universität wien

# MAGISTERARBEIT / MASTER'S THESIS

Titel der Magisterarbeit / Title of the Master's Thesis

„Modelling of low probability high impact events“

verfasst von / submitted by

Michal Majka, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magister der Sozial- und Wirtschaftswissenschaften (Mag. rer. soc. oec.)

Wien, 2018 / Vienna 2018

| | |
|---|---|
| Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet: | A 066 951 |
| Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet: | Magisterstudium Statistik UG2002 |
| Betreut von / Supervisor: | o. Univ.-Prof. Mag. Dr. Georg Pflug |

# Contents

# Acknowledgements

I thoroughly enjoyed writing this thesis. I would like to thank my supervisor, professor Pflug, for giving me a lot of freedom and supporting me throughout the process.

*Dedicated to Tadeusz and my mum.*

# 1 Introduction

There is often a need for modelling of losses incurred due to the underlying process that generates a very large number of minor occurrences, which are accompanied by very few high impact or even catastrophic events. In the real world, to name few examples, it translates to modelling of damages caused by natural disasters (e.g. earthquakes, hurricanes, tornados), operational losses in big and highly regulated institutions as banks or even claims generated by an insurance policy. In the all above mentioned cases, whether it may be policy making, regulatory capital requirements or managing risks, statistical modelling plays an important role since it makes it possible to draw rational conclusions in face of uncertainty about the distributions of aggregate losses based on the empirical data over a given time period of study. The data of interest inherits the property that is often summarized as high-frequency-low-severity (HFLS) and low-frequency-high-severity (LFHS) or high-probability-low-impact (HPLI) and low-probability-high-impact (LPHI). The latter type makes the modelling challenging as there are many outliers which are not easily captured by classical models.

One possible modelling strategy, pursued throughout this thesis, is to record all individual losses and then build the sum for the time period. Since the number of losses is not known a priori, the aggregate losses are represented as a sum of a random number of non-negative summands. The latter are assumed to be independent and identically distributed and also independent of the underlying process that governs the number of occurrences. Such modelling strategy refers to a *compound model* where the number of losses and their sizes are modelled with different distributions - counting and severity distributions, respectively.

The purpose of this thesis is to describe the compound model in detail and apply it to the real world problem characterized by the HFLS and LFHS property, where the main objective is to model the distribution of yearly losses. Such undertaking requires the development of a flexible modelling strategy that can be easily extended for the needs of particular instance of a problem. The general idea and characteristics of the compound model are introduced in the second section in a formal way. Then the structure of this thesis aims to follow natural steps in the model development process. Hence, the third section deals with the problematic of the modelling of the severity part. It first focuses on the extreme value theory that justifies the use of threshold models, in particular Generalized Pareto Distribution, to model tail events.

It then describes a modelling strategy that can be applied to losses that fall below the threshold. The fourth section shifts the focus to the frequency part of the compound model. At first, desired properties of frequency distributions are described and then few particular instances, that can be represented as members of a broader (a,b,0) class, are introduced: binomial, Poisson and negative binomal models. After the severity and frequency distributions are chosen based on theoretical considerations and explorative analysis, selected models have to be estimated using some parametric estimation techniques - which is the main focus of section 5. In general, desired properties of estimation methods are described of which the discussion is followed with the introduction to specific techniques, such as maximum likelihood and minimum distance estimation. Section 6 deals with the diagnostic of chosen frequency and severity distributions. In the last section, the theory developed throughout the thesis is consolidated and applied to model yearly damages incurred due to tornadoes in the United States over the past 30 years.

# 2 The Compound Model for Aggregate Losses

In this section the general framework of the compound model of aggregate losses is discussed and formally introduced. It is followed by the derivation of the expectation, variance and the moment generating function. Finally, a Monte Carlo algorithm that allows approximating the compound distribution is provided.

Let us suppose that during some specified period of time a set of a random number, $N$, of losses denoted by $X_1, ..., X_N$, where $N$ is a counting distribution, was observed. Furthermore, $X_i$s are continuous random variables that take on non-negative values and are mutually independent regardless of the number of losses. They also have a common distribution function $F_X(x) = P(X \leq x)$ that does not depend on any particular realisation of the counting distribution $N$. The aggregate losses can be then obtained by summing over all $X_i$s

$$S = \sum_{i=1}^{N} X_i \,, \quad \text{for } N = 0, 1, 2, ...$$

where $S = 0$ whenever $N = 0$. Due to the inherent property of heavy tails and presence of outliers in the considered modelling problems, classical statistical methods like linear regression are not appropriate tools for modelling of sizes of individual losses, denoted previously by $X_i$s, since they usually focus on predicting the conditional mean of data. Instead, a more reasonable approach is to first focus on the extrema - the tail events themselves and model their full conditional probabilistic distribution separately. Such tail is sparsely populated with data points, due to the LFHS property as well as by definition of the tail, and for this reason some distributional assumption has to be made allowing interpolating available points and further extrapolating beyond the range of the data. The other part of the distribution, a body, separated from the tail by a threshold usually contains a high number of points and for this reason, it can be modelled via empirical distribution. This results in a severity distribution that is split into two parts: body and tail via an appropriate threshold. The choice of the tail distribution as well as the problematic of the threshold selection are discussed in greater detail in section 3. The section 4 is devoted to relevant counting distributions.

A distribution function of the random sum $S$ is given by

$$F_S(x) = P(S \leq x) = \sum_{n=0}^{\infty} P(N = n)P(S \leq x|N = n) = \sum_{n=0}^{\infty} p_n F_X^n(x),$$

where, to simplify the notation, $p_n = P(N = n)$. The term $F_X^n(x)$ refers to the $n$-fold convolution of the distribution function $F_X(x)$, which is defined as

$$F_X^0(x) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$F_X^m(x) = \int_0^x F_X^{(m-1)}(x - y) f_X(y) dy.$$

The corresponding probability density function can be obtained by differentiation and is given by

$$f_X^m = \int_0^x f_X^{m-1}(x - y) f_X(y) dy.$$

Finally, the probability distribution of the aggregate distribution $F_S(x)$ is given by

$$f_S(x) = \sum_{n=0}^{\infty} p_n f_X^n(x).$$

Since $X_i$s and $N$ are independent, the expectation of the aggregate distribution can be easily obtained by conditioning

$$\mathbb{E}(S) = \mathbb{E}(\sum_{i=1}^{N} X_i) = \mathbb{E}\left(\mathbb{E}\left(\sum_{i=1}^{N} X_i|N\right)\right) = \mathbb{E}(N\,\mathbb{E}(X_1)) = \mathbb{E}(N)\mathbb{E}(X_1).$$

In order to compute the variance, the independence of $X_i$s and $N$ as well as total variance formula can be exploited

$$\begin{aligned} Var(S) &= \mathbb{E}_N(Var(S|N)) + Var_N(\mathbb{E}(S|N)) \\ &= \mathbb{E}_N(Var\left(\sum_{i=1}^{N} X_i|N\right)) + Var_N(\mathbb{E}\left(\sum_{i=1}^{N} X_i|N\right)) \\ &= \mathbb{E}_N(N\,Var(X_1)) + Var_N(N\,\mathbb{E}(X_1)) \\ &= \mathbb{E}_N(N)Var(X_1) + Var_N(N)\mathbb{E}(X_1)^2. \end{aligned}$$

The moment generating function of $S$, provided that $X_i$s are non-negative, can be calculated in the following way

$$\begin{aligned} M_S(z) &= \mathbb{E}(z^S) \\ &= \sum_{n=0}^{\infty} \mathbb{E}(z^{\sum_{i=1}^{n} X_i} | N = n) P(N = n) \\ &= \sum_{n=0}^{\infty} \mathbb{E}\left(\prod_{i=1}^{n} z^{X_i}\right) P(N = n) \\ &= \sum_{n=0}^{\infty} P(N = n) \left(M_X(z)\right)^n \\ &= \mathbb{E}_N(M_X(z)^N) \\ &= M_N(M_X(z)) \end{aligned}$$

In general, the compound distribution $F_S$ can be analytically obtained only in few cases. For instance, when the negative binomial distribution serves as a counting distribution and damages are modelled with exponential distribution (for more examples see section 4.4 in Klugman, Panjer, and Gordon, 1988 or section 6.4 in Panjer, 2006). In most cases, one has to usually resort to Monte Carlo simulations in order to find an approximate solution to $F_S$. The general simulation procedure is simple in that it only requires sampling from estimated distributions and building sums.

It is assumed that the counting and severity distributions are selected and their estimates are given by $\hat{F}_N$ and $\hat{F}_X$, respectively. Furthermore, the latter is a full distribution, i.e. it is not split into the body and tail; and the number of Monte Carlo iterations, $M$, is reasonably high. The compound distribution can be then approximated in the following way:

For each $i \in \{1, 2, 3, ..., M\}$

- Sample a number of losses $n_i$ from the estimated counting distribution $\hat{F}_N$
- Sample $X_1, X_2, ...X_{n_i}$ from the estimated severity distribution $\hat{F}_X$
- Compute $S_i = \sum_{k=1}^{n_i} X_k$ and store it

This yields $M$ different values that represent sums of losses for a predetermined time period. The higher the number of Monte Carlo iterations $M$, the more accurate the approximation of the compound distribution $F_S$.

When the severity distribution is split into the body and tail, the simulation is slightly more involved. First, the number of losses in the body is simulated from the estimate of the separate counting distribution $\mathrm{F}_{N,\text{body}}$ and damages are sampled (with replacement) from the empirical distribution $\hat{F}_{body}$. Analogically, the number of loss events in the tail is simulated from $\hat{F}_{N,\text{tail}}$ and the sizes of losses are sampled from the estimated tail distribution $\hat{F}_{N,\text{tail}}$. More formally, the simulation can be structured as follows:

For each $i \in \{1, 2, 3, ..., M\}$

- Sample a number of losses $n_i$ from $\hat{F}_{N,\text{body}}$
- Sample a number of losses $m_i$ from $\hat{F}_{N,\text{tail}}$
- Sample $X_1, X_2, ...X_{n_i}$ from $\hat{F}_{\text{X,body}}$
- Sample $Y_1, Y_2, ...Y_{m_i}$ from $\hat{F}_{\text{X,tail}}$
- Compute $S_i = \sum_{j=1}^{n_i} X_j + \sum_{k=1}^{m_i} Y_k$ and store it

# 3 Modelling of Severity

In this section the focus is put on the modelling of the severity distribution, which is split into the tail and the body. The former is modelled with an appropriate parametric distribution and addresses the low-frequency-high-severity aspect of the data. The latter is equipped with an empirical distribution that takes an advantage of the high-frequency-low-severity property of the underlying losses. This approach implies that there is a certain point that distinguishes body from the tail and this burning issue is also addressed.

In general, the discussion begins with desired properties of parametric distributions which favour successful modelling of the tail risk. Then, the general framework of the Extreme Value Theory is introduced and its application to the modelling of losses in the tail is analyzed. The discussion is then followed by an inventory of relevant parametric models that are applicable to many different instances of the right skewed data. Finally, the modelling of the body with an empirical distribution is considered.

## 3.1 Modelling of the tail

Capturing the low-frequency-high-severity (LFHS) property of the underlying dataset is a crucial part for a realistic model of sizes of losses. This issue can be directly addressed by focussing on the extreme points that can be found in the tail. There is potentially an infinite number of models that could be taken in consideration for modelling purposes, as any function after appropriate scaling can be considered to be a valid probability distribution - the only requirement is that it integrates to 1. In general, a *parametric family of probability distributions* or a *parametric model* is a collection of distribution functions whose members are indexed by a finite-dimensional parameter vector $\theta$. The family can be formally represented as

$$\{F(x;\theta) : \theta \in \Theta\},$$

where $\Theta \in \mathbb{R}^k$ is a set of all possible parameter values. Datasets relevant to the topic are heavy tailed as they are inhabited by many enormous outliers and therefore models used for modelling purposes must be able to reflect increased risk at right endpoint of the distribution, which inevitably reduces number of applicable models. First of all, they should be defined on a non-negative support, since a loss cannot be negative. Also, their tails should be sufficiently flexible to accommodate for different shapes while the

distribution should be smooth and posses relatively simple functional form.

### 3.1.1 Extreme Value Theory

In the following, the classical result in Extreme Value Theory, known as the *Fisher–Tippett–Gnedenko theorem*, is first introduced. It describes the asymptotic behaviour of the sample maxima after appropriate normalisation. Then, the more relevant theorem to this thesis, the *Pickands–Balkema–de Haan theorem* is discussed, which gives the asymptotical distribution of the excesses over high thresholds. This section follows chapters 3 and 4 in Coles, 2001.

Let $(X_1, ..., X_n)$ be a sequence of $n$ independently and identically distributed random variables from a common but unknown distribution $F$ and let $M_n = \max\{X_1, X_2, ..., X_n\}$. For each sample size $n$, the distribution of the maximum, $M_n$, is given by

$$\begin{aligned} P(M_n \leq z) &= P(X_1 \leq z, X_2 \leq z, ..., X_n \leq z) \\ &= P(X_1 \leq z)P(X_2 \leq z)...P(X_n \leq z) \\ &= [P(X_1 \leq z)]^n \\ &= F^n(z) \end{aligned}$$

The Fisher–Tippett–Gnedenko theorem states that if there exist sequences of normalizing constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$P\left(\frac{M_n - a_n}{b_n} \leq z\right) = F^n(a_n z + b_n) \overset{n\to\infty}{\longrightarrow} G(z),$$

for all $z \in \mathbb{R}$, where $G$ is non-degenerate distribution then $G$ must be either Fréchet, Gumbel or negative Weibull. It was shown by Jenkinson, 1955 that these three distributions can be represented as a single family of distributions, known since then as Generalized Extreme Value distribution. Its distribution function is given by

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

and is defined on $\{z : 1 + \xi(\frac{z-\mu}{\sigma}) > 0\}$, where $\mu \in \mathbb{R}$, $\xi \in \mathbb{R}$ and $\sigma > 0$. The parameter $\xi$ governs the type of the distribution: if $\xi > 0$ it reduces to the Fréchet, if $\xi < 0$ the negative Weibull is obtained and in case of $\xi \to 0$ the Gumbel is recovered. This theory is used in practice in the following

way: the data is split into $m$ blocks that correspond to some specified time interval, say, 1 year. Then in each block a maximum is found. Finally, the Extreme Value Distribution is fit to all $m$ maxima.

Since the modelling strategy pursued in this thesis focuses on making use of as many individual losses as possible and modelling of the tail of the severity distribution, working only with maxima could be considered as counter-productive approach. Though the main goal of introducing of the Fisher–Tippett–Gnedenko theorem, Extreme Value Distribution and block maxima method was to build theoretical foundations upon which the next important important result - the Pickands–Balkema–de Haan theorem was historically built. The latter states that the conditional distribution of exceedances, given that the random variable is greater than the threshold, is asymptotically distributed according to the Generalized Pareto Distribution (GDP). This gives rise to the so called peaks over threshold method (POT), in which the data is used more efficiently by taking all exceedances over a high threshold in consideration.

More formally, let $u > 0$ be a threshold and $X$ random variable with a distribution $F$. Let $y = X - u$ denote an excess loss. The conditional excess distribution function $F_u(y)$ is defined as follows

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

Furthermore, let $(X_1, X_2, ..., X_n)$ be a random sample from a common but unknown distribution $F$ for which the Fisher–Tippett–Gnedenko theorem holds and denote by $X$ an arbitrary $X_i$ in the sequence and $F_u(X)$ is its conditional excess function. Then the Pickands–Balkema–de Haan theorem states that for sufficiently high value of $u$

$$F_u(y) = P(X > u + y | X > u) \longrightarrow H(y),$$

where $H(y)$ is a Generalized Pareto Distribution given by

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

and is defined on the set $\{y : y > 0 \text{ and } (1 + \xi y/\tilde{\sigma}) > 0\}$, where $\tilde{\sigma} = \sigma + \xi(u - \mu)$. The quantities $\xi, \mu$ and $\sigma$ are shape, location and scale parameters, respectively. When $\xi \to 0$ then the GPD simplifies to the exponential distribution.

The mean Generalized Pareto distribution, $Y \sim GDP(\xi, \mu, \sigma)$, is defined

only for $\xi < 1$ and variance takes on finite values only for $\xi < \frac{1}{2}$. These first two central moments are given by:

$$\mathbb{E}(Y) = \mu + \frac{\sigma}{1-\xi}$$

$$Var(Y) = \frac{\sigma^2}{(1-2\xi)(1-\xi)^2}.$$

Block maxima justified by the first theorem and peaks over thresholds based on the second theorem are closely related to the mathematical point of view - if there is an approximate distribution in the former method, then in the latter method there is a corresponding approximate generalized Pareto distribution. The choice of a length of a block (1 day, 1 month, 1 year, etc.) corresponds to the difficulty of selecting a threshold value in the POT method.

Threshold selection constitutes a challenge, as there is no universally applicable methodology that would provide a correct and unique answer in every instance of the problem. It amounts to navigating through the bias-variance dilemma - the threshold has to be chosen in a way that is sufficiently high so that the asymptotical argument approximately holds and it should be small enough so that there is a substantial number of points in the tail allowing a higher degree of precision when estimating parameters.

The *mean excess plot* is useful and a popular method that helps selecting the threshold. It is based on the theoretical mean and is applicable as long as the parameter $\xi < 1$ otherwise the mean takes on infinite value. This method relies on the visual comparison of thresholds and corresponding estimates of the conditional mean of excesses. The mean excess function for the Generalized Pareto Distribution is given by

$$e(u) = \mathbb{E}(X - u | X > u) = \frac{\sigma + \xi u}{1 - \xi}$$

and it is notably linear in the variable $u$. The corresponding empirical estimate can be calculated with

$$\hat{e}(u) = \frac{\Sigma_{i=1}^{n} \max(X_i - u, 0)}{\sum_{i=1}^{n} 1(X_i > u)},$$

where $1(X > u)$ is an indicator function. If the model is valid, the empirical estimates $\hat{e}(u)$ should also increase linearly as the value of $u$ grows. This assumption can be checked by inspecting at the mean excess plot given below:

$$\{(u, \hat{e}(u)) : u < u_{\max}\}.$$

The value after which the graph is approximately linear can be regarded as a good guess for the threshold.

Another graphical method that facilitates selecting threshold relies on estimating the model to many different thresholds and checking if the parameter estimates are stable. The graph that depicts thresholds on the $x$-axis and corresponding estimates of the scale parameter $\sigma$ on the $y$-axis should be approximately linear and a similar graph of the shape parameter $\xi$ should be roughly constant. Ultimately, the tail can also be specified by selecting a threshold according to a specific quantile, for instance, 0.9 quantile.

The theory of extreme values justifies the choice of the Generalized Pareto Distribution for modelling of the tail by an asymptotical argument. In many applications, especially in the financial world, the quantiles and more specifically, the value of risk, are of main interest. Makarov, 2007 points out that for the Generalized Pareto Distribution the convergence of quantiles is not guaranteed and hence it may not always be the best choice from a practical point of view, even though supported by the theoretical foundations. Therefore, it may be reasonable to apply other heavy-tailed distribution introduced the following subsection to the tail at a cost of increased bias. The issue of model selection is addressed in the section 6.

### 3.1.2 Inventory of parametric distributions

This subsection provides a small inventory of non-negative parametric models that can be used as benchmarks for the Generalized Pareto Distribution fitted to the peaks over the high threshold. In some cases, they can possibly be used as an alternative. The list contains parametric families with at most two parameters and is not exhaustive. Included distributions vary in their shapes and behaviour at large values. An excellent discussion of the classification based on the tail behaviour can be found in Panjer, 2006, Section 4.6.

#### 3.1.2.1 Single parameter distributions:

**Exponential distribution**

$$
\begin{aligned}
f(x) &= \frac{1}{\lambda} e^{-\frac{x}{\lambda}}, \quad x \geq 0, \lambda \geq 0 \\
F(x) &= 1 - e^{-\frac{x}{\lambda}} \\
\mathbb{E}(X) &= \lambda \\
Var(X) &= \frac{1}{\lambda^2} \\
\text{Mode} &= 0
\end{aligned}
$$

**Single pareto distribution**

$$
\begin{aligned}
f(x) &= \frac{\alpha \theta^{\alpha}}{x^{\alpha+1}}, \quad x \geq \theta, \theta > 0, \alpha > 0 \\
F(x) &= 1 - \left(\frac{\theta}{x}\right)^{\alpha} \\
\mathbb{E}(X) &= \frac{\alpha\theta}{\alpha - 1}, \quad \alpha > 1 \\
Var(X) &= \frac{\theta^2 \alpha}{(\alpha-1)^2(\alpha-2)}, \quad \alpha > 2 \\
\text{Mode} &= \theta
\end{aligned}
$$

#### 3.1.2.2 Two parameter distributions:

**Gamma distribution**

$$
\begin{aligned}
f(x) &= \frac{(x/\theta)^{\alpha} e^{-x/\theta}}{x\Gamma(\alpha)}, \quad x > 0, \alpha > 0, \theta > 0 \\
F(x) &= \Gamma(\alpha, x/\theta) \\
\mathbb{E}(X) &= \alpha\theta \\
Var(X) &= \alpha\theta^2 \\
\text{Mode} &= (\alpha - 1)\theta, \ \ \alpha \geq 1
\end{aligned}
$$

**Inverse gamma distribution**

$$
\begin{aligned}
f(x) &= \frac{(\theta/x)^{\alpha} e^{-\theta/x}}{x\Gamma(\alpha)}, \quad x > 0,\, \alpha > 0,\, \theta > 0 \\
F(x) &= 1 - \Gamma(\alpha, x/\theta) \\
\mathbb{E}(X) &= \frac{\theta}{\theta - 1}, \;\; \alpha > 1 \\
Var(X) &= \frac{\theta^2}{(\alpha-1)^2(\alpha-2)}, \;\; \alpha > 2 \\
\text{Mode} &= \frac{\theta}{\alpha + 1}
\end{aligned}
$$

**Weibull distribution**

$$
\begin{aligned}
f(x) &= \frac{\tau}{\lambda}\left(\frac{x}{\lambda}\right)^{\tau-1} \exp\left(-\left(\frac{x}{\lambda}\right)^{\tau}\right), \quad x \geq 0,\, \lambda > 0,\, \tau > 0 \\
F(x) &= 1 - \exp\left(-\left(\frac{x}{\lambda}\right)^{\tau}\right) \\
\mathbb{E}(X) &= \lambda\Gamma(1 + 1/\tau) \\
Var(X) &= \lambda^2\left(\Gamma(1 + 2/\tau) - \Gamma(1 + 1/\tau)^2\right) \\
\text{Mode} &= \lambda\left(\frac{\tau - 1}{\tau}\right)^{1/\tau}, \quad \tau > 1 \;\; (\text{ mode = 0 for } \tau \leq 1)
\end{aligned}
$$

**Log-normal distribution**

$$
\begin{aligned}
f(x) &= \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right), \quad x > 0,\, \mu \in \mathbb{R},\, \sigma > 0 \\
F(x) &= \Phi\left(\frac{\ln x - \mu}{\sigma}\right), \text{ where } \Phi \text{ is the CDF of the standard normal distribution} \\
\mathbb{E}(X) &= \exp(\mu + \sigma^2/2) \\
Var(X) &= (\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2) \\
\text{Mode} &= \exp(\mu - \sigma^2)
\end{aligned}
$$

**Loglogistic distribution**

$$
\begin{aligned}
f(x) &= \frac{(\beta/\alpha)(x/\alpha)^{\beta-1}}{(1+(x/\alpha)^{\beta})^2}, \quad x \geq 0,\, \alpha > 0,\, \beta > 0 \\
F(x) &= \frac{1}{1+(x/\alpha)^{-\beta}} \\
\mathbb{E}(X) &= \frac{\alpha\pi/\beta}{\sin(\pi/\beta)},\ \beta > 1 \\
Var(X) &= \alpha^2\left(2b/\sin(2b) - b^2/\sin^2(b)\right),\ b := \pi/\beta,\ \beta > 2 \\
\text{Mode} &= \alpha\left(\frac{\beta-1}{\beta+1}\right),\ \beta > 1, \text{ otherwise undefined}
\end{aligned}
$$

**Inverse Gaussian distribution**

$$
\begin{aligned}
f(x) &= \left(\frac{\lambda}{2\pi x^3}\right)^{1/2}\exp\left(-\frac{\lambda}{2x}\left(\frac{x-\mu}{\mu}\right)^2\right), \quad x \in (0,\infty),\, \lambda > 0,\, \mu \in > 0 \\
F(x) &= \Phi\left(\frac{x-\mu}{\mu}\left(\frac{\lambda}{x}\right)^{1/2}\right) + \exp\left(\frac{2\lambda}{\mu}\right)\Phi\left(-\frac{x+\mu}{\mu}\left(\frac{\lambda}{x}\right)^{1/2}\right) \\
\text{Mode} &= \mu\left(\left(1+\frac{9\mu}{4\lambda^2}\right)^{1/2} - \frac{3\mu}{2\lambda}\right) \\
\mathbb{E}(X) &= \mu \\
Var(X) &= \frac{\mu^3}{\lambda}
\end{aligned}
$$

## 3.2 Modelling of the body

The body of the severity distribution contains all events that fall below a certain threshold. Due to the LFHS property of the underlying data, there is usually a very large number of points in the body and the shape of the data generating process can be modelled with an empirical distribution without imposing any distributional assumptions. The *Glivenko–Cantelli theorem*, which is described in the following, provides a justification for the appropriateness of the empirical distribution in the body.

Let $\{X_1, X_2, ...\}$ be a sequence of independent and identically distributed random variables with values in $\mathbb{R}$ and with common cumulative distribution function $F(x)$. If $\hat{F}_n(x)$ denotes the empirical distribution function defined as

$$\hat{F}_n(x) = \sum_{i=1}^{n} 1(X_i \leq x),$$

where $1(A)$ is an indicator funtion on a set $A$, then

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{a.s} 0.$$

Less formally, the theorem states that the empirical distribution $\hat{F}_n(x)$ uniformly converges the true distribution as the sample size gets larger.

# 4 Modelling of Frequency

This section focuses on the modelling of the number of events with a counting distribution. The counting distribution, also named frequency distribution, is a discrete probability function that assigns probabilities to non-negative integers. Developing a counting distribution helps to understand the process that governs the occurrence of events and enables predicting the number of events that may happen in future. Parametric models are usually chosen as they allow interpolating the data within the range of observed counts and also extrapolating beyond the historical maximum number of losses. In other words, parametric models allow imposing a specific functional form and assigning probabilities to arbitrary range of non-negative integers accordingly.

There is a plethora of possible models, however, only few of them have useful properties and are appropriate for a given modelling problem. In general, it is practical that a counting distribution is *infinitely divisible*. It means that if $X$ is a random variable that follows some distribution $F$, it can be represented as a sum of $n \in \{1, 2, 3, ...\}$ independently and identically distributed random variables $\{Y_1, Y_2, ..., Y_n\}$ and this sum also follows the distribution $F$. A model with such property does not depend on the length of the considered time period and the expected counts are proportional to the length of the time interval.

Throughout this section the discrete random variable $N$ represents the number of events, $p_k = P(N = k)$ denotes the probability that $k \in \{0, 1, 2, 3, ...\}$ events occur and the corresponding probability generating function is defined as

$$P_N(z) = \mathbb{E}(z^N) = \sum_{k=0}^{\infty} p_k z^k.$$

The later is a power series representation of a probability mass function and, just like a moment generating function, can be used to obtain moments. In the following, the most popular counting distributions are introduced and subsequent subsections are in big part based on Klugman, Panjer, and Willmot, 1998. Any method from the section 5 can be used to obtain estimates of the parameters.

## 4.1 Poisson distribution

Poisson distribution is a counting distribution that is widely used to model processes with a big number of possible events within a time interval but each of them has a very small probability of occurrence. The probability mass function of the Poisson distribution is given by

$$p_k = P(N = k) = \frac{e^{-\lambda}\lambda^k}{k!},$$

where $k \in \{0, 1, 2, 3, ...\}$ and $\lambda > 0$. The corresponding mean and variance are

$$\mathbb{E}(N) = Var(N) = \lambda.$$

Interestingly, both first and second central moments are equal. This property suggests that if the empirical mean and variance are more or less equal, then the Poisson distribution is an appropriate choice, provided the model is reasonable from the theoretical point of view. In order to obtain higher moments, the probability generating function given below can be used.

$$P_N(z) = e^{\lambda(z-1)}, \text{ for } \lambda > 0$$

Another useful property is related to the fact that Poisson distribution is infinitely divisible. The sum of $n$ Poisson distributed random variables follows again Poisson distribution. More rigorously, if $\{N_1, N_2, ..., N_n\}$ are independent Poisson random variables with parameters $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ then $N = N_1 + N_2 + ... + N_n$ follows a Poisson distribution with the parameter $\lambda = \lambda_1 + \lambda_2 + ... + \lambda_n$. The proof involves applying a probability generating function and can be found in section 3.2.1 in Klugman, Panjer, and Willmot, 1998. Another useful feature is that if the number of events within a specific time window follows a Poisson distribution and these events can be divided into $n$ independent categories, then the distribution in each category is independent and Poisson distributed with the scaled parameter $\lambda$. It is especially practical if, for instance, we know that counts follow a Poisson distribution but only events that exceed a certain threshold are of particular interest.

## 4.2 Negative binomial distribution

The negative binomial distribution is another popular counting distribution that is sometimes used as the alternative to the Poisson distribution. Its prob-

ability mass function is defined via

$$p_k = \binom{k+r-1}{k} \left(\frac{1}{1+\beta}\right)^r \left(\frac{\beta}{1+\beta}\right)^k,$$

where $k \in \{0, 1, 2, 3, ...\}$, $r > 0$ and $\beta > 0$. The binomial coefficient can be calculated with a well known formula

$$\binom{n}{k} = \frac{n(n-1)...(n-k+1)}{k!}.$$

It can also be expressed in terms of the gamma function $\Gamma(n)$ as follows:

$$\binom{n}{k} = \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)}$$

Here $k$ is an integer and $n$ is a real number for which $n > k-1$ holds. The mean and variance are given by

$$\mathbb{E}(N) = r\beta$$

$$Var(N) = r\beta(1+\beta)$$

Since $\beta > 0$, it can be seen that the variance is bigger then the expected value. It suggests that if the empirical variance is bigger than the empirical mean then the negative binomial distribution might be more suitable than Poisson. Higher moments can be calculated with the probability generating function that is defined as

$$P_N(z) = (1 - \beta(z-1))^{-r}.$$

The negative binomial distribution can be viewed as a generalized version of the Poisson distribution, since it can be represented as a mixture of a family of Poission distributions with gamma mixing distribution. It also generalizes the geometric distribution as it is a sum of independent and identically distributed geometric random variables. In particular, if $r = 1$, it simplifies to a geometric distribution. Interestingly, when the value of the latter parameter is smaller/bigger than 1, then the distribution has lighter/heavier tail compared to the geometric distribution.

The negative binomial distribution is also infinitely divisible and the independent sum of $n$ negative binomial random variables with the same parameter $\beta$ and different $r_i$s follows negative binomial distribution with pa-

rameters $\beta$ and $r = r_1 + r_2 + ... + r_n$.

## 4.3 Binomial distribution

The binomial distribution can also be used as a counting distribution and presents an interesting alternative to the Poisson and the negative binomial since it is defined on a finite non-negative support. It can be useful if it is known that the number of events cannot exceed some threshold or if it is reasonable to put restrictions to a maximal number of counts. Its probability mass function is defined as

$$p_k = P(N = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where $n$ represents a finite number of possible losses, $k \in \{0, 1, ..., n\}$ is the number of losses and $p \in [0, 1]$ is the probability the occurrence. The mean and variance are given by

$$\mathbb{E}(N) = np$$

$$Var(N) = np(1-p).$$

Since $p \in [0, 1]$, it can easily be seen that the variance is always smaller or equal to the mean. Hence, if the empirical variance is considerably smaller than the empirical mean, the binomial distribution might be considered. The probability generating function is given as follows

$$P_N(z) = (1 - p + pz))^n.$$

## 4.4 The (a,b,0) class

The Poisson, negative binomial and binomial described in previous subsections belong to the more general class of distributions known as $(a, b, 0)$ class. Each member of this class is characterized by the fact that it can be represented via following recursion:

$$\frac{p_k}{p_{k-1}} = a + \frac{b}{k},$$

where $k \in \{1, 2, 3, ...\}$, $p_k = P(N = k)$ and $a, b$ are real valued constants. This recursion is satisfied only for Poisson, negative binomial and binomial

distributions and the corresponding constants as well as the probability at zero, $p_0$, can be found in the table 1.

| Distribution | $p_k$ | $a$ | $b$ | $p_0$ |
|---|---|---|---|---|
| Poisson | $\frac{e^{-\lambda}\lambda^k}{k!}$ | $0$ | $\lambda$ | $e^{-\lambda}$ |
| Negative Binomial | $\binom{k+r-1}{k}\left(\frac{1}{1+\beta}\right)^r\left(\frac{\beta}{1+\beta}\right)^k$ | $\frac{\beta}{1+\beta}$ | $(r-1)\frac{\beta}{1+\beta}$ | $(1+\beta)^{-r}$ |
| Binomial | $\binom{n}{k}p^k(1-p)^{n-k}$ | $-\frac{p}{1-p}$ | $(n+1)\frac{p}{1-p}$ | $(1-p)^n$ |

TABLE 1: Table with members of the $(a, b, 0)$ class.

There is also a more broader class of counting distributions, $(a, b, 1)$, however, it is beyond of the scope of this thesis. See section 3.5.1 and 3.6.1 in Klugman, Panjer, and Willmot, 1998 for further details on $(a, b, 0)$ and $(a, b, 1)$ classes. They are of practical importance since some software used in actuarial sciences define a counting distribution in terms of the previously mentioned classes.

# 5 Parametric Estimation

This section introduces five parametric estimation methods that can be used to find optimal values of parameters that belong to distributions described in the section 3. Some assumptions and simplifications have to be made to make the estimation process tractable, while still keeping it realistic. Throughout this section, it will be assumed that there are $n$ observations available. These data points are a realisation of a sequence of real-valued random variables $(X_1, ..., X_n)$ (random sample) that are independent and follow the same probability distribution described by $\theta$. The parameter vector $\theta$ belongs to a parameter space $\Theta$ which, in turn, is contained in the $p$-dimensional Euclidean space ($\theta \in \Theta \subseteq \mathbb{R}^p$). The functional form of the model is known, however, the parameter vector $\theta$ is unknown.

Given this statistical set-up and observations $(X_1, ..., X_n)$ the goal is to estimate $\theta$. In other words, we look for an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$ that depends only on data and is as close to $\theta$ as possible. While doing so, it is possible to obtain estimates that coincide with the true values by chance. Such scenario is very desirable, however very unlikely. In almost every case, we have to take into account some error and it is important to have tools that allow us to asses the quality of an estimator in order to make a rational choice.

In general, Klugman, Panjer, and Gordon, 1988 distinguish between four different types of errors. The first error may occur when we collect data over a longer period of time and the characteristics of the population change. In such case, we draw conclusions about the population based on the sample from another population (*frame error*). It can also happen that our assumptions about the data generating process or selected model do not reflect the reality (*model error*). Klugman, Panjer, and Gordon, 1988 point out that these two errors are not related to properties of the estimators and cannot be measured. Furthermore they discuss the *estimation error* that can arise when the obtained sample is not representative of the population by chance and/or if the estimation method has some shortcomings. Quantifying this kind of error may give us an indication of the quality of the estimator and select an appropriate one for a given problem.

The first very useful and popular measure of the quality of an estimator $\hat{\theta}$ is related to its average behaviour. More specifically, it is the difference between the expected value of the estimator $\hat{\theta}$ and the true value of $\theta$. It is known as a *bias* and this concept was first introduced by David and Neyman,

1938 in the context of point estimators.

$$bias_\theta(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta} - \theta)$$

If the *bias* is equal to zero for all possible values of $\theta$ then the estimator is called to be *unbiased*. In such case, the sampling distribution of the estimator is concentrated around the true value of the parameter and, it gives us the true value on average. More formally, an estimator is *unbiased* if

$$\forall\theta \in \Theta : \mathbb{E}_\theta(\hat{\theta}) = \theta$$

There may be an estimator that is biased for a finite sample. However, with an increasing number of observations $n$, its tendency to overestimate or underestimate the true parameter becomes smaller and smaller to eventually vanish with an infinite number of observations. If such scenario is true for all possible values of $\theta$ then such estimator is called to be *asymptotically unbiased*. Formally, an estimator is *asymptotically unbiased* if

$$\forall\theta \in \Theta : \lim_{n\to\infty} \mathbb{E}_\theta(\hat{\theta}_n - \theta) = 0$$

Another desired property of an estimator is known as consistency, i.e. with and increasing sample size the sampling distribution of $\hat{\theta}$ becomes more and more concentrated around the true value of $\theta$. Putting it into mathematical terms, an estimator is consistent if

$$\forall\epsilon > 0, \forall\theta \in \Theta : \lim_{n\to\infty} \mathbb{P}(\|\hat{\theta}_n - \theta\| > \epsilon) = 0$$

Consistency can be related to the bias as it suffices for an estimator to be consistent, if it is asymptotically unbiased and its variance goes to zero with an increasing sample size. One example of the aforementioned relationship would be to estimate the population average by the sample mean penalized by the reciprocal of the sample size $n$: $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}$.

It is also possible that an estimator is unbiased but not consistent. A simple example of such case is to estimate the true mean by the last observation in the sample $X_n = x_n$ instead of the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ of $n$ independent and identically distributed observations $(X_1 = x_1, ..., X_n = x_n)$. The estimator $\hat{\theta} = x_n$ is unbiased, as all values in the sample share the same expectation. However, it does not converge to any point, as it jumps to the other value whenever $X_n = x_n$ changes.

In general, it is desirable that an estimator is unbiased and converges to

the true value of the parameter at the fastest possible rate. However, depending on the context, it may be preferable to choose an estimation method that trades these properties for other desired qualities. For instance, in the face of very rare but severe events in the dataset, one could consider taking into account slightly higher variance for robustness. It is also advantageous that the sampling distribution is (asymptotically) normally distributed, which in turn allows to test statistical hypotheses and quantify uncertainty around the parameter of interest.

## 5.1 Method of moments

The first popular estimation method is called "Method of moments". Its name comes from the very idea of setting up a system of equations that match empirical moments to theoretical moments. The number of equations has to be equal to the number of parameters and solving for the latter yields *method of moments estimator*. This simple, yet powerful, method was first introduced by a well known Russian mathematician Pafnuty Chebyshev. Following the statistical set-up, there are $n$ independent and identically distributed (i.i.d.) random variables $(X_1, ..., X_n)$ that follow some probability distribution described by $\theta \in \Theta \subseteq \mathbb{R}^p$.

The $k$-th theoretical moment of a probability distribution is given by

$$m_k(\theta) = \mathbb{E}_\theta(X^k), \quad \text{for } k \geq 1$$

and the $k$-th empirical moment can be obtained by taking a sample average of samples raised to the $k$-th power

$$\hat{m}_k(\theta) = \frac{1}{n} \sum_{i=1}^{n} X_i^k, \quad \text{for } k \geq 1$$

The $p$-dimensional parameter $\theta \in \Theta$ can be represented by a function of the first $p$ parameters. It must hold that the function $g : \mathbb{R}^p \to \mathbb{R}^p$ is bijective.

$$\theta = g(m_1(\theta), ..., m_p(\theta))$$

The method of moments estimator $\hat{\theta}_{MM}$ can be obtained in following three steps:

1. solve $g(m_1(\theta), ..., m_p(\theta))$ for $\theta$
2. for $k = 1, 2, ..., p$ estimate $m_k(\theta)$ by $\hat{m}_k(\theta)$

3. set $\hat{\theta}_{MM} = g(\hat{m}_1(\theta), \hat{m}_2(\theta), ..., \hat{m}_p(\theta))$

The biggest advantage of the method of moments estimator is that it is easy to obtain, and under some mild conditions it is consistent and asymptotically normal (Vaart, 1998). However, this method may not be robust against outliers, as it matches empirical moments to theoretical moments and the former may be easily affected by extreme values. Moment estimator is also not necessarily unbiased.

## 5.2 Maximum likelihood estimation

Method of maximum likelihood is one of the most famous statistical estimation procedures. It was introduced by R. A. Fisher in 1922. Because of the popularity and general usefulness of this method, it may be beneficiary to immerse into the long history of its development described in detail in (Aldrich, 1997). The idea behind the maximum likelihood estimation is very intuitive – it looks for a value of $\theta$ that makes the observed data most likely and chooses it as an *maximum likelihood estimator* denoted by $\hat{\theta}_{ML}$.

More formally, let $(X_1, ..., X_n)$ be a random sample from a distribution $f_\theta$ that depends on the parameter $\theta \in \Theta \subseteq \mathbb{R}^p$. For convenience, $f_\theta$ represents either continuous probability density function (p.d.f) or a discrete probability mass function (p.m.f.). The function

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$$

is called *likelihood function* and assigns probability to a particular realisation $(x_1, ..., x_n)$ of $(X_1, ..., X_n)$. Maximizing the likelihood function $\mathcal{L}(\theta)$ with respect to $\theta$ yields the maximum likelihood estimator $\hat{\theta}_{ML}$

In practice, it is more convenient and easier to optimize so called *log-likelihood function* $l(\theta)$, which is the natural logarithm of the original likelihood function.

$$l(\theta) = ln\mathcal{L}(\theta)$$

As $x \to ln(x)$ is a positive monotonic transformation, it preserves the properties of the original *likelihood function* and maximizing $l(\theta)$ is equivalent to maximizing $\mathcal{L}(\theta)$. So the maximum likelihood estimator is given by

$$\hat{\theta}_{ML} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}(\theta) = \underset{\theta \in \Theta}{\operatorname{argmax}}\, l(\theta)$$

Optimizing a (log-)likelihood function may not be an easy task. It may have multiple local maxima in addition to the global optimum. Also, the maximum may not be unique and may not even exist. Furthermore applying classical calculus tools may not be possible and one must resort to the numerical methods.

The popularity of this method lies in its large sample properties. The maximum likelihood estimators are asymptotically unbiased, consistent and have the smallest asymptotic variance among all consistent estimators. In addition, they are invariant under parameter transformation. However, they are not necessarily robust against outliers. It is also notable that the maximum likelihood estimators in an exponential family coincide with method of moments estimators (Vaart, 1998).

## 5.3 Minimum distance estimation

The *minimum distance estimation* is a parametric technique that was developed in the 1950s by Wolfowitz, 1957. It is not based on the likelihood principle, nor does it try to match some characteristics of a distribution like the method of moments does. Instead, a distance between a theoretical and an empirical distribution is minimized according to some criterion function.

The following formal definition of the minimum distance estimator is based on Drossos and Philippou, 1980. Let $(X_1, ..., X_n)$ be an i.i.d. random sample from population with a distribution function $F(x;\theta)$ that depends on some parameter $\theta \in \Theta \subseteq \mathbb{R}^p$. Let $F_n(x)$ be an empirical distribution function based upon $(X_1, ..., X_n)$ and let $d(\cdot,\cdot)$ be some criterion function that measures discrepancy between both arguments. If there exists $\hat{\theta} \in \Theta$, such that

$$d_\psi(F_n(x), F(x;\hat{\theta})) = \inf\{d_\psi(F_n(x), F(x;\theta)) : \theta \in \Theta\}$$

it is called a minimum distance estimator $\hat{\theta}_{mde}$ of $\theta$.

There are multiple choices for the discrepancy measure $d(\cdot,\cdot)$ and are discussed in detail by Parr and Schucany, 1980. Two popular alternatives are *weighted Kolmogorov distance* and *Cramér-von Mises distance*. The former is the supremum over absolute differences and the latter is the integral of the squared differences. More formally, if $K$ and $L$ denote two distribution functions with a common support and $\psi(u)$ is a weighting function, then we have:

- *weighted Kolmogorov distance*

$$d_\psi(K, L) = \sup_{x \in \mathbb{R}} |K(x) - L(x)| \psi(L(x))$$

with a *Kolmogorov distance* as a special case for $\psi(u) \equiv 1$

- *Cramér-von Mises distance*

$$d_\psi(K, L) = \int_{-\infty}^{\infty} (K(x) - L(x))^2 \psi(L(x)) dL(x)$$

With three special cases depending on the weighting function:

  - the *Cramér-von Mises statistic* for $\psi(u) \equiv 1$
  - the *Anderson-Darling statistic* for $\psi(u) = \frac{1}{u(1-u)}$
  - the *Upper tail Anderson-Darling statistic* for $\psi(u) = \frac{1}{1-u}$

The minimum distance estimator enjoys desired properties of asymptotic unbiasedness and consistency (Wolfowitz, 1957). Just like maximum likelihood estimators, they are invariant under parameter transformation, however, they are less accurate (Drossos and Philippou, 1980). In exchange, they offer excellent robustness properties (Millar, 1981). In addition, Cramér-von Mises type minimum distance estimators are asymptotically normal (Parr and Schucany, 1980).

## 5.4 Bayesian estimation

All estimation methods introduced belong to the *frequentist* approach, in which the unknown parameter $\theta \in \Theta$ is assumed to be fixed. Hence no probabilistic statements about the uncertainty of $\theta$ are made. The *Bayesian* inference is fundamentally different, as it assumes that $\theta$ is random and has its own probability structure. The initial knowledge about the $\theta$ is reflected in some prior probability distribution. After observing some data, initial beliefs are updated wtih Bayes' theorem. This leads to a posterior distribution that provides full probabilistic description of the updated beliefs about the parameter of interest.

More formally, let $X_{1:n} = (X_1, ..., X_n)$ be a random sample of size $n$ from a distribution described by a parameter $\theta \in \Theta \subseteq \mathbb{R}^p$. Let us define following quantities:

- $p(X_{1:n}|\theta)$ is the likelihood of data given under the parameter of interest $\theta$
- $\pi(\theta)$ is the prior distribution on $\theta$
- $p(X_{1:n}) = \int_\Theta p(X_{1:n}|\theta)\pi(\theta)d\theta$ is the marginal likelihood of data.

The Bayes' theorem states that the posterior probability $p(\theta|X_{1:n})$ is given by

$$p(\theta|X_{1:n}) = \frac{p(X_{1:n}|\theta)\pi(\theta)}{p(X_{1:n})}$$

The initial knowledge of the parameter of interest is incorporated into the estimation process via prior distributions. They do not have to be proper distributions, as long as the resulting posterior distribution integrates to 1 though. However, the choice of a prior can influence the inference and adds a source of subjectivity into it. They are usually chosen based on the past experience, expertise of a subject-matter-expert or common sense.

A popular choice for the prior is a distribution from a family of *conjugated priors*. Such prior, multiplied by the likelihood, leads to the posterior distribution from the same family. Although, they may not fully reflect the prior information, they are computationally convenient and make the calculation of the posterior easy. Conjugated priors were introduced by Raïffa and Schlaifer, 1961. Selecting other priors may lead to computational difficulties as it may not be possible to compute the normalizing constant $p(X_{1:n})$ analytically. In such cases, Markov Chain Monte Carlo (MCMC) techniques are usually applied to obtain the posterior. If there is little or no prior information available, an objective prior, like the Jeffrey's prior may be used.

The Bayesian estimator $\hat{\theta}_B$ is usually chosen to be the mean of the posterior distribution, which minimizes the squared error risk.

$$\hat{\theta}_B = \mathbb{E}(\theta|X_{1:n}) = \int_\Theta \theta\, p(\theta|X_{1:n})\, d\theta$$

It is asymptotically unbiased, consistent and asymptotically normal. It coincides with the maximum likelihood estimator in the limit as $n \to \infty$.

## 5.5 Quantifying the quality of point estimates with standard errors

All described techniques in previous sections have desirable large sample properties, with maximum likelihood estimator and Bayes estimator being

the most accurate estimators in the asymptotic sense. However, in the real world applications no infinite datasets are feasible and often one has to face scarcity of data, where only few data points are available. It is crucial to quantify the uncertainty about the point estimate to see how reliable a method is and if there are better alternatives. It is not impossible that the theoretically most efficient method would lead in some case to a higher uncertainty than other less efficient techniques.

One way of assessing the quality of a point estimate is to compute a frequentist standard error. Obtaining this measure of variability is analytically not possible for all described methods and we have to resort to simulation. *Bootstrap*, invented by Efron, 1979, is a powerful tool that can be used to do so. The idea is to draw $B$ independent samples with replacement from the original sample. Then for each of the bootstrap samples, we estimate parameters of interest and store them. Finally, the standard deviation is computed based on $B$ estimates, which represents the standard error of the initial parameter estimate.

More formally, if $(X_1, ..., X_n)$ is a random sample from the population with the distribution $F$ and $\hat{\theta}_n$ is a point estimate, then the standard error of $\hat{\theta}_n$ denoted by $se(\hat{\theta}_n)$ can be obtained as follows:

- For $j \in \{1, ..., B\}$, draw a sample with replacement $(X_1^{*j}, ..., X_n^{*j})$ from the original sample $(X_1, ..., X_n)$
- For $j \in \{1, ..., B\}$, compute $\hat{\theta}_n^{*j}$ based on $(X_1^{*j}, ..., X_n^{*j})$
- Compute $se(\hat{\theta}_n) = \sqrt{\frac{1}{B}\sum_{j=1}^{B}(\hat{\theta}_n^{*j} - \bar{\theta}_n^*)^2}$ , where $\bar{\theta}_n^* = \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_n^{*j}$

# 6 Model Selection

There is a plethora of data-driven statistical models that can be used to describe a data generating process of interest and to make inference about its specific aspects. As there is no single model that is able to fully represent the underlying process, it is necessary to find the best approximating model that allows to draw conclusions that are as close to the truth as possible. A possible modelling approach is to first select a subset of appropriate models for a given problem based on theoretical considerations. This collection of models should be then narrowed down by an explanatory analysis, as well as by the graphical methods that compare the data with the considered models. Finally, formal goodness-of-fit tests should be performed and their results stored in a table, which summarizes how well each model performed in each test. The best fitting model could be then selected as the model with the highest score. In the following graphical methods and formal goodness-of-fit test are introduced and a possible way of assigning scores to each model is suggested.

## 6.1 Graphical methods

Assume that a sample $(X_1, ..., X_n)$ containing $n$ independent observations comes from a population with unknown distribution $F$. Some estimation method from the section 5 was used to obtain an estimate $\hat{F}$ of $F$ and the goal is to visually assess the quality of fit. There are two popular methods that rely either on comparing empirical distribution function with a theoretical distribution function or quantiles of a sample with quantiles of a theoretical distribution function. Prior that order statistics $X_{(1)}, ..., X_{(n)}$ is created so that $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ and the empirical distribution function $F_n(x) = \frac{j}{n}$ for $x_{(j)} \leq x \leq x_{(j+1)}$ is adjusted to $F_n(x_{(j)}) = \frac{j}{n+1}$ for $x_{(j)} \leq x \leq x_{(j+1)}$ to avoid probability 1 for the maximum value in the sample.

The first method is so called probability-probability plot and can be obtained in following way

$$\left\{ \left( \hat{F}(x_{(i)}), \frac{i}{n+1} \right) : i = 1, 2, ..., n \right\}.$$

The other method is known as quantile-quantile plot. It conveys the same information as the method above, however, it is presented on a different scale.

$$\left\{ \left( \hat{F}^{-1} \left( \frac{i}{n+1} \right), x_{(i)} \right) : i = 1, 2, ..., n \right\}.$$

In both cases, any departure from a 45° indicates differences between empirical and theoretical distributions.

## 6.2 Goodness-of-fit tests

Given independent observations $(X_1, ..., X_n)$ from common population with unknown distribution, the objective is to test whether the observed data was generated by some specific distribution $F$. The null hypothesis $H_0$ usually assumes that the data comes from specified distribution and the alternative $H_1$ is its negation.

$H_0$: Data comes from $F$
$H_1$: Data does not come from $F$

The test statistic $T_n$ usually measures the discrepancy between the empirical distribution function $F_n$ and the theoretical distribution function $F$ or the difference between expected and observed counts. Under the assumption that the null hypothesis holds, the distribution of $T_n$ is specified. If $T_n$ exceeds some critical value at some specified significance level $\alpha \in (0, 1)$, the null hypothesis is rejected.

### 6.2.1 Kolmogorov-Smirnoff test

One of the most prominent goodness-of-fit test until now was developed by Kolmogorov and Smirnoff. It measures the biggest vertical difference between the empirical and the hypothesized cumulative distribution functions and relates it to probability of such discrepancy, given that the hypothesized distribution generated the data.

Let $(X_1, ..., X_n)$ be a sample of $n$ independent variables coming from the same distribution. Let denote the empirical distribution by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty, x]}(X_i)$$

where $1_{(-\infty, x]}(X_i)$ is an indicator function that takes on value 1 if $X_i \leq x$ and 0 otherwise. Assume $F_\theta$ to be the hypothesized distribution with some fixed parameter vector $\theta \in \mathbb{R}^k$. The null and alternative hypotheses are given by

$H_0 : (X_1, ..., X_n) \sim F_\theta$
$H_1 : (X_1, ..., X_n) \not\sim F_\theta$

The corresponding test statistic $K$ is obtained by computing the biggest absolute difference between $F_n(x)$ and $F(x)$ (Smirnov, 1948).

$$K = \sup_x |F_n(x) - F(x)|$$

If $F_\theta$ is continuous, then under the null hypothesis the distribution of $\sqrt{n}K$ converges the Kolmogorov distribution $L$ and it is independent of $F_\theta$ (Feller, 1948).

$$L(z) = P(K \leq z) = 1 - 2\sum_{k=1}^{\infty}(-1)^k = \frac{\sqrt{2\pi}}{z}\sum_{k=1}^{\infty} e^{-(2k-1)^2\pi^2/8z^2}.$$

### 6.2.2 Anderson-Darling test

Anderson-Darling test is another method of assessing whether a sample was generated by some specified distribution. Similar to Kolmogorov-Smirnoff test, a discrepancy between the empirical $F_n$ and theoretical $F_\theta$ distributions is of interest and it is measured by a weighted average of squared differences. This approach was developed by (Anderson and Darling, 1952).

The statistical set-up and hypotheses are identical as described in the Kolmogorov-Smirnoff test. The Anderson-Darling statistic is

$$A^2 = n\int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} f(x)dx$$

Compared to the Kolmogorov-Smirnoff statistic, which looks for the good fit in the middle of the distribution, the Anderson-Darling statistic focuses more on the tails. It is shown in Anderson and Darling, 1954 that this test statistic can be rewritten as

$$A^2 = -n - \frac{1}{n}\sum_{j=1}^{n}(2j-1)\left(\log(U_{(j)}) + \log(1 - U_{(n-j+1)})\right)$$

where the sample is ordered ($X_{(1)} < X_{(2)} < ... < X_{(n)}$) and $U_j = F_\theta(X_{(j)})$. Furthermore, the authors derived the limiting distribution, from which critical values for significance level of 5% and 1% can be obtained. They are given by 2.492 and 3.857, respectively.

### 6.2.3 Chi-square goodness-of-fit test

Chi-Square test is designed to assess whether an observed sample comes from a discrete distribution. This method is based on the idea of comparing the actual number of observations with the expected number for each category if the null hypothesis holds. After some appropriate discretization it can be also applied to continuous distributions with no major obstacles. This famous test was first developed by Karl Pearson at the beginning of the 20th century (Pearson, 1900).

The null hypothesis assumes that the data follows a specific distribution and is tested against an alternative that negates the null. The corresponding Chi-Square statistic is

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

where $O_i$ represents the observed number of observations from the sample falling into the $i$-th category and $E_i$ corresponds to the expected counts for the same category under the null hypothesis. The expected counts can be obtained with $E_i = np_i$, where $n$ is the sample size and $p_i$ is the probability of the $i$-th. If the hypothesized distribution $F$ is continuous, then some arbitrary discretization $m = d_0 < d_1 < ... < d_k = \infty$ with $k - 1$ points has to be considered and the probability $p_i$ is replaced by $p_i^* = F(d_i) - F(d_{i-1})$ for all groupings. Such procedure leads to the loss of information, so that the test is not consistent against the alternative anymore. This test statistic follows the Chi-Square distribution with $l - 1$ degrees of freedom, where $l$ is number of categories.

## 6.3 Goodness-of-fit tests for left truncated data

In this subsection modifications of goodness-of-fit statistics that account for the left truncation are considered. Such tests emerge in a natural way in many different contexts when the data is recorded if it exceeds some specified value. Also, simple hypothesis are replaced in favour of composite hypothesis. They aim to answer the question whether the empirical distribution obtained from a sample of observations belongs to a parametric family of distributions $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ or not. Such test problem is more relevant from a practical standpoint. In general, this section follows the work of Chernobai, Rachev, and Fabozzi, 2015.

More formally, the sample $(X_1, ..., X_n)$ is left-truncated at a threshold point

$H \in \mathbb{R}$ and $X_{(1)} < X_{(2)} < .... < X_{(n)}$ is the ordered statistics. The empirical distribution $F_n(x)$ of the sample is given by

$$F_n(x) = \begin{cases} 0 & x \leq x_{(1)} \\ \frac{j}{n} & x_{(j)} \leq x \leq x_{(j+1)},\ j = 1, 2, ..., n-1 \\ 1 & x \geq x_{(n)}. \end{cases}$$

A continuous truncated distribution $F_\theta$ is fitted to the sample and estimate $\hat{\theta}$ of the conditional parameter $\theta$ is obtained. The resulting truncated distribution $F^*_{\hat{\theta}}$ of the sample is given by

$$F^*_{\hat{\theta}}(x) = \begin{cases} \frac{F_{\hat{\theta}}(x) - F_{\hat{\theta}}(H)}{1 - F_{\hat{\theta}}(H)} & x \geq H \\ 0 & x < H \end{cases}$$

A composite hypothesis tests whether the empirical distribution is a member of some specific parametric family. As the particular shape of the distribution is not specified, the idea is to compare the best fitting member of the hypothesized parametric class to the empirical estimate

$H_0 : F_n \in F^*_{\hat{\theta}}$
$H_1 : F_n \notin F^*_{\hat{\theta}}$.

For such test problem the distribution under the null is not known and, hence, critical values and p-values are not available. In order to overcome this problem, one has to resort to re-sampling methods. One particular approach is to first estimate a test statistic $T$ based on the observed dataset (truncated at $H$) and then to follow steps 1-3 described below.

1. For $i \in \{1, 2, ..., I\}$:
   (a) Generate a sample $S_{i,n}$ of size $n$ from the fitted truncated distribution $F^*_{\hat{\theta}}$.
   (b) Fit truncated distribution to the sample $S_{i,n}$ and obtain conditional parameter $\hat{\theta}$.
   (c) Estimate a test statistic $T_i$ and store it in a list.
2. Obtain p-value as a fraction of times the sample statistics $T_i$ are bigger than the sample statistic $T$. In other words, p-value can be calculated as $\frac{1}{I} \sum_{i=1}^{I} 1(T_i > T)$.

3. Reject null hypothesis if the p-value is smaller than the significance level $\alpha \in (0,1)$

In order to simplify the notation, $z_H = F^*_{\hat{\theta}}(H)$ is going to denote a fitted truncated distribution evaluated at the truncation point and $z_j = F^*_{\hat{\theta}}(x_{(j)})$ is a shorthand for the fitted truncated cdf evaluated at $j$-th order statistic. All tests introduced in the following subsections can be performed using an R statistical software (R Core Team, 2017).

### 6.3.1 Kolmogorov-Smirnoff statistic

Kolmogorov-Smirnoff statistic was previously introduced in the subsection 6.2 for the simple hypothesis and full sample. Its modification for composite hypothesis and left-truncated data can be obtained as follows:

$$\begin{aligned} K^{+*} &= \sqrt{n} \sup_j \left\{ F_n(x_{(j)}) - F^*_{\hat{\theta}}(x_{(j)}) \right\} \\ &= \frac{\sqrt{n}}{1 - z_H} \sup_j \left\{ z_H + \frac{j}{n}(1 - z_H) - z_j \right\} \end{aligned}$$

$$\begin{aligned} K^{-*} &= \sqrt{n} \sup_j \left\{ F^*_{\hat{\theta}}(x_{(j)}) - F_n(x_{(j)}) \right\} \\ &= \frac{\sqrt{n}}{1 - z_H} \sup_j \left\{ z_j - \left( z_H + \frac{j-1}{n}(1 - z_H) \right) \right\} \end{aligned}$$

$$K^* = \max\{K^{+*}, K^{-*}\}$$

### 6.3.2 Anderson-Darling statistic

This test was also introduced in the previous subsection for the simple hypothesis. To test a composite hypothesis based on the left-truncated sample, the following computing formula can be used:

$$\begin{aligned} A^{*2} &= -n + 2n \log(1 - z_H) \\ &\quad - \frac{1}{n} \sum_{j=1}^{n} (1 + 2(n - j)) \log(1 - z_j) \\ &\quad + \frac{1}{n} \sum_{j=1}^{n} (1 - 2j) \log(z_j - z_H). \end{aligned}$$

### 6.3.3 Anderson-Darling upper tail statistic

Anderson-Darling upper tail test was proposed by Chernobai, Rachev, and Fabozzi, 2015. The difference to the original Anderson-Darling test is that weights are applied to the right tail. If $F$ denotes the theoretical distribution and $F_n$ is empirical distribution for a complete sample, then the test statistic is computed with

$$A^2_{up} = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{(1 - F(x))} f(x) dx$$

For left-truncated sample and composite hypothesis, the computing formula is given by:

$$A^{*2}_{up} = -2n \log(1 - z_H) + 2 \sum_{j=1}^{n} \log(1 - z_j) + \frac{1 - z_H}{n} \sum_{j=1}^{n} (1 + 2(n - j)) \frac{1}{1 - z_j}$$

## 6.4 Scoring table

Selecting a model based on a given sample may be a challenging task, especially if there are multiple models that - from a theoretical point of view - should provide a reasonable fit. Preparing a table that summarizes how well each of these models performed in goodness-of-fit tests may make the selection process easier. One possible approach is to select few tests and award each model with points depending on the value of the test statistic. There is no unique grading system and one particular example would be to assign

- 1 point for third smallest test statistic
- 2 points for second smallest test statistic
- 3 points for the smallest test statistic.

Selected tests may vary depending on the problem. For instance, for left truncated data with heavy outliers, tests with composite hypotheses described in the previous section may be a good choice:

- Kolmogorov-Smirnoff test - it tries to find a good fit in the middle of the distribution.
- Cramer-von Misses test - alternative to the KS-test.

- Anderson-Darling test - compared to the KS-test, it emphasises the fit in tails.
- the Anderson-Darling upper tail test - focuses on the fit in the upper tail.

The best performing model can be considered as a model with the highest points.

# 7 Case Study: Yearly losses incurred due to Tornadoes in the USA

This analysis focuses on the modelling of the total damages incurred due to tornadoes in the United States from 1988 till 2017 and is based on the dataset maintained by the Storm Prediction Center [1] (SPC), which is in turn the part of the National Oceanic and Atmospheric Administration [2] (NOAA). The SPC tornado archive distinguishes between the damages to property and damages to crops. Both are based on the estimates in the US Dollar values from an insurance company, and if they are not available then they are either estimated by a qualified individual or left blank. The documentation and the dataset and all data collection directives can be found on the official SPC webpage. Throughout this study, the total damages are defined as the sum of damages to property and to crops, scaled by a factor of one billion. The losses are not adjusted for inflation and other possible normalizations, as population and wealth multipliers are also not applied. The compound model with a frequency distribution and severity distribution, split into the body and tail, is applied to the underlying dataset. All computational aspects of the analysis are performed with the R statistical software (R Core Team, 2017) using own developed code as well as special extensions: *fitdistrplus* (Delignette-Muller and Dutang, 2015), *POT* (Ribatet and Dutang, 2016) and *truncgof* (Wolter, 2012).

The study begins with the explorative analysis, which aims to give a better insight into the tornado loss event collection and is intended to facilitate the development of the compound model. It then proceeds with the threshold selection, which enables selecting and calibrating tail and frequency distributions. Lastly, the compound distribution is approximated with a Monte Carlo simulation.

## 7.1 Explorative analysis

According to the dataset compiled by SPA, from 1988 till 2017 there were 38,693 tornadoes causing total losses of about 40.81 billion of US dollars in total. Notably, about 46% of occurrences did not result in any economic damage and almost 94% had individual economic impact smaller than $1M. The biggest single loss was estimated at $2.8B in 2011. As we can see in the figure

[1] https://www.spc.noaa.gov
[2] https://www.noaa.gov

1, year 2011 was the most devastating year in terms of total damages and amounted to 9.71 billion of US dollars. Two years later, in 2013, unusually high losses were again recorded at $3.65 billion. In other years, the estimates vary around the mean of $1.36 with standard deviation of $1.73. The US economy suffered the smallest damages in 2016, 2015 and 1995, of about $180M, $320M and $330M, respectively.

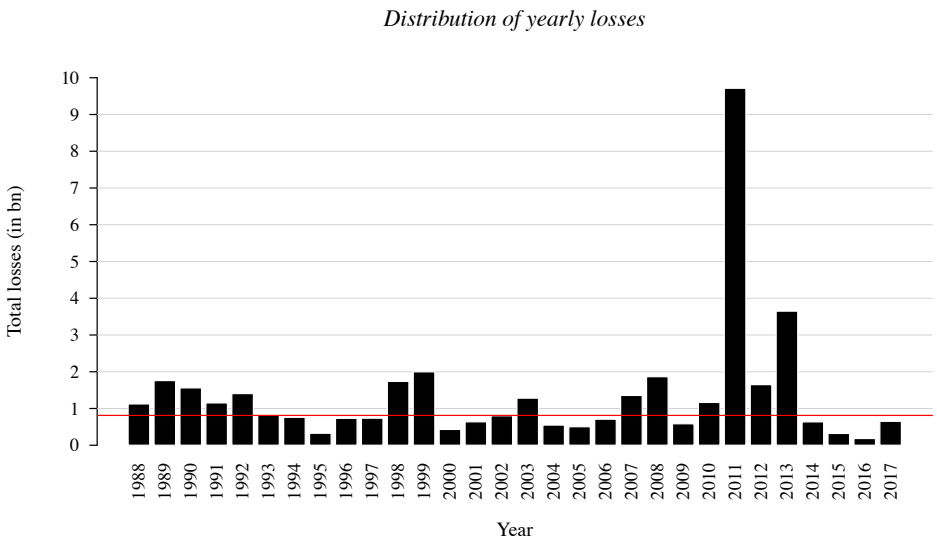



FIGURE 1: Distribution of yearly sums of damages incurred due to tornados in US from 1988 till 2017. The red horizontal represents the mean of historic yearly damages.

On average, there were 1,290 tornadoes every year with standard deviation of about 330. By inspecting the distribution of yearly occurrences, depicted on the figure 2, one can see peaks in 2004, 2008, and 2011 exceeding 1,900 events. The latter peak coincides with the highest total damage on the previously described figure 2.

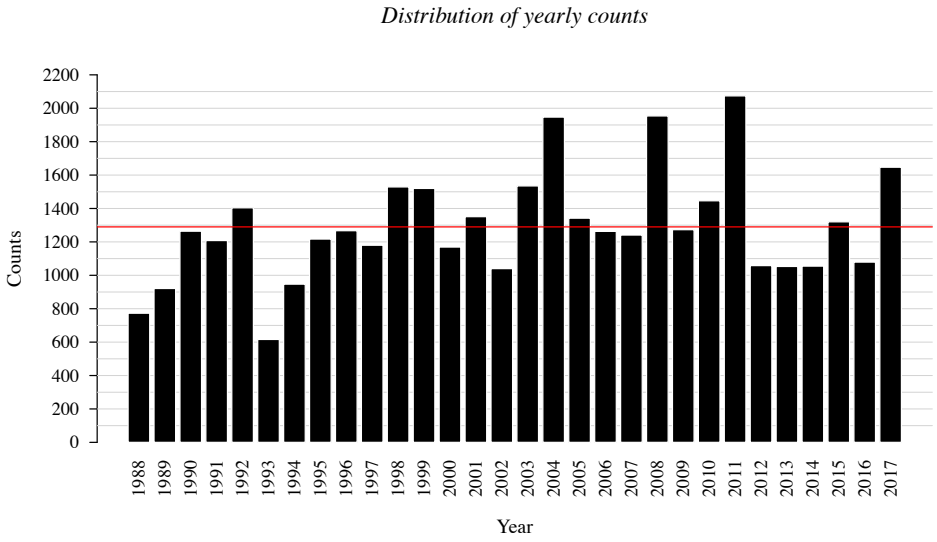



FIGURE 2: Distribution of yearly occurrences of tornadoes in USA from 1988 till 2017. Red line corresponds to the historic yearly average number of tornadoes within the considered time period.

The number of tornadoes is usually below the average and the smallest number, 616, was recorded in 1993. In general, the correlation between yearly counts and totals is 0.425 [CI: $0.076, 0.681$] and is statistically significant at the 95% confidence level. This confirms the common sense that with an increasing number of tornadoes, the economy suffers bigger total damages. All yearly totals and counts together with simple summary statistics are presented in tables 2 and 3 below.

| **Year** | Loss | Count | **Year** | Loss | Count | **Year** | Loss | Count |
|---|---|---|---|---|---|---|---|---|
| **1988** | 1.12 | 773 | **1998** | 1.74 | 1529 | **2008** | 1.87 | 1954 |
| **1989** | 1.76 | 921 | **1999** | 2.00 | 1520 | **2009** | 0.58 | 1273 |
| **1990** | 1.56 | 1264 | **2000** | 0.43 | 1169 | **2010** | 1.16 | 1446 |
| **1991** | 1.15 | 1208 | **2001** | 0.64 | 1351 | **2011** | 9.71 | 2074 |
| **1992** | 1.41 | 1404 | **2002** | 0.80 | 1040 | **2012** | 1.65 | 1058 |
| **1993** | 0.82 | 616 | **2003** | 1.28 | 1535 | **2013** | 3.65 | 1053 |
| **1994** | 0.76 | 947 | **2004** | 0.55 | 1947 | **2014** | 0.64 | 1055 |
| **1995** | 0.33 | 1217 | **2005** | 0.50 | 1342 | **2015** | 0.32 | 1320 |
| **1996** | 0.73 | 1267 | **2006** | 0.71 | 1263 | **2016** | 0.18 | 1079 |
| **1997** | 0.74 | 1180 | **2007** | 1.36 | 1241 | **2017** | 0.65 | 1647 |

TABLE 2: Yearly total losses (in bn) and counts from 1988 till 2017.

| | Loss | Count |
|---|---|---|
| Sum | 40.81 | 38693 |
| Average | 1.36 | 1290 |
| Std. dev. | 1.73 | 328 |

TABLE 3: Simple summary statistics of yearly total losses and counts from 1988 till 2017.

During the time period of the study, there were almost 40,000 tornadoes, however, only a small fraction of them inflicted damages that could be described as catastrophic. The top 10 highest losses, account for 26% of the total historic losses ($40.81B). Analogously, top 25 and 50 highest losses correspond to 36% and 45% of the total share.

The top 10 most catastrophic tornadoes are summarized in the table 4. It can be seen that 8 of these tornadoes occurred after 2010 and 5 of them only in 2011. The most devastating tornado happened in 2011 in Missouri inflicting a total damage of $2.8B. Half of the aforementioned tornadoes wreaked havoc in Alabama and Oklahoma.

| Loss | Year | State |
|---|---|---|
| 2.8 | 2011 | Missouri |
| 2 | 2013 | Oklahoma |
| 1.5 | 2011 | Alabama |
| 1 | 2011 | Alabama |
| 0.91 | 2013 | Illinois |
| 0.7 | 2011 | Alabama |
| 0.5 | 2011 | Mississippi |
| 0.5 | 2012 | Kansas |
| 0.45 | 1999 | Oklahoma |
| 0.45 | 1999 | Oklahoma |

TABLE 4: Summary of the ten most destructive individual tornadoes from 1988 till 2017 in United States. Each loss is valued in billion US dollars.

As it can be seen on figures 3 and 4, the south-central part of the USA is mostly affected by tornadoes. Alabama, Missouri and Oklahoma lost more than $4B each in the past 30 years. Most of these damages come from the catastrophic events listed in the table 4.

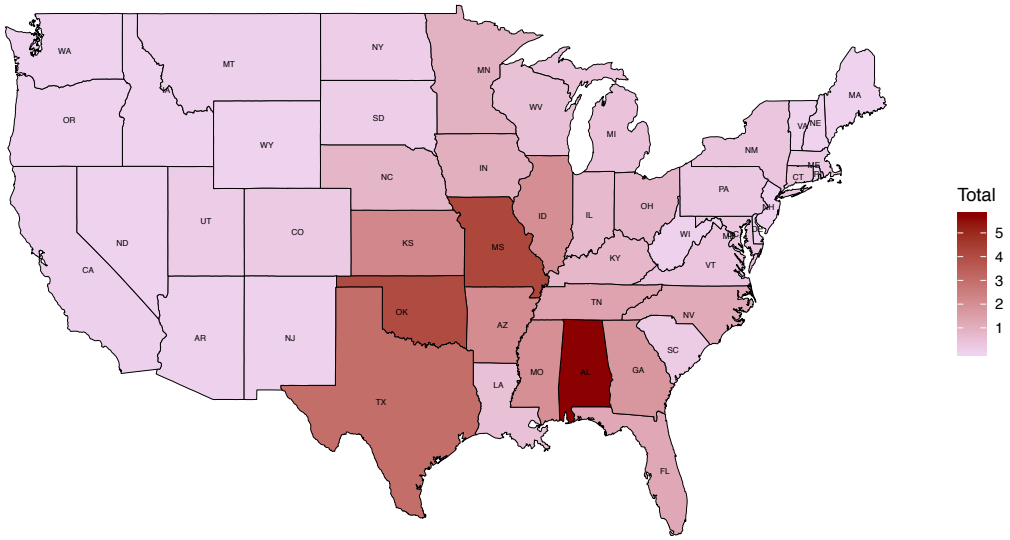



FIGURE 3: Total damages (in bn) incurred due to tornadoes in USA at state level from 1988 till 2017.

Although, Texas is the state with the most recorded tornadoes - more than 4,500 - it is on the fourth place in terms of damages at around $3B.

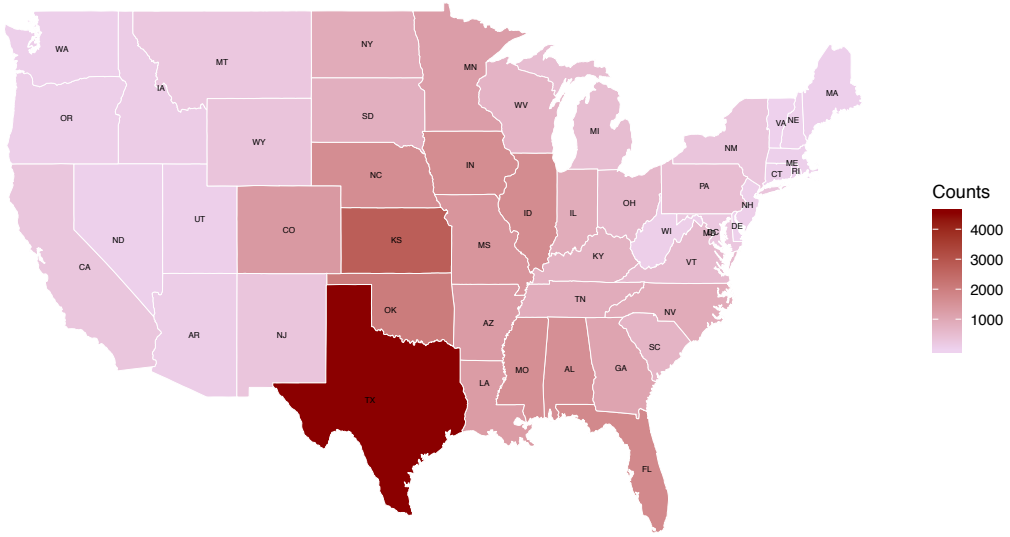



FIGURE 4: Occurrences of tornadoes in USA at state level from 1988 till 2017.

## 7.2 Selection of threshold and severity distribution

As indicated in the previous subsection, the underlying dataset contains multiple extreme events that have to be addressed with a very heavy tailed dis-

tribution. The most reasonable choice for the tail distribution, supported by the extreme value theory (as described in subsection 3.1.1) as well as by the explorative analysis, is the Generalized Pareto Distribution. However, three other distributions are going to be fit to the right tail with the same threshold - log-normal, log-logistic and inverse Gaussian. The first has light right tail, the second is similar to the log-normal but has heavier tail, and finally the inverse Gaussian is known for its heavy tail. They are going to serve merely as benchmarks and if any of them would over-perform the GPD, the choice for the tail model would need to be reconsidered.

One of the most challenging part of this analysis was to select an appropriate threshold that separates body from the tail and, hence, defines extreme events. In the following, two visual methods are going to be used - mean excess plot and stability of parameters. The strategy is to identify such value, above which the mean excess plot is approximately linear, and simultaneously above which both parameter estimates indicate stability. In both methods thresholds from 1M to 100M are considered. Values above 100M would result in a very small number of losses in the tail. Such choices would support the asymptotic argument and lower the bias, however at too high cost of increased uncertainty around the parameter estimates.

The figure 8 represents the mean excess plot. It can be seen that there are two major jumps around 25M and 50M. After the latter value, the graph seems to increase almost linearly, which suggests a threshold of 50M.

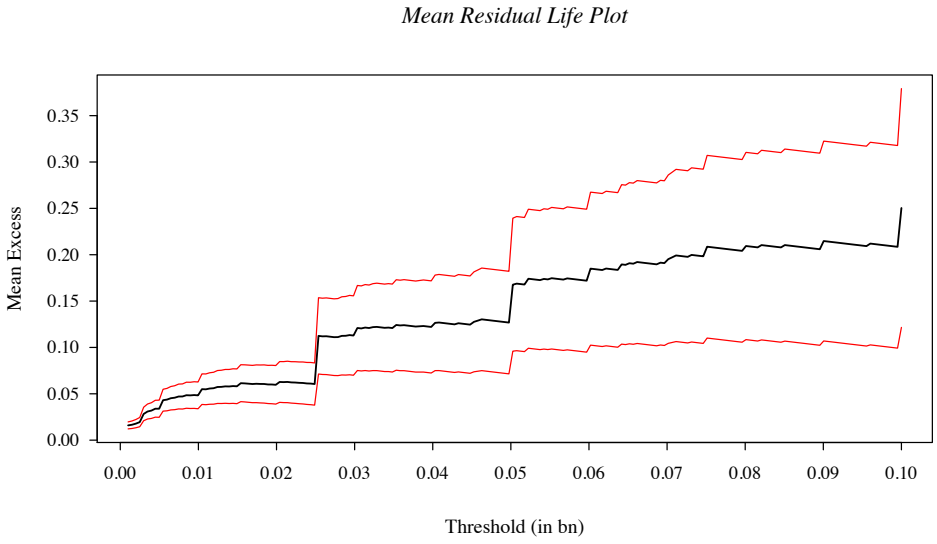
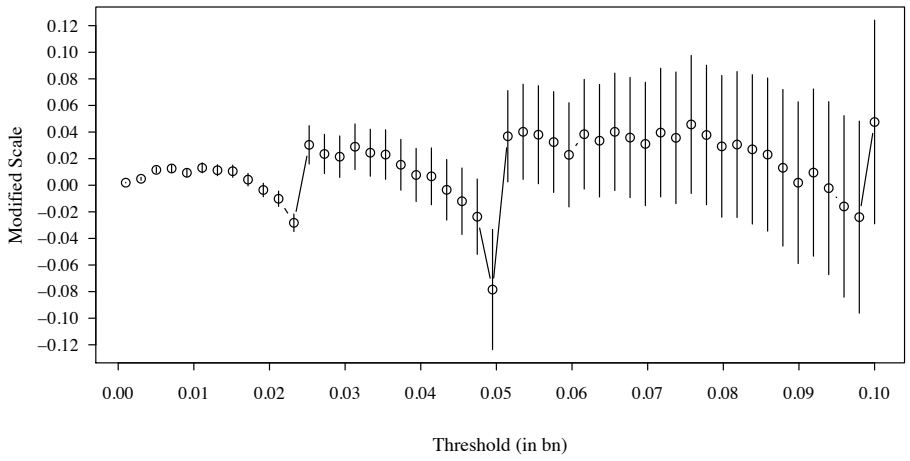



FIGURE 5: Mean excess plot for Generalized Pareto Distribution. All considered thresholds fall into the range of 1M to 100M.

By inspecting the figures 6a and 6b that show the estimates of the modified scale and shape parameters with confidence intervals over a range of pre-determined thresholds, one can see that the estimated values become more stable after the value of 50M. This supports the hypothesis that an optimal value for the threshold is about 50M. Hence, the threshold is going to be fixed at 50M and all values above are going to be considered as tail events.

(A)




(B)

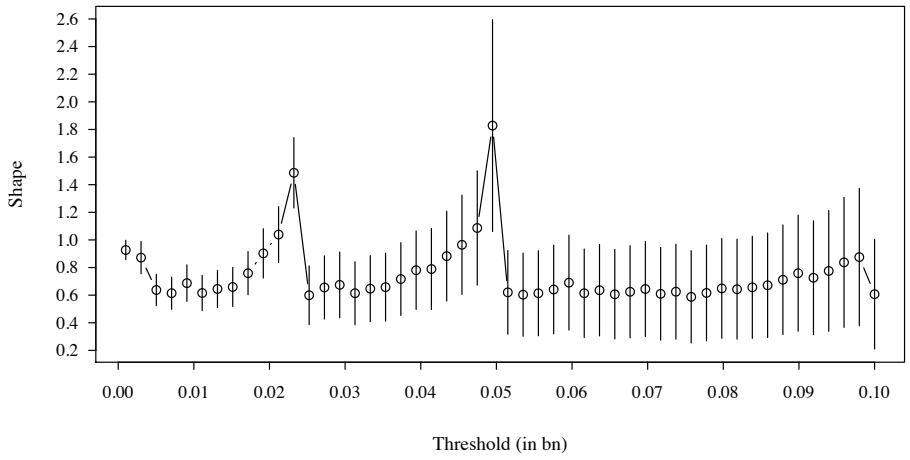


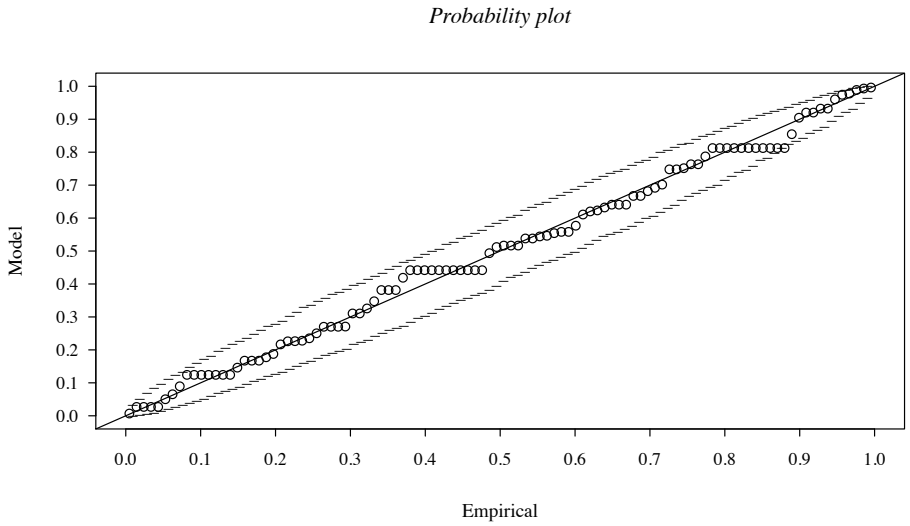

FIGURE 6: Stability of parameter estimates for thresholds between 1M to 100M. Panels (A) and (B) depict estimates of the modified scale parameter and the shape parameter, respectively.

As the threshold is fixed and the tail with 104 losses is uniquely defined, the parameters of the Generalized Pareto Distribution and all three bench-

mark models are going to be estimated using the minimum distance estimation technique with the Anderson-Darling distance. This combination offers excellent robustness properties and also a good fit in tails. The parameter estimates of the scale parameter is 0.07267 with a standard error of 0.0125 and the estimate of the shape parameter is 0.5517 with a standard error of 0.193. Both standard errors were simulated with 10.000 bootstrap samples.

The figure 7 shows the probability and quantile plots for the GPD. In case of the first diagnostic graph, all points fall within the simulated 95% confidence interval and are close to the unit diagonal. The quantile plot shows bigger departures from the theoretical line for values bigger than 0.9B, however they are still within confidence bounds. Although not perfect, the visual inspection of classical diagnostic plots indicates a reasonable fit of the model.

(A)




(B)

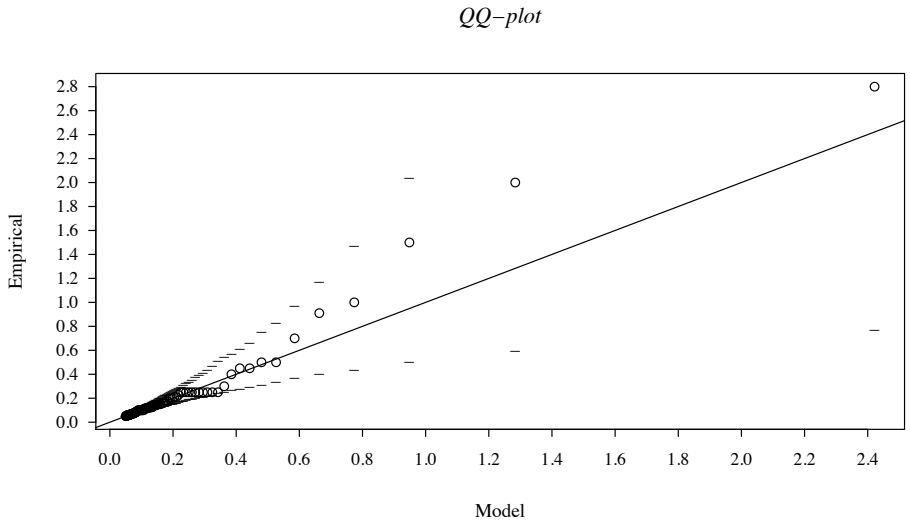



FIGURE 7: Probability plots for the Generalized Pareto Distribution. Panels (A) and (B) contains the pp-plot and qq-plot respectively. Both graphs include simulated 95% confidence intervals.

In addition to the diagnostic plots, three goodness of fit tests with composite hypotheses, previously described in the subsection 6.3, were performed using simulation with 10.000 repetitions and are summarized in the table 5. They are meant to asses the goodness-of-fit in global fashion focusing on the fit-in-the-middle of distribution (KS-test) and in tails. As it can be seen, all tests fail to reject the the null hypothesis at the confidence level of 95% and proves that the Generalized Pareto Distribution provides a reasonable fit to the tail.

| Test | T | p-value |
|---|---|---|
| Kolmogorov-Smirnoff | 0.74 | 0.60 |
| Anderson-Darling | 1.88 | 0.51 |
| Anderson-Darling upper tail | 3.66 | 0.31 |

TABLE 5: Summary of goodness of fit tests with composite hypothesis for the Generalized Pareto Distribution.

Similar goodness-of-fit tests were performed for all benchmark models. As all of these statistical tests measure the discrepancy between the empirical distribution and the fitted distribution, the smaller the test statistic, the better the fit to the dataset. The table 6 summarizes results for each test and distribution. The fitted Generalized Pareto Distribution has uniformly smallest values of test statistics, which in turn indicates that no benchmark model is performing better, given events in the tail.

| | GPD | Inv. Gaussian | Log-logistic | Log-normal |
|---|---|---|---|---|
| Anderson-Darling | 1.88 | 5.16 | 3.69 | 3.94 |
| Anderson Darling upper tail | 3.66 | 19.71 | 9.73 | 8.89 |
| Kolmogorov-Smirnoff | 0.74 | 2.47 | 1.65 | 1.79 |

TABLE 6: Summary of goodness-of-fit tests for Generalized Pareto Distribution and benchmark models.

An important aspect of a proper model is how well it is calibrated to the data. A return level plot is another visual diagnostic method that addresses this issue by depicting the level which is expected to be exceeded on average once in $k$-years. The figure 8 represents the aforementioned tool for the fitted Generalized Pareto Distribution. Yearly levels of tail events, indicated with circles on the graph, are adjusted according to the method suggested in R. M. Hosking and R. Wallis, 1995. It can be seen that they are relatively close to the solid line and all of them fall within the simulated 95% confidence bounds. Slightly bigger departure can be seen for the second and third biggest tail events. The highest loss in the tail ($2.8B from 2011) is well captured by the model according to the graph.

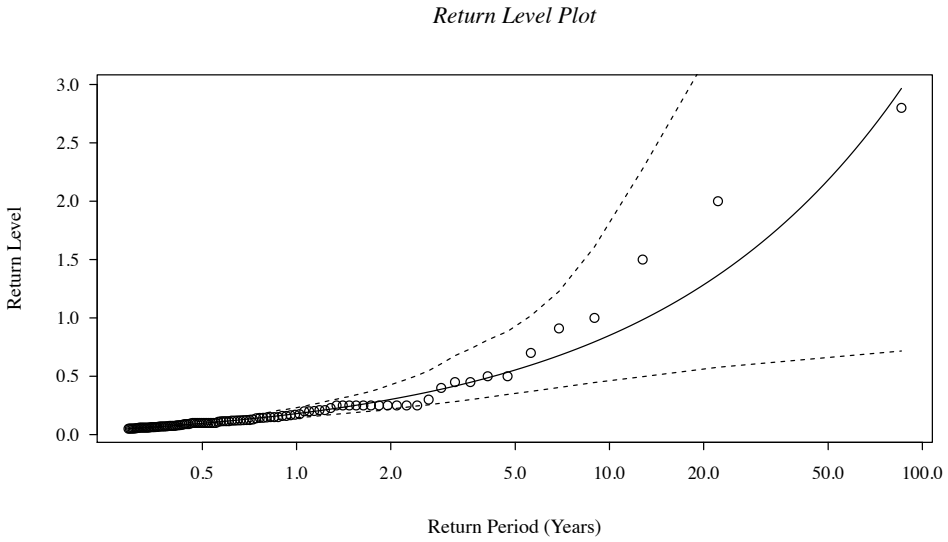



FIGURE 8: Return level plot based on the Generalized Pareto Distribution with estimated 95% confidence interval.

Based on the diagnostic plots and the statistical tests, there is no evidence speaking against the particular instance of the Generalized Pareto Distribution chosen to model the tail events. Also, benchmark models do not perform better in terms of the goodness-of-fit to the tail events. Hence, the fitted GPD (shape: 0.5517, scale: 0.07267) model is going to be used for the modelling of the tail of the severity distribution. The other part of the severity distribution, the body, is going to be modelled with the empirical distribution as described in subsection 3.2.

## 7.3 Selecting and estimating frequency distributions

For the modelling of frequency in the body and tail only Poisson and negative binomial distributions are going to be considered. The binomial law is excluded because there is no reason to assume that there is a fixed maximal number of tornadoes during a year. Selected models are going to be estimated using maximum likelihood method. Tables 7 and 8 below provide a full description of frequencies in the body and tail.

| **Year** | Body | Tail | **Year** | Body | Tail | **Year** | Body | Tail |
|---|---|---|---|---|---|---|---|---|
| **1988** | 772 | 1 | **1998** | 1523 | 6 | **2008** | 1948 | 6 |
| **1989** | 918 | 3 | **1999** | 1511 | 9 | **2009** | 1271 | 2 |
| **1990** | 1263 | 1 | **2000** | 1167 | 2 | **2010** | 1439 | 7 |
| **1991** | 1206 | 2 | **2001** | 1349 | 2 | **2011** | 2056 | 18 |
| **1992** | 1403 | 1 | **2002** | 1037 | 3 | **2012** | 1052 | 6 |
| **1993** | 616 | 0 | **2003** | 1531 | 4 | **2013** | 1047 | 6 |
| **1994** | 946 | 1 | **2004** | 1945 | 2 | **2014** | 1053 | 2 |
| **1995** | 1217 | 0 | **2005** | 1339 | 3 | **2015** | 1320 | 0 |
| **1996** | 1264 | 3 | **2006** | 1261 | 2 | **2016** | 1079 | 0 |
| **1997** | 1177 | 3 | **2007** | 1233 | 8 | **2017** | 1646 | 1 |

TABLE 7: Yearly total counts in body and tail from 1988 till 2017.

| | Body | Tail |
|---|---|---|
| Total | 38589 | 104 |
| Average | 1286.3 | 3.47 |
| Variance | 106340.8 | 13.64 |
| Std. Dev. | 326.1 | 3.69 |

TABLE 8: Simple summary statistics of yearly counts in the body and tail and counts from 1988 till 2017.

As it can be seen on the figure 9, there was only one year with more than 10 tornadoes, each causing damage of over $50M. In four other years there were no tornadoes that could be classified as tail events. On average there were 3.47 tail occurrences with a standard deviation 3.69. The empirical variance is considerably larger than the empirical mean due to 18 counts in 2011. This discrepancy suggests that the negative binomial model may be preferable over Poisson distribution, since the latter assumes the variance being equal to the mean.

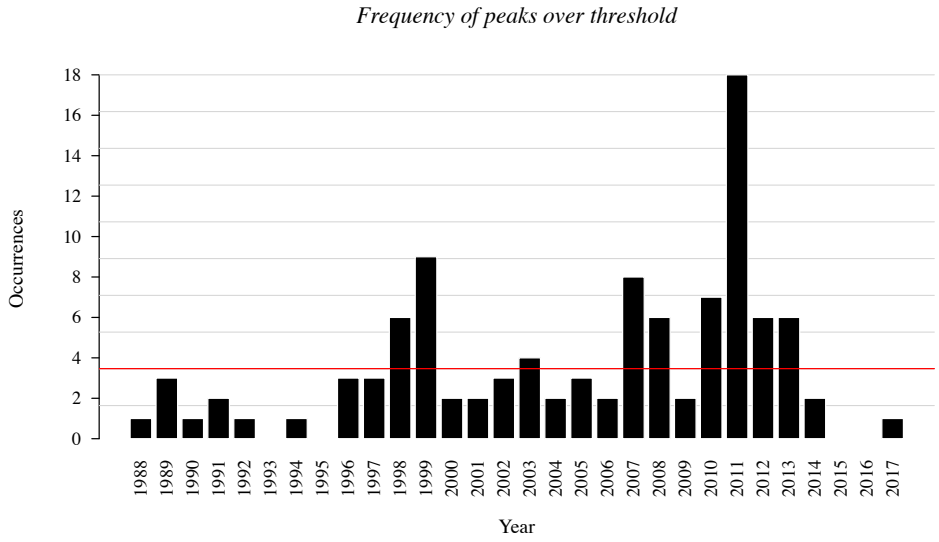
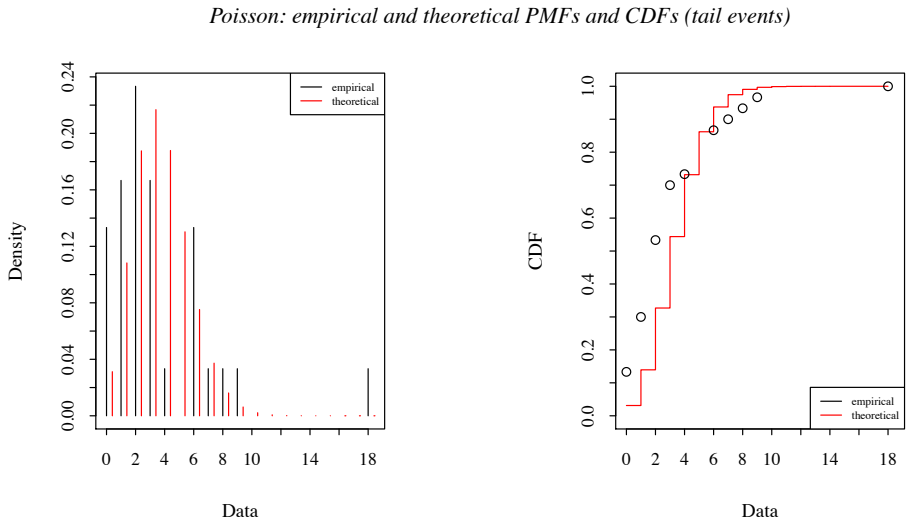



FIGURE 9: Frequency of yearly yearly occurrences of peaks over threshold.

By inspecting the figure 10 one can notice that the negative binomial provides a slightly better description of tail counts than the Poisson distribution. In addition to the visual analysis, formal chi-squared tests for both models were performed. The null hypothesis that the tail counts follow Poisson model, was rejected at 95% significance level (p-value: 0.0013) and the test failed to reject the negative binomial model (p-value: 0.489). The same analysis was performed for the counts in the body. On the figure 10, one can notice that the Poisson distribution fails to describe data well, whereas the negative binomial provides a reasonably good fit. The chi-square test does not reject the negative binomial model. All test results are summarized in table 9. Hence, the negative binomial distribution is going to be used to model body and tail occurrences.

(A)




(B)

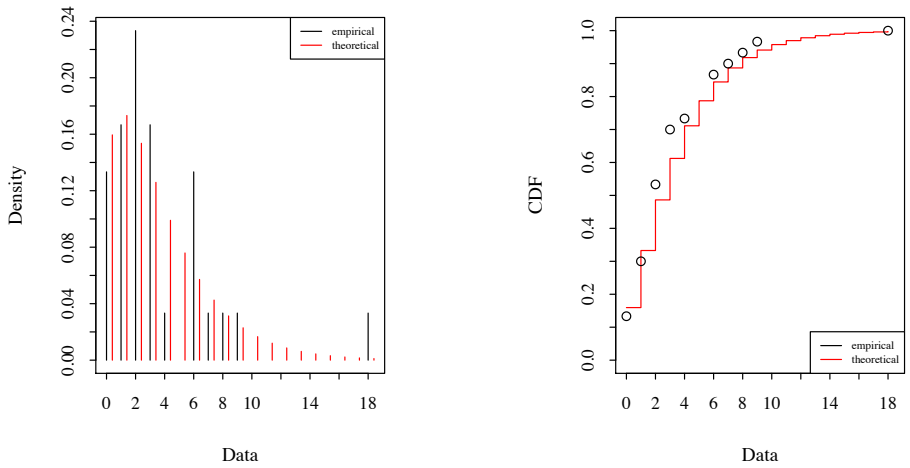
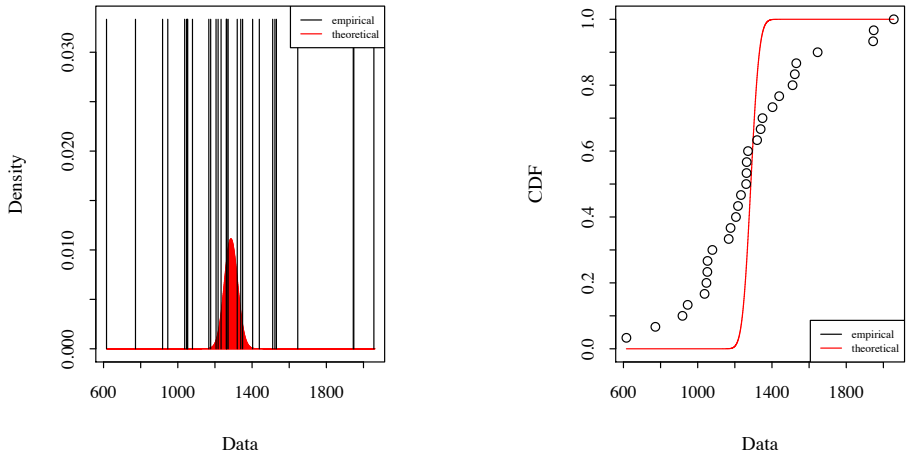



FIGURE 10: Panel (A) represents the fit of the Poisson distribution and the panel (B) represent the fit of the negative binomial distribution to the tail counts.

(A)

*Poisson: empirical and theoretical PMFs and CDFs (body counts)*




(B)

*Negative binomial: empirical and theoretical PMFs and CDFs (body counts)*

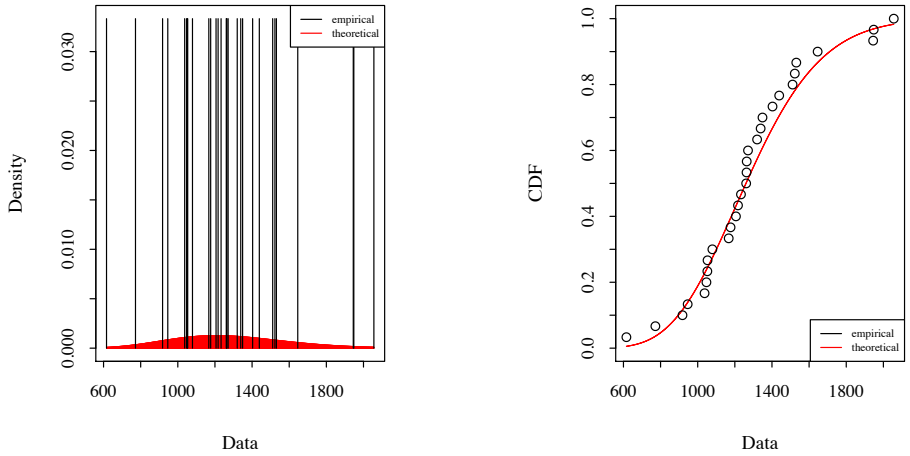



FIGURE 11: Panel (A) depicts the fit of the Poisson distribution to counts in the body. Analogously, panel (B) shows the fit of the negative binomial distribution.

| | Tail | | | Body | | |
|---|---|---|---|---|---|---|
| | Estimates | T | p-value | Estimates | T | p-value |
| Poisson | $\lambda$= 3.467 (SE: 0.34) | 17.95 | 0.001 | $\lambda$ = 1286.3 (SE: 6.55) | >1000 | 0 |
| Negative Binomial | size = 1.5817(SE: 0.613)<br>mu = 3.4663 (SE: 0.607) | 2.42 | 0.489 | size = 16.254 (SE: 4.21)<br>mu = 1286.325 (SE: 58.62) | 3.17 | 0.53 |

TABLE 9: Maximum likelihood estimates and results of chi-square tests for Poisson and negative binomial distributions.

## 7.4 Estimating compound distribution

Total yearly damage incurred due tornadoes in United States from 1988 till 2017 is going to be modelled with a compound negative binomial distribution, in which tail events are defined as losses exceeding $50M and are described by the Generalized Pareto distribution with scale parameter 0.07267 (SE:0.0125) and shape parameter 0.5517 (SE: 0.193). Estimates were obtained using the minimum distance estimator with Anderson-Darling distance. Prior the analysis, all losses were scaled by a factor of 1B. Economic damages smaller than the threshold are described via empirical distribution. Since the negative binomial distribution provides a good fit to counts in the body and tail, it is used in both cases as frequency distribution. For the body, the maximum likelihood estimates are: size=16.254 (SE: 4.21) and mu=1286.325 (SE: 58.62) and for the tail: size=1.5817 (SE: 0.61) and mu=3.4664 (SE: 0.61). The compound aggregate distribution Monte Carlo simulation with 10M iterations.

The figure 12 presents the return level plot for the approximated compound aggregate distribution. It can be seen that empirical 1 to 2 years return periods, marked with points, are slightly underestimated by the model. From 2 to 20 years return period, the model captures well the empirical behaviour with all points being close or at least within simulated 95% confidence interval. For higher return periods the model extrapolates smoothly.

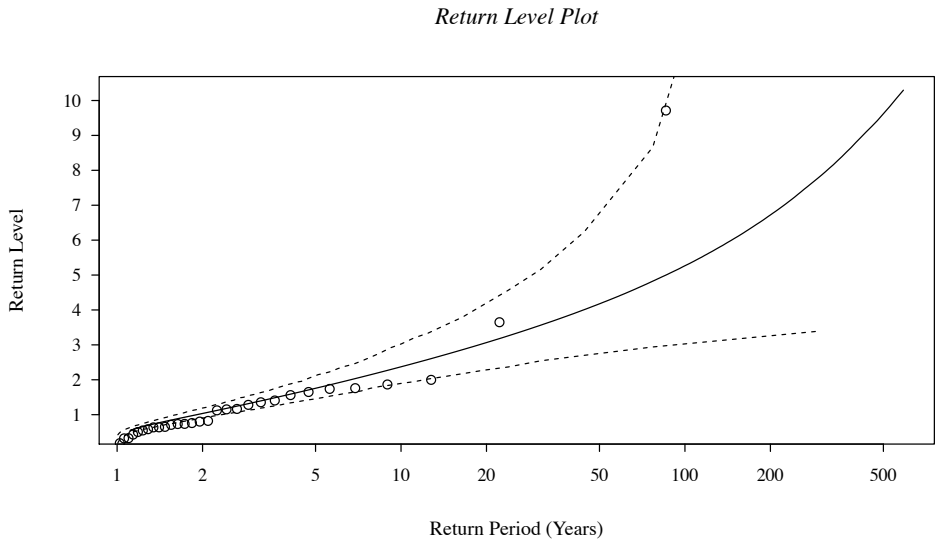



FIGURE 12: Return level plot based on the compound negative binomial model with simulated 95% confidence interval.

| Return Period (Years) | 2 | 5 | 25 | 50 | 100 | 500 | 1000 | 10000 |
|---|---|---|---|---|---|---|---|---|
| Return level (in bn) | 1.06 | 1.56 | 2.72 | 3.49 | 4.57 | 9.34 | 13.05 | 41.75 |

TABLE 10: Estimated return levels based on the developed negative binomial model.

The table 10 represents estimated return levels. 1 in 1000 years event corresponds to a \$13.05B loss. The highest historical loss of \$9.7B from 2011 is predicted to be 1 in about 500 years loss. It empirically corresponds to 1 in 85 years event but the prediction is still within 95% confidence bounds. It can be seen that the model provides a realistic description of the historical yearly losses. Extending the developed model with seasonal and spatial effects could provide even more accurate description, but such undertaking is beyond of the scope of this thesis.

# 8 Summary

The main goal of this work is to describe a modelling strategy for phenomena that generate a large number of events causing minor damage and only a small number of extreme events that manifest themselves through enormous, even catastrophic impact. Total damage incurred due to such phenomena can be estimated with a model, that is a merge of two separate distributions that describe the number of occurrences and their severities independently, elaborated throughout this thesis. The theory is then applied to a study of yearly total losses caused by tornadoes in United States from 1988 till 2017. The Generalized Pareto Distribution is chosen to model tail events separated from the body by a threshold of $50M. The negative binomial distribution is used to model frequency. The developed compound negative binomial model is shown to provide a realistic description of yearly losses and can be further extended to include seasonal and spatial effects.

# List of Figures

# List of Tables

# References

Aldrich, John (1997). "R.A. Fisher and the making of maximum likelihood 1912-1922". In: *Statist. Sci.* 12.3, pp. 162–176. DOI: 10.1214/ss/1030037906. URL: https://doi.org/10.1214/ss/1030037906.

Anderson, T. W. and D. A. Darling (1952). "Asymptotic Theory of Certain Goodness of Fit Criteria Based on Stochastic Processes". In: *Ann. Math. Statist.* 23.2, pp. 193–212. URL: https://doi.org/10.1214/aoms/1177729437.

— (1954). "A Test of Goodness of Fit". In: *Journal of the American Statistical Association* 49.268, pp. 765–769. ISSN: 01621459. URL: http://www.jstor.org/stable/2281537.

Chernobai, Anna, Svetlozar T. Rachev, and Frank J. Fabozzi (2015). "Composite goodness-of-fit tests for left-truncated loss samples". In: *Handbook of Financial Econometrics and Statistics* 1.805, pp. 575–596. ISSN: 1098-6596. DOI: 10.1007/978-1-4614-7750-1_20.

Coles, S (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer series in statistics. Springer. ISBN: 978-1-4471-3675-0. DOI: 10.1007/978-1-4471-3675-0.

David, F N. and J. Neyman (1938). "Extension of the Markoff theorem on least squares". In: *Statistical Research Memoirs* 2, pp. 105–116.

Delignette-Muller, Marie Laure and Christophe Dutang (2015). "fitdistrplus: An R Package for Fitting Distributions". In: *Journal of Statistical Software* 64.4, pp. 1–34. URL: http://www.jstatsoft.org/v64/i04/.

Drossos, Constantine A. and Andreas N. Philippou (1980). *A note on minimum distance estimates*. DOI: 10.1007/BF02480318.

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1, pp. 1–26. DOI: 10.1214/aos/1176344552. URL: https://doi.org/10.1214/aos/1176344552.

Feller, W. (1948). "On the Kolmogorov-Smirnov Limit Theorems for Empirical Distributions". In: *Ann. Math. Statist.* 19.2, pp. 177–189. DOI: 10.1214/aoms/1177730243. URL: https://doi.org/10.1214/aoms/1177730243.

Jenkinson, A.F. (1955). "The frequency distribution of the annual maximum (or minimum) values of meteorological elements". In: *Quarterly Journal of the Royal Meteorology Society* 87, pp. 145–158.

Klugman, S.A., H.H. Panjer, and G.E. Willmot (1998). *Loss Models: From Data to Decisions*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9780471238843.

Klugman, Stuart A., Harry H. Panjer, and E. Willmot. Gordon (1988). *Loss Models. From Data to Decisions*. John Wiley and Sons.

Makarov, Mikhail (2007). "Applications of exact extreme value theorem". In: 2, pp. 115–120.

Millar, P. Warwick (1981). "Robust estimation via minimum distance methods". In: *Zeitschrift fuer Wahrscheinlichkeitstheorie und Verwandte Gebiete* 55.1, pp. 73–89. ISSN: 1432-2064. DOI: 10.1007/BF01013462. URL: https://doi.org/10.1007/BF01013462.

Panjer, H.H. (2006). *Operational Risk: Modeling Analytics*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9780470051306.

Parr, William C. and William R. Schucany (1980). "Minimum Distance and Robust Estimation". In: *Journal of the American Statistical Association* 75.371, pp. 616–624. DOI: 10.1080/01621459.1980.10477522.

Pearson, K (1900). "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302, pp. 157–175. DOI: 10.1080/14786440009463897. eprint: https://doi.org/10.1080/14786440009463897. URL: https://doi.org/10.1080/14786440009463897.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

R. M. Hosking, J and J R. Wallis (1995). "A Comparison of Unbiased and Plotting-Position Estimators of L Moments". In: *Water Resources Research - WATER RESOUR RES* 31.

Raïffa, H. and R. Schlaifer (1961). *Applied statistical decision theory*. Studies in managerial economics. Division of Research, Graduate School of Business Adminitration, Harvard University. ISBN: 9780875840178.

Ribatet, Mathieu and Christophe Dutang (2016). *POT: Generalized Pareto Distribution and Peaks Over Threshold*. R package version 1.1-6. URL: https://CRAN.R-project.org/package=POT.

Smirnov, N. (1948). "Table for Estimating the Goodness of Fit of Empirical Distributions". In: *Ann. Math. Statist.* 19.2, pp. 279–281. DOI: 10.1214/aoms/1177730256. URL: https://doi.org/10.1214/aoms/1177730256.

Vaart, A. W. van der (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. DOI: 10.1017/CBO9780511802256.

Wolfowitz, J. (1957). "The Minimum Distance Method". In: *The Annals of Mathematical Statistics* 28.1, pp. 75–88. DOI: 10.1214/aoms/1177707038. URL: https://doi.org/10.1214/aoms/1177707038.

Wolter, Thomas (2012). *truncgof: GoF tests allowing for left truncated data*. R package version 0.6-0. URL: https://CRAN.R-project.org/package=truncgof.

# Appendix

## Abstract

In many real world phenomena the need for modelling of losses caused by an underlying process that generates both large number of occurrences with minor losses and occurrences with severe impact exists. Prominent examples are natural disasters (earthquakes, hurricanes, tornadoes etc.), operational losses in highly regulated institutions or claims generated by an insurance policy. In each case, the aggregated loss distribution can be modelled with a so called compound model that is a combination of two separate distributions describing the number of events and their severities. The latter can be further split into the empirical body and tail. This distinction allows the modelling of the high-probability-low impact and low-probability-high-impact aspects of the underlying data with an empirical distribution function and a continuous distribution as justified by the Glivenko–Cantelli theorem and the extreme value theory, respectively. The developed model is then applied to a study of yearly total losses caused by tornadoes in the United States from 1988 till 2017. Based on theoretical considerations and empirical evidence the Generalized Pareto Distribution is chosen to model tail events separated from the empirical body by a threshold of $50M. It is then calibrated applying the minimum distance estimation with the Anderson-Darling distance. The frequency analysis reveals that the negative binomial distribution provides a good fit of yearly tornado counts. The resulting compound negative binomial model provides a realistic description of yearly loss events and can be further extended to include seasonal and spatial effects.

## Kurzfassung

In vielen Problemstellungen ist es erforderlich Verluste zu modellieren, verursacht durch den zugrundeliegenden Prozess, der eine große Anzahl von Ereignissen mit geringer Verlusthöhe und wenige Ereignisse mit sehr hohen Verlusten erzeugt. Prominente Beispiele sind Naturkatastrophen (wie beispielsweise Erdbeben, Wirbelstürme, Tornados), Betriebsverluste in stark regulierten Institutionen oder Versicherungsfälle. In jedem Fall kann die Gesamtverlustverteilung mit einer sogenannten zusammengesetzten Verteilung modelliert werden, bei der zwei getrennte Verteilungen zusammengeführt werden, die die Anzahl der Ereignisse und ihre Verlusthöhe beschreiben. Letzteres kann weiter in das empirische Verteilungszentrum und den Verteilungsrand aufgeteilt werden. Diese Unterscheidung erlaubt die Modellierung der Aspekte hoher Wahrscheinlichkeit mit geringer Auswirkung und niedriger Wahrscheinlichkeit mit hoher Auswirkung der zugrunde liegenden Daten mittels einer empirischen Verteilungsfunktion bzw. einer kontinuierlichen Verteilung, die jeweils durch das Glivenko-Cantelli-Theorem und der Extremwerttheorie gerechtfertigt sind. Das entwickelte Model wird dann auf eine Untersuchung der jährlichen Gesamtverluste durch Tornados in den Vereinigten Staaten von 1988 bis 2017 angewendet. Basierend auf theoretischen Überlegungen und statistischen Nachweisen wird die generalisierte Pareto-Verteilung gewählt. Dies wird dann mit der minimum distance estimationMethode mittels der Anderson-Darling-Distanz kalibriert. Die Häufigkeitsanalyse zeigt, dass die negative Binomialverteilung eine gute Beschreibung der jährlichen Tornado-Zählungen liefert. Die daraus resultierende zusammengesetzte negativ Binomial Verteilung bietet eine realistische Beschreibung der jährlichen Verlustereignisse und kann um saisonale und räumliche Effekte erweitert werden.