# “Predicting motor insurance claim incidence using generalized and tree-based models: A comparative statistical approach”


| | |
|---|---|
| **AUTHORS** | Eslam Abdelhakim Seyam |



| | |
|---|---|
| **ARTICLE INFO** | Eslam Abdelhakim Seyam (2025). Predicting motor insurance claim incidence using generalized and tree-based models: A comparative statistical approach. *Insurance Markets and Companies*, *16*(2), 38-53. doi:10.21511/ins.16(2).2025.04 |
| **DOI** | http://dx.doi.org/10.21511/ins.16(2).2025.04 |
| **RELEASED ON** | Thursday, 14 August 2025 |
| **RECEIVED ON** | Saturday, 10 May 2025 |
| **ACCEPTED ON** | Thursday, 07 August 2025 |





| | |
|---|---|
| **JOURNAL** | "Insurance Markets and Companies" |
| **ISSN PRINT** | 2616-3551 |
| **ISSN ONLINE** | 2522-9591 |
| **PUBLISHER** | LLC “Consulting Publishing Company “Business Perspectives” |
| **FOUNDER** | LLC “Consulting Publishing Company “Business Perspectives” |

| NUMBER OF REFERENCES | NUMBER OF FIGURES | NUMBER OF TABLES |
|---|---|---|
| 37 | 5 | 7 |

BUSINESS PERSPECTIVES

LLC "CPC "Business Perspectives"
Hryhorii Skovoroda lane, 10,
Sumy, 40022, Ukraine

www.businessperspectives.org

**Eslam Abdelhakim Seyam** (Saudi Arabia)

# PREDICTING MOTOR INSURANCE CLAIM INCIDENCE USING GENERALIZED AND TREE-BASED MODELS: A COMPARATIVE STATISTICAL APPROACH

## Abstract

Accurate prediction of motor insurance claim frequency is necessary for efficient risk management, underwriting, and policy pricing. Predictive performance of Poisson Generalized Linear Models (GLMs), Decision Trees, and Generalized Additive Models (GAMs) is investigated using 108,699 motor third-party liability insurance contracts, representing the French Motor TPL dataset from the CASdatasets R package widely used in actuarial research. These models' predictability, explainability, and flexibility on training and testing sets are compared using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Poisson Deviance metrics. Results indicate that, although GLM offers an interpretable, accurate baseline, GAM slightly surpasses GLM and Decision Trees under all performance measures. Results demonstrate that GAM achieves superior performance across all metrics, with the lowest MSE (0.0506), RMSE (0.2251), and Poisson Deviance (36.41% training, 37.76% test), compared to GLM (MSE: 0.0509, RMSE: 0.2257, Poisson Deviance: 36.83% training, 38.08% test) and Decision Trees (MSE: 0.0582, RMSE: 0.2413, Poisson Deviance: 37.12% training, 38.31% test). The GAM model reduces prediction error by approximately 0.6% compared to GLM and 13.1% compared to Decision Trees based on MSE. Empirical findings reveal how GAMs achieve an optimum balance between model explainability and prediction flexibility, rendering them best suited for insurers who want to refine risk segmentation without compromising on regulatory compliance and business transparency. This study joins other research calling for interpretable state-of-the-art statistical techniques in insurance analytics and presents worthwhile observations for actuaries and data scientists who wish to refine motor insurance frequency modeling frameworks.

**Keywords** motor insurance, claim frequency, generalized linear models, decision trees, generalized additive models, predictive modeling

**JEL Classification** C25, C53, G22, C14, C52

**Type of the article:** Research Article

**Received on:** 10th of May, 2025
**Accepted on:** 7th of August, 2025
**Published on:** 14th of August, 2025



Eslam Abdelhakim Seyam, Ph.D., College of Business, Department of Insurance and Risk Management, Imam Mohammad Ibn Saud Islamic University (IMSIU), Saudi Arabia.



**Conflict of interest statement:**
Author(s) reported no conflict of interest

## INTRODUCTION

Claims frequency modeling for motor insurance is among the foundations of risk management, underwriting, and premium ratemaking for insurers (Wilson et al., 2024). Accurate estimation of claim numbers helps insurers stratify risk effectively, reserve sufficiently, and maintain competitive but sensible ratemaking frameworks. Conventionally, Generalized Linear Models (GLMs) under Poisson assumptions were best used for modeling claim frequency because of their interpretability and strong theory to support them (Denuit & Lang, 2004; Antonio & Valdez, 2012). Real-world insurance data, however, are usually plagued with nonlinearities, high-order variable interactions, and overdispersion that limit traditional GLMs' behavior (Clemente et al., 2023).

Although GLMs have been dependable workhorses for insurance analytics, their linear and additive nature usually cannot accommodate



http://dx.doi.org/10.21511/ins.16(2).2025.04

subtle patterns of policyholder behavior. That limits the ability of the models to generate accurate prices and risk estimates, especially for today's information-rich environment where insurers have telematics and behavioral information. More flexible modeling structures would likely accommodate such nuances and provide superior predictive accuracy for motor insurance claim frequency modeling.

The fundamental challenge for today's insurance practitioners is how to reconcile model predictiveness against interpretability. Standard GLMs are interpretable and meet regulatory approval requirements, but often fail to capture the nonlinear relationships inherent in new trends in insurance data. Modern machine learning models are typically more precise but are not explanatory enough for obtaining regulatory approval and for business decisions. The challenge poses an important operational dilemma where insurers desire models that can achieve the best predictive efficacy but are still interpretable for practical application in risk management and pricing.

The work answers insurers' practical need to identify methodological traditions capable of overcoming the accuracy-interpretability dichotomy. It consists of a retrospective comparison among three diverging methodological traditions – GLMs, Decision Trees, and Generalized Additive Models (GAMs) – to identify their relative strengths for motor insurance claim frequency prediction. Presenting empirical evidence of model behavior across key performance measures ex post, the work offers insurers and actuaries information on which to decide on model selection for risk segmentation and price setting. The findings contribute to the growing need for interpretable but flexible statistical models for insurance analytics usage, aligning with actuaries' practical needs and propelling methodological knowledge ahead (Díaz Martínez et al., 2023).

## 1. LITERATURE REVIEW

The modeling approaches for motor claims frequency have been transformed during previous decades due to new statistical techniques, abundant computing capacity, and rich sources of quality policyholder information. This revolution responds to insurers' classic call for better risk assessment tools, which harmonize predictive sophistication and regulation, and commercial accountability. An understanding of such a history of development is significant when designing new modeling approaches of the contemporary era and charting methodological trajectories for the future.

Generalized Linear Models have remained at the core of motor insurance claim frequency modeling to this day since their introduction by Nelder and Wedderburn (1972) and subsequent refinement by McCullagh and Nelder (1989). Poisson GLMs' popularity can be attributed to interpretability, well-developed theory, and conformity with rules of transparency demanded from regulators, foreshadowed early on by pioneering contributions of Denuit and Lang (2004) and Antonio and Valdez (2012). These models admit of precise parametric specifications mapping onto multiplicative rating factors well established in actuarial techniques, comprehensive illustrations recorded among others by Anderson et al. (2004), Goldburd et al. (2016), and Ohlsson and Johansson (2010). Nevertheless, inherent constraints are increasingly patent for increasingly intricate insurance data. Disavowal of an assumption of equidispersion comes all too easily for real-world claim frequency information, such that alternative corrections are needed. Pioneering early breakthroughs from Dionne and Vanasse (1989) opened up Poisson-Negative Binomial mixture formulations, which were later extended through zero-inflated and hurdle formulations by Cameron and Trivedi (1998), Hilbe (2011), and Kafková and Křivánková (2014) to address excessive zeroes from collections of insurance claims. Taken together, these extensions mirror appreciation from the actuarial community that classical GLM assumptions cannot vainly hope to mirror real-world complexity for actuarial phenomena.

The awareness of GLM constraints has prompted exploration of more adaptive methodologies capable of distinguishing intricate patterns while maintaining proper interpretability levels. Tree-based procedures are a notable methodological

advancement, expanding on the CART framework established by Breiman et al. (1984) and refining through ensemble processes envisaged by Breiman (2001) and Friedman (2001). Empirical findings from recently obtained results describe exemplary tree-based performance for problems of insurance, including research findings from Staudt and Wagner (2021) demonstrating that random forests are competent to perceive claim severity nonlinearities undetectable for linear methodologies. The advantage goes beyond single methodologies, such that a comprehensive comparison study by König and Loser (2024) determines that decision-tree-based ensemble methodologies universally surpass classical methodologies across tabular sets of insurance, including interpretability challenges consideration. The trend towards sophistication continues for higher-end applications where Clemente et al. (2023) demonstrate gradient boosting superiority over base GLMs for claim frequency prediction, and Meng et al. (2022) construct integrated telematics networks that blend driving behavior predictors and higher-end boosted trees. These collective findings suggest that tree-based methodologies offer an appealing alternative to classical methodologies, particularly when dealing with complex, high-dimensional sets of insurance.

The use of machine learning methods is confronted with significant challenges in interpretability and gaining regulatory approval, and therefore, there lies an intrinsic tension in actuarial practice. The overriding importance of model interpretability for use on insurance, for which attention was called by Kuo and Lupton (2021), creates a need for methods such as permutation feature importance and partial dependence plots. Latest developments on explainable AI, and notably SHAP values, hold exciting directions for transforming previously black-box models to interpretable frameworks without compromising predictive excellence. Generalized Additive Models yield strong middle-ground solutions that fill between GLM interpretability and such flexibility required to capture nonlinear relations based on seminal works of Hastie and Tibshirani (1986, 1990), which put GAMs on natural extensions of GLMs where linear terms are substituted by smooth functions but still possessing an additive structure for interpretational ease.

Applications of GAMs to insurance hold immense promise, and comprehensive methodological treatises by Wood (2006, 2011) make GAMs increasingly accessible to practitioners. The appeal extends to niche applications, where Denuit and Lang (2004) demonstrate Bayesian GAMs for non-life rate-making problems, and Klein et al. (2014) deploy them for treatment of location, scale, and shape parameters for integrated risk modeling paradigms. Empirical validations from recent periods give firm support for the efficacy of GAMs, especially through a decisive comparison study on Spanish motor insurance records by Díaz Martínez et al. (2023), which demonstrates that GAMs identify risk profiles through smooth curves representing patterns invisible to linear models while guaranteeing interpretability for regulation compliance. This facility for managing complex interrelations while guaranteeing interpretability receives further support from Xie and Shi (2023), which demonstrates the efficacy of GAMs for managing complex risk factor interactions for automobile applications. Advanced methodological extensions by Henckaerts et al. (2018, 2019) further enhance the usefulness of GAMs through new data-driven binning operations and merging of GAMs with tree-based machine learning processes. These cumulative advances make GAMs highly appealing tools that transcend and dominate the accuracy-interpretability compromise hampering contemporary actuarial practice.

The integration of telematics technology revolutionized motor insurance modeling through unrivaled risk assessment potential and novel methodological challenges. Verbelen et al.'s (2018) findings indicate that telematics information introduces predictive power previously unknown to traditional models, fundamentally shifting paradigms for risk classification. Their findings indicate that real-time monitoring of driving behavior provides information beyond that from static policyholder attribute-based traditional rating factors. Innovations of usage-based insurance demand new modeling frameworks capable of handling high-dimensional behavioral information with complex temporal structures. Premium recalibration on demand through telematics-enabled data is discussed by Henckaerts and Antonio (2022), and pioneer work by Ayuso et al. (2016, 2019) contributes much-needed insight on merging mile-


http://dx.doi.org/10.21511/ins.16(2).2025.04

age and driving behavior information. Pioneer telematics usage for risk differentiation is covered by Baecke and Bocca (2017) and Paefgen et al. (2013). Telematics data complexity cannot be duly supervised via traditional GLMs and therefore demands special-purpose methodologies designed by Meng et al. (2022) for handling varied driving behavior features and complex interactions thereof, fundamentally replacing traditional GLMs with boosted trees but remaining faithful to actuarial principles through combination with traditional distributional assumptions. This set of work together demonstrates how telematics information both disrupts traditional modeling paradigms and introduces novel and more advanced risk assessment opportunities.

Recent comprehensive comparisons provide illuminating information on relative effectiveness between different modeling approaches, and some notable patterns of model behavior are discerned. Systematic comparisons carried out by Wilson et al. (2024) between machine learning and GLMs for loss cost estimation indicate that integrated models that incorporate GLMs and artificial neural networks are optimal among traditional approaches. Extensive systematic comparisons between GLMs, GAMs, and tree-based approaches on extensive motor insurance databases are scarce. Much published material involves single geographic markets or narrow methodological features and omits guidance for optimal choices of models for practical usage. Geographic localization of extant literature, particularly for European and North American nations, limits transfer to non-standard motor insurance markets that are characterized by distinctive regulation environments and risk behavior.

The papers demonstrate clear trends towards more advanced methodologies deviating from classical GLMs, but comprehensive empirical comparisons that collectively consider predictive ability, interpretability, and usability of implementation are scarce. While there exists rich literature on single methodologies, there are required systematic guidelines for insurance practice on best model selection, trading off accuracy and regulation, and commercial demands. As complexity for contemporary motor insurance data continues to grow, new methodologies that achieve required interpretability for actuarial application and for handling better nonlinear relationships poorly addressed by classical methodologies become urgent. This work tries to provide a comprehensive empirical comparison between Poisson GLMs, Decision Trees, and GAMs for motor insurance claim frequency modeling and illustrate trade-offs between model interpretability and predictive capacity. The analysis bridges gaps for these methodologies through systematic comparison on comprehensive motor insurance data and comparison on several aspects, including predictive measures, explanatory capacity, and implementation ability for insurance price and risk application.

## 2. METHODOLOGY

The data used in this study belong to European Insurance Company and consist of 108,699 motor third-party liability insurance contracts, representing the French Motor TPL dataset (freMTPL2freq) from the CASdatasets R package, widely used in actuarial research (Dutang & Charpentier, 2020; Počuča et al., 2020). During cleaning, it was overall desired to have a data set that was inconsistency-free and ready for solid modeling. To preclude discrepancies when fitting models and to have analyses unaffected by incomplete information, all observations with missing information on key variables were removed initially. The data were also scanned for outliers or extreme values that would distort model estimates. For instance, Bonus-Malus, capturing a driver's claims history, was truncated at a value of 175 to diminish the influence of very high, very unusual, and perhaps untrustworthy values. The same scrutiny was also applied to variables like the driver's age and vehicle's age to have values within reasonable and expected ranges. Categorical variables were also reviewed to ensure consistency with all categories labeled and organized correctly to preserve reliability when modeling. These cleaning steps provided a solid and predictable source for additional feature engineering and model building.

This comparative analysis employed three distinct modeling approaches applied to a preprocessed motor insurance dataset to evaluate claim frequency prediction performance. The dataset was randomly partitioned using stratified sampling into training (70%), validation (15%), and test

**Table 1.** Summary of dataset variables and their measurement scales

| Column Name | Abbreviation | Description | Variable Type |
|---|---|---|---|
| Policy ID | ID | Unique identifier for the policy | Nominal (Identifier) |
| N Claims | N Claims | Number of claims | Discrete Ratio |
| EXP | EXP | Exposure time in years | Continuous Ratio (used as offset) |
| Vehicle Power | V_P | Engine power | Discrete Ratio |
| Vehicle Age | V_A | Age of the vehicle | Discrete Ratio |
| Driver Age | D_A | Age of the driver | Continuous Ratio |
| Bonus Malus | B_M | Claim history score (e.g., premium adjustment) | Ordinal |
| Vehicle Brand | V_B | Brand of the vehicle | Nominal |
| Fuel Type | F_T | Fuel category (diesel, gasoline) | Nominal |
| Population Density | P_D | People per area unit | Continuous Ratio |
| Region Code | R_C | Broad region grouping | Nominal |

(15%) sets to maintain consistent claim frequency distributions across all partitions.

The first approach utilized Decision Trees based on the Classification and Regression Trees (CART) framework, which recursively partitions the feature space into homogeneous regions. The regression tree function is expressed as

$$\hat{f}(x) = \sum_{m=1}^{M} c_m I\{x \in R_m\}, \quad (1)$$

where f($x$) is the predicted value for input $x$, $M$ is the number of terminal regions, $c_m$ is the predicted response for region $m$, and $I\{\}$ is the indicator function. At each node $t$ with observation set $A$ of size $n_t$, impurity was quantified by

$$i(t) = \frac{1}{n_t} \sum_{i \in A} (y_i - \bar{y}_t)^2, \quad (2)$$

where $y_i$ is the observed response and $\bar{yt}$ is the mean response in node $t$. For candidate splits dividing $A$ into left and right child nodes, the reduction in impurity was calculated as

$$\Delta i = i(t) - \frac{|A_l|}{n_t} i(L) - \frac{|A_r|}{n_t} i(R). \quad (3)$$

Tree growth continued until minimum node size requirements were met, followed by cost-complexity pruning using the penalized objective

$$C_\alpha(T) = \sum_m \sum_{i:x_i \in R_m} (y_i - \hat{y}_{rm})^2 + \alpha |T|, \quad (4)$$

where α controls the tradeoff between fit and simplicity.

The second approach implemented Poisson Generalized Linear Models, assuming claim counts $Y_i$ follow a Poisson distribution with probability mass function

$$P(Y_i = y_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, \quad (5)$$

where $\lambda_i$ represents the expected claim frequency. The model employed a log-link function connecting the linear predictor to expected claim frequency:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}, \quad (6)$$

ensuring

$$\lambda_i = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}\right) > 0. \quad (7)$$

Parameter estimation was conducted through maximum likelihood estimation with the log-likelihood function

$$l(\beta) = \sum_i \left[ y_i \log(\lambda_i) - \lambda_i - \log(y_i!) \right]. \quad (8)$$

Model selection employed forward selection based on Akaike Information Criterion, and the equidispersion assumption $\mathrm{Var}(Y_i) = E[Y_i] = \lambda_i$ was assessed through deviance-based diagnostics.

The third approach implemented Generalized Additive Models (GAMs), extending the GLM framework by replacing linear terms with smooth functions while maintaining additive structure. The model specification becomes

$$g(E[Y_i]) = \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + \cdots + f_p(x_{pi}), \quad (9)$$


http://dx.doi.org/10.21511/ins.16(2).2025.04

where g() is the logarithmic link function and $f_j(x_j)$ represents smooth univariate functions. For insurance applications, this ensures

$$\mu_i = \exp\left(\beta_0 + \sum_j f_j\left(x_{ij}\right)\right), \tag{10}$$

yielding a multiplicative rating structure where $\exp\{f_j(x_{ij})\}$ represents the relativistic effect of factor $j$. Smooth functions were estimated by maximizing the log-likelihood

$$l\left(\beta_0, f_1, \ldots, f_p\right) = \sum_i \log p\left(y_i \mid \mu_i\left(x_i\right)\right), \tag{11}$$

with each $f_j$ constrained to have mean zero over the data for identifiability.

Model evaluation employed a comprehensive framework comparing predictive accuracy and interpretability across the three approaches. Primary evaluation metrics included deviance

$$D = 2\sum_i \left[ y_i \log\left(y_i / \hat{\mu}_i\right) - \left(y_i - \hat{\mu}_i\right)\right], \tag{12}$$

for model fit assessment, Akaike Information Criterion ($AIC = -2l + 2k$) for model selection, and mean squared error MSE for prediction accuracy.

$$MSE = \frac{1}{n}\sum_i \left(y_i - \hat{\mu}_i\right)^2, \tag{13}$$

Statistical significance testing employed likelihood ratio tests for nested models with chi-square distributions determining significance levels. Model diagnostics included residual analysis through standardized Pearson residuals

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{V\left(\hat{\mu}_i\right)}}, \tag{14}$$

where $V(\hat{\mu}_i)$ is the variance function evaluated at the fitted mean. For Decision Trees, variable importance was calculated based on the reduction in node impurity achieved by each predictor across all splits, while for GAMs, effective degrees of freedom for each smooth term indicated the complexity of nonlinear relationships.

All computational implementations utilized *R* statistical software with appropriate packages for each modeling approach, ensuring consistent numerical optimization and cross-validation procedures. The methodology ensured fair comparison by applying identical data preprocessing, consistent evaluation metrics, and standardized validation procedures across all three modeling approaches, enabling robust statistical conclusions about relative performance for motor insurance claim frequency modeling.

## 3. RESULTS

The data contain exposure (EXP) and several rating factors for motor insurance contracts. Summary statistics for the variables are given in Table 2. The exposure, Bonus-Malus scores, and population density show high variability.

**Table 2.** Summary of descriptive statistics for key variables

| Statistic | EXP | V_P | V_A | D_A | B_M | P_D |
|---|---|---|---|---|---|---|
| Min. | 0.003 | 4 | 0 | 18 | 55 | 100 |
| 1st Qu. | 0.11 | 5 | 2 | 28 | 63 | 170 |
| Median | 0.35 | 6 | 5 | 33 | 72 | 680 |
| Mean | 0.418 | 6.07 | 6.372 | 36.87 | 75.47 | 1,441 |
| 3rd Qu. | 0.67 | 7 | 10 | 43 | 85 | 2,498 |
| Max. | 1 | 10 | 40 | 90 | 175 | 6,944 |

Next, the paper examines potential interactions among key variables. In particular, Figure 1 explores claim frequency as a function of vehicle age for different Bonus-Malus levels. The figure illustrates that claim frequency rises overall with vehicle age, particularly for Bonus-Malus level 125, while levels 75 and 100 follow a gentle upward trend then level off. Higher Bonus-Malus levels always equate to more frequent claims. There is more variability for very old and very new vehicles.

Figure 2 illustrates the association between Bonus-Malus level and claim frequency for three groups (75, 100, and 125). Main observations: Increased Bonus-Malus levels (125) have consistently higher claim frequencies across all ages. Young drivers (below 30 years old) have higher claim frequencies, especially at Bonus-Malus level 125. The claim frequency declines up to middle age (around 40-50 years) and gradually increases again among older drivers across all Bonus-Malus groups. Confidence intervals are broader at extreme ages, reflecting greater uncertainty due to fewer observations.

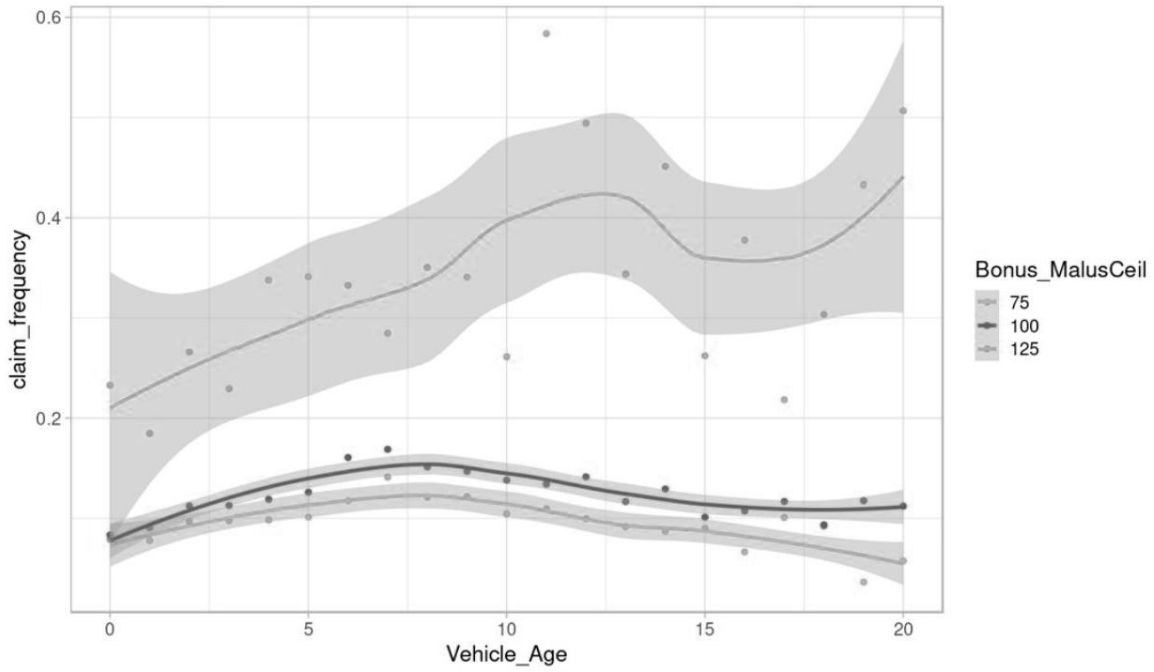


**Figure 1.** Claim frequency by vehicle age and Bonus-Malus level

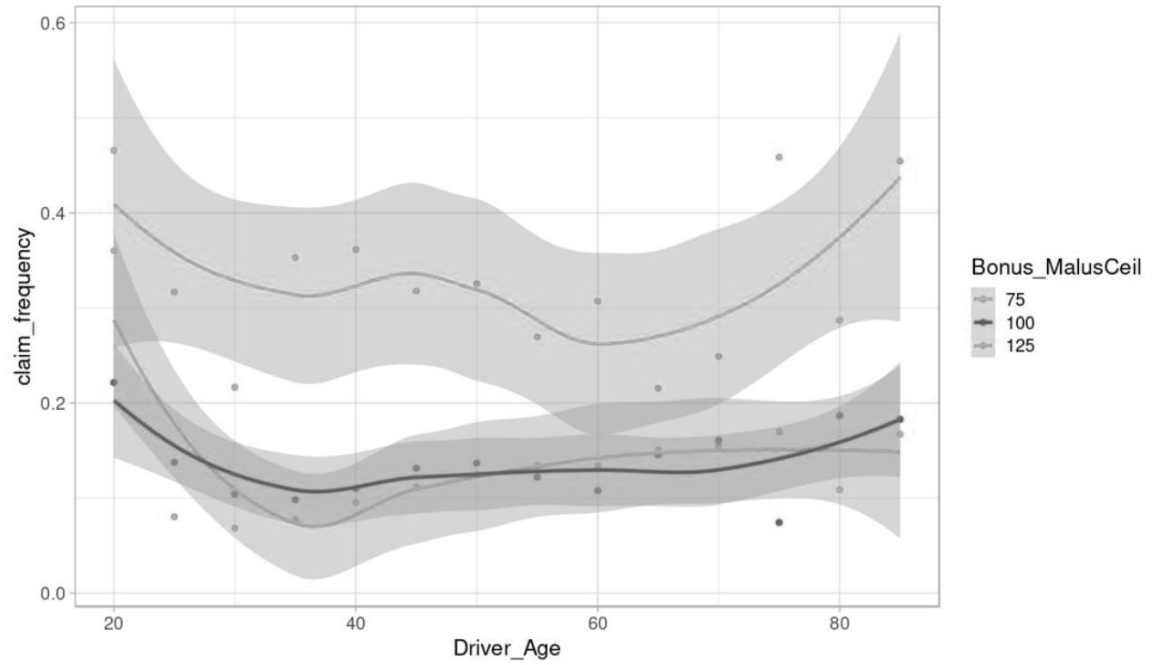


**Figure 2.** Claim frequency by driver age and Bonus-Malus level

A baseline intercept-only model (INT) was fitted to provide a benchmark for claim frequency with no predictors. The actual and predicted frequencies derived from the model's fit to the test data were 11.76% and 11.42%, respectively, indicating that the intercept-only model provides a close estimate of the population average but cannot discriminate risk between policyholders.

Weighted deviance was found to be 38.56% in the training data and 39.71% in the test data, indicating the degree of explained and unexplained variance from the intercept-only model. These results act as a baseline to compare with more sophisticated models involving policy-specific covariates. More complex models that explain in excess of about 60% of the deviance


http://dx.doi.org/10.21511/ins.16(2).2025.04

over and above the intercept-only case show distinct improvement.

A regression decision tree model was trained to predict claim frequency. The resulting tree structure is summarized in Table 3. The first and most influential split is on the Bonus Malus score at 95.5. This is intuitive: a lower Bonus-Malus score typically indicates a safer driver profile, while a higher score often reflects past claims or risky behavior.

Policyholders with Bonus-Malus < 95.5 show a significantly lower claim frequency (mean: 0.1026) compared to those with Bonus-Malus ≥ 95.5 (mean: 0.2412). Among lower-risk policyholders (Bonus-Malus < 95.5), those driving vehicles from Brand_1 or Brand_6 have the lowest average claim frequency (0.0728). Other brands (Brand_2-5, 7-8.11) are associated with a higher frequency (0.1161), indicating that vehicle type or usage might influence claim risk.

Within that higher-claim subgroup, driver age further differentiates risk: younger drivers (< 42.5 years) have a lower frequency (0.0990) compared to older drivers (≥ 42.5 years), who show a higher frequency (0.1599). In the higher-risk group (Bonus-Malus ≥ 95.5), vehicle brand again differentiates claim risk: Brands 1 and 8: moderate risk (mean frequency = 0.1350). Brands 2-7 and 11: high risk (mean frequency = 0.3095). Within this high-risk subset, region also plays a role: regions A1-A6,

**Table 3.** Decision Tree splits for claim frequency, with node sample sizes, deviances, and mean claim frequencies

| Node | Split Condition | n | Deviance | Mean Claim Frequency | Terminal Node |
|---|---|---|---|---|---|
| 1 | Root node | 108,699 | 30,356.24 | 0.1149 | No |
| 2 | B_M < 95.5 | 97,205 | 25,428.42 | 0.1026 | No |
| 4 | V_B ∈ {Brand_1, Brand_6} | 34,674 | 6,647.43 | 0.0728 | Yes |
| 5 | V_B ∈ {Brand_2-5, Brand_7-8, Brand_11} | 62,531 | 18,607.51 | 0.1161 | No |
| 10 | D_A < 42.5 | 46,919 | 12,450.53 | 0.099 | Yes |
| 11 | D_A ≥ 42.5 | 15,612 | 5,984.21 | 0.1599 | Yes |
| 3 | B_M ≥ 95.5 | 11,494 | 4,446.90 | 0.2412 | No |
| 6 | V_B ∈ { Brand_1, Brand_8} | 4,664 | 1,202.10 | 0.135 | Yes |
| 7 | V_B ∈ {Brand_2-7, Brand_11} | 6,830 | 3,112.88 | 0.3095 | No |
| 14 | R_C ∈ {A1, A10, A12, A2-A6, A9} | 3,637 | 1,313.84 | 0.2359 | Yes |

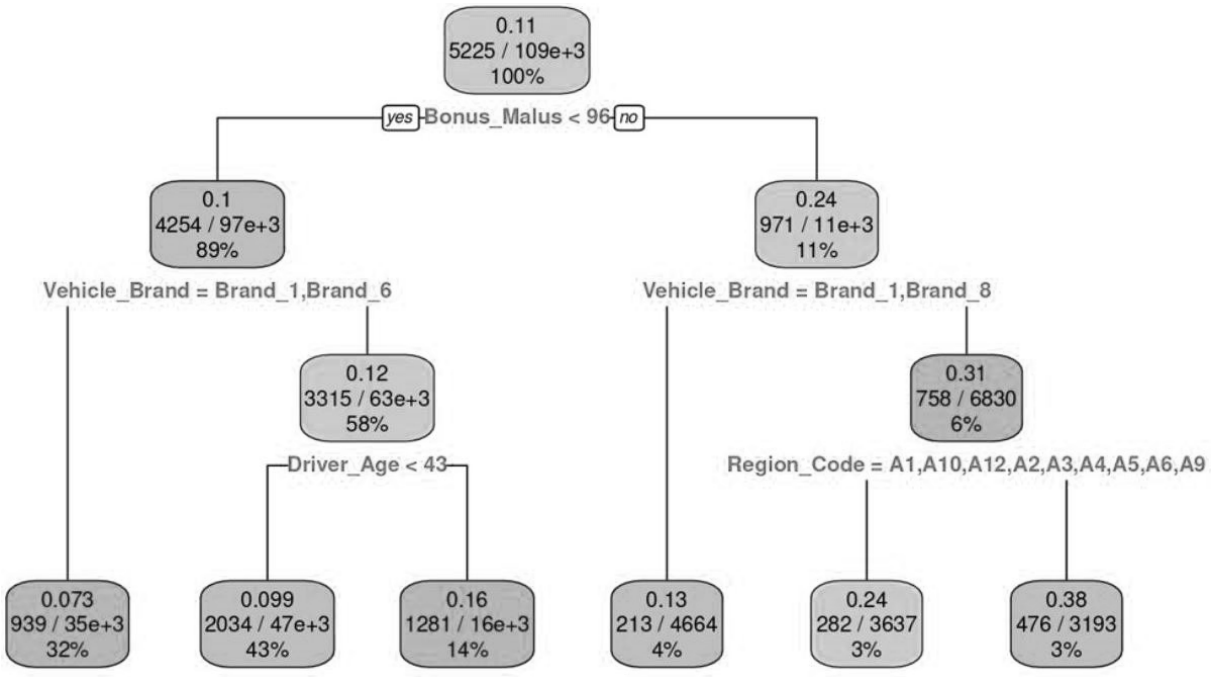


**Figure 3.** Decision Tree for claim frequency

A9, A10, and A12 show slightly lower frequencies (0.2359) than the overall high-risk average.

Figure 3 represents the Decision Tree for claim frequency prediction. At each node, the number at the top is the mean predicted claim frequency, and the fraction at the bottom (in each node's box) indicates the proportion of data in that node. Terminal nodes are colored green (lower claim frequency) or red (higher claim frequency) to indicate relative risk level.

The V_A categorizes vehicle ages into three distinct groups: Group 1 includes vehicles aged 0 years, Group 2 covers vehicles aged between 1 and 10 years, and Group 3 represents vehicles aged between 11 and 110 years, with the reference group as 1-10 years. The driver age variable was converted to a categorical variable by creating seven intervals of age, such as 18-20, 21-25, 26-30, 31-40, 41-50, 51-70, and 71-100 years, with the reference group as 41-50 years. This recording was done to enhance the interpretability of the model and illustrate the interaction between Bonus Malus and driver age. This categorization is intended to capture the nonlinear effects of vehicle age on claim frequency.

**Table 4.** Poisson GLM regression results: estimated coefficients for claim frequency

| Variable | Estimate | Std. error | z value | p-value |
|---|---|---|---|---|
| (Intercept) | −4.2042 | 0.2069 | −20.320 | < 2e-16*** |
| V_P5 | 0.0696 | 0.0517 | 1.345 | 0.1786 |
| V_P6 | 0.1119 | 0.0524 | 2.134 | 0.0328* |
| V_P7 | 0.0761 | 0.0527 | 1.443 | 0.1489 |
| V_P8 | 0.2349 | 0.0710 | 3.310 | 0.0009*** |
| V_P9 | 0.2802 | 0.0631 | 4.443 | 8.85e-06*** |
| V_A1 | −0.1763 | 0.0750 | −2.350 | 0.0188* |
| V_A3 | −0.2500 | 0.0387 | −6.456 | < 1.07e-10*** |
| D_A1 | 0.1469 | 0.0928 | 1.583 | 0.1133 |
| D_A2 | −0.1720 | 0.0589 | −2.921 | 0.0035** |
| D_A3 | −0.4636 | 0.0535 | −8.672 | < 2e-16*** |
| D_A4 | −0.2913 | 0.0479 | −6.078 | < 1.21e-09*** |
| D_A6 | 0.0585 | 0.0532 | 1.100 | 0.2712 |
| D_A7 | 0.1619 | 0.0968 | 1.672 | 0.0945 |
| B_M | 0.0176 | 0.0010 | 18.176 | < 2e-16*** |
| V_B2 | 0.5584 | 0.0553 | 10.107 | < 2e-16*** |
| V_B3 | 0.4857 | 0.0474 | 10.247 | < 2e-16*** |
| V_B4 | 0.5037 | 0.0645 | 7.813 | 5.58e-15*** |
| V_B5 | 0.6079 | 0.0962 | 6.321 | 2.59e-10*** |
| V_B6 | 0.1539 | 0.2082 | 0.739 | 0.4598 |
| V_B7 | 0.5069 | 0.0995 | 5.097 | 3.44e-07*** |
| V_B8 | 0.4532 | 0.0689 | 6.573 | 4.93e-11*** |
| F_T Gasoline | −0.1162 | 0.0330 | −3.524 | 0.0004*** |
| P_D | 0.0631 | 0.0109 | 5.764 | 8.21e-09*** |
| R_C10 | 0.3646 | 0.1955 | 1.865 | 0.0622 |
| R_C11 | 0.2931 | 0.1790 | 1.637 | 0.1016 |
| R_C12 | −0.2210 | 0.2021 | −1.094 | 0.2740 |
| R_C13 | 0.0258 | 0.2715 | 0.095 | 0.9242 |
| R_C14 | −0.0581 | 0.2044 | −0.284 | 0.7762 |
| R_C15 | 0.1371 | 0.1834 | 0.748 | 0.4546 |
| R_C2 | 0.2087 | 0.1771 | 1.178 | 0.2386 |
| R_C3 | 0.1060 | 0.1795 | 0.591 | 0.5546 |
| R_C4 | −0.0395 | 0.1842 | −0.215 | 0.8301 |
| R_C5 | 0.1543 | 0.1830 | 0.843 | 0.3991 |
| R_C6 | −0.0563 | 0.2430 | −0.232 | 0.8169 |
| R_C7 | 0.3368 | 0.1764 | 1.909 | 0.0562 |
| R_C8 | 0.3251 | 0.1844 | 1.763 | 0.0779 |
| R_C9 | 0.0990 | 0.2007 | 0.493 | 0.6217 |

*Note*: Std. error = standard error of the coefficient; z value = Wald statistic; p-value = significance level. ***p < 0.001, **p < 0.01, *p < 0.05.

The Poisson GLM regression output shows several important insights regarding the determinants of motor insurance claim frequencies. The intercept is significant and serves to identify the base log-claim frequency. Among the covariates, a consistent trend emerges: higher power levels, most notably V_P6, V_P8, and V_P9, are linked to significantly increased claim frequencies. This is consistent with the expectation that more powerful vehicles, capable of greater speeds, are more likely to engage in risk-taking driving behavior.

Vehicle age also demonstrates a significant effect. Specifically, V_A3 shows a negative correlation with claim frequency, suggesting that vehicles within this age category are less frequently involved in claims. Effects from driver age are evident as well. Middle-aged drivers (particularly D_A3 and D_A4) are much less likely to file claims, supporting the well-established industry observation that very young and very old drivers pose a greater risk compared to middle-aged drivers. Finally, the B_M variable, which reflects a policyholder's past claims record, is extremely significant and positively correlated with claim incidence. This result reinforces its critical role within experience-based pricing frameworks, where higher malus levels are appropriately associated with greater future risk.

The paper next fits a Generalized Additive Model (GAM) to allow nonlinear effects and interactions. Table 5 summarizes the estimated parametric coefficients, while Table 6 presents the significance of smooth (nonlinear) terms.


http://dx.doi.org/10.21511/ins.16(2).2025.04

**Table 5.** Parametric coefficient estimates from the GAM for claim frequency

| Variable | Estimate | Std. error | t value | p-value |
|---|---|---|---|---|
| (Intercept) | –3.2047 | 0.2406 | –13.317 | 0.001*** |
| V_P5 | 0.0577 | 0.0650 | 0.888 | 0.375 |
| V_P6 | 0.0924 | 0.0661 | 1.398 | 0.162 |
| V_P7 | 0.0608 | 0.0663 | 0.917 | 0.359 |
| V_P8 | 0.2315 | 0.0892 | 2.596 | 0.009** |
| V_P9 | 0.2667 | 0.0794 | 3.357 | 0.001*** |
| F_T Gasoline | –0.1143 | 0.0415 | –2.751 | 0.006** |
| V_B2 | 0.5377 | 0.0719 | 7.476 | 0.001*** |
| V_B3 | 0.4537 | 0.0634 | 7.152 | 0.001*** |
| V_B4 | 0.4766 | 0.0843 | 5.653 | 0.001*** |
| V_B5 | 0.5934 | 0.1224 | 4.847 | 0.001*** |
| V_B6 | 0.1352 | 0.2623 | 0.515 | 0.606 |
| V_B7 | 0.4659 | 0.1271 | 3.667 | 0.001*** |
| V_B8 | 0.4166 | 0.0894 | 4.661 | 0.001*** |
| P_D | 0.0601 | 0.0137 | 4.375 | 0.001*** |
| R_C10 | 0.38069 | 0.24520 | 1.553 | 0.1205 |
| R_C11 | 0.29981 | 0.22464 | 1.335 | 0.1820 |
| R_C12 | –0.20494 | 0.25349 | –0.808 | 0.4188 |
| R_C13 | 0.03439 | 0.34081 | 0.101 | 0.9196 |
| R_C14 | –0.03993 | 0.25639 | –0.156 | 0.8763 |
| R_C15 | 0.16310 | 0.23008 | 0.709 | 0.4784 |
| R_C2 | 0.22496 | 0.22225 | 1.012 | 0.3114 |
| R_C3 | 0.13256 | 0.22515 | 0.589 | 0.5560 |
| R_C4 | –0.01114 | 0.23114 | –0.048 | 0.9616 |
| R_C5 | 0.16872 | 0.22962 | 0.735 | 0.4625 |
| R_C6 | –0.03596 | 0.30479 | –0.118 | 0.9061 |
| R_C7 | 0.34284 | 0.22132 | 1.549 | 0.1214 |
| R_C8 | 0.35747 | 0.23139 | 1.545 | 0.1224 |
| R_C9 | 0.11210 | 0.25186 | 0.445 | 0.6563 |

*Note:* Std. error = standard error of the coefficient; t value = test statistic; p-value = significance level. ***p < 0.001, **p < 0.01, *p < 0.05.

**Table 6.** Approximate significance of smooth terms in the GAM

| Smooth term | EDF | Ref. DF | F | p-value |
|---|---|---|---|---|
| s(V_A) | 4.136 | 5.013 | 8.669 | 0.001*** |
| s(D_A) | 2.959 | 3.714 | 25.855 | 0.001*** |
| s(B_M):D_A1 | 3.094 | 3.752 | 23.909 | 0.001*** |
| s(B_M): D_A2 | 3.005 | 3.749 | 43.620 | 0.001*** |
| s(B_M): D_A3 | 1.002 | 1.005 | 65.397 | 0.001*** |
| s(B_M): D_A4 | 3.783 | 4.683 | 20.579 | 0.001*** |
| s(B_M): D_A5 | 1.906 | 2.413 | 5.962 | 0.001** |
| s(B_M): D_A6 | 4.334 | 5.290 | 3.781 | 0.002** |
| s(B_M): D_A7 | 1.565 | 1.914 | 3.593 | 0.020* |

*Note:* EDF = estimated degrees of freedom; Ref. DF = reference degrees of freedom; F = test statistic; p-value = significance level. ***p < 0.001, **p < 0.01, *p < 0.05.

The main findings from the Generalized Additive Model (GAM) analysis provide various insights regarding factors influencing motor insurance claim frequency. The parametric coefficients (Table 5) indicate that several vehicle-related factors are significantly associated with claim frequency. Increased vehicle power ratings, such as V_P8 and V_P9, are positively correlated with higher claim frequencies. Additionally, gasoline-fueled vehicles exhibit a negative coefficient, suggesting reduced claim frequency compared to the base category, possibly due to differences in vehicle usage patterns or inherent vehicle characteristics.

Several vehicle brands, particularly V_B5 and V_B2, have significantly positive coefficients, indicating brand-related risk factors. These may

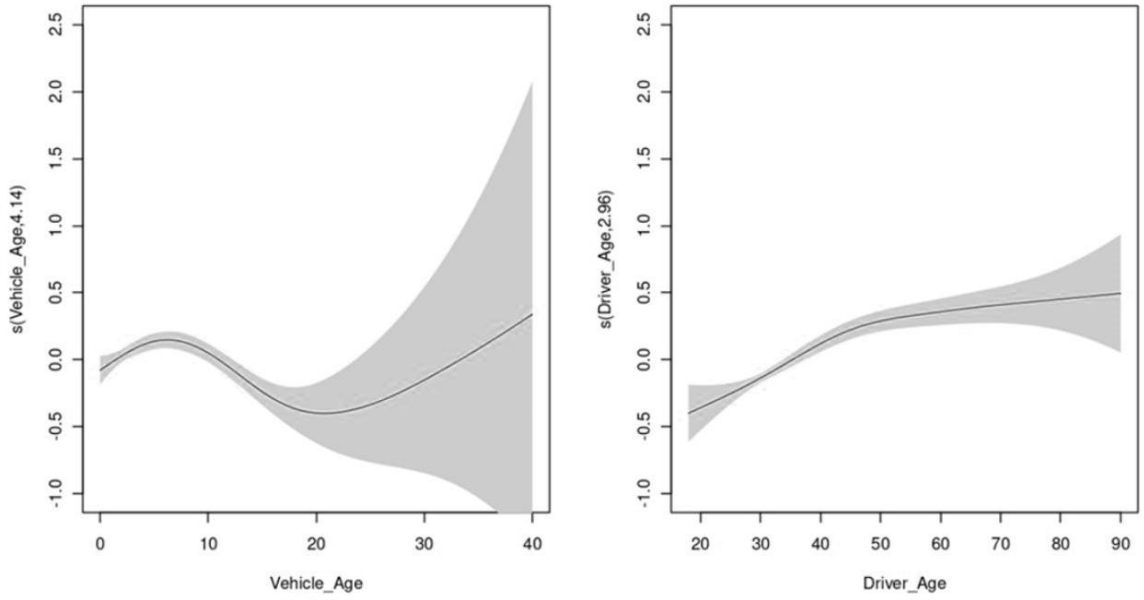


**Figure 4.** GAM smooth functions for vehicle age and driver age

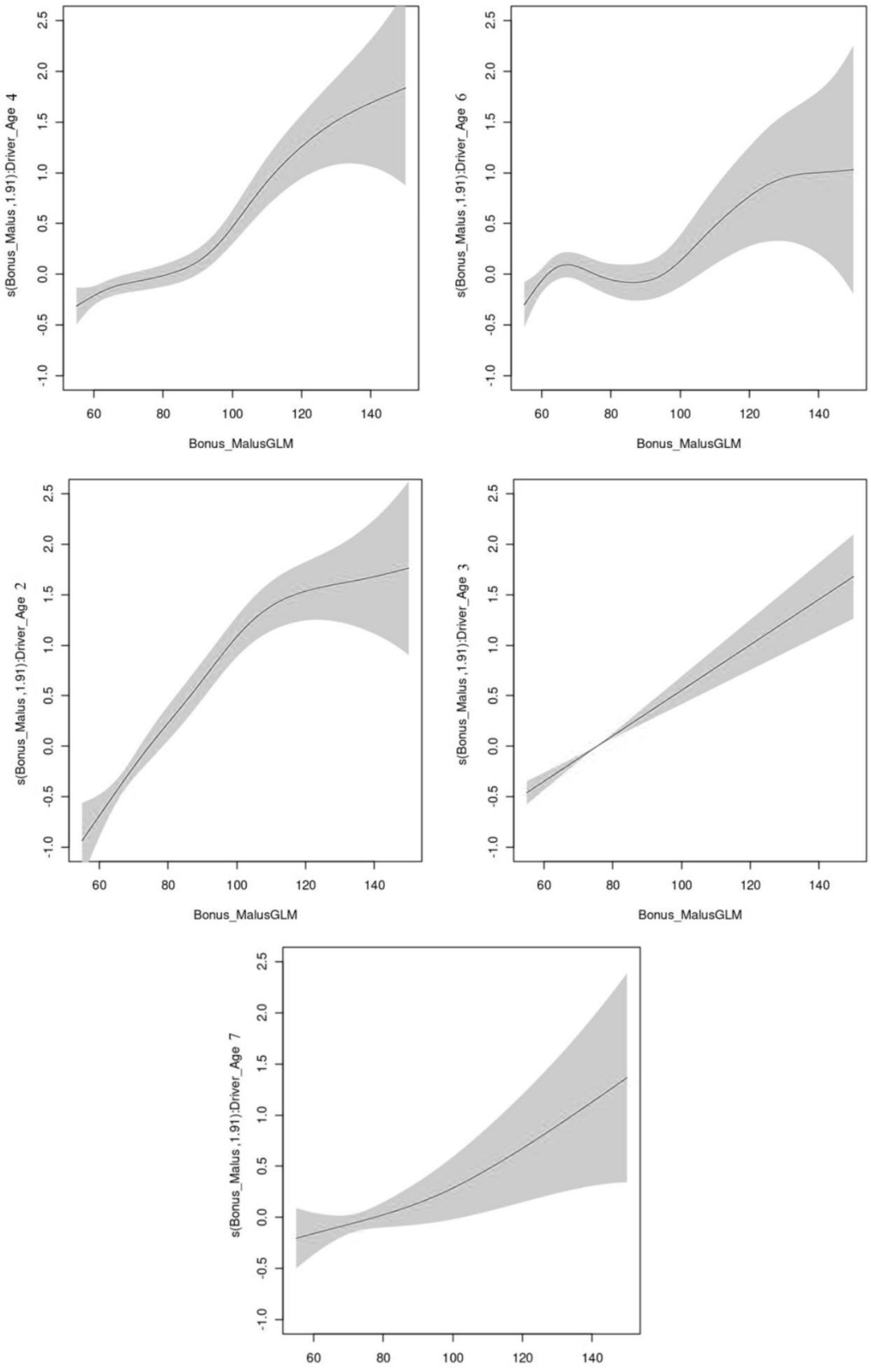


**Figure 5.** GAM smooth functions for Bonus-Malus by driver age group



http://dx.doi.org/10.21511/ins.16(2).2025.04

be influenced by vehicle design, safety features, or demographic characteristics associated with those brands. These findings reinforce the importance of incorporating detailed vehicle attributes within predictive models for claim frequency.

The smooth terms in Table 6 offer greater insight regarding nonlinear interactions and effects within the data. The vehicle age and the driver age both have significant smooth terms, indicating complex, nonlinear relationships between these variables and claim frequency. Figure 4 illustrates the smooth effects of V_A and D_A from the GAM. The chart shows the estimated smooth function (solid blue line) and 95% confidence band (gray shading). As illustrated in Figure 4, claim frequency first falls off with car age to approximately 20 years, and then it steeply rises for older cars. For driver age, claim frequency consistently rises to older ages, with a noticeable plateauing beyond about 50 years. Confidence intervals are widest at the extremes, reflecting greater uncertainty.

Figure 5 illustrates how claim frequency rises with Bonus Malus across different driver age groups. The rate of change differs: young and mid-life drivers exhibit a steeper increase in claim frequency compared to older drivers. The effect is nonlinear, and confidence intervals are broader at higher Bonus-Malus scores.

The nonlinear relationships identified through GAM analysis reveal important pricing opportunities previously invisible to linear models. The steep increase in claim frequency with Bonus-Malus scores for younger drivers suggests that traditional linear rating factors may undercharge high-risk young drivers while potentially overcharging moderate-risk profiles. This finding indicates opportunities for more nuanced risk segmentation that could improve both competitiveness and profitability. The nonlinear vehicle age effects show that very old vehicles (beyond 20 years) present significantly higher risk, suggesting current age-based rating factors may be inadequate for vintage vehicle pricing. These insights enable insurers to develop more sophisticated rating structures that better reflect true risk patterns.

The GAM's parametric coefficients are consistent with the GLM's findings. Higher vehicle power, certain brands (2, 3, 4, 5, 7, 8), and higher Bonus-Malus all increase claim frequency, while gasoline vehicles reduce it. Importantly, the smooth terms in Table 6 confirm significant nonlinear relationships. Both vehicle age and driver age have highly significant smooth functions, indicating non-monotonic risk patterns. Moreover, the interactions between Bonus Malus and driver age are significant across almost all age groups (D_A1 through D_A7), meaning the impact of the Bonus-Malus score on claim frequency varies by age. In other words, a high malus (poor history) is especially detrimental for some age groups but slightly less so for others.

Finally, the paper compares the overall performance of the Decision Tree, GLM, and GAM models with a baseline intercept-only model INT in Table 7.

**Table 7.** Model performance comparison for claim frequency prediction

| Metric | INT | DT | GLM | GAM |
|---|---|---|---|---|
| MSE | – | 0.0582 | 0.0509 | 0.0506 |
| RMSE | – | 0.2413 | 0.2257 | 0.2251 |
| Poisson Deviance (Training) | 38.56% | 37.12% | 36.83% | 36.41% |
| Poisson Deviance (Test) | 39.71% | 38.31% | 38.08% | 37.76% |

Table 7 shows a comparison of performances for claim frequency prediction. The best performances across all measures were achieved by the GAM model, which had the lowest MSE (0.0506), RMSE (0.2251), and Poisson Deviance (36.41% Learn / 37.76% Test). GLM followed and was only marginally worse than GAM. Decision Tree was more error-prone and provided a higher MSE and RMSE than GLM and GAM models. The Intercept-only model (INT), which was taken as the base case, also provided significantly higher Poisson Deviance measures, reflecting that including predictors greatly improves the explanatory power of the model.

Overall, GAM provided the best and most generalizable predictions among the compared models. In terms of interpretability and practical use, the GLM is the most straightforward to interpret,

making it preferable when model transparency is important. The GAM balances interpretability and flexibility by introducing smooth terms that can be visualized as in Figures 3, 4, and 5 to explain effects like age and Bonus-Malus interactions. Decision Trees are intuitive in structure but can become complex and unstable; in this case, the tree's poor performance suggests it may need pruning or an ensemble approach (e.g., random forest, gradient boosting) to be viable. Overall, the GAM provided a slight predictive boost and valuable insights into nonlinear patterns, while the GLM remains a well-calibrated baseline model for claim frequency.

The superior performance of GAMs over traditional GLMs has significant economic implications for insurers. The 0.6% reduction in prediction error translates to improved risk classification accuracy, potentially reducing adverse selection and enhancing portfolio profitability. More precise claim frequency predictions enable insurers to set more competitive premiums for low-risk segments while adequately pricing high-risk exposures. The ability of GAMs to capture nonlinear relationships means that traditional linear pricing models may systematically misprice certain risk profiles, leading to market share loss in profitable segments and retention of unprofitable business. For a typical insurer with €100 million in premium volume, even a 1% improvement in pricing accuracy could translate to €1 million in additional underwriting profit annually.

## 4. DISCUSSION

Better performance of GAMs over regular GLMs is in agreement with Díaz Martínez et al. (2023), who demonstrated that GAMs outline risk profiles through smooth curves representing patterns unknown to linear models. The results of this study, with GAM's minimum MSE (0.0506) and highly significant nonlinear relationships in vehicle age, driver age, and Bonus-Malus interactions, strongly concur with their findings, ratifying sector shift from GLMs to GAMs.

Decision Tree performance is quite different from that of König and Loser (2024) and Clemente et al. (2023), who saw tree-based methods outperforming traditional methods. The poor individual Decision Tree performance (MSE: 0.0582) reveals that it is necessary to employ ensemble methods such as Random Forests or Gradient Boosting, as in previous work, to realize tree-based potential, just as Staudt and Wagner (2021) found.

The fairly narrow gap in performance for GAMs and GLMs (0.6% reduction in MSE) presents a more refined perspective relative to Wilson et al. (2024), who found significant progress with hybrid models. The importance of smooth terms at the statistical level vindicates theoretical concepts by Hastie and Tibshirani (1986, 1990) and practical knowledge by Henckaerts et al. (2018, 2019) in constructing insurance tariffs.

Variable importance outputs validate prevailing actuarial wisdom while creating new knowledge. The importance of high Bonus-Malus scores supports Denuit and Lang's (2004) experience rating effectiveness, whereas nonlinear age impacts identified reveal Verbelen et al.'s (2018) nonlinear policyholder characteristic results to be more ubiquitous than classic linear assumptions.

While GAMs demonstrate superior predictive performance, insurers must consider implementation costs versus benefits. The marginal improvement over GLMs requires sophisticated statistical expertise and more complex model maintenance. However, the enhanced ability to identify profitable customer segments and reduce mispricing can justify these investments, particularly for insurers operating in competitive markets where pricing accuracy provides a strategic advantage. The interpretability advantage of GAMs over black-box machine learning approaches also reduces regulatory compliance costs compared to more complex alternatives.

These results contribute to insurance analytics literature by providing systematic empirical support for moving from GLMs to GAMs and acknowledging individual tree method constraints. The outcomes suggest that for best regulatory transparency and balance with precision, GAMs are of choice and warrant future exploration for hybrid techniques combining both machine learning flexibility and GAM interpretability.


http://dx.doi.org/10.21511/ins.16(2).2025.04

## CONCLUSIONS AND RECOMMENDATIONS

The objectives of this study were to systematically evaluate both model interpretability and predictive capacity within three various modeling methodologies for predicting motor insurance claim frequencies and to bridge a notable gap in comparative actuarial modeling literature. Empirical research demonstrates that Generalized Additive Models achieve superior prediction capacity with desirable interpretative properties for regulatory compliance and data-driven decisions conducive to businesses. The identification of strong nonlinear patterns for vehicle age, driver age, and Bonus-Malus interactions reveals important risk patterns that cannot be explained by conventional linear techniques, and so seriously call into question the continuation of GLMs as the de facto insurance pricing standard.

These findings also have important actuarial practice implications. Firstly, the consistently modest but regular outperformance of GAMs relative to GLMs, combined with their enhanced ability to reveal nonlinear risk relationships, argues powerfully for method development for insurance pricing models. Secondly, poor individual Decision Tree performance argues strongly for ensemble techniques when considering tree-based methods for insurance usage. Thirdly, smooth term statistical significance in GAMs gives strong support to the theoretical assumption that insurance risk relationships are nonlinear and require flexible modeling approaches to adequately capture their sophistication. What they reveal is that insurance organizations should prefer the implementation of GAMs to traditional GLM solutions when they are confronted with nonlinear patterns of risk in advanced policyholder portfolios where such patterns are likely to impact pricing accuracy to a great extent. The interpretability-by-design advantage of GAMs over black-box machine solutions renders them preferred solutions to regulatory environments mandating model interpretability.

Future work needs to be guided along three crucial paths. Firstly, devising hybrid modeling architectures that combine both GAM flexibility and machine learning ensemble techniques to obtain the best predictive ability and interpretability. Secondly, investigating time-varying structures for GAM so that evolving patterns of risk are learned through time-varying coefficients and online real-time updating schemes for parameters. Thirdly, investigating high-dimensional behavioral data, particularly telematics and IoT sensor data, within GAM structures to leverage new data types while maintaining interpretability of the model. These paths will enrich the insurance modeling theory foundations while practical problems for the emerging insurance marketplace are addressed.

## AUTHOR CONTRIBUTIONS

Conceptualization: Eslam Abdelhakim Seyam.
Data curation: Eslam Abdelhakim Seyam.
Formal analysis: Eslam Abdelhakim Seyam.
Funding acquisition: Eslam Abdelhakim Seyam.
Investigation: Eslam Abdelhakim Seyam.
Methodology: Eslam Abdelhakim Seyam.
Project administration: Eslam Abdelhakim Seyam.
Resources: Eslam Abdelhakim Seyam.
Software: Eslam Abdelhakim Seyam.
Supervision: Eslam Abdelhakim Seyam.
Validation: Eslam Abdelhakim Seyam.
Visualization: Eslam Abdelhakim Seyam.
Writing – original draft: Eslam Abdelhakim Seyam.
Writing – review & editing: Eslam Abdelhakim Seyam.

## REFERENCES


1. Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E., & Thandi, N. (2004). *A practitioner's guide to generalized linear models* (Casualty Actuarial Society Discussion Paper Program). Retrieved from https://www.casact.org/sites/default/files/database/dpp_dpp04_04dpp1.pdf
2. Antonio, K., & Valdez, E. A. (2012). Statistical concepts of a priori and a posteriori risk classification in insurance. *AStA Advances in Statistical Analysis, 96*(2), 187-224. https://doi.org/10.1007/s10182-011-0152-7
3. Ayuso, M., Guillén, M., & Pérez-Marín, A. M. (2016). Telematics and gender discrimination: Some usage-based evidence on whether men's risk of accidents differs from women's. *Risks, 4*(2), 10. https://doi.org/10.3390/risks4020010
4. Ayuso, M., Guillén, M., & Pérez-Marín, A. M. (2019). Using GPS data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. *Transportation Research Part C: Emerging Technologies, 68,* 160-167. https://doi.org/10.1016/j.trc.2016.04.004
5. Baecke, P., & Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems, 98,* 69-79. https://doi.org/10.1016/j.dss.2017.04.009
6. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324
7. Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees.* Wadsworth International Group. https://doi.org/10.1201/9781315139470
8. Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge University Press. Retrieved from https://scispace.com/pdf/regression-analysis-of-count-data-52t6su8ltf.pdf
9. Clemente, C., Guerreiro, G. R., & Bravo, J. M. (2023). Modelling motor insurance claim frequency and severity using gradient boosting. *Risks, 11*(9), 163. https://doi.org/10.3390/risks11090163
10. Denuit, M., & Lang, S. (2004). Non-life rate-making with Bayesian GAMs. *Insurance: Mathematics and Economics, 35*(3), 627-647. https://doi.org/10.1016/j.insmatheco.2004.08.001
11. Díaz Martínez, Z., Fernández Menéndez, J., & García Villalba, L. J. (2023). Tariff analysis in automobile insurance: Is it time to switch from generalized linear models to generalized additive models? *Mathematics, 11*(18), 3906. https://doi.org/10.3390/math11183906
12. Dionne, G., & Vanasse, C. (1989). A generalization of actuarial automobile insurance rating models: The negative binomial distribution with a regression component. *ASTIN Bulletin, 19*(2), 199-212. https://doi.org/10.2143/AST.19.2.2014909
13. Dutang, C., & Charpentier, A. (2020). *Package 'CASdatasets'*. R package. Retrieved from https://cas.uqam.ca/pub/web/CASdatasets-manual.pdf
14. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451
15. Goldburd, M., Khare, A., & Tevet, D. (2016). *Generalized linear models for insurance rating* (2nd ed.). Casualty Actuarial Society. Retrieved from https://www.casact.org/sites/default/files/database/monographs_papers_05-goldburd-khare-tevet.pdf
16. Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science, 1*(3), 297-318. https://doi.org/10.1214/ss/1177013604
17. Hastie, T., & Tibshirani, R. (1990). *Generalized additive models.* Chapman & Hall.
18. Henckaerts, R., & Antonio, K. (2022). The added value of dynamically updating motor insurance prices with telematics collected driving behavior data. *Insurance: Mathematics and Economics, 105,* 79-95. https://doi.org/10.1016/j.insmatheco.2022.03.011
19. Henckaerts, R., Antonio, K., Clijsters, M., & Verbelen, R. (2018). A data-driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal, 2018*(8), 681-705. https://doi.org/10.1080/03461238.2018.1429300
20. Henckaerts, R., Côté, M. P., Antonio, K., & Verbelen, R. (2019). Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal, 25*(2), 255-285. https://doi.org/10.1080/10920277.2020.1745656
21. Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). Cambridge University Press.
22. Kafková, S., & Křivánková, L. (2014). Generalized linear models in vehicle insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 62*(2), 383-388. https://doi.org/10.11118/actaun201462020383
23. Klein, N., Kneib, T., Klasen, S., & Lang, S. (2014). Bayesian structured additive distributional regression for multivariate responses. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 64*(4), 569-591. https://doi.org/10.1111/rssc.12090
24. König, D., & Loser, F. (2024). Claim frequency modeling in insurance pricing using GLM, deep learning, and gradient boosting. *Blätter der DGVFM, 36*(1), 45-62. Retrieved from https://aktuar.de/en/knowledge/specialist-information/detail/claim-frequency-modeling-in-insurance-pricing-using-glm-deep-learning-and-gradient-boosting/
25. Kuo, K., & Lupton, D. (2021). *Towards explainability of machine learning models in insurance pricing* (Papers 2003.10674). https://doi.org/10.48550/arXiv.2003.10674




http://dx.doi.org/10.21511/ins.16(2).2025.04

26. McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall. Retrieved from https://utstat.utoronto.ca/brunner/oldclass/2201s11/readings/glmbook.pdf

27. Meng, S., Gao, Y., & Huang, Y. (2022). Actuarial intelligence in auto insurance: Claim frequency modeling with driving behavior features and improved boosted trees. *Insurance: Mathematics and Economics, 106,* 115-127. https://doi.org/10.1016/j.insmatheco.2022.06.001

28. Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A, 135*(3), 370-384. Retrieved from https://jhanley.biostat.mcgill.ca/bios601/Likelihood/NelderWedderburn1972.pdf

29. Ohlsson, E., & Johansson, B. (2010). *Non-life insurance pricing with generalized linear models*. Springer. https://doi.org/10.1007/978-3-642-10791-7

30. Paefgen, J., Staake, T., & Fleisch, E. (2013). Multivariate exposure modeling of accident risk: Insights from pay-as-you-drive insurance data. *Transportation Research Part A: Policy and Practice, 61,* 27-40. https://doi.org/10.1016/j.tra.2013.11.010

31. Počuča, N., Jevtić, P., McNicholas, P. D., & Miljkovic, T. (2020). Modeling frequency and severity of claims with the zero-inflated generalized cluster-weighted models. *Insurance: Mathematics and Economics, 94,* 79-93. https://doi.org/10.1016/j.insmatheco.2020.06.004

32. Staudt, Y., & Wagner, J. (2021). Assessing the performance of random forests for modeling claim severity in collision car insurance. *Risks, 9*(3), 53. https://doi.org/10.3390/risks9030053

33. Verbelen, R., Antonio, K., & Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C, 67*(5), 1275-1304. https://doi.org/10.1111/rssc.12283

34. Wilson, A. A., Nehme, A., Dhyani, A., & Mahbub, K. (2024). A comparison of generalized linear modelling with machine learning approaches for predicting loss cost in motor insurance. *Risks, 12*(4), 62. https://doi.org/10.3390/risks12040062

35. Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Chapman & Hall/CRC.

36. Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B, 73*(1), 336. https://doi.org/10.1111/j.1467-9868.2010.00749.x

37. Xie, S., & Shi, K. (2023). Generalized additive modelling of auto insurance data with territory design: A rate regulation perspective. *Mathematics, 11*(2), 334. https://doi.org/10.3390/math11020334