Lisbon School of Economics & Management
Universidade de Lisboa

MASTER
ACTUARIAL SCIENCE

MASTER´S FINAL WORK
INTERNSHIP REPORT

BAYESIAN ANALYSIS FOR PREMIUM RISK: A COMPARATIVE STUDY WITH FREQUENTIST METHODS

PEDRO MIGUEL VARELA CALADO

Lisbon School of Economics & Management
Universidade de Lisboa

JUNE - 2026

Lisbon School of Economics & Management
Universidade de Lisboa

MASTER
ACTUARIAL SCIENCE

MASTER´S FINAL WORK
INTERNSHIP REPORT

BAYESIAN ANALYSIS FOR PREMIUM RISK: A COMPARATIVE STUDY WITH FREQUENTIST METHODS

PEDRO MIGUEL VARELA CALADO

SUPERVISION:
HÉLIO SILVA
RUI PAULO

JUNE - 2026

To my family and friends for
their constant love and
support, especially
throughout these two years
of my Master's degree.

GLOSSARY

AD – Anderson Darling.

AIC – Akaike Information Criterion.

BIC – Bayesian Information Criterion.

DTS – Dynamic Testing System.

K-S – Kolmogorov Smirnov.

HMC – Hamiltonian Monte Carlo.

MCMC – Markov Chain Monte Carlo.

MEL – Mean Excess Loss.

MGE – Minimum Goodness of fit Estimation.

MLE – Maximum Likelihood Estimation.

MTPL – Motor Third-Party Liability.

NUTS – No-U-Turn Sampler.

LOO-CV – Leave-One-Out Cross-Validation.

PPC – Posterior Predictive Checks.

SSQ – Sum of Squared Quantiles.

ABSTRACT AND KEYWORDS

Premium risk is a key component in non-life insurance, arising from uncertainty in both the frequency and severity of future claims. The calibration of large losses at Ageas Portugal relies on frequentist methods, but this approach provides limited quantification of the parameter uncertainty. This motivates the central objective of this work, which is to assess whether a Bayesian modelling provides better insights and predictive performance than the frequentist modelling in premium risk.

To address this question, both frequentist and Bayesian models are applied to the analysis of the frequency and severity of large losses in the line of business Motor Third-Party Liability (MTPL). The claims frequency is modelled under Poisson and Negative Binomial distributions, while claims severity is modelled under Pareto Type I, Pareto Type II and Lognormal distributions. Prior distributions are carefully specified and posterior inference is carried out using analytical solutions or simulation methods. The model adequacy is evaluated through posterior distributions, by assessing parameter uncertainty and whether the observed values are consistent with prior knowledge, as well as through Posterior Predictive Checks (PPC) and Leave-One-Out Cross-Validation (LOO-CV).

The results indicate that Bayesian models achieved higher predictive performance than their frequentist counterparts in both frequency and severity analysis. We observed limited evidence of overdispersion favouring the Poisson model in the frequency framework. On the other hand, the Pareto Type II and Lognormal models (Bayesian modelling) performed better than the Pareto Type I distribution. Overall, the Bayesian framework provides a better understanding of the uncertainty, being a complementary tool to the frequentist methods used in Ageas Portugal.

KEYWORDS: Bayesian Modelling; Frequentist Modelling; Large Losses; Predictive Performance; Premium Risk.

TABLE OF CONTENTS

Glossary i

Abstract and Keywords ii

Table of Contents iii

Table of Figures v

Acknowledgments vi

1. Introduction 7

2. Data Preparation 9

2.1. Frequency data 9

2.2. Severity data 11

3. Methodology 12

3.1. Frequentist approach 12

3.1.1. Frequency 12

3.1.2. Severity 14

3.2. Bayesian approach 16

3.2.1. Prior distribution 16

3.2.2. Sampling model 17

3.2.3. Posterior distribution 17

3.2.4. Posterior predictive distribution 18

3.2.5. Markov Chain Monte Carlo 19

3.2.6. Predictive Comparison 20

4. Results 21

4.1. Frequency – frequentist and Bayesian 21

4.1.1. Posterior Predictive Checks 25

4.2. Severity – frequentist and Bayesian 27

4.2.1. Prior and Posterior Analysis 28

4.2.2. Posterior Predictive Checks 34

4.3. Comparison 36

4.4. Conclusions and future work 38

References 40

TABLE OF FIGURES

Figure 1 – Posterior distribution of parameter θ (Poisson) 22

Figure 2 – Posterior distribution of parameter μ (Negative Binomial) 23

Figure 3 – Posterior distribution of parameter ϕ (Negative Binomial)..... 24

Figure 4 – PPC for Frequency predictive behaviour 26

Figure 5 – PPC for the Frequency variance..... 27

Figure 6 – Posterior distribution of Pareto Type I parameter..... 29

Figure 7 – Posterior distribution of Pareto Type II parameters (1st model) 30

Figure 8 – Sensitivity Analysis of parameter B..... 31

Figure 9 – Weakly and Informative prior distributions..... 32

Figure 10 – Posterior distribution of Pareto Type II parameters (final model)..... 32

Figure 11 – Posterior distribution of Lognormal parameters 33

Figure 12 – PPC for Severity predictive behaviour..... 35

Figure 13 – PPC for Severity mean 35

Figure 14 – PPC for Severity standard deviation 36

ACKNOWLEDGMENTS

First, I would like to express my deepest gratitude to my parents and brothers for their unconditional love and support, especially over the past two years. I am fully aware of the effort they made to provide me with the opportunity to move to Lisbon and pursue my Master's degree. This work is therefore dedicated to them.

I would also like to thank my girlfriend, Rosarinho, for her constant encouragement during the most challenging moments of this journey. She has always been a source of strength and a safe haven.

Furthermore, I am grateful to my friends, who have continuously lifted my spirits and supported me whenever I felt discouraged.

Finally, I would like to express my sincere appreciation to my supervisors. I am deeply grateful to Professor Rui Paulo for his guidance and availability throughout this process. Without his invaluable help, this work would not have been possible. I would also like to thank Hélio Silva for his mentorship during my professional experience at Ageas Portugal. I greatly enjoyed working with him and truly appreciate his availability and contribution to this work.

HEADNOTE

By Pedro Calado

This work analyses premium risk in non-life insurance by comparing frequentist and Bayesian approaches, with focus on large losses in the line of business Motor Third-Party Liability. The results show a better predictive performance by the Bayesian models, acknowledging uncertainty, supporting their use as good alternative to frequentist methods.

1. INTRODUCTION

First, it is important to define premium risk to better understand the motivation behind this report. Premium risk represents one of the main sources of uncertainty in non-life insurance, arising from possible differences between expected and actual claim frequency and severity, relative to the premiums charged. This makes the premium risk a very important component of non-life insurance modelling, regarding the capital requirements, pricing and internal risk management decisions. Ageas Portugal models several lines of business, but in this work we will focus on the Motor Third-Party Liability. Within this context, a proper modelling of both frequency and severity distributions with an accurate quantification of uncertainty is of major importance. This motivates the research question of this report: "Does a Bayesian method provide better insights and practical advantages in modelling premium risk, when compared with the traditional frequentist approach currently used?". Bayesian inference considers prior information and posterior distributions for parameters, thereby explicitly quantifying uncertainty about any unknown parameters. Understanding if this approach is a better alternative to model the premium risk when compared to the current firm's approach, is the main motivation for this study.

Regarding the survey of the literature used, Klugman (2008) was useful to provide theoretical background on loss distributions, goodness of fit tests and information criteria used to support model selection and validation. Hoff (2009) details the foundations of Bayesian inference, the interpretation of prior and posterior distributions and the use of simulation methods when there are no closed-form posterior distributions. Kruschke (2015) discusses practical Bayesian analysis and introduces two computational tools for posterior simulation, in particular JAGS and Stan, comparing their applicability. Simpson (2017) explains the concept of the Penalised Complexity priors, which will be very relevant for controlling model complexity and are applied to the Negative Binomial

model. Gelman (2006) provides guidance on prior distributions for variance parameters, which informed the prior distribution in the Lognormal model. Stan Development Team (2026) helped mostly with the *R* code. There we found instructions on how to build different models using Stan and reparameterisation strategies to improve sampling efficiency. Finally, Costa (2017) served as a reference for the formal structure and presentation of this Master's Final Work.

To address the research question, we will perform a comparative analysis of frequentist and Bayesian methods using Motor Third-Party Liability insurance data. The frequentist analysis follows the methodology currently used by Ageas Portugal, while the Bayesian analysis uses the same models within a probabilistic framework, focusing on uncertainty quantification. Posterior inference will be conducted using analytical solutions when possible and simulation methods otherwise. The comparison between both approaches will be carried out through predictive performance measures and model diagnostics.

The contribution of this work lies in providing a comparison between frequentist and Bayesian modelling of premium risk within the context of Ageas Portugal. It will highlight how Bayesian inference may complement the traditional frequentist methods through posterior distributions, posterior predictive checks and predictive performance. This work will explain the insights into the choice of prior distributions, sensitivity analysis and parameter identification to understand the importance of Bayesian methods in actuarial decision making.

Finally, it is important to describe the following Chapters. In Chapter 2, we present the data preparation to model frequency and severity of large losses. In Chapter 3, we describe the methodologies underlying both approaches. In Section 3.1, we present the classical inferential methods used, including parameter estimation and goodness of fit tests. In Section 3.2, we introduce the Bayesian method, covering the specification of prior distributions, sampling models, posterior inference, posterior predictive distributions. We also discuss the use of Markov Chain Monte Carlo methods for models with no closed-form and the predictive performance, explaining the Leave-One-Out Cross-Validation method to provide a common basis for evaluating both approaches. Finally, Chapter 4 displays the practical results obtained in this work, combining

parameter estimation, posterior analysis, posterior predictive checks and predictive comparison. This report concludes with Section 4.4 where we present the main findings and plans for future research.

2. DATA PREPARATION

The calibration process begins with the gathering of historic claim data from the claims database, which is then adjusted with the aim of making it representative of the losses that are expected in the forthcoming year. The main adjustments are the alignment of claim status (open/closed), checking of coherence between payments, outstandings and cost (cost must be equal to the sum of payments and outstandings), inclusion of missing claims from previous data set, and inflation adjustment.

After this initial preparation, the threshold is defined to separate attritional from large losses:

- It should be lower than the retention of the reinsurance treaty (threshold Reinsurance).
- It should be such that the number of claims that exceed the chosen threshold is large enough to fit a Generalized Pareto distribution.

The selection of the thresholds is supported by statistical methods which are: Mean Excess Plot, Hill Plot, Asymptotic MSE Plot and GertensGarbe Plot. Afterwards, four Goodness of Fit tests are performed to select the optimal threshold. It is based on the best rank averages for the following four Goodness of Fit tests: Adjusted R^2, Kolmogorov-Smirnov (K-S), Anderson Darling (AD) and Dynamic Testing System (DTS).

2.1. Frequency data

In order to be able to fit distribution functions, one must build the historical frequency series. The first step is to apply the Chain Ladder method. It begins with the establishment of the actual number of large claims at the beginning and at the end of the development year by accident years. Next, the frequency development factors are derived: the incremental deviation ratio, which measures the proportion of claims that remain classified as large from the beginning to the end of a given development year, and the cumulative deviation ratio, which reflects the proportion of claims that remain classified as large over multiple development years, relative to the initial identification. The

ultimate frequency is adjusted by three measures: no exposure, number of policies and earned premium. With no exposure, it reflects how the frequency behaves, based on historical experience, and is the statistical baseline. With the exposure equal to the number of policies, the analysis estimates the mean risk per policy. On the other hand, with earned premiums, it reflects the financial exposure. Finally, the ultimate frequency is fit by a frequency distribution, using R, for the three measures of exposure. Then the triangle of the sum of large losses is computed, which represents the total number of large claims reported until each development year, and its diagonal will be used to obtain the ultimate frequency:

(1) $Ultimate_i = Diagonal_i \times CDR_i$, for $i = 1, 2, \dots, n$,

where $Ultimate$ stands for the ultimate frequency, $Diagonal$ for the diagonal of the Chain Ladder triangle, CDR for the cumulative deviation ratio and i for the accident year. It represents an estimation of the expected number of large losses for each accident year. Finally, we obtain the final frequency distribution by dividing the ultimate frequency by the type of exposure and multiplying by the projected exposure for the next year:

(2) $Y_i = \frac{Ultimate_i}{Exp_i} \times Exp_{i+1}$,

where Exp_i stands for the observed exposure and Exp_{i+1} for the projected exposure of the next year. The frequency distribution is modelled in two different scenarios: with no exposure (where each $Exp_i = 1$) and with exposure (number of policies, $Exp1$, and premium earned, $Exp2$). So, this leads to three different series:

(3) $$\begin{cases} Y_i = Ultimate_i & , \ if\ there\ is\ no\ exposure \\ Y_i = \frac{Ultimate_i}{Exp1_i} \times Exp1_{n+1} & , \ if\ exposure\ is\ the\ number\ of\ policies \\ Y_i = \frac{Ultimate_i}{Exp2_i} \times Exp2_{n+1} & , \ if\ exposure\ is\ the\ premium\ earned. \end{cases}$$

The frequency of large claims is modelled using the Poisson and Negative Binomial distributions. Since both distributions are defined for non-negative integer values, it was necessary to apply a correction by rounding the observed data, to the nearest integer to obtain the realization of Y_i. Although there were three possible series for the random variable, Ageas Portugal is currently using the specification with no exposure so we will

only use the data $Y_i = Ultimate_i$, which represents the complete development number of observed large claims.

In the case where the ultimate number of large losses in the premium sub risk is modelled with a Poisson distribution, the parameter is the expected frequency for each year:

(4) $Y_i \sim Poisson(\lambda), i = 1, 2, \dots, n$ independently.

In case there is overdispersion, we use a Negative Binomial distribution, which will be explained with more detail in Chapter 3.

2.2. Severity data

The first step is to identify the large losses used in this analysis. There are two ways: a loss is large once it attains the large loss threshold or when the cost/incurred inflated is higher than the large threshold. After this identification, the sum of large loss incurred amounts is computed at the beginning and at the end of each development year. With these results, it is the possible to derive the severity incremental and cumulative development factors.

In this study, only claims with amounts above $800\,000$ are considered, with an additional upper bound of $20\,000\,000$. Consequently, a claim is only classified as large if its cost lies within this interval.

Finally, the incurred ultimate severity is projected and fitted to probability distributions such as Lognormal, Pareto Type I and Pareto Type II. But, before the calibration of these distributions, Ageas Portugal analyses the Mean Excess Loss (MEL), a fundamental tool to evaluate the shape of the tail:

(5) $e(u) = E[X - u | X > u]$,

where u represents the threshold and X the claim amount.

The MEL gives the mean of the claims above the threshold. An increasing MEL function indicates a heavy tailed distribution, which is the case of the Pareto distribution. On the other hand, a decreasing MEL function indicates a light tailed distribution. And finally, a constant MEL function is considered a medium tailed distribution, which is the case of the Exponential distribution. Additionally, the MEL function also helps to validate the

threshold selection, by analysing the regions where the graphic is approximately linear, which indicates where the extreme distribution behaviour starts.

3. METHODOLOGY

Both the frequentist and Bayesian methods are crucial to learn about unknown models underlying observed data. In our case, we focus on modelling the frequency and severity of future large claims. They share some similarities such as likelihood functions and probabilistic models, but differ on how they interpret the probability, the parameters and uncertainty. In Sections 3.1 and 3.2, we will explain both methods and their relevance to this report.

3.1. Frequentist approach

This Section is dedicated to the frequentist approach for both frequency and severity distributions, detailing the procedures implemented by Ageas Portugal.

3.1.1. Frequency

As explained in Chapter 2, the numbers of large losses are seen as possible realizations of a Poisson distribution with an exposure or without. To obtain the estimate of λ, we used the method of maximum likelihood, i.e., the value of λ that maximizes the likelihood function obtained from equation (4). The result is the following:

(6) $\hat{\lambda} = \frac{\sum_{i=1}^{n} Y_i}{n}$,

which is basically the mean of each series represented in equation (3). Finally, to test the assumption that the historical projected ultimate number of large losses follows a Poisson distribution we used the Pearson's Chi-Square test:

(7) $\begin{cases} H_0: data\ is\ a\ realization\ from\ a\ Poisson \\ H_1: otherwise \end{cases}$

(8) $X^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\lambda}.Exp_i)^2}{\hat{\lambda}.Exp_i}$

where in equation (7) we detail the hypothesis test and in equation (8) the test statistic. Under the null hypothesis, X^2 asymptotically follows a Chi-Square distribution with $n - 1$ degrees of freedom. The null hypothesis is rejected with p-value α, if $X^2 > X^2_{n-1}(\alpha)$, where n is the sample size and $X^2_{n-1}(\alpha)$ the critical value of the Chi-Square distribution with $n - 1$ degrees of freedom at significance level α.

In the case where there is overdispersion ($variance > mean$), another distribution could be assumed – the Negative Binomial distribution:

(9) $Y_i \sim Negative\ Binomial(k, p)$, $i = 1, 2, \dots, n$ independently,

where k is the overdispersion parameter and p the success probability parameter. Also, $E[Y] = \mu = k(1-p)/p$ and $Var[Y] = \mu + \mu^2/k$. By observing the variance, if k tends to infinity, the variance approaches that of a Poisson because the mean equals the variance. To estimate both parameters, three methods are used: Maximum Likelihood Estimation (MLE), Method of Moments (MoM) and Deviance Method.

For the MLE, the respective log-likelihood function is equal to:

(10) $l(k, p|y_{i\ \dots\ N}) = Nk\ln(p) - N\ln\big(\Gamma(k)\big) + \sum_{i=1}^{N}\ln\big(\Gamma(k + y_i)\big) + y_i\ln(1-p) - \ln(\Gamma(y_i + 1))$,

where $\Gamma(\cdot)$ represents the Gamma function. When the derivative of the log-likelihood function, with respect to k, is set to zero there is no closed-form solution, so the root must be found with numerical methods. This can be done in RStudio through the *MASS* package, using the function *theta.ml*.

As for the parameter p, is the same procedure as the previous one (derivative with respect to p, this time) and the result is the following:

(11) $p = \frac{N\hat{k}}{N\hat{k} + \sum_{i=1}^{N} y_i} = \frac{\hat{k}}{\hat{k} + \frac{\sum_{i=1}^{N} y_i}{N}} = \frac{\hat{k}}{\hat{k} + \bar{y}}$.

For all the three methods, the equation (11) is used to compute the parameter p.

The Method of Moments consists of equating the sample moments with the population moments, i.e. the mean and the variance. The sample moments are:

(12) $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and

(13) $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$,

where in equations (12) and (13) are the sample mean and the sample variance, respectively. To obtain estimates of the two parameters, one just needs to solve the equations $\bar{y} = E[y]$ and $s^2 = Var[y]$ for k and p. The result from this system of equations is the following:

(14) $\hat{k}_{MoM} = \frac{\bar{y}^2}{s^2 - \bar{y}}$ and

(15) $\hat{p}_{MoM} = \frac{\hat{k}_{MoM}}{\hat{k}_{MoM} + \bar{y}}$.

By looking at equation (14), the result is only possible if $s^2 > \bar{y}$, i.e. when there is some evidence of overdispersion.

Finally, the Deviance Method has a similar flavour as the MoM. Instead of using the populational moments mean and variance, it uses the residual deviance which is equalized to its degrees of freedom. This is done in *R*, using the function *theta.md* where it adjusts the parameter until the following equation is true: $D(\theta) = n - p^*$, where $D(\theta)$ stands for the resulting deviance for a given θ, n for the number of observations and p^* for the number of parameters (in this case only one).

After these estimations for the parameters of both Poisson and Negative Binomial distributions, some criteria are used to address the fit of the models. The first test is called Akaike Information Criterion (AIC) and is defined by:

(16) $AIC = 2k^* - 2\log p(y|\hat{\theta})$,

where k^* represents the number of parameters (1 for the Poisson and 2 for the Negative Binomial) and the log function referrers to the log likelihood. The method with the lowest AIC will be the best one.

The other criterion is called Bayesian Information Criterion (BIC) and it accounts for the number of fitted parameters with a penalty that increases with the sample size. It is given by:

(17) $BIC = k^* \log n - 2\log p(y|\hat{\theta})$,

where n represents the sample size. Again, the method with the lowest BIC will be the best one.

3.1.2. Severity

As previously explained, we use three distributions for the severity analysis: Lognormal, Pareto Type I and Pareto Type II. Given the multiple parameterisations of the Pareto distributions, we state the density functions adopted in this work, to avoid ambiguity. The Pareto Type I and Type II distributions are defined, respectively, by:

(18) $f(x) = \frac{\alpha L^{\alpha}}{x^{\alpha+1}}, x > L$ and

(19) $f(x) = \frac{\alpha B^{\alpha}}{(x+B)^{\alpha+1}}, x > 0.$

In equation (18), L corresponds to the minimum possible value of the claims severity X, i.e. the threshold, and α to the shape parameter. Additionally, in equation (19), the parameters α and B correspond, respectively, to the shape and scale parameters.

As seen in Section 2.1.2, the MEL gives the mean of the claims above the threshold. After the MEL analysis, we approach the fit of the models using three methods: Minimum Goodness of fit Estimator (MGE), MLE and Sum of Squared Quantiles (SSQ). To determine the best distribution, we used QQ-plots that allow us to evaluate the fit of the distribution to the ultimate severities. These plots compare the empirical claim quantiles with the estimated distributional quantiles. When the data follow the diagonal, where empirical and theoretical quantiles are equal, it means that the probability distribution captures efficiently the severity behaviour.

Beside these plots, we also conducted some hypothesis tests to assess the fit of each distribution, like the Kolmogorov-Smirnov (K-S) and Anderson-Darling (AD) tests. In Klugman (2008), these tests are formulated and explained. Both have the same hypothesis test:

(20) $\begin{cases} H_0: data\ came\ from\ the\ stated\ population. \\ H_1: otherwise. \end{cases}$

Let t be the left truncation point and u the right censoring point, the K-S statistic is given by $D = \max_{t \le x \le u} |F_n(x) - F^*(x)|$, which is the maximum absolute difference between the empirical distribution function ($F_n(x)$) and the cumulative distribution function ($F^*(x)$) under the null hypothesis. The critical values are $1.22/\sqrt{n}$ for $\alpha = 10\%$, $1.36/\sqrt{n}$ for $\alpha = 5\%$ and $1.63/\sqrt{n}$ for $\alpha = 1\%$. So, it has the advantage that it only depends on the sample size and not on the type of distribution being tested.

On the other hand, the AD test uses a different test statistic, given by:

(21) $A^2 = n \int_t^u \frac{[F_n(x) - F^*(x)]^2}{F^*(x)[1 - F^*(x)]} f^*(x) dx,$

which is the integrated, between the left truncation point (t) and the right censoring point (u), of the squared difference between the empirical ($F_n(x)$) and theoretical distribution functions ($F^*(x)$). The critical values are 1.933 for $\alpha = 10\%$, 2.492 for $\alpha = 5\%$ and 3.857 for $\alpha = 1\%$. According to Klugman (2008), this test is more sensitive to the tail behaviour than the K-S test, so can be seen as a more suitable test on the fit of the severity distributions.

3.2. Bayesian approach

In the Bayesian approach, probability is interpreted as a degree of rational beliefs and not just as a frequency in repeated experiments. So, available information and probability are related, and the Bayes' rule helps to put that in practice by updating a probability given that information is available. According to Hoff (2009), Bayesian Inference is the «process of inductive learning via Bayes' rule.». The Bayesian method, derived from the Bayesian Inference, is a tool that gives a good description of current data, estimates parameters, forecasts future data and provides statistical methods for model estimation and validation. Given the Bayes' rule, the information update is formalized, as seen in Hoff (2009), in the following equation:

$$(22)\ p(\theta|y) = \frac{p(y|\theta)\,p(\theta)}{\int_{\Theta} p(y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}.$$

Equation (22) represents the posterior distribution $p(\theta|y)$. In the numerator, $p(\theta)$ stands for the prior distribution and $p(y|\theta)$ for the sampling model, and the denominator stands for the prior predictive distribution. The variables are θ representing the parameter value, y the dataset and Θ the parameter space (set of all possible values for the parameter θ). Next, we will explain in more detail these distributions.

3.2.1. Prior distribution

In simple words, the prior distribution represents the initial beliefs about the unknown parameter, before observing the data. In Hoff (2009) the Gamma distribution is frequently used as the prior distribution when the data follows a Poisson distribution. For a certain positive parameter, the respective distribution is

$$(23)\ p(\theta) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta},\ \theta > 0,$$

where variables in equation (23) are b, representing the number of prior observations, a sum of counts from b and $\Gamma(\bullet)$ the Gamma function. So, it is very important to choose the variables that will be incorporated in the prior distribution because they determine the amount of prior information incorporated and directly impact the resulting posterior distribution.

In equation (22), the posterior is obtained by combining the prior with the sampling model and so the prior distribution plays a central role in the final posterior distribution. When the prior and posterior belong to the same class of prior distributions, that class is called conjugate of the statistical model. A relevant example in Hoff (2009) is the Poisson model and the Gamma prior. By combining a Poisson sampling model with a prior Gama distribution, we obtain a posterior Gamma distribution. This is very useful because it simplifies the computation, helps to interpret the posterior and avoids the use of Markov Chain Monte Carlo (MCMC) methods which will be later discussed.

3.2.2. Sampling model

In Hoff (2009), one of the relevant sampling models is the Poisson distribution, and is the same distribution used in Ageas Portugal. In equation (24) we present a Poisson distribution with mean θ:

(24) $\Pr(Y = y|\theta) = \frac{\theta^y e^{-\theta}}{y!}$, for $y \in \{0, 1, 2, \dots\}$.

In this case, with $y = y_1, \dots, y_n$ we get that the sampling model in equation (22) simplifies to:

(25) $p(y|\theta) \propto \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta}$.

3.2.3. Posterior distribution

In equation (22), the denominator can also be represented by $p(y)$ and is called the prior predictive distribution. Basically, this term is used to normalize the posterior distribution, i.e. to make sure the posterior integrates to 1. In that sense, the posterior distribution can be written in the following way:

(26) $p(\theta|y) \propto p(y|\theta)p(\theta)$

So, this distribution contains all the available information about the parameter after observing the data. It is the core of Bayesian Inference. As previously seen, when we

combine a Poisson (sampling model) with a Gamma (prior) the result would be also a Gamma distribution (posterior). In that case, the posterior distribution has parameters $a^* = a + \sum_{i=1}^{n} y_i$ and $b^* = b + n$. From equation (26), the posterior becomes

(27) $p(\theta|y) \propto \theta^{a-1+\sum_{i=1}^{n} y_i} e^{-(b+n)\theta}$, $\theta > 0$,

Hence, there is no need to use computation methods like MCMC to obtain the posterior distribution. Once this distribution is obtained, every estimation, decisions and comparisons derive from it.

3.2.4. Posterior predictive distribution

The posterior predictive distribution is one of the key elements of Bayesian Inference. While the posterior distribution describes the uncertainty about the parameter after observing the data, the posterior predictive distribution describes the uncertainty about new observations. It is the probability distribution of yet to observe data, given what has already been observed. This is very important, especially in motor insurance, in order to predict future large losses and be prepared in terms of solvency. So, the posterior predictive distribution, as the name says, depends heavily on the posterior distribution:

(28) $p(\tilde{y}|y) = \int p(\tilde{y}|\theta) p(\theta|y)\, d\theta$,

where $\tilde{y}$ stands for the additional data. Here, we assume that $\tilde{y}$ is conditionally independent of y, given θ.

Continuing with the posterior distribution shown in equation (27), if we want to predict a Poisson-distributed random variable, then the posterior predictive distribution turns out to be a Negative Binomial (after some analytical computation) with parameters $a + \sum y_i$ and $\frac{b+n}{b+n+1}$, i.e.

(29) $\tilde{Y}|y \sim NB\left(k = a + \sum y_i\,, p = \frac{b+n}{b+n+1}\right)$,

where the parameters k and p (introduced in Section 3.1.1) stand for, respectively, the number of successes and probability of success. This case, it turns out to be a well-known distribution. Otherwise, one must resort to simulation methods to compute the posterior predictive distribution.

3.2.5. Markov Chain Monte Carlo

In this report, the Bayesian inference for the models without closed-form posterior distribution will be performed using MCMC simulation methods implemented in Stan via the *RStan* interface in *R*. JAGS is another software that can also perform these computations but, according to Kruschke (2015), Stan can be more effective than JAGS, especially for large complex models, where JAGS can become too slow. These algorithms generate samples from the posterior distribution through a Markov chain whose stationary distribution corresponds to the target posterior distribution. The goal is to produce many simulations for the parameter according to the posterior distribution in a way that the set of samples represents the posterior.

To overcome some limitations, Stan applies a variation of the Metropolis-Hastings algorithm: the Hamiltonian Monte Carlo (HMC). HMC detects the direction in which the posterior distribution increases (the gradient) and uses this information to guide the proposal distribution towards regions of higher probability, so that the proposed moves follow the geometry of the posterior surface. This allows the Markov Chain to take large and informed steps through the parameter space, while maintaining a high acceptance rate. However, HMC depends on two specified parameters: the step size ε and the number of leapfrogs steps L, and the choice of either one influences heavily the efficiency, which makes it difficult to apply the method without substantial manual tuning.

To address this problem, the No-U-Turn Sampler (NUTS), introduced by Hoffman and Gelman (2014), provides a solution by eliminating the need to choose a fixed value of L. It has the capacity to identify the moment when the simulated Hamiltonian trajectory begins to retrace its path and, in that case, stops further progresses. This condition is called "U-Turn", and it occurs when the trajectory begins moving back toward earlier positions. To know when to stop running the simulation of steps, NUTS builds a binary tree that represents the Hamiltonian trajectory, doubling the number of leapfrog steps and checking the U-Turn condition in every stage. NUTS introduces a slice variable u sampled from:

(30) $u \sim Uniform\left(0, \exp\left\{\mathcal{L}(\theta) - \frac{1}{2} r \,.\, r\right\}\right)$,

where $\mathcal{L}(\theta)$ represents the log posterior and r the momentum variable. It will ensure that all candidate states lie within the same slice, i.e., $u \leq \exp\left\{\mathcal{L}(\theta) - \frac{1}{2} r \,.\, r\right\}$, simplifying the acceptance condition. It constructs a set of potential states C, from the visited ones, during the trajectory, excluding the states that violate the slice constraint or the detailed balance constraints from the recursive tree expansion. At the end of each iteration, the next sample is drawn uniformly from the set C. Finally, the step size is fixed in order to preserve the stationarity of the Markov Chain and make sure that NUTS is fully automated and does not require manual intervention.

3.2.6. Predictive Comparison

To compare the Bayesian models with the frequentist models, we will use a method called Leave-One-Out Cross-Validation (LOO-CV), which is widely used for analysing the predictive performance of statistical models. Basically, it evaluates how good a model predicts new data by iteratively removing each observation from the dataset and computing its predictive density conditional on the remaining data. Gelman (2006), defines a Bayesian estimate of out-of-sample predictive fit, called expected log pointwise predictive density (*elpd*) given by:

(31) $elpd_{loo} = \sum_{i=1}^{n} \log p(y_i|y_{-i})$,

where n represents the number of data points and $p(y_i|y_{-i})$ represents the leave-one-out predictive density given the data without the i-th observation. This measure reflects how the model is expected to perform on future data sampled from the same data generating process.

A disadvantage of this approach is the need to fit the model n times, once for each removed data point, which is computationally harsh for models using HMC. To address this problem, the Pareto-Smoothed Importance Sampling LOO (PSIS-LOO) method was proposed. It approximates the LOO method using only one single model fit and stabilizes the importance weights based on the generalized Pareto distribution. Importance sampling is a technique used to approximate expectations under a target distribution by weighting again samples from another distribution, with the weights correcting the discrepancy between the two distributions. This algorithm proceeds in the following way:

1. Compute the raw importance ratios and fit a generalized Pareto distribution to the 20% largest importance ratios.
2. Replace the largest ratios by their expected values of the order statistics of the fitted generalized Pareto distribution. These new weights will be given as $\widetilde{w}_i^s$, where s is the number of simulation draws.
3. Truncate the smoothed weights.

These steps are performed for each data point i and the result is a vector of weights w_i^s, for $s = 1, \dots, S$. Then it is possible to compute the PSIS estimate of the LOO expected log pointwise predictive density:

(32) $\widehat{elpd}_{psis-loo} = \sum_{i=1}^{n} \log\left(\frac{\sum_{s=1}^{S} w_i^s p(y_i|\theta^s)}{\sum_{s=1}^{S} w_i^s}\right)$.

The estimated shape parameter k of the fitted generalized Pareto distribution is also a key diagnostic. Its value indicates whether the importance sampling approximation is reliable:

- If $k < 0.5$, the variance of the raw importance ratios is finite, so is very reliable.
- If $0.5 < k < 0.7$, it is acceptable, but with some uncertainty.
- If $k > 0.7$, it is unreliable.

4. RESULTS

In this final Chapter of the main text, we will explain the results for both frequency and severity distributions, the comparison between each method and the respective conclusions. The results below are supported by the instructions found in Stan Development Team (2026), which was crucial in the implementation of the models in Stan within the *R* environment.

4.1.Frequency – frequentist and Bayesian

In Table I we represent the frequency inputs for this analysis.

TABLE I

FREQUENCY DATA

Observation (#)	Accident Year (i)	Frequency (Y_i)
1	2010	7
2	2011	9

3	2012	8
4	2013	6
5	2014	8
6	2015	5
7	2016	5
8	2017	7
9	2018	10
10	2019	9
11	2020	2
12	2021	6
13	2022	4
14	2023	3
15	2024	6
16	2025	0

After the construction of the initial data explained in Section 2.2.1, the next step was to define the prior distribution for the Poisson and Negative Binomial distributions. It has already been explained that the pair Poisson model and Gamma prior leads to a closed-form posterior distribution, eliminating the need to use MCMC. Given the limited amount of information available (only 16 observations, as shown in Table I), the Gamma distribution was specified to be a weakly informative prior distribution, in order to not allow prior dominance over the posterior inference. The prior used corresponds to a $Gamma(a = 0.001, b = 0.001)$. In Figure 1, we display both the posterior and prior distributions of the Poisson Bayesian model:

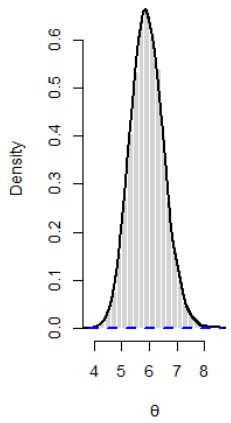

Figure 1 – Posterior distribution of parameter θ (Poisson)

By observing the blue dashed line, which corresponds to the prior distribution, we notice that inference is clearly driven by the observed data rather than by the prior specification. The black line, which corresponds to the posterior distribution, shows limited dispersion, suggesting that there is a great deal of learning about θ.

On the other hand, the Negative Binomial is more challenging due to the overdispersion parameter. There are different parameterisations of this distribution and, in this case, we used the $Negative\ Binomial(\mu, k)$, where μ stands for the mean parameter and k for the overdispersion parameter, which governs the relationship between the variance and the mean. As k increases, the distribution converges to a Poisson distribution. For the prior distribution of μ, it was logical to apply the same prior as for the case of the Poisson distribution, i.e., a $p(\mu) \sim Gamma(a = 0.001, b = 0.001)$. Regarding the overdispersion, two Negative Binomial models were considered to assess the impact of modelling uncertainty in the parameter. In the first model, the overdispersion parameter was fixed at its MLE. In the second model, k was treated as an unknown parameter and it was used a different parameterisation: $Negative\ Binomial(\mu, k = 1/\phi)$, where ϕ influences the level of overdispersion. This parameter is then reparameterised as $\phi = \psi^2$ to improve numerical stability. For ψ we used a Penalised Complexity (PC) prior via an Exponential distribution (with parameter equal to 1). According to Simpson (2017), the PC prior controls the model complexity by penalising deviations from a simpler model, in this case the Poisson model. If there is no overdispersion, the prior moves the model towards the Poisson. Figure 2 represents the posterior distribution of μ for both models:

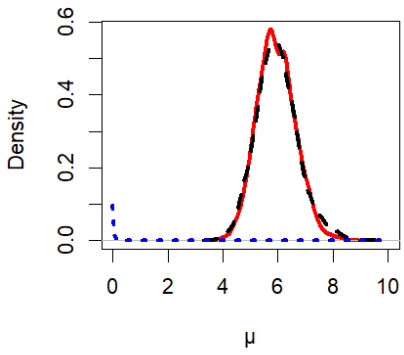

Figure 2 – Posterior distribution of parameter μ (Negative Binomial)

The blue dashed line corresponds to the prior Gamma (ϕ fixed at the MLE), the red line to the posterior distribution of the first model and the black dashed line to the posterior distribution of the second model (unknown ϕ). It can be seen that the posterior distribution of μ is identical in both models. This output can be explained by the small sample size because there is limited information regarding the frequency, and by the weak dependence between μ and k, which implies that inference on μ is robust to the modelling choice of k. Furthermore, the Gamma prior has a behaviour as expected since it has small impact on the posterior distribution.

Figure 3 shows the comparison between the PC prior distribution and the posterior distribution of the parameter ϕ:

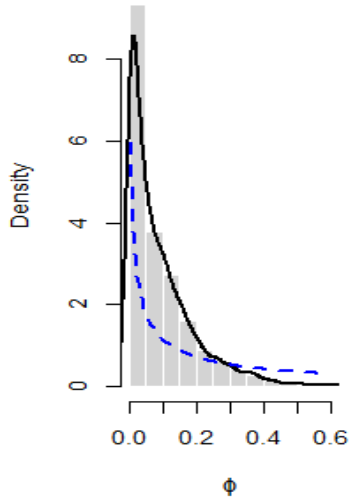

Figure 3 – Posterior distribution of parameter ϕ (Negative Binomial)

Figure 3 shows that the posterior density of ϕ is highly concentrated near zero, indicating small evidence of overdispersion. As previously explained, when the parameter $k = 1/\phi$ tends to infinity, the Negative Binomial distribution will converge to a Poisson distribution. In this case, since ϕ is close to zero, will lead to higher values of k through the expression $k = 1/\phi$. Additionally, the posterior distribution (black line) is sharper than the prior distribution (blue line) which indicates that the data does provide information about ϕ. This behaviour is consistent with the motivation of PC prior

distributions that penalize unnecessary model complexity and favour simpler models unless additional complexity is supported by the data.

From the previous results, the Bayesian approach yields full posterior distributions that quantify parameter uncertainty, whereas the frequentist approach relies on MLE. Table II aims to present posterior mean estimates alongside the corresponding MLE for both models to compare the degree of alignment between both inferential frameworks:

TABLE II

PARAMETERS COMPARISON

	θ	μ	k
MLE	5.9375	5.9375	24.2369
Posterior Mean	5.9265	5.9812	10.7213

From Table II, we observe that the posterior means are quite similar to those obtained with MLE. The only exception is the parameter k which is substantially different, which is mainly due to the approximation of the Negative Binomial to a Poisson distribution.

4.1.1. Posterior Predictive Checks

In this Section, we analyse the Posterior Predictive Checks (PPC) for the Poisson and Negative Binomial models, to provide evidence regarding the adequacy of each model in reproducing key distributional features of the data. It is based on how the models predicts certain aspects of the data and how that prediction compares to the actual observed data.

The first PPC is based on the *ppc_bars* function in R which compares the posterior distribution of replicated data (y_{rep}), by observing if is close or not to the observed the data (y). So, the observed frequencies for each outcome are represented by bars, while the model predictions are obtained by simulating replicated datasets $y_{rep} \sim p(y_{new}|y)$, where y_{new} represents a future observation drawn from the posterior predictive distribution. The plots display the proportion of occurrences in the replicated data and the corresponding credible intervals, reflecting the uncertainty regarding the model predictions. Figure 4 shows that both models produce similar predictive behaviour.

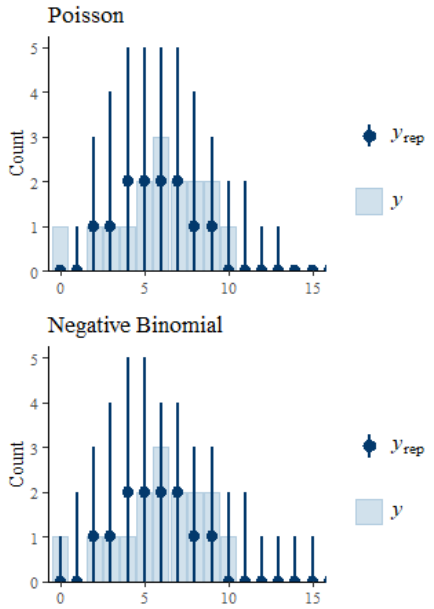

Figure 4 – PPC for Frequency predictive behaviour

The main difference is observed in the tail of the distribution, where the Negative Binomial model generates a higher frequency of extreme values compared to the Poisson model. Here, the small sample size is relevant.

Another important PPC relates to the sample variance. In theory the variance of the Poisson is equal to the mean and the Negative Binomial allows for larger values of the variance, when compared with the mean. Figure 5 confirms this:

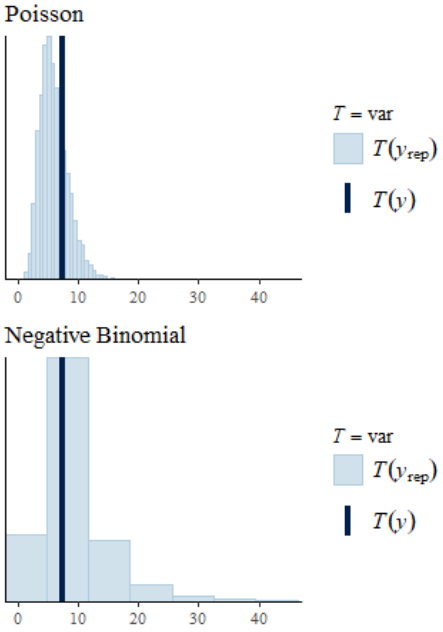

Figure 5 – PPC for the Frequency variance

Under the Poisson, the posterior predictive distribution of the variance is very concentrated, with limited variability. On the other hand, the Negative Binomial exhibits a wider posterior predictive distribution for the variance, reflecting much more uncertainty about this parameter. Both models seem adequate, but the more flexible model has more uncertainty associated with the posterior predictive distribution of the sample variance.

4.2.Severity – frequentist and Bayesian

Given the extensive set of inputs used in this analysis, Table III presents the key summary statistics in order to provide an overall context for the severity data:

TABLE III

SUMMARY STATISTICS OF TRUNCATED CLAIM SEVERITIES

Sample size	Minimum	Maximum	Mean	Standard deviation	25% percentile	Median	95% percentile
78	807 442	8 564 870	1 675 706	1 145 423	1 102 473	1 318 702	3 225 378

Since the severity data is only observed within a restricted interval, the likelihood must account for both left and right truncation. Let X be the claims severity, the observations must satisfy this condition: $L < X < U$, where L represents the lower limit

and U the upper limit. Consequently, the likelihood is given by the density of the truncated distribution:

(33) $f_T(x|\theta) = \frac{f(x|\theta)}{F(U|\theta)-F(L|\theta)}$,

where $f(x|\theta)$ represents the probability density function and $F(x|\theta)$ the cumulative distribution function.

On the other hand, the cumulative distribution function of the truncated model is given by:

(34) $F_T(x|\theta) = \frac{F(x|\theta)-F(L|\theta)}{F(U|\theta)-F(L|\theta)}$.

This expression allows the generation of replicated observations using the inverse transform sampling. For each posterior draw $\theta^{(s)}$, a random variable $u \sim Uniform(0,1)$ is generated and the replicated value is obtained by solving:

(35) $u = \frac{F\left(x_{rep}|\theta^{(s)}\right)-F\left(L|\theta^{(s)}\right)}{F\left(U|\theta^{(s)}\right)-F\left(L|\theta^{(s)}\right)} \Leftrightarrow$

$$x_{rep} = F^{-1}\left(F\left(L|\theta^{(s)}\right) + u\left[F\left(U|\theta^{(s)}\right) - F\left(L|\theta^{(s)}\right)\right]\right),$$

where x_{rep} represents the replicated value from the posterior predictive distribution. This formulation is needed in the severity framework in order to generate samples from truncated distributions.

4.2.1. Prior and Posterior Analysis

The first step in this Section is to define appropriate prior distributions for each severity model. In the case of the Pareto Type I distribution, it is well known that a Gamma prior assigned to the shape parameter is conjugate with the Pareto likelihood which gives a closed-form for the posterior distribution. Although this conjugacy no longer holds when truncation is introduced, it was nevertheless used a Gamma distribution as a prior distribution for the shape parameter α. Regarding the specification of the Gamma prior distribution, the statistical literature suggests that the value of its parameter plays an important role in the existence of moments. In particular, for a Pareto Type I, if $\alpha \leq 1$, the mean is infinite, implying an uninsurable risk. Moreover, if $\alpha \leq 2$, the variance does not exist and, for heavy tailed distribution, variance may no longer be

a meaningful risk measure (see Klugman, 2008). Based on these observations, we used a $\alpha \sim Gamma(a = 1.5, b = 1)$. The mean and variance of a Gamma distributions are given by a/b and a/b^2, respectively. In this case, the prior has a mean equal to 1.5 and a variance equal to 1.5 which gives a very plausible result, since the mean will not be infinite. In addition, several Gamma prior distributions were considered to assess the sensitivity of the posterior results to the choice of hyperparameters. In Figure 6, the posterior results are reported:

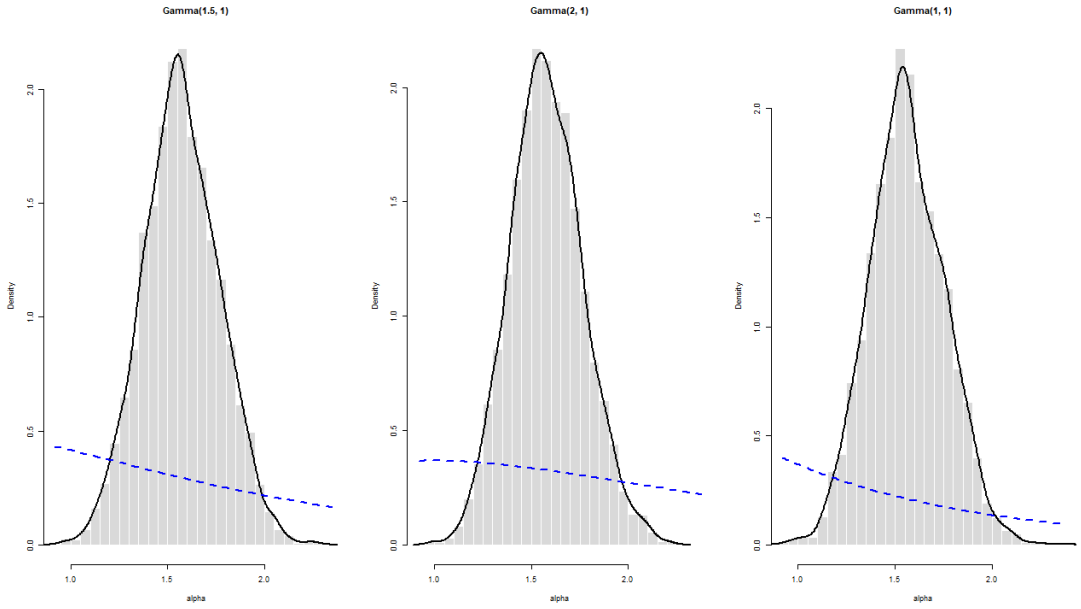

Figure 6 – Posterior distribution of Pareto Type I parameter

According to Figure 6, the posterior distributions of the Pareto Type I shape parameter are very similar across the different Gamma prior distributions. The posterior presents limited sensitivity to the choice of prior, indicating that the likelihood dominates the inference. So, comparable posterior results are obtained when using a $\text{Gamma}(1.5, 1)$ prior (left panel), a $\text{Gamma}(2, 1)$ prior (centre panel) or a $\text{Gamma}(1, 1)$ prior (right panel).

In the case of the Pareto Type II distribution, two parameters are considered: the shape parameter α and the scale parameter B, also referred to as the *Ballast* parameter. For the shape parameter, the same strategy adopted in the Pareto Type I case was followed, and a prior Gamma distribution was assigned to that parameter, namely $\alpha \sim Gamma(a = 1.5, b = 1)$. In contrast to the shape parameter, the scale parameter possesses different

statistical properties. It reflects, in this context, the typical level or magnitude of large claims, thus capturing the scale of the observed severity data. Initially, we used a prior Normal distribution for the scale parameter with mean equal to 1 000 000 and standard deviation equal to 500 000, assuming the belief that the majority of large claims is around one million while still allowing for substantial variability. However, convergence issues were observed while fitting the model, as indicated by Stan warnings, regarding poor chain mixing and lack of convergence. To address this issue, a reparameterisation of the scale parameter was adopted, by modelling it on the logarithmic scale, to improve numerical stability and sampling efficiency. So, the prior distribution was formulated as $\log B \sim Normal(\mu = 13.82, \sigma = 0.472)$, which is equivalent to $B \sim Lognormal(\mu = 13.82, \sigma = 0.472)$. The chosen hyperparameters correspond to a prior with mean of approximately 1 000 000 and standard deviation of approximately 500 000 (these results were obtained from the expressions for the moments of the Lognormal distribution). Figure 7 displays the posterior distribution as well as the prior distribution of both parameters α and B:

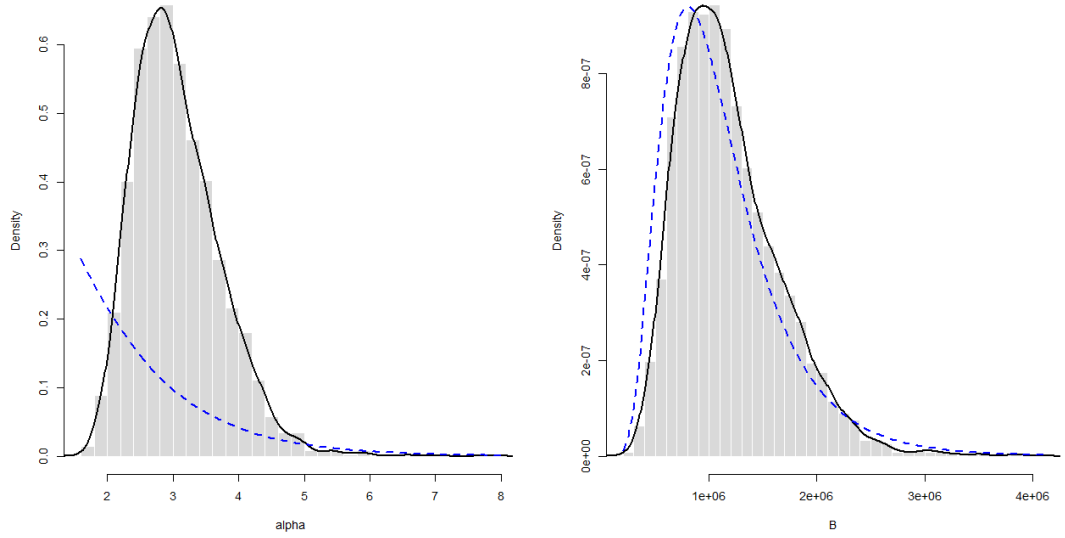

Figure 7 – Posterior distribution of Pareto Type II parameters (1st model)

Starting with the left panel, we display the shape parameter α and it is clear the update of information for this parameter. The prior (dashed line) differs substantially from the posterior (solid line), meaning that the inference for the shape parameter is highly driven by the data. This suggests that the parameter α is being learned from the observed claims,

which is a desirable feature in a Bayesian framework. In contrast, the scale parameter B (right panel) shows a posterior distribution almost identical to the prior distribution, suggesting a lower level of information in the data regarding this parameter. Because of this behaviour, a sensitivity analysis was conducted to further assess the characteristics and identifiability of the parameter B. To perform this analysis, we fixed α at the MLE and the log-likelihood function was evaluated for different values of B. Figure 8 represents these plots:

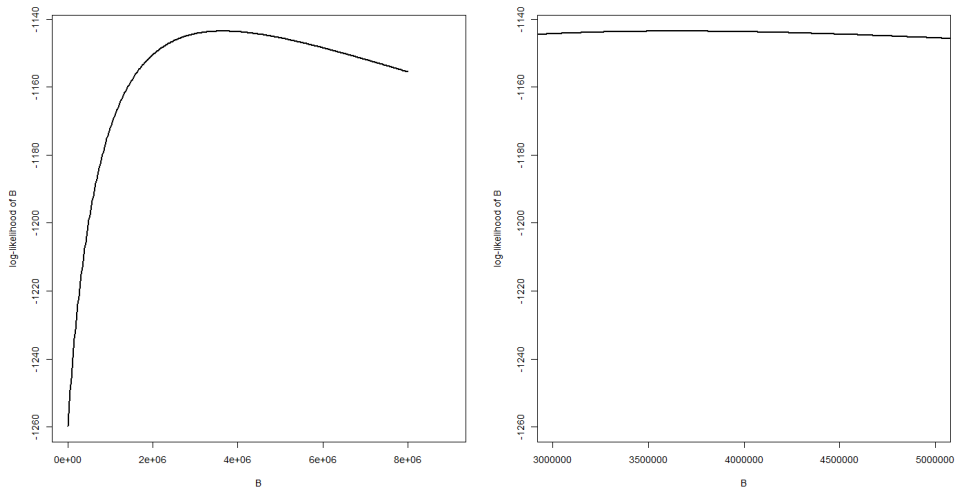

Figure 8 – Sensitivity Analysis of parameter B

In both panels, we have the log-profile likelihood of the scale parameter B. It can be observed that, between $B = 3\,000\,000$ and $B = 5\,000\,000$, approximately, the log-likelihood remains nearly constant, indicating a very flat likelihood surface over this range. In the right panel, we represent a detailed view of this region, confirming that substantial changes in the value of B does not have an impact on the log-likelihood. Consequently, the scale parameter possesses weak identifiability and a high level of uncertainty. In a Bayesian setting, this implies that the posterior distribution of B will be highly dispersed and sensitive to the prior specification. Therefore, this sensitivity analysis justifies the use of weakly informative prior for the scale parameter, as different priors lead to potentially different posterior distributions. Figure 9 compares a weakly informative prior with an informative prior distribution for the scale parameter:

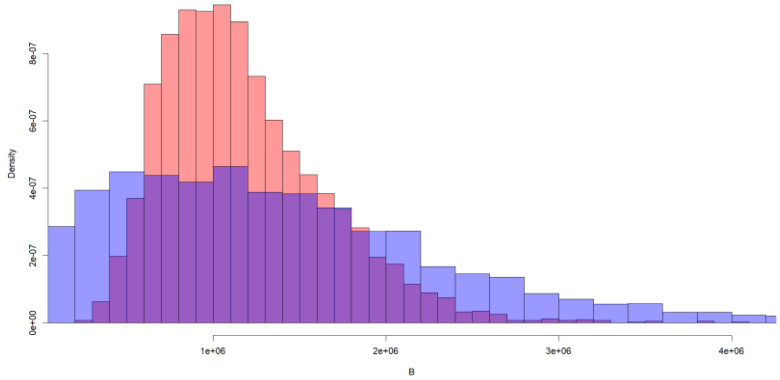

Figure 9 – Weakly and Informative prior distributions

In Figure 9, the blue region corresponds to the posterior obtained from the weakly informative prior $B \sim Lognormal(\mu = 13.82, \sigma = 2)$. The chosen hyperparameters define a prior with a mean of approximately $1\,000\,000$ and a very large standard deviation of approximately $55\,000\,000$, thus providing a very wide range of possible values for B. On the other hand, the red region corresponds to the posterior obtained from the same informative prior distribution displayed in Figure 7. It can be observed that the posterior distributions are very different regarding both tail and dispersion behaviour. In particular, the posterior obtained from the weakly informative prior exhibits a much wider spread and a heavier tail. Since the truncation of the data makes it hard to learn about B and the uncertainty surrounding this parameter is mainly driven by prior assumptions, we use the less informative (weakly) prior distribution. Figure 10 presents the final posterior distributions of both parameters of the Pareto Type II distribution:

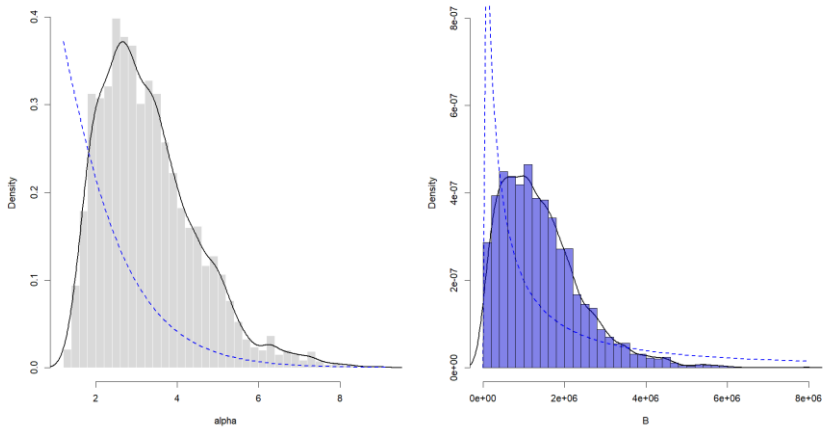

Figure 10 – Posterior distribution of Pareto Type II parameters (final model)

On the left we represent the posterior distribution of the parameter α and, on the right, the posterior distribution of the parameter B. Both corresponding prior distributions are represented by the blue dashed line. Overall, the results indicate a clear updating of prior information by the data.

The last part of this Section concerns the Lognormal distribution with parameters μ and σ, defined as $\log X \sim Normal(\mu, \sigma)$, which is the third and final severity distribution used by Ageas Portugal for the analysis of large losses. The parameter μ represents the mean of the logarithm of the claim size, therefore we used a prior specification similar to the one used for the scale parameter in the Pareto Type II (before the sensitivity analysis). Specifically, $\mu \sim Normal(\mu = 13.82, \sigma = 0.472)$, which corresponds to a Lognormal distribution for X with a prior mean of approximately one million and a standard deviation of roughly five hundred thousand. This choice reflects prior knowledge regarding the typical magnitude of large claims in this portfolio. For the parameter σ, which must be positive and represents the dispersion of the logarithm of the claim size, we used a half-Cauchy prior distribution $\sigma \sim Cauchy^{+}(0, 1)$. The use of this type of distributions has been recommended in the Bayesian literature as a weakly informative alternative. According to Gelman (2006), the half-Cauchy prior combines substantial mass near zero with heavy tails that allows the data to dominate when larger values of the parameter are supported. This helps to control inference without imposing restrictive assumptions on the variability of large losses. Figure 11 represents the posterior distribution of both parameters.

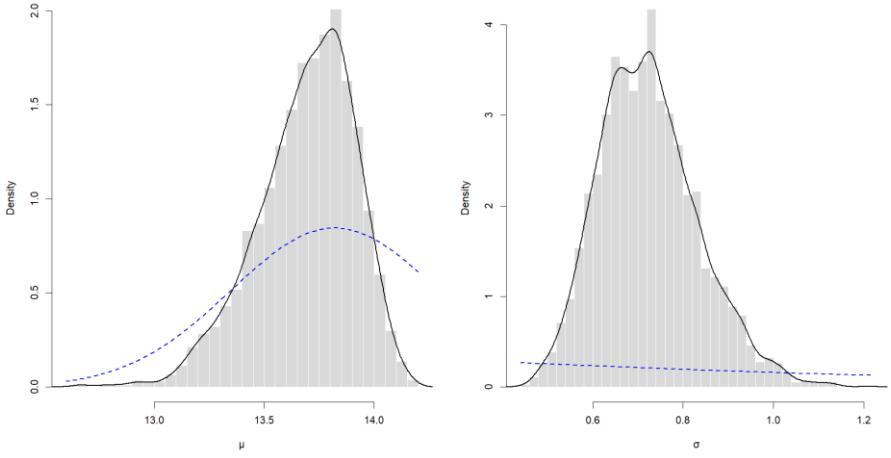

Figure 11 – Posterior distribution of Lognormal parameters

On the left we display the posterior distribution of parameter μ (note that it is represented on the logarithm scale) and the prior distribution, represented by the dashed line, is well below the posterior distribution which is a good sign of update of information. On the right, we have the posterior distribution of parameter σ and in this case the prior distribution is flat with very low density levels, representing the expected outcome of the weakly informative prior.

To conclude this Section, Table IV presents a comparison between the parameter estimates of the three severity distribution obtained using the MLE and Bayesian inference. It should be noted that, unlike the MLE approach, Bayesian inference does not yield a single point estimate for each parameter but rather a full posterior distribution. So, Table IV shows a summary of these distributions, in particular, the posterior mean of each parameter:

TABLE IV

PARAMETERS COMPARISON

	Pareto Type I	Pareto Type II		Lognormal	
	α	α	*B*	μ	σ
MLE	1.5756	6.1092	3 646 082	13.7482	0.6939
Posterior mean	1.5708	3.2917	1 407 948	13.6905	0.7240

From Table IV, the posterior means are very close to the corresponding MLE for both the Pareto Type I and the Lognormal distributions. It suggests that, for these models, the data is highly informative and the Bayesian estimates are closely aligned with their frequentist counterparts. Only in the case of the Pareto Type II distribution the parameters are much more different. It highlights the importance of careful prior specification when modelling large losses with this distribution. Since the data is censored above a threshold and B is related with the tail behaviour makes it hard to model large losses with this distribution.

4.2.2. Posterior Predictive Checks

In this Section, we analyse the PPC for the three severity models, to provide evidence regarding the adequacy of each model in reproducing key distributional features of the data.

The first PPC is based on the *ppc_dens_overlay* function in *R* which compares the empirical density of the observed data x with densities obtained from replicated datasets

generated from the posterior predictive distribution. In total, 200 posterior predictive replicates were randomly selected from $p(x_{rep}|x)$. Figure 12 shows that all three models produce similar predictive behaviour:

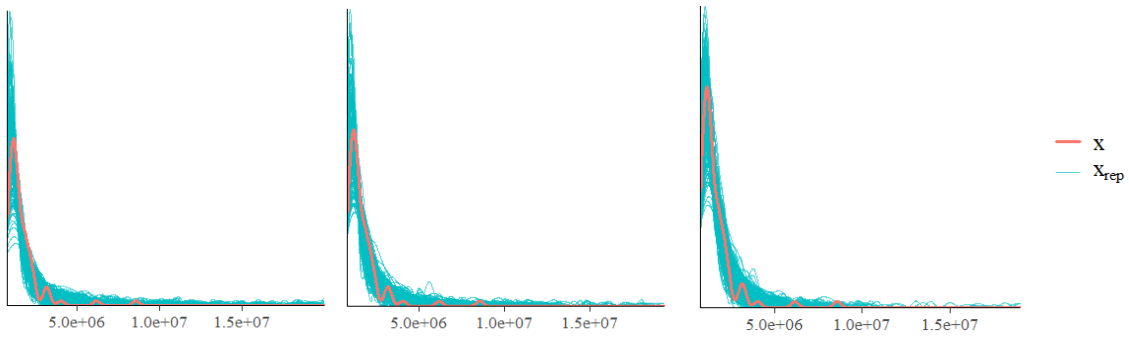

Figure 12 – PPC for Severity predictive behaviour

Here, we display the density of the posterior predictive replicated data in blue and the observed data in red for the Pareto Type I model (left panel), the Pareto Type II model (centre panel) and the Lognormal model (right panel). All the three distributions seem to have the same predictive behaviour when compared to the observed data and produce distributions with comparable shapes. This similarity may be explained by the presence of upper truncation in the severity data. By conditioning the analysis on an upper limit, some valuable information related to the extreme tail behaviour may be removed. So, potential differences beyond the truncation threshold cannot be identified, which leads to similar posterior predictive behaviour across models.

The next PPC regards the comparison between the sample mean with the posterior predictive distribution of the mean to see if the models reproduce efficiently the central tendency. This comparison is illustrated in Figure 13:

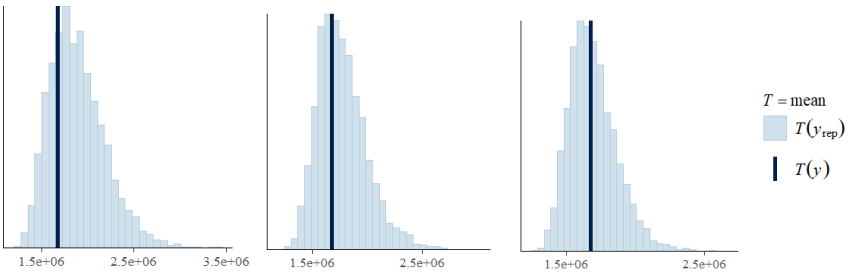

Figure 13 – PPC for Severity mean

Although the legend displays $T(y_{rep})$ and $T(y)$, these quantities correspond respectively to $T(x_{rep})$ and $T(x)$, following the internal notation adapted by the graphical functions. Only the Pareto Type I (left panel) model has an observed sample mean slightly shifted relative to its posterior predictive distribution. Both the Pareto Type II (centre panel) and the Lognormal models (right panel) present posterior predictive means around the respective mean samples, indicating a better reproduction of the central tendency of the data.

The final PPC is like the previous one, but in this case compares the sample standard deviation with the posterior predictive distribution of the standard deviation to see if the models reproduce efficiently the dispersion observed in the data. This comparison is illustrated in Figure 14:

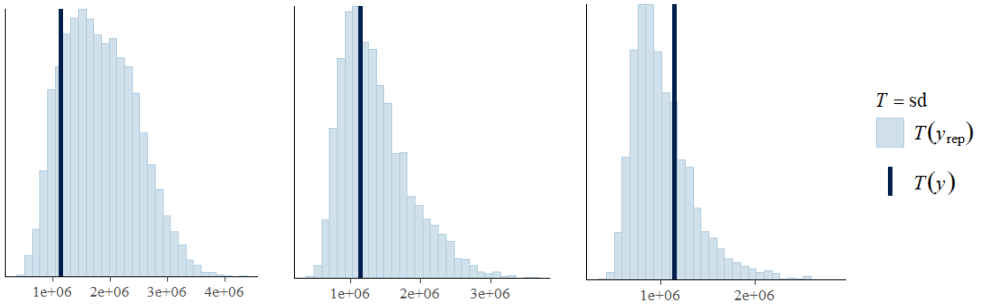

Figure 14 – PPC for Severity standard deviation

The legend follows the same convention as in Figure 13 and the ordering of the model across panels is also maintained. In this case, the Pareto Type I model has a posterior predictive distribution of the standard deviation significantly shifted to the right when compared to the observed sample, indicating an overestimation of dispersion. In the other two models, this behaviour is not as pronounced, suggesting a more adequate reproduction of the variability present in the data.

4.3. Comparison

The final step is to evaluate the predictive performance in an out of sample setting. LOO-CV will be applied to both Bayesian and frequentist formulations. Although the Bayesian LOO-CV uses the posterior predictive distribution and the frequentist LOO-CV uses predictions based on MLE, both approaches provide a common basis for comparing models in terms of expected predictive accuracy.

The model comparison, for the Bayesian models, is carried out using the log pointwise predictive density estimated via PSIS, $\widehat{elpd}_{psis-loo}$. The goal is to get the highest value of $\widehat{elpd}_{psis-loo}$, indicating a better predictive performance.

In the frequentist framework, the result measure corresponds to a frequentist Leave-One-Out predictive score, which is analogous to the Bayesian $\widehat{elpd}_{psis-loo}$. In this case, it is performed by removing each observation y_i from the dataset and re-estimating the model parameters using maximum likelihood based on the remaining data y_{-i}. The predictive performance is then evaluated by computing the log predictive density of each omitted observation using the re-estimated parameters. Then we sum these values to obtain the overall LOO-CV score. However, in this case, the model must be estimated *n* times, each time leaving out one observation from the sample, which can be more demanding for complex likelihoods. In Table V, the corresponding results are displayed for each model:

TABLE V

LOO-CV FREQUENCY RESULTS

Frequentist		Bayesian	
Poisson	Negative Binomial	Poisson	Negative Binomial
-40.9985	-41.83765	-40.8	-41.4

In the frequentist framework, the Poisson model obtains a slightly higher log predictive score than the Negative Binomial model. The same happens in the Bayesian framework, where the Poisson model attains a slightly higher value of $\widehat{elpd}_{psis-loo}$. So, the differences in the predictive performance scores are very small, indicating that both models perform similarly, with a slight advantage for the Poisson model of the Bayesian framework.

Consistent with the approach used in the frequency analysis, the results of the severity analysis are presented in Table VI.

TABLE VI

LOO-CV SEVERITY RESULTS

Frequentist			Bayesian		
Pareto Type I	Pareto Type II	Lognormal	Pareto Type I	Pareto Type II	Lognormal
-1148.286	-1146.588	-1146.369	-1148.5	-1145.6	-1145.1

Among the frequentist models the Lognormal distribution attains the highest level of the predictive criteria, despite being approximately identical to the one obtained in the Pareto Type II distribution. Additionally, in the Bayesian models, a similar pattern is observed where both the Pareto Type II and Lognormal distributions have comparable predictive performance. By comparing these two distributions across both frequentist and Bayesian approaches, the Bayesian models achieve higher LOO-CV values overall. This does not imply the Bayesian formulations are inherently the "best models", rather indicates they have superior predictive capacity with respect to large losses.

4.4. Conclusions and future work

Starting with the frequency modelling within the Bayesian framework, it was seen that the posterior distribution of the overdispersion parameter of the Negative Binomial was concentrated near zero. This result suggests limited evidence of overdispersion in the data. Additionally, the Poisson model presented a slightly higher LOO-CV than the Negative Binomial model. Taken together, these results indicate that the Bayesian Poisson model has a marginal advantage over the Bayesian Negative Binomial. However, given the small differences in predictive performance, one should take this conclusion with care. Comparing with the frequentist framework, the LOO-CV results were systematically lower than the Bayesian models, which suggests the Bayesian framework has a superior predictive performance. A possible limitation in this analysis arises from the need to round the frequency data, since claim counts must be integer numbers and Stan does not allow non-integer observations in the Negative Binomial likelihood. This formulation may have influenced inferential results and should be acknowledged when interpreting the findings.

Regarding the severity modelling, we observed that all distributions had a predictive performance like the frequentist framework, and all displayed an identical shape. In the analysis of the PPC, the Pareto Type I model showed some shortcomings in reproducing

the observed mean and variance, as indicated by the posterior predictive diagnostics. The LOO-CV results also confirmed that the Pareto Type I presented the lowest LOO-CV score and both Pareto Type II and Lognormal distributions achieved higher (and nearly identical) LOO-CV values. Like in the frequency framework, the distributions of the Bayesian models displayed higher predictive performance. However, there are some limitations in this inference, regarding the upper truncation in the severity data. As previously discussed, this action may exclude relevant information from the tail of distribution, leading to possible differences between models.

For future research, in the frequency analysis, we could explore alternative count-data formulations to avoid the need to round observations, to reduce potential distortions in the Bayesian framework. Regarding the severity analysis, removing or extending significantly the upper truncation limit, could improve the discrimination between severity models, by incorporating more information about the large losses. Also, we could explore alternative informative prior distributions to further improve the performance of the Bayesian model.

REFERENCES

Costa, L. (2017). *Rules Governing the Presentation of Written Work at ISEG*. Mimeo ISEG.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3), 515-533.

Hoff, P.D. (2009). *A First Course in Bayesian Statistical Methods*. Springer.

Klugman, S., Panjer, H. & Willmot, G. (2008). *Loss Models: From Data to Decisions, 3rd Ed.* John Wiley & Sons.

Kruschke, J.K. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS and Stan*, 2nd Ed. Academic Press.

Simpson, D., Rue, H., Riebler, A., Martins, T.G. & Sørbye, S.H. (2017). Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical Science*, 32(1), 1-28.

Stan Development Team. (2026). *Stan User's Guide (Version 2.38).* Stan Development Team.