# Motor Insurance Portfolio

Simone Deponte

This project aims to simulate a portfolio of insured individuals based on specific characteristics like vehicle type, power, and location, analyzing crucial insurance metrics such as value at risk, empirical PDF, pure premium estimation, solvency ratio, mean excess function, and reinsurance options. It comprises two main parts: data fitting and analysis, which is internal, and a publicly released GUI application for insurance professionals. The study explores compound and individual risk models, leveraging Monte Carlo methods and Generalized Linear Models (GLMs) to determine optimal distributions and parameters. It also defines key reinsurance concepts: quota-share and excess-of-loss treaties.

## I. INTRODUCTION

Motor vehicle insurance is an essential component of modern civilization, protecting individuals, businesses, and the economy at large. As the number of vehicles on the road grows worldwide, motor insurance provides critical financial protection against a variety of dangers, including accidents, damage, theft, and liability. This coverage ensures that individuals and organizations can cope with the potentially high expenses of vehicle-related catastrophes, encouraging financial stability and peace of mind.

The significance of motor vehicle insurance goes beyond personal protection. It contributes to the overall economic system by reducing the financial burden of traffic accidents and ensuring that victims are compensated for their losses. This not only benefits in healing, but also helps to keep transportation systems and the economy running smoothly.

Insurance companies must successfully manage risk. The capacity to effectively analyze and price risks enables insurers to provide fair and competitive premiums while remaining financially stable. Proper risk management enables insurers to anticipate and prepare for future claims, protecting their operations and ensuring they can satisfy their commitments to policyholders.

Understanding and resolving risks in a world where they are ever-present and continually evolving is critical for insurance firms' long-term viability. By efficiently managing risk, insurers can provide dependable coverage, enhance economic resilience, and contribute to societal safety and stability.

By investing in strong modeling and simulation techniques, the insurance business can maintain its key functions and stay resilient in the face of changing difficulties.

Unfortunately, open-access databases in the insurance industry are quite rare. Companies are often hesitant to share their data because it is the foundation of their business.

The University of Valencia was granted access to an anonymized (in order to comply with European legislation, protecting individual privacy and confidentiality) dataset of a Spanish non-life insurance company's motor vehicle insurance portfolio for a research project. The database contains 105,555 rows of data and 30 different variables. These data come from a non-life insurance firm based in Spain. This dataset contains important date-related information such as policy effective dates, insured persons' birthdates, and renewal dates. Additionally, this dataset includes key economic variables including as premiums and claim expenses. These economic indicators are required for conducting in-depth studies of the financial feasibility and profitability of automobile insurance policies, hence they are the ones evaluated.

## II. RESEARCH QUESTION

The goal of this project is to simulate a possible portfolio of insured people (following some specific characteristics, such as the vehicle they wish to insure, its power, and the location where they live) for a whole year and determine some very important and interesting statistics that are very useful in the insurance world, such as value at risk, empirical pdf of the entire portfolio, pure premium estimation, solvency ratio, plotting the mean excess function, and reinsurance option assessments. The project will be divided into two parts.

- Data Fitting and Analysis: This section, which involves fitting data and doing analyses, it's not intended to be public
- GUI Application: The application will be released with a graphical user interface, allowing insurance workers to simulate their needs.

To accomplish so, using data from a genuine insurance company and describing the real world, the first step is to select and identify the best model to replicate the portfolio.

Finding a model that accurately fits the data is a major difficulty in the motor insurance sector. Extensive studies have been undertaken in this field, resulting in the identification of several prevalent models.

The most frequently acknowledged model for a policyholder's risk is a compound model, which is defined by two main random variables. The first variable ($N(t)$) reflects the number of claims that may occur within a certain time period, and the second variable ($Y$) represents the monetary worth of a claim, assuming that it

occurs. The compound model is the most often utilized in the business.

$$S_1 = \sum_{0}^{N(t)} Y \tag{2.1}$$

In contrast, another model, the individual risk model, simplifies the approach: N represents the chance of at least one claim occurring (namely a bernoulli), and Y represents the value of the claim if at least one claim occurs. This model, while basic, has limited applicability and is unsuitable for simulations or more complex studies.

$$S_1 = N * Y \tag{2.2}$$

On the other hand, the compound model is significantly more adaptable and frequently employed in the industry. In the compound model, the random variable $N$ can be described with several distributions, including binomial ($Bin(n, p)$), Poisson ($Poi(\lambda)$), and mixed Poisson ($Poi(\Lambda)$). The Poisson distribution is particularly well-known, having given rise to the widely used compound Poisson model. In a mixed Poisson model, the parameter $\Lambda$ is a random variable itself. The compound mixed Poisson model, which uses a gamma distribution for $\Lambda$, is widely recognized. The moment-generating function shows that this results in a negative binomial distribution. Almost any continuous random variable can be used to distribute the claim amount ($Y$). A combination of these distributions is frequently believed to better depict the complexities of real-world data.

After choosing the two random variables for the compound model, the following step is to analyze the model. The Monte Carlo methodology is the most basic and extensively used tool for this research, and it is especially useful for computational simulations in the insurance industry.

Generalized Linear Models (GLMs) are commonly used in the insurance sector to pick acceptable distributions and parameters for random variables $Y$ and $N$, respectively. GLMs offer a strong framework for modeling and evaluating data, making them a popular tool for this purpose.

In practice, GLMs aid in determining the best-fitting distributions and estimating parameters by linking response variables to explanatory variables. This strategy enables insurers to adjust their models to correctly reflect the peculiarities of their data, hence improving their risk assessment and pricing strategies.

For future reference, it's important to define the following concepts:

- **Quota-Share Reinsurance**: This is a type of reinsurance arrangement where the insurer and reinsurer share premiums and losses according to a fixed percentage. For example, if a quota-share agreement specifies that the reinsurer covers 10% of all losses, the reinsurer will pay 10% of the claims and receive 10% of the premiums.
- **Excess-of-Loss Treaty**: This reinsurance agreement provides coverage for losses that exceed a specified amount. The insurer pays the reinsurer for losses above this threshold. For example, an excess-of-loss treaty with a cap of 1000 per event means that the reinsurer will cover losses beyond 1000, while the insurer retains the first 1000 of each claim.

## III. THE METHODOLOGY

To effectively present information, the methodology will be broken into five subsections where we will explore the reasons and arguments that support our conclusions.

### A. Data Cleaning

Even though there are no NA values in the database, cleaning requires certain adjustments. First, dates must be properly formatted. Additionally, the 'power' variable should be turned into a category value to facilitate comparisons, such as the power of a motorcycle vs that of a car. The portfolio will also be divided into four years: for example, 'year 2015' reflects those who purchased the policy before the end of 2015, 'year 2016' by the end of 2016, and so on.

After data cleansing is completed, the attention will shift to modeling a portfolio of individuals with comparable risk profiles. Due to predicted technological and security developments, the factors examined will be 'Type risk', 'Classification', 'Area', and 'Year'. Other characteristics, such as years of experience or the amount of claims incurred to date, while useful, will be ignored because separating them in a pool of contracts is difficult. All of them are treated as dummy variables.

### B. GLM for premiums and $N$

Because it is critical for the simulation to understand how premiums fluctuate with different client attributes, a Generalized Linear Model (GLM) is used to calculate how the premium changes. Once completed, the procedure is repeated for N. The variables stay unchanged, however in this example, a Poisson GLM is used first, followed by a Negative Binomial GLM. In both cases, because they are the most often used random variables, the best model will be chosen based on graphic visualization. The challenge with the Negative Binomial distribution is

that it has two parameters, $r$ and $p$. In a Generalized Linear Model (GLM), we need to fix the parameter $r$. To address this, the method of moments is employed to determine the optimal value for $r$.

$$r = \frac{E(X)^2}{V(X) - E(X)} \tag{3.1}$$

Once this value is established, it is used consistently throughout the entire computation.

### C. Mixture model for $Y$

Fitting the optimal distribution to Y is the most difficult aspect of the project. As is commonly seen in real-life situations and in this instance, obtaining a perfect match is frequently impossible due to an excess of hidden information. Using bootstrap techniques is one such remedy. Nonetheless, it is widely acknowledged in the insurance sector that although historical data can be utilized to fit a distribution, bootstrapping is not permitted to produce new data from it. As a result, the strategy that was selected is to divide Y into three different categories: the total amount of claims for cases where there is only one claim per year, the total amount for two claims, and the total amount for three or more claims.
Since Y should indicate the amount for a single claim rather than numerous claims, this technique does not reflect real-life settings, where the distribution of Y should be consistent regardless of the number of claims. A mixture model has been considered since graphical analysis indicates that the distributions do not fit any standard random variables.

### D. Simulation

The idea behind it is straightforward: the GLM model is used to estimate the expected value based on user-selected attributes. Once the prediction and its confidence intervals are known, it's possible to find the value of the parameter and its condifence interval of the random variable through the inverse of the link function.

We must first specify the model that was selected for a specific portfolio.

$$S_1 = \sum_{0}^{N(t)} Y^* \tag{3.2}$$

$$Y^* \sim \begin{cases} Y_1 & \text{if } N = 1 \\ Y_2 & \text{if } N = 2 \\ Y_3 & \text{if } N \geq 3 \end{cases}$$

Next, we treat the entire portfolio as the total of the individual projections for each member of the group (chosen by the user, variable $n$) so that we can consider a group of people who share the same traits.

$$S = \sum_{1}^{n} S_1 \tag{3.3}$$

Many data and statistics can be computed after the simulation is finished. More specifically, there are the following choices:

#### *1. Choosing Reinsurance Strategies*

There are various reinsurance strategies to choose from.

- Excess-of-Loss Treaty: You have the option to purchase one and set the maximum amount for every policy in the pool. In the event that a limit of 0 is selected, no excess-of-loss reinsurance is acquired.
- Quota-Share Reinsurance: This alternative is also available to you. A 0 value indicates that no quota-share reinsurance is purchased.

Using these choices, you can calculate various outcomes, namely:

- The premium that the insurance company should pay to the reinsurer based on the selected reinsurance options.
- If no excess-of-loss reinsurance is chosen and a quota-share value of 100 is selected (meaning the reinsurer takes on all the risk), you can calculate the average net premium that all the policy-holders together should pay to the insurance company (so the premium that should be paid to the reinsurance is the one the insurance should ask the whole pool).

To summarize, the reinsurance options are:

- Only excess-of-loss treaties
- Only quota-share reinsurance
- None of the above
- A combination of excess-of-loss treaties and quota-share reinsurance, namely first the excess-of-loss treaties is bought and then the remaining risk is covered by quota-share reinsurance

#### *2. Value at risk*

Once the simulation is complete, the portfolio values are sorted. The user can then select the corresponding percentile (value at risk) based on these sorted values.

#### 3. other statistics

Additional important characteristics can be computed using the following statistics:

- Variance: Measures the variability of the portfolio values.
- Amount of Premiums Received: Total premiums collected.
- Amount of Claims to Be Paid: Expected value of the claims.

It's possible to use the variance and the amount of claims to be paid (which represents the expected value of the portfolio) to calculate the coefficient of variation. A lower coefficient indicates a more diversified portfolio.

$$CoV = \frac{\sigma}{\mu} \tag{3.4}$$

Additionally, the solvency ratio can be computed. This ratio is equivalent to the combined ratio, which is the ratio of the amount of claims to be paid over the premiums received.

$$CR = \frac{claims\ to\ be\ paid}{total\ premia} \tag{3.5}$$

#### 4. Ruin Probability

The most crucial portion of the work is the calculation of the ruin probability, which is based on the following fundamental presumptions:

- Distribution of Claims: In the event for example that we are aware of two claims, $Y$ follows $Y_2$. The total sum for these two claims is generated, but it's important to figure out each claim's individual value in order to calculate the ruin likelihood. It is assumed that the ratio used to divide the total of two claims into two separate claims is Uniform (0,1). In the event that there are three claims, the entire sum is divided into each claim using the same methodology, and so on.
- Claims Allocation Over Days: We have an array representing the amount of money each single claim is worth. We assume that each claim can occur with equal probability on any day of the year. With this assumption, we split the total amount of claims to be paid evenly across the 365 days of the year.
- Premiums Allocation Over Days: Similarly, premiums are generated and are expected to be collected equally throughout the year.

To compute the ruin probability, then, We start with an initial amount of money set aside. Each day, we add the premiums received and subtract the claims to be paid for that day. If the resulting value is negative, it indicates that ruin has occurred. Otherwise, the remaining amount becomes the starting point for the next day, continuing this process until the end of the year.

### E. The user interface

Since the project's objective is to make this portion available to insurance clerks while maintaining the confidentiality of the study, all simulation features are easily accessible through a graphical user interface (GUI). Although it is more user-friendly than interacting directly with the console, the interface is still somewhat simple. In addition, a number of graphs are shown while the window is active to help with understanding and visualization.

## IV. IMPLEMENTATION OF THE METHOD

### A. Data Cleaning

The `pandas` library is used for handling dataframes, with data cleaning primarily facilitated by the `classify_values` function, which splits the vehicle's power into three categories.

### B. GLM for premiums and N

For fitting the Generalized Linear Model (GLM), the `statsmodels` library is employed for both the random variables and the GLM itself. After fitting the GLM, the `plot_real_and_simulated_histogram_kde()` function is used to compare real data and simulated values for specific database entries. Visualization is done using the `seaborn` and `matplotlib.pyplot` libraries. The characteristics of the GLM are saved to a `.pkl` file using the `pickle` library.

The `calculate_mean_and_variance()` method, implemented with `numpy`, computes variance and expected values. After fitting the Poisson GLM, the `plot_relative_frequencies()` function checks real data against the Poisson PDF. The `plot_relative_frequencies_comparison()` function then plots four different scenarios together, utilizing `pyplot`. For the Negative Binomial model, `r_definitive` is selected using the method of moments, with results saved via `pickle`. Two key functions are defined: `plot_relative_frequencies_nbinom` and `transform_array()`—the latter concatenates claim numbers higher than 7 before plotting. Additionally, the `chi_squared_test_negative_binomial()` function performs a Chi-square test to assess whether real data

with certain characteristics follow a Negative Binomial distribution with parameters $(r, p)$, where $r$ and $p$ varies within an interval. This analysis is conducted within a for loop, allowing for different year-by-year evaluations.

### C. Mixture model for Y

Fitting the model is an iterative process due to the data's complexity and hidden information. The optimal way to proceed is to visually determine the appropriate time splits and then select the best-fitting random variable based on personal preference (e.g., underestimation, overestimation). This is done using the `generate_and_plot_synthetic_data()` function, which calls the previously defined `mixture_fitting()` function. All plots are generated using `pyplot`.

### D. Simulation

The `generate_synthetic_data()` function generates a mixture with specified characteristics. This method is called within `generate_sample()`, which produces the number of claims for each pool. Similarly, `generate_premia()` serves the same purpose but focuses on premiums. Both functions rely on `scipy.stats` and `numpy`. The `plot_histogram()` and `plot_mean_excess()` functions use `pyplot`, while `seaborn_kde_plot_pdf()` uses `seaborn` for visualization. The `get_values()` function subdivides claims (which are individually represented) into days of the year.

### E. GUI

For the GUI interface, the package used is `tkinter`, specifically `ttk` and `tk`. This widely used package is utilized within the `first_page()` function to build the user interface.

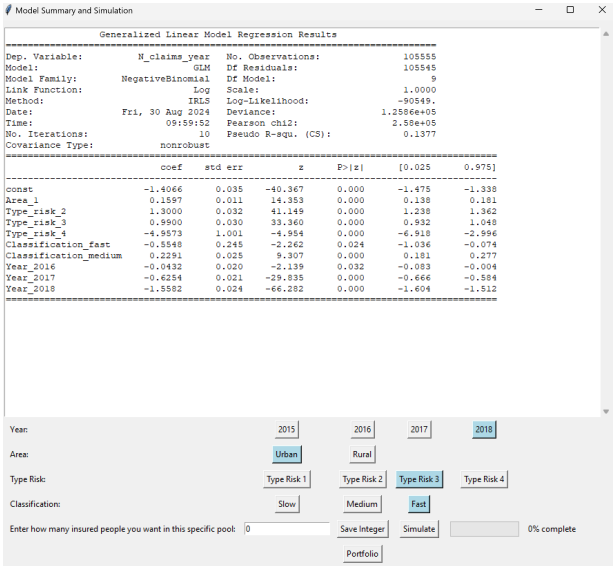


FIG. 1. Interface

- Year needs to be chosen according to which year trend of frequency of claims the user wants to follow
- Area represents if the policy-holder lives in urban ($> 30.000$ inhabitants) or rural
- Type risk is the vehicle of the policy-holder: 1 for motorbikes, 2 for vans, 3 for passenger cars, and 4 for agricultural vehicles
- Classification represents the power of the vehicle

## V. MAINTAIN AND IMPROVE THE CODE

The code could be easily updated with additional features that might require saving more information about the policyholders. First, to achieve a better distribution of $Y$, instead of saving the total yearly amount of claims (in monetary terms), it would be more effective to record the value of each individual claim. This approach would allow for a more precise determination of its distribution. Also having information about deductible would be amazing.

Additionally, the current subdivision into days assumes that claims are equally distributed over time. By also recording when each claim occurred, we could better assess whether this assumption is valid.

Another important aspect to address is the complexity of the compound model. Currently, a double loop is required, resulting in a complexity of $N^2$. Although this is manageable for small pool sizes, it becomes a problem as the number of individuals increases. To mitigate this, variance-reduction techniques can be employed, as the literature suggests these can reduce complexity. For instance, using the antithetic variates method can halve the time spent in the loop. Specifically, instead of generating $n$ random variables, we generate $n/2$ uniform

random variables, take their complements, and then use these results to generate the desired random variables by using the quantile function.

The codebases of the two global project components—the analytic model and the insurance employee's platform—are updated in their respective GitHub public repositories, enabling academics to efficiently maintain the project's code and disseminate findings to other eager developers. To make it easier for outside parties to utilize, every line of code has extensive comments.

Replicability and adaptability are essential to accomplishing the worldwide goal. As a result, new information regarding insurance policyholders needs to be incorporated into the project's code. namely, the analysis file could be updated with new important information every day.

With the inclusion of additional information, new variables may be added to enhance the analysis, necessitating updates to the analysis file. The GUI should remain largely unchanged. Improving the project's documentation should be a key priority to support this effort and guide external contributors.

## VI. RESULTS

The distribution of the claims follows a Normal distribution. Consequently, a Generalized Linear Model (GLM) with a normal distribution was applied (canonical link), and the results are as follows:

| Variable | Estimate | CI Lower | CI Upper |
|---|---|---|---|
| Intercept | 118.9392 | 114.287 | 123.592 |
| Area_1 | 22.3328 | 20.601 | 24.065 |
| Year_2016 | 8.2977 | 4.328 | 12.267 |
| Year_2017 | 9.1919 | 5.239 | 13.145 |
| Year_2018 | 3.8136 | -0.127 | 7.754 |
| Type_risk_2 | 180.7950 | 177.286 | 184.304 |
| Type_risk_3 | 201.2230 | 198.337 | 204.109 |
| Type_risk_4 | -106.5829 | -115.812 | -97.353 |
| Classification_fast | 189.8486 | 170.541 | 209.156 |
| Classification_medium | 85.3386 | 81.330 | 89.347 |

TABLE I. Regression coefficients with 95% confidence intervals.

On the other hand, a similar analysis was performed using the Negative Binomial distribution (canonical link). The characteristics of this analysis are detailed below:

| Variable | Estimate | CI Lower | CI Upper |
|---|---|---|---|
| const | -1.4066 | -1.475 | -1.338 |
| Area_1 | 0.1597 | 0.138 | 0.181 |
| Type_risk_2 | 1.3000 | 1.238 | 1.362 |
| Type_risk_3 | 0.9900 | 0.932 | 1.048 |
| Type_risk_4 | -4.9573 | -6.918 | -2.996 |
| Classification_fast | -0.5548 | -1.036 | -0.074 |
| Classification_medium | 0.2291 | 0.181 | 0.277 |
| Year_2016 | -0.0432 | -0.083 | -0.004 |
| Year_2017 | -0.6254 | -0.666 | -0.584 |
| Year_2018 | -1.5582 | -1.604 | -1.512 |

TABLE II. Regression coefficients with 95% confidence intervals.

As it's possible to see below, the fitting is pretty good, since all the proportions are equally well described.

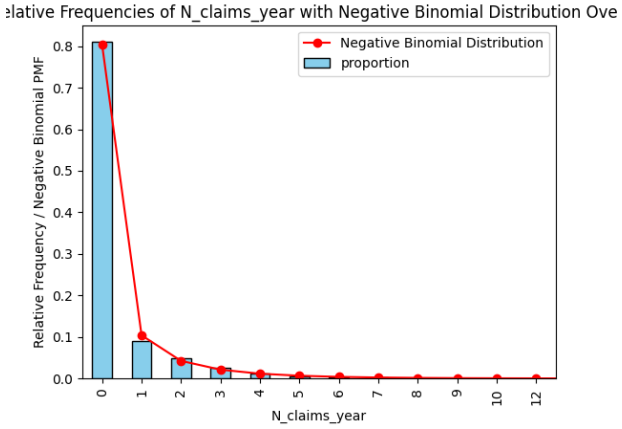


FIG. 2. First example

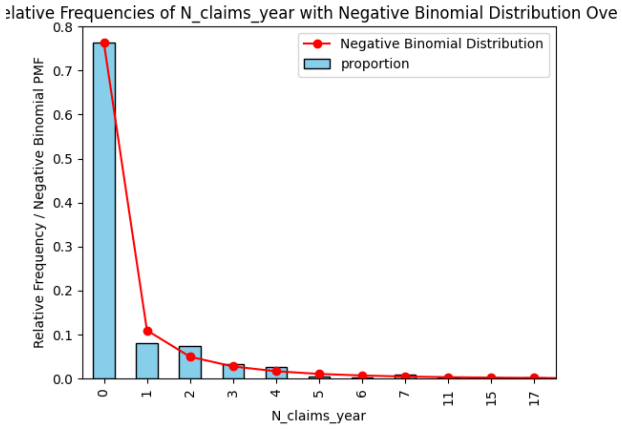


FIG. 3. Second example

The only graphical problem found is that most of the time the probability of $N = 0$ is underestimated, while the probability of $N = 1$ is overestimated.

Consider a mixture model consisting of three different random variables, $X_1$, $X_2$, and $X_3$. The overall probability density function (PDF) $f_X(x)$ of the mixture distribution is given by:

$$f_X(x) = w_1 f_{X_1}(x;\theta_1) + w_2 f_{X_2}(x;\theta_2) + w_3 f_{X_3}(x;\theta_3)$$

where:

- $f_{X_1}(x;\theta_1)$ is the PDF of the first random variable $X_1$ with parameters $\theta_1$.
- $f_{X_2}(x;\theta_2)$ is the PDF of the second random variable $X_2$ with parameters $\theta_2$.
- $f_{X_3}(x;\theta_3)$ is the PDF of the third random variable $X_3$ with parameters $\theta_3$.
- $w_1$, $w_2$, and $w_3$ are the mixture weights, such that $w_1 + w_2 + w_3 = 1$ and $w_i \geq 0$ for $i = 1, 2, 3$.

For the distribution when $N = 1$, the mixture has the following characteristics:

- $X_1$ is an Exponential random variable with rate parameter $\lambda_1 = 56.20156866464339$ shifted by $\text{shift}_1 = 40.05$.
- $X_2$ is a Gamma random variable with shape parameter $\alpha_2 = 1.44$ and scale parameter $\beta_2 = 152.78$ shifted by $\text{shift}_2 = 247.91$.
- $X_3$ is a Log-Normal random variable with shape parameter $\sigma_3 = 1.82$ and location parameter $\mu_3 = 113.5$ shifted by $\text{shift}_3 = 849.5$.

The mixture model can be expressed as:

$$X = \begin{cases} X_1 + \text{shift}_1 & \text{with probability } w_1 = 0.55 \\ X_2 + \text{shift}_2 & \text{with probability } w_2 = 0.22 \\ X_3 + \text{shift}_3 & \text{with probability } w_3 = 0.26 \end{cases}$$

For the distribution when $N = 2$, the mixture has the following characteristics:

- $X_1$ is a Gamma random variable with shape parameter $\alpha_1 = 1.08$ and scale parameter $\beta_1 = 176.18$ shifted by $\text{shift}_1 = 40$.
- $X_2$ is a Log-Normal random variable with shape parameter $\sigma_2 = 0.44$ and scale parameter $\phi_2 = 376$ shifted by $\text{shift}_2 = 584$.
- $X_3$ is a Logistic random variable with location parameter $\mu_3 = 3831$ shifted by $\text{shift}_3 = 2153$.

The mixture model can be expressed as:

$$X = \begin{cases} X_1 + \text{shift}_1 & \text{with probability } w_1 = 0.64 \\ X_2 + \text{shift}_2 & \text{with probability } w_2 = 0.23 \\ X_3 + \text{shift}_3 & \text{with probability } w_3 = 0.13 \end{cases}$$

For the distribution when $N \geq 2$, instead, the mixture has the following characteristics:

- $X_1$ is an Exponential random variable with rate parameter $\lambda_1 = 225$ shifted by $\text{shift}_1 = 40$.
- $X_2$ is a Gamma random variable with shape parameter $\alpha_2 = 1.35$ and scale parameter $\beta_2 = 284$ shifted by $\text{shift}_2 = 800$.
- $X_3$ is a Logistic random variable with location parameter $\mu_3 = 4467$ and scale parameter $s_3 = 2543$.

The mixture model can be expressed as:

$$X = \begin{cases} X_1 + \text{shift}_1 & \text{with probability } w_1 = 0.61 \\ X_2 + \text{shift}_2 & \text{with probability } w_2 = 0.27 \\ X_3 & \text{with probability } w_3 = 0.12 \end{cases}$$

Let's analyze a pool with the following characteristics: 1000 policyholders with fast cars living in an urban area, following the 2018 trend of car accidents. We specifically adjust for a higher weight on claims where $N \geq 1$, by modifying the parameters $\alpha$ and $p$.

```
Amount of people in the pool: 1000

Value of p: 0.61

Value of alpha: 0.36

Percentile analyzed for VaR: 98.0

VaR value: 164284.32703103716

Variance: 597219309.9538255

Starting amount of money of the pool: 0

Ruin probability: 1.0

Amount of premia recevied: 536187.8885460534

Amount of claims to be paid: 110849.97858484495

Premia reinsurance: 0.0

Maximum you want to pay for each claim: inf

Percentage that you, as insurance, will pay: 100
```

FIG. 4. Case with no reserve

In this scenario, the total amount of premiums received is approximately four times the theoretical pure premium required to cover the entire pool (namely the amount of money the insurance should pay). Despite this, the ruin probability remains at 1. This outcome occurs because no funds are set aside for potential claims.

```
Value of p: 0.61

Value of alpha: 0.36

Percentile analyzed for VaR: 98.0

VaR value: 163449.53827854854

Variance: 478215660.44358927

Starting amount of money of the pool: 10000

Ruin probability: 0.189

Amount of premia recevied: 536141.952146705

Amount of claims to be paid: 110151.7428772464

Premia reinsurance: 0.0

Maximum you want to pay for each claim: inf

Percentage that you, as insurance, will pay: 100
```

FIG. 5. Case with reserve

But if we set aside reserves, say 10,000, the ruin probability drops, though it takes some time to zero. This indicates that the insurance must set aside money at all times to avoid going bankrupt. It is then up to insurance regulators to determine the acceptable maximum probability of ruin that they are willing to manage.

Another significant insight can be derived from the mean excess function graph. It initially decreases and then increases again. This behavior is indicative of substantial weight in the tail of the distribution (after 200.000), a fact that is clearly observed in the corresponding probability density function (PDF) graph.

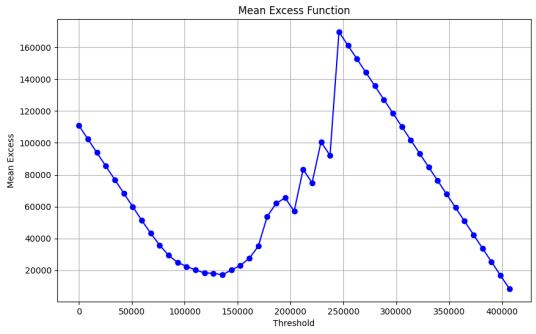


FIG. 6. Mean excess function of the pool

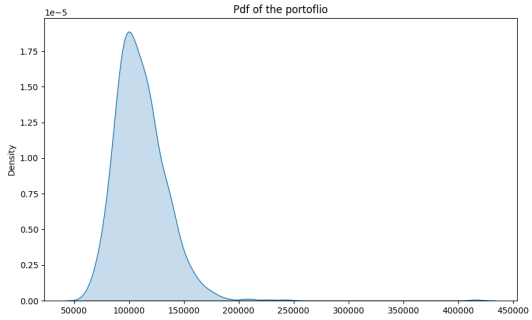


FIG. 7. PDF

Let's now examine how the statistics are affected after implementing an excess-of-loss treaty with a maximum coverage of 1000 per event, combined with a quota-share reinsurance agreement where the reinsurer covers 10% of all events. It is expected that the mean excess function would be almost entirely decreasing, or at least show a significant reduction, due to the cap imposed by the excess-of-loss treaty. This cap prevents payments beyond a certain threshold, which affects the tail behavior of the distribution.

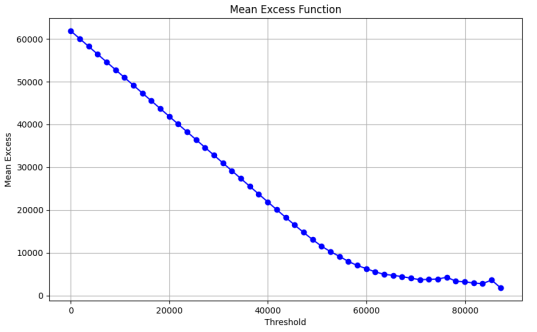


FIG. 8. Mean excess function

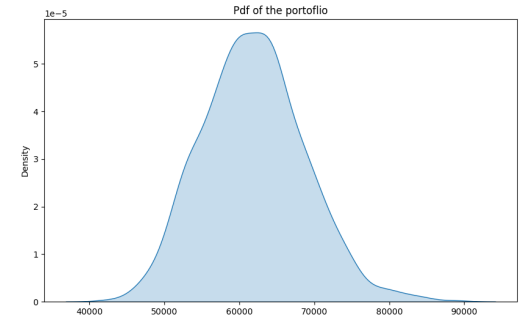


FIG. 9. PDF

Consequently, we observe a reduction in variance and a smaller value at risk. Additionally, with the same amount of money set aside, the ruin probability now drops to zero.

```
Amount of people in the pool: 1000

Value of p: 0.61

Value of alpha: 0.36

Percentile analyzed for VaR: 98.0

VaR value: 76772.80484410877

Variance: 45381652.36264235

Starting amount of money of the pool: 10000

Ruin probability: 0.0

Amount of premia recevied: 536173.714196331

Amount of claims to be paid: 62081.02163450733

Premia reinsurance: 48085.16382408496

Maximum you want to pay for each claim: 800

Percentage that you, as insurance, will pay: 90.0
```

FIG. 10. Focus on ruin probability

The key takeaway is that it is up to the insurance company to select the most suitable arrangement for each pool. It is recommended to start by using the code to generate a baseline scenario where no money is set aside and no reinsurance is in place. Subsequently, you can simulate various scenarios, such as incorporating reinsurance options or setting aside reserves, to determine their impact on the portfolio. When the analysis of a specific pool is done, the user needs to decide which of the possibilities is the best one for them. Once all decisions have been made for each pool singularly, the next step is to aggregate these pools into a portfolio and analyze the overall impact.

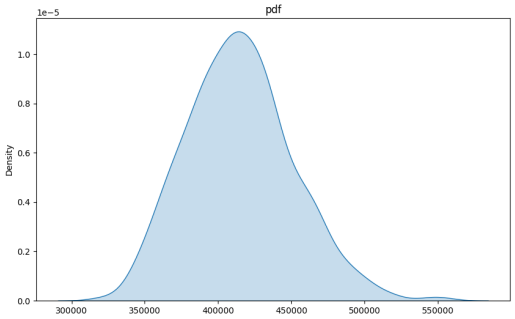


FIG. 11. PDF

This approach allows you to evaluate how the combined portfolio performs under various conditions and strategies, as we can see below and above.

**Pool 4**

```
Characteristics of the pool: This is the whole portfolio with

Total people in the pool: 2900

Percentile analyzed for VaR: 99.5

VaR value: 520309.32141825114

Starting amount of money of the pool: 0

Ruin probability: 1.0

Amount of claims to be paid: 414750.908578029

Amount of premia received: 1125592.7467169547

Variance: 1334641742.5774317

Total premia to be paid to reinsurnace: 127931.34017420953
```

FIG. 12. The whole portfolio

## VII. CONCLUSION

To conclude, this model is a preliminary framework that can be significantly improved as more data becomes available. If applied in the insurance industry, the user interface could be highly valuable for analyzing data, extracting insights, and supporting decision-making processes.

As a university project, it serves as a tool to demonstrate various theoretical concepts. For instance, it can illustrate how the coefficient of variation decreases as the pool size increases and becomes more diverse, showing the benefits of diversification. Additionally, it can help explain why premiums might appear high, even if the insurance company seems to be making substantial profits. The ruin probability analysis highlights the importance of setting aside adequate reserves and managing funds appropriately.

## VIII. REFERENCES AND HELPER-TOOLS


[1] ChatGPT, *To improve the code and write the report better*
[2] Quillbot, *To write the record better*
[3] GitHub, *To save the project*
[4] Course of Loss Models, *Master's in actuarial science*
[5] Course of Risk Theory, *Master's in actuarial science*
[6] Course of Probability and Stochastic processes, *Master's in actuarial science*
[7] Course of rate-making and claim reserving, *Master's in actuarial science*
[8] The essential guide to reinsurance, *Swiss RE*