University of Milan

Faculty of Political, Economic and Social Sciences

A Bayesian Approach to Aggregate Insurance Claim Modeling

Final Project in the Subject Bayesian Analysis

Julia Maria Wdowinska (43288A)
Edoardo Zanone (33927A)

Data Science for Economics
II Year
Master's Degree

We declare that this material, which we now submit for assessment, is entirely our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of our work. We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by us or any other person for assessment on this or any other course of study.

April 23, 2025

Contents

1 Introduction

2

The first objective of this project was to replicate the analysis conducted by Dudley. The dataset used comprises insurance claim amounts exceeding 1.5 million over a period of five years from an automobile insurance portfolio. The data, originally presented in Rytgaard (1990), is shown in Table 1.

Table 1: Insurance Claim Amounts Exceeding 1.5 Million (Data from Rytgaard, 1990)

Year	**Claim Amounts (in millions**				
1	2.495	2.120	2.095	1.700	1.650
2	1.985	1.810	1.625	–	–
3	3.215	2.105	1.765	1.715	–
4	–	–	–	–	–
5	19.180	1.915	1.790	1.755	–

Note: The threshold of 1.5 million corresponds to the retention level of an excess-of-loss insurance policy[1].

To model this data, the following mathematical notation was introduced. Let N_t denote the number of claims in year t, and let $Y_{i,t}$ represent the amount of the i-th claim in year t, where $i = 1, 2, \ldots, N_t$. The aggregate claim amount in year t is then given by

$$S_t = Y_{1,t} + Y_{2,t} + \cdots + Y_{N_t,t}.$$

To proceed, certain assumptions about the distributions of N_t and $Y_{i,t}$ were necessary. It was assumed that claims occur at a constant rate over time at random. Therefore, the number of claims N_t was modeled using a Poisson distribution. Since the individual claim amounts $Y_{i,t}$ are positive and may include large outliers, a loss distribution with a heavy tail was needed. Consequently, the Pareto distribution was selected to model $Y_{i,t}$. Hence,

$$N_t \sim \text{Poisson}(\theta), \quad 0 < \theta < \infty,$$

$$Y_{i,t} \sim \text{Pareto}(\alpha, \beta), \quad \alpha > 0, \quad 0 < \beta < y.$$

In the Pareto distribution, the parameter β determines the minimum possible value of y, meaning the support of the distribution is $[\beta, \infty)$. This is particularly appropriate in our context, as we are modeling insurance claims that exceed a specific threshold.

Additional assumptions were made: N_t was assumed to be independently and identically distributed (i.i.d.) across t; $Y_{i,t}$ were i.i.d. across both indices i and t; and N_t and $Y_{i,t}$ were assumed to be independent for all i and t. Under these assumptions, S_t follows a compound Poisson distribution, since it represents the sum of independent Pareto-distributed random variables.