Modelling catastrophe claims with left-truncated severity distributions (extended version)

Chernobai, Anna; Burnecki, Krzysztof; Rachev, Svetlozar; Trueck, Stefan and Weron, Rafal
Hugo Steinhaus Center, Wroclaw University of Technology

2005

Hugo Steinhaus Center for Stochastic Methods
Institute of Mathematics
Wrocław University of Technology

Research Report HSC/05/1

Modelling catastrophe claims with left-truncated severity distributions (*extended version*)

Anna Chernobai[1], Krzysztof Burnecki[2], Svetlozar Rachev[1,3], Stefan Trück[3], Rafał Weron[2]

[1] University of California Santa Barbara, CA 93106, USA
[2] Hugo Steinhaus Center for Stochastic Methods, Institute of Mathematics and Computer Science, Wrocław University of Technology, 50-370 Wrocław, Poland
[3] Institut für Statistik und Mathematische Wirtschaftstheorie, Universität Karlsruhe, D-76128 Karlsruhe, Germany

Abstract: In this paper, we present a procedure for consistent estimation of the severity and frequency distributions based on incomplete insurance data and demonstrate that ignoring the thresholds leads to a serious underestimation of the ruin probabilities. The event frequency is modelled with a non-homogeneous Poisson process with a sinusoidal intensity rate function. The choice of an adequate loss distribution is conducted via the in-sample goodness-of-fit procedures and forecasting, using classical and robust methodologies.

Mailing address:
The Hugo Steinhaus Center
Wrocław University of Technology
Wybrzeże Wyspiańskiego 27
50-370 Wrocław, Poland

(+48-71) 320-3530
(+48-71) 320-2654
hugo@im.pwr.wroc.pl

Modelling catastrophe claims with left-truncated severity distributions[1]

Anna Chernobai[1], Krzysztof Burnecki[2], Svetlozar Rachev[1,3], Stefan Trück[3] and Rafał Weron[2]

[1]University of California Santa Barbara, CA 93106, USA
[2] Hugo Steinhaus Center for Stochastic Methods, Institute of Mathematics, Wrocław University of Technology, 50-370 Wrocław, Poland
[3] Institut für Statistik und Mathematische Wirtschaftstheorie, Universität Karlsruhe, D-76128 Karlsruhe, Germany

Summary

In this paper, we present a procedure for consistent estimation of the severity and frequency distributions based on incomplete insurance data and demonstrate that ignoring the thresholds leads to a serious underestimation of the ruin probabilities. The event frequency is modelled with a non-homogeneous Poisson process with a sinusoidal intensity rate function. The choice of an adequate loss distribution is conducted via the in-sample goodness-of-fit procedures and forecasting, using classical and robust methodologies.

Keywords: Natural Catastrophe, Property Insurance, Loss Distribution, Truncated Data, Ruin Probability

[1]E-mail addresses: A.Chernobai: `annac@pstat.ucsb.edu`, K.Burnecki: `Krzysztof.Burnecki@pwr.wroc.pl`, S.Rachev: `rachev@pstat.ucsb.edu`, S.Trück: `stefan.trueck@statistik.uni-karlsruhe.de` and R.Weron: `Rafal.Weron@pwr.wroc.pl`.

1 Introduction

Due to increasingly severe catastrophes in the last five years the property insurance industry has paid out over \$125 billion in losses. In 2004 property insured losses resulting from natural catastrophes and man-made disasters, excluding the tragic tsunami of December 26, amounted to \$42 billion, of which 95% was caused by natural disasters and 5% by man-made incidents (SwissRe 2004). These huge billion dollar figures call for very accurate models of catastrophe losses. Even small discrepancies in model parameters can result in underestimation of risk leading to billion dollar losses of the reinsurer. Hence, sound statistical analysis of the catastrophe data is of uttermost importance.

In this paper we analyze losses resulting from natural catastrophic events in the United States. Estimates of such losses are provided by ISO's (Insurance Services Office Inc.) Property Claim Services (PCS). The PCS unit is the internationally recognized authority on insured property losses from catastrophes in the United States, Puerto Rico, and the U.S. Virgin Islands. PCS investigates reported disasters and determines the extent and type of damage, dates of occurrence, and geographic areas affected. It is the only insurance-industry resource for compiling and reporting estimates of insured property losses resulting from catastrophes. For each catastrophe, the PCS loss estimate represents anticipated industrywide insurance payments for property lines of insurance covering: fixed property, building contents, time-element losses, vehicles, and inland marine (diverse goods and properties), see Burnecki et al. (2000).

In the property insurance industry the term "catastrophe" denotes a natural or man-made disaster that is unusually severe and that affects many insurers and policyholders. An event is designated a catastrophe when claims are expected to reach a certain dollar threshold. Initially the threshold was set to \$5 million. However, due to changing economic conditions, in 1997 ISO increased its dollar threshold to \$25 million. In what follows we examine the impact of the presence of left-truncation of the loss data on the resulting risk processes.

The correct estimation of the claims frequency and severity distributions is the key to determining an accurate ruin probability. A naive and possibly misleading approach for modelling the claim magnitudes would be to fit the *unconditional* distributions. Since the lower quantiles of the actual catastrophe data are truncated from the available data set, ignoring the (non-randomly) missing data would result in biased estimates of the parameters leading to over-stated mean and understated variance estimates, and under-estimated upper quantiles, in general. Furthermore, treating the available frequency data as complete, results in under-estimated intensity of the events (for example, in compound Poisson processes for aggregated insurance claims). One serious implication of such data misspecification could be wrong (under-estimated) ruin probabilities for the compound risk process. The estimation technique for loss data truncated from below is also useful when dealing with excess-of-loss reinsurance coverage where the data generally exceeds some underlying retention, see Klugman et al. (1998) and Patrik (1981).

The paper is organized as follows. In Section 2 we give a brief overview of the insurance risk model and present a methodology of treating the loss data samples with non-randomly missing observations in which the number of missing data points is unknown. Necessary adjustments to the loss and frequency distributions are discussed. In Section 3 we examine the theoretical aspects of the effects of such adjustment procedures to the severity and frequency distributions from Section 2 on the ruin probabilities. In Section 4 we present an extensive empirical study for the 1990-1999 U.S. natural catastrophe data. In this section we model the incidence of events with a non-homogeneous Poisson process and consider various distributions to fit the claim amounts. We then conduct the goodness-of-fit tests – in-sample and out-of-sample, select most adequate models, and examine the effects of model misspecification on the ruin probabilities. Finally, an additional forecasting methodology based on the robust statistics is proposed. Section 5 concludes and states final remarks.

2 Catastrophe insurance claims model

2.1 Problem description

A typical model for insurance risk, the so-called collective risk model, has two main components: one characterizing the frequency (or incidence) of events and another describing the severity (or size or amount) of gain or loss resulting from the occurrence of an event (Panjer & Willmot 1992). The stochastic nature of both the incidence and severity of claims are fundamental components of a realistic model. Hence, claims form the aggregate claim process

$$S_t = \sum_{k=1}^{N_t} X_k, \tag{1}$$

where the claim severities are described by the random sequence $\{X_k\}$ and the number of claims in the interval $(0, t]$ is modelled by a point process N_t, often called the claim arrival process. It is reasonable in many practical situations to consider the point process N_t to be a non-homogeneous Poisson process (NHPP) with a deterministic intensity function $\lambda(t)$. We make such an assumption in our paper.

The risk process $\{R_t\}_{t\geq 0}$ describing the capital of an insurance company is defined as:

$$R_t = u + c(t) - S_t. \tag{2}$$

The non-negative constant u stands for the initial capital of the insurance company. The company sells insurance policies and receives a premium according to $c(t)$. In the non-homogeneous case it is natural to set

$$c(t) = (1+\theta)\mu \int_0^t \lambda(s)ds, \tag{3}$$

where $\mu = E(X_k)$ and $\theta > 0$ is the relative safety loading which 'guarantees' survival of the insurance company.

In examining the nature of the risk associated with a portfolio of business, it is often of interest to assess how the portfolio may be expected to perform over an extended period of time. One approach concerns the use of ruin theory (Grandell 1991). Ruin theory is concerned with the excess of the income $c(t)$ (with respect to a portfolio of business) over the outgo, or claims paid, $S(t)$. This quantity, referred to as insurer's surplus, varies in time. Specifically, ruin is said to occur if the insurer's surplus reaches a specified lower bound, e.g. minus the initial capital. One measure of risk is the probability of such an event, clearly reflecting the volatility inherent in the business. In addition, it can serve as a useful tool in long range planning for the use of insurer's funds.

The ruin probability in finite time T is given by

$$\psi(u,T) = \mathrm{P}\left(\inf_{0<t<T}\{R_t\} < 0\right). \tag{4}$$

Most insurance managers will closely follow the development of the risk business and increase the premium if the risk business behaves badly. The planning horizon may be thought of as the sum of the following: the time until the risk business is found to behave "badly", the time until the management reacts and the time until a decision of a premium increase takes effect. Therefore, in non-life insurance, it is natural to regard T equal to four or five years as reasonable (Grandell 1991). We also note that the ruin probability in finite time can always be computed directly using Monte Carlo simulations. Naturally, the choice of the intensity function and the distribution of claim severities heavily affects the simulated values and, hence, the ruin probability.

We denote the distribution of claims by F_γ and its probability density function by f_γ. The loss distribution F_γ is assumed to belong to a parametric family of continuous probability distributions. Depending on the distribution, γ is a parameter vector or a scalar; for simplicity, we will refer to it as a parameter throughout the paper. We assume that the family of distributions is sufficiently well behaved so that the parameter γ can be estimated consistently by maximum likelihood. To avoid the possibility of negative losses we restrict the distribution to be concentrated on the positive half line. Independence between frequency and severity distributions is generally assumed (we assume it also in this paper). The process N_t uniquely governs the frequency of the loss events, and the distribution F_γ controls the loss severity.

Given a sample $\mathbf{x} = (x_1, x_2 \ldots, x_n)$ containing all losses which have occurred during a time interval $[T_1, T_2]$, the task of estimating γ can be (but is not limited to) performed with the maximum likelihood estimation (MLE) principle:

$$\hat{\gamma} = \hat{\gamma}_{\mathrm{MLE}}(x) = \arg\max_{\gamma} \sum_{k=1}^{n} \log f_\gamma(x_i) \tag{5}$$

and $\lambda(t)$ can be estimated by fitting a deterministic function to the loss frequency process via, for example, a least squares procedure (see Section 4.2) Note that for a homogeneous Poisson process (HPP) we can apply the MLE and obtain $\hat{\lambda} = \hat{\lambda}_{\mathrm{MLE}}(x) = n/(T_2 - T_1)$. In reality not all insurance losses over a certain time interval are recorded accurately. In the framework of catastrophic losses, the losses of magnitudes not exceeding \$5 million (until 1996) or \$25 million (since 1997) are not recorded in the databases.

The problem of catastrophe insurance claims data thus lies in the presence of *non-randomly missing data* on the left side of the loss distribution. The question addressed in the subsequent analysis is whether ignoring the missing data has a significant impact on the estimation of the intensity function $\lambda(t)$ and the severity parameter γ, and hence the ruin probability estimates. From the statistical viewpoint, with non-randomly missing data, all estimates would be biased if the missing data is not accounted for. However in practical applications a possible rationale for ignoring the missing data would be as follows: since the major part of catastrophic insurance losses is in excess of the \$5 and even the \$25 million threshold, then losses below it can not have a significant impact on the ruin probabilities, that is largely determined by the *upper* quantiles of the loss distribution.

2.2 Estimation of loss and frequency distributions

In this section we present a procedure for a consistent estimation of the claims size distribution. Given a time interval $[T_1, T_2]$ – the sample window – the collected data which is available for estimating $\lambda(t)$ and γ, is considered *incomplete*: there exists one non-negative pre-specified threshold $H \geq 0$, that defines a partition on $\mathbb{R}_{\geq 0}$: $[0, H]$ and (H, ∞). If a random outcome of the loss distribution belongs to $[0, H]$ then it does not enter the data sample: neither the frequency nor the severity of losses not exceeding H are recorded (missing data). Realizations in (H, ∞) are fully reported, i.e. both the frequency and the loss amounts are specified. Hence, we are dealing with *truncated data*.

Let the observed sample in $[T_1, T_2]$ be of the form $\mathbf{x}^o = (x_1, x_2, \ldots, x_n) \in (H, \infty)$, where n denotes the number of observations in (H, ∞) and $x_1, x_2, \ldots, x_n$ the values of the observations. The corresponding sample space is denoted as $\mathcal{X}$. Given that the total number of observations in the complete sample is unknown, the *joint* density on $\mathcal{X}$ (with respect to the product of counting and Lebesgue measures) which is consistent with the model specification in equation (1), can be given by the following expression[1]:

$$g_{\lambda(\triangle T),\gamma}(\mathbf{x}^o) = \frac{1}{n!}\left(\int_{T_1}^{T_2} \lambda^o(t)dt\right)^n \exp\left\{-\int_{T_1}^{T_2} \lambda^o(t)dt\right\} \prod_{k=1}^{n} \frac{f_\gamma(x_k)}{1 - F_\gamma(H)} \tag{6}$$

where $F_\gamma(H)$ denotes the probability for a random realization to fall into the interval $[0, H]$. By representation (6), the Poisson process $N^o_{\triangle T}$ that counts only the losses of magnitudes greater than H is interpreted as a *thinning* of the original (complete) process $N^c_{\triangle T}$ governed by $\lambda^c(t)$, with a new intensity (rate) function $\lambda^o(t) = (1 - F_\gamma(H))\lambda^c(t)$. The time frame is $\triangle T = T_2 - T_1$. The superscripts "o" and "c" refer to "observed" (the incomplete data set), and "complete" or "conditional" (the complete data set), respectively. The maximization of the corresponding log-likelihood function of the compound process is done only with respect to γ:

$$\widehat{\gamma^c}_{\text{MLE}} = \arg\max_\gamma \log\left(\prod_{k=1}^{n} \frac{f_\gamma(x_k)}{1 - F_\gamma(H)}\right), \tag{7}$$

In this study, the estimation of the intensity function does not require MLE. It is obtained directly by fitting a deterministic function to the aggregated numbers of events per unit interval over the time frame of interest. The true intensity of the complete data set is obtained by $\widehat{\lambda^c}(t) = \widehat{\lambda^o}(t)/(1 - \widehat{F}_{\gamma^c}(H))$. The unknown value $F_{\gamma^c}(H)$ needs to be estimated from (7). Such an adjustment allows to 'add back' the fraction of the missing data. Certainly, $F_{\gamma^c}(H)$ would vary for various distributions. The crucial assumption of such an amendment is that the underlying distribution of the loss magnitudes is indeed the true distribution.

In the cases where no closed-form expression for the MLE estimate of γ in expression (7) is available, as applies to most of the distributions considered in the empirical part of this paper, we have to solve for it numerically via direct numerical optimization. However, in the cases where a closed-form expression for both the unconditional MLE estimate of γ as well as the expectations $E_\gamma(\log f_\gamma(\mathbf{x}^o))$ and $E_\gamma(\log f_\gamma(\mathbf{x}^c))$ for given value of γ are available (e.g.

[1] Other model specifications, such as a renewal process or a Cox process are also possible, but not considered in this study.

for distributions such as Gaussian, Lognormal, Exponential, 1-parameter Pareto), it might be optimal to apply the *Expectation-Maximization* algorithm (EM-algorithm), see Dempster et al. (1977). The EM-algorithm is designed for maximum likelihood estimations with incomplete data. It has been also used in a variety of applications such as probability density mixture models, hidden Markov models, operational risk, cluster analysis, survival analysis, and image processing. References include Bee (2005), Bierbrauer et al. (2004), Chernobai et al. (2005a), Chernobai et al. (2005c), Figueiredo & Nowak (2003), McLachlan & Krishnan (1997), and Meng & van Dyk (1997), among many others. Surprisingly, for the distributions under study (Lognormal and Exponential) no advantage over direct integration in either speed or accuracy was observed. Hence, only the latter method was used for calibration throughout the paper.

3 Impact of density misspecification on the ruin probability

We examine two possible approaches insurance companies may undertake for the parameter estimation and subsequently the ruin probability determination.

1 The first, and correct, approach involves finding the estimates $\widehat{\lambda^c}(t)$ and $\widehat{\gamma^c}_{\mathrm{MLE}}$ for the unknown function $\lambda^c(t)$ and parameter γ^c with the direct numerical optimization (or the EM-algorithm) and determine the ruin probability. For simulations, losses can be drawn from the distribution with the complete-data estimated parameters, and use the complete-data frequency parameter.

2 An alternative, but naive approach would be to use the observed frequency estimate $\widehat{\lambda^o}(t)$ and fit the unconditional distribution to the truncated data. Evidence from the literature indicates that this approach has been widely used in practice.

For reinsurance, when only large losses matter, a third approach could be relevant. It involves estimating the frequency function $\lambda(t)$ by the observed frequency $\widehat{\lambda^o}(t)$, and estimating the complete-data conditional distribution. For the calculation of the ruin probabilities with the Monte Carlo method (which is most commonly used), losses above the threshold should be simulated from the conditional distribution, and the observed frequency should be used.

In our subsequent analysis and estimations of the ruin probabilities, we compare approaches 1 and 2. The second oversimplified and misspecified approach will lead to biased estimates for the rate function $\lambda(t)$ and the parameter γ of the loss distribution. The bias can be expressed analytically. For a Lognormal $\mathcal{LN}(\mu,\sigma)$ loss distribution, for example, the bias is expressed as follows:

$$\begin{aligned}
E\lambda^o(t) &= \lambda^c(t)\cdot(1-F_{\gamma^c}(H)) = \lambda^c(t)+\mathrm{bias}(\lambda^o(t))\\
&= \lambda^c(t)\cdot\left(1-\Phi\left(\frac{\log H-\mu^c}{\sigma^c}\right)\right),\\
E\mu^o &= E\left(\frac{1}{n}\sum_{k=1}^{n}\log X_k \,|\, X_k>H\right) = \mu^c+\mathrm{bias}(\mu^o)\\
&= \mu^c+\sigma^c\cdot\frac{\varphi\left(\frac{\log H-\mu^c}{\sigma^c}\right)}{1-\Phi\left(\frac{\log H-\mu^c}{\sigma^c}\right)},\\
E(\sigma^o)^2 &= E\left(\frac{1}{n}\sum_{k=1}^{n}\log^2 X_k-(\mu^o)^2 \,|\, X_k>H\right)\\
&= (\sigma^c)^2+\mathrm{bias}((\sigma^o)^2)\\
&= (\sigma^c)^2\left(1+\frac{\log H-\mu^c}{\sigma^c}\cdot\frac{\varphi\left(\frac{\log H-\mu^c}{\sigma^c}\right)}{1-\Phi\left(\frac{\log H-\mu^c}{\sigma^c}\right)}\right.\\
&\quad\left.-\left(\frac{\varphi\left(\frac{\log H-\mu^c}{\sigma^c}\right)}{1-\Phi\left(\frac{\log H-\mu^c}{\sigma^c}\right)}\right)^2\right),
\end{aligned}$$

where φ and Φ denote the density and d.f. of the standard Normal law and μ^c and σ^c are the true (complete data) Lognormal parameters.

In the first expression above, the oversimplified approach leads to a misspecification bias: bias$(\widehat{\lambda^o(t)})$ that will be less than 0 always. Since the bias of the location parameter $\widehat{\mu^o}$ is always positive, then the observed $\widehat{\mu^o}$ is always overstated. For practical purposes, since $\log H < \widehat{\mu^c}$ (the threshold level is relatively low), then the bias of the scale parameter is negative, and so the true $(\sigma^c)^2$ is underestimated under the unconditional fit. The effect (increase or decrease) on the ruin probability would depend on the values of H, $\widehat{\mu^c}$ and $(\widehat{\sigma^c})^2$.

4 Empirical analysis of catastrophe data

We take for our study the PCS (Property Claim Services) data covering losses resulting from natural catastrophe events in USA that occurred between 1990 and 1999. The data were adjusted using the Consumer Price Index provided by the U.S. Department of Labor, see Burnecki et al. (2005). These events will be used for testing our estimation approaches. For the calibration and in-sample validation we consider the following data set: All claim amounts exceeding \$25 million between 1990 and 1996. For the forecasting part of the paper we consider the losses over a three year period 1997-1999.

The goal of the subsequent empirical study is three-fold: we aim at 1^0 examining the effect of ignoring the threshold (missing data) on distributional parameters, 2^0 obtain the best model via the goodness-of-fit tests, and 3^0 examine the effect of the data misspecification (from part 1^0) on ruin probability under the threshold \$25 million.

4.1 Loss distributions

The following distributions for severity are considered in the study:

Exponential	$\mathcal{E}xp(\beta)$	$f_X(x) = \beta e^{-\beta x}$ $x \geq 0,\ \beta > 0$
Lognormal	$\mathcal{LN}(\mu, \sigma)$	$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}x} \exp\ -\frac{(\log x - \mu)^2}{2\sigma^2}$ $x \geq 0,\ \mu, \sigma > 0$
Gamma	$\mathcal{G}am(\alpha, \beta)$	$f_X(x) = \frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)} \exp\{-\beta x\}$ $x \geq 0,\ \alpha, \beta > 0$
Weibull	$\mathcal{W}eib(\beta, \tau)$	$f_X(x) = \tau\beta x^{\tau-1} \exp\{-\beta x^\tau\}$ $x \geq 0,\ \beta, \tau > 0$
Burr	$\mathcal{B}urr(\alpha, \beta, \tau)$	$f_X(x) = \tau\alpha\beta^\alpha x^{\tau-1}(\beta + x^\tau)^{-(\alpha+1)}$ $x \geq 0,\ \alpha, \beta, \tau > 0$
Generalized Pareto	$\mathcal{GPD}(\xi, \beta)$	$f_X(x) = \beta^{-1}(1 + \xi x \beta^{-1})^{-(1+\frac{1}{\xi})}$ $x \geq 0,\ \beta > 0$
log-αStable	$\log \mathcal{S}_\alpha(\beta, \sigma, \mu)$	no closed-form density $\alpha \in (0, 2),\ \beta \in [-1, 1],\ \sigma, \mu > 0$

In Table 1 we demonstrate the change in the parameter values when the conditional (truncated) distribution is fitted instead of the unconditional. The location parameters are lower and the scale parameters are higher under the correct data specification. In addition, the shape parameter which is present in the relevant distributions is lower (except for log-αStable) under the conditional fit, indicating a heavier tailed true distribution for the claim size data. The log-likelihood values (denoted as l) are higher under the conditional fit, except for the Burr distribution for which parameter estimates appear highly sensitive to the initial values of the computation procedure.

	γ,$F(H)$,l	Unconditional	Conditional
Exponential	β	$2.7912{\cdot}10^{-9}$	$3.0006{\cdot}10^{-9}$
	$F(H)$	6.74%	7.23%
	l	-4594.7	-4579.2
Lognormal	μ	18.5660	17.3570
	σ	1.1230	1.7643
	$F(H)$	8.63%	42.75%
	l	-4462.4	-4425.0
Gamma	α	0.5531	$2.155{\cdot}10^{-8}$
	β	$1.5437{\cdot}10^{-9}$	$0.8215{\cdot}10^{-9}$
	$F(H)$	18.34%	≈100%
	l	-4290.6	-4245.6
Weibull	β	$2.8091{\cdot}10^{-6}$	0.0187
	τ	0.6663	0.2656
	$F(H)$	21.23%	82.12%
	l	-4525.3	-4427.1
Pareto (GPD)	ξ	-0.5300	-0.8090
	β	$1.2533{\cdot}10^{8}$	$0.5340{\cdot}10^{8}$
	$F(H)$	17.27%	32.77%
	l	-4479.2	-4423.0
Burr	α	0.1816	0.1748
	β	$3.0419{\cdot}10^{35}$	$1.4720{\cdot}10^{35}$
	τ	4.6867	4.6732
	$F(H)$	2.58%	3.87%
	l	-4432.3	-4434.3
log-αStable	α	1.4265	1.9165
	β	1	1
	σ	0.5689	0.9706
	μ	18.8584	17.9733
	$F(H)$	0.005%	23.27%
	l	-438.1	-360.6

Table 1: Estimated parameters, $F(H)$ and log-likelihood values of the fitted distribution to the PCS data. For log-αStable, l are based on log-data.

The estimated fraction of the missing data $F(H)$ is larger under the conditional fit, as expected. This could be considered as evidence for the fact that conditional estimation accounts for true 'information loss' while the unconditional fit underestimates the fraction of missing data. We point out that the estimates of $F(H)$ are explicitly dependent on the choice of the distribution (and certainly the threshold H). Further, we find that for the light-tailed Exponential distribution the estimated fraction of data below threshold H is almost negligible while for more heavy-tailed distributions like Weibull or

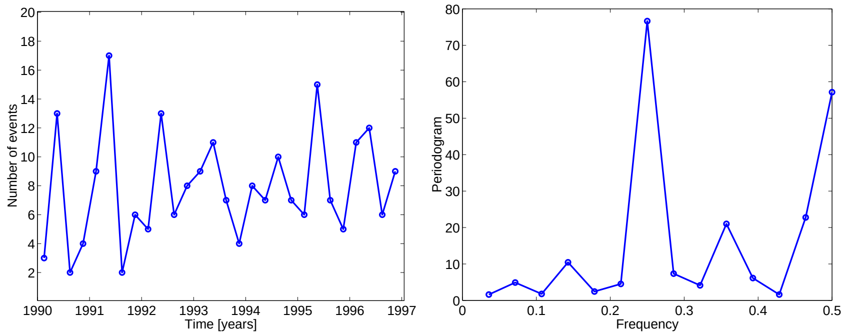

Figure 1: *Left panel*: The quarterly number of losses for the PCS data. *Right panel*: Periodogram of the PCS quarterly number of losses, 1990-1996. A distinct peak is visible at frequency $\omega = 0.25$ implying a period of $1/\omega = 4$ quarters, i.e. one year.

log-αStable the estimated fraction is significantly higher for the conditional case. The results are consistent with the findings in Chernobai et al. (2005c) with operational loss data.

For the purpose of our subsequent analysis, we decided to exclude the Gamma distribution for the following reasons. The Gamma distribution produced the true 'information loss' nearly equal to 100%, which means that if Gamma is the true distribution for the data, then nearly all data is considered missing, which is unfeasible. The true estimate of the intensity rate would blow up to infinity.

4.2 Intensity function

We model the frequency of the losses with a NHPP, in which the intensity of the counting process varies with time. The time series of the quarterly number of losses does not exhibit any trends but an annual seasonality can be very well observed using the periodogram, see Figure 1. This suggests that calibrating a NHPP with a sinusoidal rate function would give a good model. Following Burnecki & Weron (2005) we estimate the parameters by fitting the cumulative intensity function, i.e. the mean value function $\mathrm{E}(N_t)$, to the accumulated quarterly number of PCS losses. The least squares estimation is used to calibrate $\lambda(t) = a + b \cdot 2\pi \cdot \sin\{2\pi(t-c)\}$ yielding parameters a, b and c displayed in Table 2. This form of $\lambda(t)$ gives a reasonably good fit measured by the mean square error MSE $= 18.9100$, and the mean absolute error MAE $= 3.8385$. It is notable, that if, instead, a homogeneous Poisson process (HPP) with a constant intensity was considered for the quarterly number of losses, then the respective error estimates would yield MSE = 115.5730 and MAE = 10.1308. The latter values are based on the Poisson parameter estimated to be $\lambda = 33.0509$ for the data set, obtained by fitting the Exponential distribution to the respective inter-arrival times, in years. Alternatively, the mean annual number of losses, can be obtained by multiplying the quarterly number of points by four and averaging, yielding 31.7143. These result in MSE = 38.2479 and MAE = 5.3878. In either case, significantly higher values for MSE and MAE under HPP, lead to the conclusion that NHPP with the intensity rate of a functional form described above, results in a reasonably superior calibration of the loss arrival processes.

$\boldsymbol{a}$	$\boldsymbol{b}$	$\boldsymbol{c}$	MSE	MAE
30.8750	1.6840	0.3396	18.9100	3.8385

Table 2: Fitted sinusoidal function to the catastrophe loss frequency data.

To adjust for the missing data, we adjust the parameters a, b and c, according to the procedure described in Section 2.2. Using the estimates of the missing data, $F(H)$, from Table 1, straightforward calculations result in the conclusion that the true frequency of the loss events is highly underestimated.

4.3 Backtesting

In this section of our empirical study we aim at determining which of the considered distributions is most appropriate to use for the catastrophe loss data. The ultimate choice of a model can be determined via backtesting. We conduct two types of test: in-sample and out-of-sample goodness-of-fit tests.

4.3.1 In-Sample Goodness-of-Fit Tests

We test a composite hypothesis that the empirical d.f. belongs to an entire family of hypothesized truncated distributions. After necessary adjustments for the missing data, the hypotheses are summarized as:

$$\begin{array}{l} H_0: \ F_n(x) \in \widehat{F}(x) \\ H_A: \ F_n(x) \notin \widehat{F}(x), \end{array} \tag{8}$$

where $F_n(x)$ is the empirical d.f. and $\widehat{F}(x)$ is the fitted d.f. estimated for this truncated sample as:

$$\widehat{F}(x) = \begin{cases} \frac{F_{\gamma^c}(x) - F_{\gamma^c}(H)}{1 - F_{\gamma^c}(H)} & x > H \\ 0 & x \le H, \end{cases} \tag{9}$$

We consider four kinds of statistics for the measure of the distance between the empirical and hypothesized d.f.: Kolmogorov-Smirnov (D), Kuiper (V), Anderson-Darling (A^2) and Cramér-von Mises (W^2), computed as

$$D = \max(D^+, D^-), \tag{10}$$

$$V = D^+ + D^-, \tag{11}$$

$$A^2 = n \int_{-\infty}^{\infty} \frac{(F_n(x) - \widehat{F}(x))^2}{\widehat{F}(x)(1 - \widehat{F}(x))} d\widehat{F}(x), \tag{12}$$

$$W^2 = n \int_{-\infty}^{\infty} (F_n(x) - \widehat{F}(x))^2 d\widehat{F}(x), \tag{13}$$

where $D^+ = \sqrt{n} \sup_x \{F_n(x) - \widehat{F}(x)\}$ and $D^- = \sqrt{n} \sup_x \{\widehat{F}(x) - F_n(x)\}$. For numerical implementation of the above formulas consult e.g. Burnecki et al. (2005). Since the limiting distributions of the test statistics are not parameter-free, the p-values and the critical values were obtained from Monte Carlo simulations. For the truncated case, to account for the missing data the corresponding statistics were computed and the simulations were carried out according to the procedure described in Chernobai et al. (2005b). The results are presented in Table 3.

It is evident that under the data misspecification (unconditional fit), none of the considered distributions – except for Burr to some extent – appear to provide a good fit, as indicated by the near-zero p-values.[2] Note that the log-αStable distribution's fit is poor under the unconditional fit, because, despite the high p-values, the observed test statistic values are very high compared to the rest of the distributions (except the Exponential under which the test statistic values are very high and the fit is poor).

With the truncated calibration, Burr and Pareto distributions show a very good fit around both the median and the tails as indicated by the low statistic values and high p-values. The fit of the light-tailed Exponential distribution is again poor, on the basis of which we exclude it from further analysis. The Weibull distribution provides moderately low values of the test statistics, however in terms of p-values the results are rather poor. As for log-αStable and Lognormal, relatively low observed statistic values and often high p-values suggest an acceptable fit. In general, the data seems to follow a very heavy-tailed law.

Table 3 figures lead us to the conclusion that modelling claim sizes with unconditional loss distributions results in rejecting the null hypothesis of the

[2] At this point, we decided to look at whether using the correct testing methodology for the truncated samples with wrong (unconditional) parameters would result in higher p-values, which would make them more comparable to those corresponding to the conditional fit. Computational analysis indicated that the p-values were left generally almost unchanged from the left half of Table 3, and the test statistic values have increased in most cases, confirming a generally unacceptable fit of wrongly specified loss distributions.

Unconditional				Conditional			
D	V	A^2	W^2	D	V	A^2	W^2
			$\mathcal{E}xp$				
5.1234	6.1868	48.9659	10.1743	5.5543	5.9282	72.2643	13.1717
[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]
			$\mathcal{LN}$				
1.5564	2.9710	4.4646	0.7139	0.6854	1.1833	0.7044	0.0912
[<0.005]	[<0.005]	[<0.005]	[<0.005]	[0.256]	[0.315]	[0.080]	[0.133]
			$\mathcal{W}eib$				
3.2755	5.5430	14.2197	2.3859	0.8180	1.5438	1.3975	0.1965
[<0.005]	[<0.005]	[<0.005]	[<0.005]	[0.107]	[0.053]	[0.006]	[0.008]
			$\mathcal{GPD}$				
2.7084	3.9240	7.6731	1.1013	0.4841	0.8671	0.3528	0.0390
[<0.005]	[<0.005]	[<0.005]	[<0.005]	[0.795]	[0.847]	[0.487]	[0.666]
			$\mathcal{B}urr$				
0.8876	1.3149	0.8983	0.1469	0.4604	0.8668	0.2772	0.0342
[0.054]	[0.106]	[0.014]	[0.022]	[0.822]	[0.793]	[0.560]	[0.659]
			$\log\mathcal{S}$				
4.9863	4.9953	94.6918	11.3252	0.8961	1.2111	0.8062	0.1535
[0.596]	[0.616]	[0.483]	[0.575]	[0.456]	[0.470]	[0.484]	[0.444]

Table 3: Results of the in-sample goodness-of-fit tests. p-values were obtained via 1,000 Monte Carlo simulations, and are given in the square brackets.

goodness of fit, for practically all considered distributions. On the basis of such conclusion, we agree that unconditional loss distributions should not be used for forecasting purposes, and only the conditional distributions should be used. Furthermore, poor fit of the truncated Exponential distribution leads us to exclude it from further consideration as a candidate for the claim size distribution. As the next step we examine and compare the forecasting power of the remaining considered truncated loss distributions.

4.3.2 Out-of-Sample Goodness-of-Fit Tests

Examining how well or how badly various considered models predict the true future losses, is, we believe, the key to determining which of the loss distributions is the best to be used for practical purposes.

For our loss distribution estimations in Section 4.1 we used the data set from 1990 to 1996. We now analyze our models' predicting power regarding the

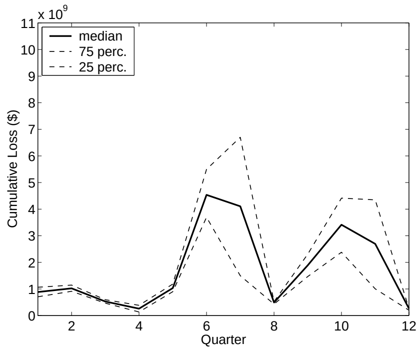

Figure 2: *Left panel*: The 25th, 50th and 75th quantiles for the PCS 1997-1999 data. *Right panel*: Actual aggregated losses (bold) and 95% bootstrap confidence interval for realized PCS 1997-1999 loss data.

data between 1997 and 1999. For the out-of-sample backtesting, we assume that our model has a one-step ahead predicting power, with one step equal to one quarter. The window length of the sample used for calibration is taken to be six years. We start with the data from the first quarter of 1990 until the fourth quarter of 1996, in order to conduct the forecasting about the first quarter of 1997. First, we estimate the unknown parameters of truncated distributions. Next, to obtain the distribution of the quarterly aggregated losses we repeat the following a large number (10,000) of times: use the estimated parameters to simulate N losses exceeding the \$25 million threshold, where N is the actual number of losses in the quarter that we perform forecasting on, and aggregate them. At each forecasting step (twelve steps total) we shift the window by one quarter forward and repeat the above procedure. Note, that in this way we test the model for the severity distribution but not for the entire risk process or the intensity itself.

We break the analysis of the forecasting results into two main parts. First, we look at how good are the different model assumptions in predicting the distribution of accumulated losses around the center, or the main body, of the actual loss distribution. Then we examine how well they predict the true aggregated losses around the tails.

At each step, we estimate the 25th, 50th and 75th percentiles of the distribution of the sum of realized losses for the corresponding quarter. This is conducted using the non-parametric bootstrapping technique. Similarly, we determine the 95% confidence interval for the quarterly sum of losses. Next, we use the previous 24 quarters of data (six years) to estimate the parameters of the conditional distributions, for the five considered cases. Based on the parameter estimates, we determine the 25th, 50th and 75th percentiles,

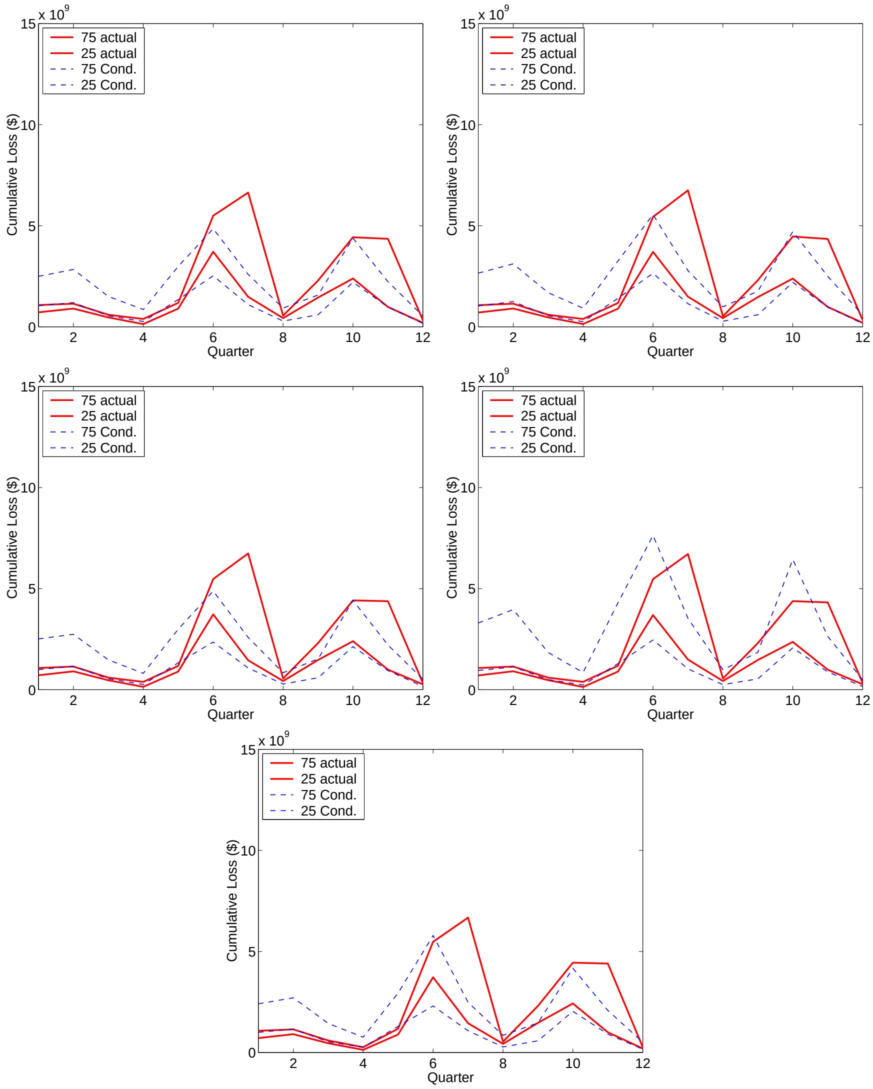

Figure 3: The 25th and 75th quantiles for conditional (dashed blue) forecasted cumulative losses and corresponding bootstrap quantiles for realized losses (solid red): Lognormal (*top left*), Weibull (*top right*), Generalized Pareto (*mid left*), Burr (*mid right*), and log-αStable (*bottom*).

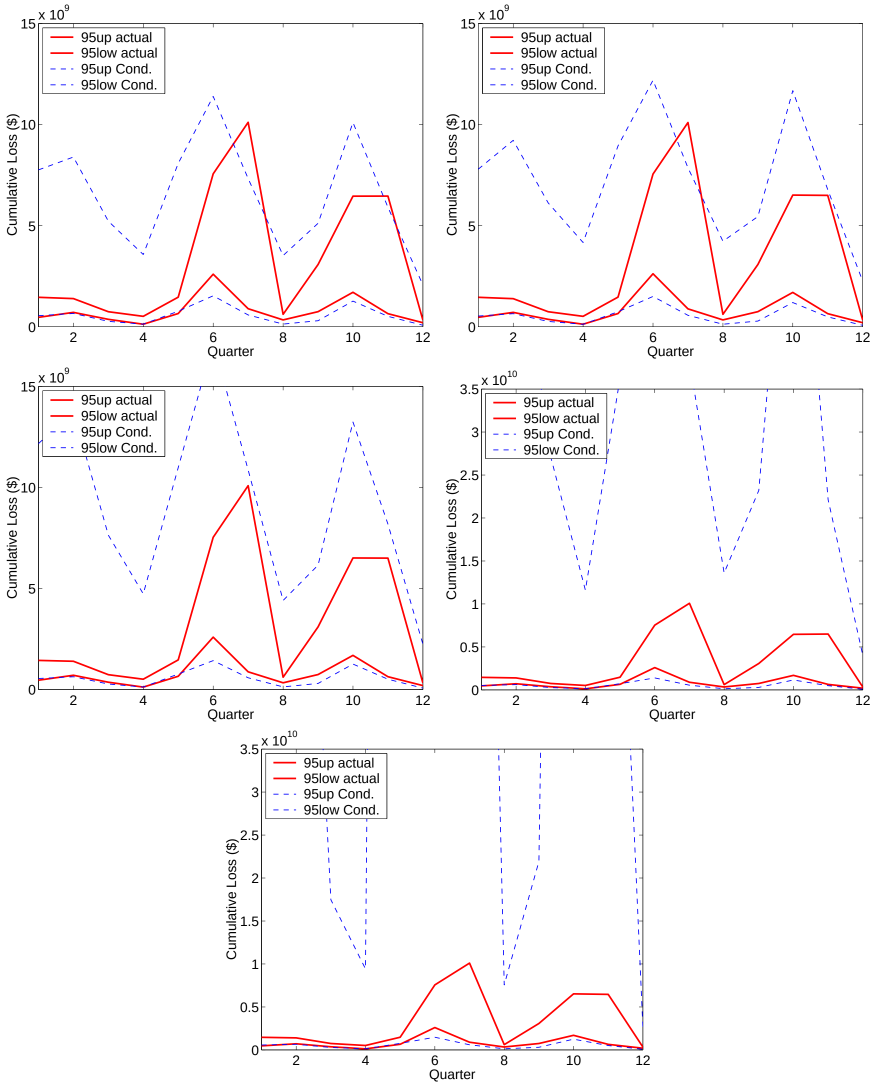

Figure 4: 95% confidence intervals for conditional (dashed blue) forecasted cumulative losses and corresponding 95% bootstrap confidence intervals for realized losses (solid red): Lognormal (*top left*), Weibull (*top right*), Generalized Pareto (*mid left*), Burr (*mid right*), and log-αStable (*bottom*).

and estimate the 95% confidence intervals (2.5th and 97.5th percentiles) for the quarterly cumulative losses, by simulating 10,000 samples (of losses that exceed the threshold of \$25 million) of the size equal to the number of realized losses in the forecasted quarter. We then compare the intervals with the bootstrapped intervals based on the actual data.

Figure 2 shows the accumulation of PCS losses above \$25 million around the center (left panel) and the tails (right panel) of their distribution, obtained via bootstrapping. Figure 3 demonstrates the 25th and 75th percentiles of the forecasted distributions, relative to the realized corresponding quantiles, obtained with the bootstrapping procedure. It is notable, that for 1997 (quarters 1-4) all models tend to overestimate the cumulative losses around the central part of the distribution, to various extents. For 1998 and 1999 (quarters 5-12) the situation changes as the empirical data seems to be more consistent with the whole sample. The Lognormal, Weibull, Pareto and Burr distributions appear to capture the spread of the central part reasonably well.

Figure 4 portrays the forecasting ability of the considered five models around the tails of the actual loss distribution. Again, in 1997, all considered models tend to overpredict the aggregated losses. As for the remaining part of the test period, Lognormal and Weibull assumptions generally result in quite accurate forecasted 95% confidence interval estimates. As for Pareto and Burr, the forecasted 95% confidence intervals' upper bounds are much higher than the true bootstrapped bounds. The effect is even more significant under the log-αStable law. Hence, the ruin probabilities may be unnecessarily high under these assumptions. Overall, the figures indicate that Lognormal and Weibull assumptions possess a reasonably good predicting power.

We also use the 10,000 Monte Carlo generated samples to compute the MSE and MAE for the estimates of the sum of losses, with respect to the true realized losses of each quarter. We present the estimates of the average MSE and MAE over the 3-year forecasting time interval in Table 4. The results confirm our conclusions. In terms of the average MSE and MAE, the Lognormal, Weibull and Pareto distributional assumptions result in relatively accurate predictions regarding the aggregated losses one quarter ahead, for the considered three year period. In particular, the Weibull model is most optimal in terms of the MSE and the Lognormal model appears most optimal in terms of the MAE. The errors under the Burr and log-αStable assumptions are considerably higher.

Summarizing, based on the in-sample goodness of fit tests, the observed test statistic values are the lowest for the GPD and Burr laws, indicating their superior in-sample fit. However, the forecasting analysis rather supports the use of the less heavy-tailed Lognormal and Weibull model, as is indicated by the smallest error estimates. Since Lognormal also provided an acceptable in-

	$\overline{\text{MSE}}$	$\overline{\text{MAE}}$
Lognormal	$6.8803 \cdot 10^{18}$	$1.2969 \cdot 10^{9}$
Weibull	$5.7869 \cdot 10^{18}$	$1.3745 \cdot 10^{9}$
Pareto (GPD)	$3.2915 \cdot 10^{20}$	$1.9123 \cdot 10^{9}$
Burr	$1.2562 \cdot 10^{24}$	$1.4986 \cdot 10^{10}$
log-αStable	$1.8654 \cdot 10^{89}$	$1.2468 \cdot 10^{42}$

Table 4: Estimates for the average MSE and MAE of forecasted aggregated losses above \$25 million, obtained via 10,000 Monte Carlo simulations.

sample fit it could be considered as the overall most appropriate distribution.

It is notable that, although Pareto, Burr and log-αStable distributions showed acceptable in-sample fit, they highly overpredict the actual losses in the forecasting three year period. This is due to the following: in the data sample used for calibration, there were few losses whose magnitude was of significantly higher order than the rest of the data. Such losses can be categorized as 'low frequency/ high severity' losses. The two largest losses and their magnitudes (after the inflation adjustment) were as follows:

1. Hurricane 'Andrew' that hit Florida and Louisiana on August 24-26, 1992; appr. \$18.5 billion losses;
2. Northridge earthquake that struck in California on January 17, 1994; appr. \$14.4 billion losses.

The latter loss entered all twelve shifted sample windows used for forecasting the quarterly losses, while the former entered the first seven. As a consequence very heavy-tailed distributions yielded good in-sample fits. However, since losses of a comparable magnitude were not present in the 1997-1999 sample, this resulted in an overestimation of the confidence intervals' upper bounds.

4.4 Ruin Probability

We consider a hypothetical scenario where the insurance company insures losses resulting from catastrophic events in the United States. The company's initial capital is assumed to be $u = \$10$ billion and the relative safety loading used is $\theta = 30\%$. We choose different models of the risk process whose application is most justified by the statistical results described above. We decided to exclude both the Burr and the log-αStable distributions as they

	Unconditional	Conditional	Increase (appr. # times)
Lognormal	0.00545	0.10443	19
	[0.00122]	[0.00197]	
Weibull	0.00754	0.10785	14
	[0.00799]	[0.00317]	
Pareto (GPD)	0.07938	0.15997	2
	[0.00238]	[0.00418]	

Table 5: 5-year ruin probability estimates and the degree of increase for $u = \$10$ billion and $\theta = 30\%$, based on 10×10,000 Monte Carlo simulations. The standard errors are indicated in the square brackets.

	Unconditional	Conditional	Increase (appr. # times)
Lognormal	0.00669	0.13137	20
	[0.00092]	[0.00205]	
Weibull	0.00874	0.13077	15
	[0.00056]	[0.00142]	
Pareto (GPD)	0.10376	0.20434	2
	[0.00197]	[0.00428]	

Table 6: 10-year ruin probability estimates and the degree of increase for $u = \$10$ billion and $\theta = 30\%$, based on 10×10,000 Monte Carlo simulations. The standard errors are indicated in the square brackets.

produce an infinite first moment, making them inapplicable for the purpose of estimating the ruin probability (recall the premium formula (3)). This leaves us with three models: Lognormal, Weibull and GPD (Pareto).

In this paper, the ruin probability is approximated by means of Monte Carlo simulations. For the Monte Carlo method purposes we generated 10×10,000 simulations. Estimates for the 5-year and 10-year ruin probabilities are demonstrated in Tables 5 and 6. We recall that the 'unconditional' case refers to the naive approach when neither the severity nor the frequency distributions were adjusted for the 'information loss' due to the missing data. 'Conditional' case refers to the correct data specification, in which truncated loss distributions were used and the frequency was adequately adjusted, leading to more accurate estimates of the finite-time ruin probabilities. The standard errors were computed from all 10×10,000 simulations.

Tables 5 and 6 demonstrate the effects on the ruin probability due to the data (mis)specification. Based on the results with the considered three distributions, the ruin probability tends to be significantly underestimated when the naive approach is used (i.e. when unconditional distribution is fitted to the loss magnitudes, and the frequency rate function estimates are left unadjusted) instead of the correct approach, in which both the frequency and severity account for the missing data.

4.5 Robust Estimation Approach

We briefly discussed the reasons why in-sample goodness of fit tests favor heavy-tailed distributions (e.g. Pareto has high p-values and relatively low test statistic values), while more moderately heavy-tailed distributions such as Weibull or Lognormal have a better forecasting power as heavier-tailed distributions tend to overestimate the true losses. As was pointed out, two losses of magnitudes \$18.5 and \$14.4 billion are present in the data used for the in-sample testing, while comparable magnitude losses did not enter the data used for forecasting. We refer to such events as "low frequency/ high severity" events.

In recent years outlier-resistant or so-called robust estimates of parameters are becoming more wide-spread in risk management. Such models – called robust (statistics) models – were introduced by P.J.Huber in 1981 and applied to robust regression analysis, more recent references on robust statistics methods include Huber (2004), Rousseeuw & Leroy (2003), Martin & Simin (2003), Knez & Ready (1997) and Hampel et al. (1986). Robust models treat extreme data points as outliers (or some standard procedure is used to detect outliers in the data) which distort the main flow of the loss process. Under the robust approach, the focus is on modelling the major bulk of the data that is driving the entire process. Robust models help protect against the outlier bias in parameter estimates and provide with a better fit of the loss distributions to the data than under the classical model. Moreover, outliers in the original data can seriously drive future forecasts in an unwanted (such as worst-case scenario) direction, which is avoided by the robust approach models.

Following the idea of robust statistics, for the forecasting purposes we offer a second methodology that involves determining outliers and trimming the top 1-5% of the data. This data adjustment results in a more robust outlook regarding a general future scenario. Excluding the outliers in the original loss data noticeably improves the forecasting power of considered loss distributions, and can be used for forecasting of the generic (most likely) scenario of future losses within reasonable boundaries. The resulting ruin probabilities will be more optimistic than otherwise predicted by the classical model.

	γ,$F(H)$,l	Unconditional	Conditional
Exponential	β	$4.7177{\cdot}10^{-9}$	$5.3486{\cdot}10^{-9}$
	$F(H)$	11.13%	12.52%
	l	-4437.8	-4411.9
Lognormal	μ	18.5210	17.7897
	σ	1.0236	1.4611
	$F(H)$	7.32%	30.26%
	l	-4391.9	-4362.5
Gamma	α	0.8983	$1.6475{\cdot}10^{-6}$
	β	$4.2378{\cdot}10^{-9}$	$1.8241{\cdot}10^{-9}$
	$F(H)$	13.18%	≈100%
	l	-3829.7	-3798.6
Weibull	β	$7.8535{\cdot}10^{-8}$	$0.8615{\cdot}10^{-3}$
	τ	0.8579	0.4121
	$F(H)$	16.01%	61.84%
	l	-4431.8	-4363.1
Pareto (GPD)	ξ	-0.3717	-0.6655
	β	$1.3220{\cdot}10^{8}$	$0.6083{\cdot}10^{8}$
	$F(H)$	16.70%	30.46%
	l	-4415.7	-4362.8
Burr	α	0.2022	0.2046
	β	$6.6114{\cdot}10^{33}$	$0.1774{\cdot}10^{33}$
	τ	4.4573	4.2746
	$F(H)$	2.66%	4.25%
	l	-4374.7	-4376.9
log-αStable	α	1.4822	1.7796
	β	1	-1
	σ	0.5681	1.2292
	μ	18.7586	17.1497
	$F(H)$	0.09%	51.09%
	l	-382.5	-291.8

Table 7: Estimated parameters, $F(H)$ and log-likelihood values of the fitted distribution to the *trimmed* PCS data. For log-αStable, l are based on log-data.

In the context of this paper, the pitfall of using only the classical approach, that makes use of the entire data set, is that it leads to an upward bias in the forecasted losses and overly conservative estimates of the ruin probabilities – both phenomena observed in this paper's empirical study. The adjustments the insurers would have to incorporate into their policies (such as raising the safety loading factor θ) on the basis of such estimates may not be reasonable for the practical purposes. Practitioners are more likely to be searching for a

Unconditional				Conditional			
D	V	A^2	W^2	D	V	A^2	W^2
Exponential							
2.8508	4.5958	13.5090	2.6223	3.3324	4.0820	23.2179	4.3263
[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]	[<0.005]
Lognormal							
1.3512	2.5616	3.7007	0.5581	0.5693	0.9528	0.4288	0.0502
[<0.005]	[<0.005]	[<0.005]	[<0.005]	[0.545]	[0.685]	[0.334]	[0.460]
Weibull							
2.4880	4.5218	9.6317	1.5862	0.6615	1.1256	0.7045	0.0906
[<0.005]	[<0.005]	[<0.005]	[<0.005]	[0.334]	[0.433]	[0.102]	[0.145]
Pareto (GPD)							
2.6100	3.8319	7.0758	0.9959	0.4696	0.9133	0.3363	0.0331
[<0.005]	[<0.005]	[<0.005]	[<0.005]	[0.831]	[0.766]	[0.529]	[0.751]
Burr							
0.8914	1.3929	1.0414	0.1600	0.4799	0.9472	0.3972	0.0466
[0.043]	[0.058]	[0.005]	[0.010]	[0.757]	[0.607]	[0.274]	[0.399]
log-αStable							
4.2920	4.3149	65.8189	8.5308	1.3800	1.4265	4.4837	0.7335
[0.587]	[0.603]	[0.459]	[0.549]	[0.730]	[0.753]	[0.699]	[0.711]

Table 8: Results of the in-sample goodness-of-fit tests for the *trimmed* data. p-values in square brackets were obtained via 1,000 Monte Carlo simulations.

stable model that would capture the mainstream tendency of the loss process.

We emphasize, however, that we are not recommending the use of only one of the two approaches – classical or robust – instead of the other. Rather, in the presence of outliers, we encourage the use of both models for the analysis, and use the robust model as the complement to the classical.[3]

Here we exclude the top 1% of the data, which corresponds to the two highest losses described earlier as outliers. We reproduce Tables 1, 3, 4, 5 and 6, and Figures 3 and 4, now for the trimmed data. They are presented respectively in Tables 7, 8, 9, 10 and 11, and Figures 5 and 6. We would like to briefly comment on the results.

Table 8 provides in-sample goodness of fit results. Under the conditional ap-

[3]We wish to stress that the reason for trimming the top several percent of the data is independent from the consideration that, since the highest claims are transferred to the reinsurers, they need to be taken out from the insurers' database. This transfer of risk from the insurers to the reinsurers is not treated in this paper.

	$\overline{\text{MSE}}$	$\overline{\text{MAE}}$
Lognormal	$2.6103{\cdot}10^{18}$	$1.0635{\cdot}10^{9}$
Weibull	$2.0225{\cdot}10^{18}$	$1.0135{\cdot}10^{9}$
Pareto (GPD)	$2.3701{\cdot}10^{19}$	$1.2851{\cdot}10^{9}$
Burr	$5.9534{\cdot}10^{23}$	$1.0456{\cdot}10^{10}$
log-αStable	$2.6091{\cdot}10^{18}$	$1.0701{\cdot}10^{9}$

Table 9: Estimates for the average MSE and MAE of forecasted aggregated losses above $25 million, obtained via 10,000 Monte Carlo simulations, for the *trimmed* data.

proach, even after excluding the two highest points, the Pareto model results in the lowest distance statistics and highest p-values. Also for the heavy-tailed log-αStable p-values are quite high, followed by Burr and Lognormal, which contrary to the former also give good results for the test statistics. The fit of the Weibull distribution is not so good while the results for the Exponential remain very poor for the trimmed data. Another main observation is that the forecasts for all distributions show clearly lower MSEs and MAEs for all distributions (Table 9). While the Weibull distribution provides the best forecasts in terms of the MSE and MAE, the log-αStable and Lognormal give only slightly worse results. Like in the case of the original data the forecasts of the Burr distribution are very conservative and give the highest MSE and MAE error. As regards the finite time ruin probabilities, the figures are much lower than under the classical model with untrimmed data, even in the "conditional" case (we note that the gap between the figures under unconditional and conditional models is remarkably big), because the robust methods put more probability to the medium and small size losses and thus increase the probabilities of ruin resulting from them (Table 10 and 11).

Overall, evidence confirms the classical model result that the Lognormal assumption can be considered as plausible providing both an acceptable in-sample and good out-of-sample results. For the trimmed data, however also the log-αStable provides good results for both criteria. The Pareto model provides the best in-sample fit but higher forecast errors than Lognormal and log-αStable while Weibull is convinving in terms of forecast MSE and MAE but doesn't provide good in-sample results.

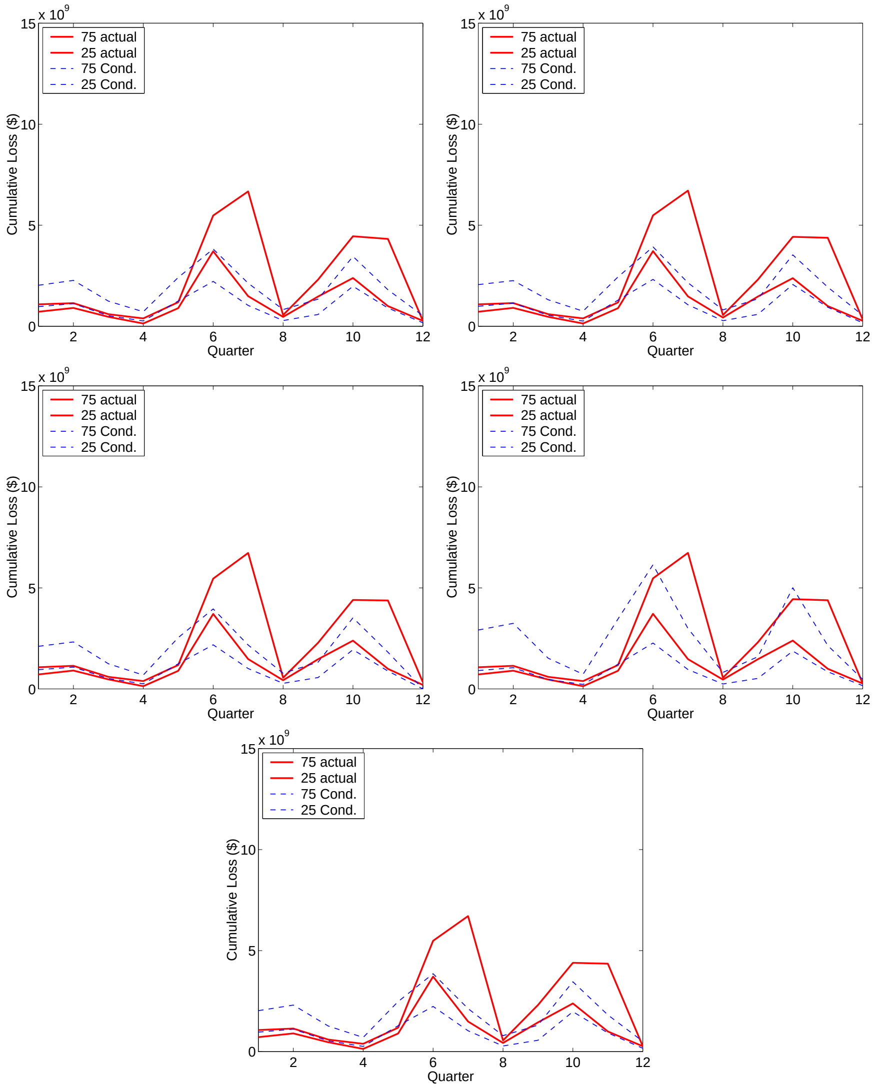

Figure 5: The 25th and 75th quantiles for conditional (dashed blue) forecasted cumulative losses and corresponding bootstrap quantiles for realized losses (solid red), for the *trimmed* data: Lognormal (*top left*), Weibull (*top right*), Generalized Pareto (*mid left*), Burr (*mid right*), and log-αStable (*bottom*).

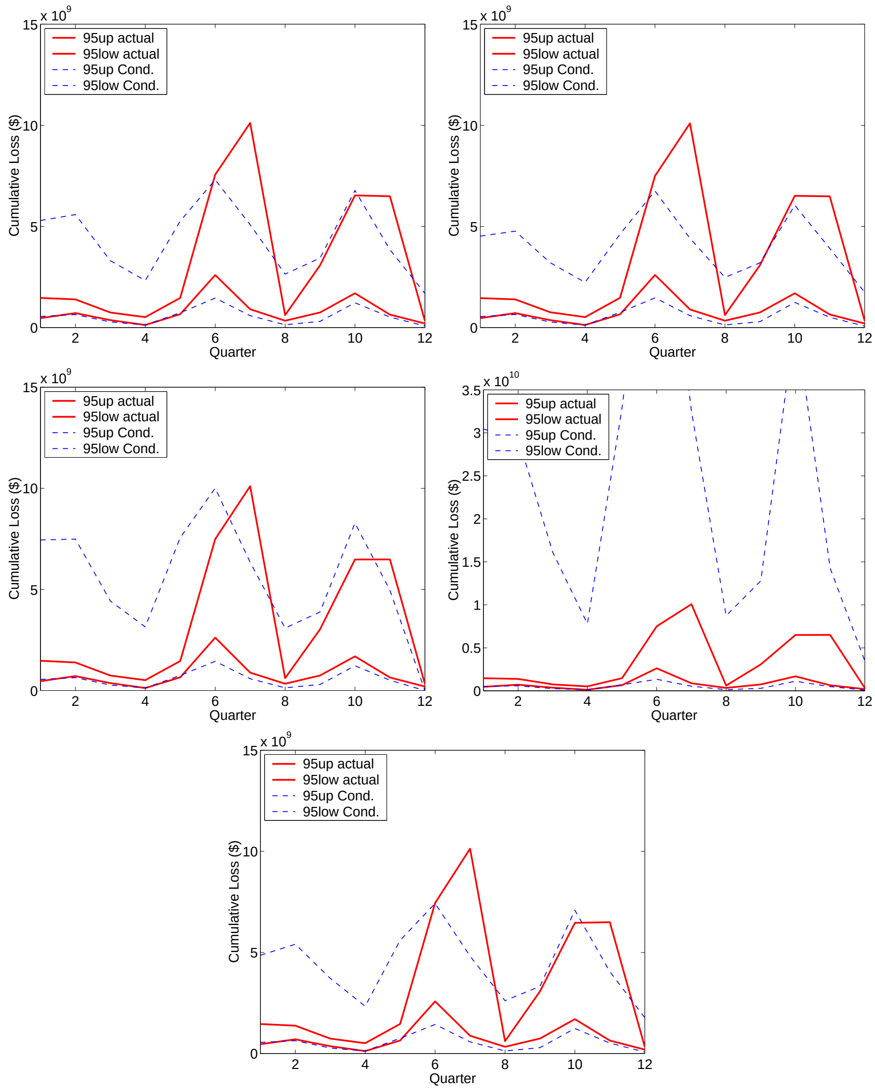

Figure 6: 95% confidence intervals for conditional (dashed blue) forecasted cumulative losses and corresponding 95% bootstrap confidence intervals for realized losses (solid red), for the *trimmed* data: Lognormal (*top left*), Weibull (*top right*), Generalized Pareto (*mid left*), Burr (*mid right*), and log-αStable (*bottom*).

	Unconditional	Conditional	Increase (appr. # times)
Lognormal	0.00077	0.03040	39.5
	[0.00035]	[0.00223]	
Weibull	0.00004	0.00702	175.5
	[0.00007]	[0.00082]	
Pareto (GPD)	0.01487	0.10434	7
	[0.00095]	[0.00281]	

Table 10: 5-year ruin probability estimates and the degree of increase for $u =$ \$10 billion and $\theta = 30\%$, based on 10×10,000 Monte Carlo simulations, for the *trimmed* data. The standard errors are indicated in the square brackets.

	Unconditional	Conditional	Increase (appr. # times)
Lognormal	0.00103	0.03719	36
	[0.00092]	[0.00205]	
Weibull	0.00006	0.00835	139
	[0.00010]	[0.00087]	
Pareto (GPD)	0.01911	0.14079	7
	[0.00132]	[0.00337]	

Table 11: 10-year ruin probability estimates and the degree of increase for $u =$ \$10 billion and $\theta = 30\%$, based on 10×10,000 Monte Carlo simulations, for the *trimmed* data. The standard errors are indicated in the square brackets.

5 Conclusions

This paper focused on analyzing the effects of data misspecification, under which the loss data available from a loss database is truncated from below at a predetermined threshold level. Such thresholds are often ignored in practice. We examined the consequences on the estimates of the ruin probabilities and the choice of claim size distribution for the catastrophic claims model. The theoretical study proposed a practical solution to the problem and suggested that using truncated (conditional) distributions instead of regular (unconditional) distributions provides for more accurate distributional parameters. In-sample goodness-of-fit tests confirmed that fitting a wrong distribution results in an unacceptable fit and the rejection of the null hypothesis that the data is drawn from a hypothesized family of distributions.

Fitting truncated distributions and using a correct testing procedure, considerably improved the goodness-of-fit. The truncated Lognormal distribution showed a good fit – both in-sample and out-of-sample (forecasting). Pareto, Burr and log-αStable showed a good in-sample fit but a poor forecasting power. The Weibull distribution provided good forecasting results but the in-sample fit of the distribution was rather poor. Overall, Lognormal was concluded to be superior on the basis of the in-sample and out-of-sample goodness of fit testing procedures.

A methodology for performing necessary adjustments to the frequency parameter of the loss events was also proposed. It was demonstrated that given the estimated fraction of missing data, the true total number of claims is significantly underestimated, if the missing data is left unaccounted for. The paper then argued and provided empirical evidence that the finite-time ruin probabilities are always seriously underestimated under the model misspecification, i.e. when unconditional distributions are wrongly used. For the several considered distributions, the true ruin probability is up to 20 times higher for 5 and 10 years, than what one would obtain using a wrong model. This has a variety of serious implications for the insurance purposes, and can be of high interest for both insurers and reinsurers. As a counter measure, for a more solid protection against the risk of ruin, they might need to consider increasing the safety loading parameter θ.

This paper finally advocates the use of the "robust" approach, as a complement to the classical approach, under which top 1% of the loss data was treated as outliers and was excluded from the database, and all model parameters were reestimated. Such models are more pertinent to the mainstream loss events and possess an advantage of a more credible forecasting capacity. The use of the robust methodology improved remarkably the forecasts. It also improved the fit of the log-αStable distribution (making it roughly the second-best candidate), and confirmed the good results of the Lognormal law.

Acknowledgements

We are thankful to S. Stoyanov of the FinAnalytica Inc. for partial computational assistance.

References

Bee, M. (2005), On maximum likelihood estimation of operational loss distributions, Technical Report 3, University of Trento.

Bierbrauer, M., Trück, S. & Weron, R. (2004), 'Modeling electricity prices with regime switching models', *Lecture Notes in Computer Science* **3039**, 859–867.

Burnecki, K., Härdle, W. & Weron, R. (2004), An introduction to simulation of risk processes, *in* J. Teugels & B. Sundt, eds, 'Encyclopedia of Actuarial Science', Wiley, Chichester.

Burnecki, K., Kukla, G. & Weron, R. (2000), 'Property insurance loss distributions', *Physica A* **287**, 269–278.

Burnecki, K., Misiorek, A. & Weron, R. (2005), Loss distributions, *in* P. Cizek, W. Härdle & R. Weron, eds, 'Statistical Tools for Finance and Insurance', Springer, Berlin.

Burnecki, K. & Weron, R. (2005), Modeling of the risk process, *in* P. Cizek, W. Härdle & R. Weron, eds, 'Statistical Tools for Finance and Insurance', Springer, Berlin.

Chernobai, A., Menn, C., Trück, S. & Rachev, S. (2005a), 'A note on the estimation of the frequency and severity distribution of operational losses', *Mathematical Scientist* **30**(2).

Chernobai, A., Rachev, S. & Fabozzi, F. (2005b), Composite goodness-of-fit tests for left-truncated loss samples, Technical report, University of California Santa Barbara.

Chernobai, A., Trück, S., Menn, C. & Rachev, S. (2005c), Estimation of operational Value-at-Risk with minimum collection thresholds, Technical report, University of California Santa Barbara.

Dempster, A., Laird, N. & Rubin, D. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society, Series B (Methodological)* **39**(1), 1–38.

Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997), *Modeling Extremal Events for Insurance and Finance*, Springer-Verlag, Berlin.

Figueiredo, M. A. T. & Nowak, R. D. (2003), 'An EM algorithm for wavelet-based image restoration', *IEEE Transactions on Image Processing* **12**(8), 906–916.

Grandell, J. (1991), *Aspects of Risk Theory*, Springer-Verlag, New York.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, R. J. & Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, John Wiley & Sons.

Huber, P. J. (2004), *Robust Statistics*, John Wiley & Sons, Hoboken.

Klugman, S. A., Panjer, H. H. & Willmot, G. E. (1998), *Loss Models: From Data to Decisions*, Wiley, New York.

Knez, P. J. & Ready, M. J. (1997), 'On the robustness of size and book-to-market in cross-sectional regressions', *Journal of Finance* **52**, 1355–1382.

Kremer, E. (1998), Largest claims reinsurance premiums for the Weibull model, *in* 'Blätter der Deutschen Gesellschaft für Versicherungsmathematik', pp. 279–284.

Madan, D. B. & Unal, H. (2004), Risk-neutralizing statistical distributions: with an application to pricing reinsurance contracts on FDIC losses, Technical Report 2004-01, FDIC, Center for Financial Research.

Martin, R. D. & Simin, T. T. (2003), 'Outlier resistant estimates of beta', *Financial Analysts Journal* **59**, 56–69.

McLachlan, G. & Krishnan, T. (1997), *The* EM *Algorithm and Extensions*, Wiley Series in Probability and Statistics, John Wiley & Sons.

Meng, X.-L. & van Dyk, D. (1997), 'The EM algorithm - an old folk-song sung to a fast new tune', *Journal of the Royal Statistical Society, Series B (Methodological)* **59**(3), 511–567.

Mittnik, S. & Rachev, S. T. (1993a), 'Modelling asset returns with alternative stable distributions', *Econometric Reviews* **12**, 261–330.

Mittnik, S. & Rachev, S. T. (1993b), 'Reply to comments on modelling asset returns with alternative stable distributions and some extensions', *Econometric Reviews* **12**, 347–389.

Panjer, H. & Willmot, G. (1992), *Insurance Risk Models*, Society of Actuaries, Schaumburg.

Patrik, G. (1981), Estimating casualty insurance loss amount distributions, *in* 'Proceedings of the Casualty Actuarial Society', Vol. 67, pp. 57–109.

Rousseeuw, P. J. & Leroy, A. M. (2003), *Robust Regression and Outlier Detection*, John Wiley & Sons, Hoboken.

SwissRe (2004), 'Sigma preliminary report'.