NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão da Informação

Master Program in Statistics and Information Management

An optimal deductible evaluation under an Excess of Loss Reinsurance Treaty for an Automobile Insurance portfolio

Henrique da Cunha Alcaide

Dissertation proposal submitted in fulfillment of the requirements for the degree of

Master of Science in Statistcs and Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão da Informação

Universidade Nova de Lisboa

NOVA INFORMATION MANAGEMENT SCHOOL

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

AN OPTIMAL DEDUCTIBLE EVALUATION UNDER AN EXCESS OF LOSS REINSURANCE TREATY FOR AN AUTOMOBILE INSURANCE PORTFOLIO

by

Henrique da Cunha Alcaide

Dissertation proposal presented as requirement for obtaining the Master's degree in Information Management, with a specialization in Analysis and Risk Management

Advisor: Maria de Lourdes Belchior Afonso

May 2022

Abstract

Reinsurance has become a widely used solution for insurance companies to protect themselves from risks.

For this study it is considered a simulated portfolio consisted by policies that are covered by an own damage vehicle insurance, in which an Excess of Loss reinsurance will be applied.

The main concern is to build a practical tool that allows to determine the insurer's optimal retention level of risk.

Several techniques from non-life actuarial insurance, data science and mathematics are used to determine the optimal reinsurance deductible:

Data simulation, to create a portfolio; Pricing, while applying Generalized Linear Models to a simulated automobile portfolio; Reinsurance, in order to reduce the insurer's risk; Decision Theory, to obtain an optimal deductible for reinsurance, and the collective risk model to have a theoretical model to compare with the results.

For this type of portfolios, this approach provides a baseline to decide the insurer's retention limit in case of signing an Excess of Loss per risk treaty.

All the work in this study will be achieved through the R tool (Version 4.1.0).

Keywords: Data Simulation, Pricing, Reinsurance, Decision Theory.

Resumo

O Resseguro tornou-se numa solução amplamente utilizada pelas companhias de seguros para se protegerem contra os riscos tomados.

Para este estudo é considerada uma carteira simulada constituida por apólices cobertas pelo seguro automóvel de danos próprios, ao qual será aplicado um resseguro de Excesso de Perdas (Excess of Loss).

A principal preocupação é construir uma ferramenta prática que permita determinar o nível óptimo de retenção do risco para uma seguradora.

Para determinar o dedutível ótimo de resseguro são utilizadas várias técnicas dos ramos de atuariado Não-Vida, Ciência de Dados e Matemática :

Simulação de dados, para criar uma carteira. Aplicação de modelos Lineares Generalizados para obtenção de uma tarifa para uma carteira automóvel simulada. Resseguro, para reduzir o risco da seguradora. A teoria de decisão, para obter o dedutível ótimo de resseguro e o modelo de risco colectivo, para a comparação dos resultados com um modelo teórico.

Para este tipo de carteira, a abordagem usada permitirá às seguradoras determinarem o limite de retenção a ser tomado em caso de aplicação de um resseguro Excesso de Perdas.

Todo o trabalho desenvolvido neste estudo será efetuado através da ferramenta R (Versão 4.1.0).

Palavras-chave: Simulação de Dados, Tarifação, Resseguro, Teoria de Decisão.

Contents

List of Figures

List of Tables

1 | Introduction

Exploitation and identification of risks are increasingly crucial for businesses. Companies that operate their framework under financial uncertainty are always looking for reasonable solutions to maintain a healthy situation.

Insurance companies are permanently exposed to risk, and there is the need to develop efficient and precise mechanisms that can provide them with the required protection.

Reinsurance is increasing its importance as a risk management tool in this dimension. As a result, insurers are betting on it mainly due to its structured and organized workflow.

In general, the concept of reinsurance is not very recognized outside of the insurance sector. However, the concept is self-explanatory, as it can be easily described with a single statement: "Reinsurance is an insurance of an insurance contract".

With this in mind, the reader can quickly feel that reinsurance is organized in layers. That is why companies consider it one of the best ways to share risk.

Fundamentally, reinsurance in many situations can be the critical point to avoiding a ruinous situation in a company's short, medium, or long term. It can help an insurance company stay solvent in a few words.

1.1 Problem Identification

As we are working with a contract between two parties, it is reasonable to infer that a reinsurance owner will share risk and the profit that would be potentially obtained from the portfolio. Therefore, it is a contractual consequence that results from risk-sharing.

Facing this and taking into account that an insurance company is obliged to fulfil the capital requirements, it is possible to identify a critical issue for the insurance company:

There is the need to find a proper balance between the risk taken by the reinsurance owner and the amount lost when sharing risk.

If we pretend to be more rigorous, this problem is underlying the requirement of precision a fixed amount (retention limit) of risk that an insurance company is willing to keep/share.

Mathematically speaking, this study may remind a problem of optimization, regularly identified as an objective function problem. This means that the scientist intends to maximize and/or minimize a group of constraints over feasible solutions to the problem.

1.2 Objectives

Through the following chapters, we are intended to solve the main problem that can be formulated as: *"What is the best retention limit to be set by an insurance company when signing up a reinsurance contract?"*.

We are focused on simulating a particular reinsurance contract in the automobile sector, the Excess of Loss reinsurance contract.

This agreement will be applied under a facultative Non-Life Insurance product, an Own Damage Car Insurance.

The main goal of this project is to construct a practical tool that allows this particular type of portfolio to determine the optimal insurer's level of risk that should be retained in case of signing up for a reinsurance contract.

By considering a practical workflow, the reader will be able to quantify the optimal amount of risk that an insurer should be responsible for.

2 | State of the Art

Actuarial mathematics originated towards the end of the 17^{th} century. Then, for the first time, E. Halley produced a mortality table that allowed the calculation of annuity values (Bühlmann (1970)). During the 1930s, there was another decisive advance in actuarial mathematics with advances in probability theory, statistics, and economics. Since then, Bühlmann (1970), Bowers et al. (1997) and more recently Tse (2009) and Wuthrich (2019), among many others, have dedicated several pages to non-life actuarial mathematics.

Regarding the insurance tariff construction, much literature is available. More recently, in Huang and Meng (2019), an automobile insurance classification ratemaking was developed, recurring to Generalized Linear Models.

As a pioneer in the reinsurance subject, Finetti (1940) presents a study about insurance risks. The primary approach was dedicated to the mean-variance efficient frontiers for a constraint set.

More recently, these studies were improved by Kalluszka, M. Kaluszka (2004), which presented optimal reinsurance by capturing the reinsure's risk recurring to mean-variance premium principles.

Nowadays, optimizing reinsurance tools in non-life insurance is a common theme for an actuary.

The identification of an optimal reinsurance treaty and the derivation of an optimal retention cedent limit, also called a deductible, are considered an advantage to the insurer. This can make the company resolve one of its main concerns: the fulfilment of the capital requirements imposed by *Solvency II* (see Delegated (2015)).

The primary concern in an insurance company is finding the optimal reinsurance to a specific line of business, for which there is a vast literature, especially regarding the two topics mentioned above. However, the methodologies for determining an optimal deductible are discussed at a secondary level, both in published studies and books about reinsurance.

Such the methodology presented by Moro and Rita (2016), a similar analysis was made during the research made about this topic. By using Google Scholar as a searching tool for this topic - "an optimal reinsurance" – it is apparent that the number of publications made on this topic grew up over the past decades. The most significant growth has taken place in the new millennium, especially in the last twenty years,

where the number of publications has doubled every five years.

Centeno and Simões (2009) present a complete summary with relevant results about this generic topic. Similar to a literature overview of this theme, the reports reviewed in this paper are based on aggregated claims portfolio cases. Only a few authors are focusing on the individual claim case (see Dickson and Waters (2006) and Centeno and Guerra (2010)).

Remaining in Centeno and Simões (2009), the authors state some critical words about the nonexistence of an ideal and generic reinsurance treaty for all the cases: *"... Indeed, there is no ideal type of reinsurance applicable to all the cases. Each kind of reinsurance offers protection against only certain factors that affect the claim distribution"*.

Refining the bibliographic search for "optimal retention level", the total number of results sharply decreases, confirming the initial idea of this literature review that the evaluation of reinsurance optimal retention limits is a subsequent study after defining an optimal reinsurance treaty to the insured risks.

Reviewing the optimal deductible point, most of the results focus on setting a single constraint to achieve the target, making the multi-setting of constraints more unusual. The last approach can be addressed, for instance, in Karageyik and Dickson (2016) and in Bulut Karageyik and Şahin (2017)

Following these considerations, there are lots of findings based on the usage of different criteria that might be useful in this study.

A widely used method is the maximization of the adjustment coefficient. According to Centeno and Simões (2009), this target is a proxy for minimizing the probability of ruin related to the time variable. Whether setting time as a continuous variable to an infinite horizon scheme, as in Straub (1988) or considering time as a continuous variable to a finite horizon, as in Centeno et al. (1997), the deductible output is distinct for the reinsurance contract applied (see De Vylder and Goovaerts (1988)). The results of the retention limit also differ when time is assumed as a discrete variable.

Another retention limit optimization technique is the variance minimization of an insurance portfolio. The insurers prefer this methodology to consider expected profits against its variance stability. For the first time in De Finetti (1940), this approach is presented for a non-life portfolio of independent sub-portfolios with a fixed expected profit. This case considers two types of contracts: an excess of loss and a quota-share treaty.

Meeting with the insurer's mission of staying solvent, researchers usually consider capital requirements as a possible criterion to determine the optimal reinsurance. For an insurance company, these measures, such as Value at Risk (VaR), Credit Value at Risk ($CVaR$), and the reserve amount, are based on extreme cases. In Cai and Tan (2007), it is applied an application of a stop-loss reinsurance treaty. The authors show that the optimal deductible obtained either using a (VaR) or ($CVaR$) is the same.

Regarding more theoretical results, a utility function study is also frequent. However, the application to the industry is not accessible due to the assumption of conditions that are hard to interpret.

As mentioned before, most published papers are developed considering the aggregate claim case.

Kaas et al. (2008) present a simple technique for the total claim cost determination, the Central Limit Theorem (*CLT*) application. This approach is based on the Normal approximation, and it involves the knowledge of two moments from the distribution: the expected value of the total claim cost and its variance.

Improving the requirement for three moments requires more rigorous methods as alternatives. For example, in Mourik (2018), the Normal Power approximation is used, and in Dickson and Waters (1996), the Translated Gamma approximation.

Another alternative to find the aggregate claim distribution is the *Panjer* recursive method, which calculates the total claim cost by recursion. In this approach, the individual claim random variable needs to be a discrete distribution that must be carefully addressed, for instance, by applying a proper Chi-Squared test. This method is explained in detail in Holtsmark (2015).

When reviewing the literature of different methodologies to achieve the same goal, there is an attempt to decide which one is the best. In this case, it is reasonable to infer that it depends on the insurer's goal. In an ideal case, the actuary would try to construct a model that optimizes various constraints simultaneously.

3 Background - Literature Review

3.1 Insurance and Reinsurance - Study Relevance

A Non-Life insurance policy is a contract between the insured entity (the policyholder), who is interested in transferring risk, and the insurer (or the cedent), who underwrites this uncertainty.

In this type of contract, the insured party agrees to pay an *a priori* amount of money, named as premium, in exchange for benefits explicit in the contract. The following figure intends to present the workflow of an insurance contract.

The following figure intends to present the workflow of an insurance contract.

Figure 3.1: Flow Chart of an insurance contract. **Source:** Authors.

Briefly explaining, that the cedent bears the policyholder's risks by conferring them protection against future events. In return, the policyholder agrees to pay money during the contractual period. On the policyholder's side, this event can be summarized as an exchange of uncertainty for certainty.

On the cedent's side, the reverse happens. Therefore, this exchange needs to be well monitored by the insurer. The decision to carry a specific risk should be made with the utmost rigor, using available tools and experience from the industry.

Facing this and always with the ultimate purpose of complying with the regulatory requirements of a financial institution, one of the main concerns on the insurer's side

is to define an insurable risk.

This definition is not unanimous and varies significantly for different industries. In insurance, this concept is usually defined as a random variable representing the likelihood of an unexpected event occurrence. This definition implies that the risk taken must fulfil the following requirements: Incorporate uncertainty in the future and be completely accidental, measurable, and quantifiable.

Another important topic on the cedent's side is the fulfillment of the capital requirements. With the uprising of *Solvency II* since 2016, see Delegated (2015), companies that are covered by this regime are obliged to fulfil capital measures to guarantee a healthy financial situation.

Ideally, an insurance company should be self-sufficient, meaning that premiums resorting from policyholders should be enough to guarantee the solvency status. Moreover, suppose the company strategy is correctly done. In that case, this amount of money should be sufficient to provide a reserve fund (commonly named as reserve) to the company and pay dividends to the shareholders.

In practical terms, this goal is hard to be reached. Therefore, a possible strategy for companies to meet the imposed requirements is to protect themselves from the eventuality of accepting risks that they cannot support.

Reinsurance provides a simple approach to solve it, or at least partially. It allows transferring the insurer's technical provisions to a reinsurer, who firms the contract under the condition of receiving a premium.

This liability division reduces the insurer's exposure to the variability of the aggregate claims process during a given period. Additionally, this tool can be considered an excellent option to increase the insurer's financial capacity. The risk released becomes available to make profit-generating investments for the insurance company, which indicates the company's viability.

3.2 The Theoretical Continuous Risk Collective Model

An actuary is always eager to find new and accurate ways to explain actual events in mathematical terms.

Focusing on the Non-Life side of an insurance company, a usual problem is understanding the collection risk process of automobile insurance products.

Usually, these types of events are described by probabilistic models. And in this dissertation, we will focus on analyzing the applicability of the collective risk model.

This model has been improved since its genesis, which dates back to the 19^{th} century. Originated by Ernst Filip Lundberg, a Swedish mathematician with several relevant results in risk theory, the model in analysis has been redesigned and improved with several developments made by authors such as Kolmogoroff, Bartlett, Cramér, Doob, Feller, Gnedenko, Khintchine, and many others.

Nowadays, the collective risk model continues to be widely used in the insurance sector, and as explained in Bowers et al. (1997), its formulation can be presented as a parcel of a more holistic process named a surplus process over an extended period t.

The latter stochastic process is used to be recognized in this industry as a model that allows the actuary to evaluate the random fluctuations of the reserves by taking into account the initial surplus of insurance, the premiums collected during the period evaluated and the aggregate claim process.

Let us then start with the formulation of the surplus process at time t. Denoted as $U(t)$, the mathematical equation (as it is done in Bowers et al. (1997)) is given by the following expression:

$$U(t) = u + ct - S(t), t \geq 0 \tag{3.1}$$

where,

- u represents the initial surplus;
- ct stands for the premiums collected until time t at a constant rate $c > 0$ ([1]);
- $S(t)$ stands for the aggregate claim's amount paid until time t (The aggregate claim process).

For this particular study, we will assume that:

- An insurance company starts with a non - negative surplus;
- The premiums grow up at a constant rate;
- The company is still paying claims at time t;
- The time window considered is t equal to one year.

Facing the last item pointed out, the equation (3.1) can be reformulated by "ignoring" the time structure of S as:

$$U = u + c - S \tag{3.2}$$

where,

- u represents the initial surplus;
- c stands for the premiums collected during the year considered at a constant rate $c > 0$;
- S stands for the aggregate claims amount paid during the year considered.

[1] As explained in Bowers et al. (1997), ct is a deterministic model. Meaning that the unique stochastic component is S(t).

3.2.1 The Aggregate Claim Process

The Aggregate Claim process or the collective risk model intends to be focused on the explanation of S, which is part of the equation (3.2) presented in section (3.2). It constitutes itself a particular statistical/probabilistic behavior that can be formulated in mathematical terms as:

$$S = \sum_{i=1}^{N} X_i \tag{3.3}$$

where,

- N corresponds to the claims number process;
- X_i the $i-th$ claims amount with $i = 1, ..., N$;
- X_i are independent and identically distributed (i.i.d) and independent from N.
- X_i are non-negative real number and if X_i=0, there is no claim occurrence.

Quoting Bowers in Bowers et al. (1997), this model assumes *"a random process that generates claims for a portfolio of policies"*, meaning that it considers the entire portfolio as a whole. Consequently, it underlies the case where multiple losses can be derived from a single contract.

From the model description in equation (3.3), while N stands for the Frequency random variable, X stands for the Severity random variable.

Thus, N which can be understood as the claims frequency random variable, and X as the amount of the claims random variable, are the variables of primary interest when exploring S. In a more formal context, they are called exploratory variables.

As stated in the model specifications, N and X are independent, which means that modeling S can be performed by modeling these two processes individually.

3.2.2 The Expected Value and Variance of the Aggregate Claims Amount

In practical terms, modeling S is usually performed by estimating its probabilistic moments. A probabilistic distribution can be uniquely explained with these quantitative measures (see Feller (2008)).

In fact, from the actuary's point of view, more than understanding the exact behavior of S, it would be more interesting to anticipate the future behavior of the amount of the aggregate claims. In a few words, there is the need to determine the expected value of S, which in mathematical terms corresponds to figuring out the first moment of its probabilistic distribution.

Although the expected value of losses is an important metric to unravel the behavior of S, it is not sufficient to define accurately the final target. Therefore, there is a need to estimate other moments of the probabilistic distribution of S.

From the insurer's point of view, determining how the expected value of S varies significantly its global behavior can be an essential metric for this analysis. This objective can be achieved by estimating the second moment of the probabilistic distribution of S.

As said before, the aggregate losses would be uniquely modeled by estimating all their moments.

However, due to the high increase in computational complexity, it is pretty standard in this industry sector to only base S modulation by resorting to these two moments.

As such, the final target will consist of estimating the expected value and the variance of S, $E[S]$ and $V[S]$, respectively.

Following Bowers et al. (1997), the definition of a $k-th$ element of S can be easily achieved through a particular statistical function called a moment generating function (mgf).

For the particular case of S, let's denote its mgf as φ_S, which can be formulated by using the law of iterated expectations combined with the assumption of X_i being i.i.d:

$$\varphi_S(r) = \varphi_N(ln(\varphi_X))(r), \forall r \in \mathbb{R}^+ \tag{3.4}$$

The determination of the expected value of S, $E[S]$, can be obtained from the first order derivative of the mgf of S at point zero, $\varphi'_S(0)$:

$$E[S] = \varphi'_S(0) = \varphi'_N(ln(\varphi_X))(ln(\varphi_X))'(0) = E[N] \times E[X]. \tag{3.5}$$

Using simple algebra calculations (considering the law of total variance), the variance of the aggregate loss can be expressed at the expense of $E[S]$ and $E[S^2]$ ([2]) as:

$$V[S] = E[S^2] - E[S]^2 = V[X] \times E[N] + E[X]^2 \times V[N]. \tag{3.6}$$

3.3 Estimating $E[S]$

From the business point of view, the meaning of the estimation of $E[S]$ can be seen as an attempt to know the necessary amount of money that an insurer should collect to cover future losses resulting from its clients' claims.

It is possible to translate this target into the need for finding the insurance premium. Knowing the expected value of the losses in the portfolio allows the cedent to anticipate the monetary amount they should charge the client to cover the losses caused.

According to Guerreiro (2016), the estimation of $E[S]$ can be obtained by modeling the Pure Endowment, and quoting this author, *"the premium should be determined*

[2] $E[S^2]$ corresponds to the second order derivative of S at point zero and it can be denoted as $\varphi''_S(0)$.

in accordance with the characteristics of the risk to be insured. Being at the same time remunerative for the insurer and fair and equitable for the insured".

Following the last quoted author, the calculation of $E[S]$ can be formalized by the following expression:

$$PureEndowment(t) = Frequency(t) \times ClaimMeanCost(t) \tag{3.7}$$

As presented above, this concept is also a stochastic process formulated at the expense of two other stochastic variables: the claims frequency and the claims mean cost. Therefore, we will take a deep look into both components in the following sections.

3.3.1 The Claims Frequency

The concept of claims frequency considering a range period of $[0,t)$ is defined by the losses observed during the time of exposure of each policy.

Due to the inclusion of information related to risk exposure covered by the insurer, the claims frequency can be seen as an accurate measure to monitor the number of policies collected.

It is a concept that depends on time, which means that the process can be stochastically defined as:

$$F(t) = \frac{N(t)}{Exp(t)} \tag{3.8}$$

where,

- $N(t)$ corresponds to the number of claims registered in period $[0,t)$;
- $Exp(t)$ corresponds to number of units exposed to risk in a range period of $[0,t)$.

3.3.1.1 The Claims Number Process

Looking particularly into the nominator of the equation (3.8) presented in section (3.3.1), $N(t)$ is also assumed as a stochastic process. Generally, $N(t)$ is considered as a mixed Poisson process that it's obtained from the homogeneous Poisson process, with parameter λ. Additionally, it is assumed that the Poisson parameter is assumed to be a random variable, Λ, varying from zero to infinite, $(0,\infty)$.

In the automobile insurance sector, Λ is commonly defined as following a *Gamma* distribution with $E(\Lambda) = \frac{\alpha}{\beta}, \alpha, \beta > 0$ with a generating moment function defined in general as $\varphi_\lambda(r) = \left(\frac{\beta}{\beta - r}\right)^\alpha, r < \beta$.

In this particular case study, the moment generating function can be expressed based on the previous assumptions as:

$$\varphi_{N(t)}(r) = E[e^{rN(t)}] = \int_0^\infty E[e^{rN(t)}|\Lambda = \lambda] f_{\lambda(\lambda)} d\lambda = \varphi_\lambda(t(e^r - 1)) = \left(\frac{1 - \frac{t}{\beta+t}}{1 - \frac{t}{\beta+t}e^r}\right)^\alpha, r < ln\left(\frac{\beta+t}{t}\right) \tag{3.9}$$

As we can observe, this is the moment generating function of a negative binomial, with parameters α and $p = \frac{t}{\beta+t}$.

For the particular case of t equaling one year ($t = 1$ year), the equation (3.9) can be rewritten as:

$$\varphi_N(r) = \left(\frac{1 - \frac{1}{\beta+1}}{1 - \frac{1}{\beta+1}e^r} \right)^{\alpha}, r < ln(\beta + 1) \tag{3.10}$$

3.3.2 The Claims Mean Cost/Claims Severity

The claims mean cost corresponds to the total amount paid by the insurer during the time of exposure considered. It includes the amount already paid by the insurance company; The amount already paid on claims that have not been closed; The expected claim amount that has already been reported but not fully regularised; The expected amount of money corresponding to future claims that have occurred but at the moment t are not reported.

In this study, we will consider the final loss already paid by the insurance company for the claims to reduce the calculation and simulation complexity. This final consideration originates from the concept of severity, which can be expressed as:

$$Severity(t) = \frac{Ind(t)}{N(t)} \tag{3.11}$$

where,

- Ind(t) corresponds to the total lost expected to be paid by the insurance company for the claims observed;
- $N(t)$ corresponds to claims Frequency.

3.3.3 Individual Records: The impact on Claims Severity Formulation

Working in insurance means constant contact with data. As such, the actuary is entirely dependent on its structure and content.

Facing the possibility of finding different scenarios, the actuary needs to adapt to them and work with the provided data properly. This adaptation can sometimes lead to minor or significant changes in the models.

The severity process is one of these cases. The information available to work may be organized individually. Each row of the data set is seen as an individual policy. In the aggregated format, each row correspondent to a group of policies with the same profile.

In this dissertation, we will follow the first approach. Therefore, there is no need to treat the severity process as presented in (3.11) in this particular case. Briefly explaining that the denominator, which corresponds to the exposure to the risk, "disappears" since it is defined as equal to one:

$$Severity(t) = \frac{Ind(t)}{1} = Ind(t) \tag{3.12}$$

where,

- Ind(t) corresponds to the total amount already paid by the insurance company for the claims observed.

In the end, the calculation of the severity process will rely on the evaluation of each claims with the total of policies in the portfolio.

3.3.3.1 The Claims Amount Process

The insurance claims amount corresponds to the numerator of the equation (3.11) in section (3.3.2). As previously presented in (3.3.2), it corresponds to the total amount already paid by the insurance company for the observed claims.

There are several distributions for modeling this variable. While doing our research, the authors usually use a *Gamma*, a *LogNormal*, a *Pareto*, or a *Weibull* to fit a probabilistic distribution to the claims amount process.

This study will rely on the *Gamma* distribution possibility, which presents positive support and positive *skewness*, which are two critical aspects when simulating claims on which it is intended to apply a reinsurance contract.

Considering the process (3.3), if $X_i \sim Gamma(\alpha, \beta)$ ([3]), the probability density function formulation can be written as:

$$g(x) = \frac{1}{\Gamma(\alpha)} \beta^{\alpha} e^{-\beta x} x^{\alpha-1} I_{x>0}(x); \alpha, \beta > 0 \tag{3.13}$$

The expected value and the variance of this distribution are: $E[X] = \frac{\alpha}{\beta}$ and $V[X] = \frac{\alpha}{\beta^2}$, with $a, b > 0$.

The moment generating function (mgf) of X_i, it is given by:

$$\varphi_{X_i}(r) = \left(1 - \frac{r}{\beta}\right)^{-\alpha_i}, r < \beta. \tag{3.14}$$

[3] For all $i = 1, ..., N(t)$ and α corresponding to the shape parameter and β to the rate parameter

3.4 Ratemaking Models

Determining the correct premium to be charged by each policy is considered a big target for an actuary.

This process consists of building a tariff based on a set of techniques, with the final purpose of determining the policyholder's premiums.

In practical terms, the elaboration of a tariff is nothing more than aggregating the insured population of a given insured product and establishing a set of rules built on a population basis and their relationship with the insured product.

To be more precise, the actuary should analyze the whole portfolio of policies and cluster the records into groups with the same profile. The characteristics that allow this separation should be the ones that cause the most significant impact regarding determining the risk incurred by the insurer. These groups are named tariff levels that have to be unique and mutually exclusive. Then, without further detail, the premiums for each of the tariff levels are built based on the premium of a standard insured level.

Summing up, a tariff must identify the group of rules that will allow the insurer to calculate, for each policy, the premium required to be charged in order to fulfill the losses originated.

Ultimately, if the actuary does a good job, the construction of this tariff should cover the future expenses of possible claim occurrences. This will inevitably revert to a sustainable situation for the insurer, providing the solvency stage (the self-sufficient stage).

The ratemaking process can be categorized based on two phases:

- ***A priori*** Ratemaking: In this initial step, the insurance premium calculation is based on the characteristics of the insured. Each policyholder can be allocated into similar groups based on the aggregating features. There is no need to claim event occurrence to determine the final target (premium).

- ***A posteriori*** Ratemaking: After the premium calculation *a priori*, the initial premium can be adjusted based on additional information. In this step a *Bonus-Malus* system is taken into account. However, this is a topic out of the scope of this dissertation ([4]).

In this dissertation, we will only address a *priori* tariff since the final objective is to focus on the determination of the optimal reinsurance retention limit for the optimal reinsurance for this specific insurance portfolio.

[4]For further information of *Bonus-Malus* systems, we suggest the following reference: Lemaire (1995).

3.4.1 An *a Priori* Ratemaking

As presented in section (3.4), the concept of a *priori* ratemaking is almost self-explanatory. The term a *priori* is linked to the fact that the premiums calculation process is made without extensive knowledge of the policyholder's ability to originate a claim. The elaboration of an *a priori* tariff by modeling the expected value of the aggregate claim variable (S) directly, does not allow us to understand the impact of the variables that influence each of the dependent variables: the claims frequency and the severity of the claim.

The common alternative consists of modeling each of these two random variables individually. I.e., once we prove that there is no collinearity between N and X, it is possible to estimate them individually. Therefore, during our research, we tried to find an individual model that could fit a function that could relate (individually) the dependent variables to the independent variables.

3.4.1.1 Generalized Linear Models

A traditional tool to build an a priori ratemaking is the Generalized Linear Models (GLMs). Generally speaking, a GLM can be described as a group of models that allows the user to estimate the impact of the explanatory variables over the exploratory variables.

In this case, a GLM can be explained as a mechanism that quantifies the insurable risk by estimating the frequency of claims and the amount that should be paid. Consequently, with both of these outputs, the insurance company can obtain the expected insurance premium to be received.

This set of models is considered an extension of the Multiple Linear Regression Model (see Carsey and Harden (2013)). In addition to the classical model, GLMs assume that each $i-th$ element of the responsive variable belongs to the Exponential Family distribution. The link function between the expected value of the exploratory variable and its covariates can be any differentiable function.

From the results presented in McCullagh and Nelder (1989), for each policy belonging to the insurance portfolio, it is possible to formulate the probability density function of each exploratory variable (Y_i) as:

$$f(y_i;\theta_i,\phi,\omega_i) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y_i,\phi,\omega_i)\right\} \tag{3.15}$$

where,

- ω_i corresponds to the risk exposure;
- $b(\cdot)$ is characterized by the parameters θ_i and ϕ;
- The function $c(\cdot,\cdot,\cdot)$ is not relevant for GLM's.

Compared with the classic linear regression model, this extension allows several improvements. Those are mainly related to the characteristics of the dependent variable (variable's distribution, heteroscedasticity, ...), with the independent vs. dependent variable's relation, among other factors.

3.5 Portfolio Simulation: The Monte Carlo Method

Data simulation is a widely used tool to understand specific business behaviors in insurance. It gives the scientist the possibility of building hypothetical scenarios that can be very similar to the real world.

The Monte Carlo (MC) simulation techniques are considered efficient for obtaining an insurance portfolio.

Theoretically speaking, this method consists of algorithms used to build a deterministic model.

Once the distribution for each exploratory variable is defined, the MC technique uses the pseudo-random number generator (PRNG) method to obtain a sequence of numbers from an initial seed. In this case study, the final goal is quite simple. Following CCarsey and Harden (2013), we will generate our benchmark focused on the utilization of Monte Carlo Techniques in order to simulate Generalized Linear Models for the claims frequency and claims severity.

3.5.1 Simulating a Count Generalized Linear Model for N

Starting with N. This random variable only takes integer values as final output ($N = 0, 1, 2, 3, ...$). Therefore, it can be modeled by a count generalized linear model.

Following it, N is usually modeled by using an $(a, b, 0)$ class of distributions. In academic terms, the preferred distributions for their modeling are the Poisson regression model and a Negative Binomial regression model.

According to Guerreiro (2016), the Negative Binomial approach is more flexible than the Poisson. This is derived from the fact that it allows the inclusion of an overdispersion factor. With this, the actuary can simulate the number of claims assuming that the mean of N is different from its variance.

We will focus on this variable's simulation process following a Negative Binomial modeling regression based on this information.

Using the distribution reparametrization ([5]) presented in section 3.3.4 in Guerreiro (2016), we can derive the expected value and the variance of N as:

$$E[N] = \mu; \tag{3.16}$$

[5]The reparamterization explicit in the reference allows the reader to define the distribution of N by using a mean parameter and an overdispersion factor, i.e., $N \sim BN(\mu, \nu)$, where μ and θ are defined as in (3.16) and (3.17.

and

$$V[N] = \mu + \frac{\mu^2}{\theta}. \tag{3.17}$$

Where,

- μ stands for the mean of N;
- θ stands for the overdispersion factor.

As explained in Carsey and Harden (2013), the critical step to simulate N consists in linking the model's systematic portion ([6]) to the mean of the Negative Binomial probability distribution. After this, we are able to draw the exploratory variable N.

Fixing the overdispersion parameter as constant, we can build the mean regression structure of N as $g(\mu_i) = \eta_i = x_i'\beta$, where g is the link function, $\beta = (\beta_0, ..., \beta_p)'$, corresponds to the vector of the mean regression parameters; x_i is the $i-th$ vector value of the explanatory variables, and η_i is a linear predictor. In this case $g(.) : (0, ..., \infty) \mapsto \mathfrak{Re}$. A usual mean link function used in this type of study is the logarithm, i.e., $g(\mu) = log(\mu)$ ([7]).

In practical terms, this procedure can be obtained by equating the model's systematic part of N to its mean. By doing it, we can infer that:

$$E[N] = exp(\beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + ... + \beta_k Y_k), \tag{3.18}$$

and

$$V[N] = exp(\beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + ... + \beta_k Y_k) + \frac{(exp(\beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + ... + \beta_k Y_k))^2}{\theta}. \tag{3.19}$$

With,

- Y_i, corresponding to the explanatory variables of N;
- β_i, corresponding to the linear predictors;
- θ corresponding to the overdispersion parameter.

3.5.2 Simulating a Count Generalized Linear Model for X

The claim's amount distribution simulation is more complex than the claim's number process. Since we are dealing with a reinsurance contract built based on a simulated insurance portfolio, there are several aspects to be considered.

As it is shown in (Guerreiro (2016)), the distributions that are commonly used to explain X are the *Gamma* distribution, the *LogNormal* distribution, or in more rare cases, a *Weibull* or a *Pareto* distribution.

[6] Systematic Component: It refers to the explanatory variables $(X_1, X_2, ..., X_k)$ as a combination of linear predictors; e.g. $\beta_0 + \beta_1 X_1 + \beta_2 X_2$. Using our nomenclature, it corresponds to η_i.

[7] The logarithmic function confers to the actuary a multiplicative ratemaking structure for N (see section 3.6 in Guerreiro (2016)).

These are the most commonly used distributions by the great majority of the references at an academic level. In order to simplify the studies, X is usually fitted as one of the previous distributions.

However, in practical terms, it is tough to fit a single distribution to X. Particularly in portfolios for reinsurance purposes, it is widespread to observe the existence of costly claim sizes. Usually, the values are higher than observed under a regular insurance agreement.

A widespread solution adopted by actuaries when doing the fitting process is to break the distribution of X into smaller pieces. Then after it, there is tailored work on allocating a distribution to each part.

Let us take a stand on the simulation side. Facing the problem previously identified, we can break X into two parts.

To avoid an excessive increase of complexity, a simple combination that can be performed is to mix two *Gamma* probability distributions. In mathematical terms: In mathematical terms: $X = \gamma_1 \times X_1 + \gamma_2 \times X_2$, where X_1 corresponds to the left part (lower claim sizes) of X and X_2 to the right part (higher claim sizes) of X. γ_1 and γ_2 correspond to the portfolio's proportion of each part, X_1 and X_2.

This particular choice anticipates the next step: the elaboration of two *Gamma* ratemakings.

Recalling the simulation of X_1 and X_2, the first step is to reparametrize each one of them, as shown in equation (15) of section 3.3.7 in Guerreiro (2016).

With this, it is possible to derive the expected value and variance of each X as:

$$E[X_1] = \frac{\alpha_1}{\beta_1}; E[X_2] = \frac{\alpha_2}{\beta_2}. \tag{3.20}$$

and

$$V[X_1] = \frac{\alpha_1}{\beta_1^2}; V[X_2] = \frac{\alpha_2}{\beta_2^2}. \tag{3.21}$$

With,

- α_1 and α_2 being the shape parameters of X_1 and X_2 correspondingly;
- β_1 and β_2 stand for the rate parameters of X_1 and X_2 correspondingly.

Applying the same theory used for the simulation of N, we can set the shape (*Gamma* parameter) of X_1 and X_2 as fixed. After that, we replicate the mean regression's structure with a logarithmic link function to ensure the applicability of a multiplicative ratemaking to each part of X.

We match the model's systematic part of X_1 and X_2 to the correspondent means. By doing it, we can infer that:

$$E[X_1] = exp(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_k Z_k) \tag{3.22}$$

and

$$E[X_2] = exp(\beta_0' + \beta_1' W_1 + \beta_2' W_2 + ... + \beta_k' W_k), \tag{3.23}$$

with,

- Z_i and W_i corresponding to the variables that explain the dependent variable X_1 and X_2 correspondingly;
- Each i-th element of β_i and β_i' are the linear predictors of X_1 and X_2 correspondingly.

3.5.3 Simulating the Explanatory variables

As previously explained, we are interested into simulating the mean regression structure of N, X_1 and X_2, generally summarized by the following three points:

1. The link function: $g(\mu_i) = \eta_i$;
2. The mean regression parameters, $\beta = (\beta_0, ..., \beta_p)'$;
3. The vector value of the explanatory variables, $Y = (Y_1, Y_2, ..., Y_k)$ in case of N, $Z = (Z_1, Z_2, ..., Z_k)$ in case of X_1 and $W = (W_1, W_2, ..., W_k)$ in case of X_2.

Relying on the third point, we are now focused into simulating Y, Z, and W.

In the following sections, we are concerned about the considerations when simulating Y in the case of N, Z in the case of X_1 and W in the case of X_2.

3.5.3.1 Considerations on the Explanatory variables Simulation

When dealing with simulated data, the actuary must be careful about the existing relationships between the variables. Recalling that we are simulating fictitious variables that specific models will use, there is a need to be aware of the premises of each one of the models.

In the collective risk model, we have that N and X are independent of each other in this case. Consequently, the variables that explain these two primary variables should not be the same, or at least, they should not be highly correlated.

This premise is shared when calculating insurance premiums. A Generalized Linear Model assumes that the explanatory variables selected as significant to each dependent variable should not be correlated.

With these considerations in mind, we will get into more detail in the following sections to simulate Y in the case of N, Z, and W in the case of X.

3.5.3.2 Simulating the Explanatory variables of N

The definition of the explanatory variables of N should follow an objective criteria for the driver's characteristics.

Recalling what is presented in (3.1), "*the concept of insurable risk is considered to be a random variable and it must follow these conditions: incorporate uncertainty in future and be completely accidental, measurable, and quantifiable.*"

Therefore, the driver's ability should be discarded in this particular study. As such, the simulation of N should not consider variables such as Sex, Education Level, or other variables related to the policyholder's characteristics.

As a final decision, the choice relies on variables mainly associated with the vehicle's driver.

3.5.3.3 Simulating the Explanatory variables of X

Usually, it is said that the damage itself explains the claim's amount.

To understand this statement, let us consider a hypothetical practical scenario in which a policyholder intends to do a dangerous car overtaking. Then, accidentally, he/she hits the car in front of them. What would define the price's claim?

Since we ignore the driver's ability, the answer is not immediate.

However, we know that an impact on a one-year Porsche car will be more expensive than an impact on a ten-year Renault car. Therefore, a more "powerful" vehicle is taken as more likely to originate a more severe claim.

With that in mind, a possible answer to the previous rhetoric question is: The price's claim is defined by the vehicle's characteristics.

Recalling the way we defined the claim's amount ($X = \gamma_1 \times X_1 + \gamma_2 \times X_2$), we will have to consider two different simulation processes. A first one for the simulation of less severe claims, X_1, and a second one for X_2 which considers more expensive claims.

A reader framed with data science problems will immediately identify the need to simulate clustered data for X by doing this split.

In the following section, we will briefly explain the core steps to simulate this type of data ([8]).

3.5.3.4 Generating Clustered Data through Gaussian Mixture Models

In order to simulate data that can be classified into similar groups (clusters), a commonly used tool is Gaussian Mixture Models (GMMs).

A GMM is a probabilistic model that requires that all data points are simulated using a mixture of a finite number of Gaussian distributions. In addition, these models incorporate information about the data covariance structure and the centers of the latent Gaussians.

A GMM is built based on the Expectation-Maximization (EM) algorithm. This approach can be seen as an alternative that extends the k-means algorithm.

[8] For more detailed information we suggest the reader to consult *Gaussian mixture models: k-means on steroids*.

Briefly explaining, the EM algorithm intends to find the parameters that fits best the data. The mathematical formulation of this algorithm can be reached in (Xuan et al. (2001)).

In this dissertation, we are interested into simulating the GLM of X_1 and X_2 through a GMM.

Recalling (Hamerly and Elkan (2003)), a k-means algorithm is based on the notion of similarity of each point to a *centroid* of a cluster. By saying this, an immediate problem emerge. Usually, the concept of similarity implies the usage of distance definitions. This means that a k-means algorithm only allows the usage of continuous variables.

That is one of the main fundamentals of using Gaussian distributions. These distributions are continuous, and there are no constraints when using them, at least at the first sign.

However, it is frequently observed that the explanatory variables are categorical in a real business. Moreover, as was reviewed previously, the variables used in GLMs are split into categories.

In practical terms, a recurring solution to solve this barrier is to separate the categorical variable's simulation from the continuous variables.

For continuous variables, the strategy consists of applying GMMs. The design of this type of variable is fundamentally based on a covariance structure between the variables identified to simulate each GLM.

Once this structure is defined, depending on the inputs to the covariance setup, there is an infinite set of variable relationships that can be obtained from it. ([9]).

Categorical variables can be simulated based on simple combinatorial sampling methods. Considering the scientist's expertise, it is possible to generate this type of variable by assigning each record's probability belonging to a pretended category.

3.5.4 Premiums Calculation

In the context of ratemaking calculations, the two main variables used for modeling premiums are the claims frequency and the claims severity.

Although these two concepts are obtained from N and X, the usage of these new concepts is more accurate than the latter ones. Let us see the value of this statement.

The modeling of these two variables is possible due to intermediate steps that are fully explained in Guerreiro (2016). In order to avoid an information overload, we will not expose all the concepts behind the theory of a tariff construction.

Based on this, we only want to highlight that the premium's calculation method used in this study will rely on the multiplicative model.

[9] In *Gaussian mixture models: k-means on steroids* it is possible to do a more extensive review on these possibilities

To apply it, the continuous explanatory variables, such as: the driver's age, the vehicle's age, among others are discretized by considering an equal (or almost equal) number of observations for each of the categorical variable's bins.

Again, to get more information about the common nomenclature used in generalized linear models it is recommended to recall Guerreiro (2016).

3.5.5 Insurance Profit Calculation

After describing an event through mathematical language, an actuary can start playing with the concepts and build useful metrics to explain the business.

Following it, an important measure to monitor is the insurance profit of each line of business.

The formulation of this concept can be defined by the difference between the insurance gains obtained by the policyholder's premiums and the losses resulting from the claims observed during the period considered:

$$E[P] = ct - E[S(t)], \tag{3.24}$$

where,

- c stands for the premiums collected until time t at a constant rate $c > 0$;
- $E[S(t)]$ denotes the expected value of the aggregate claims amounts paid until time t.

3.6 Reinsurance

According to Carter (2013), there are two main types of reinsurance contract: the facultative versus the obligatory and the proportional versus the non-proportional.

Facing a facultative reinsurance contract, the reinsurer has the power to deny, case-by-case, some or all of the policies that the insurer freely decides to reinsure. In the obligatory one, also called automatic reinsurance, the reinsurer is obliged to share all the risks subscribed by the treaty.

The name automatic is originated by the fact that the acceptance of the risks, do not need any approval from the reinsurance party. In this case, the insurer intends to reinsure all the policies of their portfolio.

The facultative approach is commonly used to either complement the obligatory form or used in cases where the obligatory one is not available to a specific type of risk. It means that the facultative form enables to underwrite risks in addition to those covered by the obligatory reinsurance. Both previous types can be separated into proportional and non-proportional.

As the name suggests, the proportional reinsurance consists into sharing equally the premiums and losses and agreed proportional percentage share, for each written policy by the primary insurer. The non-proportional reinsurance is only concerned

with losses. In this type, an upper limit is set for the insurer's losses. Above the established limit, the amount of claims are carried by the reinsurer.

Regarding the proportional form, we have Quota Share and Surplus reinsurance treaties. Within the non-proportional form are the Excess of Loss and Stop Loss treaties.

3.6.1 The Excess of Loss reinsurance contract (XLS)

The present study intends to address the non-proportional Excess of Loss (XLS) type of reinsurance contract.

Under an XLS, it is set a maximum claim amount, M, for the insurer's losses covered by the contract. This threshold is named as deductible, priority or retention limit.

The covered claims that exceed M are transferred to the reinsurer's side. Usually, this second party of the contract also imposes an annual threshold based on the total claim's amount, m.

As such, this type of contracts can be defined as an $m \times M$ excess of loss reinsurance contract, that can be read as an m in excess of M reinsurance contract.

In many cases, m is defined as infinite, meaning that the reinsurer will keep with all losses above the threshold defined for the cedent part, M. In this particular case, the contract is only identified based on M.

Regarding compensation definition, this type of contracts can be distinguished into three main categories: The Excess of Loss per risk, the Excess of Loss per event and the aggregate Excess of Loss.

Since we are interested into a type of reinsurance that gives the insurer a protection against a single loss or risk incurred at a specified amount, we will only do a review of an XLS per risk.

Taking this scenario, let's consider a reinsurance XLS treaty. As such, we will denote X as the claim's amount variable.

Following this contract, the parameterization of the insurer ceded and retained claim amount defined as:

$$X^{ced} = max(0, X - M) = \begin{cases} 0, & \text{if } X \leq \text{M} \\ X - M, & \text{if } X > \text{M} \end{cases} \tag{3.25}$$

$$X^{ret} = min(X, M) = \begin{cases} X, & \text{if } X \leq \text{M} \\ M, & \text{if } X > \text{M} \end{cases}; \tag{3.26}$$

The retained claim distribution function is given by:

$$F_X^{ret}(x) = \begin{cases} F_X(x), & \text{if } x < \text{M} \\ 1, & \text{if } x \geq \text{M} \end{cases} \tag{3.27}$$

Considering the aggregate claims process for a given period defined as previously in section (3.3), we can define the aggregate ceded and retained claim amount as:

$$S^{ced}(t) = \sum_{i=1}^{N(t)} X_i^{ced}; \tag{3.28}$$

$$S^{ret}(t) = \sum_{i=1}^{N(t)} X_i^{ret}. \tag{3.29}$$

The main advantage of an excess of loss per risk is the risk mitigation against large single losses.

3.7 Optimal Reinsurance

Following the third chapter of Holtsmark (2015), in this dissertation we pretend to maximize the profit retained by the insurer and minimize the variance of the aggregate claim process retained by the insurer.

In this last reference, it is proved that an Excess of Loss contract is the optimal reinsurance agreement for the combination of these two specific criteria.

3.7.1 Premium Principles

As it can be addressed in Bühlmann (1970) or Young (2006), an insurance premium is performed by using a premium principle. Which corresponds to a rule for assigning a premium to an insurance risk.

The previous literature suggests that premiums can be computed through a Pure Premium Principle, an Exponential principle , an Expected value principle, a Variance principle, among others.

To obtain the reinsurer's premium we will consider the Expected Value Premium Principle. By definition, it adds a safety loading ([10]), $\eta > 0$ to the expectation of the insurer's claim costs.

In this particular case, it is possible to derive the following expression for the expected value of the reinsurance premium: $C_{Reins} = (1+\eta) \times E(N) \times E(X^{ced})$.

[10] Loading for risk is desirable because a general requirement for a premium rule is to charge at least the expected payout of the risk S in exchange of the insured risk.

3.7.2 Insurance Profit Calculation after Reinsurance

For this case-study, the insurance profit is considered as one of the main criteria to obtain the pretended optimal retention limit.

Similarly to what we have presented in section (3.5.5), it is crucial to define this measure in mathematical terms, after applying the pretended reinsurance contract to the portfolio.

After the application of the reinsurance contract to the simulated portfolio, the expected value for the insurer retained profit can be expressed as:

$$E[P_{ret}] = c - C_{Reins} - E[S^{ret}], \tag{3.30}$$

with $E[S^{ret}] = E[X^{ret}] \times E[N]$.

3.7.3 The Variance Calculation of the Aggregate Claims Amount retained after Reinsurance

After applying the excess of loss reinsurance contract, it is pretended that the variability of the aggregate claim's amount retained by the insurer should not fluctuate too much.

Following the same approach as in equation (3.6) in section (3.2.2), the variance of the Aggregate Claim's retained by the insurer after reinsurance can be expressed by: $V[S^{Ret}] = V[X^{Ret}] \times E[N] + E[X^{Ret}]^2 \times V[N]$.

3.8 Decision Theory

Decision theory is a cross-cutting subject to several social sciences. The insurance business is not an exception. Companies that are always subject to risky conditions need to evaluate accurately the decisions that are performed. It is important for the insurer to feel confident about the decisions made.

In mathematical terms, this concern can be carried by a specific area, so-called multi-criteria decision making (MCDM).

Briefly explaining, it considers situations in which a decision produced by a decision agent can be re-evaluated by different perspectives, according to different criteria.

In fact, the same decision to solve a problem can lead to different decisions by different agents: while some can opt for the decision that maximizes the result (for example, the insurance profit), others can decide to select the decision that minimizes this factor. The latter are decision-makers who prefer not to opt for the decision that would maximize the profit, due to the future unpredictability of facing a bad result.

After some research in this topic, according to Kahraman (2008), there are two main categories when considering multi-criteria decision methods:

- Multiple-Attribute Decision Making (MADM);

- Multiple-Objective Decision Making (MODM).

Since we pretend to select the best alternative out of a set of finite possibilities, we will only review the Multiple-Attribute Decision Making (MADM).

As it is presented in Karageyik and Dickson (2016), a MADM is a family of algorithms that generally can be explained easily based on a matrix structure procedure (commonly named as decision matrix or evaluation matrix).

As such, let's consider A as a matrix constituted by M rows and N columns. Based on this, we can allocate to each row an element of $O = (O_1, ..., O_m)$, corresponding to a vector with m alternatives for the scenario in study. For each column, we can define a vector with n criteria considered by the analyst to perform the final choice, $C = (C_1, ..., C_n)$.

An element belonging to $A = \{a_{ij}\}$, can be described based on the pair (O_i, C_j), where $i = 1, ..., M$ and $j = 1, ..., N$. The following table summarizes the presented matrix formulation.

Table 3.1: Decision matrix for MADM methods. **Source:** Karageyik and Dickson (2016)

	Alternatives (O_i)	**Attributes (Criteria)** (C_j)			
		C_1	C_2	...	C_n
A=	O_1	a_{11}	a_{21}		a_{n1}
	O_2	a_{12}	a_{22}		a_{n2}
	⋮	⋮	⋮	⋯	⋮
	O_m	a_{1m}	a_{2m}		a_{nm}

Considering a MADM, the optimal solution is obtained depending on the selected algorithm.

In Velasquez and Hester (2013) and Karageyik and Dickson (2016), several methods are presented, however based on the advantages and disadvantages of each method, this study-case will do a review on the TOPSIS-MADM method.

3.8.1 TOPSIS

TOPSIS stands for **T**echnique for **O**rder of **P**reference by **S**imilarity to **I**deal **S**olution. As explained in Qin et al. (2008) it is "an approach used to identify an alternative which is closest to the ideal solution and farthest to the negative ideal solution in a multi-dimensional computing space".

Simply explaining, TOPSIS is a simple decision-making algorithm that intends to provide to scientist the possibility of finding an equilibrium between an outperforming criteria and a criteria that is performing worse than expected.

The main goal of the following section is to provide a practical explanation of this method using a fictitious scenario.

3.8.1.1 TOPSIS - Graphical Explanation

In order to explain the TOPSIS algorithm, we will consider a two-dimension hypothetical scenario:

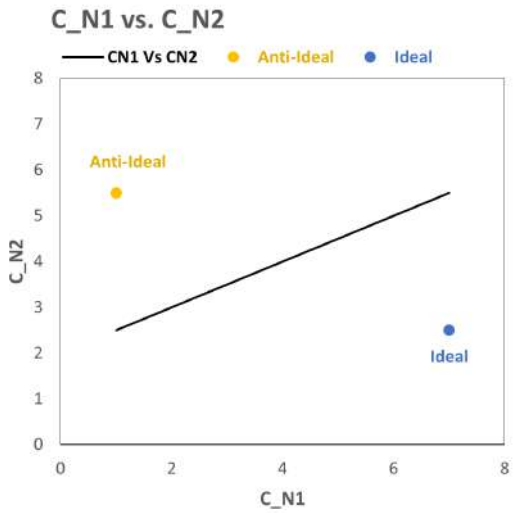
> Let's suppose that an actuary intends to determine the optimal solution for a set of alternatives of a particular research scenario, let's call it scenario A.
>
> Considering A, he/she establishes two general criteria, C_1^A and C_2^A.
>
> Without loss of generality, he/she decides to:
>
> - Maximize $(C_1^A)_N$;
> - Minimize C_2^A.

A logical initial step to achieve the optimal point, the one that finds the best balance between both criteria, is to visualize C_1^A against C_2^A.

In this step, it is important that both criteria can be comparable. To do it, an usual approach is to normalize the variables.

Let's then assume that C_{N1}^A and C_{N2}^A are directly obtained by normalizing C_1^A and C_2^A respectively. Additionally we can suppose that C_{N1}^A against C_{N2}^A behavior may be presented by the following chart:

Figure 3.2: Hypothetical Scenario of C_1^A Vs. C_2^A. **Source:** Authors.



Looking to the presented chart, it is possible to conclude that higher values of C_1^A imply the same type of behaviour for C_2^A. Therefore, it is not straight to fulfill the constraints for both variables (maximize C_1^A and minimize C_2^A).

If the choice relies on a point belonging to the segment located on the right side of the chart, it will fulfil the requirement of maximizing C_1^A, however it will fail the second constrain of minimizing C_2^A. The opposite occurs when picking a point located on the left side of the black segment.

Facing it, the problem will be solved if the decision-maker pre-establishes a win-lose balance for both variables.

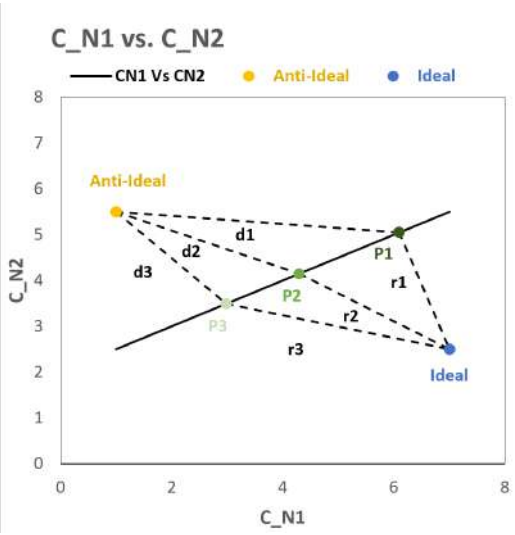
A natural first step to achieve the final goal would rely on the identification of two particular solutions in the previous chart:

- The best solution, named as Ideal Solution;
- The worst solution (named as Nadir or Anti-Ideal).

After doing this, the actuary can try to identify some points belonging to the curve (Pareto Curve) that can enable the discovery of the minimum distance of C_1^A and C_2^A, simultaneously.

By doing it, it is possible to re-draw the previous chart as:

Figure 3.3: Hypothetical Scenario of C_1^A Vs. C_2^A. **Source:** Authors.



Recalling the established constraints for both criteria, it is expected that the Ideal solution point would be drawn in the lower right corner (blue point) of the chart. On the other hand, the Anti - Ideal point is placed on the opposite corner (yellow point).

Looking to the black line, it represents the set of feasible solutions to the problem formulated by the actuary. With this in mind, a second step is to evaluate what would be the point belonging to this segment that is closer to the Ideal point and farthest to the Anti-Ideal point.

The rational behind this algorithm is to find the solution that maximizes the distance between each Pi and the Anti-Ideal point, given by di, and minimizes the distance between each Pi and the Ideal point, ri. In common language, TOPSIS is intended to attain the solution that is less similar to the Anti-Ideal point and more similar to the Ideal point.

This process is obtained by the calculation of a similarity score based on both distances, di and ri ([11]).

[11] $P_i \in P = (P_1, P_2, P_3)$, $r_i \in r = (r_1, r_2, r_3)$ and $d_i \in d = (d_1, d_2, d_3)$.

With more mathematical rigor, each step of the TOPSIS algorithm, for the specified constraints (maximize C_1 and minimize C_2), can be summarized by the following table:

Figure 3.4: TOPSIS Algorithm, Step-by-step workflow. **Source:** Authors.

TOPSIS Algorithm | Step – by – Stepj

Step 1: Create a matrix $\boldsymbol{A} = \{\boldsymbol{a_{ij}}\}_{\boldsymbol{C\times N}}$.

It consists of creating a matrix with **C rows** and **N columns**.
C corresponds to the number of alternatives and **N** to the number of criteria

Step 2: Normalize A.

The normalization method used in this study is obtained for each element of A as:

$$\alpha_{ij} = \frac{a_i}{\sqrt{\sum_{\mathrm{i}=1}^{M}(a_{ij})}}$$

Step 3: Assign the importance to each criterion

It consists of creating a vector $\omega_j, with\ \mathrm{j} = 1, \dots, N$.
It represents the weight og each criterion.
A second step is to multiply ω_j by A, with the constraint of: $\sum_{j=1}^{N} w_j = 1$.
It originates $A^* = \omega_j$ x A.

Step 4: Determination of best and worst alternatives of A^*

The **Ideal** (best) solution is defined as: $\boldsymbol{I} = \left(\mathbf{Max}(\boldsymbol{A_i^*})_{\boldsymbol{C_1}}, \mathbf{Min}(\boldsymbol{A_j^*})_{\boldsymbol{C_2}}\right)$,
The **Anti-Ideal** (worst) solution is defined as: $\boldsymbol{AI} = \left(\mathbf{Min}(A_i^*)_{\boldsymbol{C_1}}, \boldsymbol{Max}(A_j^*)_{\boldsymbol{C_2}}\right)$.

Step 5: Calculation of each alternative distance to the best and worst solutions.

In this study we will consider the Euclidean distance for calculations.
The distance to the Ideal solution is given by: $\mathrm{d}^I = \sqrt{\Sigma_{\mathrm{i}=1}^{N}(A_{ij} - I_{ij})^2}$
The distance to the Anti-Ideal solution is given by: $\mathrm{d}^{AI} = \sqrt{\Sigma_{\mathrm{j}=1}^{N}(A_{ij} - I_{ij})^2}$

Step 6: Calculation of TOPSIS scores.

The TOPSIS method allows the agente to rank the similarity of each alternative to the Ideal solution and to the Anti-Ideal Solution.
It is obtained based on both distances calculated in Step 5: $S_i = \frac{\mathrm{d}^{AI}}{d^{AI}+\mathrm{d}^I}$.

Step 7: Rank the scores in descending order.

4 | Methodology

To achieve the objective of finding the optimal level of reinsurance to be retained in a simulated automobile portfolio, the workflow will be divided into five main steps.

Firstly, a database will be obtained through data simulation. Using Monte Carlo techniques performed in R-tool, an insurance automobile portfolio will be generated by doing several assumptions in multiple dimensions, such as:

- The identification of the variables that are commonly used in this type of studies;
- The choice of the distributions to be used for each random variable taken as of interest for the study;
- The choice of the parameters to be used as inputs for the variables distributions.

The aforementioned points are related to the insurance business and the way in which information in this area is reflected in terms of data. Consequently, the actuary must be careful with the type of assumptions such that the inputs used to perform the simulation do not bias the results of the study.

In general, the actuary who is involved in the insurance sector has a great access to the business knowledge. Assuming the advantage of being able to extract historical data from internal datamarts, it's very common to resort to the application of resampling methods to perform the simulation of the pretended data. With this methodological technique, the actuary can mimic existing information by considering samples of the databases and generate new data containing the information of the initial sources.

Another way to achieve this goal is through a more theoretical approach. By doing an intensive study about the business, the simulation can be performed based on reliable references. As an example, if there is a particular interest of assessing the accident rate of a country/city/region, it's possible to search for information developed by statistical entities dedicated to this type of studies. Experimentally, the actuary can replicate the scenario based on the evidences observed.

In this project, the selection of the portfolio variables to be used was mainly done following two references in non-life insurance: Kaas et al. (2008) and Guerreiro (2016). Due to the absence of an online database with the desired characteristics for this study and also due to the new data protection policies of the insurance companies, the selection of the variables and their parameters were subjectively done. However, a conditional subjectivity was conducted.

To identify the variables that are usually associated with the occurrence of claims in Portugal, an exhaustive search on the national institute of statistics website about the accidents occurred in Portugal during the period from 2000 to 2019 was made. An example of the webgraphy approached for this analysis can be addressed by the statistical yearbooks available in: *Estatísticas dos Transportes e Comunicações 2004* or *Estatísticas dos Transportes e Comunicações 2019*.

Along with the identification of the general claim patterns evidenced in the references, the simulation process has progressed to a specific insurance product usually named as Own Damage car insurance. As such, it is deemed necessary to take into account its characteristics.

An Own Damage (OD) car insurance coverage is an optional contract, where the insured expects a reimburse from the cedent in case of an automobile self-damage unexpected event. Which means that the simulation process should consider that:

- The OD products are related with the type of client's car. It is reasonable that more expensive cars should have a higher coverage, when compared with the cheapest ones.
- An OD product requires some important variables to consider. For instance, the claims cost and the number of claims is crucial, because are considered as key variables when calculating the amount of indemnities in the company's portfolio.
- A claim originated by a car/vehicle under an OD product shall not be higher than the car's/vehicle's value.
- It is deemed important to have a holistic view of the features and patterns of the OD type of clients. With that in mind, variables such as zone of residence, driver's age, years of driving license and others are also relevant.

Based on the points evidenced previously, it is increasingly clear that the simulation process for this study is in fact more complex than what is taken as usual in a simple variable statistical simulation context.

In this project, the usage of Monte Carlo techniques will be seen as a vehicle to perform the simulation of a specific model structure, the Generalized Linear Models. Along with it, the dimension of this simulation procedure will increase by introducing a clustering structure to our variables dataset at severity level ([1]).

Broadly explaining, the framework starts based on the premise of simulating the two components of the regular aggregate claims process:

- The claims frequency process.
- The severity claims process.

[1] The reader may perceive that an abusive use of language is being incurred. With this, it is intended to emphasize that the cluster structure will be introduced when simulating the severity process.

And as usual, for each regression model a response variable is defined and described through the relationships verified with a set of independent variables, named as explanatory variables. In this particular case, for the claims frequency process the response variable corresponds to the number of claims and for the case of the severity process, it corresponds to the claims amount.

Particularly in this study, the claims frequency process is explained by a Negative Binomial count model, which implies that the exploratory variable (the claims number) follows a Negative Binomial distribution.

Regarding the severity process, this is where the clustering framework comes in. We will simulate two independent generalized linear models considering each exploratory variable following a *Gamma* distribution. By taking advantage of the Gaussian Mixture Models, it will be possible to design both GLM's in such a way that one of the response variables will only constitute cheaper compensations and the other one will only constitute more expensive compensations, giving rise to two mutually exclusive groups (clusters).

This approach is defined based on the requirement of using the portfolio information for reinsurance implementation. With this, it's intended that both the distribution of the number of claims and the distribution of the severity of claims should have a higher density of observations in the left tail of the distribution and a lower density in the right tail of the distribution (right skewness).

After simulating our database based on the previous assumptions, we will be able to perform the **second step** of this dissertation. It corresponds to the *a priori* analysis of the simulated database. In an uncompromising way, the actuary will perform several analyses of the information in the database, in order to extract patterns, trends, relationships of the variables.

In this step, a quantitative and qualitative analysis of the variables considered will be performed based on the application of hypothesis tests, analysis of statistics, indicators, among others.

In a very broad way, we will include a graphical analysis through the elaboration of line charts, histograms, box plots, bar plots among others, with the ultimate purpose of identifying the main characteristics hidden in the data. Initially, this approach will be done taking into consideration only the explanatory variables, but gradually we will follow the path in order to focus on the variables that are taken as the main ones for our project (Response variables: Number of claims/claims frequency and claims amount/severity) .

Nowadays, this step is considered as a fundamental and transversal step for all industries and in particular for insurance companies. Looking at the non-life sector, the step in point may prove to be fundamental to understand the portfolio's behaviour. This knowledge will undoubtedly be crucial in terms of the premiums calculation that better suits the insured profiles.

Proceeding to the **third step** a set of rules will be established to calculate the premiums that the insurer should charge for the product provided. For this purpose, the usage of Generalized Linear Models will be a crucial tool to achieve it. By using it, we will be able to calculate the expected value of the claims number and the expected value of the claims amount of each policy. Consequently, it will be possible the expected value of the aggregate claims process.

In this step, we will also include a clustering analysis (k-means). This action will facilitate a lot the perception of possible patterns/groups that may be "hidden" in the data. However, as this procedure can be considered as an independent dissertation topic, we will not explore this topic too deeply. As such, we will keep with a very broad approach, but always with the ambition of extracting as much information as possible.

These initial steps consist of preparing all the premises needed to build a non-proportional reinsurance treaty to the portfolio and, in a final stage, in determining the optimal retention limit to be applied.

Moving towards the end of the study, the **fourth step** will consist on the application of an Excess of Loss reinsurance treaty to the simulated portfolio. In this step, the non-proportional contract will be performed based on section (3.6) of chapter (3).

The **last step** relies on answering to the initial goal of finding the optimal reinsurance retention limit. To achieve this objective, we will first proceed to a comparative study of the expected value of the profit of the business line when calculated with and without the application of an Excess of Loss reinsurance contract.

To do it, we will start by considering the simple version, i.e., the profit calculation without applying the reinsurance contract, such as: $E(P) = E[c - S]$, where C and S stand for premiums collected during one year time period and the aggregate amount of the claims also during one year time period, respectively.

In the case of excess of loss reinsurance application, the profit estimate will be calculated, such as: $E[P^{ret}] = c - C_{Reins} - E[S^{ret}]$, where P^{ret} and S^{ret} are the insurer's retained premiums and aggregate loss retained by the insurer in one year time period.

Secondly, we will look at the retained Variance of the Aggregate Claim distribution, $V[S^{Ret}]$, before and after applying the excess of loss reinsurance contract.

To accomplish it, we will consider the application of a multi-criteria decision analysis method, named as Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS). As it is presented in our literature review, this technique evaluates a set of alternatives by considering criteria constraints defined by the actuary.

In this dissertation, we will rely on two criteria:

- Minimizing the retained Variance of the Aggregate Claim distribution;
- Maximizing the retained insurance Profit.

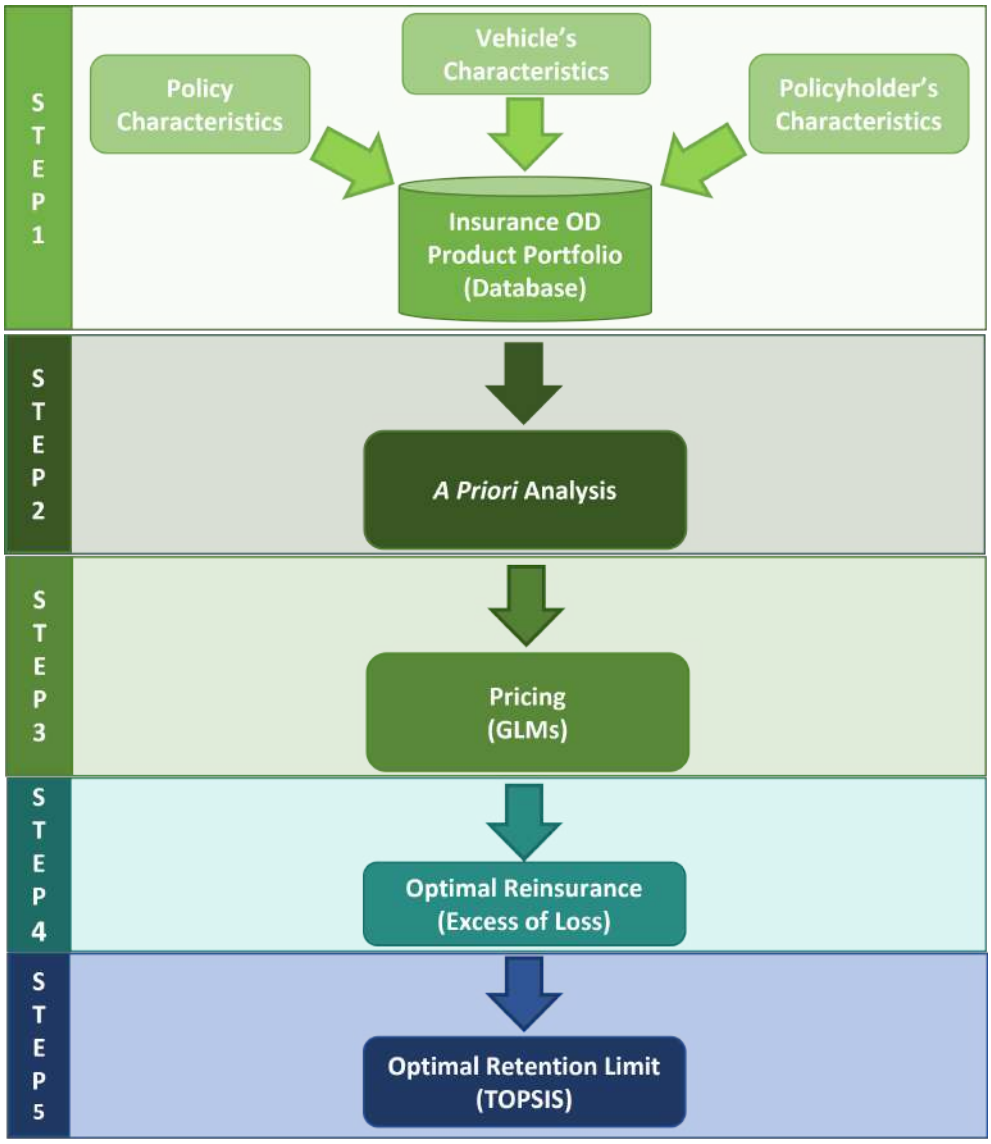
In this study we will fix the entire calculation for a safety loading of ten percent that is usually assigned for compensation reasons to the reinsurer.

Facing the results, it's planned to perform two sensitivity analysis on the results

obtained. A first one related to the uniqueness condition of the value obtained, and a second one related to the assignment of weights to the selected decision criteria.

The following diagram intends to summarize our dissertation workflow:

Figure 4.1: Methodology Flow Chart. **Source:** Authors.



5 | Portfolio Data Description

For an actuary, the layout of the information is very important. The necessary work to be done is undoubtedly affected by the way the database is structured.

In this particular case-study, the information is split into two files. Although they contain complementary information, the structure of both is clearly distinct:

1. The ***BASE_DADOS_N.txt*** file, corresponds to the claims number information. In this file, four hundred thousand observations are presented, with each row corresponding to a single policy. In conclusion, this file presents all the information that characterizes the number of claims of each policy, in terms of the insured product and also in terms of the characteristics of the insured.
2. The ***BASE_DADOS_X.txt*** file, corresponds to the information of each claim originated within a policy. In this case, each row corresponds to the breakdown of the policy taking into account the number of claims observed in the exposure period. The number of observations in this file increases to four hundred and one thousand and seventy-eight.

For both scenarios, the period of exposure considered is one year. Being the starting date, the 30^{th} of November 2020 and the end date, the 30^{th} of November 2021 (1-year time-window).

Since all the records were observed at the same period of time, which implies no dependency between time and the portfolio variables, the data available in this dataset can be classified as a cross-sectional data.

Generally, the two datasets are mostly made up of non-categorical variables. However, the representation of the categorical variables are also significant. Below, it's presented a summary of this split in absolute and percentage terms:

Table 5.1: Absolute and Relative Count of Categorical and Non-Categorical Variables. **Source:** Authors.

	BASE_DADOS_N.txt		**BASE_DADOS_X.txt**	
	#	**%**	**#**	**%**
Categorical Variables	8	42%	8	47%
Non-Categorical Variables	11	58%	9	53%

Remaining with a more qualitative analysis, when comparing the two data files it's possible to find that while the *"BASE_DADOS_X.txt"* dataset is composed of nineteen variables, the *"BASE_DADOS_N.txt"* dataset only presents seventeen. The two missing variables in the latter database corresponds to the information of the amount originated by each policy (the individual amount and the aggregate amount by policy). As such, we can conclude that the *"BASE_DADOS_N.txt"* can bee seen as an aggregation of the information of the *"BASE_DADOS_X.txt"* dataset by policy.

Facing this last conclusion, depending on the type of the pretended analysis, the actuary can choose to focus the efforts in one database only, or both. For example, if there is only interest in analysing the frequency of claims by policy, the option may be to only use the document *"BASE_DADOS_N.txt"*. Whereas, if the interest is to analyse the amount of money insured, then the analysis will fall on the *"BASE_DADOS_X.txt"* document.

Below we can assess the discrimination of the variables available in the latter database ([1]):

Table 5.2: Portfolio variable's description. **Source:** Authors.

	Variable Name (Portuguese)	Variable Name (English)	Variable Type	Description
ID	ID	ID	Key Variable	It identifies uniquely the policies.
Policyholder'sCharacteristics	N_sin	N_clm	Discrete	Number of reported Claims
	Exposicao	Exposure	Continuous	Portion of days that the policy is in force.
	Indem_Indiv	Ind_Clm	Continuous	It corresponds to the individual claim's amount.
	Indem_Agreg	Agg_Clm	Continuous	It corresponds to the aggregated claim's amount.
	Sexo	Sex	Categorical	It corresponds to the policyholder's sex. It is split into two categories: Male and Female.
	Idade_Condutor	Driv_Age	Continuous	It corresponds to the policyholder's age.
	Estado_civil	Civil_Status	Categorical	It corresponds to the policyholder's civil status.
	Hab_Lit	Literacy	Categorical	It corresponds to the policyholder's literay level.
	Regiao	Region	Categorical	It corresponds to the policyholder's region of residence.
	Distrito	District	Categorical	It corresponds to the policyholder's district of residence.
	Anos_Carta	Driv_Lic	Continuous	It corresponds to the number of years of the policyholder's driving license.
Vehicle'sCharacteristics	Combustível	Fuel	Categorical	It corresponds to the vehicle's fuel.
	Tipo_veic	Veic_Type	Categorical	It corresponds to the vehicle's type.
	Idade_Veiculo	Veic_Age	Continuous	It corresponds to the vehicle's age.
	Cilindrada	Displacement	Continuous	It corresponds to the vehicle's displacement.
	Cavalos	Horse_Power	Continuous	It corresponds to the vehicle's horsepower.
	Marca	Veic_Brand	Categorical	It corresponds to the vehicle's brand.
	Valor_veiculo	Veic_Val	Continuous	The current price of the insured vehicle.

[1] All these variables are also available in *"BASE_DADOS_N.txt"*, with the exception of the individual claim amount and the policy claim amount (*Ind_Clm* and *Agg_Clm*, respectively).

6 | Exploratory Data Analysis

This chapter introduces an exploratory portfolio analysis presented previously in the chapter (5). This step will consist of a more quantitative assessment of the databases, performed uncompromisingly. Without an in-depth knowledge or preconceived ideas in this research area, we will do an *a priori* analysis to understand what is "hidden" in the data.

As usual in this type of data science study, the very first step consists of creating a graphical analysis of the variables' outputs. After this step, another standard action in this type of investigation resorts to a statistical testing phase where various hypothesis tests are formulated and analyzed by the actuary. By doing so, there is gain in two dimensions:

- Univariate knowledge of the portfolio under study;
- Multivariate knowledge of the portfolio (relationships between the exploratory / explanatory, exploratory / exploratory, explanatory / explanatory variables).

In the following sections, at a first phase, we will present a univariate analysis that consists of an individual study of the portfolio variables, considering a division of these into two large groups:

- Quantitative Variables;
- Qualitative Variables.

In a second phase, we will deep dive into the relationships between the principal and secondary variables (multivariate study of the portfolio).

This last step can be seen as anticipation for what was identified as the third step of the methodology: The Non-Life Insurance Pricing (see chapter (4)).

6.1 *A priori* assessment of the *"BASE_DADOS_N"* variables

From this section onwards, we will assume that the simulated database is provided by a real *XYZ* insurance company. Therefore, all the assumptions made for the database simulation will be forgotten. The main goal of this action corresponds to the replication of a true-life scenario of an insurance company.

Without any loss of generality, we will start with the analysis by taking the database that contains the frequency of claims per policy, i.e., the "*BASE_DADOS_N*" database.

6.1.1 First impression of the portfolio variables

Based on the dataset made available by the insurance company, it is possible to get a first statistical impression of the quantitative and qualitative portfolio variables.

Since the type of analysis to be produced for each type is significantly different, we can split the workflow into two steps:

- Firstly, we will carry about table (6.1), which provides a summary of the basic statistics computed for all the **non-categorical** portfolio variables;
- Secondly, we will consider table (6.2), which provides a frequency count of all **categorical** variables, by each constituent level.

Both steps were performed through R using the *summary()* function, available in *stats* R-package (version 3.3.5) and below, we can find the results for the first approach:

Table 6.1: Continuous variables summary of *BASE_DADOS_N*. **Source:** Authors.

Statistic	N_clm	Exposure	Driv_Age	Driv_Lic	Vehic_Age	Displacement	Horse_Power	Vehic_Val
Min.	0	0.055	19	0.3	0	0	78	2908
1st Q.	0	0.828	34	15.5	4.169	1.591	136.4	12458
Median	0	0.91	45	24.8	8.337	2.213	196.9	21891
Mean	0.03605	0.8504	44.88	24.89	8.415	2.225	197.9	28151
3rd Q.	0	0.965	55	34.4	12.65	2.855	258.4	38746
Max.	4	1	83	59.2	17	4.4	357	125520

Regarding the qualitative variables, the output is given by the following table:

Table 6.2: Categorical variables summary of *BASE_DADOS_N*. **Source:** Authors.

Sex		Region		Marital_Status		Literacy		Fuel		District		Vehic_Type		Brand	
Level	**Count**	**Level**	**Count**	**Level**	**Count**	**Level**	**Count**	**Level**	**Count**	**Level**	**Count**	**Level**	**Count**	**Level**	**Count**
0	200,316	Center	160,000	Married	88,035	(,9]	184,034	Electric	21,776	Beja	29,913	1	124,197	Audi	25,885
1	199,684	North	120,000	Divorced	69,800	(9,12]	159,978	Diesel	139,901	Coimbra	15,958	2	83,490	BMW	26,198
		South	120,000	Single	105,972	Superior	55,988	Gasoline	208,371	Évora	11,886	3	68,633	Citroen	28,939
				Widow	136,193			Hybrid	29,952	Faro	30,104	4	13,694	Fiat	26,945
										Guarda	11,886	5	8,027	Ford	25597
										Leiria	15,944	7	59,946	Jaguar	25,700
										Lisboa	128,098	8	42,013	Mercedes	26,049
										Porto	83,845			Nissan	27,080
										Setúbal	48,097			Opel	27,159
										Vila Real	24,269			Peugeot	28,839
														Porsche	25,856
														Renault	25,818
														Seat	28,171
														Tesla	25,893
														Volkswagen	25871
Sum	**400,000**	-	**400,000**	-	**400,000**	-	**400,000**	-	**400,000**	-	**400,000**	-	**400,000**	-	**400,000**

6.1.2 Outliers, Missing Values and Wrong Values Analyses

Before going into the quantitative exploration of the portfolio variables (as said in the previous section), we first suggest a short analysis of the quality of the information in the database.

In fact, before starting to make a more quantitative approach, one of the first concerns of an actuary is to assess the quality of the data to be worked on. In this sense, the search for missing values, incorrect data and outliers is a natural and relevant step when the database is analyzed for the first time.

6.1.2.1 Missing Values Analysis

Working with the existence of missing values is always a big concern for an actuary. It is natural that when performing detailed work on a database, there is the need to have as much information as possible. This is what allows to obtain results more adjusted to the reality in the study.

In this project, the search for missing values for the categorical variables is immediate. Looking at the table (6.1), we can conclude that there are no missing values.

This assertion is made based on the sum of the absolute frequencies of each of the portfolio categorical variables. Since they all sum up to four hundred thousand records, which corresponds to the total number of policies, we can be confident that the portfolio at the policy level does not present any missing value.

In fact, this analysis can be done at once for all variables. Combining the *summary()* and the *is.na.data.frame()* ([1]) R-functions, it's possible to confirm that for both types of variables, the logical value to test for missing values is always equal to *FALSE* ([2]).

The output from R that shows the non-existence of missing values is summarized in the following table:

Table 6.3: Test of outliers existence in *BASE_DADOS_N*. **Source:** Authors.

	N_Clm	Exposure	Region	Sex	Marital Status	Literacy	Fuel	Vehic_Type	District	Driv_Age	Driv_Lic	Vehic_Age	Displacement	HorsePower	Brand	Vehic_Val
Mode	logical	logical	logical	logical	logical	logical	logical	logical	logical	logical	logical	logical	logical	logical	logical	logical
False	400,000	400,000	400,000	400,000	400,000	400,000	400,000	400,000	400,000	400,000	400,000	400,000	400,000	400,000	400,000	400,000
True	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

6.1.2.2 Wrong Values Analysis

The identification of erroneous values ([3]) is done locally and, in some cases, can be considered as a subjective process.

This topic is transversal to both types of variables. For the qualitative type, this data treatment is done using advanced methods, such as decision trees.

[1] The *is.na.data.frame()* is available in the *base* package of R (version 4.1.0).

[2] The logical value *FALSE* indicates that the function *is.na.data.frame()* does not return any evidence of missing values existence.

[3] With wrong values, we intend to refer to records outputs that may not make sense given the nature of the variables. For example, the presentation of a policy whose number of claims is negative.

For the quantitative variables, the treatment can be as simple as replacing the missing value with the mean/mode/median of the observed values for the given variable. In more extreme cases, where there is a high occurrence of this phenomenon, it is deemed necessary to eliminate these records or even exclude the entire variable.

In this particular study, considering the summary of table (6.1), it was assumed subjectively that there are no strange values in the population of each variable.

We can say that there will be no need to use techniques to get around this type of data problem.

For the categorical variables, all the levels presented can be considered possible to occur, given the nature of each variable. Therefore, we can conclude that there are no erroneous values in the database taken.

6.1.2.3 Outliers Analysis

Considering that an *outlier* corresponds to a point in the distribution that falls outside the common pattern of the population, we can start by examining the amplitude intervals of each of the continuous variables in the portfolio.

Recalling once more the table (6.1), all continuous variables do not present large amplitudes, which raises the possibility of the non-existence of *outliers*.

However, given the nature of this project, this more crude conclusion without consistent support is not enough. Therefore, we do not want to make this decision hastily without first carrying out a more rigorous analysis that can assess the presence of *outliers*.

A *boxplot* is often used when assessing this type of data quality. Essentially, it can present a visual output that aggregates each variable distribution into four large groups, called *quartiles*. The population that does not fall into these groups is identified as an *outlier*.

In figure (6.1) below presented, we can find this analysis reflected. We performed for each continuous variable a *boxplot* by using the *boxplot()* function available in the *graphics* R-package (version 4.1.0),

Looking more closely at the overall picture of the quantitative variables, only two of them show a behavior susceptible to the presence of *outliers*:

- The *Exposure* variable that presents points at the left of the *boxplot* left whisker ([4]).
- The *Vehicle Value* presents points at the right of the *boxplot* right whisker ([5]).

As in both cases, the points that fall outside the most populated range of each of the variables are theoretically considered as *outliers*. For these cases, the need of identify and quantify them was felt, and it was obtained through a deeper analysis

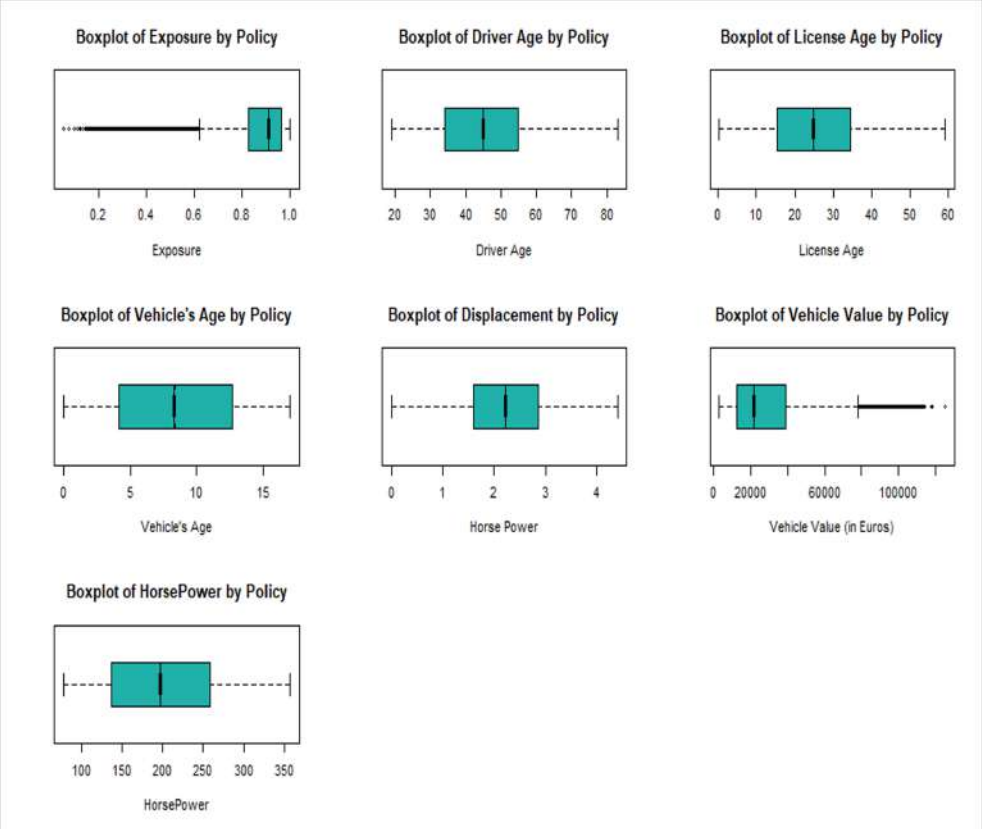
[4]The left whisker of the *boxplot* indicates that the observations have a distance smaller or equal than the metric given by: $1^{st}Quartile - 1.5 \times Inter - QuartileRange$.

[5]The right whisker of the *boxplot* indicates that the observations have a distance higher or equal than the metric given by: $3^{rd}Quartile + 1.5 \times Inter - QuartileRange$.

in R. Using the *boxplot.stats()* R-function, available in *grDevices* package (version 4.1.0), it was possible to identify exactly which points were considered as outliers, and subsequently count them.

The results of these analyses are presented in the figure below:

Figure 6.1: Portfolio Outliers Assessment. **Source:** Authors.



From the outputs (6.1) it is possible to conclude that:

- The *Exposure* variable accounts for 58,040 outliers (about 15% of the total of policies). These records can be explained as the policies with the least exposure time in the portfolio over the year under consideration;
- The *Vehic_Value* variable accounts for 12,023 outliers (around 3% of the total of policies), which can be interpreted as the most expensive vehicles in the portfolio. Consequently, these records can be identified as the policies that can give rise to the most expensive claims.

Although these extreme observations are significantly distant from the core of the remaining distribution of each variable, the action determined for this project took a more conservative perspective, meaning that all of the *outliers* will be maintained.

Since we want to apply a reinsurance treaty to this portfolio, the decision will preserve the policies that can originate regular or severe claims.

After carrying out this first qualitative analysis of the data in general, we can return to the initial plan of evaluating the qualitative and quantitative variables.

In the following sections, we will start with the qualitative variable analyses, and then we will move on to the quantitative variables.

6.1.3 Qualitative Variable Analysis

Generally, the assessment of categorical variables can be more complex to be conducted when compared to quantitative variables.

This difficulty is associated with the univariate analysis in terms of quality and materiality and when understanding the interactions with other variables.

Due to the impossibility of calculating correlation coefficients for qualitative variables, the alternative for constructing relationship indicators is the analysis of contingency tables. Then, based on the observed frequencies, it is possible to build tabular elements that put the variables "face to face" in terms of frequency.

This section will only perform a univariate analysis of the variables, with a commitment to proceed with a multivariate analysis later. This last step will allow us to understand the interactions between the portfolio and response variables (Number of Claims and Claims Amount).

6.1.3.1 Variables Assessment

Proceeding to the univariate analysis, we can start with the "*Sex*" variable.

Belonging to the binary type, it takes two distinct possible outputs:

- *0*, if the policyholder is male;
- *1*, if the policyholder is female.

In the following table, we present the absolute and relative frequencies for each constituent level of this variable:

Table 6.4: Absolute and Relative frequencies of *Sex* Variable. **Source:** Authors.

Sex	0	1
# of Population	200,316	199,684
% of Population	50.1%	49.9%

The table (6.4), shows that the study population is equally distributed in terms of sex. Approximately 50% of the population are males, while the remaining 50% are females.

Focusing on the variable *Region*, it is presented in the database as a categorical divided into three levels:

- *Center*, if the policyholder belongs to the Center of Portugal;
- *North*, if the policyholder belongs to the North of Portugal;
- *South*, if the policyholder belongs to the South of Portugal.

Out of these three output possibilities the category correspondent to the *Center* region stands out slightly (with a percentage of 40%) from the two remaining regions (30%).

Below, it's possible to find the frequency distribution as well as its weight on the total of the variable *Region*:

Table 6.5: Absolute and Relative frequencies of *Region* Variable. **Source:** Authors.

Region	Center	North	South
# of Population	160,000	120,000	120,000
% of Population	40%	30%	30%

Regarding the *"Marital_Status"* variable, it's a categorical variable divided into four levels, split by:

- *Married*, if the policyholder presents a married marital status;
- *Divorced*, if the policyholder presents a divorced marital status;
- *Single*, if the policyholder presents a single marital status;
- *Widow*, if the policyholder presents a widow marital status.

Below, we can find a table that presents a relative and absolute frequencies of the results breakdown by each marital status level:

Table 6.6: Absolute and Relative frequencies of *Marital Status* Variable. **Source**: Authors.

Marital_Status	Married	Divorced	Single	Widow
# of Population	88,035	69,800	105,972	136,193
% of Population	22%	17%	26%	34%

Looking at the results presented, it is possible to conclude that the population is more or less equally distributed along with its categories. However, we can still verify that the level with the highest percentage corresponds to the *"Widow"* class (about 34% of the population).

The *"Literacy"* variable is classified as a particular qualitative variable. Since assigning an order to each constituent level is possible, this variable belongs to the ordinal type.

In terms of possible outputs are divided into three different variables:

- *(,9]*, if the policyholder presents a scholarship between the first and the ninth class;
- *(9,12]*, if the policyholder presents a scholarship between the tenth and the twelfth class;
- *Superior*, if the policyholder presents college or higher graduation.

Below, we can find a table that presents a relative and absolute frequencies of the results breakdown by each literacy level:

Table 6.7: Absolute and Relative frequencies of *Literacy* variable. **Source:** Authors.

Literacy	(,9]	(9,12]	Superior
# of Population	184,034	159,978	55,988
% of Population	46%	40%	14%

This variable is completely dominated by a population with the least advanced levels of education. It accounts for 86% of the population allocated to levels *"(,9]"* and *"(9,12]"*. The *"Superior"* level, only registered 14% of the population under study. The results can be addressed in the table below:

The *"Fuel"* variable is split into four different levels and it is related to the type of fuel of each insured vehicle:

- *Electric*, if the vehicle is powered by electricity;
- *Diesel*, if the vehicle is powered by diesel;
- *Gasoline*, if the vehicle is powered by gasoline.
- *Hybrid*, if a mixed fuel system powers the vehicle.

The following table summarizes these findings, in relation to the *Fuel* variable.

Table 6.8: Absolute and Relative frequencies of *Fuel* Variable. Source: Authors.

Fuel	Electric	Diesel	Gasoline	Hybrid
# of Population	21,776	13,9901	208,371	29,952
% of Population	5%	35%	52%	7%

Facing the results, it is possible to highlight the behavior of two levels in particular. In this case, the policies are massively concentrated in the categorical levels *Diesel* and *Gasoline*, with a percentage of 87%.

The variable *"Vehic_Type"* it is a categorical variable with seven levels. Of these, the levels *1*, *2* and *3* stand out as the top three, with the first two accounting for more than half of the population (52%).

Below, we can find the summary table for this variable.

Table 6.9: Absolute and Relative frequencies of *Vehic_Type* variable. **Source:** Authors.

Vehic_Type	1	2	3	4	5	7	8
# of Population	124,197	83,490	68,633	13,694	8,027	59,946	42,013
% of Population	31%	21%	17%	3%	2%	15%	11%

The *District* variable is also a categorical type of variable, and it can be split into ten levels. We highlight *Lisbon* and *Porto* which are the two districts with the highest number of policies in the portfolio. The remaining levels present a low frequency when compared with these two levels.

Focusing on the less populated cities, it is possible to verify a homogeneous distribution between the levels (around 3% and 8%).

The summary table can be found below:

Table 6.10: Absolute and Relative frequencies of *District* variable. **Source**: Authors.

District	Beja	Coimbra	Evora	Faro	Guarda	Leiria	Lisboa	Porto	Setubal	Vila Real
# of Population	29,913	15,958	11,886	30,104	11,886	15,944	128,098	83,845	48,097	24,269
% of Population	7%	4%	3%	8%	3%	4%	32%	21%	12%	6%

Finally, the last categorical variable of the portfolio under study corresponds to the variable "*Brand*", which contains fifteen distinct levels.

This variable stands out for its uniformity in terms of frequency along all of its categorical levels (between 6% and 7%).

Below, we can find its summary table:

Table 6.11: Absolute and Relative frequencies of *Brand* Variable. **Source**: Authors.

Brand	Audi	BMW	Citroen	Fiat	Ford	Jaguar	Mercedes	Nissan	Opel	Peugeot	Porsche	Renault	Seat	Tesla	Volkswagen
# of Population	25,885	26,198	28,939	26,945	25,597	25,700	26,049	27,080	27,159	28,839	25,856	25,818	28,171	25,893	25,871
% of Population	6%	7%	7%	7%	6%	6%	7%	7%	7%	7%	6%	6%	7%	6%	6%

6.1.4 Quantitative Variable Analysis

The study of the quantitative variables has already started in section (6.1.2.3). In fact, The visualization through *boxplots* is beneficial for a first quality data check.

However, when looking at the information presented in these charts, there is a relevant loss in terms of each variable distribution/density. This is mainly because the data is hidden behind an aggregation into *quartiles*.

To make up for this deficiency, a general view of the distribution of each quantitative variable in the database was performed. As such, instead of visualizing *boxplots*, it was more useful to have a look on alternative plots, such as histograms, density plots and *violin* plots.

In the next figure (6.2), the visuals presented result from a combination of histograms and *violin* plots that were performed by combining two principal R-functions:

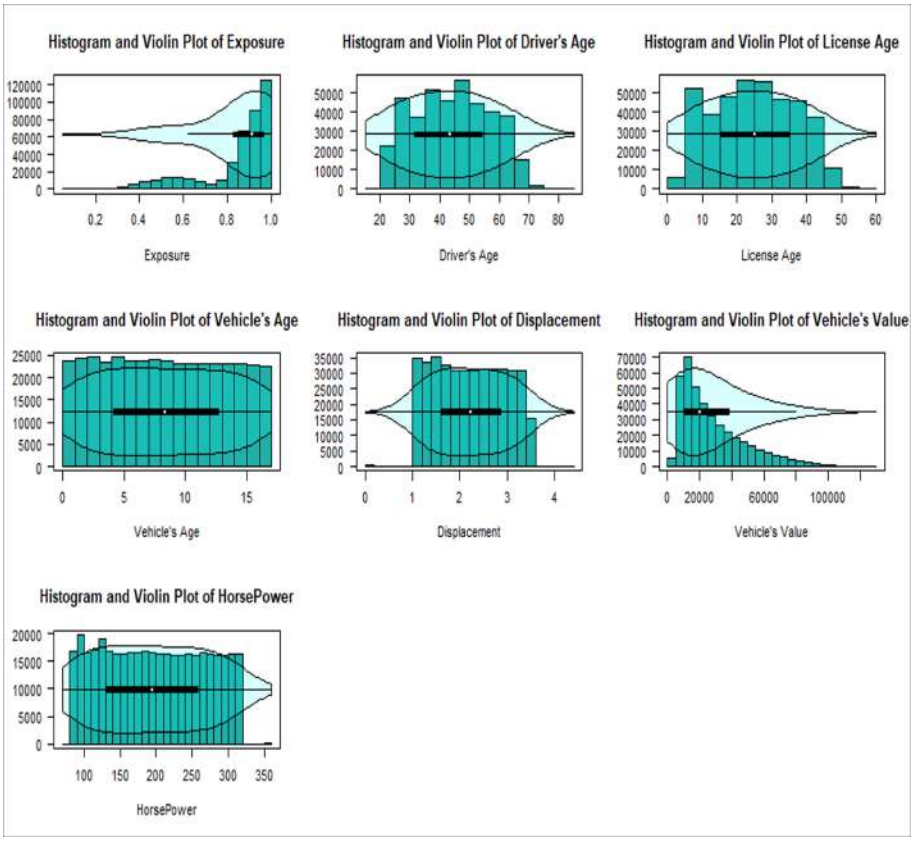
- The *hist()* function, available in the *graphics* R-package (version 4.1.0);
- The *violplot()* function, available in the *violplot* R-package (version 4.1.0)

Given the results obtained from each plot of figure (6.2), it's possible to draw some important conclusions for each of the seven variables presented:

- The histogram of the **vehicles' age** and the **vehicles' horsepower** distributions show us that the portfolio has the population uniformly ([6]) distributed along the various bins considered for the construction of each graph variable. This means that no groups or patterns can be highlighted for these two variables.
- The **vehicles' displacement** distribution is almost similar to the remaining variables. Although the whole policyholders' population is equally concentrated in the intermediate value range of this variable, the only difference is that there is a tiny group with a displacement value equal to zero. This group will have to be analyzed in further chapters. As a first approach, it can be inferred that this group may correspond to the electric vehicles that are already known to exist, given the initial picture of the categorical variable: *Displacement*.

[6]The uniform concept of the sentence should not be taken as the statistical concept of uniform distribution, but rather in its wider sense of the word, where it means having a constant frequency between bins.

Figure 6.2: *Violin* Plots and Histograms of the Portfolio Continuous Variables. **Source:** Author.



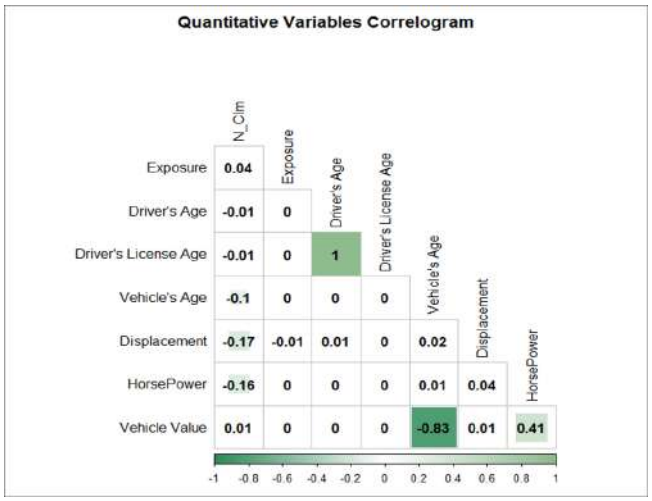
- The **driver's license age** age histogram is similar to the driver's age representation. This conclusion may give rise to some correlation between both variables.
- The **claims exposure** variable reflects a significantly different behavior from the variables pointed above. In fact, it shows a high concentration towards the highest values of its density probability domain. With this chart, there is no doubt that we are dealing with a left-skewed variable.
  In terms of business, this characteristic may be interpreted as the significant predominance of policies exposed throughout the whole year.
- The **vehicles' value** variable also presents a peculiar distribution. In this case, most observations are concentrated in lower values of the variable. It means that clients of an OD insurance coverage opt for cheaper vehicles. However, this conclusion may be misleading since the tail on the right appears to be heavy.

In line with the previous discussion, more than evaluating the individual behavior, it is also essential to assess the interactions between them and the remaining portfolio variables. For this purpose, in figure (6.3), we present a correlogram consisting of a tabular representation of the correlations observed for the whole portfolio at policy level.

Once again, this correlation matrix was obtained from the R tool by using the

version 0.92 of the *corrplot* package.

Figure 6.3: Correlogram of the Portfolio Quantitative Variables. **Source:** Authors.



In general, when addressing the results presented in figure (6.3) we can see no significant relationships between the variables. However, it is essential to point out some exceptions, such as the perfect and positive correlation between driver's age and years of license variables. This conclusion corroborates our aforementioned hypothesis about the possibility of the existence of a correlation behavior between these two variables.

Still facing the output of the figure (6.3), we can note that the *Vehicle's Value* is also correlated with two other variables of the portfolio: The *Vehicle's Age* and the *HorsePower*.

In the case of *Vehicle's Age*, the correlation coefficient indicates a negative correlation. That means that older cars are cheaper. This output reflects the expected age depreciation factor commonly observed in the automobile industry sector.

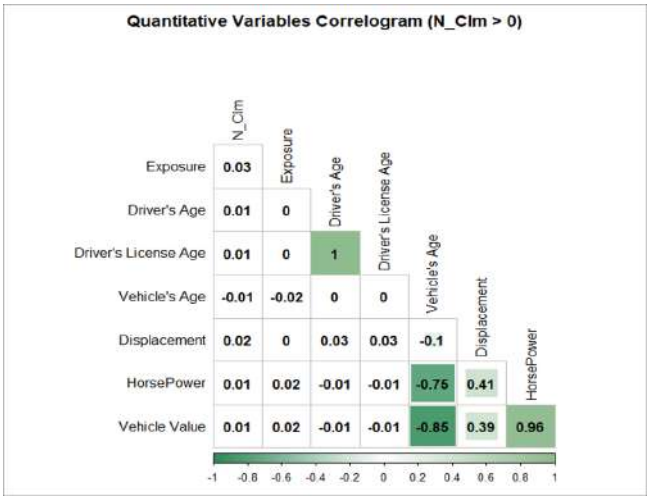
Regarding the *HorsePower* relationship, it is possible to conclude a positive correlation coefficient, which ([7]) indicates the opposite scenario. I.e., cars with higher horsepower are associated with the more expensive vehicle values.

Furthermore, given the proximity to zero of the remaining correlation coefficients, we do not consider as relevant the results obtained for each pair of the remaining variables. Nevertheless, we know that this scenario can change substantially when a filter is applied to variables of particular interest.

Let us take the following case in particular. Assuming that we are interested in evaluating the relationship of the variables considering only those policies that generated at least one claim. I.e., in cases where *N_Clm* is strictly greater than zero, the output for the correlation table is given by figure (6.4).

[7] The correlation outputs are automatically computed as Pearson coefficients computed in R.

Figure 6.4: Correlogram of a subset of the Portfolio Quantitative Variables. **Source:** Authors.



It is worth noting that the correlation between the variables HorsePower/Vechle's Age, HorsePower/Displacement and Vehicle Value/Displacement increased significantly, going from a scenario with no correlation (presented in the correlogram (6.3)) to a different one, in which they are negatively correlated in the first case and positively correlated in the others.

With this example, it is possible to understand that depending on the scenario to be analyzed, and it is necessary to consider an additional caution with the variable relationships. This care will minimize the bias when performing the modeling phase of the variables in the study.

Anticipating the next step, an important aspect to highlight from these two figures ((6.3) and (6.4)) is related to the interaction between the number of claims and the other quantitative portfolio variables. As it can be seen in both representations, the correlation coefficients are very close to zero, which indicates that possibly, these variables may not have much relation with the variable Number of Claims (N_Clm), which will be taken as the response in one of the processes to be modeled ([8]).

The following two sections will focus our analysis on disassembling each of these principal variables, the claims number N_Clm and the claim amount Ind_Clm.

Initially, the process will start by exploring the number of claims, identified as N_Clm (see chapter (5)) and then we will move to the exploitation X that underlies the information of each claim cost (named as Ind_Clm in chapter (5)).

At this stage, these are the two exploratory variables that will allow us to assemble the models for the tariff construction:

- The exploitation of N will enable the claims frequency model formulation ([9]);

[8]This conclusion is only to be seen as a first analysis against the results, which can obviously be changed in the course of the more detailed study.

[9]In this model the variable exposure, identified as "Exposure" (see chapter 5), is also determinant

- The exploitation of X will be crucial to assemble the claims severity model.

6.2 Exploring the Number of Claims, N

Exploring directly the claims frequency process is not an easy job. As presented in equation (3.8) in section (3.3.1), this variable corresponds to the ratio between the number of claims, N, and the time exposure of each policy, $Exposure$, that in conceptual and statistical terms are completely different from each other.

Facing this, an initial good solution is to restrict our exploratory analysis to the numerator which consists on the number of claims registered in the portfolio for the own damage type of product, N. Afterwards, we can move gradually to the claims frequency definition ([10]).

For the study of N, we can calculate at an initial stage the basic statistics of this variable and represent its behaviour in graphical terms.

Therefore, in the table below we can find the statistics of N, performed through R using the *summary()* function, available in *stats* R-package (version 3.3.5):

Table 6.12: Basic Statistics of N. Source: Authors.

Basic Statistics of N								
Min	Max	1st Qua.	Median	Mean	3rd Qua.	Skew	Kurtosis	Variance
0.00	4.00	0.00	0.00	0.036	0.00	6.24	49.03	0.041

As we can see, the minimum number of claims registered in the portfolio is equal to zero, while the maximum value for this variable is equal to four. With this, we can conclude that N, as an integer variable, belongs to the discrete range of $\{0, 1, 2, 3, 4\}$.

From the statistics presented in table (6.12), the 1^{st} *quartile*, the mean and the 3^{rd} *quartile* are equal to zero or close to this value. This fact allows us to conclude that the great majority of the records belonging to N are concentrated around zero. Therefore, there is the possibility of N following a zero inflated distribution.

The previous conclusion for N is also reflected into two different dimensions:

1. The average number of claims in the portfolio is very low. With a percentage value of 3.6%, we can infer that on an average basis there is only a record of $14,200$ claims out of the $400,000$ observed policies (during the year under review).
2. The positive value of 6.24 for the *skewness* parameter indicates a high concentration of the records in the left tail of the distribution of N;

With all of these conclusions, an interesting step to be performed by the actuary consists of drawing the distribution of N.

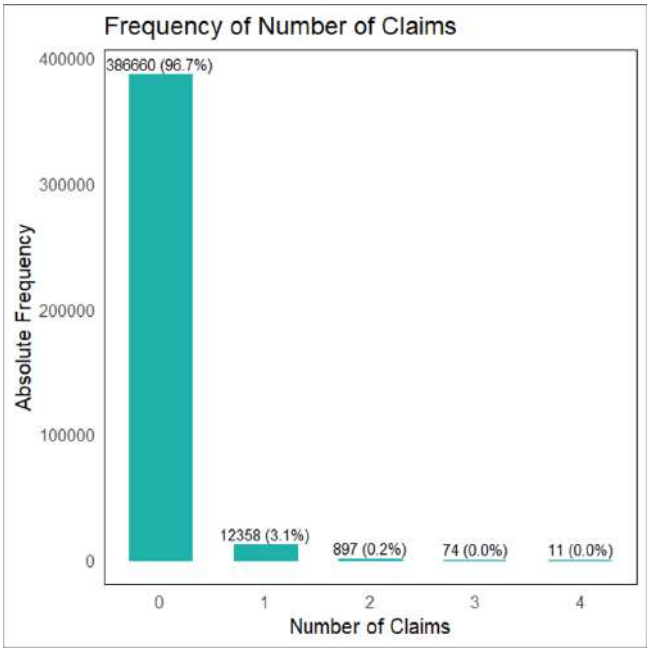
to build the frequency model. In this particular study, this variable is considered as an offset when applying the GLM to N. Consequently N is considered as the principal variable used to model the claims Frequency model.

[10] In our specific study with time equaling to one year, we can just denote as N.

For this, an easy step to take in the first instance is to produce a graph of absolute frequencies. With this the actuary will be more conscious about the behaviour/shape of the distribution of N.

The barplot presented in the following figure (6.5) corresponds to the distribution of N that was performed in R using the *ggplot()* function available in the *ggplot2* R-package (version 3.3.5):

Figure 6.5: Frequency of number of claims. **Source:** Authors.



Considering the output of this frequency bar plot, the characteristics of N are clearer:

- N follows a count distribution;
- N is a zero inflated distribution (97% of the portfolio policies are concentrated in the zero claims bucket);
- N is a right-skewed distribution.

Before starting "guessing" what is the effective behaviour of N, i.e., before trying to fit a distribution to this variable, the actuary can also develop initial indicators that can extract important information.

An example of these possible indicators is the comparison between the average and the variance of the number of claims observed in the portfolio.

In theoretical terms, if N follows a Poisson distribution, we expect to verify an equality between both statistics, meaning that the variance and the mean of N would be very close to each other. On the other hand, if there is an inequality between both measures, namely if the variance of the number of claims is higher than its mean,

there is the possibility of N following a Negative Binomial distribution. This fact is explained due to the presence of the overdispersion factor.

The statistical results presented in table (6.12), shows that the variance of N is higher than its mean, suggesting the presence of great variability in the distribution of N.

6.2.1 Fitting the Distribution of N

After conducting a preliminary statistical analysis of N, a crucial step for the prosecution of our study corresponds to fit a distribution to the number of claims reported to the insurance company.

As usual, to accomplish this need the actuary starts to formulate a statistical hypothesis test. For this, there is the need of defining the null hypothesis and consequently the alternative hypothesis to be tested, H_0 and H_1, respectively.

In this particular case-study, we will follow Carsey and Harden (2013), by starting the testing step with two different null assumptions:

1. H_0: N follows a Poisson distribution;
2. H_0: N follows a Negative Binomial distribution.

The aforementioned tests were performed through the *vcd* R-package, version 1.4-2. This package has the function *goodfit* available to perform a Chi-Squared goodness-of-fit test, that allows the user to choose what is the preferred method to estimate the distribution parameters. In this particular case, the choice relied on the Maximum Likelihood method and the results are summarized in the following table:

Table 6.13: Poisson and Negative Binomial Fitting Tests for N. **Source:** Authors.

Goodness-of-fit test for:	P(> χ^2)	df	Chi-Squared Statistic
1. H_0: N follows a Poisson distribution	0	3	1617.864
2. H_0: N follows a Negative Binomial distribution	0.1795413	2	3.4347

Facing the results presented in table (6.13), it's possible:

- To reject the null hypothesis of N following a Poisson distribution.
- Not to reject the null hypothesis of N following a Negative Binomial distribution.

These decisions were made by taking a level of confidence of 95% (i.e, a level of significance of 5%).

6.2.1.1 Observed vs. Fitted frequencies of N

According to the tests presented in table (6.13) of section (6.2.1), N is more likely to follow a Negative Binomial distribution than a Poisson distribution. A fact that is corroborated by the closeness between the observed and the fitted frequencies of the tests performed.

In our analysis, we found interest in plotting the results of this comparison. We used the *plot()* function available in the *base* R-package to have more detailed information about the observed and the fitted objects of both distributions.

In figure (6.6), we can find a set of three graphs that enables us to answer what we proposed in the previous paragraph. In this particular case, the plots correspond only to the Poisson scenario.

As a term of comparison, we decided to replicate the same visualizations considering a Negative Binomial distribution (see figure 6.7).

Let us first focus on figure (6.6). The graph at the top left of this figure is named as "Standing Plot" and has the particular feature of plotting the square root of the observed frequency data instead of the raw frequency. By applying this modification to the data, we can get a better sensitivity to the data behavior, especially when the frequency of the bars is minimal (when N is higher than two claims).

The graph at the top right of the figure is named as "Hanging Plot" and it aims to compare the observed and the fitted objects of the Poisson distribution. The difference in this second chart consists of changing the bars' position.

This action allows the actuary to understand how to analyze the gap between the y-axis and the bar levitating or overlapping this same axis. Explaining this further, in this chart, we "glue" the bars to the red dots (fitted values) to understand if the observed population is/isn't close to the fitted data. This particular chart is named as *hanging rootogram*.

Finally, on the bottom left of the figure, we can find the "*Deviation Plot*" graph, which combines both type of observations, the observed and the fitted data. This combination is achieved by considering the difference of both frequencies. I.e., we subtract the fitted frequencies to the observed, and then we plot them.

The analysis of these additional charts allows the actuary to build his/her strategic delineation to assemble a model that can explain this variable.

Along with the results for the Poisson distribution, we also performed the same type of analysis when considering the Negative Binomial distribution (see the results in figure (6.7) ([11]).

[11] In both cases, we used the goodness-of-fit models, when the parameters were estimated by the maximum likelihood method. In a more intensive approach it's also advised to perform the same results by switching the estimation type of approach.

Figure 6.6: Standing, Hanging and Deviation Poisson plots. **Source:** Authors.

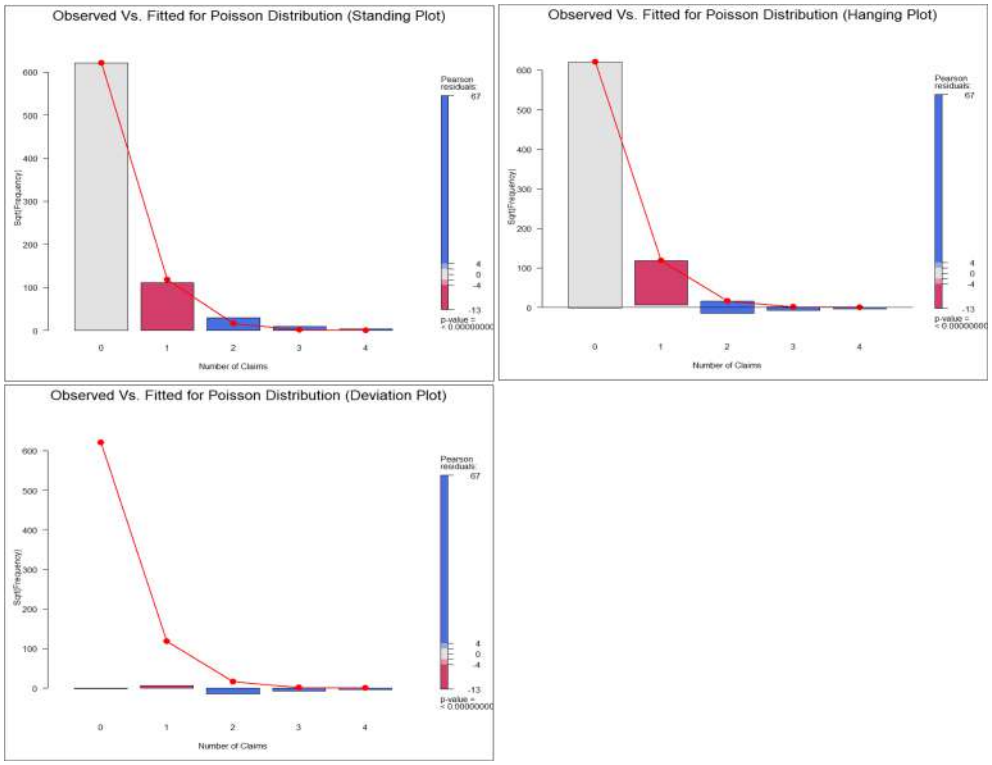
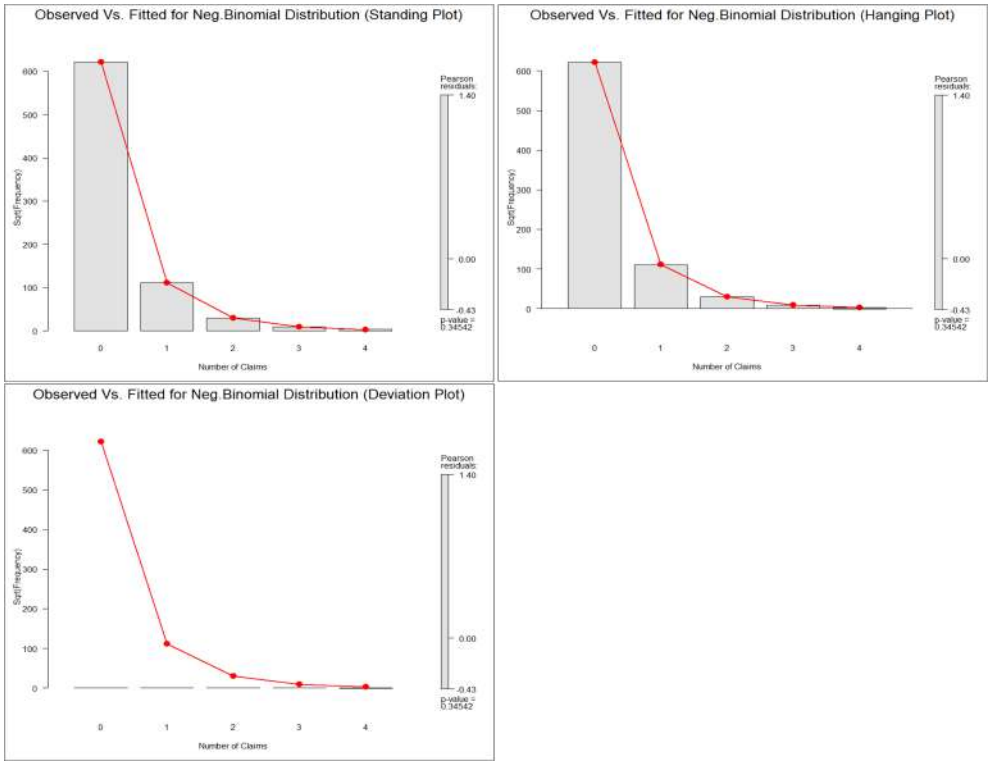


Figure 6.7: Standing, Hanging and Deviation Negative Binomial plots. **Source:** Authors.



From the visualizations presented before, it's possible to extract some interesting aspects. Focusing on the comparison of both *hanging rootograms*, we can see that:

- The Poisson model slightly overestimates the frequency of the number of claims **equaling to one**. This conclusion is given to the fact that the bar of the observed distribution moved **upward** in relation to the reference axis ($y = 0$).
- The Poisson model (figure (6.6)) slightly underestimates the frequency of the number of claims being **greater than one**. This conclusion is obtained by the observable **downward** shift in relation to the reference axis ($y = 0$) of the bars corresponding to the observed distribution.
- The Poisson model presents bars that are shaded with a grade of colours from blue to red. The red coloured bars correspond to a negative contribution to the Pearson chi-square (negative standard Pearson residuals), the blue coloured bars correspond to a positive contribution to the Pearson chi-square (positive standard Pearson residuals).
- The Negative Binomial model (figure (6.7)) seems to estimate accurately the frequency of the number of claims, for each bin considered (number of claims being equal to one to four).

Explaining in a light and simplistic way, to complete our study on the best distribution to be fitted to the variable number of claims, we decided to analyse for each class (i.e. for N=1,2,3,4) the observed deviations from the two distributions that we are considering as possible fits to our distribution (Poisson and Negative Binomial distributions).

To do so, our analyses were performed based on the references presented in Friendly and Meyer (2015) by constructing the two graphs visible in figure (6.8), the "*Poissoness*" and "Negative Binomialness" plots.

At this stage it's important to note that the performance of the red lines presented in the two graphs below were performed based on a Maximum Likelihood estimation for each distribution parameter. I.e, the count metameter was calculated assuming that:

- The mean parameter of the Poisson distribution was estimated from the claim variable distribution through the Maximum Likelihood method.
- The mean and the scale parameter of the Negative Binomial distribution were estimated from the claim variable distribution through the Maximum Likelihood method.

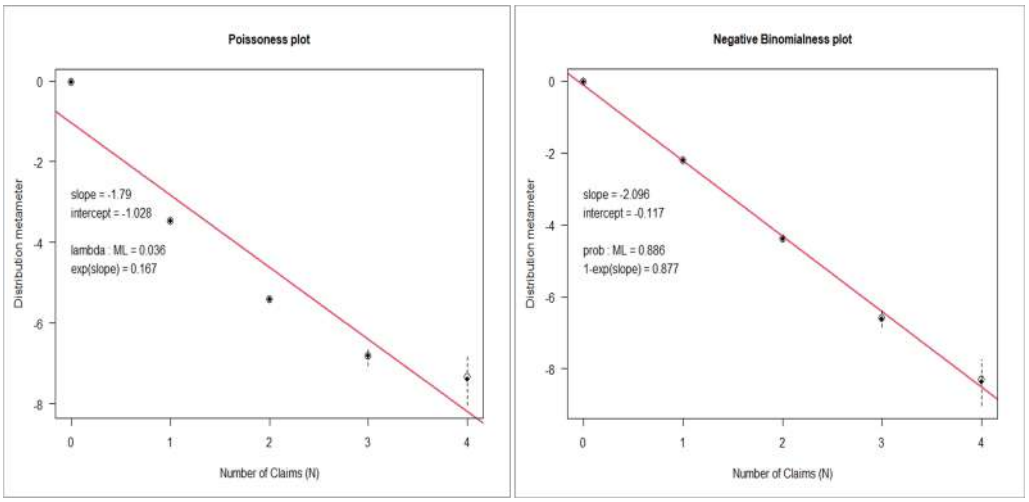
The outputs obtained below were generated through the R-package *vcd* at the expense of the *distplot()* R-function.

From the "Poissoness"plot available in figure (6.8), it's possible to see that the black dots are not on the red line. This means that the points obtained are not "accurately linear" enough, so that the Poisson distribution may be taken as not being the most appropriate to explain the claims data.

In contrast, looking to the right chart of figure (6.8), there is a graphical evidence that all the black points are on the red line. As such, it means that the Negative Binomial distribution can be a better option to fit the claims data.

The conclusion obtained from the goodness-of-fit tests are compliant with the information retrieved from the analyses of the previous charts. By adding all these results together, the actuary can be more certain that it's possible not to reject that N follows a Negative Binomial distribution.

Figure 6.8: *Poissoness* and *Negative Binomialness* plots. **Source:** Authors.



6.3 Exploring the Claims Frequency, F

When building an automobile tariff to a specific portfolio, the knowledge of the claims frequency behaviour reveals to be a critical step.

To achieve this goal, we will continue with the study of the *"BASE_DADOS_N"* database, that contains the aggregated information per policy.

The ultimate reason for this action is justified by the database structure at policy level, which allows a proper modeling of the claims frequency observed for the insured own damage line of business portfolio.

Assessing once more the database, it's noticeable that the claims frequency variable is not directly withdrawable from the benchmark. However, as explained in section (3.3.1), this variable (F) can be built as the ratio between the number of claims, N, and the exposure risk (named as $Exposure$).

By doing it, a first assessment of the new variable was performed and as it was done in the case of the claims number variable, N, the process of analyzing the claims frequency variable will start by presenting its basic statistics (see table (6.14) presented below).

Table 6.14: Summary of Claims Frequency. **Source:** Authors.

Min.	Max	1st Qua.	Median	Mean	3rd Qua.	Skewness	Kurtosis	Variance
0.00	6.54	0.00	0.00	0.042	0.00	6.87	63.01	0.056

The analysis of the claims frequency is undoubtedly different from the one performed regarding the number of claims N.

Starting from the fact that we are dealing with two different type of variables, the number of claims is represented by a discrete random variable, whereas the frequency of claims is transformed into a continuous variable due to the combination of the exposure factor with the number of claims. As such, the latter concept under analysis can be seen as a weighted average over the unit exposure of each policy in the portfolio.

Although it's not very rigorous to analyse both concepts against each other, we allowed ourselves to compare the results obtained in table (6.12) and table (6.14).

As it's possible to verify that the conclusions of F presented in table (6.14) are quite different from those obtained in table (6.12) for N. These differences are especially noticeable because of the change in the statistical maximum, skewness, kurtosis and variance of F. All of those modifications are directly related with the inclusion of the exposure variable in the denominator of the frequency.

However, the final conclusions that can be evidenced are mostly the same as the ones observed in the case of the number of claims:

- The minimum, the maximum, the 1^{st} *quartile*, the median and the 3^{rd} *quartile* are all equal to zero.
- High concentration of the distribution of F at the zero point;
- F is a right-skewed distribution;
- The variance of F is greater than its mean.

Despite this, it's important to identify and quantify the behaviour of the frequency of claims. This was already seen in the literature review that this more general notion (F) will be mainly modeled from the number of claims (N). As such, the search for a distribution that fits the claims frequency is not something the actuary should be worried about.

Nevertheless, in terms of process it doesn't imply that there is no interest in understanding in great detail the behaviour of this variable. In alternative, even if we don't focus directly on the variable distribution, we can extract interesting information indirectly through the multivariate analysis between this variable and the others in the portfolio.

After proceeding with the recognition of the main characteristics, patterns and relationships that may exist, we will continue with the next step of the framework: The calculation of the insurance premiums.

In order to avoid a large extension of this assessment, we have decided to present only some analyses carried out for the multivariate study of the claims frequency.

A very important point to note is that the database investigation varies from actuary to actuary. The same is applicable to the construction of a tariff for determining premiums. It's very difficult for two different actuaries to reach exactly the same type of tariff. More than that, it's quite likely that the same actuary performs different results/analyses with the same data in different moments.

In the next sections, we will illustrate some of the observed results, based on the same strategy adopted so far. I.e., we will separate the process into two parts, taking into account the qualitative and quantitative variables.

6.3.1 Claims Frequency Vs. Qualitative Variables Analysis

In this section we aim to present the relationships between the claims frequency and the categorical variables of the portfolio. With this, the actuary can start understanding which are the variables that may or may not better explain F.

Let's start by looking at the case of *Fuel*. As presented in section (6.1.3), this is a qualitative variable that can be explained by its four categorical levels: *Electric*, *Gasoline*, *Diesel* and *Hybrid*.

Briefly recalling the principal characteristics of this variable, we have that the category with the highest materiality observed corresponds to the *Gasoline* level, which represents fifty-two percent of the total number of policies. The one with the least representation corresponds to the *Electric* vehicle level, which takes only five percent of the total number of policies insured.

Focusing on the relationship detection between *Fuel* and Claims Frequency, the first investigation consists of computing the average number of claims by *Fuel* type. The results are presented in the table below:

Table 6.15: Average Number of Claims by *Fuel* type. **Source:** Authors.

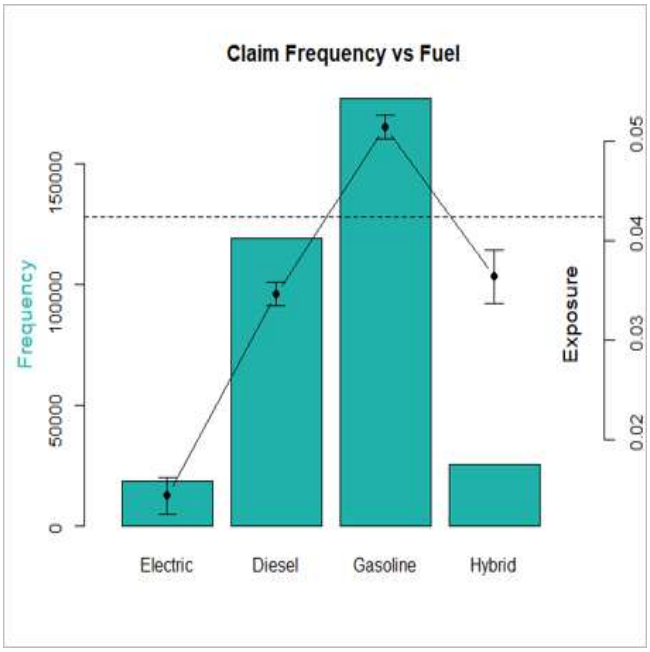
Type of Fuel	Gasoline	Diesel	Electric	Hybrid
Average Number of Claims	5.14%	3.46%	1.44%	3.64%

Taking into account the results presented, it's possible to verify that each category does not constitute a high level of claims frequency risk. The values are all between one percent and five percent, this being an expected scenario, given the high observation of policies with no claims reported over the exposure time.

It's also worth noting that, despite the fact that *Hybrid* vehicles are not one of the most relevant for this portfolio, in terms of materiality is the one that presents the second highest value of claims.

The figure (6.9) shows an efficient representation of the conclusions mentioned, through a graphical visualization of the empirical frequency of claims for each one of the tariff levels belonging to the portfolio categorical variables.

Figure 6.9: Claims risk by type of *Fuel*. **Source:** Authors.



In terms of layout, the bars coloured in sea-green represent the observed frequency of policies associated with each type of fuel. Confidence intervals for the estimated claims frequency are also visible by the whiskers. And finally, the dashed line shows the output corresponding to the portfolio's overall mean claims rate.

Deeply exploring, we can conclude that the *Gasoline* categorical level presents a claims frequency above the observed global average for the portfolio. Furthermore, it's also visible that the two categorical levels *Diesel* and *Hybrid* appear to have a claims frequency very close to each other.

Given this last conclusion, we can formulate a hypothesis test on the equality of the frequency of claims for both categorical factors.

For this purpose, we used the function *wtd.t.test()* available in the *weights* package of the R tool (version 1.0.4), in order to perform a *"Two Sample Weighted T-Test (Welch)"*.

The aim of this test is to evaluate if the frequency of claims observed for the *Diesel* categorical level is the same as the observed for the *Hybrid* categorical level:

$$H_0: \mu_{Diesel} = \mu_{Hybrid}$$

Where,

- μ_{Diesel} sands for the average of claims frequency observed within the categorical value *Diesel* of the *Fuel* variable.
- μ_{Hybrid} stands for the average of claims frequency observed within the categorical value *Hybrid* of the *Fuel* variable.

The output obtained for the computation of the t-test is presented below in table (6.18):

Table 6.16: Output of means t-test. **Source**: Authors.

Two Sample Weighted T-Test (Welch)		
t.value	df	p.value
-1.290485	42841.73908	0.19689

The results indicate that at the five percent significance level, there is the evidence of not rejecting the null hypothesis. This means that it's possible to conclude that the risk observed in *Diesel* and *Hybrid* categorical factors is the same.

Although a test of means is not sufficient to say that the distributions of both categorical levels are equal, it's indeed a good indicator that at least the *Diesel* and the *Hybrid* factors of the *Fuel* variable, present the same type of risk in terms of claims frequency. Given the similarity of the claims behaviour, it's a good motivation for the actuary to combine these two levels into a single one.

Analogously, the same test was performed to the remaining *Fuel* tariff levels and all outputs indicated a rejection of the null hypothesis. Therefore, a merge of more *Fuel* levels is not advisable.

An analysis that is also very important for the knowledge of the interactions between the *Fuel* variable and the claims frequency is the elaboration of a classical linear regression. This way the actuary can estimate the representation of the average number of claims per policy among the various categorical levels of the *Fuel* variable.

In this case, the analysis was once more performed in R, through the usage of the *lm()* function available in the *stats* package (version 3.6.2). The outputs are presented in the following table (6.17).

Table 6.17: Output of *lm()* linear regression application - *Fuel* Variable. **Source:** Authors.

Formula: lm(formula = N_Clm/Exposure ~ Fuel, weights = Exposure)				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.051	0.0005	99.29	<2.00e-16***
Electric	-0.03	0.0016	-22.02	<2.00e-16***
Diesel & Hybrid	-0.016	0.0007	-21.36	<2.00e-16***

To perform this regression we have already joined both *Diesel* and *Hybrid* factors. As a main conclusion, we see that all levels of the *Fuel* variable are very significant for the explanation of the Claims Frequency, which means that in general, the *Fuel* variable is important to detect the behaviour of *F*.

This searching process was also carried out for all other categorical variables in the portfolio. The results are summarized in the following table (6.18).

Table 6.18: Summary of *lm()* Regressions - Categorical Variables. **Source:** Authors.

Variable	Lowest Materiality Frequency Level	Highest Materiality Frequency Level	Lowest Claims Frequency Level	Highest Claims Frequency Level	Levels to be joined	p-value (t-test)
Fuel	Hybrid	Gasoline	Electric	Gasoline	Diesel and Hybrid	0.1968893
Sex	1	0	0	1	None	Always below 5%
Region	South and North (same materiality)	Center	Center	North	None	Always below 5%
Marital Status	Divorced	Widow	Divorced	Widow	None	Always below 5%
Literacy	Superior	(,9]	Superior	(,9]	None	Always below 5%
Vehicle Type	5	1	5	7	2 + 3 3 + 4 2 + 4 7 + 8	0.05324163 0.7458181 0.6830864 0.4728702
District	Evora and Guarda	Lisbon	Setubal	Guarda	None	Always below 5%
Brand	Ford	Citroen and Peugeot	Volkswagen	Citroen	None	Always below 5%

As it's possible to infer from the table (6.27), along with the *Fuel* variable, the t-tests carried out indicate also the aggregation of certain levels of the *Vehicle Type* variable.

Once more, these levels are advised to be viewed as an aggregate level when modeling the frequency of claims.

As we did for the *Fuel* variable, we also performed linear regressions individually for the remaining categorical variables and the results obtained in R are summarized in the following table:

Table 6.19: Summary of P-Values Results - Categorical Variables. **Source:** Authors.

Variable	P-values <0.05	P-values >0.05
Sex		X
Region	X	
Marital_Status	X	
Literacy	X	
Vehicle Type	X	
District	X	
Brand	X	

From the table (6.19), it's possible to conclude that all categorical variables (with the exception of Sex) can significantly explain the claims frequency.

This assertion is made based on the fact that, at the significance level of five percent, all the categories of each variable were considered as significant to explain the behaviour of F.

6.4 Claims Frequency - An *a priori* Tariff

As stated in the literature review, given the independence between the frequency and severity of claims, the actuary can work on the elaboration of the tariff individually for each model. This means that the estimation of the aggregated expected losses can be obtained through the independent modeling of frequency and severity of claims.

In this section, we are completely focused on estimating the claims frequency model. To do so, we will use generalized linear models to estimate the impact of the tariff factors on the expected value of the Claims Frequency.

Similarly to what we have done up to this point, all the results presented in this section are obtained using the R tool.

6.4.1 modeling the Claims Frequency

The usage of generalized linear models is very recurrent when modeling the claims frequency. In this specific case, it's considered a very useful tool, especially because it allows the estimation of its expected value.

This is one of the fundamental steps for this study, namely for estimating the aggregate losses of the insurer and consequently to calculate the premiums to be charged to the insured parties.

Therefore, it's necessary to proceed to the most adequate choice of the model to be used for the modeling process.

According to the literature review, the models that are often used in problems where the exploratory variable is from the integer domain are identified as count models. Out of the possibilities, the Poisson and the Negative Binomial models are included.

Based on the a *priori* analyses performed in section (6.2.1), we have seen that the Negative Binomial distribution is the one that better explains the observed claims number. Although we are aware that the application of the model should not be based only on the fact that the response variable follows a certain distribution, we allowed ourselves to proceed this study with the application of a Negative Binomial count model to the claims frequency.

To reduce the project's length, the realization of a claims frequency tariff will not proceed with a comparative analysis between different models. Namely, the Poisson model, Hurdle models, among others.

However, we advise the actuary to carry out this analysis, as it is essential to compare/evaluate the overall performance of the generalized linear models applied for the construction of the tariff and subsequent premiums calculation.

In this dissertation, the construction of a tariff is only a necessary step towards the ultimate purpose of obtaining an optimal retention level for the reinsurance treaty applied over the portfolio. I.e., fixing the tariff construction process, we will assume that the optimal deductible will correspond to the best monetary value that is possible to achieve, given the premiums allocated to each policy. It's natural that this value is modified when the tariff is adjusted.

Focusing on the Negative Binomial assumption, the first major mission is to ensure that we are in the perfect conditions to proceed with the application of the model. With this, we want to highlight that there may be some variable adjustments before we actually start with the modeling process.

6.4.1.1 Variables Transformation

For the construction of a tariff it's essential to take into consideration several fundamental aspects in terms of data structure. One of the critical points corresponds to the definition of the type of variables to be considered at the modeling phase.

Generally in insurance, it's quite common to observe that only categorical variables are considered as part of the tariff construction, which means, the variables that will explain the response variable. This assumption is duly justified by the need to discriminate each variable into groups that represent the same level of claims frequency risk.

However, the databases do not always present the information completely prepared for the type of study intended to be performed. It is up to the actuary to know the procedures that will be fundamental to adapt to the available reality.

In this sense, given the presence of continuous variables, as is the case observed in the database currently under study, the actuary is advised to resort to variable discretization methods application.

In this dissertation project, the approach taken is quite simple. For each of the continuous variables in the portfolio, n quantiles were subjectively calculated for each variable and used to segregate its distributions into the corresponding n equal or nearly equal parts (called as rate factors of the tariff).

In the following two tables ((6.20) and (6.21)) all the variables subjected to this type of transformation are summarized:

Table 6.20: Summary of Variables Transformations. **Source:** Authors.

Vehic_Val_CAT			Vehic_Age_CAT			Driv_Age_CAT		
Number of Levels: 10			Number of Levels: 8			Number of Levels: 7		
Level	**Absolute Freq.**	**Relative Freq.**	**Level**	**Absolute Freq.**	**Relative Freq.**	**Level**	**Absolute Freq.**	**Relative Freq.**
(2.91e+03,8.32e+03]	40 013	10%	**(-1,2.09]**	50 000	13%	**(18,29]**	58 275	15%
(8.32e+03,1.11e+04]	39 997	10%	**(2.09,4.17]**	50 000	13%	**(29,37]**	63 841	16%
(1.11e+04,1.39e+04]	39 997	10%	**(4.17,6.25]**	50 000	13%	**(37,42]**	55 341	14%
(1.39e+04,1.75e+04]	39 998	10%	**(6.25,8.34]**	50 000	13%	**(42,48]**	59 974	15%
(1.75e+04,2.19e+04]	40 005	10%	**(8.34,10.5]**	50 000	13%	**(48,53]**	51 718	13%
(2.19e+04,2.73e+04]	39 990	10%	**(10.5,12.7]**	50 000	13%	**(53,60]**	55 591	14%
(2.73e+04,3.44e+04]	40 005	10%	**(12.7,14.8]**	50 000	13%	**(60,83]**	55 260	14%
(3.44e+04,4.39e+04]	40 000	10%	**(14.8,17]**	50 000	13%	-	-	-
(4.39e+04,5.82e+04]	39 995	10%	-	-	-	-	-	-
(5.82e+04,1.26e+05]	40 000	10%	-	-	-	-	-	-

Table 6.21: Summary of Variables Transformations. **Source:** Authors.

Driv_Lic_CAT			HorsePower_CAT			Displacement_CAT		
Number of Levels: 6			Number of Levels: 6			Number of Levels: 4		
Level	**Absolute Freq.**	**Relative Freq.**	**Level**	**Absolute Freq.**	**Relative Freq.**	**Level**	**Absolute Freq.**	**Relative Freq.**
(0,10.8]	67203	17%	**(77,118]**	66 667	17%	**(-1,1.59]**	100 000	25%
(10.8,19.3]	67508	17%	**(118,157]**	66 667	17%	**(1.59,2.21]**	100 000	25%
(19.3,24.8]	65763	16%	**(157,197]**	66 666	17%	**(2.21,2.86]**	100 000	25%
(24.8,30.8]	66516	17%	**(197,238]**	66 667	17%	**(2.86,4.4]**	100 000	25%
(30.8,38.3]	67175	17%	**(238,279]**	66 666	17%	-	-	-
(38.3,59.2]	65835	16%	**(279,357]**	66 667	17%	-	-	-

As it's possible to confirm, the six variables present different number of levels. In fact, this was a subjective choice that was made to have all levels uniformly distributed in terms of policy count, without having any notion about the materiality of the claims frequency occurred for each level.

In this way, we ensure that we do not bias the final results of the frequency of claims modeling.

6.4.1.2 Choice of the tariff variables

As mentioned earlier in the previous section, we don't want to observe the collinearity phenomenon in the final model used to explain the claims frequency. Therefore, it's important to identify which variables should be used as a starting point for the modeling study.

When it comes to the analysis of existing relationships between continuous variables, it's enough to resort to the computation of correlation coefficients (for example the Pearson's coefficients) for each pair of the portfolio variables.

Regarding the categorical type, this approach is not applicable. Given the impossibility of calculating correlation coefficients, the identification of collinearity between this type of variables is much more complex.

Because of this complexity, we will only consider categorical variables for the claims frequency modeling, the solution found to mitigate this constraint consisted into performing independence tests on the variables identified as likely to present relationships.

This type of tests evaluate the null hypothesis that the variables of interest are independent. In case of rejection, the conclusion is that there is a relationship between them. Therefore, there is no need to take them simultaneously in the generalized linear model.

Let's take, for example, the evaluation of the independence between the variables *Region* and *Brand*.

To do so, we built a contingency table to perform a chi-square test, assuming the following null hypothesis:

H_0: *Region* and *Brand* are independent variables.

The test begins by performing a simple contingency table between both variables:

Table 6.22: Contingency Table between *Region* and *Brand*. **Source:** Authors.

Contingency Table	Audi	BMW	Citroen	Fiat	Ford	Jaguar	Mercedes	Nissan	Opel	Peugeot	Porsche	Renault	Seat	Tesla	Volkswagen	Total (n_r)
Center	10,416	10,568	11,182	10,550	10,396	10,427	10,612	10,685	10,704	11,226	10,609	10,470	11,212	10,461	10,482	160,000
North	7,688	7,694	8,898	8,190	7,595	7,554	7,642	8,290	8,262	8,946	7,699	7,640	8,442	7,823	7,637	120,000
South	7,781	7,936	8,859	8,205	7,606	7,719	7,795	8,105	8,193	8,667	7,548	7,708	8,517	7,609	7,752	120,000
Total (n_c)	25,885	26,198	28,939	26,945	25,597	25,700	26,049	27,080	27,159	28,839	25,856	25,818	28,171	25,893	25,871	400,000

Based on this count, the next step consists in computing the expected frequency for each pair of Region/Brand levels.

This calculation can be obtained by the following ratio:

$$E_{r,c} = \frac{(n_r \times n_c)}{n}, \tag{6.1}$$

where,

- n_r corresponds to the total by row;
- n_c corresponds to the total by column;

- n corresponds to the global total.

Applying it, results are given by:

Table 6.23: Expected Frequency of Region/Brand variables. **Source:** Authors.

E_rc	Audi	BMW	Citroen	Fiat	Ford	Jaguar	Mercedes	Nissan	Opel	Peugeot	Porsche	Renault	Seat	Tesla	Volkswagen
Center	10354	10479.2	11575.6	10778	10238.8	10280	10419.6	10832	10863.6	11535.6	10342.4	10327.2	11268.4	10357.2	10348.4
North	7765.5	7859.4	8681.7	8083.5	7679.1	7710	7814.7	8124	8147.7	8651.7	7756.8	7745.4	8451.3	7767.9	7761.3
South	7765.5	7859.4	8681.7	8083.5	7679.1	7710	7814.7	8124	8147.7	8651.7	7756.8	7745.4	8451.3	7767.9	7761.3

The next step is crucial and it consists of determining the chi-squared statistic, which is obtained by the following sum:

$$\chi^2 = \sum \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}}, \tag{6.2}$$

where,

- $O_{r,c}$ corresponds to the observed frequency;
- $E_{r,c}$ corresponds to the expected frequency;

The result (χ^2) obtained from the equation (6.4.1.2) is then compared with the critical value of the Chi-square distribution (χ^{*2}) for the correspondent number of degrees of freedom (DF).

The final conclusion is made based on this comparison and if the statistic obtained from the test is greater than the theoretical one, then the hypothesis is rejected, meaning that the variables used do present some type of relationship. Otherwise, it's possible not to reject that the variables are independent.

As it's suggested in the previous paragraph, to finish this test is also necessary to calculate the degrees of freedom. To do so, it's enough to perform the following calculation:

$$DF = (r-1) \times (c-1) \tag{6.3}$$

where,

- r corresponds to the number rows in the contingency table;
- c corresponds to the number columns in the contingency table.

In the specific case we are evaluating, the results of these two last steps are summarized below:

- χ^2 Calculation

Table 6.24: χ^2 Results. **Source:** Authors.

χ^2	Audi	BMW	Citroen	Fiat	Ford	Jaguar	Mercedes	Nissan	Opel	Peugeot	Porsche	Renault	Seat	Tesla	Volkswagen
Center	0.37	0.75	13.38	4.82	2.41	2.10	3.55	1.99	2.34	8.30	6.87	1.97	0.28	1.04	1.72
North	0.77	3.48	5.38	1.40	0.92	3.15	3.81	3.39	1.60	10.01	0.43	1.43	0.01	0.39	1.99
South	0.03	0.74	3.62	1.82	0.69	0.01	0.04	0.04	0.25	0.027	5.62	0.18	0.51	3.25	0.01

- DF Statistic of the Test and Reference Chi-Squared

Table 6.25: Statistic of the Test and Reference Chi-Squared. **Source:** Authors.

χ^2	χ^{*2}	**DF**
107.02	50.99	28

Since χ^2 is greater than χ^{*2}, we are in the condition of rejecting the null hypothesis at a five percent significance level. Therefore, *Region* and *Brand* are not independent.

Based on this conclusion, we have extended the analysis by performing a set of independence tests between all the categorical variables of the portfolio.

As output, the following table (6.26) summarizes all the cases in which there is an evidence of rejecting the null hypothesis, i.e., the pairs of variables for which there is not an independent relation between them.

Table 6.26: Summary of Portfolio Variables Not Independent. **Source:** Authors.

Variable 1	**Variable 2**
District	*Region*
Driv_Lic_CAT	*Driv_Age*
Brand	*Region*
Vehic_Age_CAT	*Region*
Displacement	*Region*
HorsePower_CAT	*Region*
Vehic_Val_CAT	*Vehic_Age_CAT*

Given the finding of no independence between these variables, we conclude that there is an evidence that some of the constituent levels of the pairs of variables (Variable 1 and Variable 2) may be related to each other.

In the model we want to build, it's fundamental that this kind of relationships / patterns do not exist.

To ensure that we will not have problems of robustness, at this stage the major decision is to choose only one of the variables as a model input.

Although it's intuitive to take the choice randomly, it's not very rigorous. The solution lies in the iterative evaluation of an important indicators obtained as output of the model: The Akaike's information criterion (*AIC*). The model that do present the lowest value for this indicators corresponds to the best option to be considered.

Once the used tariff variables are determined, it's still important to define the characteristics of the standard insured.

6.4.1.3 Definition of the Standard Insured

As the name suggests, the standard insurer defines the combination of all the standard scales of the tariff variables.

In order to maintain the robustness of the model, these scales are established by taking into account the categories with the highest representation in the portfolio variables.

In the particular case-study, the standard insured is defined as following:

Table 6.27: Standard Insured of the tariff of N. **Source:** Authors.

Variable	Standard Insured
Region	Center
Sex	0
Marital_Status	Widow
Literacy	(,9]
Fuel	Gasoline
Vehic_Type	1
District	Lisbon
Driv_Age_CAT	(29,37]
Driv_Lic_CAT	(10.8,19.3]
Vehic_Age_CAT	(4.17,6.25]
HorsePower_CAT	(77,118]
Displacement_CAT	(-1,1.59]
Vehic_Val_CAT	(2.91e+03,8.32e+03]
Brand	Citroen

Since the model was built with categorical variables only, it is recalled that the estimates obtained for the remaining tariff levels corresponds to the discounts/surcharges over the estimate obtained for the Standard Insured.

In fact, the Standard Insured characterises the base premium that will be aggravated or discounted, taking into account the risk of each level of the tariff variables.

6.4.1.4 Negative Binomial GLM Application

After carrying out an exhaustive analysis of the portfolio presented in the database *BASE_DADOS_N.txt*, the main objective of this section is to show the results obtained for the claims frequency tariff.

In the first release, whose outputs are presented in table (6.28), the model was performed based on the following assumptions:

- The model ran assuming the number of claims (N_Clm) as the response variable and the Exposure as a model offset.
- The model was ran using the *glm.nb()* R-function available in the *MASS* package (version 7.3-55).
- The model ran with all categorical variables in the database excluding the continuous variables, which were also transformed into categorical variables.
- All the variables are re-leveled such that the *glm.nb()* R-function recognizes the Standard Insured as the intercept of the model.
- The model ran in R, assuming the following input:

```
1 model_full=glm.nb(N_Clm ~ Sex + District + Region + Literacy + Brand + Driv_Lic_
    ↪ CAT + Marital_Status + Vehic_Type + Fuel + Driv_Age_CAT + Vehic_Age_CAT +
    ↪  HorsePower_CAT + Displacement_CAT + offset(log(Exposure)),data=FREQ)
2 summary(model_full)
```

Overall, the results of the first run show that the great majority of the variables can successfully predict the claims frequency. For this majority, the conclusion is drawn based on the fact that the p-values show an output lower than five percent.

Table 6.28: Output of Negative Binominal GLM Application. **Source:** Authors.

	Estimate	**Std.Error**	**z value**	**Pr(>|z|)**	Significance
(Intercept)	3.358039	0.354505	9.472	<2E-16	***
Sex1	0.01	0.054101	0.185	0.85335	
DistrictCoimbra	0.913565	0.220082	4.151	3.3098E-05	***
DistrictEvora	2.152318	0.201077	10.704	<2E-16	***
DistrictFaro	1.158519	0.177615	6.523	6.90748E-11	***
DistrictGuarda	-0.288338	0.283418	-1.017	0.308983	
DistrictLeiria	0.748387	0.220164	3.399	0.000676	***
DistrictLisboa	1.700744	0.157483	10.8	<2E-16	***
DistrictPorto	2.524327	0.16174	15.607	<2E-16	***
DistrictSetubal	2.539429	0.155019	16.381	<2E-16	***
DistrictVilaReal	1.793493	0.188592	9.51	<2E-16	***
RegionNorth	NA	NA	NA	NA	
RegionSouth	NA	NA	NA	NA	
Literacy(9,12]	0.009892	0.059248	0.167	0.867409	
LiteracySuperior	-0.026848	0.084836	-0.316	0.751649	
BrandAudi	-2.160075	0.148343	-14.561	<2E-16	***
BrandBMW	-1.70791	0.136372	-12.524	<2E-16	***
BrandFiat	0.757769	0.108462	6.986	2.81872E-12	***
BrandFord	-4.691395	0.218868	-21.435	<2E-16	***
BrandJaguar	-2.959768	0.171223	-17.286	<2E-16	***
BrandMercedes	-1.810811	0.141331	-12.813	<2E-16	***
BrandNissan	-0.886528	0.115413	-7.681	1.5739E-14	***
BrandOpel	-0.634208	0.109873	-5.772	7.82435E-09	***
BrandPeugeot	0.460893	0.105892	4.352	1.34599E-05	***
BrandPorsche	-3.030403	0.177222	-17.099	<2E-16	***
BrandRenault	-4.731163	0.213151	-22.196	<2E-16	***
BrandSeat	-0.311935	0.107078	-2.913	0.003578	**
BrandTesla	-4.047822	0.192088	-21.073	<2E-16	***
BrandVolkswagen	-4.920341	0.22915	-21.472	<2E-16	***
Driv_Lic_cat(-0.7,10.8]	-0.046313	0.17397	-0.266	0.790073	
Driv_Lic_cat(19.3,24.8]	-0.006126	0.163396	-0.037	0.970091	
Driv_Lic_cat(24.8,30.8]	0.076196	0.219306	0.347	0.728259	
Driv_Lic_cat(30.8,38.3]	-0.015577	0.264653	-0.059	0.953066	
Driv_Lic_cat(38.3,59.2]	0.067472	0.307907	0.219	0.826547	
Marital_StatusMarried	-0.017167	0.074544	-0.23	0.81786	
Marital_StatusDivorced	0.016314	0.081686	0.2	0.841706	
Marital_StatusSingle	0.016054	0.069805	0.23	0.818108	
Vehic_Type2	-0.329128	0.078607	-4.187	2.82677E-05	***
Vehic_Type3	-0.301923	0.088382	-3.416	0.000635	***
Vehic_Type4	-0.224182	0.162104	-1.383	0.166681	
Vehic_Type5	-1.099813	0.253904	-4.332	1.48025E-05	***
Vehic_Type7	0.346279	0.089184	3.883	0.000103	***
Vehic_Type8	0.284087	0.099743	2.848	0.004397	**
FuelElectric	0.822137	0.120477	6.824	8.85342E-12	***
FuelDiesel	-0.094757	0.062058	-1.527	0.126784	
FuelHybrid	0.327879	0.112891	2.904	0.00368	**
Driv_Age_cat(18,29]	0.090626	0.173036	0.524	0.600457	
Driv_Age_cat(37,42]	-0.06901	0.161883	-0.426	0.669892	
Driv_Age_cat(42,48]	-0.113243	0.211928	-0.534	0.593102	
Driv_Age_cat(48,53]	-0.079684	0.250801	-0.318	0.750697	
Driv_Age_cat(53,60]	-0.34537	0.284249	-1.215	0.224355	
Driv_Age_cat(60,83]	-0.898355	0.318701	-2.819	0.00482	**
Vehic_Val_cat(8.32e+03,1.11e+04]	-2.270537	0.209324	-10.847	<2E-16	***
Vehic_Val_cat(1.11e+04,1.39e+04]	-3.120808	0.255794	-12.2	<2E-16	***
Vehic_Val_cat(1.39e+04,1.75e+04]	-4.188671	0.288816	-14.503	<2E-16	***
Vehic_Val_cat(1.75e+04,2.19e+04]	-5.537144	0.323894	-17.096	<2E-16	***
Vehic_Val_cat(2.19e+04,2.73e+04]	-6.177107	0.351439	-17.577	<2E-16	***
Vehic_Val_cat(2.73e+04,3.44e+04]	-7.616743	0.383643	-19.854	<2E-16	***
Vehic_Val_cat(3.44e+04,4.39e+04]	-7.259129	0.410123	-17.7	<2E-16	***
Vehic_Val_cat(4.39e+04,5.82e+04]	-6.310863	0.44499	-14.182	<2E-16	***
Vehic_Val_cat(5.82e+04,1.26e+05]	-3.976943	0.489459	-8.125	4.47E-16	***
Vehic_Age_cat(-1,2.09]	2.265213	0.150985	15.003	<2E-16	***
Vehic_Age_cat(2.09,4.17]	1.161184	0.119169	9.744	<2E-16	***
Vehic_Age_cat(6.25,8.34]	-1.305141	0.141025	-9.255	<2E-16	***
Vehic_Age_cat(8.34,10.5]	-2.931557	0.200305	-14.635	<2E-16	***
Vehic_Age_cat(10.5,12.7]	-5.668797	0.28953	-19.579	<2E-16	***
Vehic_Age_cat(12.7,14.8]	-6.383202	0.333306	-19.151	<2E-16	***
Vehic_Age_cat(14.8,17]	-8.007621	0.378864	-21.136	<2E-16	***
HorsePower_cat(118,157]	-0.435926	0.089524	-4.869	1.11933E-06	***
HorsePower_cat(157,197]	-1.388998	0.128396	-10.818	<2E-16	***
HorsePower_cat(197,238]	-2.139333	0.167732	-12.754	<2E-16	***
HorsePower_cat(238,279]	-2.362101	0.188814	-12.51	<2E-16	***
HorsePower_cat(279,357]	-4.033669	0.208851	-19.314	<2E-16	***
Displacement_cat(1.59,2.21]	-0.685738	0.063462	-10.805	<2E-16	***
Displacement_cat(2.21,2.86]	-6.427997	0.232163	-27.687	<2E-16	***
Displacement_cat(2.86,4.4]	-1.72149	0.073725	-23.35	<2E-16	***

As exception, we point out that the variables corresponding to *Sex*, *Region*, *Literacy*, *Driv_License_CAT*, *Marital_Status* and *Driv_Age_CAT* do not verify it. Therefore, they cannot predict the response variable accurately.

Also from the results, we can conclude that the standard deviations are not very high (close to zero), which is a good indicator of a small variability of the estimates obtained for each level of the explanatory variables.

As a final note, we don't want to overlook the fact that two of the coefficients, corresponding to the variable *Region*, do not present any results. The output obtained shows that the coefficients of *North* and *South* levels were not calculated, due to singularity problems. This means that, these levels may present an exact linear relationship with another variable used for modeling purpose.

This finding emphasises the need already pointed out in this dissertation: All collinear variables should be eliminated from the model.

To address this deficiency, which has already been seen to impact the model results, we have retrieved the topic that generated the table (6.26) presented in section (6.4.1.2).

Recovering the mentioned table, we proceeded with a new run of the *model_full*, assuming the presence/exclusion of the variables identified as collinear.

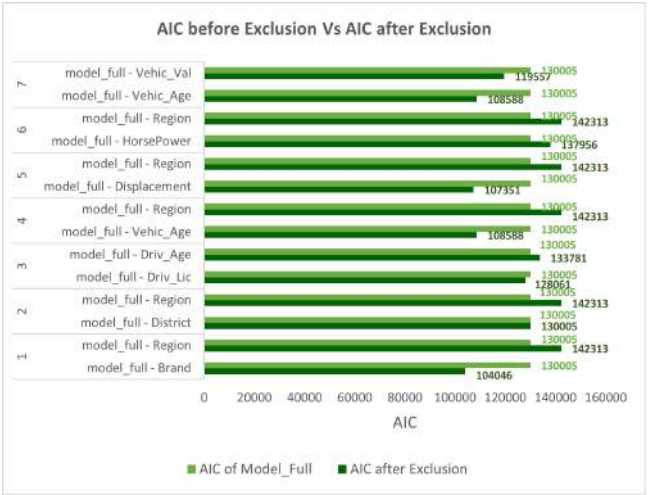
In this context, seven experiments were performed in order to understand what is the impact generated if one of the variables is maintained/excluded in detriment of the collinear variable.

In each experiment, two new runs of the *model_full* were performed assuming:

- In the first new run, the exclusion of one collinear variable;
- In the second new run, the exclusion of the other collinear variable.

The figure (6.10) is intended to summarize the six experiments performed.

Figure 6.10: *AIC* before and after Model Exclusions. **Source:** Authors.



The graph obtained above is quite enlightening in terms of the criterion chosen as decisive to maintain or exclude one of the variables that shows a collinearity behaviour.

In the end, the new run that presents the lowest AIC denounces that the output of this variable improves the discriminatory power of the model.

The following table presents a summary of the results and actions performed based on the outputs of the previous graph.

Table 6.29: *AIC* Variables Comparison. **Source:** Authors.

Experiment	Variable Maintained	Variable Excluded	Reason of Exclusion	Criteria
1	**Region**	**Brand**	After testing the independence Chi-Squared Test, it's possible not to reject that the Region and Brand are not independent.	***AIC* of *Region* < *AIC* of *Brand***
2	**Region**	**District**	After testing the independence Chi-Squared Test, it's possible not to reject that the District and Region are not independent.	***AIC* of Region < *AIC* of District**
3	**Driv_Age_CAT**	**Driv_Lic_CAT**	The variables Driv_Age and Driv_Lic are perfectly correlated.	***AIC* of Driv_Age_CAT < *AIC* of Driv_Lic_CAT**
4	**Region**	**Vehic_Age_CAT**	After testing the independence Chi-Squared Test, it's possible not to reject that the Region and Vehic_Age_CAT are not independent.	***AIC* of Region < *AIC* of Vehic_Age_CAT**
5	**Region**	**Displacement_CAT**	After testing the independence Chi-Squared Test, it's possible not to reject that the Region and Displacement are not independent. Consequently, Region and Displacement_CAT are not independent.	***AIC* of Region < *AIC* of Displacement_CAT**
6	**Region**	**HorsePower_CAT**	After testing the independence Chi-Squared Test, it's possible not to reject that the Region and HorsePower are not independent. Consequently, Region and HorsePower_CAT are not independent.	***AIC* of Region < *AIC* of HorsePower_CAT**
7	**Vehic_Age_CAT**	**Vehic_Value_CAT**	The variables Vehic_Val and Vehic_Age are directly correlated.	***AIC* of Vehic_Age_CAT < *AIC* of Vehic_Value_CAT**

Considering this study, the running model in a first instance underwent several transformations and the variables that were maintained when excluding the collinearity factor are:

Table 6.30: Maintained Variables List. **Source:** Authors.

Variable
Region
Sex
Marital_Status
Literacy
Fuel
Vehic_Type
Driv_Age_CAT
Exposure

It is irrefutable that this type of study is a constant search for a better model. As such, it is fundamental that the final model, i.e., the model taken as the one that best explains the response variable and that best fits the data under study, does not present variables that are not significant for this purpose.

In response to this need, actuaries specialised in this type of analysis developed sophisticated methods that allow the elimination of variables not considered as important to explain the response variable.

One of the methods that is widely used in bibliographical references is the *Stepwise Backwards* model. This method starts from an initial model with a significant number of variables in which is made an evaluation of *p-values* resulting from *Wilks* Maximum Likelihood tests. Based on these outputs, the variables are decided to be removed significant or maintained in the final model.

In practical terms, the method was applied directly in R through the *MASS* R-package (version 7.3-55).

The following table shows the process run in the R project, which highlights all the steps performed by the Stepwise model. Along with this, we also show the evolution of the *AIC* value:

Table 6.31: *Stepwise* Steps for the tariff of N. **Source:** Authors.

Stepwise Step	**Model**	***AIC***
Initial Model	*N_Clm ~ Sex + Region + Literacy + Marital_Status + Vehic_Type + Fuel + Driv_Age_CAT + offset(log(Exposure))*	*121,011*
First Step	*Initial Model - Martial_Status*	*121,006*
Second Step	*Model of First Step - Sex*	*121,004*
Third Step	*Model of Second Step - Literacy*	121,002
Final Model	*Model of Third Step = N_Clm ~ Region + Vehic_Type + Fuel + Driv_Age_CAT+ offset(log(Exposure))*	

The estimates obtained for the ***Final Model*** can be seen in the following table:

Table 6.32: Final Model Estimates. **Source:** Authors.

VariableLevel	Estimate	Std.Error	z value	Pr(>|z|)	Significance
(Intercept)	-3.14874	0.030293	-103.942	<2E-16	***
RegionNorth	0.667871	0.027882	23.954	<2E-16	***
RegionSouth	0.362029	0.027557	13.138	<2E-16	***
Vehic_Type2	-0.34168	0.027095	-12.61	<2E-16	***
Vehic_Type3	-0.41205	0.030815	-13.372	<2E-16	***
Vehic_Type4	-0.33323	0.058384	-5.708	1.15E-08	***
Vehic_Type5	-1.32787	0.133348	-9.958	<2E-16	***
Vehic_Type7	0.090506	0.02601	3.48	0.000502	***
Vehic_Type8	0.081087	0.028879	2.808	0.004988	**
Fuel Electric	-1.04827	0.063983	-16.384	<2E-16	***
FuelDiesel	-0.32917	0.019832	-16.598	<2E-16	***
FuelHybrid	-0.04887	0.037682	-1.297	0.194655	
Driv_Age_CAT(18,29]	0.086164	0.032339	2.664	0.007713	**
Driv_Age_CAT(37,42]	-0.05807	0.032114	-1.808	0.070556	.
Driv_Age_CAT(42,48]	-0.00853	0.032302	-0.264	0.791673	
Driv_Age_CAT(48,53]	-0.04302	0.03307	-1.301	0.193259	
Driv_Age_CAT(53,60]	-0.25429	0.033521	-7.586	3.31E-14	***
Driv_Age_CAT(60,83]	-0.61374	0.037506	-16.364	<2E-16	***

According to the results obtained in table (6.32), it's possible to discriminate that not all factors of the variables *Fuel* and *Driv_Age_CAT* are effectively significant for the model considered. However, we have that the variables *Fuel* and *Dri_Age_CAT* are overall important for the model, which means that after reflection of the results, one should try to aggregate the non-significant factors to another that were identified as such.

For this study, it was decided not to aggregate classes, but this is an advisable action to take in future projects.

In terms of results, it's possible to see that the vast majority of the tariff factors present negative parameter estimates. This indicates that the standard insured is the one who contributes the most in terms of severity risk.

To highlight that, we have only have a few categories that effectively present positive estimated coefficients. Meaning that the risk carried is higher than the observed for the standard insured.

Before ending this section, we would like to point out that by applying this method of eliminating variables, the levels characterising the standard insured were also changed. Thus, in the following table we characterise the standard insured for the Final Model:

Table 6.33: Standard Insured of the Final Model. **Source:** Authors.

Variables	Level
Sex	0
Region	Center
Vehic_Type	1
Fuel	Gasoline
Driv_Age_CAT	(29,37]

6.4.1.5 E[N] Calculation

Once we have chosen the model that best characterises the claims frequency, it's time to estimate its expected value.

There are two ways of achieving this goal. I.e., the way we can set up the tariff structure can be done assuming an additive or multiplicative model. The difference between the two is only in the way in which discounts and surcharges to the standard premium are calculated. While in the additive model these are done in absolute value, in the multiplicative model they are done proportionally.

In this dissertation, we decided to proceed with the multiplicative model. As such, we consider that the expected value of our variable of interest, the frequency of claims is obtained as follows:

$$E[F] = \mu_{ij} = \gamma_0 \gamma_{i1} \gamma_{i2} ... \gamma_{ik} = exp(\beta_0 + \beta_{i1} Y_{i1} + \beta_{i2} Y_{i2} + ... + \beta_{ik} Y_{ik}), \tag{6.4}$$

where,

- The parameter γ_0, corresponds to the estimated claims frequency for policyholders with the characteristics of the standard insurer, identified in table (6.33) in the previous section.

- $\gamma_{i1}\gamma_{i2}...\gamma_{ik}$ represent the estimates of the frequency model for all policyholders not included in the standard policyholder category. These estimates are always calculated on the basis of the standard level.

In the following table (6.34) we present a global summary regarding the tariff structure of the claims frequency.

For each of the levels considered to be part of this tariff, we calculate the expected value of F. We recall that the calculation was made taking into account the estimate obtained for the standard insured party.

Table 6.34: Tariff structure of claims frequency. **Source:** Authors.

	E[N]
(Intercept)	0.042906
Sex1	0.042906
RegionNorth	0.0836702
RegionSouth	0.0616233
Vehic_Type2	0.030488
Vehic_Type3	0.0284164
Vehic_Type4	0.0307466
Vehic_Type5	0.0113719
Vehic_Type7	0.0469704
Vehic_Type8	0.0465301
FuelElectric	0.0150404
FuelDiesel	0.0308716
FuelHybrid	0.0408596
Driv_Age_cat(18,29]	0.0467669
Driv_Age_cat(37,42]	0.0404853
Driv_Age_cat(42,48]	0.0425415
Driv_Age_cat(48,53]	0.0410992
Driv_Age_cat(53,60]	0.0332723
Driv_Age_cat(60,83]	0.0232259

Given the results presented, it can be seen that almost all tariff levels present an expected claims frequency value lower than that of the standard insured. As an exception to the rule, only the *North* and *South* regions present a higher value for this estimate.

6.5 Exploring the Claims Severity

Starting from this section, we will move our study to the analysis of the second database, named as ***BASE_DADOS_X.txt***.

As previously explained in chapter (5), this database presents a granular vision of the insured policies. In practical terms, each row corresponds to an individual record, which may or may not correspond to a claim of each policy. Along with the information about the individual claim amount, the aggregate claims cost per policy is also presented by row.

As mentioned before, the modeling of claims severity should be carried out independently from the modeling of claims frequency. However, comparing with the claim frequency process, modeling the claims severity proves to be a much more complex process in several dimensions.

This increase of complexity is essentially related to the fact that the amount of claims is a process that is difficult to explain using variables known as explanatory variables. Usually, there are very few variables that can be seen as significant to model the claims severity. And, for this study we add the condition that all driving ability as well as driver characteristics have to be set aside for severity modeling. This decision aims to exclude any kind of subjectivity on the actuary side, that may generate bias in the study.

A second dimension that should be considered is the need to segregate individual indemnities arising from policyholders. This split aims to provide us at least two large groups of claims: The regular claims and severe claims.

It's not always very easy to make this division. As has been noted, we are very dependent on the data we have in our possession and in the face of this, we have to come up with well-structured strategies that allow us to effectively monitor behaviors, patterns and trends in the data.

With regard to the recognition of identical patterns between observations of variables, the methods that are most used nowadays are the clustering models. Among this large group, the *k-means* model stands out, which will be the model selected to perform our clustering research.

Focusing on its construction, the application of this model has a number of assumptions that have to be met. One of them is that the model only "accepts" working with continuous variables, due to the need of calculating distances between observations.

Thus, combining this restriction with the fact that we are not considering the characteristics of the driver to explain the severity process, we are left with a limited set of variables to be used. The table (6.35) summarizes the list of variables considered as in conditions to be used for modeling purpose.

In addition to these variables, we also know that sometimes the *Vehic_Type*, the *Region* and/or the *District* of the policyholder can be considered as a discriminatory information to determine the rate associated with the severity of the claims. As such,

Table 6.35: Summary of the Input Variables List. **Source:** Authors.

Variable
Vehic_Value
Vehic_Age
Displacement
HorsePower

for the study of the data clusters they will not be included, however, when modeling, these variables should still be considered.

6.5.1 Claims Severity - A Clustering Model application

Generally, the application of the *k-means* method assumes a number of initial and intermediate steps with regard to data cleaning and processing. Steps such as identifying *outliers*, *missing values*, wrong data, data transformation, creating new metrics, variable standardization and many more.

In this project, all these steps were performed and no wrong or missing values were identified. As for the identification of *outliers*, it should be noted that the variable *Vehic_Value* shows the presence of values that are not similar to those observed in the remaining population.

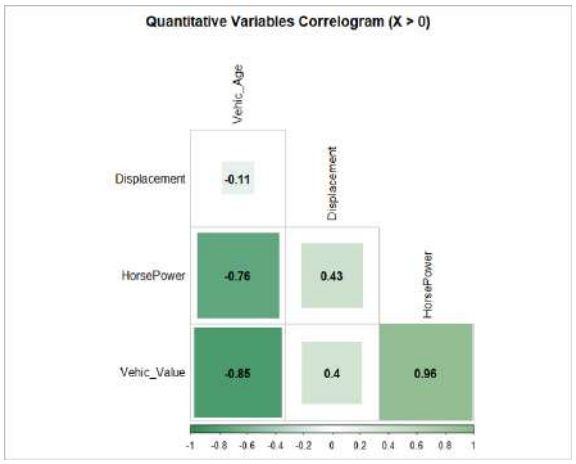
Under penalty of the results being slightly biased, it was decided not to remove these observations from the study.

Another important aspect when applying a clustering model, consists on the evaluation of the relationships between variables. I.e., the assessment of the correlations between the variables is deemed a very important step to be done.

Similarly to what we did in figure (6.3) of section (6.1.4), we compute the (*Pearson's*) correlation coefficients in order to identify the correlated variables in the set presented by table (6.35).

The results are shown in the figure (6.11):

Figure 6.11: Correlogram of Clustering Input Variables. **Source:** Authors.



Given these results, it can be seen that the variables *Vehic_Age*, *HorsePower* and *Vehic_Value* are highly correlated between each other.

We know that there are two types of variables that can cause collinearity problems in clustering algorithms. One of them, corresponds to the presence of irrelevant variables and the other to the presence of redundant variables.

In theoretical terms, the claims severity can be explained at the expense of the age of the vehicle, the power of the vehicle and the cost of a vehicle. As such, they can be considered as relevant variables.

Since we are actually dealing with a redundancy problem, one possible approach to mitigate this problem consists into maintain only one collinear variable.

The major concern that emerges by doing it is reflected on the decision of which variable should be maintained/excluded.

The variable to be retained should be the one considered most useful for the study or the one with the highest potential for action.

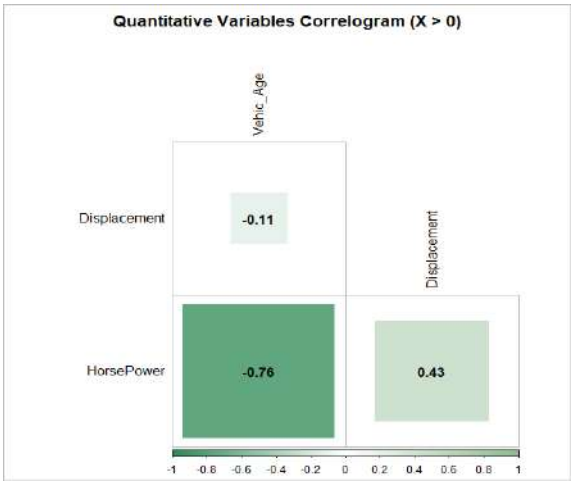
In practical terms, we have taken in figure (6.11), the case that shows the highest correlation Pearson coefficient. Looking to the figure mentioned previously, this case corresponds to the *Vehic_Value vs. HorsePower* that presents a Pearson coefficient value equal to 0.96, which is very close to the perfect correlation relationship.

Adding the expertise factor to decision making, it's common sense that more powerful vehicles have a higher propensity greater propensity to reach higher speeds and cause more damage when an impact occurs.

As such, this variable, *HorsePower*, was considered to be maintained as input for the model. Therefore, the variable *Vehic_Value* was was excluded from the set of variables to be considered for the clustering model.

Applying this action, we performed a new analysis of the correlation coefficients for the variables considered as input. The results are shown in the following correlogram:

Figure 6.12: Correlogram of Clustering subset of Variables. **Source:** Authors.



Similarly to what was done in the previous step, we took the case in which there

is a higher correlation indicator (positive or negative). Based on figure (6.12), it corresponds to the coefficient between the variables *Vehic_Age* and *HorsePower*.

The solution adopted in this step was exactly the same as in the previous step and therefore we kept the variable *HorsePower* as input of the algorithm.

This leaves us in the end with the variables *Displacement* and *HorsePower*. Recalculating the correlation only for these two variables the output is equal to 0.41, which can be considered as no correlation between the two variables.

Thus, we will proceed to build the clustering model using only these two variables. At this stage, it's important to realize that the clustering model is going to be applied by using the claims per policy instead of the individual claims registered in each claim.

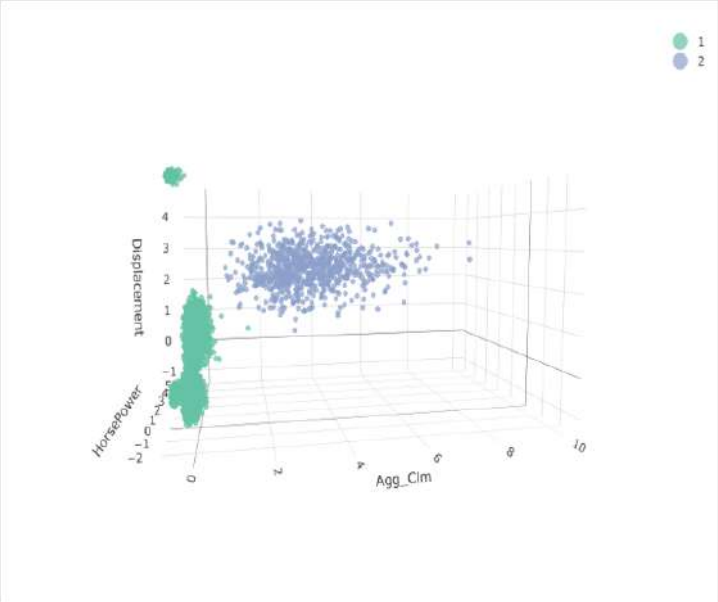
Using the *kmeans()* function available in the *stats* package (version 3.6.2), we were able to apply the clustering model in the desired data.

For this, the only inputs that had to be assigned were the number of *centroids* to be performed, the number of random sets that should be chosen and the data corresponding to the two selected variables (*HorsePower* and *Displacement*).

An important detail is that the data have been standardized using the *scale* R-function available in the *base* package of the R tool.

Given the assumptions used, the following chart shows the segregation of the thirteen thousand three hundred and forty policies that originated claims into two groups presented in the following figure:

Figure 6.13: K-means Clustering output. **Source:** Authors.

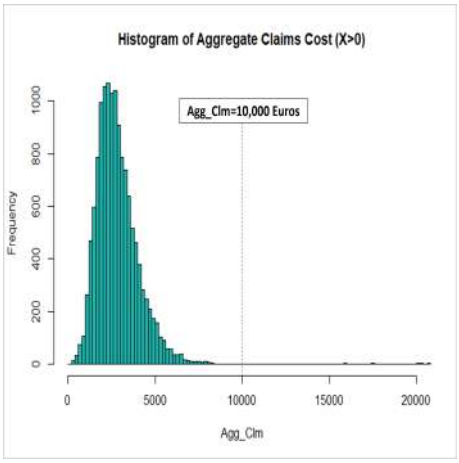


The figure (6.13) shows us that it's indeed possible to separate two large groups for the aggregate claims variable.

In terms of results, it can be seen that the group one, painted in green, corresponds to policies whose aggregate indemnity is less or equal than ten thousand euros. In the

fact, it's not possible to verify this value in the figure due to the standardisation performed in the data used for the model. However, in the next histogram corresponding to the *Agg_Clm*, this assertion becomes more evident:

Figure 6.14: Aggregate Claims Cost Histogram. **Source:** Authors.



All that is painted in blue in figure (6.13) corresponds to the second group which has the particularity of verifying all aggregate claims above ten thousand euros.

In conclusion, what really matters is that in general we were able to identify that the policies that gave rise to claims, can be divided into two large groups:

- The group one, corresponding to **regular** claims;
- The group two corresponding to **severe** claims.

Given this analysis of our portfolio, there are several questions that may emerge. One of them corresponds to the possibility that the number of clusters initially defined is not in fact the most appropriate for our data.

Without going too deep, we proceeded to an analysis that will not be presented, which consists in drawing the *elbow* graph. This plot, evaluates a metric called *Within-Cluster Sum of Square*, that measures the distance between each observation and the *centroid* for several *k's* (number of clusters used as input). The point at which there is not a large variation from $k-1$ to k is defined as the optimal number of clusters to be used for the segregation of the data used. In this case, the optimal output corresponded to six clusters.

However, from the point of view of this dissertation it doesn't make sense to split into so many groups. As such, we kept with the information obtained in the graph presented previously (for k equaling to two).

Facing these conclusions, we can start formulating the overall strategy for determining the claims severity tariff.

Given that we have the aggregate claims separated into two large groups, below and above ten thousand euros, it's of immediate conclusion that the individual claims belonging to group one (below ten thousand euros) are also contained in the group of claims below ten thousand euros.

Perhaps, the greatest concern corresponds to the group identified as having the aggregate indemnities above ten thousand euros, which when split into their individual indemnities may fall into the group identified as one in chart (6.13).

Performing a quick analysis of the cases identified as possible problems, we conclude that there are only sixteen cases, which is not a very significant materiality in terms of changing the distribution for both groups.

6.5.2 modeling Regular Claims

Given the separation of the two types of claims, we have that the random variable identified as X, in section (3.2.1), for the expected loss of the portfolio must be split into two.

The way to proceed to perform this separation corresponds to the simple implementation of the criterion identified for the variable Ind_Clm in the previous section. More specifically, all rows whose individual claim is less than or equal to ten thousand euros will form a new subset of the parent database. The same happens for records whose individual indemnity is greater than ten thousand euros. With this action, we can work on both sub-groups and model them differently.

Let's start by looking at all the individual claims **below** ten thousand euros.

In order to make the process lighter, we will associate these claims to a random variable that we will call X_1.

6.5.3 Distribution of X_1

As it was done in the first process, the first step is to get to know the response variable a little better.

For this, we will proceed to a brief exploratory analysis made *a priori* and only then we will start the modeling itself. When we work with databases, the first intuition is clearly to start determining the basic statistics of the main variables, as well as to start drawing graphs that show us the statistical behavior of the variables.

The following table shows the basic statistics corresponding to variable X_1:

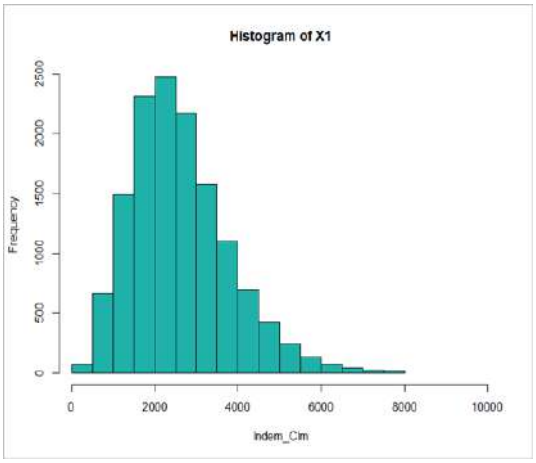
Table 6.36: Basic Statistics of X_1. **Source:** Authors.

Min.	1st Qu	Median	Mean	3rd Qu.	Max.	Skew	Kurtosis
110.8	1,770.5	2,447.8	2,611.8	3,293.4	9,900	0.83135	3.960986

From the summary table presented above we can conclude that X_1 corresponds to claims between one hundred and ten euros and nine thousand nine hundred euros. On average, for the portfolio considered, the cost of claims is around two thousand and six hundred euros. Given that the skewness value is less than 1 and greater than 0.5, we can consider that the distribution is moderately skewed. This skewness is to the right given the positive kurtosis value.

Below, we can find this representation in figure (6.15) that reflects the previous conclusions in graphical terms:

Figure 6.15: Individual Claims Cost Histogram. **Source:** Authors.



As we can see, the minimum claims amount registered in the portfolio is equal to one hundred euros, while the maximum value for this variable is equal to nine thousand nine hundred euros.

From the statistics presented in table (6.36), it's possible to conclude that seventy five percent of the population is concentrated in the claims below three thousand and three hundred euros. That, when combined with a positive skewness, clearly indicates right skewed behaviour.

With these characteristics, there are many possible distributions to fit X_1. The most common are *Gamma*,*Weibull* and *LogNormal* distributions.

6.5.4 Fitting the Distribution of X_1

In this section we aim to fit a distribution for X_1.

To do it, we started our analysis by the creation of a *Cullen and Frey graph*, that consists on graphing the skweness and kurtosis "face to face", for each of the distributions to be evaluated as possible to be fitted.

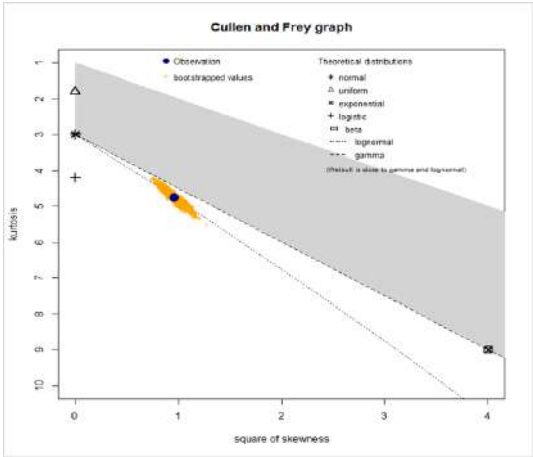
Although we know that these two distribution moments are not very robust due to the presence of a high level of uncertainty associated to them, we decided to exhibit the analysis.

The following chart was obtained through a R-function named as *descdist()* available in the R-package *fitdistrplus* (version 1.1-8).

The outputs are presented in the following figure (6.16).

To compensate for the identified uncertainty of the skewness and kurtosis parameters, it's noticeable that the graph shows a set of yellow points corresponding to the

Figure 6.16: Cullen and Frey graph for X_1. **Source:** Authors.



results obtained when a bootstrap, of two thousand and five hundred trials, is performed. This representation allows several scenarios "around" the observed point to be drawn, allowing a better justification of the decision to be taken.

The results presented in the figure shows a clear approximation of the observation (identified with a blue dot in the graph) to the line identified as belonging to the *Gamma* distribution.

Again, it should be noted that, this analysis should be seen as a distribution indicator and not as a criterion for the final selection of the distribution of X_1.

More rigorously, we formulated three robust statistical tests by assuming for each one of them a null hypothesis that will be tested using R.

The hypothesis are given by:

- H_0^1: X_1 follows a *Gamma* distribution;
- H_0^2: X_1 follows a Weibull distribution;
- H_0^3: X_1 follows a *LogNormal* distribution.

Starting by analysing the first hypothesis H_0^1, we performed a *Kolmogorv-Smirnov* test by using the function *ks.test()* available in *stats* R-package. This function was used together with the *fitdist()* that allowed us to estimate the observed parameters of the data by using the moment matching method.

Therefore, the test was applied and the following table summarizes the test results as well as the estimated parameters:

Table 6.37: Test Results of the Estimated Parameters for X_1. **Source**: Authors.

Parameters	**Estimate**	***p-value***
shape	5.2597	0.06907
rate	0.001956596	

Since the *p-value* is greater that five percent, it's possible not to reject the null hypothesis, i.e., X_1 follows a *Gamma* distribution.

Being said that, in fact there is no need to proceed to the remaining tests. Although, we performed them and the results provided were always to reject the null hypothesis H_0^2 and H_0^3.

6.5.5 A *Gamma* GLM Application for X_1

In terms of process, the path to be taken for the construction of a tariff for X_1 is very similar to what was performed in the case of the frequency of claims. As such, we will restrict ourselves to presenting the main results for this tariff.

Before doing so, we want to recall that the intermediate steps were performed by:

1. Transforming the continuous variables into categorical ones, through the discretization quantile method;

2. Choice of tariff variables. This process is similar to the one in section (6.20), where a selection of the variables to be included in the modeling phase of X_1 was made. The levels generated for each of the variables in this conditions will be explicit when presenting the final results of the tariff.

3. Definition of the Standard Insured. As it was performed in the case of the claims frequency, the standard insured was defined based on the count observed in each level of the tariff variables. To guarantee robustness, the level with the high number of claims (without taking into account if the record constitutes in fact a claim).

The assumptions used to perform this model were:

- The model ran assuming the claims amount below ten thousand euros (X_1) as the response variable.
- The model was ran using the *glm()* R-function available in the *stats* package (version 3.6.2).
- With the exception of the driver's age, all variables taken as starting point for the model run were taken as categorical.
- All the categorical variables are re-leveled such that the *glm()* R-function recognizes the Standard Insured as the intercept of the model.
- The linkage function used was the logarithmic function, to provide us a multiplicative structure to calculate the premiums.
- The model ran in R, assuming the following input:

```
1 model_gamma_full=glm(Idem_Ind ~ Sex + Vehic_Type + HorsePower_CAT + Marital_
      ↪ Status + Literacy + Driv_Age,family=Gamma(link="log"),data=SEV_withlim)
2 summary(model_gamma_full)
```

The outputs of this release are summarized in the following table:

Table 6.38: Output of *Gamma* GLM Application for X_1. **Source:** Authors.

	Estimate	**Std.Error**	**t value**	**Pr(>|t|)**	
(Intercept)	7.634804	0.018316	416.83	<2E-16	***
Sex0	0.002447	0.007166	0.341	0.7328	
Vehic_type2	0.017584	0.011109	1.583	0.1135	
Vehic_type3	0.007458	0.012222	0.61	0.5417	
Vehic_type4	0.019077	0.023764	0.803	0.4221	
Vehic_type5	0.119783	0.05664	2.115	0.0345	*
Vehic_type7	-0.02271	0.009979	-2.276	0.0229	*
Vehic_type8	-0.02199	0.011128	-1.976	0.0481	*
HorsePower_CAT(95,116]	0.169548	0.011493	14.753	<2E-16	***
HorsePower_CAT(116,126]	0.187248	0.010842	17.27	<2E-16	***
HorsePower_CAT(126,168]	0.2303	0.010878	21.171	<2E-16	***
HorsePower_CAT(168,357]	0.510039	0.010838	47.062	<2E-16	***
Marital_StatusDivorced	-0.00188	0.011967	-0.157	0.8752	
Marital_StatusSoSingle	0.007881	0.010402	0.758	0.4486	
Marital_StatusWidow	0.00077	0.009783	0.079	0.9373	
Literacy(9,12]	-0.00491	0.007704	-0.637	0.5243	
LiteracySuperior	-0.02099	0.011523	-1.821	0.0686	.
Driv_Age	0.000336	0.000311	1.08	0.2799	

As can be seen, not all variables taken as significant in explaining the cost of claims below ten thousand euros. For example, all levels of the variables *Marital_Status*, *Literacy*, *Sex* and the variable *Driv_Age* do not show any indication of significance in explaining the response variable. As such, all of these can be viewed by the actuary as possible variables to be excluded.

At this stage there would be several possibilities to carry out this study: The construction of maximum likelihood ratio tests considering fitted models, the construction of Wald tests assuming that the estimates of these variables are equal to zero, among others.

The truth is that, similarly to what was done for the frequency of accidents, for the modeling of X_1, we applied directly the *Stepwise Backwards* method for the elimination of variables.

The following table shows the process run in the R project, which highlights all the steps performed by the *Stepwise* model. Along with this, we also show the evolution of the *AIC* value:

Table 6.39: *Stepwise* Steps along *AIC* value for X_1. **Source:** Authors.

Stepwise Step	**Model**	***AIC***
Initial Model	*Idem_Ind ~Sex + Vehic_Type + HorsePower_CAT + Marital_Status + Literacy + Driv_Age*	*22,7343*
First Step	***Initial Model*** *- Martial_Status*	*22,7338*
Second Step	***First Step*** *- Sex*	*22,7336.9*
Third Step	***Second Step*** *- Driv_Age*	*22,7336.0*
Fourth Step	***Third Step*** *- Literacy*	22,7335.3
Final Model	***Fourth Step*** *= Idem_Ind ~Vehic_Type + HorsePower_CAT*	

Taking into consideration the ***Final Model*** explicit in table (6.39), the estimates obtained are given by:

In the *Final Model*, it's intended to have only the variables that are significant to explain our response variable. In fact, from the outputs of the previous table, with the

Table 6.40: Final Model estimates for X_1. Source: Authors.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.647452	0.008981	851.49	<2E-16	***
Vehic_Type2	0.018456	0.011096	1.663	0.0963	.
Vehic_Type3	0.006821	0.012197	0.559	0.576	
Vehic_Type4	0.017378	0.023724	0.732	0.4639	
Vehic_Type5	0.1178	0.056556	2.083	0.0373	*
Vehic_Type7	-0.02008	0.009862	-2.036	0.0418	*
Vehic_Type8	-0.01934	0.011048	-1.75	0.0801	.
HorsePower_CAT(95,116]	0.169459	0.011487	14.753	<2E-16	***
HorsePower_CAT(116,126]	0.187162	0.010837	17.27	<2E-16	***
HorsePower_CAT(126,168]	0.230126	0.010874	21.164	<2E-16	***
HorsePower_CAT(168,357]	0.510014	0.010835	47.072	<2E-16	***

exception of levels *2* and *3*, all the categorical factors of the two variables do present a *p-value* lower than five percent.

Another important aspect to be pointed out is that all the estimates for the levels of the significant levels are positive. As such, cases for which these conditions are met reflect surcharges on the estimate obtained for the standard insured.

The two levels corresponding to vehicle types *7* and *8* are distinguished from the remaining, because the estimates obtained by the model are negative. Therefore, they represent a discount from the estimate provided by the standard insured.

In view of the results obtained, it's always good to be aware that they can be improved.

To this end, pooling categorical levels of the significant variables is recommended. However, this aggregation of classes must have some basis, since we are changing the final model obtained. I.e., it would be a benefit if the aggregation performed considered levels with the same risk profile.

There are two ways to proceed with this action:

1. One is the more rigorous one, in which a *Wald* test is performed on the equality of estimates between the two categorical levels;

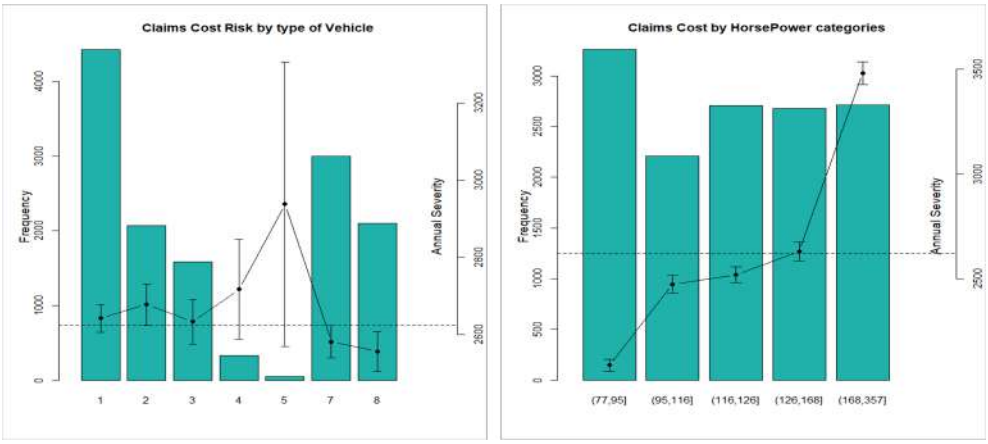
2. The second approach corresponds to a lighter view in which for each of the levels the observed risk is calculated and if the distance between both is reduced, in a subjective way the actuary can aggregate the levels under consideration.

In the case presented in this dissertation, we proceeded with the second approach. As such, we calculated for each of the two variables the risk that each of the constituent levels represents in terms of cost of claims and represented them in figure (6.17).

The graphs presented below are similar to the ones produced in (6.9), but in this case we are evaluating claims severity as risk. Considering the observable outputs, we have that the variable *HorsePower_CAT* presents relatively well distinguishable levels, in terms of risk.

With respect to the variable *Vehic_Type*, we concluded in a first phase, that level *3* should be aggregated to the standard insured (level *1*), since both are above the average

Figure 6.17: Claims Cost Risk by *Vehic_Type* and *HorsePower*. **Source:** Authors.



annual severity and very close to each other as is visible in the figure (only costs below ten thousand euros are considered).

Based on the same rationale, we merged the levels *2* and *4* of the *Vehic_Type*.

After performing this aggregation we run the model again and the observed estimates are given by:

Table 6.41: Model Estimates for X_1. **Source:** Authors.

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	
(Intercept)	7.649273	0.008362	914.782	<2e-16	***
Vehic_Type2 & 4	0.016503	0.010057	1.641	0.1009	
Vehic_Type5	0.115992	0.056458	2.054	0.0399	*
Vehic_Type7	-0.02188	0.009321	-2.348	0.0189	*
Vehic_Type8	-0.02114	0.010566	-2.001	0.0454	*
HorsePower_CAT(95,116]	0.169486	0.011485	14.757	<2e-16	***
HorsePower_CAT(116,126]	0.187126	0.010835	17.271	<2e-16	***
HorsePower_CAT(126,168]	0.230165	0.010872	21.17	<2e-16	***
HorsePower_CAT(168,357]	0.509916	0.010831	47.079	<2e-16	***

The adjustments conducted were found to be beneficial only for the addition of level *3* to the insured level. The p-value remained below five percent. However, the same was not true for levels *2* and *4*, where, after aggregating the levels (level *2 & 4*, a p-value above the five percent reference value was maintained.

Given the observed results, we have that with two exceptions, all levels of the two variables generate aggravations to the premium calculated for X_1. This conclusion is given by the fact that the estimates are positive. The two exceptions correspond to types *7* and *8* of the *Vehic_Type* variable. In these cases, we verify a discount on the estimate of the standard insured.

In the next section, we will calculate the expected value observed for X_1 for each tariff level of the variables.

6.5.5.1 E[X1] Calculation

For the portfolio under consideration, the calculation of the expected value of the cost of regular claims is based on the same assumptions used for the expected value of claims frequency.

In other words, we are considering that the construction of the estimates is done at the expense of the standard insured.

Given a multiplicative structure, the premiums were calculated for X_1, and the results are presented in the following table:

Table 6.42: Expected Value of Claims Frequency for X_1. **Source:** Authors.

	E[X1]
(Intercept)	2,099.118
Vehic_Type2 & 4	2,134.047
Vehic_Type5	2,357.282
Vehic_Type7	2,053.681
Vehic_Type8	2,055.202
HorsePower_CAT(95,116]	2,486.818
HorsePower_CAT(116,126]	2,531.073
HorsePower_CAT(126,168]	2,642.386
HorsePower_CAT(168,357]	3,495.351

The results observed in this table are in line with what was concluded from the results obtained in the table (6.41), seen in the previous section.

6.5.6 modeling Severe Claims

A severe claim, as the name indicates corresponds to a very high claim cost and, it depends from dataset to dataset.

In this project, the threshold that sets the separation between a regular claim and a severe claim is ten thousand euros, as concluded in figure (6.13) and (6.14) of the section previously presented in this project. Additionally, in these two figures, it's possible to confirm that, the probability distribution of severe claims is different from the regular claims.

As such, there is a need to model these costs differently. Taking into account the bibliographical references, the most common model to be used corresponds to the construction of a logistic regression.

However, for this dissertation and similarly to what was done for X_1, we started the study by analysing a *priori* X_2, where X_2 corresponds to the random variable used to identify this type of claims).

It is only after we have carried out this initial study that, we will move on to the action plan for modeling these claims.

First of all, it should be noted that we are dealing with a very small materiality of the portfolio. After counting the cases (at the claim level), we have registered eight

hundred and seventy two cases, representing only around six percent of the total population with at least one claim registered.

Nevertheless, since they are considered severe claims, all of these observations represent a major impact on the premiums calculation.

6.5.7 Distribution of X_2

As we have done so far, we begin the exploration of the variable by performing a basic statistical analysis.

The results obtained are given by the following table (6.43), once again performed in R using the *summary()* function, available in *stats* R-package (version 3.3.5)

Table 6.43: Summary of X_2. **Source:** Authors.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Skew	Kurtosis
10,083	27,482	35,535	37,244	46,236	99,540	0.6	3.39

From the summary table presented above we can conclude that X_2 corresponds to claims between ten thousand and eighty three euros and one hundred thousand euros. On average, for the portfolio considered, the cost of claims is around thirty seven thousand euros. Given that the skewness value is less than 1 and greater than 0.5, we can consider that the distribution is moderately skewed. This skewness is to the right given the positive kurtosis value.

6.5.8 Fitting the Distribution of X_2

Given the characteristics presented in table (6.43) of section (6.5.7), the actuary may notice several similarities to what has been pointed out in the case of X_1. It is not unreasonable then to perform an adjustment to this distribution assuming as null hypothesis that:

$$H_0\text{: } X_2 \text{ follows a } \textit{Gamma} \text{ distribution.}$$

To test the null-hypothesis a *Kolmogorv-Smirnov* test was performed by using the function *ks.test()* available in *stats* R-package. This function was used together with the *fitdist()* that allowed us to estimate the observed parameters of the data by using the moment matching method.

Therefore, the test was applied and the following table (6.44) summarizes the test results as well as the estimated parameters:

Table 6.44: Summary of *Kolmogorv-Smirnov* test results for X_2. **Source:** Authors.

Parameters	Estimate	*p-value*
shape	6.41046	0.441
rate	0.0001721	

Since the *p-value* is greater that five percent, it's possible not to reject the null hypothesis, i.e., X_2 follows a *Gamma* distribution.

6.5.9 A *Gamma* GLM Application for X_2

Since there is no evidence to reject that X_2 follows a *Gamma* distribution, we can take advantage of this conclusion. As mentioned before, severe claims should be modelled differently since they usually follow a different distribution from the one observed for regular claims.

For our database, we have indeed, that the data fits a distribution commonly used for modeling regular claims. Therefore, we proceeded with the construction of a tariff also for X_2.

In order to avoid repetition, we will present only the main results of our analysis. Nevertheless, we recommend the actuary to retrieve the intermediate steps until obtaining them. These are broken down in general steps 1-3 of section (6.5.5).

A point that really cannot be ignored consists on the explanation of the assumptions made to run the model for X_2. As such, we present them in the following points:

- The model ran assuming the claims amount above ten thousand euros (X_2) as the response variable.
- The model was ran using the *glm()* R-function available in the *stats* package (version 3.6.2).
- With the exception of the driver's age, all variables taken as starting point for the model run were taken as categorical.
- All the categorical variables are re-leveled such that the *glm()* R-function recognizes the Standard Insured as the intercept of the model.
- The linkage function used was the logarithmic function, to provide us a multiplicative structure to calculate the premiums.
- The model ran in R, assuming the following input:

```
1 model_gamma_full_2=glm(Indem_Indiv ~ Region + Sex + Vehic_Type + HorsePower_CAT
      ↪ + Marital_Status + Literacy + Fuel + Driv_Age,family=Gamma(link="log"),
      ↪ data=SEV_withlim_2)
2 summary(model_gamma_full_2)
```

The outputs of this release are summarized in the following table:

Table 6.45: Output of *Gamma* GLM Application for X_2

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	10.39344	0.080989	128.332	<2e-16	***
RegionCenter	0.089073	0.044806	1.988	0.047145	*
RegionSouth	-0.03634	0.033854	-1.073	0.283408	
Sex0	-0.01967	0.027043	-0.727	0.467258	
Vehic_Type2	0.08032	0.042557	1.887	0.059465	.
Vehic_Type3	0.064762	0.047596	1.361	0.173988	
Vehic_Type4	0.039838	0.112111	0.355	0.722423	
Vehic_Type5	0.121684	0.199179	0.611	0.541415	
Vehic_Type6	0.013386	0.039206	0.341	0.732861	
Vehic_Type7	0.044946	0.044637	1.007	0.314267	
HorsePower_CAT(228,241]	0.011782	0.042513	0.277	0.78174	
HorsePower_CAT(241,254]	0.124641	0.041262	3.021	0.0026	**
HorsePower_CAT(254,269]	0.150513	0.042321	3.556	0.000397	***
HorsePower_CAT(269,319]	0.112146	0.041958	2.673	0.007671	**
Marital_StatusDivorced	0.04402	0.045724	0.963	0.335972	
Marital_StatusSingle	0.00546	0.039825	0.137	0.890995	
Marital_StatusWidow	-0.06421	0.037497	-1.713	0.087175	.
Literacy(9,12]	-0.01997	0.029468	-0.678	0.498059	
LiteracySuperior	0.034274	0.043192	0.794	0.42771	
FuelDiesel	0.001707	0.03013	0.057	0.954836	
FuelHybrid	-0.00423	0.055175	-0.077	0.938973	
Driv_Age	0.000869	0.001289	0.675	0.500181	

As can be observed, for the case of X_2, most of the variables are not being considered as significant to explain the response variable. This conclusion is supported by the fact that the vast majority of p-values are all greater than five percent.

This way, and similarly to what we have done for the other tariffs, we applied the *Stepwise Backwards* model in R. The results of this application were also obtained using R, and are presented in the table (6.46) presented below:

Table 6.46: *Stepwise* Steps for X_2

Stepwise Step	**Model**	***AIC***
Initial Model	*Idem_Ind sim Region + Sex + HorsePower_CAT + Marital_Status + Vehic_Type + Literacy + Fuel + Driv_Age*	*18,571.29*
First Step	*Initial Model - Vehic_Type*	*18,564.36*
Second Step	*First Step - Fuel*	*18,560.4*
Third Step	*Second Step - Literacy*	*18,557.83*
Fourth Step	*Third Step - Driv_Age*	*18,556.25*
Fifth Step	*Fourth Step - Sex*	*18554.74*
Sixth Step	*Fifth Step - Marital_Status*	*18,554*
Final Model	*Fourth Step = Idem_Ind sim Region + HorsePower_CAT*	*18,554*

The *Stepwise* results indicate that only two variables are significant in explaining X_2. We are talking about the variables *Region* and *HorsePower_CAT*.

For the model identified as ***Final Model*** in the previous table presented, we run the model again, to highlight the estimates obtained for the variables in question. The results are summarized in the following table:

Table 6.47: Final Model Estimates for X_2. **Source:** Authors.

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**	
(Intercept)	10.42892	0.03242	321.683	<2e-16	***
RegionCenter	0.09495	0.03477	2.731	0.006449	**
RegionSouth	-0.03944	0.03067	-1.286	0.198867	
Horsepower_CAT(228,241]	0.02035	0.04225	0.482	0.630178	
Horsepower_CAT(241,254]	0.12863	0.04104	3.134	0.001783	**
Horsepower_CAT(254,269]	0.15986	0.04188	3.817	0.000145	***
Horsepower_CAT(269,319]	0.12043	0.04149	2.903	0.003797	**

As it can be seen, all the categorical levels present positive estimates. Therefore, all of them are considered to be worsening in relation to the estimate observed for the standard insured ([12]).

In the next section, we will calculate the expected value observed for X_2 for each tariff level of the variables.

6.5.9.1 E[X2] Calculation. Source: Authors.

Given a multiplicative structure, the premiums were calculated for X_2, and the results are presented in the following table:

Table 6.48: Expected Value of Claims Frequency for X_2

	E[X2]
(Intercept)	33,823.73
RegionCenter	37,192.81
RegionSouth	32,515.8
Horsepower_CAT(228,241]	34,519.03
Horsepower_CAT(241,254]	38,466.57
Horsepower_CAT(254,269]	39,686.82
Horsepower_CAT(269,319]	38,152.55

The results coincide with what was expected. In fact only the *South Region* which implies a premium discount, when compared to the premium obtained for the standard policyholder.

At this stage, we are able to calculate the estimated aggregate losses. The following chapter aims to demonstrate how to do it.

[12] Note that the insured insurer is simply defined for the model built as: *HorsePower_CAT* equal to "*(152,228]*" and *Region* equal to *North*

6.6 Calculating E[S]

As presented in equation (3.5) of chapter (3), we have that the estimation of the aggregate losses of the portfolio is made at the expense of the expected value of the frequency and severity of claims, as follows:

$$E[S] = E[N] \times E[X]. \tag{6.5}$$

The best solution at this point is to look at each element of the factors separately and only at a later stage move forward with the product calculation.

Let us start with the $E[N]$.

The calculation of the expected value of the claims frequency is immediate. We already have the tariff set up in (6.34) and so, it's sufficient to make the allocation of the values obtained for each observation in the database.

Regarding the calculation of $E[X]$, the process is more complex. The approach takes into consideration an important result related to the calculation of expected values.

To be more precise, this calculation is obtained taking into account a breakdown of the variable X, conditional on the threshold from which this breakdown occurs. Therefore, we are talking about the Conditional Expected Value result.

In the specific case of this dissertation, the identified limit is ten thousand euros. As such, applying the result we have that the estimate of X is given by:

$$E[X] = E[X|X \leq 10,000]P[X \leq 10,000] + E[X|X > 10,000]P[X > 10,000] \tag{6.6}$$

As presented in section (6.5.6), we have that the probability of a claim above ten thousand euros being observed is only six percent. And, since in our case X is only split into two parts, the complementary probability is equal to ninety four percent.

Looking at what is presented in equation (6.6) and taking the process developed so far, it is possible to ascertain that this equality can be rewritten as follows:

$$E[X] = E[X_1] \times 0.94 + E[X_2] \times 0.06 \tag{6.7}$$

At this stage, the process becomes direct. By considering each tariff of X_1 and X_2 presented in table (6.42) of section (6.5.5.1) and (6.48) of section (6.5.9.1) respectively, the final work consists on the individual allocation of the values obtained for each claim in the database.

After it, the actuary will only have to do the work of calculating for each claim its expected value using the equation (6.7).

The final calculation of the expected value of aggregate losses will be obtained by summing each of the aggregate losses obtained by computing the equation (6.5).

In the case of the portfolio we are considering in this study, the value obtained is 73,451,440€.

The average premium per policy is obtained by the simple coefficient between the expected value of aggregate losses and the total number of policies in the database. For the case of the problem under study it's given by the 183.62€.

7 | Results and Discussion

7.1 Reinsurance and TOPSIS Results

In this section, we aim to present the final step of the proposed workflow. As such, we will show the results obtained for the determination of the optimal reinsurance retention limit, under the assumptions of maximizing the insurance profit (after reinsurance) and minimizing the variance of the retained aggregated claim amount.

Below in section (7.1.1), we will find the results observed from the first step highlighted in section (3.4), that consists of listing all the alternatives belonging to the set of feasible solutions. Recalling this step, we present the results on a matrix $M \times N$, where C corresponds to the number of feasible options to the problem and N to the number of criteria.

In order to evaluate the pretended set of possible solutions, we have to define what would be the initial and ending points for the retention limit (following the notation in section (3.6.1), m and M respectively). In this step, it is also underlied the need of defining the increment value between options, which corresponds to the amount of money that is added from one alternative to the following one.

Within a range of the various options, the definition of the initial value can be performed in two possible ways:

- **Subjectively**: The actuary complies with a standard limit imposed by the insurance company that results from the industry expertise.
- **Objectively**: By considering the insurance portfolio characteristics, the actuary can opt to perform a clustering analysis in order to have a full understanding of the insured population profile and define the a more adjusted initial deductible to the specific situation.

7.1.1 Results

After the clustering construction, the portfolio could be segregated into two parts. Therefore, we could initialize the process of getting the pretended solution with M being equal to 8,000€. Meaning that, this was the starting point that was defined based on the value in which the claims amount is split. Translating this into a business

language, we are saying that at first stage, the insurer will be in charge of all losses below 8,000€.

Regarding the ending point, this value was defined based on the given portfolio, subjectively. We identified the most expensive portfolio's claim (99,540€) and decided to round it to 100,000€.

Finally, the choice of the 100€ increment was done based on a sensitivity analysis, that will be featured later in this section. This step is usually hidden in the procedure, but it reveals an important part in this type of studies.

Below, we can find a subset of our evaluation matrix, with the explicit alternatives and the correspondent profit and variance results:

Table 7.1: Evaluation Matrix. **Source:** Authors.

AlternativeID	M(€)	Variance	Profit/100(by policy)(€)
1	8,000	0.455816	0.00963356
2	8,100	0.459485	0.009655331
3	8,200	0.463198	0.009677098
4	8,300	0.466955	0.00969886
5	8,400	0.470756	0.009720618
		⋮	
919	99,800	22.0625	0.015839612
920	99,900	22.10598	0.01583964
921	100,000	22.14951	0.015839667

Regarding the presented outputs for both criteria, it seems reasonable to infer that for higher values of M, there is also an increase in the retained variance and profit.

However, an actuary can be unsatisfied only with this conclusion. Considering this primary observations he/she can be interested into get in more detail about the shape of both criteria.

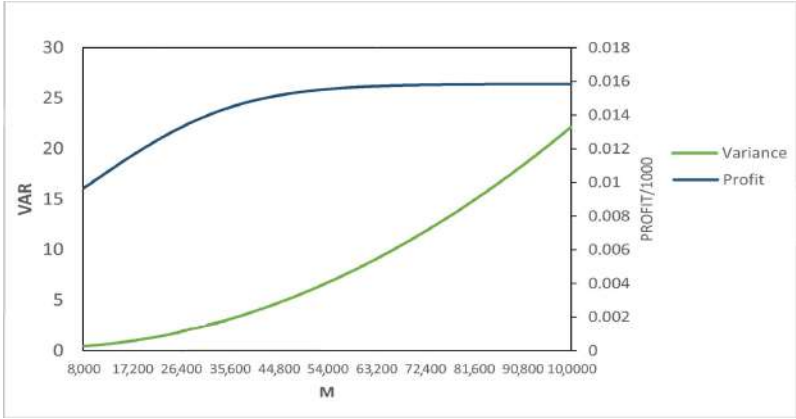
In order to clarify this need, it is possible to build a chart that considers these two metrics for each value of M.

From this graphic, in fact the greater the retention limit considered, the greater the profit and variance observed.

However, the increasing speed of both criteria is different. When comparing with profit, it seems that variance starts increasing faster for greater values of M. More than that, the profit's behaviour slows down to a point where it starts looking to be flat, for greater values of M.

In visual terms, it is noted by the concavities of each variable shape. While profit (blue line) presents a downwards concavity shape, variance (green line) presents an upwards concavity.

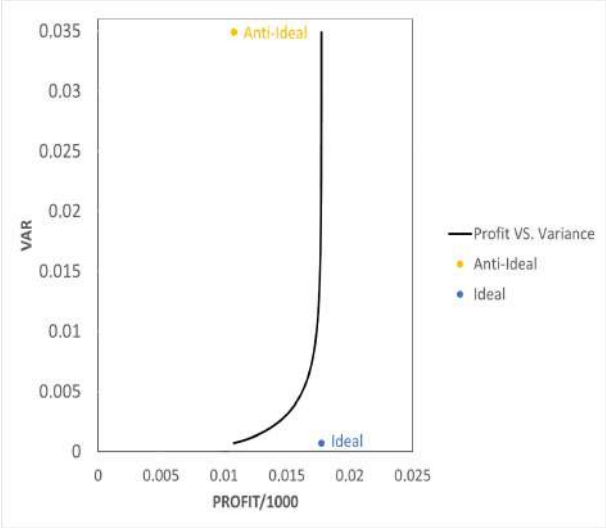
Figure 7.1: M vs. Variance of the Aggregate Claim Amount. **Source:** Authors.



Now that we are more aware about the behaviour of all alternatives, it is time to get into the final solution decision.

Similarly to what was presented in section (3.8.1.1), let's then plot the insurance profit (after reinsurance) against the retained variance of the aggregated claim amount. The following chart corresponds to the results of the pretended analysis when considering both criteria normalized ([1]).

Figure 7.2: Profit vs Variance of the Aggregate Claims Amount**Source:** Authors.



The scenario presented above is not particularly friendly in terms of decision making. As we can see, we have that the output from variance increases drastically when considering higher values of profit.

Since we pretend to minimize the variance and maximize as much as possible the profit, we have to find a proper balance between both criteria.

As we have been mentioning throughout this project, the proper way to do it is to

[1] The standardized evaluation matrix can be addressed in the annexes.

apply a decision theory technique. Out of a group of possibilities reviewed, the choice fell on TOPSIS.

This tool, as said before intends to find the optimal solution, the one that minimizes the distance to the best scenario and maximizes the distance to the worst scenario.

With this in mind, we also present in this previous chart two points that are drawn based on the information of our set of solutions:

- The blue point, corresponding to the ideal solution.
- The yellow point, corresponding to the anti-ideal solution.

Both points are characterized based on the criteria constraints defined for this study. Since we want to maximize profit and minimize variance, the hypothetical best solution is explained at the expense of the maximum standardized profit (0.0133€ by policy) observed in the set of possible solutions (black line). At the same time, for the ordinate it captures the minimum value for the standardized variance, 1.91.

As the name suggests, the anti-ideal solution (yellow point) is characterized by the minimum value for the profit variable and the maximum for the variance.

At this stage, we just need to define what would be the importance allocated for each variable. Meaning that through the application of TOPSIS, this algorithm allows the user to define if a specific variable will contribute more or less to the final decision.

In order to avoid the increasing of complexity, we decided to set the same importance for each criterion. I.e., we allocated to the profit and variance the same importance weight of fifty percent. In this section, we will also present a sensitivity analysis over this factor. We will vary the allocated weight and plot for each scenario the optimal solution obtained by *TOPSIS*.

Let's finally present the optimal solution to this insurance portfolio. After applying the TOPSIS algorithm, that is fully explained in section (3.8.1.1) we end up with a list of scores that are performed by the algorithm.

Recalling the way this score is defined, we have the following expression for each alternative:

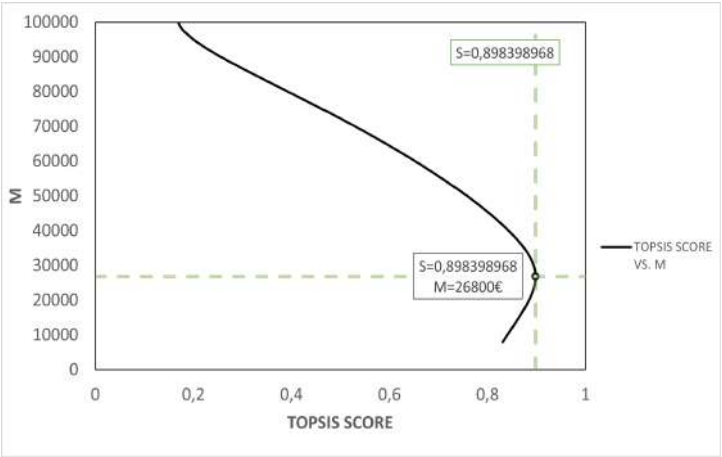
$$s_i = \frac{d^{AI}}{d^{AI} + d^{I}}, \tag{7.1}$$

with d^{AI} corresponding to the Euclidean distance between the anti-ideal and a possible final solution (belonging to the black line of chart (7.2)). d^{I} corresponds to the Euclidean distance between the ideal and possible final solution to the problem.

Following the results, it is possible to plot each $i-th$ TOPSIS-score against the retention limit used to perform this result:

From the chart previously presented, we can see that the optimal retention limit corresponds to the maximum TOPSIS-score, that in this specific graphic is given by the inflection point in the vertical axis.

Figure 7.3: TOPSIS Score vs M **Source:** Authors.

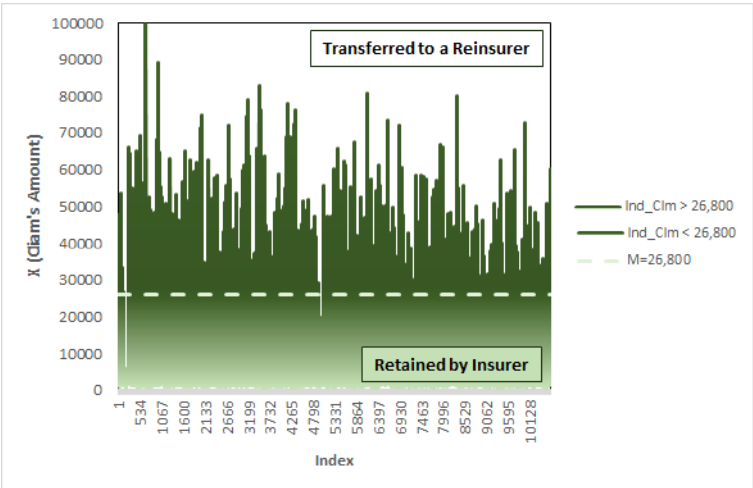


That being said, for the simulated insurance portfolio the optimal retention limit is equal to 26,800€.

Facing this value, the insurer should retain all the risks bellow this threshold. Above this value, it is advised to transfer the risk to a reinsurance company.

The chart below intends to represent this threshold applied to the claims of our simulated portfolio.

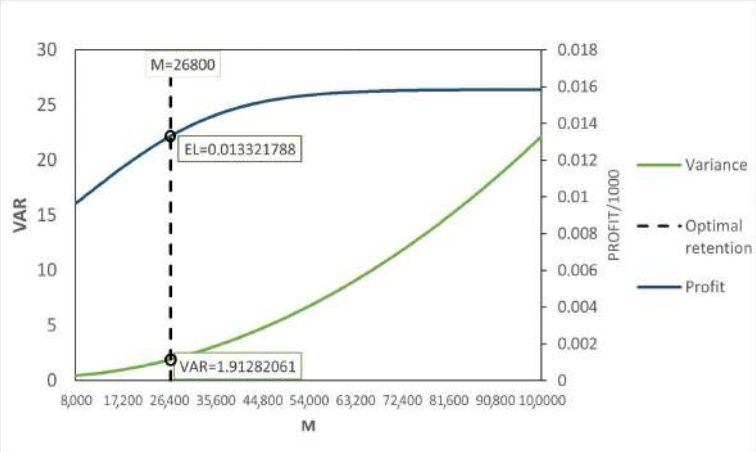
Figure 7.4: Claims Threshold **Source:** Authors.



After obtaining this result, it is expected to evaluate how much profit is expected to be obtained after applying the reinsurance contract to the portfolio. To do it, we can simply recall section (7.2) and see what is the correspondence between the optimal deductible and the expected profit and variance:

As we can see, for each policy the insurer will expect to gain a profit around 13€. In this specific portfolio, with 400,000 policies, it is expected a final profit of 5,200,000€.

Figure 7.5: M vs. Variance of the Aggregate Claim **Source:** Authors.



7.2 Sensitivity Analysis

This section intends to present a group of analysis regarding the factors that can vary when computing the optimal retention limit.

We will start presenting the results and discuss about the impacts on the final reinsurance retention limit calculation.

7.2.1 Increment Choice Analysis

The choice of the increment when defining the evaluation matrix can become a problem to calculate the optimal retention limit.

In fact, if this value is not tight enough, we can rely on a sub-optimal retention limit to the portfolio considered.

Since we are working with reinsurance, we tend to define a very thick partition. As such, the decision of an increment of 100€ was not immediate. In order to obtain it, we implemented TOPSIS in a cyclical manner. Briefly explaining, keeping all other TOPSIS factors fixed, we started to vary the increment of our evaluation matrix starting with values of 10,000€ to 10€. With it, we could determine for each matrix the optimal value through the TOPSIS scores.

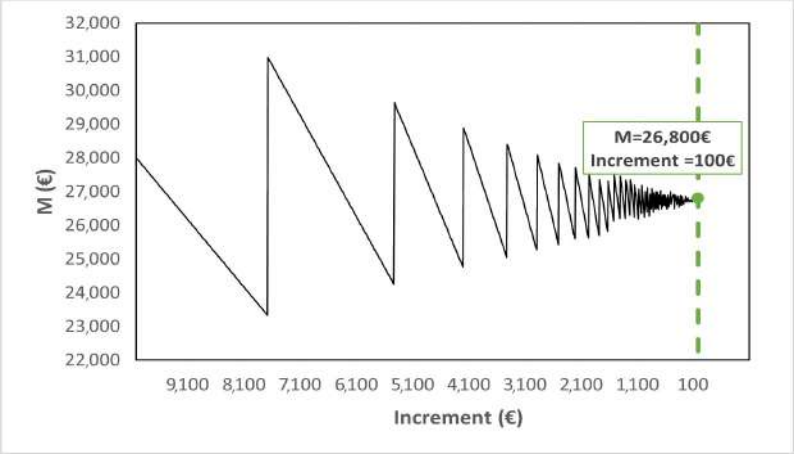
The final result can be plotted in a chart that considers the increment in the horizontal axis and the correspondent optimal retention limit in the vertical axis:

As we can see, the smaller the increment, the smaller the variation in the optimal retention limit. As such we defined the value of 100€ as a reasonable value to be defined as our increment value.

7.2.2 Weight allocation on each criterion

As mentioned in section (7.1), the TOPSIS algorithm allows the user to define the importance of each criterion to achieve the optimal solution.

Figure 7.6: Increment Sensitivity Analysis **Source:** Authors.



In our procedure, we decided to allocate the same importance for both criteria. However, the analyst can be interested into change this.

As such, we present below the optimal retention limits obtained when considering a set of weight combinations for both criteria. This combination set is presented in the following table. In the first column, we present the weight allocated to the retained variance and the second column corresponds to the retained profit:

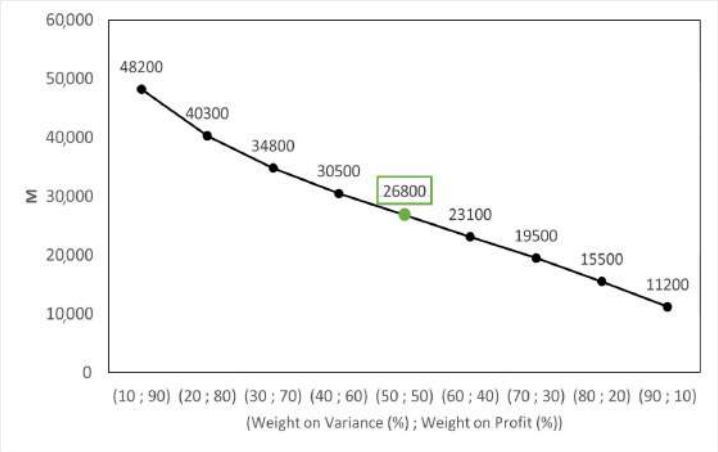
Table 7.2: Optimal Retention Limits. **Source**: Authors.

Weight on Variance (%)	Weight on Profit (%)	Optimal Retention Limit (€)
10	90	48,200
20	80	40,300
30	70	34,800
40	60	30,500
50	50	26,800
60	40	23,100
70	30	19,500
80	20	15,500
90	10	11,200

As we can see, there is a clear decreasing pattern of the optimal retention limit. If the actuary is more interested into increasing the profit obtained in the end (after reinsurance), it means that the he/she can define a greater optimal retention limit (48,200€). On the other hand, if the principal constraint relies on minimizing the retained variance as much as possible, then the final retention limit should be smaller (11,200€).

This table can also be presented as a chart in order to understand better the decreasing pattern:

Figure 7.7: Evolution of M **Source:** Authors.



8 | Conclusion

The application of reinsurance is increasingly frequent in the insurance sector. From a financial point of view, this type of contract is undoubtedly becoming a fundamental tool for meeting the requirements under the *Solvency II* regime.

Therefore, it is essential for an insurer, once entering into a contract of this type, to e sure that the results will be favorable. It is deemed necessary that the agreed treaty would correspond to the optimal.

The great challenge that we set ourselves in this dissertation corresponds precisely to the elaboration of a line of thought that would allow answering to this need.

Given the lack of online sources/databases that would allow us to carry out this work on the desired dimension, considering the automobile sector taking an of Own Damages portfolio, it was decided to carry out the study from scratch.

Lightly speaking, generalized linear model simulation concepts were introduced, allowing us to build a database. Although we have not performed a more detailed study regarding the performance of the simulation, we noticed, through the exploration of the data, several patterns, connections, and dependencies that were simulated.

In the second phase of our study, we delved into the topic of calculating rates that would allow us to calculate the expected value of aggregate losses and the value of the premium that should be allocated per policy. The procedure of this phase was entirely carried out through generalized linear models to proceed with the frequency of claims and severity of the claims modeling phase.

This step proved to be crucial when developing the main objective that motivated the preparation of this dissertation, the calculation of the optimal deductible to be taken in an automobile reinsurance contract.

Taking an Excess of loss reinsurance as optimal reinsurance for the portfolio in question, we developed an objective problem in which, assuming a set of conditions, we applied a tool associated with the decision theory that became central to the determination of the full. The tool used was the *TOPSIS* algorithm.

In global terms, we can conclude that, for the criteria defined in the scope of this dissertation, the application of this method has allowed us to obtain a full that is considered optimal and that can be extrapolated if more decision criteria are taken. However, this work was not carried out in this project.

Furthermore, the generalization of this framework to other insurance products is

undoubtedly another dimension to be considered when taking the approach of this dissertation.

Bibliography

Bowers, N., H. Gerber, J. Hickman, D. Jones, and C. Nesbitt (1997). *Actuarial Mathematics*. 2nd. The Society of Actuaries.

Bulut Karageyik, B. and Ş. Şahin (2017). "Determination of the optimal retention level based on different measures." In: *Journal of Risk and Financial Management* 10(1), p. 4.

Bühlmann, H. (1970). *Mathematical Methods in Risk theory*. Springer. ISBN: 0072-7830.

Cai, J. and K. S. Tan (2007). "Optimal retention for a stop-loss reinsurance under the VaR and CTE risk measures." In: *ASTIN Bulletin: The Journal of the IAA* 37(1), pp. 93–112.

Carsey, T. M. and J. J. Harden (2013). *Monte Carlo simulation and resampling methods for social science*. Sage Publications.

Carter, R. L. (2013). *Reinsurance*. Springer Science & Business Media.

Centeno, M. d. L. and M Guerra (2010). "The optimal reinsurance strategy—the individual claim case." In: *Insurance: Mathematics and Economics* 46(3), pp. 450–460.

Centeno, M. d. L. et al. (1997). "Excess of loss reinsurance and the probability of ruin in finite horizon." In: *Astin Bulletin* 27(1), pp. 59–70.

Centeno, M. L. and O. Simões (2009). "Optimal reinsurance." In: *RACSAM-Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas* 103(2), pp. 387–404.

De Finetti, B. (1940). *Il problema dei pieni*. Istituto italiano degli attuari.

De Vylder, F. and Goovaerts (1988). "Recursive calculation of finite-time ruin probabilities." In: *Insurance: Mathematics and Economics* 7(1), pp. 1–7.

Delegated, C (2015). "Commission delegated regulation (EU) 2015/35 of 10 October 2014 supplementing directive 2009/138/EC of the European parliament and of the council on the taking-up and pursuit of the business of insurance and reinsurance (Solvency II)." In: *Official Journal of European Union*.

Dickson, D. C. and H. R. Waters (1996). "Reinsurance and ruin." In: *Insurance: Mathematics and Economics* 19(1), pp. 61–80.

Dickson, D. C. and H. R. Waters (2006). "Optimal dynamic reinsurance." In: *ASTIN Bulletin: The Journal of the IAA* 36(2), pp. 415–432.

Estatísticas dos Transportes e Comunicações 2004.

Estatísticas dos Transportes e Comunicações 2019. URL: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=71883472&PUBLICACOESmodo=2,urldate={2021-12-30}.

Feller, W. (2008). *An introduction to probability theory and its applications, vol 2*. John Wiley & Sons.

Finetti, B. de (1940). "The Problem of 'Full-Risk Insurances'." In: *Journal of Investment Management* 4, pp. 19–43.

Friendly, M. and D. Meyer (2015). *Discrete data analysis with R*. Vol. 120. CRC Press.

Gaussian mixture models: k-means on steroids. URL: https://smorbieu.gitlab.io/gaussian-mixture-models-k-means-on-steroids/?fbclid=IwAR0Db-3ygi31hgF0yCZaxyRXqB5N6e1dtJyzb2J8IpGIJ0-RyIUxGCQLya4#gaussian-mixture-models-k-means-on-steroids (visited on 10/27/2021).

Guerreiro, G. (2016). *Manual de Construção de Tarifas com R – O Exemplo do Seguro Automóvel*. 2nd. FCT-UNL.

Hamerly, G. and C. Elkan (2003). "Learning the k in k-means." In: *Advances in neural information processing systems* 16, pp. 281–288.

Holtsmark, C. A. J. P. (2015). "Optimal Reinsurance Per Event." Master's thesis.

Huang, Y. and S. Meng (2019). "Automobile insurance classification ratemaking based on telematics driving data." In: *Decision Support Systems* 127, p. 113156.

Kaas, R., M. Goovaerts, J. Dhaene, and M. Denuit (2008). *Modern actuarial risk theory: using R*. Vol. 128. Springer Science & Business Media.

Kahraman, C. (2008). "Multi-criteria decision making methods and fuzzy sets." In: *Fuzzy multi-criteria decision making*. Springer, pp. 1–18.

Kaluszka, M. (2004). "Mean-variance optimal reinsurance arrangements." In: *Scandinavian Actuarial Journal* 2004(1), pp. 28–41.

Karageyik, B. B. and D. C. Dickson (2016). "Optimal reinsurance under multiple attribute decision making." In:

Lemaire, J. (1995). *Bonus-malus systems in automobile insurance*. Vol. 19. Springer science & business media.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models II*.

Moro, S. and P. Rita (2016). "Forecasting tomorrow's tourist." In: *Worldwide Hospitality and Tourism Themes*.

Mourik, T (2018). "Mortality risks, reinsurance and risk-based supervision." In: *South African Actuarial Journal* 18(1), pp. 1–15.

Qin, X.-S., G. H. Huang, A Chakma, X. Nie, and Q. Lin (2008). "A MCDM-based expert system for climate-change impact assessment and adaptation planning–A case study for the Georgia Basin, Canada." In: *Expert Systems with Applications* 34(3), pp. 2164–2179.

Straub, E. (1988). *Non-life insurance mathematics*. 517/S91n. Springer.

Tse, Y.-K. (2009). *Nonlife Actuarial Models, Theory, Methods and Evaluation*. Cambridge University Press. ISBN: 978-0-521-76465-0.

Velasquez, M. and P. T. Hester (2013). "An analysis of multi-criteria decision making methods." In: *International journal of operations research* 10(2), pp. 56–66.

Wuthrich, M. V. (2019). "Non-life insurance: mathematics & statistics." In: *Available at SSRN 2319328*.

Xuan, G., W. Zhang, and P. Chai (2001). "EM algorithms of Gaussian mixture model and hidden Markov model." In: *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*. Vol. 1. IEEE, pp. 145–148.

Young, V. R. (2006). "Premium Principles." In: *Encyclopedia of Actuarial Science*. DOI: https://doi.org/10.1002/9781118445112.stat04727.