An Extreme Value Approach to Modelling Construction Defect Insurance Claims

Matilda Ekermann, Ida Swartling

Bachelor's thesis
2022:K2

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Centrum Scientiarum Mathematicarum

Contents

Abstract

Predicting future large claims, as well as the total cost, of a specific insurance is essential for insurance companies, for example when setting premium levels or purchasing reinsurance coverage. The purpose of this thesis is to investigate if extreme value theory can be applied to construction defect insurance claims.

Data is provided by an insurance company offering construction insurance and two approaches are tested; the block maxima method using the generalized extreme value distribution and the peaks over threshold method using the generalized Pareto distribution. For both approaches, estimates for 10 and 50 year return levels, as well as 95% confidence intervals for the estimates, are calculated. Due to large variances for long periods of predictions, the confidence intervals are rather wide for both methods and hence the estimates need to be updated when more data become available in the future.

Additionally, a model to estimate the expected total annual payout for the construction defect insurance of this specific insurance company is proposed. The estimated total annual payout should also be used as an indication of how large buffers the insurance companies need to build up in order to have enough coverage for possible large payouts in the future.

Acknowledgement

We would express our gratitude towards our supervisor Nader Tajvidi at the Center of Mathematical Sciences at Lund University. Thank you for challenging us and for being so encouraging and supportive throughout the entire process.

We also want to express our gratitude towards Erik Landén and Axel Weckström at Gar-Bo Försäkring AB, without you this thesis would not have been possible to write. A special thank you to Axel, for your interest in the project and for providing us with the necessary data.

1 Introduction

A historical background of the theory used in this thesis as well as general information about insurance can be found in this section. More specific information regarding the insurance considered in this thesis, as well as the most common types of damage, is also described.

1.1 Historical Background Extreme Value Theory

The subject of extreme value analysis was first developed in the early 20th century in the form of an abstract study [1]. One of the key results, developed in 1928 by R. Fisher and L. Tippett, was the theory about the possible limit laws of the sample maximum. Other important contributions were made by M. Fréchet (1927), R. von Mises (1936) and B. Gnedenko (1942) [2]. The methodology was later, starting in the 1950's, applied to physical phenomena, initially mainly in the field of civil engineering [1]. The first book dedicated exclusively to extreme value theory, also considered to be the main referential work for applications within the field of engineering [3], was written by E.J. Gumbel in 1958. The content of the book is presented at an elementary level, with the purpose of making the results more accessible [4].

1.2 Insurance

Insurance is a contract in which a risk is transferred from the insured to an insurer. The agreement that is made between the insured and the insurance company is documented in an insurance letter as well as in the terms and conditions. These documents include details regarding the insurance coverage, the premium and other relevant details. An example of such a risk is the possible financial loss following construction defects on new production of single-family houses.

1.2.1 Juridical History on Construction Defect Insurance

In 1993 the Swedish parliament established a new law, the building defect insurance act, to resolve problems with moisture damage, problems that are usually caused during the construction period and discovered years later. When detected, the damage is generally extensive and is thus also expensive to remedy. If not repaired, the constructional damage might have negative

effects on the health of the residents. According to the building defect insurance act, a construction defect insurance must be in place when a building intended, partly or fully, for permanent residence is constructed. Initially the acct applied only to apartment buildings, but was in 2005 changed to include also detached and semi-detached houses. In addition to providing a financial protection for the residents, the aim of the act was to eliminate the uncertainties regarding which party is responsible to pay for the necessary repairs. Consequently the period between discovery and repair would be shortened, preserving the health of the residents [5].

The law was abolished in 2014 [6]. In the ministry memorandum investigating the consequences of an abolition it was for example argued that the building permit process would be more efficient without the act in place. It was also argued that the insurance was offered by too few insurance companies, causing an ineffective competitive situation unfavorable for the customers. At that time the two companies providing the insurance for single-family houses were Försäkrings AB Bostadsgaranti and Gar-Bo Försäkring AB, for short Gar-Bo [5]. Although the law was abolished, similar insurances are still offered, today only to private individuals by Gar-Bo [6].

1.2.2 Gar-Bo's Construction Defect Insurance for Newly Built Houses

Gar-Bo, founded in 1989, is a Swedish insurance company that offers several types of insurance related to construction of houses. One of the insurance types offered is the construction defect insurance (Nybyggnadsförsäkring in Swedish) which is the one that will be studied in this thesis [7].

The construction defect insurance covers damage that is caused from flaws in material and execution during the construction period, as well as the error that caused the damage. The insurance period is 10 years starting from the day that the final inspection is approved. The insurance applies to the building specified in the insurance letter, regardless of changes in ownership.

There are a few exceptions to the insurance and three of them are highlighted here. Firstly, the insurance does not cover any aesthetic damage. If the damage does not affect the functionality or utilization of the building, it will not be covered by the insurance. Secondly, the insurance does

not cover payments where other parties carry the responsibility for financial compensation. Thirdly, there is a Force Majeure clause meaning that the insurance does not cover damage caused by, for example, natural disasters, war or government action [8].

The premium is based on several aspects, including choice of contractor, building costs and the amount of risk involved in the construction. The deductible is set to be 0.5 times the price base amount (prisbasbelopp) – a fixed amount which reflects the price tendency in society and is decided by the government each year. There is also an option to pay a higher premium and get a deductible of 0.2 times the price base amount. The insured must pay one deductible per damage claim, regardless of how many claims are made during the 10 year period [9].

1.2.3 Common Flaws Causing Extensive Damage

A survey with 822 respondents from the construction industry showed that the cause of the three most common damages to houses are all related to water. They are, in order, water penetrating different types of roofs, water escaping pipes not including wetrooms and kitchens, and lastly, construction flaws in wetrooms. It is also the case that the three most costly damages are the same as the three most common ones. In addition to this, the survey shows that these types of damages are all largely discovered after the warranty period has expired, and are caused in the production state. That means that unless the owner of the house has insurance, they will have to pay the full repair cost themselves.

It was established that the type of damages that were most common a few decades ago are still the most common ones today. The frequency of the most common damages is also the same today as it was a few decades ago, but since there are more houses being build today the total amount of damages have increased [10].

2 Description of Data

The data that has been analyzed in this thesis is provided by Gar-Bo. The original data contains a set of claims made for the construction defect insurance that has led to, or will possibly lead to, a payout. It includes the date the claim was made, the start date of the insurances, the finalized payments, the estimated total payout, the contract sum for the construction and the amount of apartments in each building as well as the type of insurance; Nybyggnad or Nybyggnad Flerbostad. The amount of insurance policies written each year as well as the total contract sum for these constructions were also provided.

2.1 Cleaning of Data

The data provided by Gar-Bo contains more information than needed to conduct the different analyzes in this thesis. Therefore the data needs to be cleaned, and in this section it is described how this is done.

2.1.1 Type of Insurance

In the data, there are two different types of insurances for newly built houses; Nybyggnad and Nybyggnad Flerbostad. The former is the insurance for single family houses and the latter for houses that contains two or more apartments. The analysis in this thesis is, for two different reasons, restricted to the insurance for single family houses. The first reason being that the data for the amount of apartments in each construction is not complete, and hence the division of apartment buildings into single family households cannot be made. Secondly, the apartments in the same building cannot be argued to be independent from one another. If one of the apartments was built incorrectly chances are that several of them were. In the theory used in this thesis independence of the variables is preferable and thus the theory would be more difficult to apply to the claims made on houses with more than one apartment.

2.1.2 Years

The original data set includes data on insurance policies written between 1999 and 2021. In the analysis of the data, insurances starting before 2002

are removed. In the years up until and including 2001 not as many insurance policies were written as in the following years and thus it can not be assumed that the probability of an extreme event happening will be the same for insurances starting in these years. Since a stable amount of insurances over time is preferred, these claims are disregarded. In parts of the analysis only claims on insurances starting before and including 2011 are considered. Insurances starting after 2011 might still receive claims since claims can be made up until ten years after the final inspection. Parts of the analysis requires the final amount of the claims to be known for each year, and thus the insurance policies written after 2011 are in these cases disregarded.

2.1.3 Investigation costs

Sometimes when the insurance company investigates a claim there are investigation costs, most commonly for inspection of the house. An assumption is made that posts in the data corresponding to a payout lower than the deductible are investigation costs and are thus disregarded in the analysis of this thesis. The deductible is assumed to be half of the price base amount of the year the claim is made. There is an option to pay a higher premium and get a lower deductible, but this is very rare.

2.2 Sorting of Data

When the data was analyzed, the payout was connected to the underwriting year, the year when the insurance policy was written, and not to the year when the claim was registered. Different amounts of insurance policies are written each year and there might be minor changes in the terms and conditions. When connecting the payout to the underwriting year, the analysis takes into consideration both changes in volume and in risk between different years.

2.3 Inflation

Since the payments have been made during various years, the amounts are converted to the current monetary value using statistics from SCB [11]. Due to investigation and remedy of the damage, the assumption that the payments were finalized one year after Gar-Bo received the claim is made. The contract sum for the construction is also used in some cases, and hence this

too is converted to the current monetary value. The cleaned and sorted data where the payouts have been converted to the current monetary value can be seen in Figure 1.

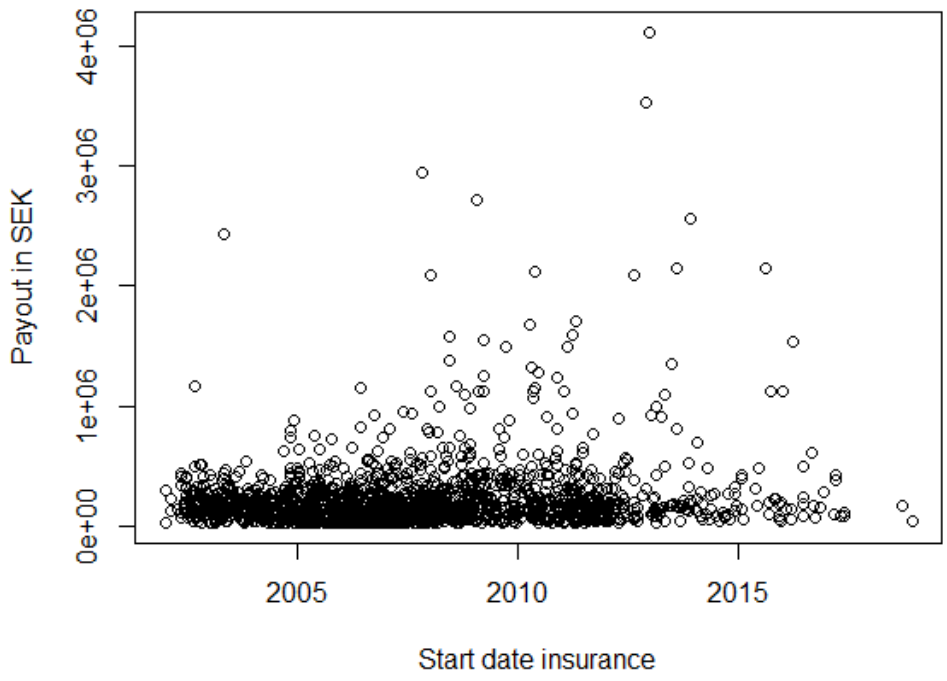

Figure 1: Scatter plot of sorted and cleaned data converted to the current monetary value

2.4 Assumption of Equal Distribution

For the insurance for newly built houses, the maximum insurance coverage is the same as the contract sum of the construction of the house. In other words it varies with each insurance. To be able to affirm that each of the claims could assume to be equally distributed it is important that the size of the payout is not correlated with the cost of building the house. In Figure 2 below the contract sum of the construction is plotted against the payout

made. A Pearson correlation test is also made. The test gives a correlation coefficient of $6.62 \cdot 10^{-2}$. We therefore assume that there is no correlation between the construction cost and the payout. Hence the claims are assumed to be equally distributed.

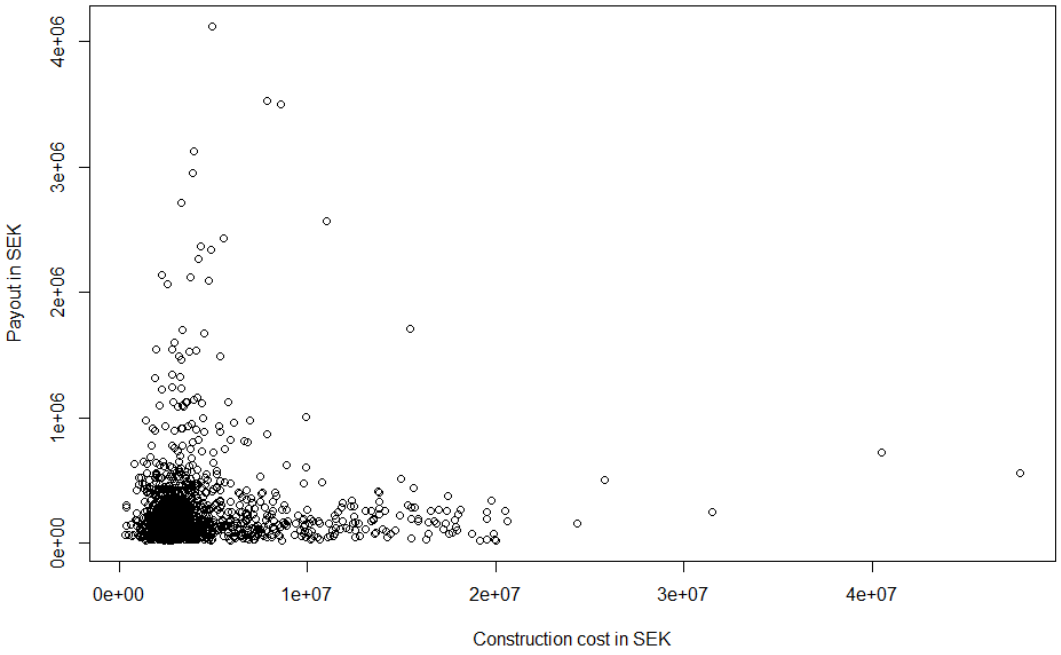

Figure 2: Plot of construction cost against payout.

3 Theory

When an insurance company is interested in modelling and predicting large financial losses due to, for instance, construction errors when a house was built, it is important to look at those events that caused the great pay-outs. The events that cause an insurance company to pay significantly more money than usual, the extreme events, are of interest to predict since they can possibly cause bankruptcy if the probability of these extreme events happening is not taken into account in setting insurance premiums. The question of what an extreme event is in a mathematical sense is discussed and answered in this section.

3.1 Generalized Extreme Value Distribution (GEV)

Suppose that $X_1, ..., X_n$ is a sequence of independent random variables with common distribution function F. Let

$$M_n = \max\{X_1, ..., X_n\}$$

denote the maximum over the n variables. For all values of n, the exact distribution of M_n is

$$\begin{aligned} \mathrm{P}(M_n \leq z) &= \mathrm{P}(X_1 \leq z, \ldots, X_n \leq z) = \\ &= \mathrm{P}(X_1 \leq z) \times \cdots \times \mathrm{P}(X_n \leq z) = \\ &= F^n(z). \end{aligned}$$

In practice, the distribution function F is often unknown and hence the theory cannot be directly applied. Therefore, the behaviour of F^n as $n \to \infty$ is investigated. Note that $F^n(z) \to 0$ as $n \to \infty$ for any $z < z_+$, with $z_+ = \inf\{z : F(z) = 1\}$. Thus the distribution of M_n becomes concentrated at z_+. To avoid obtaining a degenerate distribution let

$$M_n^* = \frac{M_n - b_n}{a_n},$$

where $\{a_n > 0\}$ and $\{b_n\}$ are sequences of constants. If $\{a_n\}$ and $\{b_n\}$ are chosen appropriately, the distribution of M_n^* will be stabilized. The possible limiting distributions of M_n^* are presented in the theorem below.

Theorem 1 [1] If there exists sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\mathrm{P}((M_n - b_n)/a_n \leq z) \to G(z) \qquad \text{as } n \to \infty,$$

where G is a non-degenerate distribution function, then G belongs to one of the following families:

$$\mathrm{I} : G(z) = \exp\Big\{ -\exp\Big[-\Big(\frac{z-b}{a}\Big)\Big]\Big\}, \qquad -\infty < z < \infty;$$

$$\mathrm{II} : G(z) = \begin{cases} 0, & z \leq b, \\ \exp\big\{ -(\frac{z-b}{a})^{-\alpha}\big\}, & z > b; \end{cases}$$

$$\mathrm{III} : G(z) = \begin{cases} \exp\big\{ -\big[-(\frac{z-b}{a})^{-\alpha}\big]\big\}, & z < b; \\ 1, & z \geq b, \end{cases}$$

for parameters $a > 0$, b and, in the case of families II and III, $\alpha > 0$.

In all of these so called extreme value distributions, a is the location parameter and b the scale parameter. In family I and II there is a shape parameter α as well. These three families, to which M_n^* converges in distribution, are called the Gumbel (I), Fréchet (II) and Weibull (III) families.

To avoid the problem of having to choose the model that best suits the data, the generalized extreme value (GEV) family of distributions is generally used in statistical analysis. The GEV distribution combines the three families of extreme value distributions into a single family of distributions defined in the following theorem.

Theorem 2 [1] If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\mathrm{P}((M_n - b_n)/a_n \leq z) \to G(z) \qquad \text{as } n \to \infty$$

for a non-degenerate distribution function G, then G is a member of the GEV family

$$G(z) = \exp\Big\{ -\Big[1 + \xi\Big(\frac{z-\mu}{\sigma}\Big)\Big]^{-1/\xi}\Big\}, \tag{1}$$

defined on $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$, where, $-\infty < \mu < \infty, \sigma > 0$ and $-\infty < \xi < \infty$.

In this model, μ is the location parameter, σ is the scale parameter and ξ is the shape parameter. The Gumbel distribution is attained by letting $\xi \to 0$,

while the Fréchet distribution corresponds to the case when $\xi > 0$ and the Weibull distribution to the case when $\xi < 0$.

3.1.1 Block Maxima

Fitting a GEV distribution to a sequence of independent observations $X_1, X_2, ...$ can be done by using a block maxima approach. The data is divided into a number of blocks. Typically one block corresponds to the number of observations, n, in a year. In each of the blocks there is a maxima, giving rise to a sequence of block maxima, $M_{n,1}, M_{n,2}, ..., M_{n,m}$, where each of the m maxima corresponds to the maximum observation in that particular block. A GEV distribution can be fitted to these maxima. There is however a trade-off when choosing block size. Too many observations in a block means that there are fewer block maxima which can lead to large variance. Too few observations in a block can lead to large bias since the limiting distribution might not be a good fit if the maxima is not taken over sufficiently many observations.

3.1.2 Max-Stability

One property of the GEV distribution is the max-stability property. It means that taking the maxima over already existing maxima with a GEV distribution will result in a new set of maxima, also GEV distributed.

Definition 1 A distribution G is said to be max-stable if, for every $n = 2, 3, ...$, there are constants $\alpha_n > 0$ and β_n such that $G^n(\alpha_n z + \beta_n) = G(z)$, $\forall z \in \mathbb{R}$.

Theorem 3 [1] A distribution is max-stable if, and only if, it is a generalized extreme value distribution.

Let M_{nk} be the maximum of k maxima and let M_n be such that $(M_n - b_n)/a_n$ has the limit distribution G. Hence, if nk is large enough,

$$\mathrm{P}(M_{nk} - b_{nk})/a_{nk} \leq z \approx G(z).$$

Since M_{nk} and M_n have the same distribution

$$\mathrm{P}(M_{nk} \leq z) \approx G\Big(\frac{z - b_{nk}}{a_{nk}}\Big)$$

and

$$\mathrm{P}(M_{nk} \leq z) \approx G^k\Big(\frac{z - b_n}{a_n}\Big),$$

where the distribution functions G and G^k are equal except for the location and scale parameters. It follows that

$$G^k = \left(\exp\left\{ -\left[1+\xi\Big(\frac{z-\mu}{\sigma}\Big)\right]^{-1/\xi}\right\}\right)^k = \\ = \exp\left\{ -\left[1+\xi^*\left(\frac{z-\mu^*}{\sigma^*}\right)\right]^{-1/\xi^*}\right\}$$

where $\mu^* = \mu + \frac{\sigma}{\xi}(k^\xi - 1)$, $\sigma^* = \sigma k^\xi$ and $\xi^* = \xi$.

Confidence intervals for the theoretical estimates, with a significance level of α, are calculated using the delta method as

$$I_{\mu^*} = \mu^* \pm \lambda_{\alpha/2}\sqrt{\mathrm{Var}(\mu^*)}, \qquad I_{\sigma^*} = \sigma^* \pm \lambda_{\alpha/2}\sqrt{\mathrm{Var}(\sigma^*)}$$

and

$$I_{\xi^*} = \xi^* \pm \lambda_{\alpha/2}\sqrt{\mathrm{Var}(\xi^*)}.$$

The variance for μ^* is calculated using the gradient as $\mathrm{Var}(\mu^*) = \nabla{\mu^*}^T V \nabla\mu^*$ with V as the covariance matrix for $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ and

$$\nabla{\mu^*}^T = \left(\frac{\partial\mu^*}{\partial\mu}, \frac{\partial\mu^*}{\partial\sigma}, \frac{\partial\mu^*}{\partial\xi}\right) = \left(1, \frac{1}{\xi}(k^\xi - 1), -\frac{\sigma}{\xi^2}(k^\xi - 1) + \frac{\sigma}{\xi}\log(k)k^\xi\right).$$

Lastly, the variance for σ^* is given by $\mathrm{Var}(\sigma^*) = \nabla{\sigma^*}^T V \nabla\sigma^*$ with

$$\nabla{\sigma^*}^T = \left(\frac{\partial\sigma^*}{\partial\mu}, \frac{\partial\sigma^*}{\partial\sigma}, \frac{\partial\sigma^*}{\partial\xi}\right) = (0, k^\xi, \sigma k^\xi \log(k)).$$

The variance for $\xi^* = \xi$ is calculated as $\mathrm{Var}(\xi^*) = \mathrm{Var}(\xi)$.

3.1.3 Return Level

Assuming blocks correspond to observations in a year, the return level is the level which is expected to be exceeded once every $\frac{1}{p}$ years, where $\frac{1}{p}$ is the return period corresponding to the return level z_p. The return can be calculated by inverting Eq.(1) as follows:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}[1 - \{\log(1-p)\}^{-\xi}], & \xi \neq 0 \\ \mu - \sigma\log\{-\log(1-p)\}, & \xi = 0, \end{cases}$$

where $G(z_p) = 1 - p$. By defining $y_p = -\log(1-p)$, a simplified expression for the return level is obtained:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}[1 - y_p^{-\xi}], & \xi \neq 0 \\ \mu - \sigma\log(y_p), & \xi = 0. \end{cases}$$

The return level z_p can also be expressed as the level which is exceeded by the yearly maximum in any given year with probability p.

3.2 Generalized Pareto Distribution (GPD)

An event is regarded as extreme if it exceeds a certain threshold u. Denote the points that exceed u by $\{x_i : x_i > u\}$ and label them as $x_{(1)}, x_{(2)}, ..., x_{(k)}$. Define the exceedances above u as $y_j = x_{(j)} - u$ for $j = 1, ..., k$. Each y_j can be considered an independent realization of a random variable that can be approximated by a member of the generalized Pareto family.

Theorem 4 [1] Let $X_1, X_2, ...$ be a sequence of independent random variables with common distribution function F and let

$$M_n = \max\{X_1, ..., X_n\}.$$

Denote an arbitrary term in the X_i sequence by X, and suppose that F satisfies Theorem 2, so that for large n,

$$\mathrm{P}(M_n \leq z) \approx G(z)$$

where

$$G(z) = \exp\Big\{ -\Big[1 - \xi\Big(\frac{z-\mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}} \Big\}$$

for some $\mu, \sigma > 0$ and ξ. Then, for large enough u, the distribution function of $(X - u)$, conditional on $X > u$ is approximately

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}} \tag{2}$$

defined on $\{y : y > 0 \text{ and } (1 + \xi y/\tilde{\sigma}) > 0\}$, where

$$\tilde{\sigma} = \sigma + \xi(u - \mu).$$

The variables which are members of the family of distributions described by Eq.(2), are members of the generalized Pareto family.

The distribution limit of the exceedances above the threshold depends greatly on the value of the shape parameter ξ.

If $\xi < 0$: the upper bound of the distribution of the exceedances is $u - \frac{\tilde{\sigma}}{\xi}$.

If $\xi > 0$: the distribution of exceedances has no upper limit.

If $\xi = 0$: the distribution is unbounded and by taking the limit of Eq.(2) as $\xi \to 0$ the obtained distribution is exponential with parameter $\frac{1}{\tilde{\sigma}}$,

$$\lim_{\xi \to 0} H(y) = \lim_{\xi \to 0} 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}} = 1 - e^{-\frac{y}{\tilde{\sigma}}}, \qquad y > 0.$$

If a sequence of independent and identically distributed random variables is denoted by $X_1, X_2, ...$, and an arbitrary term in the sequence X_i is denoted X, a conditional probability describing the stochastic behaviour of extreme events is as follows

$$\mathrm{P}(X > u + y | X > u) = \frac{1 - F(u + y)}{1 - F(u)}, \qquad y > 0.$$

How a threshold is chosen is discussed below.

3.2.1 Threshold Selection

When choosing a threshold u there is a trade-off between bias and variance. By choosing a threshold that is too low, there will be more exceedances which may lead to large bias. Since the GPD is an approximate distribution for

observations over a large enough u, choosing a threshold that is too low can cause the GPD to be a bad fit for the data. Choosing too high of a threshold will result in few exceedances and there is a risk for high variance. The standard practice is to choose a threshold that is as low as possible, as long as the limit model provides a good approximation. There are two methods for selecting an appropriate threshold, one is to study the mean residual life plot, a graphical method. The other is to fit models across a range of different thresholds, a model based method.

An important property of the GPD is that it is stable under the change of threshold. A lemma and proof is provided below.

Lemma 1 Suppose $(X - u_0|X > u_0) \sim \text{GPD}(x : \sigma, \xi)$. Then, given $u > u_0$, $(X - u|X > u) \sim \text{GPD}(x : \sigma_u, \xi_u)$ with $\sigma_u = \sigma + \xi(u - u_0)$ and $\xi_u = \xi$.

Proof. Case 1: $\xi \neq 0$

$$\text{P}(X - u > +x|X > u) =$$

$$= \text{P}(X - u_0 > x + u - u_0|X - u_0 > u - u_0) =$$

$$= \frac{\left(1 + \xi \cdot \frac{x+u-u_0}{\sigma}\right)_+^{-\frac{1}{\xi}}}{\left(1 + \xi \cdot \frac{u-u_0}{\sigma}\right)_+^{-\frac{1}{\xi}}} = \left(1 + \xi \cdot \frac{x}{\sigma + \xi(u - u_0)}\right)_+^{-\frac{1}{\xi}}.$$

Case 2: $\xi = 0$

$$\text{P}(X - u > x|X > u) =$$

$$= \text{P}(X - u_0 > x + u - u_0|X - u_0 > u - u_0) =$$

$$= \frac{e^{-\frac{x+u}{\sigma}}}{e^{-\frac{u}{\sigma}}} = e^{-\frac{x}{\sigma}}.$$

□

The distribution obtained from Case 2 is in accordance to the lack of memory property of the exponential distribution.

Threshold Selection - a Graphical Method

First, note that if $Y \sim \text{GPD}(y : \sigma, \xi)$, then

$$\text{E}(Y) = \frac{\sigma}{1-\xi}$$

if $\xi < 1$ and infinite if $\xi \geq 1$.

Let $X_1, X_2, ..., X_n$ be a sequence where the GPD is a valid model for the exceedances over the threshold u_0, with corresponding scale parameter σ_{u_0}. Denote an arbitrary term of that sequence X. This implies

$$\text{E}(X - u_0 | X > u_0) = \frac{\sigma_{u_0}}{1-\xi}, \qquad \xi < 1.$$

Since the GPD is stable under the change of threshold, it means that if the GPD model is valid for exceedances over u_0, it must also be valid for exceedances over $u > u_0$. The expectation of the exceedances over u, where σ_u corresponds to the threshold u, is given by:

$$\text{E}(X - u | X > u) = \frac{\sigma_u}{1-\xi} = \frac{\sigma_{u_0} + \xi u}{1-\xi} = \frac{\sigma_{u_0}}{1-\xi} + u \cdot \frac{\xi}{1-\xi}, \tag{3}$$

which is a linear function in u. An empirical estimate of the mean of the excesses over u, is given by the sample mean of the exceedances over u. From Eq.(3), if the value of the threshold $u > u_0$ increases, the expected value of the excesses change linearly in u. Thus the plot of the excesses should also be linear in u. Such a plot is called the mean residual life plot. The plot consists of the points positioned at

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) \right) : u < x_{\max} \right\},$$

where $x_{(1)}, ..., x_{(n_u)}$ are the n_u exceedances above u and $x_{\max}$ is the largest of the exceeding values.

Threshold Selection - a Model Based Method

The second method for estimation of threshold consists of checking for stability of the estimates σ and ξ for a range of different thresholds fitted to the GPD. The smallest threshold u for which the parameters are constant is

chosen. As previously stated in Lemma 1, if the GPD is a good model for exceedances over a threshold u_0, then it is also a good model for exceedances over a new threshold $u > u_0$ with $\xi_u = \xi$ and

$$\sigma_u = \sigma_{u_0} + \xi(u - u_0). \tag{4}$$

This implies that σ_u is a linear function in u unless $\xi = 0$. Instead, for $\xi = 0$ the scale parameter is defined as:

$$\sigma^* = \sigma_u - \xi u,$$

which, by Eq.(4), is constant with respect to u. Due to the stability of the GPD under the change of threshold, it can be stated that if exceedances above u_0 is GPD, then for exceedances above $u > u_0$ estimates of ξ and σ^* should be approximately constant, or at least stable.

3.2.2 Return Level

It is of interest to know what value x_m will be exceeded once in every m observations, x_m being the m-observation return level. To find the return level in the case where the GPD is a good model for exceedances over a threshold u, basic knowledge of conditional probability can be used.

Note that:

$$\mathrm{P}(X > x | X > u) = \frac{\mathrm{P}((X > x) \cap (X > u))}{\mathrm{P}(X > u)} = \frac{\mathrm{P}(X > x)}{\mathrm{P}(X > u)}.$$

It follows that:

$$\mathrm{P}(X > x | X > u) = \Big[1 + \xi\Big(\frac{x - u}{\sigma}\Big)\Big]^{-1/\xi} \iff \mathrm{P}(X > x) = \zeta_u \Big[1 + \xi\Big(\frac{x - u}{\sigma}\Big)\Big]^{-1/\xi},$$

where $\zeta_u = \mathrm{P}(X > u)$. The m-observation return level is given by solving the following equation for x_m:

$$\zeta_u \Big[1 + \xi\Big(\frac{x_m - u}{\sigma}\Big)\Big]^{-1/\xi} = \frac{1}{m}$$

Solving for x_m:

$$x_m = \begin{cases} u + \frac{\sigma}{\xi}[(m\zeta_u)^{\xi} - 1], & \xi \neq 0 \\ u + \sigma \log(m\zeta_u), & \xi = 0 \end{cases}$$

given that m is large enough to guarantee that $x_m > u$.

It might be interesting to know the expected level of exceedance on an annual scale instead, in other words the level that is expected to be exceeded once every N years. Suppose the number of observations each year is n_y, let z_N denote the N-year return level, with $m = N \cdot n_y$:

$$z_N = \begin{cases} u + \frac{\sigma}{\xi}[(Nn_y\zeta_u)^\xi - 1], & \xi \neq 0 \\ u + \sigma \log(Nn_y\zeta_u), & \xi = 0. \end{cases}$$

When estimating return levels, the parameters σ and ξ are replaced by their maximum likelihood estimates. The estimate of ζ_u is simply the proportion of observations exceeding the threshold u, given as follows:

$$\hat{\zeta}_u = \mathrm{P}(X > u) = \frac{k}{n},$$

where k is the number of observations exceeding u and n is the total number of observations. In this case, note that that $\hat{\zeta}_u$ is the maximum likelihood estimate of ζ_u since the number of exceedances above the threshold u follows the binomial distribution $\mathrm{Bin}(n, \zeta_u)$.

Confidence intervals for the x_m return level can be calculated using the delta method as

$$I_{\hat{x}_m} = \hat{x}_m \pm \lambda_{\alpha/2}\sqrt{\mathrm{Var}(\hat{x}_m)}.$$

Where $\mathrm{Var}(\hat{x}_m) \approx \nabla x_m^T V \nabla x_m$ with

$$\nabla x_m^T = \left(\frac{\partial x_m}{\partial \zeta}, \frac{\partial x_m}{\partial \sigma}, \frac{\partial x_m}{\partial \xi}\right) =$$

$$= \left(\sigma m^\xi \zeta_u^{\xi-1}, \xi^{-1}((m\zeta_u)^\xi - 1), -\sigma\xi^{-2}((m\zeta_u)^\xi - 1) + \sigma\xi^{-1}(m\zeta_u)^\xi \log(m\zeta_u)\right)$$

evaluated at $(\hat{\zeta}, \hat{\sigma}, \hat{\xi})$, and V as the covariance matrix for $(\hat{\zeta}_u, \hat{\sigma}, \hat{\xi})$ as

$$V = \begin{bmatrix} \mathrm{Var}(\hat{\zeta}_u) & 0 & 0 \\ 0 & \mathrm{Var}(\hat{\sigma}) & \mathrm{Cov}(\hat{\sigma}, \hat{\xi}) \\ 0 & \mathrm{Cov}(\hat{\xi}, \hat{\sigma}) & \mathrm{Var}(\hat{\xi}) \end{bmatrix}.$$

Note that since ζ_u is binomial distributed, $\mathrm{Var}(\hat{\zeta}_u) \approx \hat{\zeta}_u(1 - \hat{\zeta}_u)/n$.

3.3 Poisson Process

In a Poisson process, events occur randomly and independently of each other. Only the average rate of the occurrences is known, denoted by λ. Since the events occur randomly it might be the case that two events occur close in time and then there is a longer time interval until the next event occurs.

The number of events X that occur in a unit time interval, depending on the time t and with an average rate of $\mu = \lambda t$, has a Poisson distribution with parameter μ and the probability mass function

$$\mathrm{P}(X = x) = \frac{e^{-\mu}\mu^x}{x!}, \qquad x \in \Omega = \{0, 1, 2, ...\}.$$

The properties of the Poisson process makes it a good model for some events occurring in nature, such as the decay of radioactive particles, *or perhaps the occurances of constructions errors in newly built houses.*

Theorem 5 [12] Let $\{N_k(t), t \geq 0\}, k = 1, ..., n,$ be n independent inhomogeneous Poisson processes with intensity functions $\lambda_k(t), k = 1, ..., n$ and define the process $\{M(t), t \geq 0\}$ by $M(t) = N_1(t) + N_2(t) + ... + N_n(t)$. Then $\{M(t)\}$ is an inhomogeneous Poisson process with intensity function $\lambda(t) = \lambda_1(t) + ... + \lambda_n(t)$. Specifically, if $\{N_k(t)\}, k = 1, ..., n$ are Poisson processes with intensities $\lambda_1, ..., \lambda_n$ then $\{M(t)\}$ is a Poisson process with intensity $\lambda = \lambda_1 + ... + \lambda_n$.

3.4 Sums of a Random Number of Random Variables

The following theorem describes how the expectation and variance of a sum of a random number of random variables are calculated.

Theorem 6 [13] Let $X_1, X_2, ...$ be i.i.d nonnegative, integer-valued random variables, and let N be a nonnegative, integer-valued random variable, independent of $X_1, X_2, ...$ Set $S_0 = 0$ and $S_n = X_1 + X_2 + ... + X_n$, for $n \geq 1$. Then:

I) If $\mathrm{E}(N) < \infty$ and $\mathrm{E}(|X|) < \infty$, it holds that

$$\mathrm{E}(S_N) = \mathrm{E}(N) \cdot \mathrm{E}(X).$$

II) If also $\mathrm{Var}(N) < \infty$ and $\mathrm{Var}(X) < \infty$, then

$$\mathrm{Var}(S_N) = \mathrm{E}(N) \cdot \mathrm{Var}(X) + (\mathrm{E}(X))^2 \cdot \mathrm{Var}(N).$$

3.5 Probability and Quantile Plots

Definition 2 Given an ordered sample of independent observations

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$$

from a population with estimated distribution function $\hat{F}$, a probability plot consists of the points

$$\left\{ \left(\hat{F}(x_{(i)}), \frac{i}{n+1} \right) : i = 1, ..., n \right\}.$$

Definition 3 Given an ordered sample of independent observations

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$$

from a population with estimated distribution function $\hat{F}$, a quantile plot consists of the points

$$\left\{ \left(\hat{F}^{-1}\left(\frac{i}{n+1} \right), x_{(i)} \right) : i = 1, ..., n \right\}.$$

For the estimated distribution function $\hat{F}$ to be considered a reasonable model for the data, the points of the probability and quantile plots should lie close to the unit diagonal which represents the theoretical distribution function F. For probability and quantile plots corresponding to the GEV distribution and GPD specifically, see Section 8.1 and 8.2 in the Appendix.

4 Method

In this section it is described how the block size used in the GEV analysis is chosen, how the threshold for the GPD analysis is decided, as well as how the intensity measure of the payouts is calculated.

4.1 Generalized Extreme Value Distribution (GEV)

When fitting a GEV distribution to the data, only insurances which are no longer active are considered, in other words the ones with a start date between 2002-2011. The data from which the maxima are taken consists of 1889 insurances where a claim has lead to a payout. If several payouts were made for the same insurance, those payouts are summed and regarded as one data point.

4.1.1 Block Size Selection

When determining the block size for which the GEV model provides the best fit for the data, three different options are compared. The options are choosing maxima on an annual basis obtaining 10 maxima, on a six month basis obtaining 20 maxima, and on a quarterly basis obtaining 40 maxima, see Figure 3 below and Figure 14 and 15 in the Appendix. Comparing the probability and quantile plots in each of the three figures it can be concluded that the model with the best fit is the one in Figure 3 with a block size of six months. The figure shows that the pp- and qq-plot does not deviate much from the unit diagonal. The return level plot shows that the empirical return levels correspond well to the theoretical ones just as the empirical density correspond well to the theoretical one in the density plot. Figure 4 shows the scatter plot of the maxima for the six month block size selection

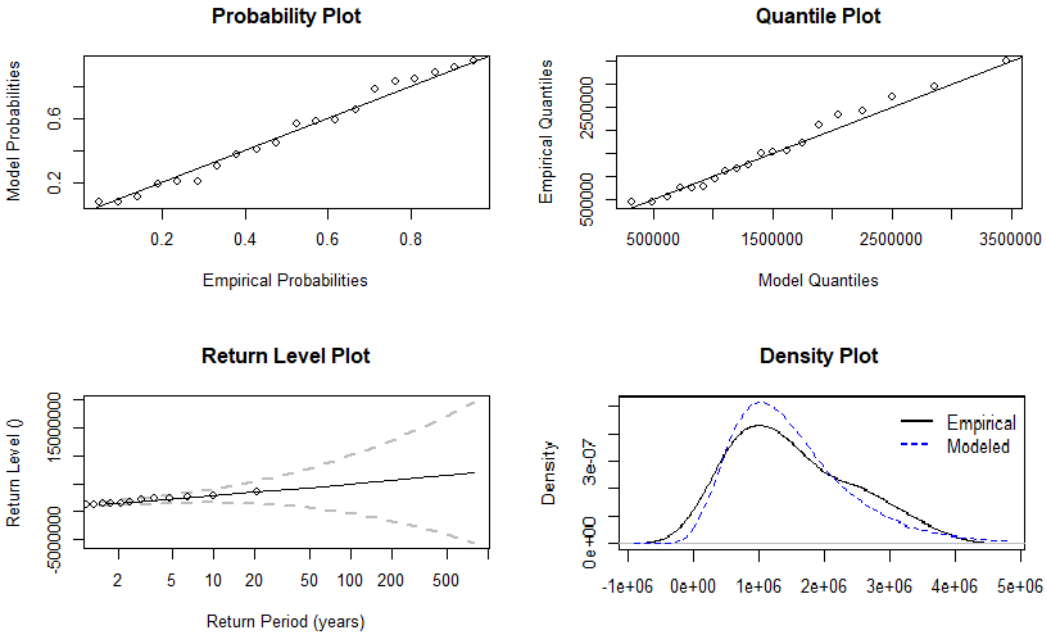

Figure 3: Diagnostic plots for six month block size.

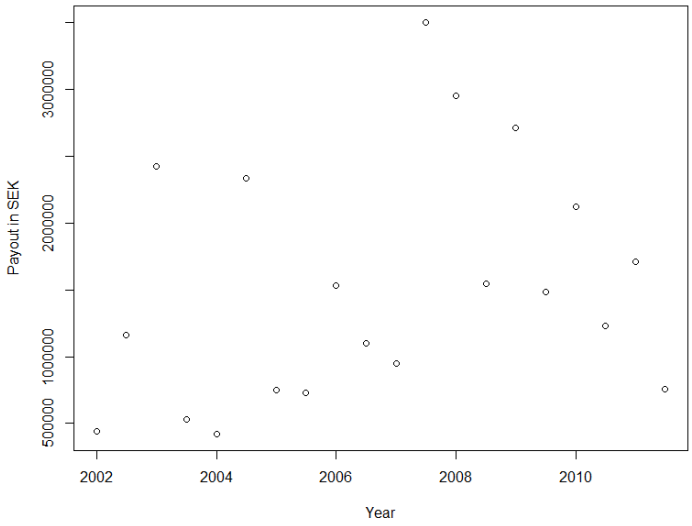

Figure 4: Scatter plot for six month block size.

A reason as to why the block sizes of three months does not give a very good fit can be that there are fewer observations in each block in this model compared to the other two, which can lead to large bias. This is discussed in Section 3.1.1. The model with a block size of twelve months does not give a bad fit, however the pp- and qq-plots are not as good as the model with six month block size. This might be because it only has 10 maxima.

4.1.2 Max Stability

Due to the max-stability property of the GEV distribution, see Section 3.1.2, the parameters of well fitted GEV models with different block sizes should be in relation to each other. The confidence intervals for the theoretical estimates of the different block size models can be calculated, using parameters from a model with a smaller block size. If the estimated parameters lie within that confidence interval, it strengthens the possibility that the GEV model is a good fit for the data.

It is concluded in the result section, see Tables 2 and 3, that the parameters of the GEV models agree well with the max-stability property for the most part. It is only the estimated shape parameter for the twelve month block size that does not lie within its designated confidence interval in the case when the confidence interval is created using the parameters estimated from the quarterly block maxima model. In all other cases the original estimates lie within the confidence intervals created using the estimates of one of the other two models.

4.1.3 Time Dependent Models

It is of interest to know whether the GEV model depends on time or not to obtain the best fit. To determine this a likelihood-ratio test is performed, see 8.4 in the Appendix, between the model where the GEV parameters does not depend on time and a range of models where the parameters depend more and less on time. If the likelihood-ratio test between two models gives a resulting p-value that is less than the chosen α-level, the conclusion is that the simpler model is the better one. The simplest model is the one where the parameters does not depend on time. Tests are performed for all three block sizes with $\alpha = 0.05$ to determine if the models depend on time, and also to see if there is a difference to whether the GEV parameters depend on time or not depending on the block size.

4.2 Generalized Pareto Distribution (GPD)

When fitting a generalized Pareto distribution to the exceedances above a threshold, data between the years 2002-2021 is used. This results in 1937 payouts, both from insurances that are no longer active and ones that still are.

4.2.1 Jitter

The probability that the payout for one insurance is the exact same as for another is in theory zero. The houses have different building costs, possibly different deductibles and most likely damages with different repair costs. Despite this, in the data received from Gar-Bo, some of the payout sums are the exact same. There might be plenty of reasons for this, but to avoid possible errors when performing the analysis a small jitter is added to the data. That is, different amounts which are small in comparison to the payout sums, are added or subtracted to every entry so that there are none that are identical.

4.2.2 Threshold Selection

Methods on how to choose a threshold is discussed in Section 3.2.1. It is not always clear what the threshold should be by only fitting models to different thresholds or by just looking at the mean residual life plot. Therefore, the two methods are combined to obtain a threshold appropriate to the data.

It is fair to assume that no future payout can be infinitely large. It is therefore sensible, according to what is discussed in Section 3.2, to choose a threshold such that the shape parameter ξ is negative. Using the model based method, after trying a range of different thresholds, one at u=600 000 with 86 exceedances is chosen. Fitting a GPD distribution to the 86 exceedances, the qq-plot and density plot looks good meaning that both the empirical quantiles and density correspond well to the theoretical ones, see diagnostic plots in Figure 5 below. Despite ξ not being negative, this threshold is chosen for three main reasons. The first one being that the threshold which gives a negative value for ξ is very high and has few exceedances. Possible consequences of this is discussed in Section 3.2.1. The second reason is that the diagnostic plots in Figure 5 show a very good fit, better than for the other thresholds that were tried. Thirdly, for thresholds greater than u=600 000

the parameters are relatively stable, indication that exceedances above this threshold are indeed GPD.

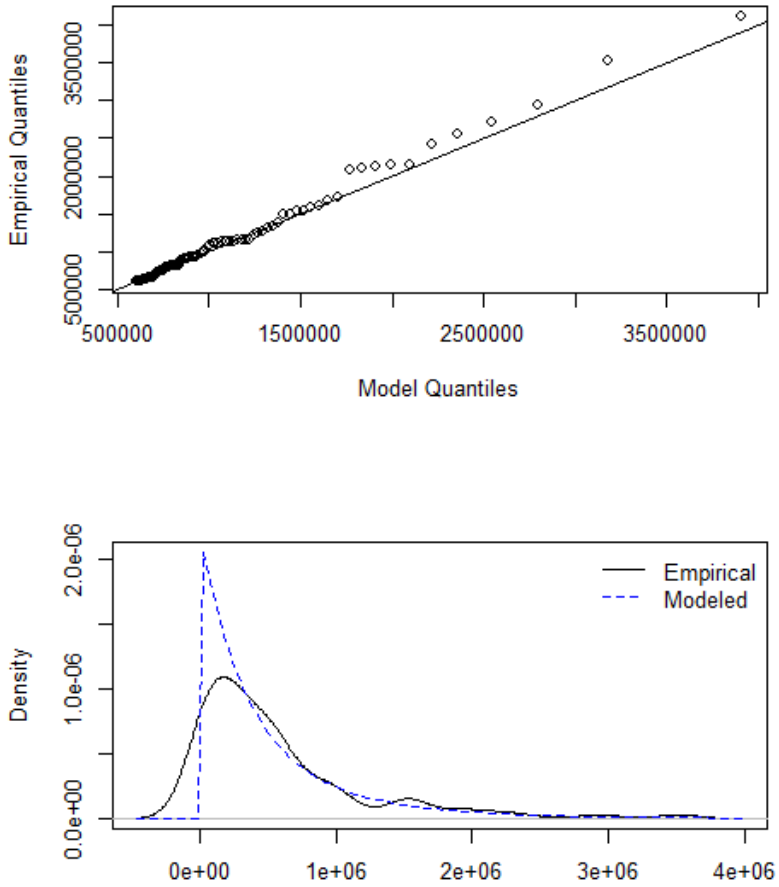

Figure 5: Diagnostic plots for threshold $u = 600\ 000$.

As for the graphical method of finding a threshold the mean residual life plot in Figure 6 below is analyzed. There is evidence for linearity above u=500 000, strengthening the choice of threshold at u=600 000 using the model based method.

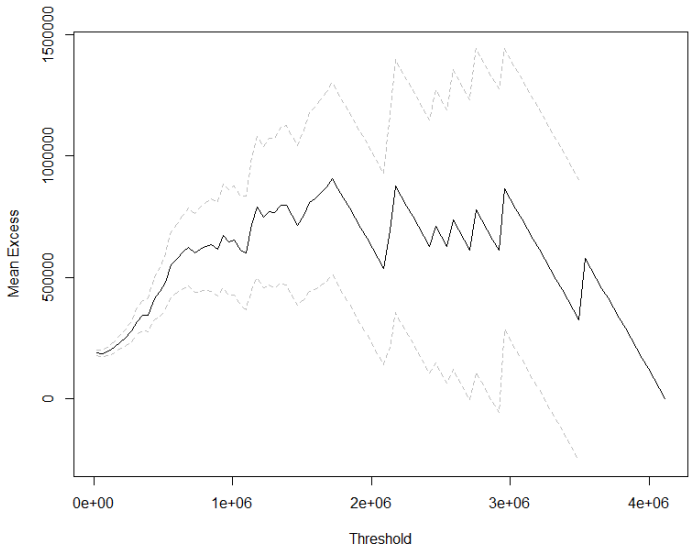

Figure 6: Mean residual life plot.

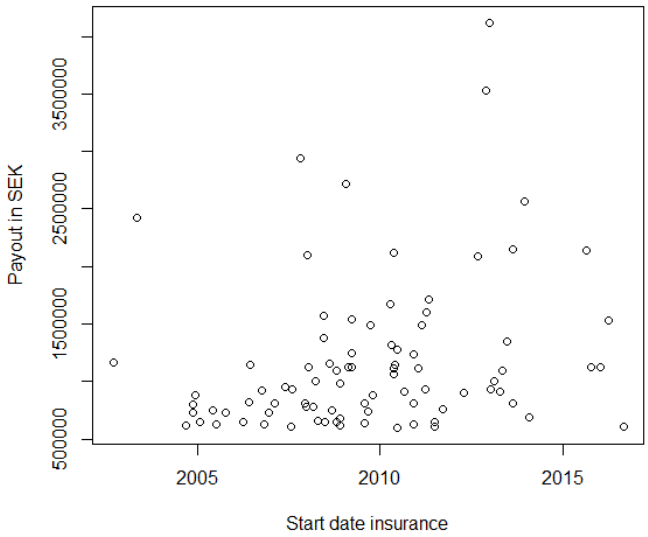

Figure 7: Scatter plot for threshold u=600 000.

4.3 Total Annual Payout

In this section, the method for estimating the total annual payout is discussed.

4.3.1 Gamma Distribution

To estimate the total annual payout a distribution needs to be fitted to the entire set of payouts. Since the exceedances have already been modeled, a distribution needs to be fitted only to the payouts below 600 000 SEK. To find a suitable distribution a histogram is plotted, see Figure 7 below.

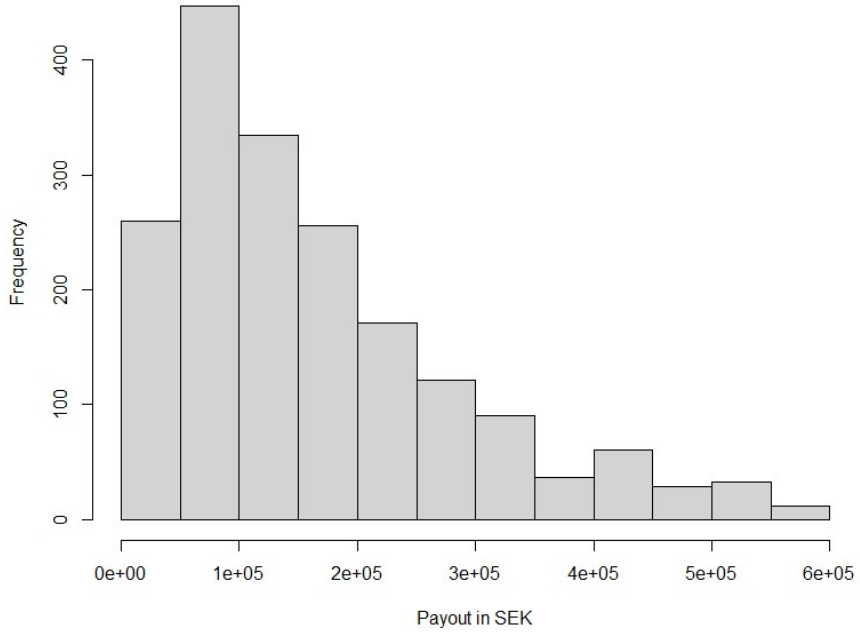

Figure 8: Histogram payouts under 600 000 SEK.

From this plot it is fair to assume that a gamma distribution is a good fit. Several other distributions are also fitted to the data, but in the end the best model is given by the gamma distribution. The diagnostic plots for the fitted gamma distribution are shown in Figure 8 below.

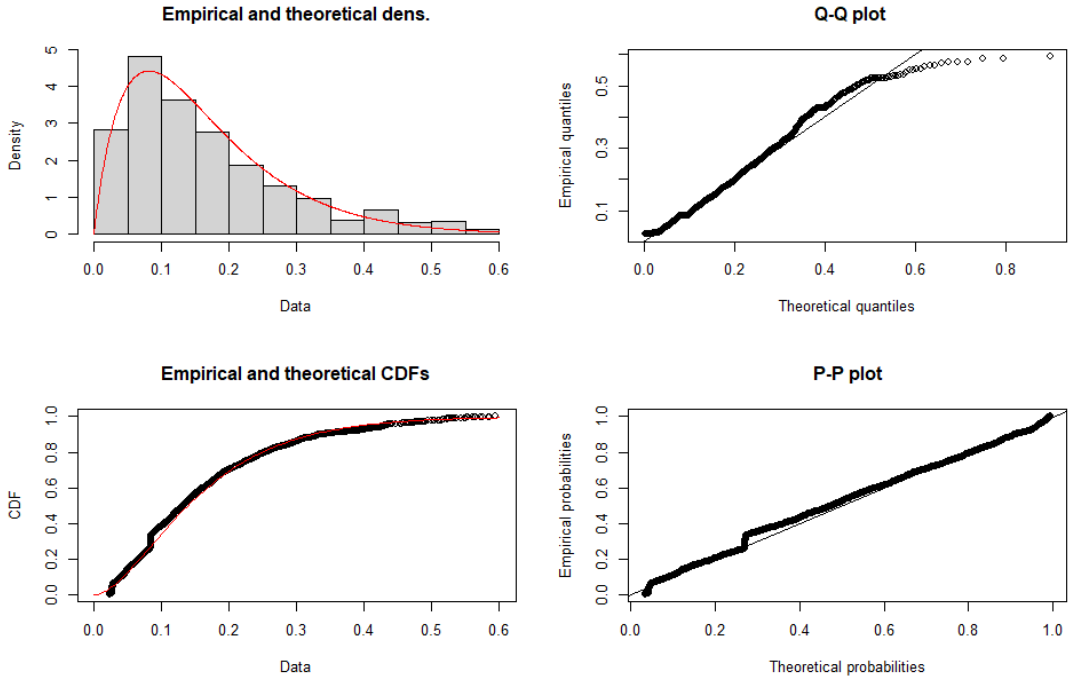

Figure 9: Diagnostic plots gamma distribution.

4.3.2 Estimation of Rate of Occurrence

To estimate the total payout per year, the amount of payouts made in a year needs to be estimated. The frequency of the payouts was modeled by a Poisson process with a rate of $\mu, 2\mu, \ldots, 10\mu$, where μ is the total rate of occurring payouts for insurances with a start date in 2021, 2μ the rate for insurance policies written in 2020 and so on. The rate 10μ corresponds to the insurances with a start date in 2002-2012, in other words the ones that are no longer active or that will no longer be active after 2022.

To obtain the estimate for μ, let N_k be the amount of payouts made for insurances that are k years old, $k = 1, ..., 10$. Assume that $N_k \sim \text{Poi}(\mu_k)$ where are $\mu_k = k\mu$. Each year N_k there are i_k payouts and hence μ can be estimated using MLE as follows:

$$\hat{\mu} = \underset{\mu}{\operatorname{argmax}} \prod_{k=1}^{10} \prod_{i=1}^{i_k} \frac{(\mu_k)^i}{i!} e^{-\mu_k}.$$

Two methods are tested to obtain a final estimate for the rate of payouts per year, μ_{total}. In the first method,

$$\mu_{total} = \mathrm{P}_1 \cdot \mu + \mathrm{P}_2 \cdot (2\mu) + \cdots + \mathrm{P}_{10} \cdot (10\mu)$$

where $\mathrm{P}_i, 1 \leq i \leq 10$, is the amount of payouts in the period i divided by the total amount of payouts for the whole time period. Here 2021 is referred to as period 1, 2020 as period 2 and so on, until period 10 which consists of the years 2002-2012.

In the second method, the average is taken over the estimates for the different time periods:

$$\mu_{total} = \frac{\mu + 2\mu + \cdots + 10\mu}{10} = 5.5\mu.$$

When comparing the two estimates, the second one gives a better model of our data and is hence used in further analysis.

4.3.3 Expectation and Variance of Total Annual Payout

The expectation of the sum of total payouts per year is calculated using Theorem 6 in Section 3.4. Two sums are calculated and combined, one for total payouts below 600 000 SEK where the payouts are gamma distributed, and one for total payouts above 600 000 SEK where the payouts are GPD.

Note that the probability that a payout, X, is either above or below 600 000 SEK is one. If the number of payouts below and above the threshold are Poisson processes, let N_1 and N_2 represent the number of payouts per year below and above 600 000 SEK respectively. Denote their intensity measures by λ_1 and λ_2. By Theorem 5 in Section 3.3, the intensity measures can be deduced from the following equation

$$\lambda_1 + \lambda_2 = \mathrm{P}(X < 600000) \cdot \mu_{total} + \mathrm{P}(X > 600000) \cdot \mu_{total} = \mu_{total},$$

where μ_{total} is the mean of the total amount of payouts per year from the second method in Section 4.3.2.

Let Y denote the payouts which are gamma distributed and let Z denote the ones that are GPD. Then, the sums of total payouts above and below 600 000 SEK is given by

$$S_{N_1} = \sum_{i=1}^{N_1} Y_i \quad \text{and} \quad S_{N_2} = \sum_{i=1}^{N_2} Z_i.$$

Finally, from Theorem 6 in Section 3.4, the expectation of the total annual payout is

$$\mathrm{E}(S_{N_1} + S_{N_2}) = \mathrm{E}(N_1) \cdot \mathrm{E}(Y) + \mathrm{E}(N_2) \cdot \mathrm{E}(Z).$$

The variance of the total annual payout is also calculated using Theorem 6 in the following way

$$\mathrm{Var}(S_{N_1} + S_{N_2}) =$$

$$= \mathrm{E}(N_1) \cdot \mathrm{Var}(Y) + (\mathrm{E}(Y))^2 \cdot \mathrm{Var}(N_1) + \mathrm{E}(N_2) \cdot \mathrm{Var}(Z) + (\mathrm{E}(Z))^2 \cdot \mathrm{Var}(N_2).$$

Note that

$$P(X < 600000) = \frac{\text{Number of payouts less than 600 000 SEK}}{\text{Total number of payouts}}$$

and

$$\mathrm{P}(X > 600000) = \frac{\text{Number of payouts above 600 000 SEK}}{\text{Total number of payouts}}.$$

Also note that if N_1 and N_2 are Poisson distributed, then

$$\mathrm{E}(N_1) = \mathrm{Var}(N_1) = \lambda_1$$

and

$$\mathrm{E}(N_2) = \mathrm{Var}(N_2) = \lambda_2.$$

5 Results

In this section the estimated parameters and return levels with corresponding confidence intervals is presented for the GEV and GPD models. In addition to this, the max-stability property of the GEV is checked and the result is presented below. This section also contains the result of the estimated amount of payouts per year as well as the expected total annual payout in SEK together with the confidence interval.

5.1 Generalized Extreme Value Distribution (GEV)

The model chosen as the best fit for the data is the one with a six month block size where the parameters do not depend on time. The independence of time was concluded from performing likelihood-ratio tests with significance level $\alpha = 0.05$.

5.1.1 Parameters

The parameters for the six month block size selection as well as their respective 95% confidence intervals and standard deviations are presented in Table 1 below. The shape parameter for the model is positive which means that the fitted distribution is Fréchet. However, since the confidence interval contains both positive and negative values as well as zero, the Weibull and Gumbel distributions might also be appropriate fits for the data. The estimations of the parameters are calculated using the maximum likelihood method and the confidence intervals are found using the delta method.

Table 1: GEV parameter estimates for six month block size.

Parameter	**Estimate**	**Lower CI**	**Upper CI**	**Standard deviation**
μ	$1.082 \cdot 10^6$	$7.282 \cdot 10^5$	$1.434 \cdot 10^6$	$1.806 \cdot 10^5$
σ	$7.158 \cdot 10^5$	$3.336 \cdot 10^5$	$1.098 \cdot 10^6$	$1.950 \cdot 10^5$
ξ	$6.271 \cdot 10^{-2}$	$-6.162 \cdot 10^{-1}$	$0.742 \cdot 10^{-1}$	$3.464 \cdot 10^{-1}$

5.1.2 Max-Stability

In order to check the validity of the model, parameters and confidence intervals is also calculated for the three and twelve month block size selection.

If the model is a good fit for the data it should be max-stable, and hence the estimated parameters, Table 2, should lie within the confidence intervals, Table 3, created using estimates calculated from a model with a different block size selection, Table 4.

As can be seen from Tables 2 and 3, the only confidence interval that does not contain its designated parameter is the one for the shape parameter of the annual block size calculated from the parameters of the three month block size. As argued before, neither the three nor the twelve month block size selections provides the best fit for the data. For the three month block size some of the time periods, namely the first quarters of 2002, contains too few data points from which the maximum is taken. If there are not sufficiently many observations in each block the asymptotic properties of the GEV model is lost and the provided fit might not be a very suitable one. For the annual block size since the model was fitted to too few maxima. That both models are somewhat flawed might be the explanation as to why the connection between these two models is not max-stable. The max stability property holds for the connection between the three and six month block size selection as well as for the one between the six and twelve month block size selection, indicating that the six month block size selection might be an appropriate fit for the data.

Here the parameters μ, σ and ξ are calculated using the maximum likelihood method and the 95% confidence interval are calculated using the delta method.

Table 2: GEV parameters for different block sizes.

block size	**Annual**	**Six months**	**Quarterly**
μ	$1.855 \cdot 10^6$	$1.082 \cdot 10^6$	$7.344 \cdot 10^5$
σ	$9.260 \cdot 10^5$	$7.158 \cdot 10^5$	$4.641 \cdot 10^5$
ξ	$-4.549 \cdot 10^{-1}$	$6.271 \cdot 10^{-2}$	$5.390 \cdot 10^{-1}$

Table 3: Confidence interval calculated from theoretical GEV parameters.

	Six months $\longrightarrow$ Annual	**Quarterly $\longrightarrow$ Annual**	**Quarterly $\longrightarrow$ Six months**
I_{μ^*}	$(1.112 \cdot 10^6, 2.067 \cdot 10^6)$	$(9.912 \cdot 10^5, 2.391 \cdot 10^6)$	$(7.971 \cdot 10^5, 1.452 \cdot 10^6)$
I_{σ^*}	$(4.134 \cdot 10^5, 1.082 \cdot 10^6)$	$(9.036 \cdot 10^4, 1.869 \cdot 10^6)$	$(2.674 \cdot 10^5, 1.081 \cdot 10^6)$
I_{ξ^*}	$(-6.162 \cdot 10^{-1}, 7.416 \cdot 10^{-1})$	$(-7.281 \cdot 10^{-3}, 1.085)$	$(-7.281 \cdot 10^{-3}, 1.085)$

Table 4: Theoretical GEV parameter estimation.

Max-stability parameters	k	$\mu^* = \mu + \frac{\sigma}{\xi}(k^\xi - 1)$	$\sigma^* = \sigma k^\xi$	$\xi^* = \xi$
Six months → Annual	2	$1.589 \cdot 10^6$	$7.476 \cdot 10^5$	$6.271 \cdot 10^{-2}$
Quarterly → Annual	4	$1.691 \cdot 10^6$	$9.799 \cdot 10^5$	$5.390 \cdot 10^{-1}$
Quarterly → Six months	2	$1.124 \cdot 10^6$	$6.7441 \cdot 10^5$	$5.390 \cdot 10^{-1}$

5.1.3 Return Levels

The return levels in Table 5 are calculated using maximum likelihood estimation, see Section 3.1.3, and the 95% confidence intervals are calculated using both the delta method and the profile likelihood method. The 10 year return level, in other words the level that will be exceeded on average once during a ten year period, is approximately 2.8 million SEK. This agrees somewhat with the data, where this level was exceeded twice in a ten year period. However, since the 10 year return level is only an average, it is possible for this level to be exceeded more that once during a ten year period or to not be exceeded at all. It is also worth mentioning that the smaller one of the two exceedances only exceeded the ten year return level by approximately 100 000 SEK.

Since the data does not contain information from a 50 year period, it is more difficult to analyze the plausibility of this estimate. The 50 year return level has hitherto not been exceeded since the highest value during the ten years used for this analysis is approximately 3.5 million SEK. According to these estimates the level that on average will be exceeded once every 50 years is 4.2 million SEK.

Table 5: Return levels for block size six months.

Method	**Return Period**	**Return Level**	**Lower CI**	**Upper CI**
Delta	10 year	$2.812 \cdot 10^6$	$1.701 \cdot 10^6$	$3.924 \cdot 10^6$
	50 year	$4.246 \cdot 10^6$	$7.174 \cdot 10^5$	$7.776 \cdot 10^6$
Profile Likelihood	10 year	$2.812 \cdot 10^6$	$2.267 \cdot 10^6$	$5.979 \cdot 10^6$
	50 year	$4.246 \cdot 10^6$	$3.001 \cdot 10^6$	$2.458 \cdot 10^7$

Comparing the confidence intervals for the two different methods, it is the

intervals calculated using the profile likelihood method that should be the most accurate, see plots in Figure 9 and Figure 10 below. Note that the confidence interval for the 50 year return period is large and does therefore not give a very precise description of future payouts. The confidence interval for the 100 year return period is also very large and gives a very poor picture, it is omitted for this reason.

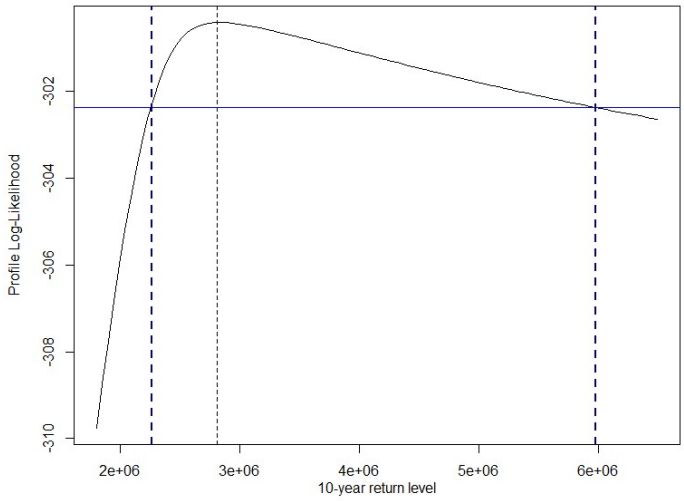

Figure 10: Confidence interval for 10 year return level using profile likelihood.

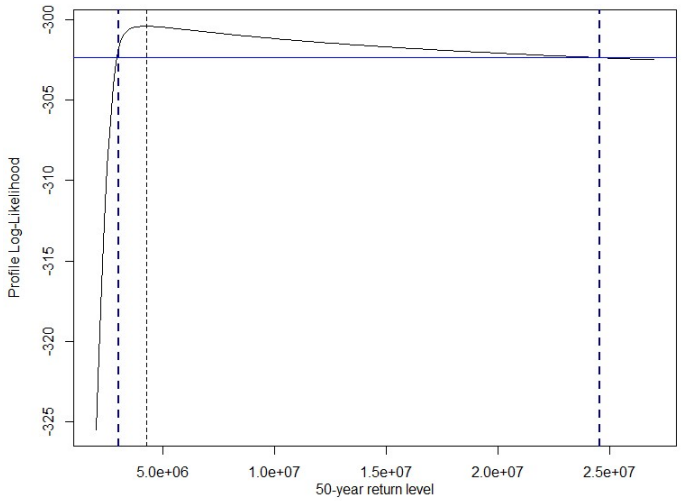

Figure 11: Confidence interval for 50 year return level using profile likelihood.

5.2 Generalized Pareto Distribution (GPD)

The results in this section is based on the threshold u=600 000 with 86 exceedances.

5.2.1 Parameters

The estimated parameters for the GPD model and their 95% confidence intervals are presented in Table 6 below. The parameters are calculated using the maximum likelihood method and the confidence intervals are found using the delta method.

Table 6: Theoretical GPD parameter estimation.

Parameter	**Estimate**	**Lower CI**	**Upper CI**	**Standard deviation**
σ	$4.539 \cdot 10^5$	$4.481 \cdot 10^5$	$4.597 \cdot 10^5$	$2.968 \cdot 10^3$
ξ	$2.043 \cdot 10^{-1}$	$1.454 \cdot 10^{-3}$	$4.071 \cdot 10^{-1}$	$1.035 \cdot 10^{-1}$

5.2.2 Return Levels

Return levels for 10-, 50-, and 100 year return period is calculated using maximum likelihood estimation, the 95% confidence intervals are calculated using

the profile likelihood method. They are presented in Table 7 below.

Table 7: Return levels for GPD.

Return Period	**Return Level**	**Lower CI**	**Upper CI**
10 year	$4.662 \cdot 10^6$	$3.180 \cdot 10^6$	$1.048 \cdot 10^7$
50 year	$7.108 \cdot 10^6$	$3.978 \cdot 10^6$	$2.429 \cdot 10^7$
100 year	$8.435 \cdot 10^6$	$4.390 \cdot 10^6$	$3.486 \cdot 10^7$

When comparing the return levels in Table 7 to what can be seen in the data, it agrees relatively well. During the 17 years which are included in this analysis, the largest payout is approximately 4.1 million SEK. This means that there are no payouts which exceed the 10 year return level of approximately 4.6 million SEK. In other words, the level that should be exceeded once on average every 10 years has not been exceeded in 17 years. It can be argued that this might still agree with the property that it should be on average exceeded once every 10 years since, first of all, it is only an average. Second of all, some of the insurances are still open, meaning that Gar-Bo can still receive claims exceeding the 10 year return level. There is no data from a 50 year period, however, it can be seen that the 50 year return level has not been exceeded as of yet. The same holds for the 100 year return level.

It should be noted that the confidence intervals of the return levels are very large implying that there is an uncertainty around how large future payouts might be. An explanation to this is that the shape parameter ξ is not negative, see Table 6, meaning that there is no upper limit to how large a payout can be. Figures 11, 12 and 13 shows plots of the different confidence intervals using profile likelihood method.

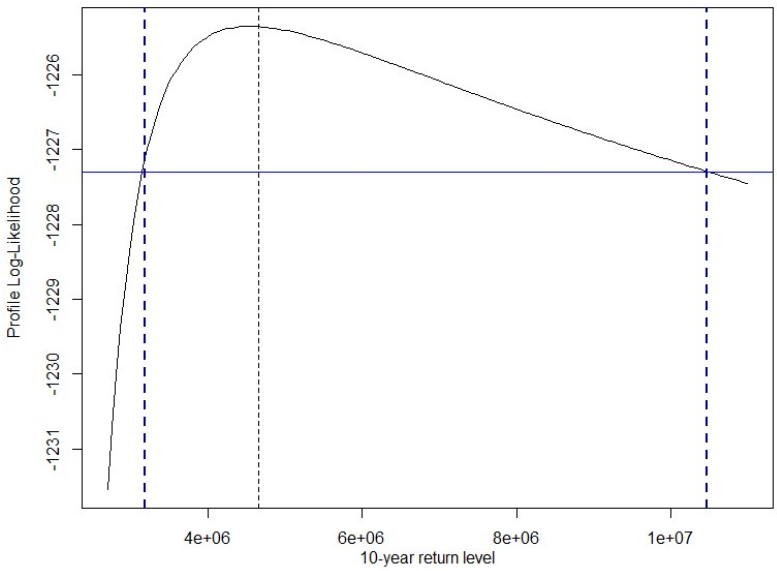

Figure 12: Confidence interval for 10 year return level.

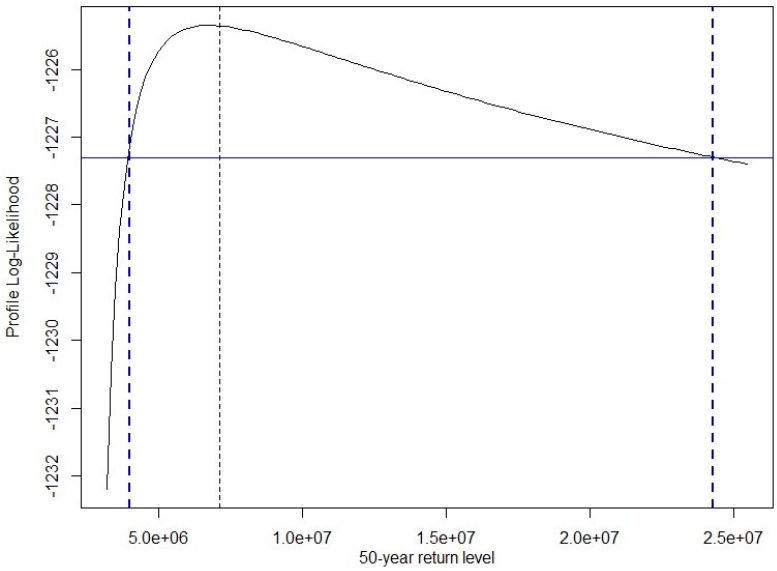

Figure 13: Confidence interval for 50 year return level.

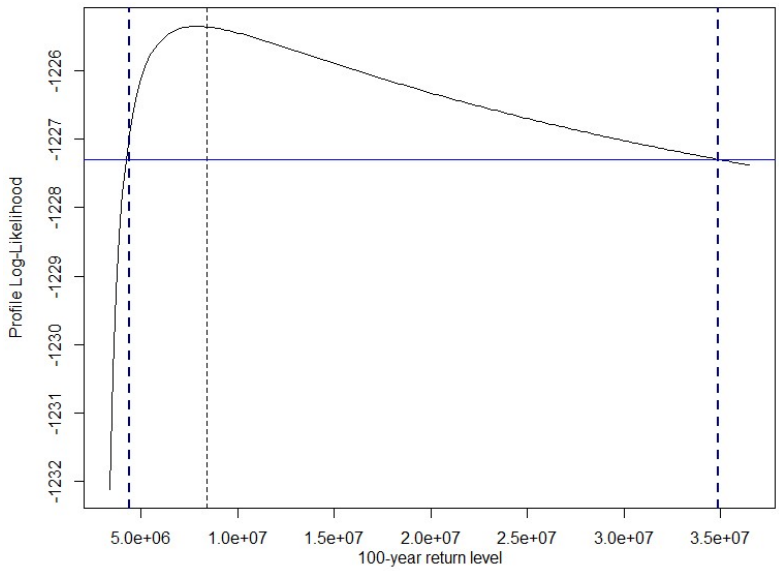

Figure 14: Confidence interval for 100 year return level.

5.3 Total Annual Payout

The expected cost of the total amount of payouts each year is presented in this section. The result is based on the second method from Section 4.3.2 where $\mu_{total} = 5.5\mu$.

5.3.1 Parameters and Confidence Interval

The intensity parameter is estimated as $\mu \approx 35.27$. The amount of payouts per year is given by $5.5\mu = \mu_{total} \approx 194$, meaning that there are, according to this model, approximately 194 payouts per year made by Gar-Bo for this specific insurance.

How large the total sum of the 194 payouts is expected to be is calculated using the method in Section 4.3.3. The result is presented in Table 8 below and the values used to obtain this result are presented in Table 9 in the Appendix.

Table 8: Expected total payout per year in SEK.

Total Payout	**Lower CI**	**Upper CI**	**Standard deviation**
$40.903 \cdot 10^6$	$31.780 \cdot 10^6$	$50.026 \cdot 10^6$	$4.654 \cdot 10^6$

When comparing this result to the data it seems like a reasonable estimate. Again the confidence interval is rather large, however the total annual payout differs quite a bit and thus a large confidence interval is natural.

6 Conclusions and Further Research

As can be seen in the result section, the GEV and the GPD models give different results for the 10 and 50 year return levels. The return levels for the GEV are a bit lower than what has been observed so far and the GPD model gives higher return levels than what can be seen in the data. From a theoretical perspective, the more accurate result should in this case be the return level obtained from the GPD model. This is because of the fact that, when the maxima is taken over every six months, some of the larger payouts might be disregarded if they are not the maxima in that time interval. For the GPD model, all exceedances over the chosen threshold are considered, and thus this model gives higher return levels. Since this model gives a better accuracy, a 100 year return level can also be calculated.

For both the GEV and the GPD models, the confidence intervals for the return levels are rather wide, making it difficult to draw exact conclusion on which levels will be exceeded on average once every 10, 50 and 100 years. The calculated return levels can however be seen as indications for how large these levels might be.

As for the estimated total annual payout, this too should be considered as an indication, and not an exact result, of the expected total payout made by Gar-Bo in a year. The confidence interval could be seen as quite wide, which might be natural in this case since there are a lot of factors that affect the size of the total payout.

One such aspect worth discussing regarding the type of insurance that is dealt with in this thesis is that new building techniques are constantly developed and implemented. The new techniques might not be seen as problematic as of today, but they might cause extensive damage discovered further on in time. This would perhaps cause insurance companies as Gar-Bo to receive claims which could not be predicted from the data we received when writing this thesis.

The construction business is of course dependent on certain hardware and raw material. If a large manufacturer can no longer provide these types of supplies, alternative solutions must be developed. This could lead to that new solutions which have not been tested for very long are implemented and might result in damage further on in time as discussed in the previous section.

For further research it would be interesting to divide the different types of damages into groups and perform the same analysis as in this thesis but within the different groups. For example, damages in wetrooms and damages caused by water penetrating roofs would be in two separate groups. In this way the return levels and the expected annual payout per damage category could be investigated. It would give a great overview of the short- and long-term differences between the categories, and could perhaps even give more exact results.

7 References

[1] Coles. S. *An Introduction to Statistical Modeling of Extreme Values*. London: Springer, 2004.

[2] Haan. L and Ferreira. A. *Extreme Value Theory An Introduction*. New York: Springer, 2006.

[3] Beirlant. J. *Statistics of Extremes Theory and Applications*. Chichester: J. Wiley, 2004.

[4] Gumbel. E.J. *Statistics of Extremes*. New York: Columbia University Press, 1958.

[5] Miljödepartementet. *Avskaffande av den obligatoriska byggfelsförsäkringen*. (Ds 2011:2). Stockholm: Fritzes. URL: https://www.regeringen.se/49bbd3/contentassets/03872e54a36d4729b8c92e6e2a43de14/avskaffande-av-den-obligatoriska-byggfelsforsakringen-ds.-20112.

[6] Konsumenternas. *Färdigställandeskydd och byggfelsförsäkringar*. URL: https://www.konsumenternas.se/forsakringar/boendeforsakringar/fardigstallandeskydd/.

[7] Gar-Bo Försäkring AB. *Om oss*. URL: https://www.gar-bo.se/om-gar-bo.

[8] Gar-Bo Försäkring AB. *Villkor Nybyggnadsförsäkring*. URL: https://www.gar-bo.se/sites/default/files/uploads/Rapporter/villkor_nybyggnadsforsakring_nbf_1-3_2022-01-01.pdf.

[9] Gar-Bo Försäkring AB. *För- och efterköpsinformation Nybyggnadsförsäkring*. URL: https://www.gar-bo.se/sites/default/files/uploads/for-_och_efterkopsinformation_nybyggnadsforsakring_nbf_1-2_2020-05-01.pdf.

[10] Boverket. *Kartläggning av fel, brister och skador inom byggsektorn*. Karlskrona: Boverket, 2018. Rapportnummer: 2018:36. URL: https://www.boverket.se/globalassets/publikationer/dokument/2018/kartlaggning-av-fel-brister-och-skador-inom-byggsektorn.pdf.

[11] SCB. *Inflation i Sverige 1831-2020*. URL: https://www.scb.se/hitta-statistik/statistik-efter-amne/priser-och-konsumtion/konsumentprisindex/konsumentprisindex-kpi/pong/tabell-och-diagram/konsumentprisindex-kpi/inflation-i-sverige/.

[12] Rydén. T and Lindgren. G. *Markovprocesser*. Lund: Lund University with Lund Institute of Technology, 2000.

[13] Gut. A. *An Intermediate Course in Probability.* 2nd ed. New York: Springer, 2009.

8 Appendix

8.1 Probability and Quantile Plots - GEV

Let $z_{(1)} \leq z_{(2)} \leq ... \leq z_{(m)}$ be the ordered block maxima to which a GEV model is fitted. The estimated GEV model is denoted $\hat{G}$. Then the probability plot, pp-plot, is created using the following pair:

$$\Big\{\Big(i/(m+1), \hat{G}(z_{(i)}\Big), \qquad i = 1, ..., m\Big\},$$

where

$$\hat{G}(z_{(i)}) = \exp\Bigg\{ - \Big[1 + \hat{\xi}\Big(\frac{z_{(i)} - \hat{\mu}}{\hat{\sigma}}\Big)\Big]^{-\frac{1}{\hat{\xi}}}\Bigg\}.$$

The quantile plot, qq-plot, is created using the following pair:

$$\Big\{\Big(\hat{G}^{-1}(i/(m+1)), z_{(i)}\Big), \qquad i = 1, ..., m\Big\},$$

where

$$\hat{G}^{-1}(z) = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\Big[1 - \Big(\log(z)\Big)^{\hat{\xi}}\Big].$$

If the pp- and qq-plot consist of points which lie close to the unit diagonal which corresponds to the theoretical GEV model, then the estimated GEV model $\hat{G}$ can be considered a reasonable model [1].

8.2 Probability and Quantile Plots - GPD

Let $y_{(1)} \leq y_{(2)} \leq ... \leq y_{(k)}$ denote the exceedances over a threshold u and let $\hat{H}$ denote the estimated generalized Pareto model.Then the probability plot (pp-plot) is created using the following pair:

$$\Big\{\Big(i/(k+1), \hat{H}(y_{(i)}\Big); \qquad i = 1, ..., k\Big\}.$$

and, assuming $\hat{\xi} \neq 0$, the quantile plot (qq-plot) is created using the following pair:

$$\Big\{\Big(\hat{H}^{-1}(i/(k+1)), y_{(i)}\Big), \qquad i = 1, ..., k\Big\}.$$

where

$$\hat{H}(y) = \begin{cases} 1 - \left(1 + \frac{\hat{\xi} y}{\hat{\sigma}}\right)^{-\frac{1}{\hat{\xi}}}, & \hat{\xi} \neq 0 \\ 1 - \exp\left(\frac{y}{\hat{\sigma}}\right), & \hat{\xi} = 0. \end{cases}$$

and

$$\hat{H}(y)^{-1} = u + \frac{\hat{\sigma}}{\hat{\xi}}\left[y^{-\hat{\xi}} - 1\right].$$

For the estimated generalized Pareto model $\hat{H}$ to be considered reasonable for modelling excesses over a threshold u, both the pp-plot and qq-plot should consists of points that are close to the unit diagonal, corresponding to the theoretical model [1].

8.3 Block Size Selection

Figures with pp- and qq-plots, return level plots and density plots for block size of three and twelve months.

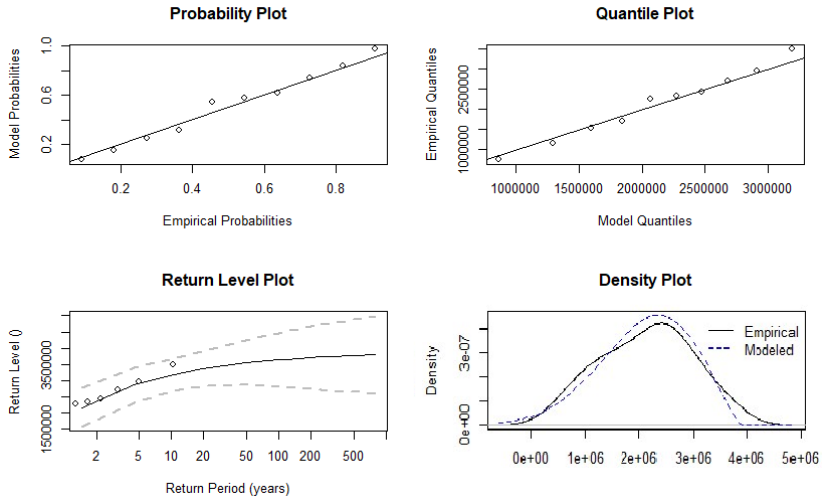

Figure 15: Twelve month block size.

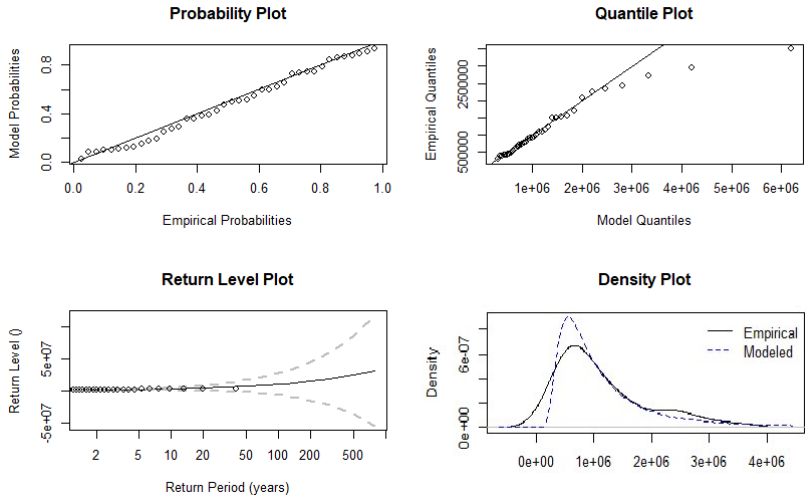

Figure 16: Three month block size.

8.4 Likelihood Ratio Test

Let $x_1, x_2, ..., x_n$ be independent realizations of a random variable X with probability density function $f(x, \theta)$. The likelihood function is defined as

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Since the log-function is monotonic the log-likelihood function has its maximum at the same point as the likelihood function. It is sometimes more convenient to use and is defined in the following way:

$$\ell(\theta) = L(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta).$$

The likelihood function gives the probabilities of the observed data as a function of θ. Different values of θ corresponds to different models, with different probabilities, to the observed data. Naturally, it is of interest to find the θ which generates the greatest value for the likelihood function since that corresponds to the model with the highest probability to the observed

data. The method of estimating $\hat{\theta}$, the θ which maximizes the likelihood function, is called maximum likelihood estimation.

When two maximized log-likelihood models should be tested against each other to determine which one is the best suited model, the deviance function can be used. Defined as:

$$D(\theta) = 2\{\ell(\hat{\theta}_0) - \ell(\theta)\}.$$

The deviance function satisfy $D(\theta_0) \sim \chi^2_d$, where d is the dimension of the model parameter θ_0. Let c_α be the $(1-\alpha)$ quantile of the χ^2_d distribution, then

$$C_\alpha = \{\theta : D(\theta) \leq c_\alpha\}$$

is the $(1-\alpha)$ confidence region for θ_0.

When testing two models, $\mathcal{M}_0$ and $\mathcal{M}_1$, against each other the deviance statistics is used, defined as:

$$D = 2\{\ell_1(\mathcal{M}_1) - \ell_0(\mathcal{M}_0)\}.$$

Note that model $\mathcal{M}_0$ is a subset of model $\mathcal{M}_1$ where k of the components of θ are restricted to be zero. Thus it is sufficient to check if 0 lies in C_α or not, equivalent to checking if $D < c_\alpha$, to know if it is possible to use model $\mathcal{M}_0$ as a reduced model of $\mathcal{M}_1$.

The test is called a likelihood ratio test, explained in the following theorem.

Theorem 7 [1] Suppose $\mathcal{M}_0$ with parameter $\theta^{(2)}$ is the sub-model of $\mathcal{M}_1$ with parameter $\theta_0 = (\theta^{(1)}, \theta^{(2)})$ under the constrains that the k-dimensional sub-vector $\theta^{(1)} = \mathbf{0}$. Let $\ell_0(\mathcal{M}_0)$ and $\ell_1(\mathcal{M}_1)$ be the maximized values of the log-likelihood for models $\mathcal{M}_0$ and $\mathcal{M}_1$ respectively. A test of the validity of model $\mathcal{M}_0$ relative to $\mathcal{M}_1$ at the α level of significance is to reject $\mathcal{M}_0$ in favor of $\mathcal{M}_1$ if $D = 2\{\ell_1(\mathcal{M}_1) - \ell_0(\mathcal{M}_0)\} > c_\alpha$, where c_α is the $(1-\alpha)$ quantile of the χ^2_k distribution.

8.5 Total Annual Payout

The parameters and their values used to calculate the total annual payout per year and the confidence interval in Section 5.3.1.

Table 9: Values used to calculate the total annual payout.

Total number of payouts	$= 1937$
$\mathrm{E}(N_1) = \mathrm{Var}(N_1) = N_1$	$= 1.854 \cdot 10^2$
$\mathrm{E}(N_2) = \mathrm{Var}(N_2) = N_2$	$= 8.613$
$\mathrm{P}(X < 600000)$	$= 0.956$
$\mathrm{P}(X > 600000)$	$= 0.044$
$\mathrm{E}(Y)$	$= 1.663 \cdot 10^5$
$\mathrm{E}(Z)$	$= 1.170 \cdot 10^6$
$\mathrm{Var}(Y)$	$= 1.411 \cdot 10^2$
$\mathrm{Var}(Z)$	$= 5.502 \cdot 10^{11}$

Bachelor's Theses in Mathematical Sciences 2022:K2
ISSN 1654-6229

LUNFMS-4062-2022

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lu.se/