Pure Mathematics 理论数学, 2023, 13(3), 593-605
Published Online March 2023 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2023.133063

Hans汉斯

保险索赔金额的上尾建模：来自丹麦火灾保险数据的证据

边同鑫

南京审计大学统计与数据科学学院，江苏 南京

收稿日期：2023年2月21日；录用日期：2023年3月22日；发布日期：2023年3月30日

摘　要

为了对保险索赔数据进行建模，即找到最适合的分布，本文使用最大似然框架对丹麦火灾保险数据进行建模和分析，并在幂律分布，对数正态分布和威布尔分布这三种候选分布中找到每组数据的最适合分布。研究发现，建筑损失组数据符合对数正态分布；内容物损失组数据符合对数正态分布；利润损失组的数据很好地拟合了幂律分布，而总损失组的数据也很好地符合幂律分布。此外，我们还证明了下界的估计影响模型参数的估计和用于测试分布的*p*值。最后，我们使用拟合分布模型来获得相应的VaR，并将其与经验VaR进行比较，结果相似。这也意味着最大似然框架有助于保险索赔数据的建模。

关键词

丹麦火灾保险数据，幂律分布，对数正态分布，威布尔分布，参数估计，假设检验，阈值选择

Modelling the Upper Tail of Insurance Claim Amount: Evidence from Danish Fire Insurance Data

Tongxin Bian

School of Statistics and Data Science, Nanjing Audit University, Nanjing Jiangsu

Received: Feb. 21st, 2023; accepted: Mar. 22nd, 2023; published: Mar. 30th, 2023

Abstract

In order to model the insurance claim data, that is, to find the most suitable distribution, this paper uses the maximum likelihood framework to model and analyze the Danish fire insurance data,

文章引用: 边同鑫. 保险索赔金额的上尾建模: 来自丹麦火灾保险数据的证据[J]. 理论数学, 2023, 13(3): 593-605.
DOI: 10.12677/pm.2023.133063

and finds the most suitable distribution for each group of data among the three candidate distributions of Power law distribution, Lognormal distribution and Weibull distribution. It is found that the data of Building group is well fitted by Lognormal distribution; the data of Contents group is well fitted by Lognormal distribution; The data of Profits group is well fitted by the Power law distribution, and the data of Total group is well fitted by the Power law distribution. Moreover, we also prove that the estimation of the lower bound affects the estimation of the model parameters and the *p*-values used to test the distributions. Finally, we use the fitted distribution model to obtain the corresponding VaR and compare it with the empirical VaR, and the results are similar. This also means that the maximum likelihood framework is helpful to the modeling of insurance claim data.

Keywords

Danish Fire Insurance Data, Power Law Distribution, Lognormal Distribution, Weibull Distribution, Parameter Estimation, Hypothesis-Testing, Threshold Selection

1. 引言

在现实生活中，非人寿保险索赔的情况相对复杂，很难用共同分布对所有索赔金额数据进行建模，因为可能存在不止一个模型或未知分布,可以参见 Agnes 等人(2023) [1]。在风险研究中，保险公司不关注整体数据，而只关注可能带来不可逆转风险的部分。在非人寿保险中，保单造成的巨额损失通常占索赔总额的很大比例如。如果发生重大损失，将带来严重后果。大损失的数量很少，这使得很难找到这部分数据的适当分布。非常有必要对保险索赔金额进行建模，特别是大额损失部分，这可以帮助保险公司做出决策，以计算保费，衡量尾部风险和找到最优再保险方案。保险索赔金额的分配在金融，保险和精算领域得到了广泛的研究，可以参见，Embrechts 等人(1997) [2]，Han and Jiang (1997) [3]，McNeil (1997) [4]等。

定义 $(\Omega,\mathcal{F},P)$ 是损失随机变量的概率空间。对于任何损失变量 X，用 $F_X(x):=P(X\le x)$ 表示其分布函数，以及其中

$$P(X>x)=\bar{F}(x)=1-F(x)$$

其中 $P_X(x)$ 被认为是索赔严重性分布。

为了避免在某一特定风险承受水平下在规定的期限内破产，监管机构要求金融机构持有一定数量的风险准备金以满足其未来负债。保险公司通常使用风险度量来管理风险，例如风险价值，指在一定的置信水平下，金融资产(或投资组合)在未来特定时间段内的最大可能损失。对于 VaR 的研究，可以参见 Xia 等人(2023) [5]、Peng 等人(2023) [6]。定义 $\mathrm{VaR}_q(X)$，表示在水平 $q\in(0,1)$ 下，

$$\mathrm{VaR}_q(X):=\inf\{x:F_X(x)\ge q\},$$

我们可以发现，$\mathrm{VaR}_q(X)=F^{\leftarrow}(q)$；另外，其右端点为 $x_F=\sup\{x\in\mathbb{R}:F(x)<1\}\le\infty$。可以看出，索赔严重程度分布估计的准确性也会影响风险价值的估计。

Smith (1989) [7]提出了一种利用极值理论(EVT)和广义 Pareto 分布(GPD)的理论构建复合模型的方

法，称为超额阈值(EOT)或峰值阈值(POT)方法，可以参见 Barlow 等人(2023) [8]。该方法广泛用于拟合非寿险数据的分布。该复合模型由两部分组成，分别对低于和高于阈值的索赔进行建模，可以参见 Ghaddab 等人(2023) [9]。当然，这个阈值是通过一些方法预先确定的，包括图形诊断和启发式方法，可以在 Davison 和 Smith (1990) [10]、Gerstengarbe 和 Werner (1989) [11]、Coles (2001) [12]和 DuMouchel (1983) [13]、Ferreira 等人(2003) [14]、Loretan 和 Phillips (1994) [15]、Stoev 等人(2011) [16]指出，这些图形诊断方法显然是主观的，并且可能对分布尾部的噪声或波动敏感。Wang 等人(2020) [17]认为，这些方法没有理论依据，但易于实施并在实践中频繁使用。Morever，Clauset 等人(2009) [18]提出超阈值分布对该阈值的选择非常敏感。因此，寻找合适的阈值选择方法尤为重要。此外，Scollnik (2007) [19]推广了具有相应估计方法的复合模型，该方法允许同时估计阈值和其他模型参数。然而，在这种方法中，批量分布固定为对数正态分布。Wong 和 Li (2010) [20]提出了一种复合模型，阈值仅在其中确定，参数通过最大间距乘积法(MPS)估算。不幸的是，他们考虑的复合模型是 EOT 模型。

在实践中，我们倾向于更关注上尾部的保险索赔。换言之，我们希望找到一种更简洁和基于统计的方法来选择阈值并估计保险索赔的上尾部的分布参数。Clauset 等人(2009) [18]提出了一个最大似然框架，该框架估计参数以及所选分布的下限，即分布的阈值。此外，Campolieti (2018) [21]认为，最大似然框架允许模型选择和与替代分布的比较，以及评估拟合分布的拟合优度。这也是本篇文章使用最大似然框架方法的原因所在。该方法有助于我们在候选分布中找到最适合的分布来拟合保险索赔。我们希望使用一个通用且简单的分配模型来模拟我们关心的保险索赔。

本文的其余部分组织如下。第 2 节回顾了本文主要使用的三种候选分布模型和最大似然框架方法。第 3 节详细介绍了丹麦火灾保险数据集，并对数据集进行了描述性统计。第 4 节使用第 2 节中提到的方法对丹麦火灾保险数据进行建模和测试，并对结果进行分析和讨论。第 5 节给出了一些结论。

2. 模型和统计方法

回顾 Kleiber 和 Kotz (2003) [22]、Klugman 等人(2012) [23]、Andr 和 Bermudez (2020) [24]，以及 Bazyari (2023) [25]文献，我们发现对保险索赔数据建模的研究通常集中于重尾分布。本文选取了大多数研究者广泛关注和研究的幂律分布、对数正态分布和威布尔分布。同时，这三种分布在许多重尾分布中也具有代表性。此外，在实践中，重要的是只考虑数据的上尾部，即超过阈值的数据部分。在一个世界中，我们考虑幂律分布、对数正态分布和在 $x_{\min}$ 处截断的威布尔分布，表示为

$$p(x)=\frac{\alpha-1}{x_{\min}}\left(\frac{x}{x_{\min}}\right)^{-\alpha},\alpha>0 \tag{2.1}$$

$$p(x)=\frac{1}{\left(1-\Phi\left(\frac{\ln x_{\min}-\mu}{\sigma}\right)\right)\sigma x\sqrt{2\pi}}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right),\mu\in\mathbb{R},\sigma>0 \tag{2.2}$$

$$p(x)=\beta\lambda\exp\left(\lambda x_{\min}^{\beta}\right)x^{\beta-1}\exp\left(-\lambda x^{\beta}\right),\lambda>0,0<\beta<1 \tag{2.3}$$

对于 $x\geq x_{\min}>0$，这里 $x_{\min}$ 是阈值，等价于使得模型成立的 x 值中最小值，以及 $\Phi(\cdot)$ 代表着正态分布的 CDF。此外，α,μ,σ,β 和 λ 是各自模型的相应参数，需要通过统计方法进行估计。根据概率分布的规律性计算系数。不难发现，这三种分布模型都是右偏和重尾的。研究发现，每个分布的尾部衰减率都不同，其中正态分布大于对数正态分布和韦伯分布，这两种分布的尾部衰退率都大于幂律分布，但对数正态和韦伯分布的尾部衰变率相差不大。这意味着幂律分布的尾部更重，其次是对数正态分布和韦伯分

布，最轻的是正态分布。

在本文中，我们使用最大似然(ML)估计模型参数。如 Clauset 等人(2009) [18]所示，方程(2.1)中幂律分布指数 α 的最大似然估计量(MLE)可估计为

$$\hat{\alpha}=1+n\left(\sum_{i=1}^{n}\ln\frac{x_i}{x_{\min}}\right)^{-1}, \tag{2.4}$$

其中，$x_i, i=1,\cdots,n$ 是 x 的观测值，并且 $x_i \geq x_{\min}$。显然，n 是 $\geq x_{\min}$ 的观测数，这也是上尾部的观测数。还可以计算其他两个模型参数 μ,σ,β 和 λ 的 MLE。为了简单起见，这里只给出幂律分布指数的 MLE 表达式。在方程(2.4)中，我们发现指数 α 的 MLE 是 $x_{\min}$ 的函数。换句话说，$x_{\min}$ 的值对指数 α 的 MLE 有重要影响。Clauset 等人(2009) [18]证实，如果未正确选择 $x_{\min}$，幂律指数的 MLE 将与其真实值急剧偏离。这也是直觉上可以理解的。如果 $x_{\min}$ 小于实际 $x_{\min}$，则在估计时考虑不属于幂律分布的数据；如果 $x_{\min}$ 大于实际 $x_{\min}$，则在估计过程中将考虑较少的数据，这两者都会使估计不准确。不仅如此，我们还可以从方程(2.1)中判断。

Clauset 等人(2007) [26]提出了一种估计 $x_{\min}$ 的方法，该方法被认为更加客观和有原则。该方法选取 $x_{\min}$，以使观测数据与拟合的幂律分布之间的差异尽可能小，

$$D(x)=\max_{x\geq x_{\min}}\left|F_n(x)-F(x)\right| \tag{2.5}$$

其中 $F_n(x)$ 是值至少为 $x_{\min}$ 的观测数据的经验 CDF，$F(x)$ 为最适合 $x\geq x_{\min}$ 数据的幂律分布的 CDF。因此，$x_{\min}$ 的估计值，被表示为 $\hat{x}_{\min}$，使方程(2.5)中的 D 最小化。

仅估计模型参数是不够的，还需要测试 MLE 方法的拟合优度，该方法生成一个 p 值，用于量化假设的合理性。本文使用了一种经典的测试方法，称为 Kolmogorov-Smirnov 测试(KS 测试)，适用于基础分布固定的情况。然而，在我们的例子中，由于 $x_{\min}$ 的不确定性，基础分布在不同的数据集之间有所不同。为了量化 $x_{\min}$ 估计中的不确定性，我们选择了 Efron 和 Tibshirani (1993) [27]详细介绍的非参数“bootstrap”方法。在计算 KS 测试的 p 值之后，我们需要使用 p 值来判断是拒绝幂律分布假设还是认为幂律假设分布是真的。如果 $p\leq 0.05$，则排除幂律分布假设。相反，人们认为幂律分布能够很好地拟合数据。我们认为，当用不同的分布拟合数据时，$x_{\min}$ 的值是不同的，因此对其他两个分布进行了相同的 KS 检验。最后，我们根据 KS 测试的 p 值判断哪个分布更适合数据集。具有最高 p 值的分布被认为是候选分布中适合数据集的最佳分布。

3. 数据集和描述性统计

该数据集可以在 R 语言的 SMPracticals 包中找到。它由四个部分组成：建筑损失、内容物损失、利润损失、总损失，其中这四组的数据量分别为 1990、1679、616、2167，对以丹麦克朗为单位的火灾损失索赔的观察结果，以 1985 年的价格来计算的。自 Embrechts 等人(1997) [2]和 McNeil (1997) [4]首次考虑最后一组数据以来，许多学者对其进行了广泛研究，主要是关于哪种分布更适合这组数据。我们让读者参考 Wong 和 Li (2010) [20]、Lee 等人(2012) [28]、Nadarajah 和 Bakar (2014) [29]、Wang 等人(2020) [17]，仅列举文献中的几个。

为了更直观地探索数据趋势，图 1 给出了丹麦火灾保险索赔的密度图。核密度估计是一种估计随机变量概率密度函数的非参数方法。核密度图本质上是一条直方图拟合曲线，可以看作是一个概率密度图。从图 1 中，我们可以清楚地观察到这四组数据中有很多右尾数据，并且有一些较大的值。这显然是一个右偏分布。但很难看出它是否是重尾分布。我们可以通过描述性统计进一步分析数据。此外，从图 1 中，我们还可以看到这些损失是非负的。

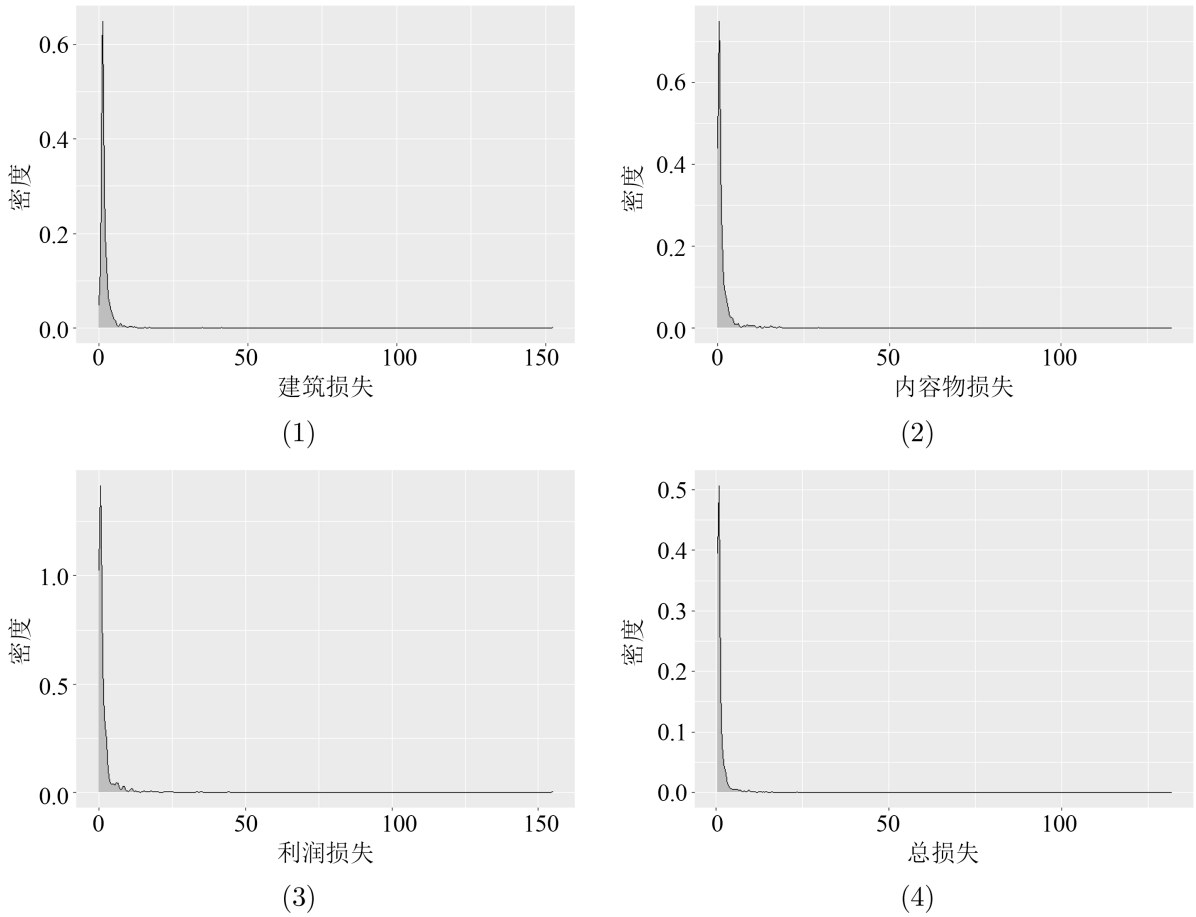


Figure 1. Kernel density plot
图 1. 核密度图

此外，表 1 给出了丹麦火灾数据集中四组数据的一些描述性统计数据。根据这组数据的平均值、中值、标准差、最大值和最小值，我们发现数据中存在较大的值，并且频率较低。这组数据的偏度和峰度表明它服从偏斜和重尾分布。偏度为 0，两侧的尾部长度是对称的，例如，正态分布是对称分布。如果偏度小于 0，则分布具有负偏差。它也被称为左偏。左尾巴比右尾巴长。如果偏度大于 0，则分布具有正偏差。它也被称为右偏。右尾巴比左尾巴长。偏差的绝对值越大，分布的偏差越严重。峰度可以描述分布模式的陡度。如果峰度小于 3，则称分布峰度不足，且分布平缓。如果峰度大于 3，则该分布称为峰度过大，且分布陡峭。峰度的绝对值越大，其分布形式与正态分布之间的差异就越大。

Table 1. Descriptive statistics for Danish fire insurance claims, millions (of Danish Kroner)
表 1. 丹麦火灾保险索赔的描述性统计，百万(丹麦克朗)

类别	均值	中值	标准差	偏度	峰度	最小值	最大值
建筑损失	1.99	1.32	4.51	23.84	714.93	0.0232	152.41
内容物损失	1.70	0.58	5.35	15.06	309.10	0.0008	132.01
利润损失	0.85	0.27	2.94	15.40	303.68	0.0041	61.93
总损失	3.39	1.78	8.51	18.76	483.76	1.0000	263.25

为了用右偏和厚尾分布建模和分析数据，我们需要选择一个具有相同两个财产的分布模型。如第 2 节所述，三个分布模型是右偏和重尾的。它们可以用于丹麦火灾保险数据的建模，这是合理的。

4. 实证结果和讨论

在第 2 节和第 3 节中，我们分别详细描述了丹麦火灾保险索赔数据集和三个候选分布模型的特征，并发现它们都具有右偏和重尾的特征。因此，将这三种候选分布作为丹麦火灾保险索赔数据的拟合分布是合理的。然而，仍然需要通过统计方法获得特定的参数值。表 2 给出了使用最大似然框架估计和测试每组数据分布的结果。从表 2 中，我们发现，对于建筑损失数据，幂律分布和对数正态分布的 p 值均是大于 0.05 的，并且对数正态分布的 p 值是大于幂律分布的的 p，这表明对数正态分布比幂律分布更适合；对于内容物损失，幂律分布、对数正态分布以及韦布尔分布的 p 值均是大于 0.05 的，并且对数正态分布的 p 值是大于韦布尔分布的 p 值，这表明三种分布中对数正态分布最多；对于利润损失数据，幂律分布的 p 值是大于 0.05 的，对数正态分布以及韦布尔分布的 p 值均是小于 0.05 的，这意味着幂律分布可以描述这组数据；对于总损失组数据，幂律分布和对数正态分布的 p 值均是远远大于 0.05，并且幂律分布的 p 值是大于对数正态分布的的 p 值的，这表示幂律分布可以描述这组数据。使用动态阈值选择方法，每组数据的尾部长度(即 n_{tail} 相对合理。Clauset 等人(2009) [18]提到，当 n_{tail} 较小时，应谨慎对待高 p 值。

Table 2. Estimates of power law, lognormal and weibull distributions for Danish fire data of four groups
表 2. 对四组丹麦火灾数据进行幂律、对数正态和韦布尔分布估计

类别	分布	参数 1[1]	参数 2[1]	$x_{\min}$	n_{tail}[2]	$\hat{p}$ [3]
建筑损失	幂律分布	$\hat{\alpha}=2.7663$	-	1.24223602	1126 (1990)	0.1232
	对数正态分布	$\hat{\mu}=-18.5883$	$\hat{\sigma}=3.3733$	1.21507197	1159 (1990)	0.2128
	韦布尔分布	-	-	-	-	-
内容物损失	幂律分布	$\hat{\alpha}=2.2402$	-	1.965924	310 (1679)	0.0412
	对数正态分布	$\hat{\mu}=-2.4674$	$\hat{\sigma}=1.9443$	0.8787129	661 (1679)	0.4780
	韦布尔分布	$\hat{\beta}=0.1869$	$\hat{\lambda}=5109.529$	0.8873114	659 (1679)	0.3890
利润损失	幂律分布	$\hat{\alpha}=2.1265$	-	0.577557756	186 (616)	0.0820
	对数正态分布	$\hat{\mu}=-1.2372$	$\hat{\sigma}=1.3753$	0.013554	600 (616)	0.0048
	韦布尔分布	$\hat{\beta}=0.3456$	$\hat{\lambda}=16.1784$	0.081287045	520 (616)	0.0000
总损失	幂律分布	$\hat{\alpha}=2.4037$	-	1.375	1564 (2167)	0.4294
	对数正态分布	$\hat{\mu}=-16.6967$	$\hat{\sigma}=3.6801$	1.90264	976 (2167)	0.3092
	韦布尔分布	-	-	-	-	-

[1]它表示每个分布的参数估计。[2]它表示≥$x_{\min}$的观察次数；括号中的数字表示数据总量。[3]这是 2500 次复制 KS 检验得出的 p 值。

我们根据图 2~5 中双对数尺度的数据绘制了三个候选分布(即，基于表 2 中的估计)。这些图片为四组丹麦火灾数据与我们估计的三个候选分布的拟合提供了视觉证据。从图 2 中，我们可以发现，建筑组数据的上尾部的大部分观测值都在图中虚线附近，只有少数观测值远离这条线。图 2(1)和(2)都显示了这一特性，这也表明幂律分布和对数正态分布与建筑损失组数据非常吻合。对于图 3，与(2)和(3)中的虚线相比，(1)中偏离虚线的观察值更多，这意味着对数正态分布和韦伯分布对内容物损失组数据的拟合优于幂律分布。对于图 4，很明显，有许多观察结果偏离了(2)中的虚线。和(3)中存在明显的偏差点。相比之下，(1)中的观察结果更好地分布在虚线周围。这意味着幂律分布更符合利润损失组的数据。对于图 5，

只有(2)中的一些观察结果偏离虚线，(1)中的情况类似。这也意味着幂律分布和对数正态分布都能很好地拟合总损失组的数据。

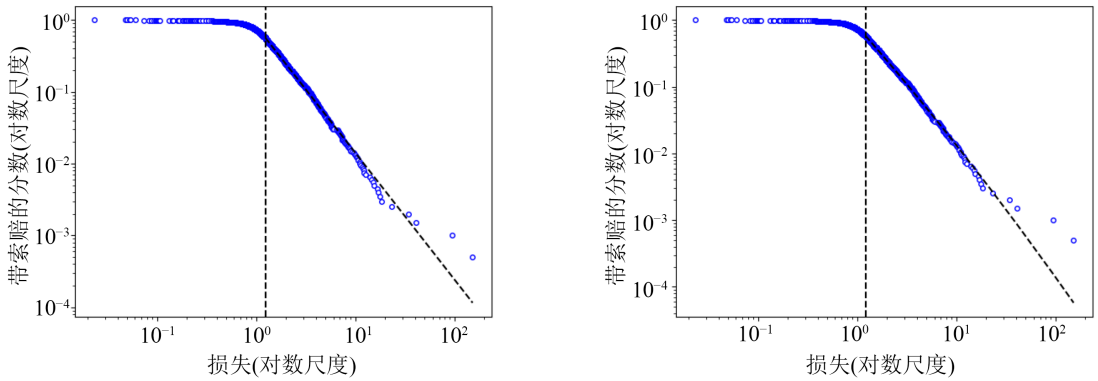


Figure 2. Plot of the data for the Building group and fitted Power law and Lognormal distribution. Notes: vertical dashed line denotes estimate of $x_{\min}$

图 2. 对建筑损失组数据图拟合幂律分布、对数正态分布。注：垂直虚线表示 $x_{\min}$ 的估计值

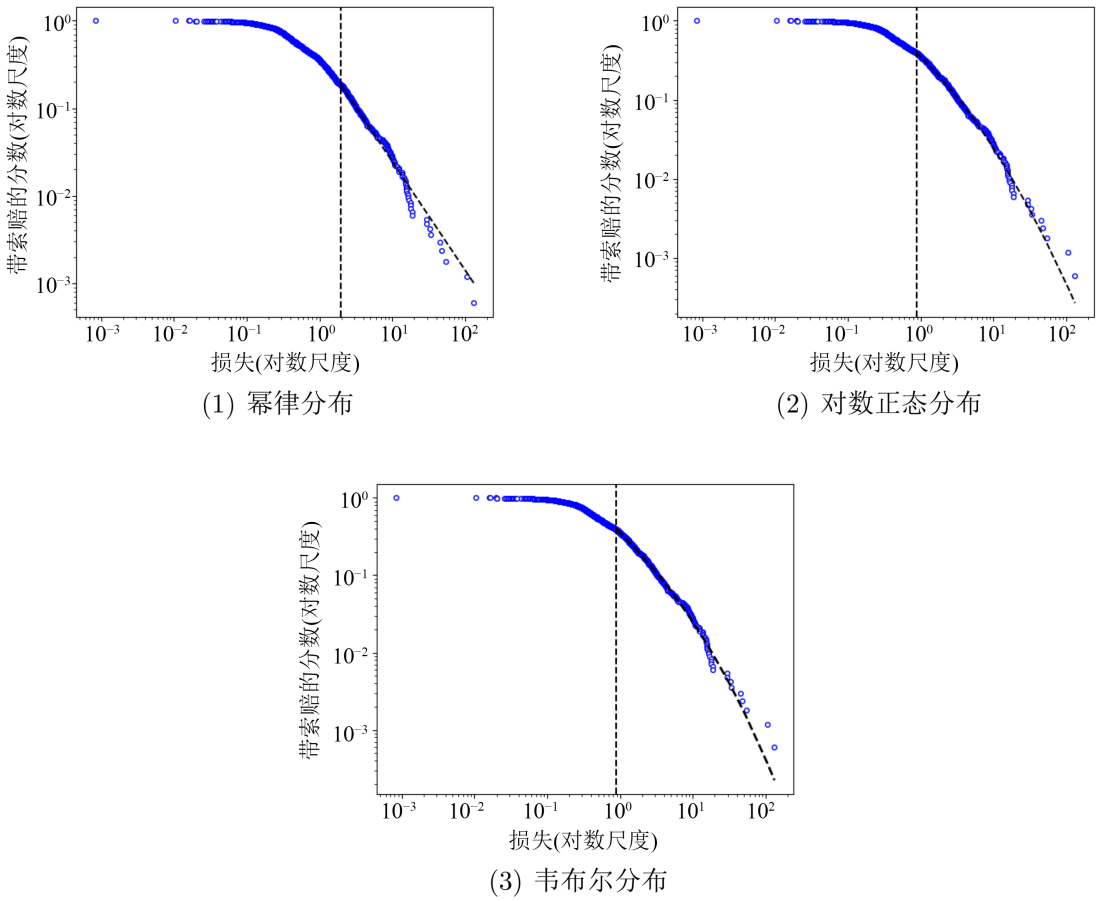
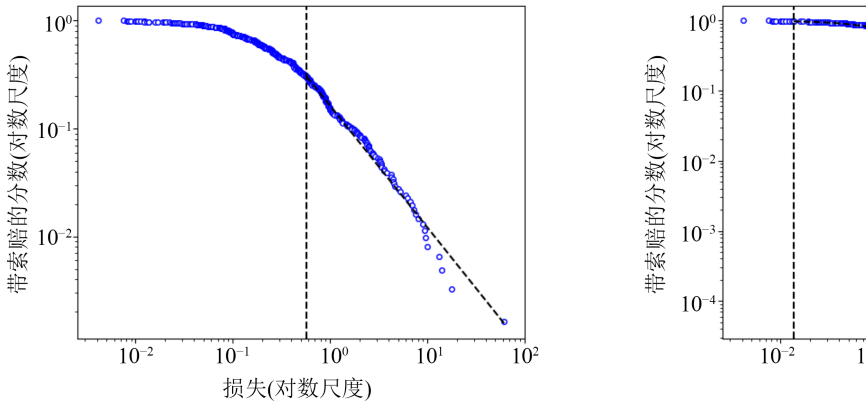
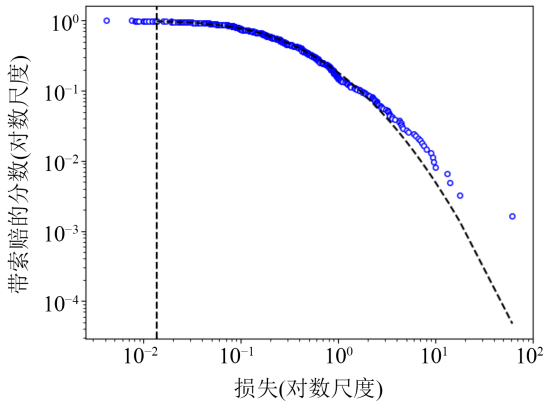


Figure 3. Plot of the data for the Contents group and fitted Power law, Lognormal and Weibull distribution. Notes: vertical dashed line denotes estimate of $x_{\min}$

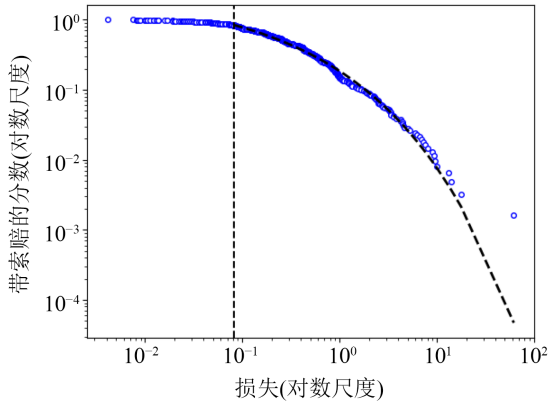
图 3. 对内容物损失组数据图拟合幂律分布、对数正态分布和韦布尔分布。注：垂直虚线表示 $x_{\min}$ 的估计值



(1) 幂律分布



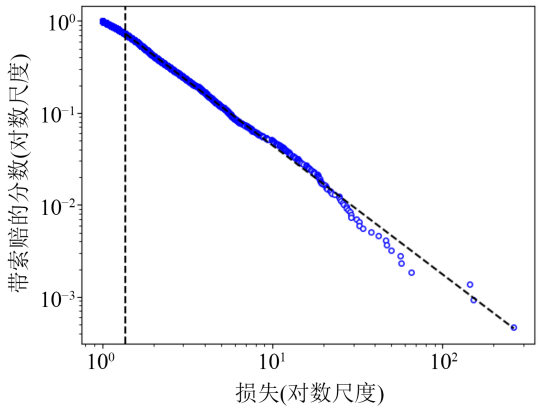
(2) 对数正态分布



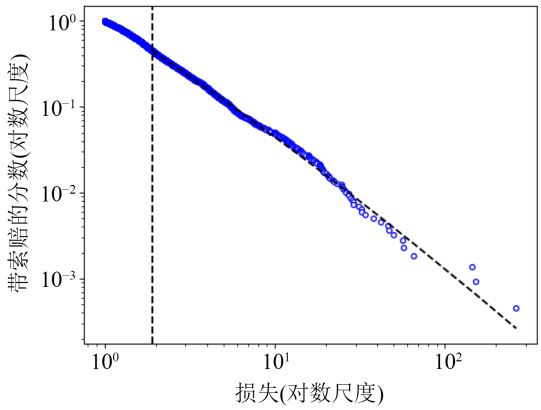
(3) 韦布尔分布

Figure 4. Plot of the data for the Profits group and fitted Power law, Lognormal and Weibull distribution. Notes: vertical dashed line denotes estimate of $x_{\min}$

图 4. 对利润损失组数据图拟合幂律分布、对数正态分布和韦布尔分布。注：垂直虚线表示 $x_{\min}$ 的估计值



(1) 幂律分布



(2) 对数正态分布

Figure 5. Plot of the data for the Total group and fitted Power law and Lognormal distribution. Notes: vertical dashed line denotes estimate of $x_{\min}$

图 5. 对总损失组数据图拟合幂律分布、对数正态分布和韦布尔分布。注：垂直虚线表示 $x_{\min}$ 的估计值

如上所述，阈值的选择对分布的估计和测试有一定的影响。我们想验证保险索赔数据是否也会产生这样的结果。因此，我们将数据中的最小值作为阈值，即对整个数据进行建模，结果如表 3 所示。我们可以发现，如果没有选择阈值来建模数据，测试的 p 值小于 0.05，这拒绝了原始假设，也就是说，建立的模型不能很好地拟合丹麦火灾数据。表 2 中每组数据的每个分布的参数与表 3 中的值不同。这也证明，对于保险索赔数据，在建模之前选择阈值是有效的。

Table 3. Estimates of power law, lognormal and weibull distributions for Danish fire data of four groups, $x_{\min}$ equals minimum claim amount in each group
表 3. 对四组丹麦火灾数据进行幂律、对数正态和韦布尔分布估计，$x_{\min}$ 等于每组的最低损失金额

类别	分布	参数 1[1]	参数 2[1]	$x_{\min}$	n_{tail}[2]	$\hat{p}$ [3]
建筑损失	幂律分布	$\hat{\alpha}=1.2438$	-	0.02319109	1990 (1990)	0.0000
	对数正态分布	$\hat{\mu}=0.3384$	$\hat{\sigma}=0.7438$	0.02319109	1990 (1990)	0.0000
	韦布尔分布	$\hat{\beta}=1.0306$	$\hat{\lambda}=1.9994$	0.02319109	1990 (1990)	0.0000
内容物损失	幂律分布	$\hat{\alpha}=1.14984$	-	0.000825083	1679 (1679)	0.0000
	对数正态分布	$\hat{\mu}=-0.4263$	$\hat{\sigma}=1.2700$	0.000825083	1679 (1679)	0.0000
	韦布尔分布	$\hat{\beta}=0.7074$	$\hat{\lambda}=1.2274$	0.000825083	1679 (1679)	0.0000
利润损失	幂律分布	$\hat{\alpha}=1.2369$	-	0.004084	616 (616)	0.0000
	对数正态分布	$\hat{\mu}=-1.2874$	$\hat{\sigma}=1.4261$	0.004084	616 (616)	0.0204
	韦布尔分布	$\hat{\beta}=0.6080$	$\hat{\lambda}=0.4854$	0.004084	616 (616)	0.0000
总损失	幂律分布	$\hat{\alpha}=2.2707$	-	1.0000	2167 (2167)	0.0000
	对数正态分布	$\hat{\mu}=-4.6232$	$\hat{\sigma}=2.1843$	1.0000	2167 (2167)	0.0000
	韦布尔分布	$\hat{\beta}=0.1322$	$\hat{\lambda}=7.7715$	1.0000	2167 (2167)	0.0000

[1] 它表示每个分布的参数估计。[2] 它表示$\geq x_{\min}$ 的观察次数；括号中的数字表示数据总量。[3] 这是 2500 次复制 KS 检验得出的 p 值。

表 2 中选择阈值的方法是动态的，这使我们想知道如果预先固定阈值，分布的估计和测试将如何受到影响。有几种方法可以预先确定阈值，包括固定分位数规则、平方根规则、经验规则、AMSE 希尔估计量的最小值、指数检验、Gertensgarbe 图。有关这些方法的详细信息，请参考 Wang 等人(2020) [17]。使用上述六种方法的阈值选择结果如表 4 所示。我们可以发现，不同方法选择的阈值是非常不同的。

Table 4. Thresholds of each group of data under different threshold selection methods. Note: Numbers in parentheses indicate the number of observations $\geq x_{\min}$
表 4. 不同阈值选择方法下每组数据的阈值。注意：括号中的数字表示$\geq x_{\min}$ 的观察次数

类别	固定分位数规则	平方根规则	经验规则	AMSE 希尔估计量的最小值	指数检验	Gertensgarbe 图
建筑损失	3.3869602 (199)	7.32064422 (45)	5.41727672 (78)	1.780105 (595)	12.1522 (15)	-
内容物损失	3.206442 (168)	10.5 (41)	7.493649 (70)	3.567985 (148)	16.50165 (16)	5.401557 (92)
利润损失	1.668520578 (62)	3.549245785 (25)	2.540220152 (39)	1.330967 (71)	2.303665 (47)	4.638219 (18)
总损失	5.561735 (217)	18.322083 (47)	12.376238 (82)	2.956522 (546)	2.05937 (84)	4.500423 (297)

我们选择了一组尾部数据较多的数据进行验证，以确保分布估计和测试的可靠性的稳定性。通过比较分析，我们选择了根据 AMSE 希尔估计量的最小值方法获得的每组丹麦火灾数据的阈值，结果见表 5。

Table 5. Estimates of power law, lognormal and weibull distributions for danish fire data of four groups, $x_{\min}$ is determined by this method Minimum AMSE of the Hill estimator in each group
表 5. 对四组丹麦火灾数据进行幂律、对数正态和韦布尔分布估计，各组的 $x_{\min}$ 通过 AMSE 希尔估计量的最小值方法所确定

类别	分布	参数 1[1]	参数 2[1]	$x_{\min}$	n_{tail}[2]	$\hat{p}$ [3]
建筑损失	幂律分布	$\hat{\alpha}=2.7641$	-	1.780105	595 (1990)	0.4300
	对数正态分布	$\hat{\mu}=-13.4585$	$\hat{\sigma}=2.9288$	1.780105	595 (1990)	0.3840
	韦布尔分布	$\hat{\beta}=0.0754$	$\hat{\lambda}=2.1930\text{e}-18$	1.780105	595 (1990)	0.5796
内容物损失	幂律分布	$\hat{\alpha}=2.2660$	-	3.567985	148 (1679)	0.0160
	对数正态分布	$\hat{\mu}=-0.5082$	$\hat{\sigma}=1.5780$	3.567985	148 (1679)	0.0744
	韦布尔分布	$\hat{\beta}=0.2443$	$\hat{\lambda}=0.0089$	3.567985	148 (1679)	0.0784
利润损失	幂律分布	$\hat{\alpha}=2.0864$	-	1.330967	71 (616)	0.0280
	对数正态分布	$\hat{\mu}=0.0893$	$\hat{\sigma}=1.2389$	1.330967	71 (616)	0.8160
	韦布尔分布	$\hat{\beta}=0.3583$	$\hat{\lambda}=0.1370$	1.330967	71 (616)	0.8128
总损失	幂律分布	$\hat{\alpha}=2.4281$	-	2.956522	546 (2167)	0.6312
	对数正态分布	$\hat{\mu}=-19.4915$	$\hat{\sigma}=3.9193$	2.956522	546 (2167)	0.1620
	韦布尔分布	$\hat{\beta}=0.0619$	$\hat{\lambda}=5.4561\text{e}-22$	2.956522	546 (2167)	0.1172

[1] 它表示每个分布的参数估计。[2] 它表示$\geq x_{\min}$的观察次数；括号中的数字表示数据总量。[3] 这是 2500 次复制 KS 检验得出的 p 值。

从表 5 可以看出，对于通过该方法选择的阈值，对于四组数据，建立的分布不是很稳定。对于具有大尾部数据的建筑损失组和总损失组数据，结果与动态阈值选择的结果相差不大。但是，对于尾部较小的内容物损失组和利润损失组数据，结果与动态阈值选择的结果有很大不同。内容物损失组数据中对数正态分布和韦布尔分布的测试结果(即 p 值)相对较低，这与表 2 中的值大相径庭。对于利润损失组数据，对数正态分布和韦布尔分布的 p 值与表中的 p 值相比非常大。我们可以观察到，利润损失组数据中 n_{tail} 的值仅为 71，这使得大 p 值的可信度较低。通常，动态选择阈值的方法更稳定。

如第 1 节所述，保险公司通常更关注偿付能力，例如风险价值。因此，我们还希望看到通过该方法估计的分布模型的性能，以预测 VaR，结果如表 6 所示。对于建筑损失组数据，幂律分布预测的 VaR 值高于对数正态分布。显然，对数正态分布预测的 VaR 值更接近于经验分布。对于内容物损失组数据，对数正态分布和韦伯分布对 VaR 的预测相似，优于幂律分布。对于利润损失组数据，对数正态分布的预测 VaR 值低于经验 VaR 值。韦布尔分布预测的 VaR 值与经验 VaR 值非常接近，但仍有一个较低的值。幂律分布预测的 VaR 值更安全。对于总损失组数据，幂律分布对 VaR 有更好的预测，并且更接近经验 VaR 值。

Table 6. Comparison of predicted VaR and empirical VaR
表 6. 比较预测的 VaR 和经验的 VaR

类别	分布	95%	96%	97%	98%	99%
建筑损失	真实	4.8	5.3147877	6.04259095	7.72889417	11.08968177
	幂律分布	6.7731	7.6852	9.0446	11.3785	16.8467
	对数正态分布	6.581035	7.424278	8.664238	10.75149	15.47456

Continued

内容物损失	真实	5.92718	7.913031	9.424084	12.28218	16.4026
	幂律分布	22.0099	26.3487	33.2277	46.0773	80.5778
	对数正态分布	11.57234	13.45102	16.24485	21.00229	31.87839
	韦布尔分布	11.74653	13.65395	16.47717	21.24533	31.95976
利润损失	真实	3.217612193	3.549245785	4.537953795	6.654835847	9.276437848
	幂律分布	8.2505	10.0578	12.9840	18.6089	34.4299
	对数正态分布	2.811268	3.250539	3.885861	4.926929	7.163149
	韦布尔分布	3.651128	4.256908	5.130178	6.548524	9.526237
总损失	真实	10.011123	11.801242	14.293194	18.628281	26.214641
	幂律分布	11.6187	13.6206	16.7187	22.3178	36.5683
	对数正态分布	15.9362	18.5092	22.4135	29.2682	45.8367

丹麦火灾数据是一组经典数据。关于总损失组的数据已有很多文献，这是丹麦火险数据中数据最多的一组火灾保险数据集。这组数据已被许多学者广泛研究，其结果与本文的研究结果一致，见 Scollnik (2007) [19]、Wong 和 Li (2010) [20]。数据集中剩余三组数据的建模结果也与经验判断一致。因此，在之后的研究中可以对重尾分布族的其他分布进行估计和检验，这样有助于研究保险索赔数据的重要性质以帮助解决实际问题。除此之外，还可以为保险索赔理论提供实证方面的研究。本文目前所研究的是单个保险索赔随机变量所服从的分布，未来希望研究如何能够估计和检验出多个具有某种相依结构的保险索赔随机变量的联合分布。

5. 结束语

在本文中，使用最大似然估计器来估计候选分布的参数和下限(即阈值)。此外，当对四组丹麦火灾数据进行建模时，该方法可以通过比较 p 值在三个候选分布中选择最佳拟合分布。这些结果与其他研究人员的结果一致，也与第 4 节中提到的经验判断相结合。众所周知，重尾分布中不仅有这三种分布，未来还可以估计和检验其他分布。本文可以作为保险索赔数据建模的启发式探索。此外，本文所述方法也可以为保险索赔的理论研究提供实证支持。我们关注的是，基于保险索赔数据建立的模型是否能够通过这样一种动态选择的阈值方法准确预测其风险度量指标。

基金项目

本文受 2022 年江苏省研究生科研创新计划项目资助(项目批准号：KYCX22_2212)，项目名称：金融扭曲风险测度的渐近行为研究，类别：自然科学。

参考文献

[1] Agnes, M., Koestoer, R.H. and Sodri, A. (2023) Social and Environmental Risks Integration into Underwriting of Non-Life Insurance: A Review of Sustainable Finance in Indonesia. *Jurnal Ilmu Lingkungan*, **21**, 125-131.

[2] Embrechts, P., Klppelberg, C. and Mikosch, T. (1997) Modelling Extremal Events for Insurance and Finance. Springer, Berlin. https://doi.org/10.1007/978-3-642-33483-2

[3] Han, T. and Jiang, H. (1997) The Distribution of Amount on Insurance Claim and Its Application. *Journal of East China Normal University* (*Natural Science*), No. 4, 30-33.

[4] McNeil, A.J. (1997) Estimating the Tails of Loss Severity Distributions Using Extreme Value Theory. *ASTIN Bulletin*: *The Journal of the IAA*, **27**, 117-137. https://doi.org/10.2143/AST.27.1.563210

[5] Xia, Z., Zou, Z. and Hu, T. (2023) Inf-Convolution and Optimal Allocations for Mixed-VaRs. *Insurance*: *Mathematics*

and Economics, **108**, 156-164. https://doi.org/10.1016/j.insmatheco.2022.12.001

[6] Peng, S., Yang, S. and Yao, J. (2023) Improving Value-at-Risk Prediction under Model Uncertainty. *Journal of Financial Econometrics*, **21**, 228-259. https://doi.org/10.1093/jjfinec/nbaa022

[7] Smith, R.L. (1989) Extreme Value Analysis of Environmental Time Series: An Application to Trend Detection in Ground-Level Ozone. *Statistical Science*, 367-377. https://doi.org/10.1214/ss/1177012400

[8] Barlow, A.M., Mackay, E., Eastoe, E. and Jonathan, P. (2023) A Penalised Piecewise-Linear Model for Non-Stationary Extreme Value Analysis of Peaks over Threshold. *Ocean Engineering*, **267**, Article ID: 113265. https://doi.org/10.1016/j.oceaneng.2022.113265

[9] Ghaddab, S., Kacem, M., de Peretti, C. and Belkacem, L. (2023) Extreme Severity Modeling Using a GLM-GPD Combination: Application to an Excess of Loss Reinsurance Treaty. *Empirical Economics*, 1-23. https://doi.org/10.1007/s00181-023-02371-4

[10] Davison, A.C. and Smith, R.L. (1990) Models for Exceedances over High Thresholds. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **52**, 393-425. https://doi.org/10.1111/j.2517-6161.1990.tb01796.x

[11] Gerstengarbe, F.W. and Werner, P.C. (1989) A Method for the Statistical Definition of Extreme-Value Regions and Their Application to Meteorological Time Series. *Zeitschrift fuer Meteorologie* (*German Democratic Republic*), **39**, 224-226.

[12] Coles, S., Bawa, J., Trenner, L. and Dorazio, P. (2001) An Introduction to Statistical Modeling of Extreme Values. Springer, London. https://doi.org/10.1007/978-1-4471-3675-0

[13] DuMouchel, W.H. (1983) Estimating the Stable Index α in Order to Measure Tail Thickness: A Critique. *The Annals of Statistics*, **11**, 1019-1031. https://doi.org/10.1214/aos/1176346318

[14] Ferreira, A., de Haan, L. and Peng, L. (2003) On Optimising the Estimation of High Quantiles of a Probability Distribution. *Statistics*, **37**, 401-434. https://doi.org/10.1080/0233188021000055345

[15] Loretan, M. and Phillips, P.C. (1994) Testing the Covariance Stationarity of Heavy Tailed Time Series: An Overview of the Theory with Applications to Several Financial Datasets. *Journal of Empirical Finance*, **1**, 211-248. https://doi.org/10.1016/0927-5398(94)90004-3

[16] Stoev, S.A., Michailidis, G. and Taqqu, M.S. (2011) Estimating Heavy-Tail Exponents through Max Self-Similarity. *IEEE Transactions on Information Theory*, **57**, 1615-1636. https://doi.org/10.1109/TIT.2010.2103751

[17] Wang, Y., Haff, I.H. and Huseby, A. (2020) Modelling Extreme Claims via Composite Models and Threshold Selection Methods. *Insurance*: *Mathematics and Economics*, **91**, 257-268. https://doi.org/10.1016/j.insmatheco.2020.02.009

[18] Clauset, A., Shalizi, C.R. and Newman, M.E.J. (2009) Power-Law Distributions in Empirical Data. *Society for Industrial and Applied Mathematics Review*, **51**, 661-703. https://doi.org/10.1137/070710111

[19] Scollnik, D.P. (2007) On Composite Lognormal-Pareto Models. *Scandinavian Actuarial Journal*, **2007**, 20-33. https://doi.org/10.1080/03461230601110447

[20] Wong, T.S.T. and Li, W.K. (2010) A Threshold Approach for Peaks-over-Threshold Modeling Using Maximum Product of Spacings. *Statistica Sinica*, **20**, 1257-1272.

[21] Campolieti, M. (2018) Heavy-Tailed Distributions and the Distribution of Wealth: Evidence from Rich Lists in Canada, 1999-2017. *Physica A*: *Statistical Mechanics and Its Applications*, **503**, 263-272. https://doi.org/10.1016/j.physa.2018.02.057

[22] Kleiber, C. and Kotz, S. (2003) Statistical Size Distributions in Economics and Actuarial Sciences. John Wiley & Sons, Hoboken. https://doi.org/10.1002/0471457175

[23] Klugman, S.A., Panjer, H.H. and Willmot, G.E. (2012) Loss Models: From Data to Decisions. John Wiley & Sons, Hoboken. https://doi.org/10.1002/9781118787106

[24] Andr, L.M. and de Zea Bermudez, P. (2020) Modelling Dependence between Observed and Simulated wind Speed Data Using Copulas. *Stochastic Environmental Research and Risk Assessment*, **34**, 1725-1753. https://doi.org/10.1007/s00477-020-01866-1

[25] Bazyari, A. (2023) Infinite Time Ruin Probability in the Individual Risk Model with Dependent Structure for Light and Heavy Tailed Distributions. *Journal of Statistical Sciences*, **16**, 309-330. https://doi.org/10.52547/jss.16.2.309

[26] Clauset, A., Young, M. and Gleditsch, K.S. (2007) On the Frequency of Severe Terrorist Attacks. *Journal of Conflict Resolution*, **51**, 58-87. https://doi.org/10.1177/0022002706296157

[27] Efron, B. and Tibshirani, R.J. (1993) An Introduction to the Bootstrap. Chapman and Hall, New York. https://doi.org/10.1007/978-1-4899-4541-9

[28] Lee, D., Li, W.K. and Wong, T.S.T. (2012) Modeling Insurance Claims via a Mixture Exponential Model Combined with Peaks-over-Threshold Approach. *Insurance*: *Mathematics and Economics*, **51**, 538-550.

https://doi.org/10.1016/j.insmatheco.2012.07.008

[29] Nadarajah, S. and Bakar, S. (2014) New Composite Models for the Danish Fire Insurance Data. *Scandinavian Actuarial Journal*, **2014**, 180-187. https://doi.org/10.1080/03461238.2012.695748