Danish fire insurance data: a review and additional analysis

Sandile Charles Shongwe[1*] and *Walena Anesu* Marambakuyana[1]

[1]Department of Mathematical Statistics and Actuarial Science, Faculty of Natural and Agricultural Sciences, University of the Free State, Bloemfontein, 9301, South Africa

Abstract. The Danish fire insurance data is one of the most recognised and well-known datasets in the empirical insurance claims literature. This dataset is used in many textbooks and articles to illustrate the analysis of fire insurance claims, more specifically in the application of heavy-tailed loss distributions and extreme value theory. In this paper, we provide a short review of publications that used the Danish fire insurance data and conduct an additional analysis. Our additional analysis on the Danish fire insurance data involves investigating the: (i) modality issue using appropriate statistical tests and software, (ii) k-means clustering pattern using different techniques, (iii) effect of using a splicing model on the data, and (iv) differences in results that we obtained as compared to what other cited researchers reported in their earlier publications. In short, the objective of this paper is to highlight the importance of the Danish fire claims dataset by showcasing different models where it has been used to verify certain hypotheses in the empirical actuarial field. Additional analyses are also conducted to illustrate its dense usefulness in the actuarial and extremes field, where real-life datasets are scarce because they are often subjected to a lot of proprietary and privacy laws.

1 Introduction

There are very few real-life datasets from the insurance sector that are freely available – this is due to the fact that private companies' datasets are subject to a lot of proprietary laws. The Danish fire loss data, which was collected by Copenhagen Reinsurance in Denmark, is one of the most studied datasets as it is built in some of the data analysis software, including R. This dataset covers losses from fire due to buildings, contents, and profits from January 1980 to December 1990. The Danish data is reported in millions of Danish Krones (DKKs) and adjusted for inflation to reflect 1985 values. Most of the earlier publications cite [1] as the first statistical publication to use this dataset. Table 1 provides a summary of some of the applications of the Danish fire loss data with 2492 observations (or a subset of it with 2167 observations that are greater than or equal to 1 million DKKs) that we found in the literature. This dataset has received a lot of attention from researchers in the actuarial and extreme value fields, mainly because insurers rarely give public access to their data which spans over a decade.

*Corresponding author: shongwesc@ufs.ac.za

Table 1. Summary of some applications of the Danish fire loss data.

Reference	Study
McNeil (1997)	Estimating the tails of loss severity using extreme value theory
Resnick (1997)	Statistical techniques and plotting devices in extreme value theory
Cooray & Ananda (2005)	Composite lognormal-Pareto distribution with fixed mixing weights
Scollnik (2007)	Composite lognormal-Pareto with unrestricted mixing weights
Drees & Muller (2008)	Bivariate model
Brazauskas & Kleefeld (2009)	Fitting of the generalized Pareto
Ausin et al. (2009)	Bayesian estimation of finite time ruin probabilities
Carreau & Bengio (2009)	Hybrid Pareto models
Esmaeili & Kluppelberg (2010)	Bivariate compound Poisson process
Charpentier & Oulidi (2010)	Beta kernel quantile estimation
Pigeon & Denuit (2011)	Composite lognormal-Pareto model with a random threshold
Guillotte et al. (2011)	Non-parametric Bayesian inference on bivariate extremes
Eling (2012)	Skew-normal and skew-student distribution
Bernadi et al. (2012)	Bayesian approach for skew mixture models
Scollnik & Sun (2012)	Composite Weibull-Pareto with unrestricted mixing weights
Ruckdeschel & Horbenko (2013)	Robust estimators in generalized Pareto models
Maghsoudi et al. (2014)	Composite Weibull-Inverse transformed gamma
Nadarajah & Abu Bakar (2014)	Composite lognormal-Burr model
Abu Bakar et al. (2015)	Composite Weibull models
Calderin-Ojeda (2015)	Composite Weibull-Burr with mode-matching procedure
Calderin-Ojeda and Kwok (2016)	Composite lognormal-Stoppa & Weibull-Stoppa
Miljkovic & Grün (2016)	K component non-Gaussian mixture models
Reynkens et al. (2017)	Splicing: Mixed Erlang for the body and a Pareto for the tail
Abu Bakar et al. (2017)	Two-component mixture models with a focus on the Burr
Grün & Miljkovic (2019)	256 composite models
Abu Bakar & Nadarajah (2019)	17 Two-component mixture models
Liu & Ananda (2023)	Exponentiated composite inverse Gamma-Pareto model
Marambakuyana & Shongwe (2024a)	256 mixture models
Marambakuyana & Shongwe (2024b)	19 standard heavy-tailed distributions

Table 2. Descriptive statistics for the Danish fire claims data with 2492 observations.

Descriptive measure	Amounts (in mil DKK)
Minimum	0.3134
1st quartile	1.1572
Median	1.6339
3rd quartile	2.6455
Mean	3.0627
Maximum	263.2504
Sum	7623.2460
Standard deviation	7.9767
Coefficient of variation	2.6045
Skewness	19.8961
Kurtosis	549.5736

In this paper, we focus on the larger dataset of 2492 observations and the detailed descriptives of this dataset are given in Table 2. Overall, the Danish loss data is similar to other data of the same nature in that it has a positive support, a high frequency of low severity claims and a low frequency of high severity claims, it is strongly skewed to the right, and it is leptokurtic (it has fat, heavy tails). The Danish dataset is available in the 'SMPracticals' package in R [30]. Other add-on packages in R – 'evir' [31], 'fitdistrplus' [32] and 'QRM'

[33], etc – have a subset of the dataset which consists of 2167 losses recorded in millions of DKKs (which are > 1 million DKKs). The 'fitdistrplus' package contains the 'danishmulti' dataset which provides more details about the claims greater than 1 million DKKs. The dataset consists of a column for date, losses due to buildings, contents, profits, and total losses. For the Danish losses which exceed 1 million DKKs, 54% of the total losses were due to losses to the building, 39% of the losses were due to total losses to the contents and 7% of the total losses were due to the losses of profits.

The rest of the paper is structured as follows: In Section 2, a discussion on the modality and cluster analysis of the data is presented, while Section 3 presents a comparative analysis of the trimmed Danish data using two-component mixture models. Section 4 presents a comparison of two composite models. Section 5 presents a splicing model for modelling the body and the tail of the data, separately. Finally, concluding remarks are given in Section 6.

2 Modality and clustering

2.1 Modality

Multimodality of the Danish fire loss data is discussed in [22,24,26], where the authors deduced that the dataset is multimodal, i.e., that it has at least two modes. There are many tests for unimodality/multimodality available in R, of which two are explored in this study. The first test is found in the corrected 'diptest' add-on package [34] and the second test in the 'LaplacesDemon' add-on package [35]. Abu Bakar et al. [24] stated that using the Hartigans dip test [36], the Danish fire loss data is at least bimodal. To verify this assertion, we used the two abovementioned packages, and the results are reported in Figures 1 and 2. Based on this, we can conclude that the Danish data used in this research work is unimodal, i.e., that is has exactly one mode.

```
library(diptest)
dip.test(d)

        Hartigans' dip test for unimodality / multimodality
data:  d
D = 0.0055055, p-value = 0.9416
   alternative hypothesis: non-unimodal, i.e., at least bimodal
```

Fig. 1. Multimodality test using 'diptest' package for the Danish fire loss data.

```
library(LaplacesDemon)
is.amodal(d)
[1] FALSE
is.unimodal(d)
[1] TRUE
is.bimodal(d)
[1] FALSE
is.trimodal(d)
[1] FALSE
is.multimodal(d)
[1] FALSE
```

Fig. 2. Multimodality test using 'LaplacesDemon' package for the Danish fire loss data.

2.2 Clustering

The aim of cluster analysis is to find distinct groups or 'clusters', where observations in a similar group will, in general, have similar characteristics. In this sub-section, 3 different k-means clustering methods are implemented to determine the number of optimal clusters on

the Danish dataset. Figure 3 illustrates the variance within clusters method, where the total within sum of squares decreases as k increases, and an elbow (or bend) is observed at $k = 3$. This means that the clusters beyond this value have very little value for the data. So, it can be said that using the within sum of squares method, the observations for Danish data can be classified into 3 clusters. Figure 4 illustrates the average silhouette width method for the clusters and a high average silhouette width is indicative of good clustering. Figure 4 suggests that for $k = 2$, the average silhouette is maximized, and the average silhouette width decreases as k increases for $k \geq 2$. So, we can say that using the average silhouette width method, the observations for Danish data can be classified into 2 clusters. Finally, Figure 5 illustrates the total intra-cluster variation for different values of k, where it is observed that using the gap statistic method, the number of optimal clusters is $k = 1$.

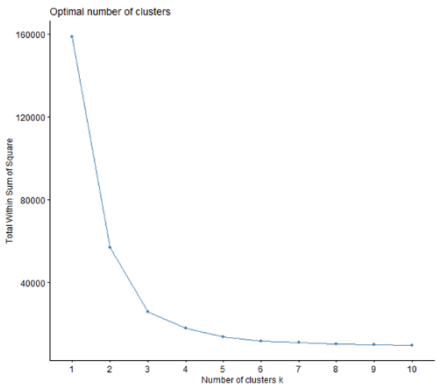

Fig. 3. Number of clusters using the within sum of squares method.

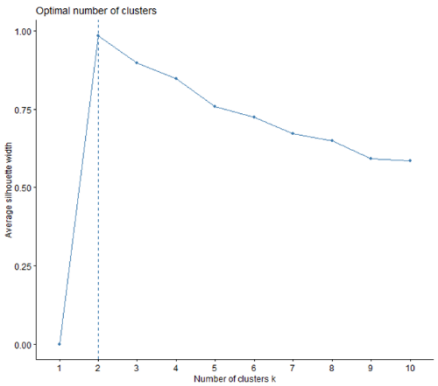

Fig. 4. Number of clusters using the average silhouette width method.

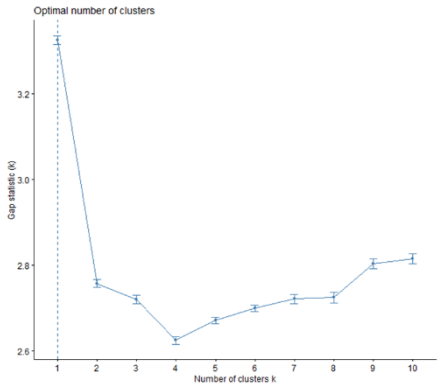

Fig. 5. Number of clusters using the gap statistic method.

For k-means clustering, we observe that the last 3 ordered (largest ones) observations of the Danish data are grouped into one cluster for $k = 2$, 3 and 4 – see Table 3. This is interesting because when [26] proposed a data trimming technique for the Danish data, these observations were also excluded from the new trimmed data, indicating that these are extremely high compared to the rest of the data (see Section 3).

Table 3. k-means clustering method.

		Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	
k	**2**	3 *186.7737*	2489 *2.8413*			Number of obs. per cluster *Cluster mean*
	3	3 *186.7737*	79 *22.3693*	2410 *2.2011*		Number of obs. per cluster *Cluster mean*
	4	3 *186.7737*	24 *35.5797*	119 *12.7706*	2346 *2.0027*	Number of obs. per cluster *Cluster mean*

3 Mixture models on trimmed data

Abu Bakar and Nadarajah [26] proposed a data trimming technique which excludes 'extreme values' that satisfy the condition

$$\ln\left[\frac{x_i}{\max(\boldsymbol{x})}\right] > d \tag{1}$$

where $\boldsymbol{x}$ denotes the complete dataset, i.e., $\boldsymbol{x} = \langle x_1, x_2, \ldots, x_n \rangle$ and $d = -1$**.** Let us consider the last six observations of the ordered Danish data to see which values satisfy (1) and should be excluded. The last six observations of the dataset are: 56.22543, 57.41064 65.70749, 144.65759, 152.41321, and 263.25037, respectively. By substituting the values of the observations into (1), we get -1.543736, -1.522876, -1.387893, -0.5987361, -0.5465102, and 0, respectively. Therefore, the observations 144.65759, 152.41321, and 263.25037 are excluded from the trimmed data because they satisfy the condition set in (1). Abu Bakar and Nadarajah [26] proposed 17 two-component mixture models with the inverse transformed gamma (ITG) as the first component and the second component being either the distributions from the transformed beta family or the transformed gamma family. The 17 distributions considered for the second component can be classified into two categories:

(i) The transformed beta family (9 in total)

- 4-parameter: Transformed beta distribution.
- 3-parameter: Generalized Pareto, Burr, and inverse Burr distributions.

- 2-parameter: Pareto, inverse Pareto, loglogistic, paralogistic, and inverse paralogistic distributions.

(ii) The transformed gamma family (8 in total)

- 3-parameter: Transformed gamma and ITG distributions
- 2-parameter: Gamma, inverse gamma, Weibull and inverse Weibull distributions
- 1-parameter: Exponential and inverse exponential distribution.

Table 4 is adopted from [26], while Table 5 is based on our calculations, where the values in brackets are for the complete data (2492 observations) and the other are for the trimmed data (2492 – 3 = 2489 observations). Let p denote the number of parameters of the model. Three information criteria – the negative log-likelihood (NLL), the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC) are considered for model selection. In general, the lower the value for the information criterion, the better the fit to the data and/or the simpler the model, i.e., the less parameters the model has. Let $\ell(\theta)$ denote the maximised log-likelihood function of a model, then the NLL is defined as NLL $= -\ell(\theta)$. The AIC is defined as AIC = 2NLL + $2p$ and the BIC is defined as BIC = 2NLL + $p\log(n)$, where n is the number of observations. Abu Bakar and Nadarajah [26] concluded that their top three models were the mixture of ITG and Burr distributions, the mixture of ITG and paralogistic distributions and the mixture of ITG and inverse paralogistic distributions (these are boldfaced in Table 4).

The values in Table 5 provide slightly larger likelihood values (lower NLL values) which tells us that the results in Table 4 are sub-optimal solutions. If we also take a closer look at the NLLs for the complete data and trimmed data in Table 4, it is observed that some of the NLLs for the complete data are smaller than the NLLs for the trimmed data, which may imply the method of parameter estimation used is not efficient. Contrary to [26], based on Table 5, using the BIC as a model selection criterion, the models which provide the best fit are the mixture of ITG and transformed beta distributions, the mixture of ITG and Burr distributions and the mixture of ITG and loglogistic distributions, respectively. The 'nlm ()' function and the 'nlminb ()' function in R package 'stats' were used as optimization functions to find the maximum likelihood parameter estimates. For most of the results in Table 5 convergence was reached. The likelihood ratio tests, p parameter estimates for each of the mixture models and risk measures for the complete data and the trimmed data have been left for future research due to the space constraint. Such an analysis would require its own separate manuscript as its length would be approximately the size of [26].

Table 4. Summary of NLL, AIC, BIC values of the mixture models which are provided in [26] for the complete Danish data (in parenthesis) and the trimmed data.

First component	Second component	p	NLL	AIC	BIC
ITG	Transformed beta	8	(3802.128) 3793.856	(7620.256) 7603.712	(7666.823) 7650.279
ITG	**Burr**	**7**	**(3788.123) 3785.672**	**(7590.246) 7585.344**	**(7630.991) 7626.090**
ITG	Inverse Burr	7	(3924.832) 3932.253	(7863.664) 7878.507	(7904.409) 7919.252
ITG	Generalized Pareto	7	(3793.655) 3795.445	(7601.310) 7604.891	(7642.056) 7645.637
ITG	Pareto	6	(3846.816) 3845.978	(7705.632) 7703.956	(7740.557) 7738.881
ITG	Inverse Pareto	6	(3800.299) 3800.884	(7612.598) 7613.768	(7647.523) 7648.693
ITG	**Paralogistic**	**6**	**(3788.902) 3788.873**	**(7589.804) 7589.746**	**(7624.729) 7624.671**
ITG	**Inverse paralogistic**	**6**	**(3789.921) 3789.672**	**(7591.843) 7591.343**	**(7626.768) 7626.268**
ITG	Loglogistic	6	(3854.597) 3854.665	(7721.195) 7721.330	(7756.120) 7756.255
ITG	Transformed gamma	7	(3825.021) 3821.532	(7666.042) 7659.064	(7712.609) 7705.609
ITG	ITG	7	(3797.648) 3805.058	(7609.297) 7624.115	(7650.043) 7664.861
ITG	Gamma	6	(3802.169) 3860.177	(7616.339) 7732.353	(7651.264) 7767.278
ITG	Inverse gamma	6	(3799.366) 3799.518	(7610.732) 7611.035	(7645.657) 7645.96
ITG	Weibull	6	(3972.072) 3821.375	(7956.144) 7654.750	(7991.069) 7689.675
ITG	Inverse Weibull	6	(3819.280) 3820.417	(7650.56) 7612.274	(7685.485) 7687.76
ITG	Exponential	5	(3801.137) 3801.088	(7612.274) 7612.175	(7641.379) 7641.28
ITG	Inverse exponential	5	(3809.623) 3805.569	(7629.245) 7621.138	(7658.349) 7650.243

Table 5. Summary of NLL, AIC, BIC values of the mixture models based on *our calculations* for the complete Danish data (in parenthesis) and the trimmed data.

First component	Second component	p	NLL	AIC	BIC
ITG	**Transformed beta**	**8**	**(3778.845) 3742.260**	**(7573.691) 7500.520**	**(7620.257) 7547.087**
ITG	**Burr**	**7**	**(3783.060) 3746.480**	**(7580.121) 7506.960**	**(7620.867) 7547.706**
ITG	Inverse Burr	7	(3786.849) 3751.927	(7587.698) 7517.854	(7628.444) 7558.600
ITG	Generalized Pareto	7	(3792.268) 3756.866	(7598.536) 7527.731	(7639.282) 7568.477
ITG	Pareto	*6*	(3801.026) 3765.294	(7614.053) 7542.587	(7648.978) 7577.512
ITG	Inverse Pareto	6	(3798.931) 3763.505	(7609.862) 7539.009	(7644.787) 7573.934
ITG	Paralogistic	6	(3788.788) 3754.107	(7589.577) 7520.213	(7624.502) 7555.138
ITG	Inverse paralogistic	6	(3789.419) 3753.816	(7590.838) 7519.631	(7625.763) 7554.556
ITG	**Loglogistic**	**6**	**(3786.852) 3751.998**	**(7585.704) 7515.997**	**(7620.629) 7550.922**
ITG	Transformed gamma	7	(3788.307) 3753.570	(7592.615) 7521.139	(7639.181) 7561.885
ITG	ITG	7	(3794.448) 3760.151	(7602.897) 7530.302	(7643.643) 7575.047
ITG	Gamma	6	(3792.502) 3757.557	(7597.005) 7527.113	(7631.930) 7562.038
ITG	Inverse gamma	6	(3798.563) 3762.558	(7609.126) 7537.117	(7644.051) 7572.042
ITG	Weibull	6	(3788.999) 3754.317	(7589.998) 7520.633	(7624.923) 7555.558
ITG	Inverse Weibull	6	(3798.622) 3763.135	(7609.244) 7538.269	(7644.169) 7573.194
ITG	Exponential	5	(3800.993) 3765.319	(7611.986) 7540.638	(7641.090) 7569.742
ITG	Inverse exponential	5	(3798.926) 3763.498	(7607.852) 7536.996	(7636.957) 7566.100

4 Best composite model

Grün and Miljkovic [25] stated in their study that the composite Weibull-Inverse Weibull model is the best composite model for the Danish data out of the 256 fitted composite models they proposed, while Maghsoudi et al. [17] stated that their proposed composite Weibull-ITG model is the best composite model for the Danish data. In this section, we investigate whether the composite Weibull-ITG model with 5 parameters provides a significant improvement in fit compared to the composite Weibull-Inverse Weibull model with 4 parameters. The likelihood ratio test is performed using the values in Table 6.

Table 6. Summary of the information criteria for the composite Weibull-Inverse Weibull and the Weibull-inverse transformed gamma models.

Head	Tail	p	NLL	AIC	BIC
Weibull	Inverse Weibull	4	3820.010	7648.020	7671.300
Weibull	ITG	5	3817.939	7645.878	7674.982

The classical likelihood ratio test can be used to assess the goodness-of-fit of two models, where one is a subset of the other model (or is nested in the other model). That is, the null model is a special case of the alternative model. The test statistic is defined as $D = -2[\ell(\theta_0) - \ell(\theta_1)] = 2[NLL_0 - NLL_1]$, where $\ell(\theta_0)$ and $\ell(\theta_1)$ are the maximised log-likelihood values of the non-nested model and the nested model, respectively and NLL_0 and NLL_1 are the corresponding negative log-likelihood values. Under the null hypothesis, the assumed model is the composite Weibull-Inverse Weibull model, i.e., the composite Weibull-ITG model does not provide a significant improvement in fit over the composite Weibull-Inverse Weibull model for the Danish data. The test statistic is $D = 2[3820.01 - 3817.939] = 4.142$ and the critical value at a 5% significance level is 3.841459. Therefore, the null hypothesis is rejected, and the alternative model is a significant improvement over the null model, i.e., the composite Weibull-ITG model does provide a significant improvement in fit over the composite Weibull-Inverse Weibull model for the Danish data.
Risk measures are essential for actuaries, investors, and financial institutions to make informed decisions about investments and risk management strategies. The appropriateness of the proposed models is also validated by comparing empirical risk estimates and theoretical risk estimates. Two popular risk measures, Value at Risk (VaR) and Tail Value at

Risk (TVaR) were estimated at 99% security levels. VaR is defined as a quantile risk measure and TVaR is defined as the expected loss in the event that the loss exceeds VaR. Let $F(\cdot)$ and $F^{-1}(\cdot)$ denote the cumulative distribution function (cdf) and inverse cdf of a continuous random variable X, respectively. Then, the VaR of X at a $100p\%$ security level denoted by $\mathrm{VaR}_{\mathrm{p}}(X)$, is the $100p\%$ quantile of F such that

$$P\left(X < \mathrm{VaR}_{\mathrm{p}}(X)\right) = p, \qquad F^{-1}(p) = \mathrm{VaR}_{\mathrm{p}}(X).$$

The TVaR of X at a $100p\%$ security level denoted by $\mathrm{TVaR}_{\mathrm{p}}(X)$, represents the average of all VaR values exceeding security level, p, such that

$$\mathrm{TVaR}_{\mathrm{p}}(X) = \frac{1}{1-p}\int_{p}^{1} \mathrm{VaR}_{\mathrm{u}}(X)du = \mathbb{E}\left[X|X > \mathrm{VaR}_{\mathrm{p}}(X)\right].$$

Note that VaR can be interpreted as the lower bound for the capital required to avoid insolvency, whereas TVaR can be interpreted as the expected value of total loss, given that the loss exceeds VaR. The composite Weibull-Inverse Weibull model closely matches the empirical TVaR at a 99% security level whereas the composite Weibull- ITG model closely matches the empirical VaR at a 99% security level, see Table 7. More importantly, unlike the composite Weibull-Inverse Weibull model, the composite Weibull- ITG model does not underestimate the VaR at a 99% security level.

Table 7. Summary of the risk estimates for the composite Weibull-Inverse Weibull and the Weibull-ITG models and their percentage deviation with respect to empirical estimates in parenthesis.

		$\mathbf{VaR_{0.99}}$	$\mathbf{TVaR_{0.99}}$
Empirical estimates		**24.613**	**54.603**
Head	**Tail**		
Weibull	Inverse Weibull	22.770 (-7.5%)	63.860 (17%)
Weibull	ITG	25.107 (2.0%)	81.031 (48.4%)

5 Splicing

For the subset of the Danish data with 2167 observations greater than 1 million DKKs, [23] proposed a splicing model with a mixture of Erlang distributions (ME) for the body and a Pareto distribution for the tail. In this section, we apply the splicing model to the complete dataset of 2492 observations. This model merges the flexibility of the ME distribution with the ability of the Pareto distribution to model the extreme events. The extreme value theory (EVT) approach of splicing differs from the approach taken when constructing composite models. The main difference is that on the EVT approach, the threshold is determined first, whereas with the composite models, the threshold is solved simultaneously as a function of the model parameters. The mean excess plot [37] and the Hill plot [38] were used to estimate a suitable threshold, where applicable. For the mean excess plot, the point beyond t for which the points are increasing is where the Pareto distribution is appropriate for modelling the behaviour of the losses given $X > t$. The R add-on package 'ReIns' (see [39]) is used for statistical computation.

The mean excess plot in Figure 6 is ultimately increasing for the Danish data (except the two extreme values) which suggests that the underlying distribution is heavy-tailed. Chernobai et al. [40] state that a mean excess plot that has an upward slope suggests a distribution like that of the Pareto distribution. The underlying distribution in the tail is heavier than the exponential distribution and lighter than the Pareto distribution as seen from the QQ plots in Figure 7.

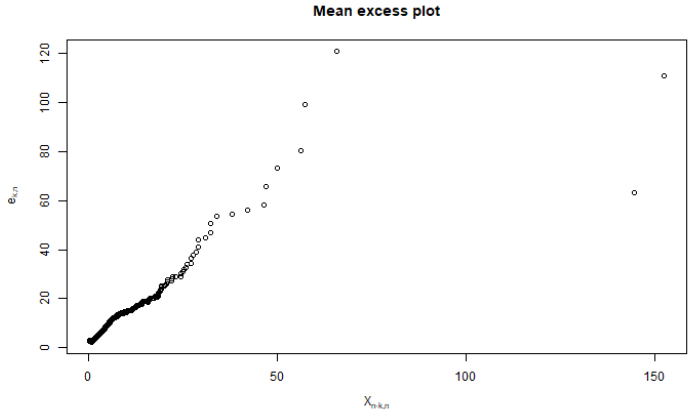

Fig. 6. The mean excess plot for the Danish data.

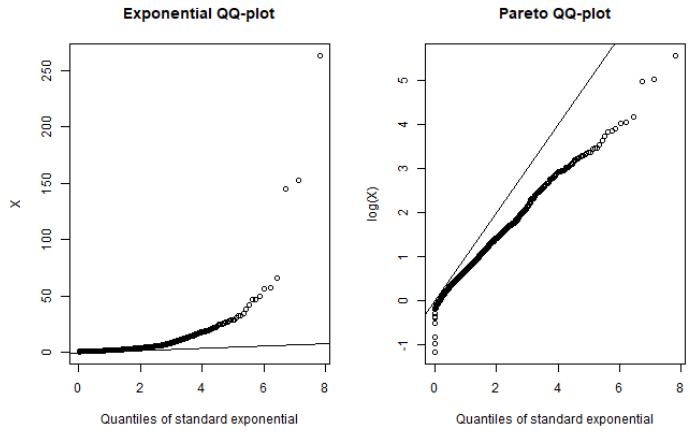

Fig. 7. The exponential QQ plot and the Pareto QQ plot for the Danish data.

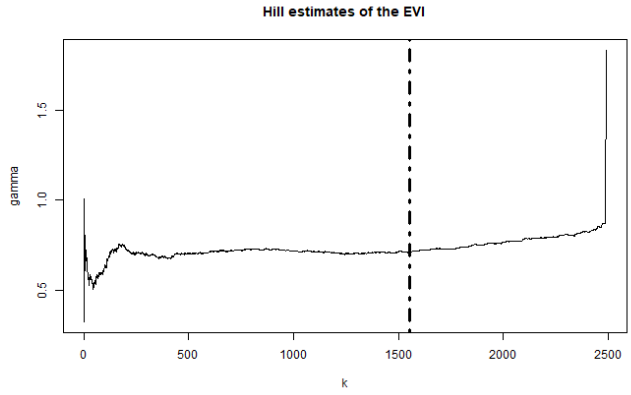

Fig. 8. The Hill plot with k that minimizes the AMSE.

The splicing point or the threshold point is chosen by considering the k value on the Hill plot where the Asymptotic Mean Squared Error (AMSE) is at its minimum, i.e. AMSE = Variance + Bias^2. Therefore, the values for which the AMSE is minimised are $k = 1552$ and $\gamma = 0.71$. Therefore, the threshold is the $(n-k)$th observation, where n is the number of observations. The threshold is the $(2492-1552)$th observation which is 1.384293 – see Figure 8. Therefore, 37.7% of the observations are modelled by the mixture of Erlangs, and 62.3% of the observations are modelled by the Pareto distribution. The proposed spliced model is optimal for capturing the body of the data – see Figure 9. From the QQ plot, one can see that the spliced model is slightly heavier (lighter) than the underlying model of the data where the plotted points are below (above) the reference line. Overall, the model seems adequate even in the tail area.

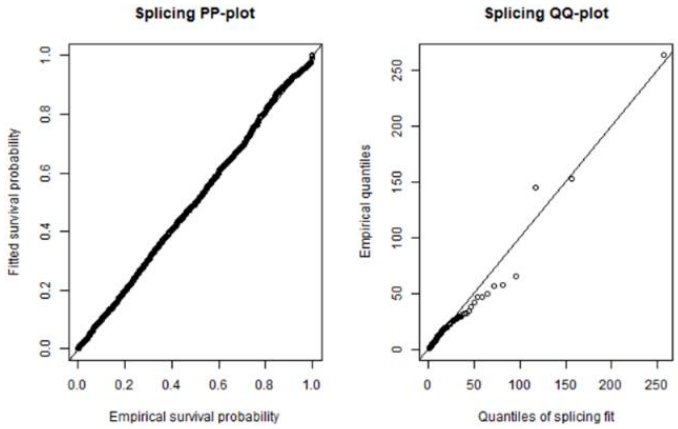

Fig. 9. The splicing PP plot and the splicing QQ plot.

The composite Weibull-Inverse Weibull model has a NLL of 3820.01 and a BIC of 7671.3 (see Table 6 above) which are both less than the ones for spliced model in Table 8. It seems that when it comes to overall fit, this particular spliced model is not as good as some of the composite models. Proper investigation into splicing with different mixtures for the body may show that other spliced models perform better for modelling insurance claims – this can be studied further in the future.

Table 8. Summary of the information criteria for the spliced model.

Head	Tail	NLL	AIC	BIC
Mixture of Erlangs	Pareto	3850.607	7709.215	7732.498

Table 9. Summary of the risk estimates for the spliced model and percentage deviation with respect to empirical estimate in parenthesis.

		$\mathbf{VaR_{0.95}}$	$\mathbf{VaR_{0.99}}$	$\mathbf{TVaR_{0.95}}$	$\mathbf{TVaR_{0.99}}$
Empirical estimates		8.406	24.614	22.155	54.604
EVT methods					
Body	**Tail**				
Mixture of Erlangs	Pareto	8.316 (-1.1%)	26.102 (6.0%)	28.745 (29.7%)	90.225 (65.2%)

The main difference with the risk measures obtained for the composite Weibull-Inverse Weibull and the spliced model in Table 9 is that the spliced model does not underestimate the VaR at a 95% and 99% security level. However, the composite Weibull-Inverse Weibull

model provides risk estimates for the TVaR that closely match the empirical estimates when compared to the spliced model.

6 Concluding remarks

Based on the discussion herein, it is suspected that some of the minor difference in analytical results of the Danish fire claims data is due to the use of either different software or different packages. The short analysis performed in this work was done using R (readers can request these codes from the authors). The varying number of clusters being identified by different methods of clustering used here justifies why many authors have experimented with this data using composite and mixture models. Finally, more research work on splicing techniques in the context of this dataset is required. Overall, the purpose of this paper is to highlight the importance of the Danish fire claims and illustrate its usefulness in the empirical actuarial and extremes fields to validate certain hypotheses because real-life datasets are very scarce, as they are subject to a lot of proprietary and privacy laws.

References

1. A.J. McNeil, ASTIN Bulletin: J. IAA, **27**(1), 117 – 137 (1997)
2. S.I. Resnick, ASTIN Bulletin: J. IAA, **27**(1), 139 – 151 (1997)
3. K. Cooray, M.M. Ananda, Scand. Actuar. J., **2005**(5), 321 – 334 (2005)
4. D.P. Scollnik, Scand. Actuar. J., **2007**(1), 20 – 33 (2007)
5. H. Drees, P. Muller, Insur. Math. Econ., **42**(2), 638 – 650 (2008)
6. V. Brazauskas, A. Kleefeld, Insur. Math. Econ, **45**(3), 424 – 435 (2009)
7. C.M. Ausin, M.P. Wiper, L.E. Rosa, Appl. Stoch. Model Bus. Ind., **25**(6), 787 – 805 (2009).
8. J. Carreau, Y. Bengio, Extremes, **12**, 53 – 76 (2009)
9. H. Esmaeili, C. Kluppelberg, Insur. Math. Econ., **47**(2), 224 – 233 (2010)
10. A. Charpentier, A. Oulidi, Stat. Comput., **20**(1), 35 – 55 (2010)
11. M. Pigeon, M. Denuit, Scand. Actuar. J., **2011**(3), 177 – 192 (2011)
12. S. Guillotte, F. Perron, J. Segers, J. R. Stat. Soc. Ser. B Stat. Method., **73**(3), 377 – 406 (2011)
13. M. Eling, Insur. Math. Econ., **51**(2), 239 – 248 (2012)
14. M. Bernardi, A. Maruotti, L. Petrella, Insur. Math. Econ., **51**(3), 617-623 (2012)
15. D.P. Scollnik, C. Sun, N. Am. Actuar. J., **16**(2), 260 – 272 (2012)
16. P. Ruckdeschel, N. Horbenko, Statistics, **47**(4), 762 – 791 (2013)
17. M. Maghsoudi, S.A. Abu Bakar, N.A. Hamzah, *Composite Weibull-Inverse Transformed Gamma distribution and its actuarial* application, in AIP Conference Proceedings, **1605**(1), 1007 – 1012 (2014)
18. S. Nadarajah, S.A. Abu Bakar, Scand. Actuar. J., **2014**(2), 180 – 187 (2014)
19. S.A. Abu Bakar, N.A. Hamzah, M. Maghsoudi, S. Nadarajah, Insur. Math. Econ., **61**, 146 – 154 (2015)
20. E. Calderin-Ojeda, Commun. Stat. Case Stud. Data Anal. Appl., **1**(1), 59 – 69 (2015)
21. E. Calderin-Ojeda, C.F. Kwok, Scand. Actuar. J., **2016**(9), 817 – 836 (2016)

22. T. Miljkovic, B. Grün, Insur. Math. Econ., **70**, 387 – 396 (2016)
23. T. Reynkens, R. Verbelen, J. Beirlant, K. Antonio, Insur. Math. Econ., **77**, 65 – 77 (2017)
24. S.A. Abu Bakar, S. Nadarajah, Z.A. Adzhar, Empir. Econ., **54**, 1503 – 1516 (2017)
25. B. Grün, T. Miljkovic, Scand. Actuar. J., **2019**(8), 642 – 660 (2019)
26. S.A. Abu Bakar, S. Nadarajah, J. Appl. Stat., **46**(5), 835 – 852 (2019)
27. B. Liu, M.M. Ananda, Commun. Stat. - Theory Methods, **52**(21), 7618 – 7631 (2023)
28. W.A. Marambakuyana, S.C. Shongwe, Mathematics, **12**(2), 335 (2024)
29. W.A. Marambakuyana, S.C. Shongwe, J. Stat. Appl. Probab., **13**(3), 1031 – 1044 (2024)
30. A. Davison, SMPracticals: R package version 1.4-3 (2019)
31. B. Pfaff, A. McNeil. evir: Extreme values in R. R package version 1.7-4 (2018)
32. M.L. Delignette-Muller, C. Dutang, J. Stat. Softw., **64**, 1 – 34 (2015)
33. B. Pfaff, A. McNeil, QRM: R package version 0.4-31 (2020)
34. M. Maechler, diptest: R package version 0.76-0 (2021)
35. Statisticat. LaplacesDemon: R package version 16.1.6 (2021)
36. J.A. Hartigan, P.M. Hartigan, Ann. Stat., 70 – 84 (1985)
37. J. Beirlant, Y. Goegebeur, J. Teugels, J. Segers, Statistics of Extremes, (John Wiley & Sons, 2004)
38. B.M. Hill, Ann. Stat., **3**, 1163 – 1174 (1975)
39. T. Reynkens, R. Verbelen, ReIns: R package version 1.0.11 (2023)
40. A.S. Chernobai, S.T. Rachev, F.J. Fabozzi, John Wiley & Sons (2007)