Risk Analysis and Pricing in Non-Life Insurance

Non-Life Insurance Mathematics

Björn Thor Stefánsson
University of Iceland
bts5@hi.is

Viktor Már Guðmundsson
University of Iceland
vmg3@hi.is

April 25, 2025

Abstract

This report explores pricing and risk assessment in non-life insurance through two parts. The first models pure premium using French motor insurance data, combining generalized linear models and a spliced severity distribution to capture both typical and extreme claims. The second uses Danish fire insurance data to evaluate the financial impact of large claims under different reinsurance strategies. A hybrid empirical-lognormal model is used to simulate annual losses via a compound Poisson process. Risk measures such as VaR and CTE help assess profitability and downside protection across reinsurance contracts.

Keywords Non-Life Insurance · Premium Modeling · Reinsurance · Generalized Linear Models · Tail Risk

1 Introduction

Pricing and risk management are core challenges in non-life insurance. Insurers must set premiums that reflect underlying risks while protecting against large, infrequent losses. This requires models that can capture both everyday claim behavior and extreme events.

This report focuses on two complementary parts. The first estimates pure premiums using French automobile insurance data, combining frequency and severity models, including a spliced Gamma-GPD approach. The second analyzes Danish fire insurance claims using simulations to compare the impact of excess-of-loss reinsurance contracts on overall revenue and tail risk. Together, these analyses highlight practical methods for data-driven decision-making in insurance pricing and risk mitigation.

2 Pure Premium Estimation and Risk-Based Pricing

The goal of the first part of the project is to estimate the so called pure premium and thereon to propose a way to price the insurance i.e. to calculate the premium a customer has to pay.

$$p_{Net}(t) = \mathbb{E}[S(t)],$$

We can formulate the pure premium as following

$$\mathbb{E}[S(t)] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{N(t)} X_i \middle| N(t)\right]\right] = \mathbb{E}[N(t)\mathbb{E}[X]] = \mathbb{E}[N(t)] \cdot \mathbb{E}[X]$$

This decomposition allows us to model claim frequency $\mathbb{E}[N(t)]$ and severity $\mathbb{E}[X]$ either separately or jointly. In this project, we model them separately, but an alternative is to use a single model—such as the Tweedie distribution—that captures both components simultaneously. Both approaches make use of the generalized linear model (GLM) framework, which provides flexibility in distributional assumptions and facilitates statistical inference through its well-defined likelihood structure.

2.1 Background

In this section, we review the structure of GLMs and explain why the Tweedie distribution is particularly useful for modeling pure premiums in insurance. GLMs provide a flexible way to model non-negative, skewed, and overdispersed data—common features of insurance claims. The Tweedie family is especially relevant because it can simultaneously capture both claim frequency and severity through a compound Poisson-Gamma formulation, while remaining within the GLM framework.

2.1.1 Generalized Linear Models (GLM)

A generalized linear model relates relates a linear predictor

$$\eta = X\beta$$

to the dependent variable Y though a link function $g(\mu)$ such that

$$g\left(\mathbb{E}[Y]\right) = X\beta$$

Further the GLM framework has a mean-variance framework linking the the variance to the mean. For example an exponential dispersion model has the variance function of

$$\mathrm{Var}(Y) = \phi \cdot \mathrm{Var}(\mu)$$

This allows for different distributional assumptions other than the normal distribution. The normal distribution is a special case of a GLM model where the base measure is the Lebesgue measure, then

$$\mathrm{Var}(Y) = \phi \cdot \mathrm{Var}(\mu) = \phi = \sigma^2$$

Generalized linear models must fulfill the form of an exponential distribution model:

$$P_{\theta,\phi}(Y \in A) = \int_A \exp\left(\frac{y\theta - \kappa(\theta)}{\phi} + c(y,\phi)\right) d\nu(y)$$

Where θ are the model parameters, $\kappa(\theta)$ the cumulant function, ϕ the dispersion parameter which acts as a variance parameter and $\nu(y)$ a base measure.

How can we construct or identify a distribution that fits within the GLM framework? If a distribution admits exponential tilting—meaning its moment generating function is finite in a neighborhood of zero—then the tilted version belongs to the exponential dispersion family and can serve as the basis for a generalized linear model.

Exponential Tilting: Given a random variable X with a probability distribution P_0, density f and a moment generating function $M_X(\theta) = \mathbb{E}[e^{\theta X}] < \infty$, the exponentially tilted measure P_θ is defined as

$$P_\theta(X \in dx) = \frac{\mathbb{E}[e^{\theta X}\mathbb{I}_{X\in dx}]}{M_X(\theta)} = e^{\theta x - \kappa(\theta)} P_0(X \in dx),$$

where $K(\theta)$ is the cumulative generating function (CGF), which ensures that the tilted measure integrates to one, making it a valid probability distribution:

$$K(\theta) = \log \mathbb{E}[e^{\theta X}] = \log M_X(\theta).$$

Then $P_\theta(X \in dx) = f_\theta(x) \propto e^{\theta x} f_0(x)$ is the $\theta-$titled density.

2.1.2 The Tweedie Model

Any model within the GLM framework can be uniquely identified by its variance function $\mathrm{Var}(Y)$. For example:

Power p	**Variance Function**	**Distribution**	**Family**
0	$\mathrm{Var}(Y) = \phi$	$Y \sim \mathcal{N}(\mu,\phi)$	Gaussian
1	$\mathrm{Var}(Y) = \phi\mu$	$Y \sim \mathcal{P}(\mu)$	Poisson
2	$\mathrm{Var}(Y) = \phi\mu^2$	$Y \sim \mathcal{G}(\mu,\phi)$	Gamma
3	$\mathrm{Var}(Y) = \phi\mu^3$	$Y \sim \mathcal{IG}(\mu,\phi)$	Inverse-Gaussian

Table 1: Distributions in the exponential dispersion family by variance function and power p.

Hence, the Tweedie model $\mathrm{Tw_p}(\mu,\phi)$ is defined by its variance function $\mathrm{Var}(Y)=\phi\mu^p$ where p is a pre-defined parameter. For $1<p<2$, this corresponds to a compound Poisson-Gamma mixture:

$$S(t)=\sum_{i=1}^{N(t)} X_i \quad \text{where} \quad X\sim\mathcal{G}(\alpha,\beta) \quad \text{and} \quad N(t)\sim\mathrm{Poisson}(\lambda t).$$

We can show that the Tweedie distribution with $1<p<2$ is equal in distribution to the compound Poisson-Gamma process $S(t)$ (and can therefore be used to model it) by demonstrating that both have the same cumulant generating function (CGF) when the parameters λ, α, and β are chosen appropriately.

The moment generating function for $S(t)$ is

$$M_{S(t)}(\theta)=\mathbb{E}[e^{\theta S(t)}]=\mathbb{E}\left[\mathbb{E}\left[e^{\theta\sum_{i=1}^{N(t)}X_i}\middle|N(t)\right]\right]=\mathbb{E}\left[M_X(\theta)^{N(t)}\right]=e^{\lambda t(M_X(\theta)-1)}=\exp\left(\lambda t\left[\left(1-\frac{\theta}{\beta}\right)^{-\alpha}-1\right]\right)$$

Therefore, the cumulant generating function of $S(t)$ is given by:

$$K_{S(t)}(\theta)=\log M_{S(t)}(\theta)=\lambda t\left[\left(1-\frac{\theta}{\beta}\right)^{-\alpha}-1\right]$$

Next, consider a random variable $Y\sim\mathrm{Tw}_p(\mu,\phi)$, which belongs to the exponential dispersion family $\mathrm{ED}(\mu,\phi)$. The CGF of such a variable is

$$K_Y(\theta)=\frac{\kappa(\theta_0+\theta\phi)-\kappa(\theta_0)}{\phi} \quad \text{where } \mu=\kappa'(\theta_0).$$

To derive the form of $\kappa(\theta)$, we use the variance function $V(\mu)=\mu^p$, which implies

$$\frac{d}{d\theta}\kappa'(\theta)=[\kappa'(\theta)]^p.$$

This differential equation can be solved as follows:

$$\frac{[\kappa'(\theta)]^{1-p}}{1-p}=\theta+C_1.$$

Assuming $C_1=0$ and integrating gives:

$$\kappa(\theta)=\frac{1}{2-p}[(1-p)\theta]^{\frac{2-p}{1-p}} \quad \text{for } p\neq 1,2$$

We now show that the Tweedie CGF matches the CGF of the compound Poisson-Gamma distribution when the following parameter relationships hold:

$$\lambda=\frac{\mu^{2-p}}{(2-p)\phi}, \quad \alpha=\frac{2-p}{p-1}, \quad \beta=\frac{\phi(p-1)}{\mu^{p-1}}$$

Substituting these into the CGF of $S(t)$ and simplifying yields:

$$K_{S(t)}(\theta)=\frac{1}{\phi}\left[\frac{1}{2-p}\left[(1-p)(\theta_0+\theta\phi)\right]^{\frac{2-p}{1-p}}-\frac{1}{2-p}[(1-p)\theta_0]^{\frac{2-p}{1-p}}\right].$$

This expression can be recognized as:

$$K_{S(t)}(\theta)=\frac{1}{\phi}[\kappa(\theta_0+\theta\phi)-\kappa(\theta_0)].$$

This confirms that the CGFs of $S(t)$ and Y are identical, and therefore $S(t)\sim\mathrm{Tw}_p(\mu,\phi)$.

The equivalence holds regardless of which link function is used in a GLM context. Note that $\alpha=\frac{2-p}{p-1}$ must be positive since it represents the shape parameter of the Gamma distribution. This requires $\frac{2-p}{p-1}>0$, which is satisfied when $1<p<2$. Additionally, $\lambda>0$ requires $(2-p)>0$, again satisfied when $p<2$. Therefore, the compound Poisson-Gamma representation of the Tweedie distribution is valid specifically when $1<p<2$.

2.2 Modeling Approaches

In this section, we discuss the various statistical models employed to estimate claim frequency and severity, which together determine the pure premium.

2.2.1 Poisson Model for Claim Frequency

The claim frequency $N(t)$ is typically modeled using a Poisson distribution, where the number of claims in a period t follows

$$N(t) \sim \text{Poisson}(\lambda t)$$

with λ representing the claim rate. Under the GLM framework, we model the expected number of claims as

$$\mathbb{E}[N(t)] = \lambda t = \exp(X\beta)$$

where X is the design matrix of covariates and β the coefficient vector. The Poisson model assumes the variance equals the mean:

$$\text{Var}(N(t)) = \mathbb{E}[N(t)] = \lambda t$$

2.2.2 Quasi-Poisson and Overdispersion

Insurance claim data frequently exhibits overdispersion, where the variance exceeds the mean, violating the Poisson assumption. The quasi-Poisson model addresses this by introducing a dispersion parameter ϕ:

$$\text{Var}(N(t)) = \phi \cdot \mathbb{E}[N(t)] = \phi\lambda t$$

The parameter estimation remains unchanged, but standard errors are adjusted by the dispersion factor ϕ, calculated as

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

where n is the sample size and p is the number of parameters.

2.2.3 Gamma Model for Claim Severity

For claim severity X, we employ a Gamma distribution which naturally models positive, right-skewed data:

$$X \sim \text{Gamma}(\alpha, \beta)$$

with shape parameter α and rate parameter β. The density function is

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

In the GLM framework, we parametrize the expected severity as

$$\mathbb{E}[X] = \frac{\alpha}{\beta} = \exp(Z\gamma)$$

where Z represents severity-related covariates and γ the corresponding coefficients. The variance follows

$$\text{Var}(X) = \frac{\alpha}{\beta^2} = \frac{(\mathbb{E}[X])^2}{\alpha} = \phi(\mathbb{E}[X])^2$$

with dispersion parameter $\phi = 1/\alpha$.

2.2.4 Offset Term

To account for varying exposure periods, we incorporate an offset term in our frequency model. Let e_i represent the exposure duration for policy i, measured in years. The expected number of claims becomes

$$\mathbb{E}[N_i] = \lambda_i e_i = e_i \exp(X_i\beta)$$

which we implement in the GLM framework as

$$\log(\mathbb{E}[N_i]) = \log(e_i) + X_i\beta$$

where $\log(e_i)$ enters as a fixed offset. This adjustment ensures that policies with different exposure periods are weighted appropriately in the model.

2.2.5 Generalized Additive Models (GAMs)

While GLMs assume linear relationships between covariates and the response (on the link scale), Generalized Additive Models extend this framework by allowing non-linear relationships through smooth functions:

$$g(\mathbb{E}[Y]) = \beta_0 + \sum_{j=1}^{p} f_j(X_j)$$

where f_j are smooth functions typically represented as penalized splines. GAMs maintain the distributional assumptions of GLMs but offer greater flexibility in capturing complex relationships, particularly for continuous variables like age or policy duration, without requiring explicit polynomial terms or categorical transformations.

2.2.6 Randomized Residuals

For count data models like Poisson, ordinary residuals often fail to follow a normal distribution, complicating model diagnostics. Randomized residuals address this by introducing uniform random noise:

$$r_i^{\text{rand}} = \Phi^{-1}\left(F(y_i - 1|\hat{\mu}_i) + u_i \cdot [F(y_i|\hat{\mu}_i) - F(y_i - 1|\hat{\mu}_i)]\right)$$

where

- Φ^{-1} is the inverse standard normal CDF
- $F(y|\mu)$ is the Poisson CDF evaluated at y with mean μ
- $u_i \sim \text{Uniform}(0, 1)$ is a random variable
- $\hat{\mu}_i$ is the predicted mean for observation i

These residuals approximately follow a standard normal distribution when the model is correctly specified, enabling standard diagnostic tools like QQ-plots and facilitating detection of overdispersion and other model inadequacies.

2.3 Data Exploration and Preprocessing

This analysis explores automobile insurance claims in France, examining both frequency and severity across various risk factors. The dataset encompasses over 413,000 policies. Insurance data typically exhibits very long tails and is overly dispersed, presenting challenges for generalized linear models which struggle with these extreme distributions. To address this limitation, we utilize the 95th percentile data for fitting our models, as shown in Table 2. This percentile-based split was determined specifically by the severity data's much heavier tail distribution, while the claim number distribution remains unaffected by this methodological choice.

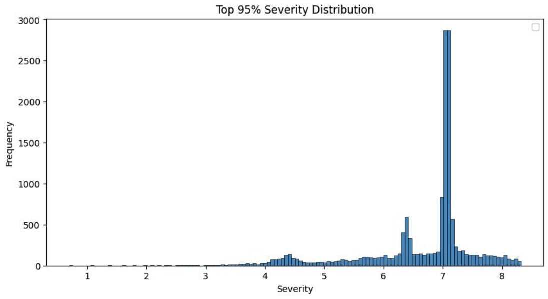

Figure 1: Non-Extreme Severity (Bottom 95%)

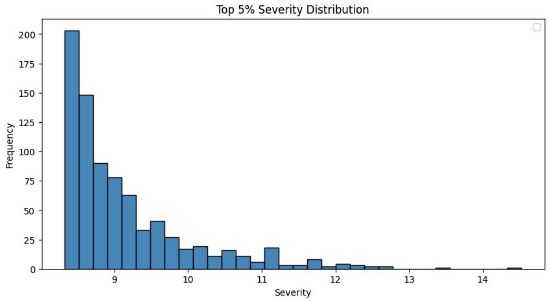

Figure 2: Log-Transformed Extreme Severity (Top 5%)

This figure illustrates the effect of truncating the top 5% of severity values. While the bulk of the data becomes more homogeneously distributed and easier to model, a substantial portion of monetary value remains concentrated in the tail. Therefore, we retain both segments—the 95% subset for modeling purposes and the tail subset for separate analysis. Table 2 summarizes key descriptive statistics for the full dataset, the truncated 95% portion, and the extreme tail.

Statistic	Whole Data	95% Data	Tail Data
Policy Information			
Total policies	413,960	-	-
Policies with no claims	397,779 (96.09%)	-	-
Policies with claims	16,181 (3.91%)	15,371 (3.71%)	810 (0.20%)
Claim Frequency (Claimnb)			
Count	16,181	15,371	810
Mean	1.10	1.10	1.06
Median	1.00	1.00	1.00
Standard deviation	0.33	0.33	0.25
Minimum	1.00	1.00	1.00
Maximum	4.00	4.00	3.00
Skewness	3.37	3.34	3.80
Kurtosis	12.64	12.46	13.63
Claim Severity			
Count	16,181	15,371	810
Mean	1,985.71	1,052.40	19,696.76
Median	1,143.00	1,134.00	6,761.00
Standard deviation	18,679.41	699.41	81,429.29
Minimum	2.00	2.00	4,110.00
Maximum	2,036,833.00	4,106.00	2,036,833.00
Skewness	84.86	1.32	19.84
Kurtosis	8,845.14	3.11	471.70
Cutoff Information			
Severity threshold	4,110.00	-	-
Percentile cutoff	95%	-	-

Table 2: Statistical comparison of insurance claim data: complete dataset vs. below and above 95% percentile cutoff.

2.3.1 Categorical Features

Before analyzing categorical predictors, we must consider the driverage variable. Using it as a numerical variable in our Poisson model implies an exponential relationship between driver age and claim frequency due to the log link function. However, driver age likely has a more complex, non-monotonic relationship with claim frequency. Risk typically decreases sharply from newly licensed drivers to those 25+, remains relatively stable during middle age, then increases again for older individuals. By treating driver age as categorical, we can capture this complex pattern while maintaining interpretability, enabling more accurate and justifiable insurance pricing adjustments based on driver age groups. To determine optimal categories, we employed a tree-based method that maximizes splits based on the target variable. Decision trees identify optimal cut points in continuous variables by creating the most homogeneous groups with respect to claim frequency, allowing for data-driven age categories rather than arbitrary brackets.

Group	Driver Age Range	Potential Mergers (p > 0.05)
1	17.0 - 22.5	None
2	22.5 - 26.5	Group 7 (p = 0.3108)
3	26.5 - 30.5	Group 6 (p = 0.5205), Group 7 (p = 0.1255)
4	30.5 - 55.5	Group 6 (p = 0.0548)
5	55.5 - 61.5	None
6	61.5 - 72.5	Group 3 (p = 0.5205), Group 4 (p = 0.0548)
7	72.5 - 100.0	Group 2 (p = 0.3108), Group 3 (p = 0.1255)

Table 3: Driver age risk groups and statistically similar groups ($p > 0.05$) that could potentially be merged.

After identifying meaningful breakpoints (e.g., ages 25, 45, and 65), we incorporated them as categorical levels in our GLM. This approach captures true risk patterns while preserving the GLM framework's interpretability advantages for insurance pricing. We further validated these groups using a χ^2-test to determine statistical differences. Despite finding some statistically insignificant differences, we chose not to combine any groups since no two insignificant groups were adjacent. See Table 3 and Figure 3.

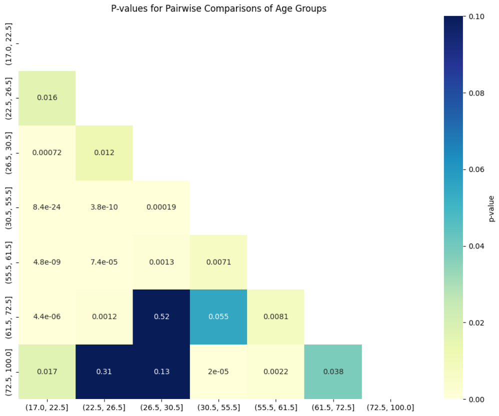

Figure 3: χ^2–test contingency table test for driver age groups.

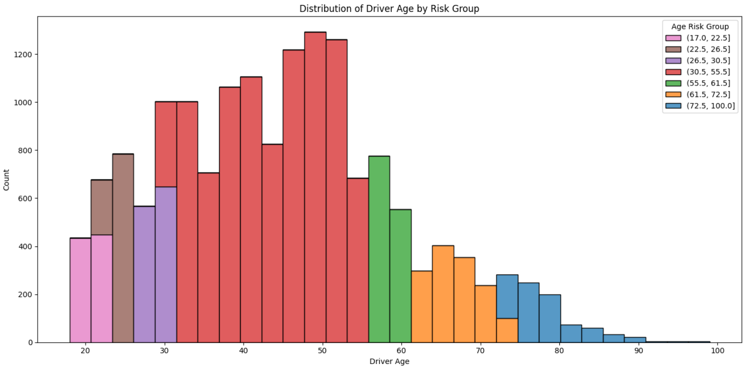

Figure 4: Driver age groups distribution.

Figure 5: Analysis of driver age groups: significance testing and distribution.

It is useful to plot the our categorical variables to see if it is prudent to aggregate any further on the basis of statistical power concerns or similarity in severity. See Figure 6 for refernce.

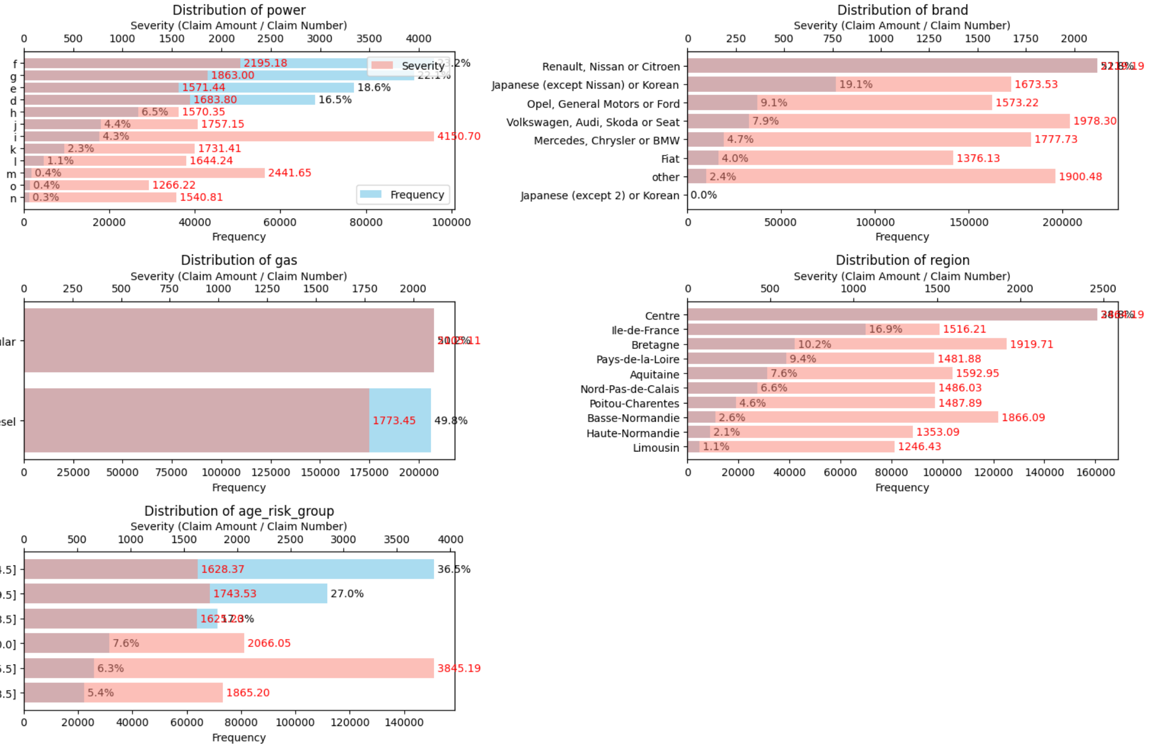

Figure 6: Distributions of categorical variables by severity and number of contracts.

Based on these distributions, we make the following adjustments:

1. We propose merging power bands J through O, which correspond to the most powerful vehicles. These bands are close in range and aggregating them is reasonable. Additionally, we merge bands H and I to create more balanced group sizes across the remaining categories. We avoid combining power bands that differ significantly in magnitude.
2. The brand categories are poorly structured: they are not clearly based on price, region, or buyer profile. Since the current groupings do not offer meaningful segmentation, we combine all brands that individually account for less than 5% of the data into a single "Other" category. This results in a category of roughly 10%, which is more balanced and statistically viable.
3. For regional grouping, we will conduct a separate analysis to determine whether certain regions can be merged based on similar behavior.
4. The fuel type variable (gas vs. diesel) appears to be well-balanced and requires no modification.
5. The age variable has already been grouped appropriately and will remain unchanged.

Our region groupings were primarily based on geographic proximity, though a few were informed by specific characteristics. For example, the Centre region, which includes Paris, was treated separately due to its urban density and higher frequency of claims. The final suggested groupings are illustrated in Figure **??**, which overlays them on a map of France for geographic context. These groupings form the basis for our final set of categorical variables, which will be subject to backward selection during the modeling phase.

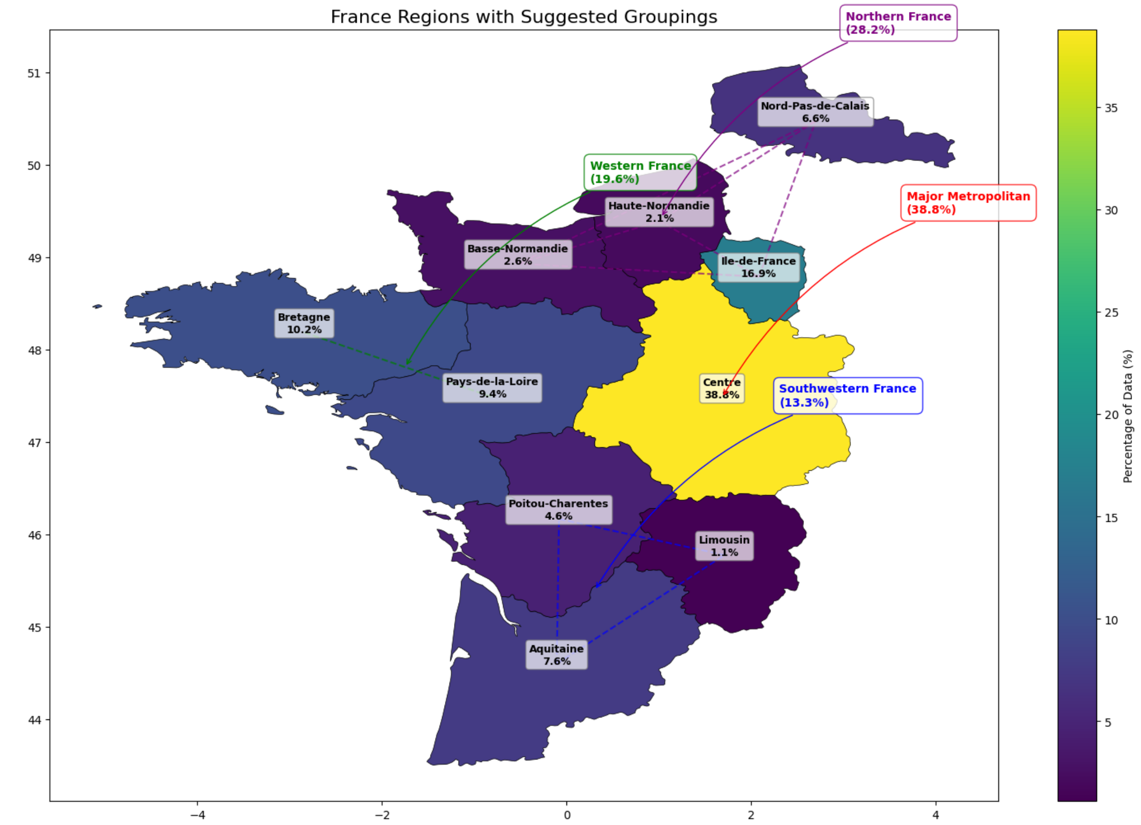

Figure 7: France map with suggested regional groupings based on geographic and claim characteristics.

To further justify our groupings, Figure 7 displays the original regional distribution alongside the proposed groupings. The aggregation helps balance the data across regions and reduces noise in low-volume areas, enabling more robust modeling.

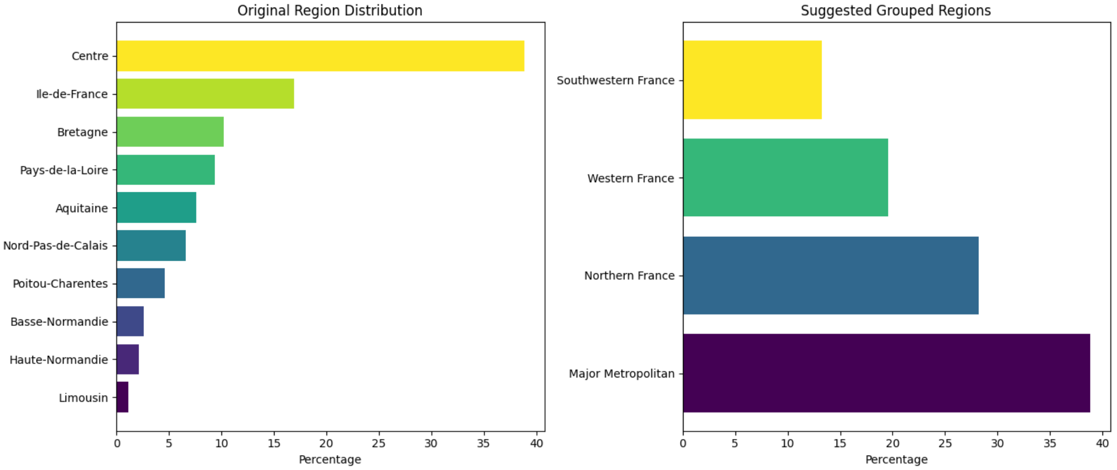

Figure 8: Bar plots comparing the original region distribution (left) with the suggested grouped regions (right).

Finally, we summarize the distributions of all categorical variables after preprocessing. These will be available for likelihood ratio tests during feature selection in the modeling stage. See Figure 9.

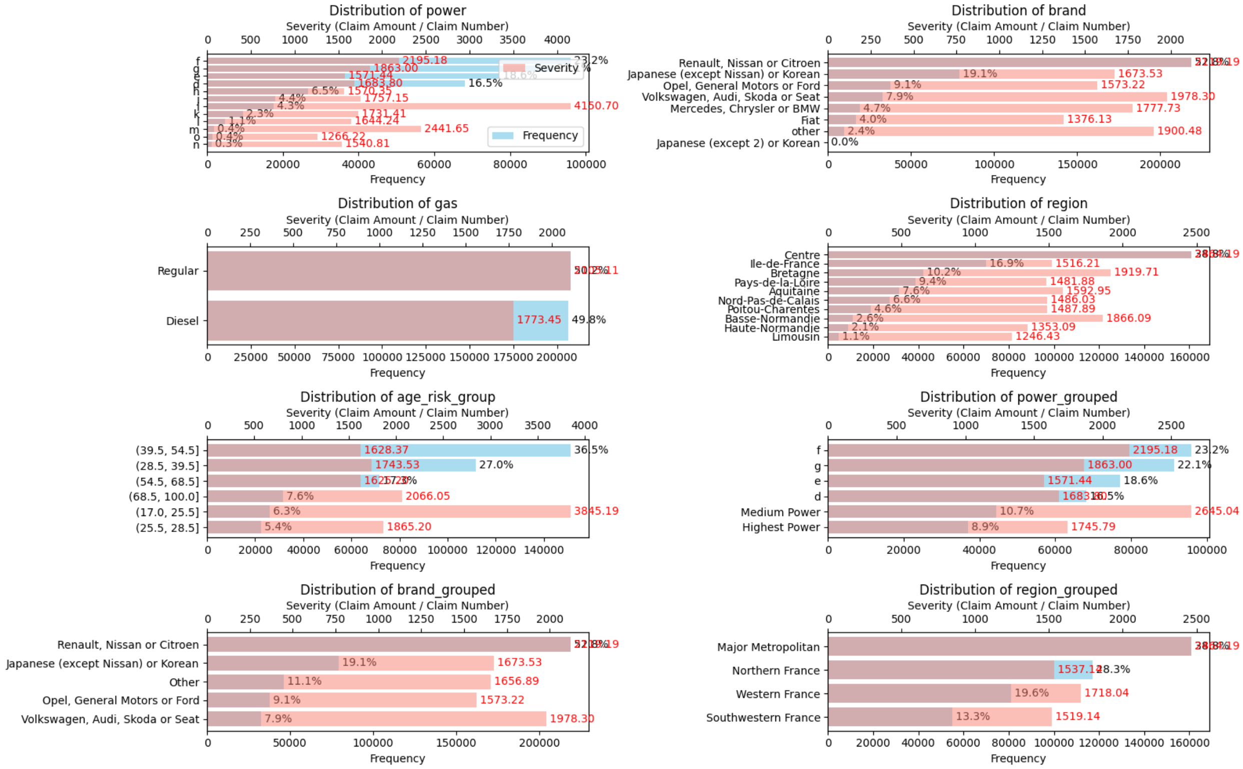

Figure 9: Final categorical variable distributions after preprocessing.

2.3.2 Numerical Features

We begin by examining the pairwise relationships and correlation structure of the numerical variables in the dataset. As shown in Figure 10, none of the variables appear to exhibit strong correlation with claim frequency or severity. This suggests that the numeric features offer limited predictive power on their own, particularly in relation to key outcomes of interest in pricing models.

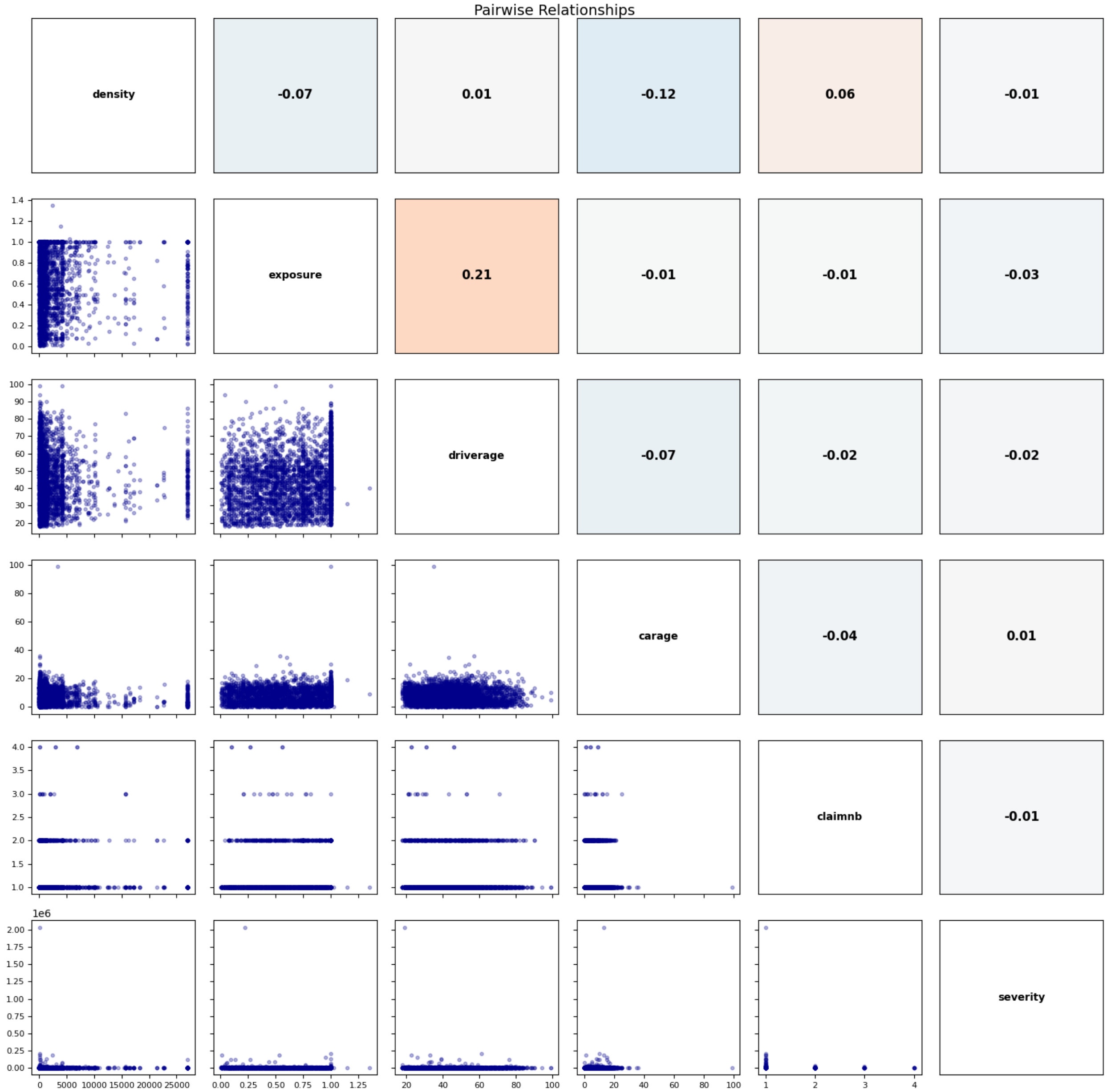

Figure 10: Pairwise plot and correlation of numerical features.

To further understand the role of each variable, we plot the marginal distributions along with claim severity overlays in Figure 11. Most variables are heavily skewed, especially density, which has a long right tail. Interestingly, a large subset of contracts (5085) have density below 1000 and are concentrated in the Centre region, suggesting that high claim counts are not confined to densely populated urban areas. Moreover, the severity of claims appears relatively stable across different density levels.

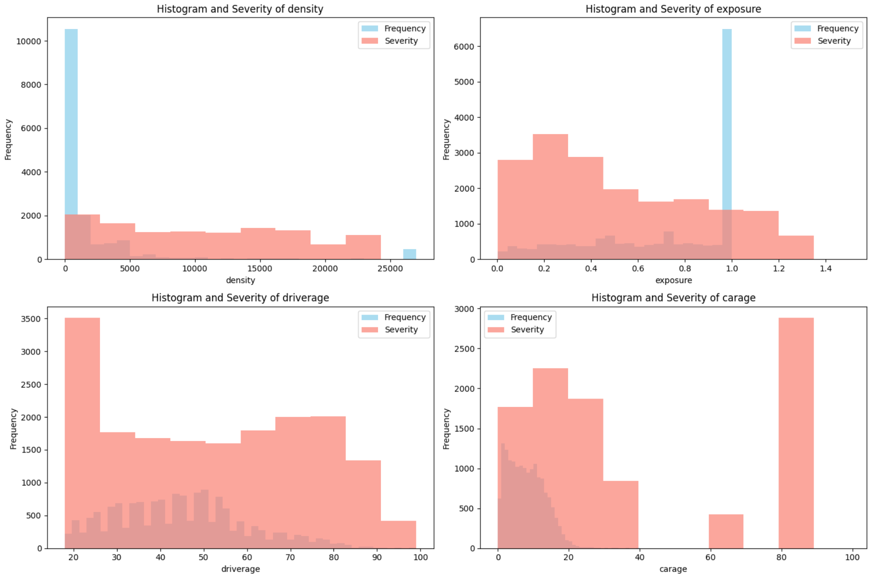

Figure 11: Histograms of numerical features with overlaid claim severity.

2.4 Model Selection and Implementation

First we will model $\mathbb{E}[N(t)]$ and $\mathbb{E}[X]$ separately and then we propose a merger of the two models which can model $\mathbb{E}[S(t)]$ on its own. Generalized linear models allow for greater modeling flexibility while also being interpretable making this class of models a great choice for heavily regulated industries like insurance.

2.4.1 Modelling $\mathbb{E}[X]$

To model the expected claim size, $\mathbb{E}[X]$, we follow a standard three-step process:

1. Choosing an appropriate exponential distribution family for the response.
2. Choosing a link function.
3. Selecting the predictors.

We begin by selecting a distribution based on log-likelihood maximization applied to the untransformed claim sizes. Alternative transformations were tested but did not improve the results meaningfully. As shown in Table 4, the Gamma distribution performs well and ranks second, just behind the Weibull minimum. However, given the Gamma distribution's widespread use in the insurance industry and its desirable properties, we choose it as our working model. This choice is also illustrated in Figure 12.

Rank	Distribution	Log-likelihood
0	Weibull Min	-120,826.38
1	Gamma	-120,973.80
2	Lognormal	-120,991.68
3	Pareto	-122,305.89
4	Inverse Gaussian	-122,437.09

Table 4: Comparison of distribution fits for claim sizes ranked by log-likelihood (higher values indicate better fit).

Although the differences in log-likelihood between the top distributions are relatively small, the Gamma distribution remains a practical and theoretically sound option for modeling claim severity. It naturally accommodates right-skewed, positive data, and its interpretability makes it especially suitable for industry applications.

We use the log-link function for this model, as it complements the Gamma distribution and ensures a positive mean prediction. While alternative link functions, such as the inverse, could be considered, the log-link is both standard and effective in this context.

1. The response variable Y is positive and hence the mean μ is also positive which is a requirement for gamma.
2. Our distribution is right skewed which matches the gamma distribution.

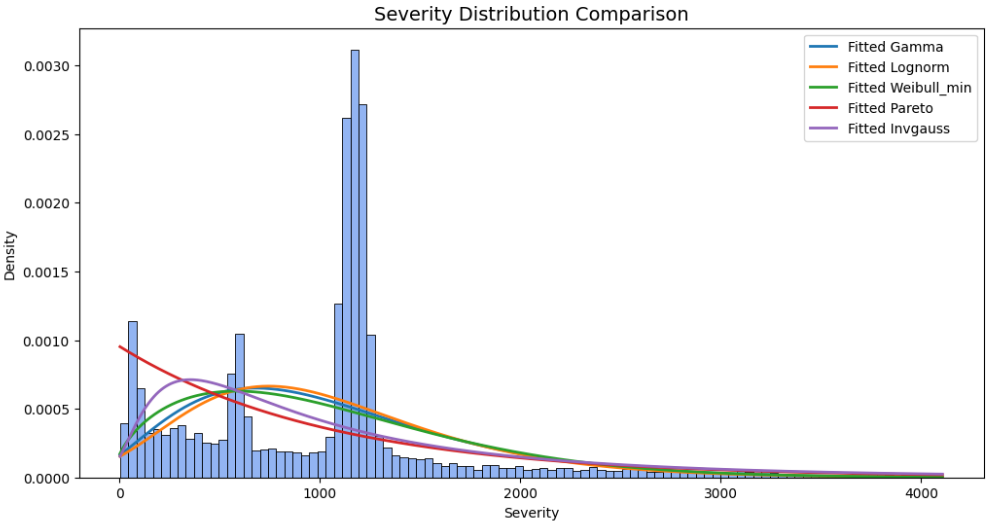

Figure 12: Comparison of Candidate Distribution Family Fits for Claim Severity

After developing grouped features for our categorical variables based on domain knowledge and decision trees, we proceeded with model selection. Beginning with a full model containing all predictors, we performed backward selection using the likelihood ratio test to systematically eliminate non-significant variables. The exposure column was incorporated as an offset term to account for varying policy durations.

Model Information	
Dependent Variable	severity
Model	Gamma
No. Observations	15371
Df Residuals	15355
Df Model	15
Log-Likelihood	-1.25×10^5
Deviance	12473
Scale (Dispersion)	1.5474
Pearson χ^2	2.38×10^4
Pseudo R-sq. (CS)	0.5198

Table 5: Full Generalized Linear Model Specification for Claim Severity

Table 5 summarizes our final Gamma model with log link for claim severity. Notably, our custom feature groupings outperformed the original categorical encodings, validating our intuition behind the groupings. This was particularly evident for driver age groupings, which were constructed using decision trees that minimized Gini impurity at each split.

We proceed by applying backward selection using the likelihood ratio test to remove insignificant predictors. The exposure column is used as an offset to account for varying levels of risk exposure. The final model retains the most relevant predictors. Table 6 presents the resulting coefficient estimates along with their standard errors, z-values, and p-values.

Variable	Estimate	Std. Error	z-value	p-value
Intercept	9.5365	0.053	180.861	0.000
Age Group (25.5, 28.5]	0.1025	0.054	1.906	0.057
Age Group (28.5, 39.5]	-0.0035	0.039	-0.089	0.929
Age Group (39.5, 54.5]	0.0608	0.037	1.655	0.098
Age Group (54.5, 68.5]	0.0660	0.042	1.590	0.112
Age Group (68.5, 100.0]	0.1537	0.049	3.133	0.002
Power Group (Medium)	-0.0067	0.046	-0.145	0.885
Power Group (d)	0.0327	0.043	0.761	0.447
Power Group (e)	0.0273	0.041	0.666	0.505
Power Group (f)	0.0763	0.040	1.921	0.055
Power Group (g)	-0.0047	0.040	-0.115	0.908
Region (Northern France)	-0.0710	0.028	-2.501	0.012
Region (Southwestern France)	-0.0966	0.033	-2.973	0.003
Region (Western France)	-0.0273	0.027	-1.011	0.312
Density	$-5.2e{-6}$	$2.3e{-6}$	-2.304	0.021
Exposure	-2.8162	0.034	-83.909	0.000

Table 6: GLM Coefficient Estimates for Claim Severity

Note: Baseline categories are the lowest age group (under 25.5), the lowest power group, and the reference region. Power groups (d, e, f, g) represent specific vehicle power bands. Standard errors are non-robust.

The final model indicates that some of our custom features improved model performance over the original ungrouped data, particularly the age groups derived from decision tree splits. These effects are visible in the older age brackets, where severity increases and coefficients become statistically significant. Regional effects and exposure are also clearly influential, while some power groups were less conclusive. This reinforces the value of incorporating structured domain-driven grouping into model development.

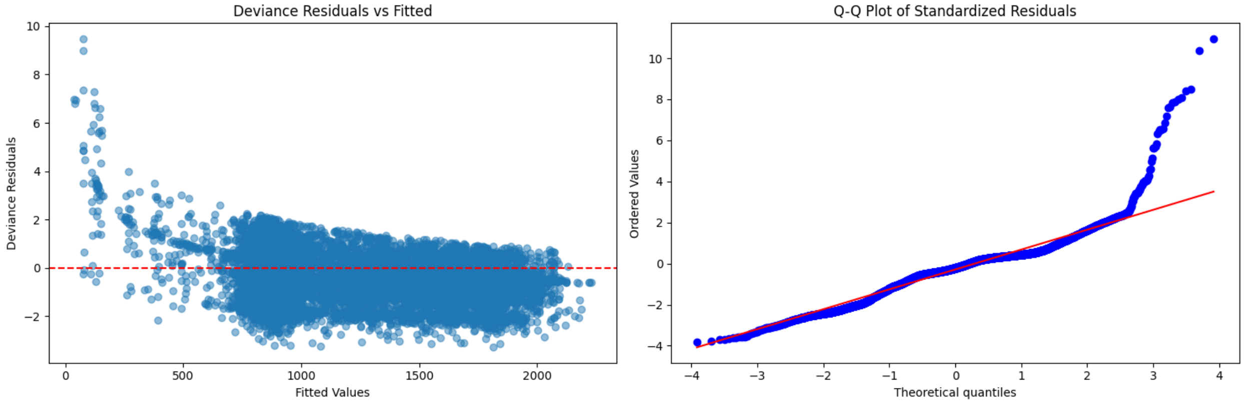

Figure 13: Deviance Residuals vs. Fitted Values for the Gamma GLM Severity Model

2.4.2 A GMM-Enhanced Gradient Boosting Approach for Severity Prediction

This model leverages a two-stage approach to predict severity values in a regression task. First, a Gaussian Mixture Model (GMM) with $K = 5$ components captures the underlying multimodal structure in the data, functioning as a soft clustering mechanism. Formally, the GMM models the data distribution as

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where π_k represents mixture weights, $\boldsymbol{\mu}_k$ cluster centers, and $\boldsymbol{\Sigma}_k$ covariance matrices. The GMM assigns each observation to its most likely component, generating cluster labels that serve as additional features.

In the second stage, a Gradient Boosting Machine (GBM) learns the relationship between features (including cluster assignments) and log-transformed severity. The GBM constructs an ensemble of decision trees sequentially:

$$F_M(\mathbf{x}) = \sum_{m=1}^{M} \gamma_m h_m(\mathbf{x}),$$

where each new tree h_m focuses on the residuals of previous trees. The logarithmic transformation of the target variable helps manage skewness in the severity distribution, while the cluster information allows the GBM to learn specialized prediction patterns for different data subpopulations. This approach effectively handles heterogeneity in the data that a single regression model might miss, improving predictive performance particularly for complex, multimodal datasets.

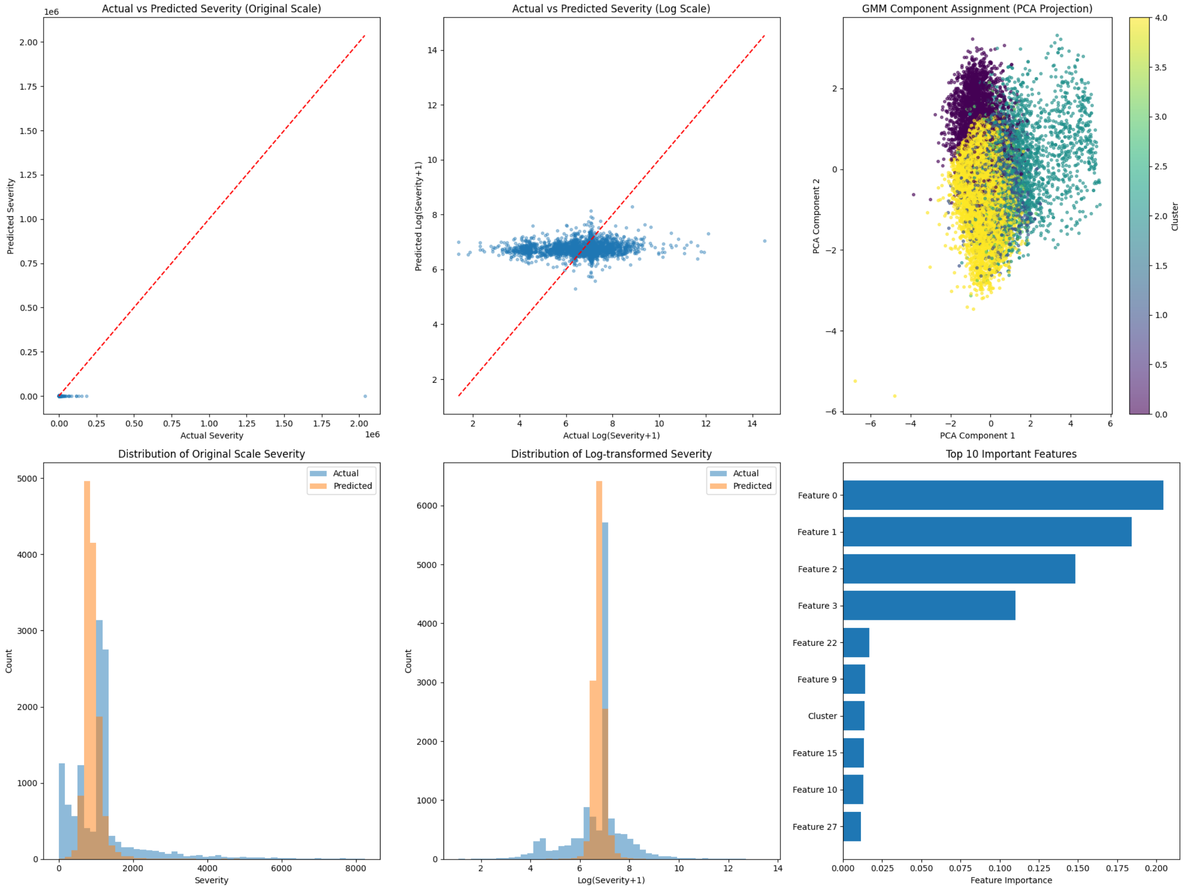

Figure 14: Predicted vs. actual severity, distributions (original and log-transformed), and feature importance from the GBM.

As seen in Figure 14, the model is able to capture the bulk distribution of severity, particularly in the central mass of the data. The log transformation appears to stabilize the distribution, and the GMM-based features seem to contribute meaningfully based on the importance chart. However, performance in the tail remains poor.

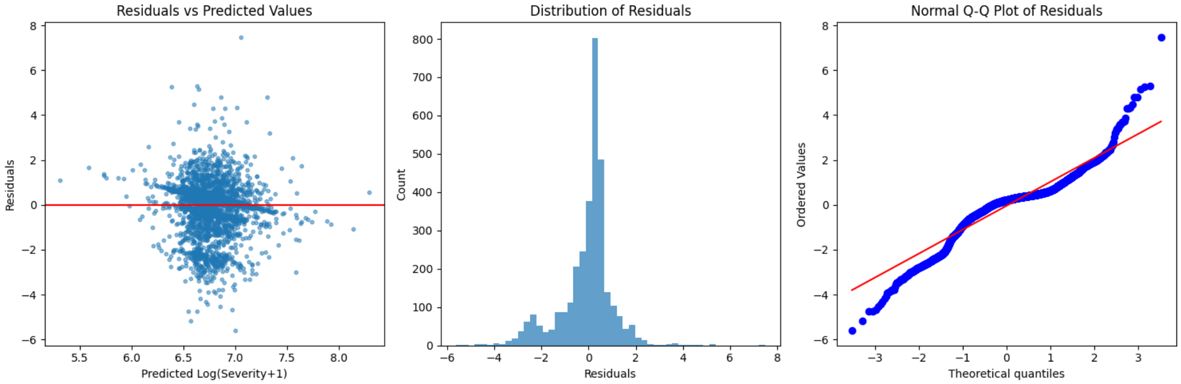

Figure 15: Residual analysis: residuals vs. predicted values, histogram of residuals, and Q–Q plot.

Figure 15 provides a residual analysis. While the residuals are roughly centered, they show clear deviation from normality, especially in the Q–Q plot. This supports the idea that, although the model fits the central tendency well, its error distribution is still quite non-Gaussian.

Overall, this model heavily prioritizes the peak of the distribution where most observations lie. This illustrates that complex models do not necessarily perform better—especially when the data lacks strong predictive signals. In practice, this model performed poorly and should be viewed primarily as an exploratory exercise.

2.4.3 Modelling $\mathbb{E}[N(t)]$

We first propose a Poisson regression model, as it is commonly used for modeling count data such as the number of claims. Consistent with standard practice, we include the logarithm of exposure, $\ln(\text{Exposure})$, as an offset in the model. To assess model appropriateness, we calculate the dispersion parameter:

$$\phi = \frac{\text{Pearson Residuals}}{n - p}$$

The dispersion parameter was measured at $\phi = 1.018751$ for all the data so we will stick with the normal Poisson instead of a quasi-poisson model or another model which adjusts the variance structure to accommodate the over dispersion in the data. The Poisson model isn't best suited for zero inflated data like this. Other models which can be tested are hurdle and zero-inflated models. Table 7 shows that the backwards selection process favored all groupings except for the regional groupings.

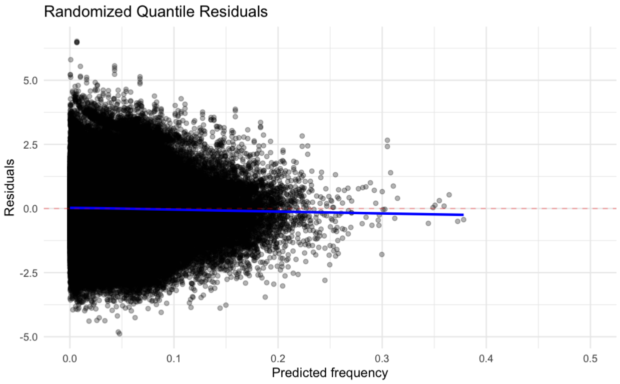

Figure 16: Randomized deviance residuals for the Poisson counting model.

Figure 16 displays the randomized deviance residuals from the fitted Poisson model. No clear anomalies are observed in the residual structure, suggesting a reasonable fit, though the model naturally emphasizes average behavior and struggles with structural zeros.

A generalized additive model (GAM) was also fitted using smoothing terms for the numerical predictors. However, the improvement in performance was marginal, likely due to the low correlation between numerical variables and claim frequency observed earlier (see Figure 10).

Table 7 presents the coefficient estimates from the final model selected via backward elimination. All categorical variables were retained except for some regional groupings, indicating limited predictive power from regional segmentation.

Variable	**Estimate**	**Std. Error**	**z-value**	**p-value**	**Sig.**
(Intercept)	-1.805e+00	4.602e-02	-39.215	< 2e−16	***
Driver Age Group 2	-5.731e-01	3.939e-02	-14.552	< 2e−16	***
Driver Age Group 3	-8.749e-01	2.820e-02	-31.025	< 2e−16	***
Driver Age Group 4	-7.987e-01	2.660e-02	-30.025	< 2e−16	***
Driver Age Group 7	-9.609e-01	3.026e-02	-31.758	< 2e−16	***
Driver Age Group 8	-9.148e-01	3.553e-02	-25.750	< 2e−16	***
Region: Basse-Normandie	-1.397e-01	5.340e-02	-2.617	0.00888	**
Region: Bretagne	-1.501e-01	3.666e-02	-4.094	4.24e-05	***
Region: Centre	-2.257e-01	3.174e-02	-7.112	1.14e-12	***
Region: Haute-Normandie	-1.097e-01	7.047e-02	-1.557	0.11957	
Region: Île-de-France	7.832e-02	3.759e-02	2.083	0.03722	*
Region: Limousin	1.453e-01	7.270e-02	1.999	0.04558	*
Region: Nord–Pas-de-Calais	7.563e-02	4.180e-02	1.809	0.07040	.
Region: Pays-de-la-Loire	-6.185e-02	3.771e-02	-1.640	0.10099	
Region: Poitou-Charentes	-6.733e-02	4.456e-02	-1.511	0.13082	
Power Group (e)	6.374e-02	2.644e-02	2.411	0.01590	*
Power Group (f)	8.013e-02	2.586e-02	3.099	0.00194	**
Power Group (g)	5.595e-02	2.572e-02	2.176	0.02958	*
Power Group: Highest	1.886e-01	3.280e-02	5.750	8.93e-09	***
Power Group: Medium	1.401e-01	3.076e-02	4.553	5.29e-06	***
Brand: Opel, GM, or Ford	3.123e-01	3.345e-02	9.338	< 2e−16	***
Brand: Other	2.625e-01	3.182e-02	8.249	< 2e−16	***
Brand: Renault, Nissan, Citroën	1.904e-01	2.682e-02	7.099	1.25e-12	***
Brand: VW, Audi, Skoda, Seat	2.777e-01	3.462e-02	8.022	1.04e-15	***
Car Age	-1.288e-02	1.524e-03	-8.452	< 2e−16	***
Fuel Type: Regular	-1.314e-01	1.614e-02	-8.143	3.86e-16	***
Density	1.623e-05	1.772e-06	9.162	< 2e−16	***

Note: Significance codes — $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$
Null deviance: 115,188 on 413,959 degrees of freedom
Residual deviance: 113,344 on 413,933 degrees of freedom
AIC: 146,750
Dispersion parameter for Poisson family taken to be 1

Table 7: Poisson Regression Model Results with Log Link

2.5 Prediction Results

The total expected loss is the pure premium. That is the theoretical amount that clients should be expected to pay if the expected value of the insurance is 0.

$$p_{Net}(t) = \mathbb{E}[S(t)] = \mathbb{E}[N(t)] \cdot \mathbb{E}[X]$$

Now we two models predicting the counting and the severity. Lets investigate our results.

2.6 Severity

It can be good to look at some statistics about the predicted and actual values.

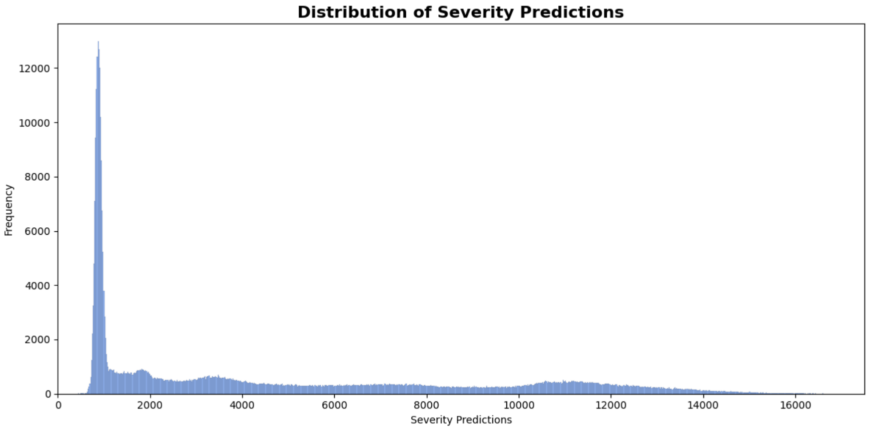

Figure 17: Distribution of predicted severity.

Table 8: Combined (Frequency × Severity) Model Validation Statistics

Metric	Value	Interpretation
Basic Error Metrics		
Mean Error	-971.72	Model underpredicts on average
Mean Absolute Error	2,871.36	Higher absolute deviation
Mean Relative Error	737.70%	Very high percentage error
RMSE	18,845.44	High error dispersion
Distribution Comparison		
Actual Mean / Predicted Mean	1,985.71 / 2,957.43	Model overpredicts by 48.94%
Actual Median / Predicted Median	1,143.00 / 1,485.36	Higher median predictions
Actual Std Dev / Predicted Std Dev	18,679.41 / 3,029.39	Much lower variance in predictions
Insurance Specific Metrics		
Actual Total Claims / Predicted	32.13M / 47.85M	48.94% overestimation
Combined Ratio	148.94%	Significantly unprofitable pricing
Probability of Loss	34.17%	Moderate risk of underpricing
Correlation Metrics		
Pearson Correlation	0.0344	Very weak linear relationship
Spearman Correlation	0.0994	Very weak rank correlation

2.6.1 Counting process

Note that our model fullfils the balance property i.e. the sum of counts is the same as the actual sum of counts.

Table 9: Frequency Model Validation Statistics

Metric	**Value**	**Interpretation**
Basic Error Metrics		
Mean Error	0.0000	No systematic bias
Mean Absolute Error	0.0816	Low absolute deviation
Mean Relative Error	40,808,534,158.30%	Extremely high (due to zero claims)
RMSE	0.2223	Low error dispersion
Distribution Comparison		
Actual Mean / Predicted Mean	0.0431 / 0.0431	Perfect match of means
Actual Median / Predicted Median	0.0000 / 0.0423	Higher median predictions
Actual Std Dev / Predicted Std Dev	0.2232 / 0.0314	Lower variance in predictions
Insurance Specific Metrics		
Actual Total Claims / Predicted	17,837 / 17,837	0.00% error in total claims
Combined Ratio	100.00%	Break-even pricing model
Probability of Loss	3.91%	Low risk of underpricing
Correlation Metrics		
Pearson Correlation	0.0975	Very weak linear relationship
Spearman Correlation	0.0975	Very weak rank correlation

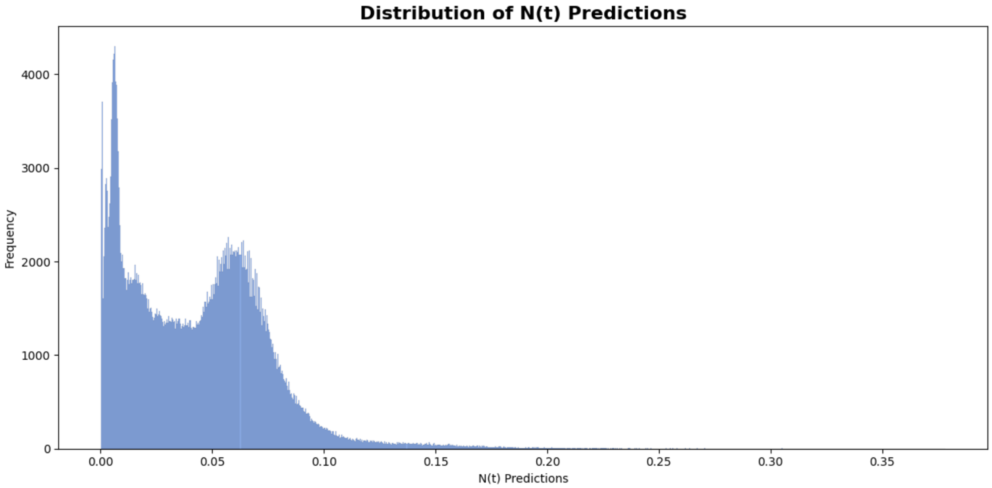

Figure 18: N(t) predictions.

2.7 Pure Premium

We observe that our model slightly overestimates the pure premium and thus apply a scaling to the severity by the ratio of actual claims to predicted pure premium. Hence our pure premium estimation can be seen in Table 10. However a pure premium per policy of 83.3 is lower than the average of france which according to LeLynx is €545 annually. That of course includes a risk premium. However it is prudent to note that our data is heavily skewed to exposures less than 1 and thus may be prone to under represent risk at an annual scale even though it is used as an offset in our models, see Figure 20.

Metric	Value
Total Pure Premium	39,634,808.37
Total Actual Claims	34,465,077.00
Scaling Factor	0.8696
Total Adjusted Losses	34,465,076.99
Premium Per Policy	83.3

Table 10: Pure Premium vs Actual Claims

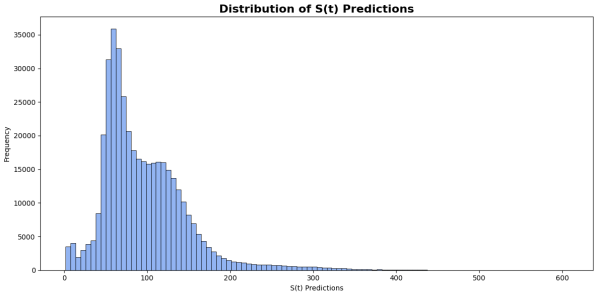

Figure 19: S(t) pure premium predictions.

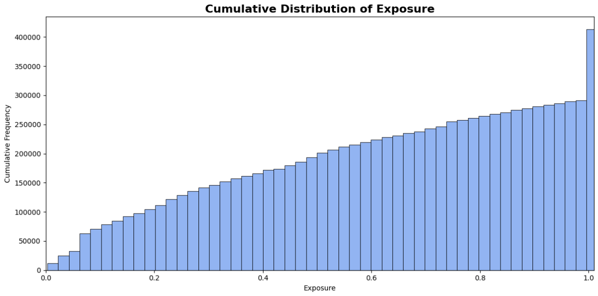

Figure 20: Cumulative distribution of Exposure.

2.8 Risk Segmentation and Pricing Adjustments

2.8.1 Expected Value Method

This method adds a safety loading to the pure premium:

$$p_{EV}(t) = (1 + \rho)E[S(t)]$$

where $\rho > 0$ is the safety loading factor. This approach is widely used due to its simplicity and intuitive interpretation, where the loading directly represents a percentage increase over expected claims.

2.8.2 Standard Deviation Method

This method incorporates risk through the standard deviation of claims:

$$p_{SD}(t) = E[S(t)] + \alpha\sqrt{\text{var}(S(t))}$$

The premium includes a loading proportional to the standard deviation, reflecting the uncertainty in claim amounts. This approach is particularly useful when the distribution approximately follows the Central Limit Theorem.

2.8.3 Combined Ratio Method

The combined ratio is a key performance indicator used by regulators and insurers to assess profitability:

$$\text{Combined Ratio} = \text{Loss Ratio} + \text{Expense Ratio} + \text{Reinsurance Ratio}$$

where:

$$\text{Loss Ratio} = \frac{\text{Paid Claims} + \text{Change in Claims Reserve}}{\text{Net Written Premiums}}$$
$$\text{Expense Ratio} = \frac{\text{Operating Expenses}}{\text{Net Written Premiums}}$$
$$\text{Reinsurance Ratio} = \frac{\text{Reinsurance Premiums} - \text{Reinsurance Recoveries}}{\text{Net Written Premiums}}$$

A combined ratio below 100% indicates underwriting profit, while above 100% signals a loss. Premium adjustments maintain this ratio at target levels (typically 75-95%).

2.8.4 Solvency-Based Premium Adjustment

Based on regulatory capital requirements like Solvency II:

$$p_{Solv}(t) = E[S(t)] + \delta \cdot \text{SCR}$$

where SCR is the Solvency Capital Requirement and δ is a factor representing return on regulatory capital. Premium adjustments ensure sufficient capital to maintain the required solvency ratio:

$$\text{Solvency Ratio} = \frac{\text{Available Capital}}{\text{Required Capital}} \geq 100\%$$

2.8.5 Risk-Based Capital Approach

Premium adjustments based on risk-based capital (RBC) requirements:

$$p_{RBC}(t) = E[S(t)] + \gamma \cdot \text{CoC} \cdot \text{EC}$$

where EC is the economic capital allocation, CoC is the cost of capital (typically 6-10%), and γ is a business factor. The premium ensures that the return on risk-adjusted capital meets or exceeds targets:

$$\text{RAROC} = \frac{\text{Risk-Adjusted Return}}{\text{Economic Capital}} \geq \text{Hurdle Rate}$$

Insurance regulators often require this ratio to exceed minimum thresholds, driving premium adjustments.

2.9 Pricing

We choose the expected value loading factor with $\rho = 0.2$ and the combined ratio method of 0.75. See Figure 21.

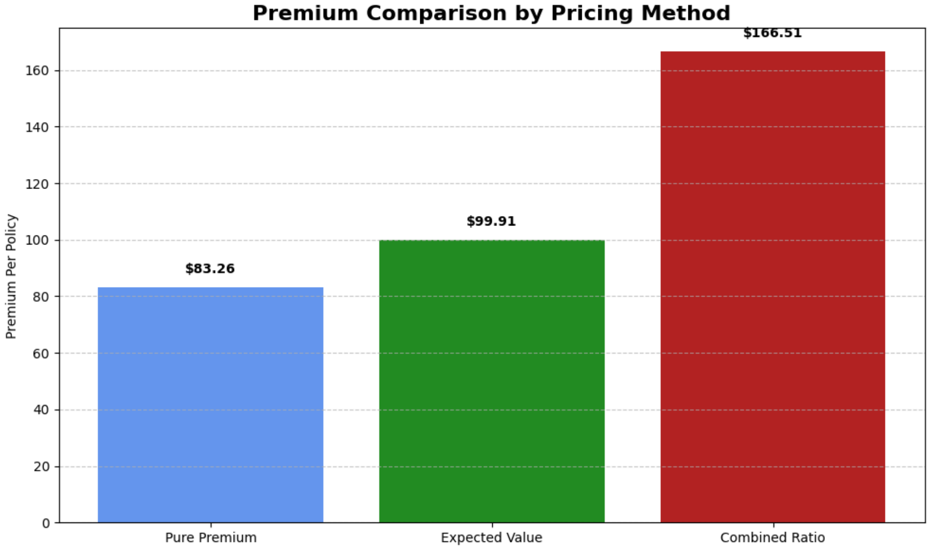

Figure 21: Insurance cost per policy.

Analysis of Feature Effects on Insurance Risk Score S(t)

Risk Factor Analysis Report

2.10 Overview

This analysis examines how different factors influence the insurance risk score S(t), which has an overall average of 83.26. Features are ranked by impact score, revealing which characteristics most strongly affect risk assessment.

2.11 Key Findings

2.11.1 Age Risk Group (Impact: 107.47)

Age emerges as the most influential factor with stark variations:

- Young drivers (17-26.5 years) dramatically increase risk scores by up to 107.9%
- Middle-aged drivers (30.5-55.5), representing 252,624 cases, reduce risk by 6.2%
- Older drivers systematically decrease risk, with the 72.5+ group reducing S(t) by 21.1%

2.11.2 Geographic Region (Impact: 26.58)

Significant regional disparities exist:

- Urban areas (Ile-de-France) show elevated risk (+21.4%)
- Northern France broadly carries higher risk (+16.1%)
- Centre region demonstrates the largest risk reduction (-10.5%)

2.11.3 Vehicle Brand (Impact: 18.80)

Clear patterns emerge across manufacturers:

- German automotive groups (Volkswagen/Audi) increase risk most significantly (+16.2%)
- American brands (Opel/GM/Ford) also substantially increase risk (+11.1%)
- French brands (Renault/Nissan/Citroen), the largest segment (218,591 cases), uniquely decrease risk (-6.4%)

2.11.4 Vehicle Power (Impact: 16.86)

Power categories show consistent patterns:

- Highest power vehicles increase risk substantially (+12.5%)
- Lower power categories ('d' and 'g') decrease risk by approximately 7.8%
- 9 of 12 power categories increase risk, with only 3 showing decreases

2.11.5 Fuel Type (Impact: 12.03)

A clear division exists between fuel types:

- Diesel vehicles increase risk by 7.2%
- Regular fuel vehicles decrease risk by 7.2%
- The dataset is balanced between these categories (206,350 vs. 207,610)

2.12 Strategic Implications

1. **Age-based pricing:** The dominant impact of driver age (particularly for young drivers) should guide premium calculation and risk mitigation programs.
2. **Geographic targeting:** Insurance offerings should be tailored by region, with different approaches for Northern France and major metropolitan areas.
3. **Vehicle-specific factors:** Brand, power, and fuel type significantly impact risk assessment. French-manufactured, lower-powered, and regular fuel vehicles present more favorable risk profiles.

4. **Portfolio optimization:** The data reveals lower-risk categories often contain larger numbers of policyholders, suggesting opportunities for targeted expansion in these segments.
5. **Refined segmentation:** The clear risk differentials across all features justify sophisticated segmentation in pricing and underwriting practices.

These findings enable more precise risk-based pricing, strategic marketing initiatives targeting lower-risk segments, and specialized risk mitigation programs for higher-risk categories.

Table 11: Analysis of Feature Effects on S(t) (Overall average: 83.26)

Feature	**Group**	**Mean S(t)**	**% Diff**	**Count**	**Effect**
7***age_risk_group**	(17.0, 22.5]	173.12	+107.9%	11,322	INCREASE
	(22.5, 26.5]	157.37	+89.0%	21,182	INCREASE
	(72.5, 100.0]	65.65	-21.1%	19,591	DECREASE
	(61.5, 72.5]	67.35	-19.1%	38,168	DECREASE
	(55.5, 61.5]	69.20	-16.9%	36,427	DECREASE
	(26.5, 30.5]	88.54	+6.3%	34,646	INCREASE
	(30.5, 55.5]	78.09	-6.2%	252,624	DECREASE
10***region**	Ile-de-France	101.08	+21.4%	69,989	INCREASE
	Nord-Pas-de-Calais	99.39	+19.4%	27,357	INCREASE
	Limousin	97.13	+16.7%	4,580	INCREASE
	Centre	74.50	-10.5%	160,814	DECREASE
	Haute-Normandie	76.48	-8.1%	8,795	DECREASE
	Bretagne	77.16	-7.3%	42,200	DECREASE
	Basse-Normandie	77.37	-7.1%	10,916	DECREASE
	Poitou-Charentes	78.70	-5.5%	19,081	DECREASE
	Aquitaine	85.54	+2.7%	31,399	INCREASE
	Pays-de-la-Loire	84.62	+1.6%	38,829	INCREASE
4***region_grouped**	Northern France	96.62	+16.1%	117,057	INCREASE
	Major Metropolitan	74.50	-10.5%	160,814	DECREASE
	Western France	80.73	-3.0%	81,029	DECREASE
	Southwestern France	84.13	+1.0%	55,060	INCREASE
8***brand**	Volkswagen, Audi, Skoda or Seat	96.72	+16.2%	32,707	INCREASE
	Opel, General Motors or Ford	92.46	+11.1%	37,477	INCREASE
	Japanese (except 2) or Korean	91.00	+9.3%	1	INCREASE
	Mercedes, Chrysler or BMW	89.10	+7.0%	19,314	INCREASE
	Renault, Nissan or Citroen	77.92	-6.4%	218,591	DECREASE
	Fiat	87.50	+5.1%	16,757	INCREASE
	other	86.06	+3.4%	9,886	INCREASE
	Japanese (except Nissan) or Korean	85.40	+2.6%	79,227	INCREASE
6***brand_grouped**	Volkswagen, Audi, Skoda or Seat	96.72	+16.2%	32,707	INCREASE
	Opel, General Motors or Ford	92.46	+11.1%	37,477	INCREASE
	Mercedes, Chrysler or BMW	89.10	+7.0%	19,314	INCREASE
	Renault, Nissan or Citroen	77.92	-6.4%	218,591	DECREASE
	Other	86.97	+4.5%	26,643	INCREASE

Continued on next page

Table 11: Analysis of Feature Effects on S(t) (Overall average: 83.26)

Feature	**Group**	**Mean S(t)**	**% Diff**	**Count**	**Effect**
	Japanese (except Nissan) or Korean	85.40	+2.6%	79,227	INCREASE
12***power**	l	93.66	+12.5%	4,689	INCREASE
	j	89.97	+8.1%	18,074	INCREASE
	d	76.80	-7.8%	68,150	DECREASE
	g	76.82	-7.7%	91,351	DECREASE
	o	89.47	+7.5%	1,511	INCREASE
	k	89.24	+7.2%	9,554	INCREASE
	n	89.01	+6.9%	1,311	INCREASE
	f	88.81	+6.7%	95,902	INCREASE
	m	87.49	+5.1%	1,835	INCREASE
	h	86.26	+3.6%	26,748	INCREASE
	e	85.42	+2.6%	77,189	INCREASE
	i	83.07	-0.2%	17,646	DECREASE
6***power_grouped**	Highest Power	90.07	+8.2%	36,974	INCREASE
	d	76.80	-7.8%	68,150	DECREASE
	g	76.82	-7.7%	91,351	DECREASE
	f	88.81	+6.7%	95,902	INCREASE
	e	85.42	+2.6%	77,189	INCREASE
	Medium Power	85.00	+2.1%	44,394	INCREASE
2***gas**	Diesel	89.29	+7.2%	206,350	INCREASE
	Regular	77.26	-7.2%	207,610	DECREASE

3 Large Claim Analysis and Reinsurance Impact

A Danish insurance company has collected data on fire damage claims from 2010 to 2020. The data contains 2164 claims which are measured in millions of Danish kroner (DKK). The company has 630 000 fire insurance policies, with an annual claim frequency of 900 per million policies. The current premium per policy is 3 331 DKK.

The company is evaluating two Excess-of-Loss (XL) reinsurance contracts:

1. **Contract 1:** The insurer retains up to 10 million DKK per claim, and the reinsurer covers excess losses up to 40 million DKK. The reinsurance premium is 609 DKK per policy.
2. **Contract 2:** The insurer retains 10 million DKK, and the reinsurer covers all excess losses without an upper limit. The reinsurance premium is 872 DKK per policy.

In the following sections, we evaluate the financial impact of the two reinsurance contracts based on historical fire claim data. We begin by modeling the severity of individual claims, then use this to simulate the distribution of total annual claims via a compound Poisson process. From there, we analyze the revenue implications of each reinsurance strategy, comparing key risk metrics such as Value-at-Risk and Conditional Tail Expectation. This allows us to assess which reinsurance contract provides the best trade-off between profitability and risk reduction.

3.1 Loss Distribution Analysis

To assess the risk profile of the insurance portfolio, we analyze the severity distribution of claims. The dataset reveals a heavy-tailed loss distribution, indicating a small number of high-cost claims.

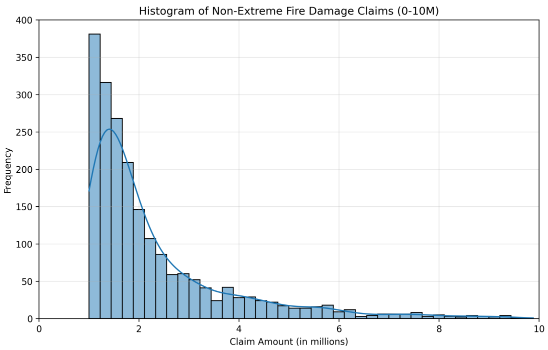

Figure 22: Non-Extreme Fire Damage Claims ($\leq$ 10M)

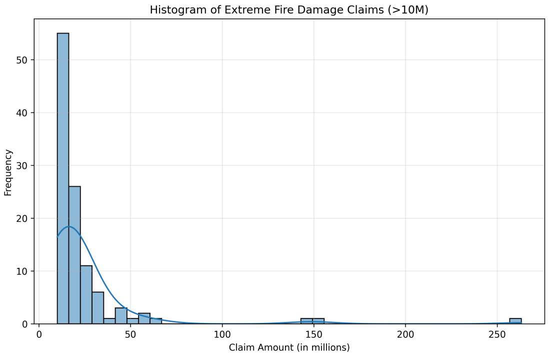

Figure 23: Extreme Fire Damage Claims ($>$ 10M)

The histograms in Figure 22 and Figure 23 illustrate the distribution of claim amounts, separated into non-extreme and extreme values. Claims below 10 million DKK make up the majority of cases (95%), while the extreme claims, though rare, contribute significantly to overall losses.

To better understand claim patterns over time, Figure 24 provides a yearly breakdown of total claims.

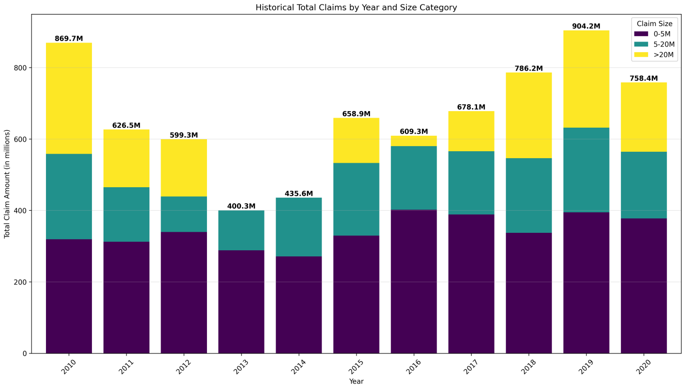

Figure 24: Total Fire Damage Claims by Year (2010-2020)

The historical data reveals that claim amounts fluctuate over time, with some years experiencing significantly fewer large claims (e.g., 2013 and 2014). While these variations are visible, we do not explicitly incorporate time dynamics in our modeling due to limited data. However, temporal trends could be explored in future work if a longer time series becomes available.

3.1.1 Expected Loss Calculation

Given that claim frequency follows a Poisson distribution with an expected number of claims per year:

$$\lambda = \frac{900}{10^6} \times 630\,000 = 567$$

The total expected loss each year without reinsurance is computed as:

$$E[S] = \lambda \cdot E[X] = 567 \times 3.39 = 1\,919.67 \text{ million DKK}$$

where the expected loss per claim $E[X]$ is calculated empirically from the dataset.

The variance of total losses per year is then computed as:

$$\mathrm{Var}(S) = \lambda E[X^2] = 567 \times 83.90 = 47\,569.83 \text{ million DKK}^2$$

where the expected loss of individual claims squared $E[X^2]$ is also derived empirically.

3.2 Distribution Approximation

The accurate modeling of claim severity is crucial for risk assessment, particularly when estimating the impact of reinsurance on total annual losses. Since this modeling directly influences our understanding of both expected claims and tail risk, we evaluate several approaches to approximate the claim distribution as accurately as possible before proceeding to simulate total claims and analyze revenue outcomes.

3.2.1 Modeling Approaches

We split the data into 50% training and 50% test sets to evaluate different modeling approaches:

1. **Pure Empirical:** Using the empirical distribution from the training set with no parametric assumptions.
2. **Hybrid Model:** Empirical distribution for common claims, theoretical distribution for extreme claims.
3. **Two-Part Theoretical:** Different theoretical distributions for non-extreme and extreme claims.
4. **Single Theoretical:** One parametric distribution for all claims.

For the hybrid model, we tested different percentile thresholds to determine the optimal split point between empirical and theoretical components. Performance was evaluated using multiple metrics including Wasserstein distance, KL divergence, and relative errors in key statistics (mean, standard deviation, and high quantiles).

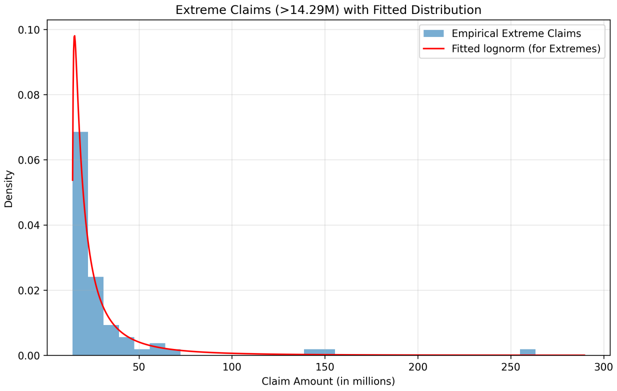

Figure 25: Extreme claims (>14.29M) fitted with a log-normal distribution.

3.2.2 Optimal Model Selection

Cross-validation consistently showed that the hybrid approach outperformed other methods, with the configuration:

- 97% of claims (below 14.29M DKK) modeled using the empirical distribution
- 3% of claims (above 14.29M DKK) modeled using a fitted log-normal distribution

This 97/3 hybrid approach provides the best of both worlds: data-driven modeling for common claims where we have abundant observations, and theoretical extrapolation for extreme events where data is sparse.

Figure 25 shows how the log-normal distribution fits the extreme claims. While different theoretical distributions (Pareto, generalized Pareto) were competitive for modeling extreme values, the log-normal consistently performed well across different splits of the data.

The resulting total loss distribution from this hybrid approach is shown in Figure 27.

3.3 Reinsurance Modeling and Financial Impact

Reinsurance reduces risk exposure by capping individual claim payments. The reinsured loss per claim, X_{XL}, follows:

$$X_{XL} = \begin{cases} 0 & \text{if } X \leq M \\ X - M & \text{if } M < X \leq M + L \\ L & \text{if } X > M + L \end{cases}$$

Figure 26 compares the total claim amounts retained by the insurer under different reinsurance scenarios.

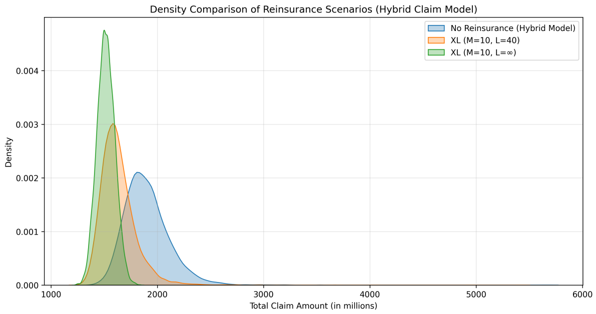

Figure 26: Density comparison of retained claims under different reinsurance scenarios.

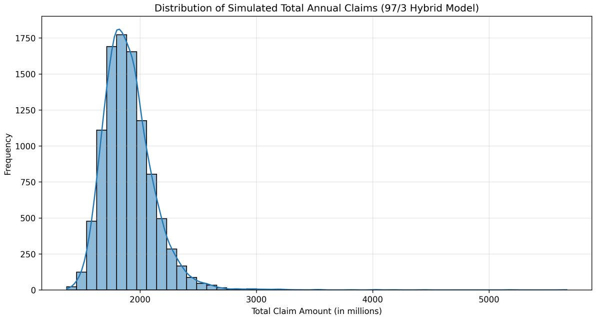

Figure 27: Distribution of total annual claims generated using the 97/3 hybrid model.

The density plot shows that Contract 1 ($M = 10, L = 40$) shifts the distribution to the left while maintaining a similar shape to the no-reinsurance scenario. Contract 2 ($M = 10, L = \infty$), however, dramatically narrows the distribution, virtually eliminating the tail risk.

Using the hybrid claim simulation model, we derive risk metrics for each reinsurance scenario, shown in Table 12.

Metric	**No Reinsurance**	**XL (M=10, L=40)**	**XL (M=10, L=∞)**
Expected Value	1 903.24	1 632.83	1 516.29
Standard Deviation	223.69	179.64	83.02
Coefficient of Variation	0.118	0.110	0.055
99% VaR	2 552.92	2 196.76	1 713.60
Interquartile Range	469.04	479.01	98.41

Table 12: Comparison of risk metrics under different reinsurance scenarios.

Table 13 provides a detailed comparison of key extreme quantiles across the reinsurance scenarios.

Quantile	No Reinsurance	XL (M=10, L=40)	XL (M=10, L=∞)
90%	2 168.56	1 827.41	1 624.83
95%	2 280.37	1 927.63	1 655.93
99%	2 556.76	2 203.20	1 716.43
99.9%	3 213.13	2 896.35	1 787.87

Table 13: Quantiles of simulated total losses under different reinsurance structures (50,000 simulations).

3.4 Revenue and Profit Analysis

When selecting reinsurance contracts, insurers must weigh the trade-off between risk reduction and profitability. To evaluate this balance, we analyze the revenue implications of each contract, factoring in both premium income and the cost of claims and reinsurance.

While we do not include a separate figure for the revenue density, it closely mirrors the distribution of total annual claims shown in Figure 26. Each revenue distribution is simply a horizontally shifted version of the corresponding claims distribution, meaning the shapes and variability remain comparable. In particular, the no-reinsurance option displays the widest spread, reflecting greater financial volatility.

To better understand the risk-return tradeoff, we use a quantile plot of simulated net revenue, shown in Figure 28. The figure presents the full distribution of revenue under each reinsurance scenario.

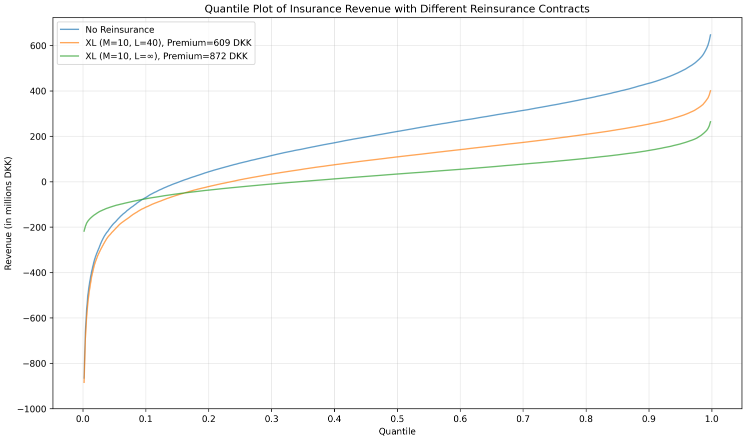

Figure 28: Quantile plot of insurance revenue with different reinsurance contracts.

Contract 2 ($M = 10, L = \infty$) clearly dominates in the lowest 9% of quantiles, offering significantly higher revenue in the most adverse outcomes. In contrast, the no reinsurance option performs better in approximately 91% of cases, especially in higher quantiles where upside potential is retained. Contract 1, however, is consistently outperformed by the no reinsurance option across almost the entire quantile range, indicating that it provides limited value. The plot spans from the 0.2% to the 99.8% quantiles, offering a detailed view of both tails while excluding only the most extreme, unlikely events. This visual comparison underscores the trade-off between profit potential and tail risk protection and helps identify which contracts are truly viable.

The revenue analysis in Table 14 shows a crucial insight: while Contract 2 has the lowest expected revenue (32.07M DKK), it provides the best protection against catastrophic outcomes. The 1% VaR (Value-at-Risk) under Contract 2 is -167.26M DKK, compared to -458.23M DKK with no reinsurance and -488.34M DKK under Contract 1. However, VaR only tells us the threshold value for the worst 1% of cases — it does not reflect the average outcome in such scenarios. For that, the Conditional Tail Expectation (CTE) is more informative. The 1% CTE under Contract 2 is -198.08M DKK, while it reaches -720.39M DKK without reinsurance and -755.07M DKK under Contract 1, showing that even the average outcomes in extreme cases are significantly more favorable with Contract 2.

To understand the practical impact of these figures, it helps to compare them with the expected revenue. Under no reinsurance, the expected annual revenue is 194.04M DKK, meaning that the average outcome in the worst 1% of years (1% CTE) corresponds to a loss of 3.7 times the expected revenue. For Contract 2, that multiple increases to 6.2 times — a higher ratio, but with significantly smaller absolute losses. This reflects the trade-off in stability: while Contract

Metric	No Reinsurance	XL (M=10, L=40)	XL (M=10, L=∞)
Expected Revenue	194.04	81.53	32.07
Standard Deviation	215.87	170.86	83.19
Coefficient of Variation	1.11	2.10	2.59
Loss Ratio	90.75%	96.11%	98.47%
5% VaR (Loss)	-181.84	-212.77	-106.76
5% CTE (Loss)	-367.98	-396.01	-143.92
1% VaR (Loss)	-458.23	-488.34	-167.26
1% CTE (Loss)	-720.39	-755.07	-198.08

Table 14: Revenue metrics under different reinsurance scenarios.

2 offers lower profits in good years, it limits how bad the worst years can get. Contract 1, however, performs poorly on both fronts: it reduces expected revenue to 81.53M DKK while producing a higher average loss in extreme years (-755.07M DKK) than even no reinsurance — a result that makes it a clearly suboptimal choice.

The loss ratio increases across the three scenarios, reflecting the cost of transferring risk to reinsurers. Contract 2 has the highest loss ratio at 98.47%, compared to 90.75% with no reinsurance. This reflects a strategic trade-off: a smaller retained profit in exchange for much lower earnings volatility and improved protection against extreme events. While Contract 2 may require the company to save over several years to prepare for rare adverse outcomes, its more stable risk profile makes it better suited for long-term solvency planning.

3.5 Strategic Recommendations

Based on our comprehensive analysis of the reinsurance options, we recommend:

- **Rejection of Contract 1:** Contract 1 underperforms across all key metrics. It delivers lower expected revenue than no reinsurance while increasing exposure to large losses. Its 1% and 5% CTE values are worse than operating without reinsurance, and as shown in Figure 28, its revenue curve is consistently below the no reinsurance option. This indicates it introduces cost without providing meaningful protection and should be ruled out.
- **Preferred Contract:** Contract 2 ($M = 10$, $L = \infty$) provides the best protection against extreme losses and offers the lowest risk of financial distress. Although it reduces expected annual revenue to 32.07M DKK, the reduction in catastrophic downside makes it a strong choice for long-term stability. Its 1% VaR is 63.5% lower than the no reinsurance scenario, and it cuts average losses in the worst 1% of years from -720.39M DKK to -198.08M DKK — a significant improvement in solvency protection.
- **Risk of Financial Distress:** In a 1-in-100 year scenario, the company could face a net revenue shortfall of at least 458.23M DKK (1% VaR) and on average 720.39M DKK (1% CTE) without reinsurance — nearly 3.7 times the expected annual revenue. Under Contract 2, this is reduced to at least 167.26M DKK and on average 198.08M DKK — 6.2 times the expected revenue. Although the relative hit is larger, the absolute losses are much more manageable.
- **Profitability Considerations:** While Contract 2 reduces expected annual profit to 32.07M DKK (compared to 194.04M DKK without reinsurance) and increases the loss ratio to 98.47% (from 90.75%), these costs should be viewed as strategic investments in financial stability rather than simply lost profit. The higher loss ratio reflects the premium paid for protection against catastrophic losses, which brings substantial benefits in terms of reduced earnings volatility and potentially lower regulatory capital requirements.
- **Model Limitations:** This analysis is based on Danish fire insurance data from 2010–2020, a period covering only 11 years. It may not capture low-frequency, high-severity events that occur every 50+ years. Furthermore, the model assumes constant claim frequency and that the statistical distribution of claim severity remains unchanged in the future. If climate risks or population growth increase future claim volume or size, actual losses could exceed our estimates. Finally, net revenue metrics do not account for administrative costs like staff salaries, which further reduce the company's retained earnings and recovery capacity.

The optimal choice between no reinsurance and Contract 2 ultimately depends on the company's risk appetite, regulatory capital requirements, and competitive environment. For companies prioritizing financial stability, trustworthiness, and long-term solvency, Contract 2 is the most prudent and resilient option.

4 Conclusion

This report demonstrates how statistical modeling and simulation can support more accurate pricing and risk control in non-life insurance. In the pure premium estimation task, we found that generalized linear models combined with a spliced distribution effectively captured key claim patterns, enabling more nuanced pricing based on risk characteristics. While some model limitations remained, such as limited explanatory power in the severity model, the structure still supports fair and data-aligned pricing.

In the reinsurance analysis, a hybrid empirical-lognormal model provided a strong fit for large loss modeling. Simulations showed that an unlimited excess-of-loss contract significantly reduced exposure to rare but severe losses, despite lower average profit. Contract 1, however, offered limited protection and was outperformed across almost all metrics. These findings underscore the importance of matching risk strategy with financial goals—whether that's maximizing expected revenue or minimizing extreme downside risk.

A Code Implementation

This appendix provides a concise overview of the key computational methods used in our analysis. We focus on the most critical aspects of our implementation rather than common visualization or data loading functions.

A.1 Finding the Optimal Claim Distribution Model

Our analysis required comparing different approaches for modeling the claim severity distribution. We implemented a testing framework in the `compare_claim_models.py` script to evaluate:

1. **Pure empirical resampling**
2. **Pure theoretical models**
3. **Hybrid approaches with different threshold percentiles**

The key comparison code that evaluated competing models against a test set is shown below:

```python
def compare_hybrid_percentages(percentiles):
    """
    Compare hybrid models with different percentage thresholds.
    """
    # Load data and create train-test split
    data = load_data()
    claims = data["Loss"].values
    train_idx, test_idx = create_random_split(claims)
    train_claims = claims[train_idx]
    test_claims = claims[test_idx]

    # Create bootstrap reference distribution from test set
    bootstrap_samples = []
    for _ in range(bootstrap_size):
        sampled_claims = np.random.choice(test_claims, size=len(test_claims),
                                          replace=True)
        bootstrap_samples.append(np.sum(sampled_claims))
    bootstrap_samples = np.array(bootstrap_samples)

    # Evaluate empirical baseline and different hybrid models
    results = {}
    hybrid_results = {}

    # Test empirical model first
    empirical_totals = simulate_empirical(train_claims, num_claims_test)
    empirical_metrics = evaluate_distributions(empirical_totals, bootstrap_samples,
                                              test_claims, "Empirical")
    results["empirical"] = empirical_metrics

    # Test hybrid models with different percentiles
    for percentile in percentiles:
        model_name = f"hybrid_{percentile}"
        hybrid_totals, hybrid_details = simulate_hybrid(train_claims,
                                                       num_claims_test,
                                                       percentile)
        hybrid_metrics = evaluate_distributions(hybrid_totals, bootstrap_samples,
                                               test_claims, f"Hybrid {percentile}%")
        hybrid_results[model_name] = hybrid_metrics

    # Find the best hybrid model based on a composite score
    best_hybrid = None
    best_score = float('inf')
    for model_name, metrics in hybrid_results.items():
        # Calculate balanced score (lower is better)
        score = (abs(metrics['mean_rel_error']) * 0.2 +
                abs(metrics['std_rel_error']) * 0.2 +
                abs(metrics['q99_rel_error']) * 0.2 +
                metrics['kl_full'] * 0.2 +
                metrics['wasserstein']/100 * 0.2)

        if score < best_score:
            best_score = score
            best_hybrid = model_name
```

This code systematically evaluates different percentile thresholds for hybrid models (94-99%) and identifies the optimal split point based on a balanced score that considers mean error, standard deviation error, 99% quantile error, KL divergence, and Wasserstein distance.

A.2 Hybrid Simulation Approach

After identifying that the 97/3 hybrid model performed best, we implemented this approach in the `simulate_claims_composite` function:

```
def simulate_claims_composite(data, num_simulations=NUM_SIMULATIONS,
                             lambda_=567, threshold_percentile=97):
    """
    Simulate claims using a hybrid approach:
    - Use empirical sampling for claims below the threshold
    - Use fitted parametric distribution for claims above the threshold
    """
    # Calculate threshold (97th percentile by default)
    claims = data["Loss"].values
    threshold = np.percentile(claims, threshold_percentile)

    # Split claims into non-extreme and extreme
    non_extreme = claims[claims <= threshold]
    extreme = claims[claims > threshold]

    # Find best distribution for extreme claims
    best_dist, best_params = find_best_distribution(extreme)

    # Simulate claims
    total_claims = []
    all_individual_claims = []

    for _ in range(num_simulations):
        # Generate Poisson number of claims
        num_claims = np.random.poisson(lambda_)

        # For each claim, decide if it's extreme
        p_extreme = len(extreme) / len(claims)
        claim_types = np.random.choice(['non-extreme', 'extreme'],
                                      size=num_claims,
                                      p=[1-p_extreme, p_extreme])

        # Generate individual claims
        individual_claims = []
        for claim_type in claim_types:
            if claim_type == 'non-extreme':
                # Sample from empirical distribution of non-extreme claims
                claim = np.random.choice(non_extreme)
            else:
                # Sample from fitted distribution for extreme claims
                claim = best_dist.rvs(*best_params)
                # Ensure claim is at least at the threshold
                claim = max(claim, threshold)

            individual_claims.append(claim)

        # Calculate total claim amount
        total_claim = sum(individual_claims)
        total_claims.append(total_claim)
        all_individual_claims.append(individual_claims)

    return np.array(total_claims), all_individual_claims
```

This function implements our 97/3 hybrid approach by:

1. Splitting claims at the 97th percentile threshold
2. Using direct resampling for claims below the threshold
3. Fitting a log-normal distribution for extreme claims
4. Generating a realistic mix of normal and extreme claims based on their empirical proportions
5. Simulating the Poisson frequency distribution for total annual claims

A.3 Reinsurance Impact Modeling

To evaluate reinsurance contracts, we implemented the `apply_xl_reinsurance_to_individual` function that applies Excess-of-Loss terms to individual claims:

```
def apply_xl_reinsurance_to_individual(claim, M, L):
    """
    Apply Excess-of-Loss (XL) reinsurance to an individual claim.
    Args:
        claim: Individual claim amount
        M: Retention limit (insurer's retention)
        L: Upper limit (ceding limit)

    Returns:
        float: Net claim retained by the insurer
    """
    if L == np.inf:
        # With unlimited cover, insurer retains min(claim, M)
        return min(claim, M)
    else:
        # With limited cover, insurer retains min(claim, M) + max(0, claim - (M+L))
        return min(claim, M) + max(0, claim - (M + L))
```

This function implements the mathematical formula for the XL reinsurance structure, handling both capped and unlimited reinsurance contracts.

A.4 Evaluation Metrics

We used several metrics to compare distribution similarity, with Kullback-Leibler (KL) divergence and Wasserstein distance playing central roles:

```
def calculate_kl_divergence(p_samples, q_samples, bins=20, smoothing=1e-10):
    """
    Calculate the KL divergence between two sets of samples.
    Uses adaptive binning and smoothing to avoid numerical issues.
    """
    # Determine bin edges based on combined range
    min_val = min(np.min(p_samples), np.min(q_samples))
    max_val = max(np.max(p_samples), np.max(q_samples)) * 1.1

    # Create bins and calculate histograms
    bin_edges = np.linspace(min_val, max_val, bins+1)
    p_hist, _ = np.histogram(p_samples, bins=bin_edges, density=True)
    q_hist, _ = np.histogram(q_samples, bins=bin_edges, density=True)

    # Apply smoothing and normalize
    p_hist = p_hist + smoothing
    q_hist = q_hist + smoothing
    p_hist = p_hist / np.sum(p_hist)
    q_hist = q_hist / np.sum(q_hist)

    # Calculate KL divergence: KL(p||q)
    return np.sum(p_hist * np.log(p_hist / q_hist))
```

We also used the Wasserstein distance (Earth Mover's Distance) from SciPy, which measures how much "work" is needed to transform one distribution into another. These metrics allowed us to quantitatively assess which model most accurately reproduced the true claim distribution, especially in the crucial tail regions.