossRef]
21. Guillen, M.; Nielsen, J.P.; Pérez-Marín, A.M.; Elpidorou, V. Can automobile insurance telematics predict the risk of near-miss events? *N. Am. Actuar. J.* **2020**, *24*, 141–152. [CrossRef]
22. Zahi, J. Non-life insurance ratemaking techniques. *Int. J. Account. Financ. Audit. Manag. Econ.* **2021**, *2*, 344–361.
23. Schumaker, L. *Spline Functions: Basic Theory*; Cambridge University Press: Cambridge, UK, 2007.
24. Bett, N.; Kasozi, J.; Ruturwa, D. Temporal Clustering of the Causes of Death for Mortality Modelling. *Risks* **2022**, *10*, 99. [CrossRef]
25. Gan, G.; Valdez, E.A. Data clustering with actuarial applications. *N. Am. Actuar. J.* **2020**, *24*, 168–186. [CrossRef]
26. Peters, G. Statistical Machine Learning and Data Analytic Methods for Risk and Insurance. 2017. Available online: https://ssrn.com/abstract=3050592 (accessed on 1 November 2021).
27. Xie, S.; Gan, C. Fuzzy Clustering and Non-negative Sparse Matrix Approximation on Estimating Territory Risk Relativities. In Proceedings of the 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Hyderabad, India, 7–10 July 2013; pp. 1–8.
28. Xie, S.; Gan, C.; Chua-Chow, C. Estimating Territory Risk Relativity for Auto Insurance Rate Regulation using Generalized Linear Mixed Models. In Proceedings of the 10th International Conference on Data Science, Technology and Applications (DATA 2021), Online, 6–8 July 2021; pp. 329–334.
29. So, B.; Boucher, J.P.; Valdez, E.A. Synthetic Dataset Generation of Driver Telematics. *Risks* **2021**, *9*, 58. [CrossRef]
30. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
31. Hastie, T.J. Generalized additive models. In *Statistical Models in S*; Routledge: London, UK, 2017; pp. 249–307.
32. Hastie, T.; Tibshirani, R. Generalized additive models: Some applications. *J. Am. Stat. Assoc.* **1987**, *82*, 371–386. [CrossRef]
33. Ostertagová, E. Modelling using polynomial regression. *Procedia Eng.* **2012**, *48*, 500–506. [CrossRef]
34. Bemporad, A. Piecewise linear regression and classification. *arXiv* **2021**, arXiv:2103.06189.
35. Wuthrich, M.V.; Buser, C. Data Analytics for Non-Life Insurance Pricing. Swiss Finance Institute Research Paper. 2021. Available online: https://ssrn.com/abstract=2870308 (accessed on 1 November 2021).

36. Maindonald, J. Smoothing Terms in GAM Models. 2010. Available online: https://maths-people.anu.edu.au/~johnm/r-book/xtras/autosmooth.pdf (accessed on 1 November 2021).
37. Xie, S.; Lawniczak, A.T. Estimating major risk factor relativities in rate filings using generalized linear models. *Int. J. Financ. Stud.* **2018**, *6*, 84. [CrossRef]

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.