# Statistical Tools for Quantitative Risk Management

Assignment II: Extreme Value Analysis

**Vincent Buekers**
r0754046


G0Q24A
2019-2020

# Fire Claims: European Fire Insurance Portfolio

The data to be analyzed in this assignment consists building and content losses related to fire claims. Question A entails an extreme value analysis with the goal of obtaining reliable quantile estimates. Question B outlines full models for both building and content losses, thus including all claims. Question C is dedicated to examining the tail dependence between the building and content losses. Finally, Question D concludes the analysis with a somewhat exploratory examination of the risk over time.

## Question A

Of interest are the quantiles $Q(1-\frac{1}{1500})$ and $Q(1-\frac{1}{2000})$, i.e. the values exceeded only 1 in 1500 cases and 1 in 2000 cases respectively. In order to obtain these estimates, identifying an appropriate tail distribution is key. As a preliminary means of comparison, the exponential distribution turned out inappropriate since the tails of both building and content losses have much heavier tails than that of the exponential.

In order to obtain quantile estimates, it thus becomes necessary to rely on a different distribution. For heavy right tails, Pareto type distributions might be more appropriate. Indeed, a Pareto-type tail seems plausible for both building and contents since the Pareto QQ-plot is roughly linear for large $X_{n-k,n}$ (Fig. 1). Depending on an appropriate value of $k$, this seems to be the case for both building and content losses. The estimates for the linear slopes $\gamma$ follow from the Hill estimator $H_{k,n} = \hat{\gamma}_k$ or its bias reduced version $\hat{\gamma}_k^{BR}$. This, in turn, could allow to identify an optimal choice of $k$.

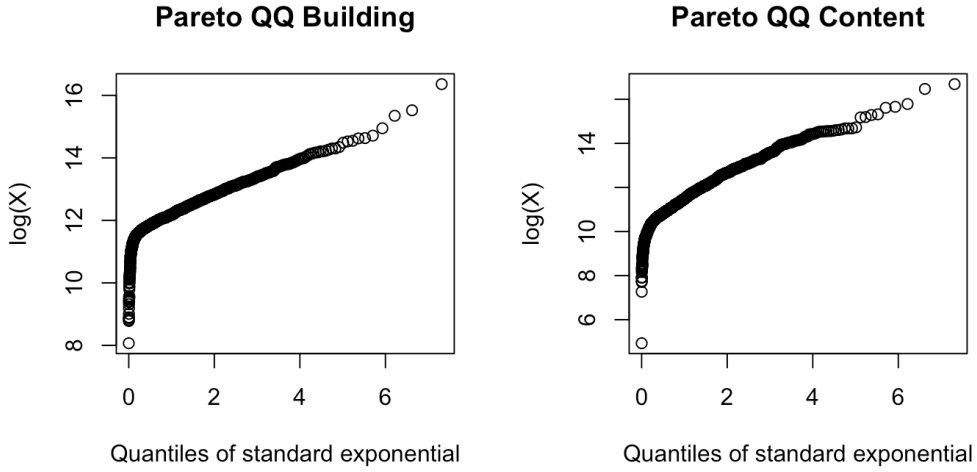


Figure 1: Pareto-plot

The slope parameters have been estimated using both the standard Hill procedure as well as its bias reduced version, resulting in the estimates displayed in Fig. 2. The estimates w.r.t. building losses are quite stable for a relatively large interval of $k$, whereas that interval for content losses is much narrower. The `Hill.kopt()` function has determined the optimal $k$ to be 515 and 55 for building and content respectively. The corresponding slope estimates are $\hat{\gamma}_{515} = 0.5995$ and $\hat{\gamma}_{515}^{BR} = 0.5965$ for building, while they are $\hat{\gamma}_{55} = 0.5950$ and $\hat{\gamma}_{55}^{BR} = 0.5907$ for content. This is also visualized in Fig. 2.

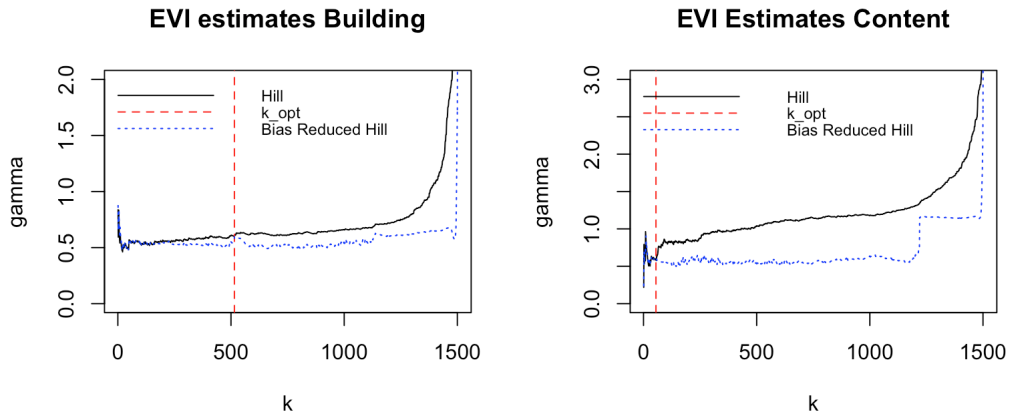


Figure 2: EVI Estimates

The quantile estimates following from the slope parameters obtained above are listed in below (Table 1) for the optimal k (vertical dashed line). Moreover, the Weissman estimates and their bias reduced versions are displayed in function of $k$ (Fig. 3), along with a non-parametric estimate (horizontal dashed line). As can be seen, these estimates are to a large extent in agreement for the $k$ as selected by `Hill.kopt()`. Fortunately, this is reassuring w.r.t. to that particular choice of $k$. The estimates are particularly volatile for the content losses, in which a larger $k$ would quickly result in much higher quantiles. In contrast, the quantiles for the building losses seem somewhat more flexible w.r.t. the choice of $k$.

(a) Building

| | $p = 1/1500$ | $p = 1/2000$ |
|---|---|---|
| $\hat{q}^{+}_{k,p}$ | 8865941 | 10534687 |
| $\hat{q}^{BR}_{k,p}$ | 8730569 | 10364948 |

(b) Content

| | $p = 1/1500$ | $p = 1/2000$ |
|---|---|---|
| $\hat{q}^{+}_{k,p}$ | 12418773 | 14737353 |
| $\hat{q}^{BR}_{k,p}$ | 12270698 | 14543618 |

Table 1: Quantile Estimates

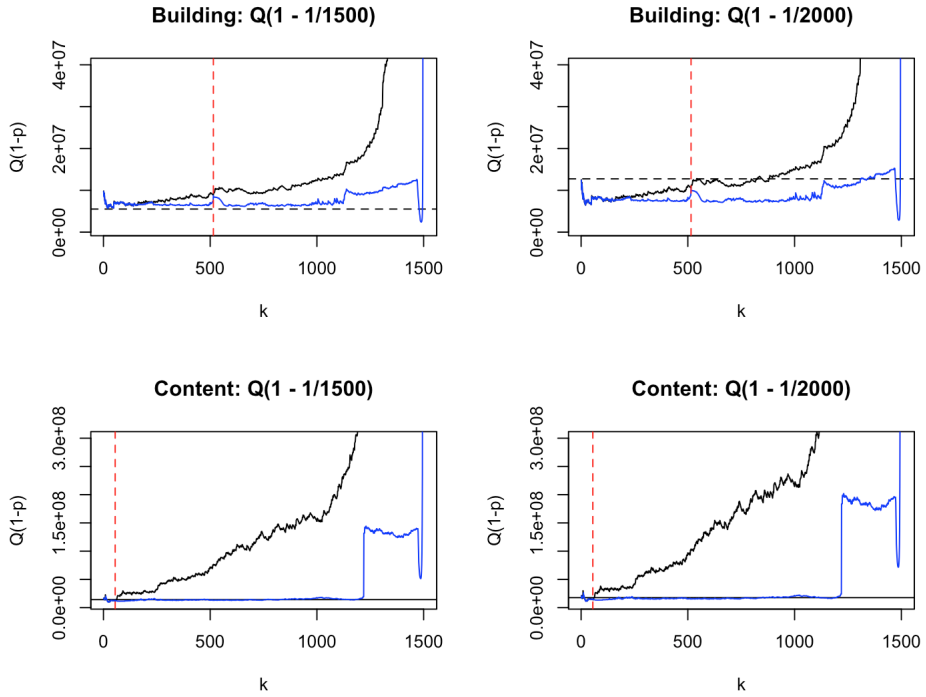


Figure 3: Quantile Estimates

# Question B

As opposed to the tail modelling conducted in the previous section, it is now of interest to provide a model for all claims, for both building and content losses. To that end, splicing is implemented in order to separately model 'small' and 'large' claims by means of two component densities. This requires the selection of a certain cut-off between these components. when considering the optimal $k$ values from the preceding section, the corresponding cut-offs seem somewhat restrictive (Fig. 4, red dashed line). Consequently, a more appropriate cut-off is heuristically selected (Fig. 4, black dashed line). These correspond to $X_{n-20,n}$ and $X_{n-25,n}$ for building and content losses respectively.

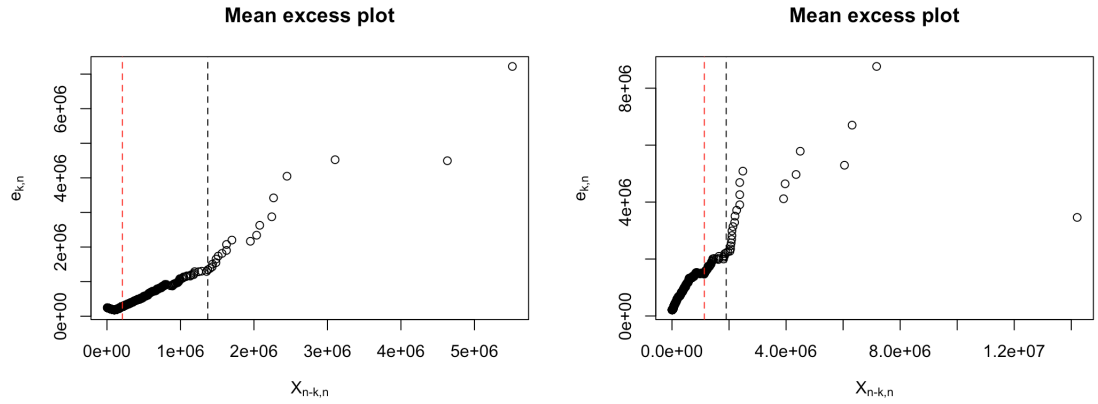


Figure 4: Mean excess plots: Splicing cut-offs

Based on these cut-offs, several full models are estimated, all in the form Mixed Erlang and Pareto Splicing. For both building and content losses, Model 1 corresponds to a 3-component mixture consisting of 2 mixed Erlangs and a Pareto using both cut-offs. Model 2 represents a 2-component mixture using the cut-off based on the optimal $k$ from the extreme value analysis. Lastly, Model 3 is also a 2-component mixture, yet based on the heuristic cut-off that was chosen. As indicated by both Information Criteria, Model 2 is selected for both variables. Consequently, the splicing models using the optimal $k$ splice should be preferred.

(a) Building

| | AIC | BIC |
|---|---|---|
| Model 1 | 39494.94 | 39537.46 |
| Model 2 | 39492.87 | 39524.76 |
| Model 3 | 39652.23 | 39684.11 |

(b) Content

| | AIC | BIC |
|---|---|---|
| Model 1 | 38863.83 | 38906.34 |
| Model 2 | 38860.44 | 38892.32 |
| Model 3 | 38916.86 | 38948.75 |

Table 2: Model selection: Information Criteria

- The model parameters characterizing model 2 for building losses are the following:
  - EVA parameters: $\gamma = 0.599$, $k = 515$, $\pi = 0.3429$
  - ME parameters: $t^l = 209941$ $M = 2$, $\mathbf{r} = (2, 9)$, $\alpha = (0.0658, 0.934)$, $\lambda^{-1} = 16195.68$
- The model parameters characterizing model 2 for content losses are the following:
  - EVA parameters: $\gamma = 0.595$, $k = 55$, $\pi = 0.0366$
  - ME parameters: $t^l = 1133422$ $M = 2$, $\mathbf{r} = (2, 6)$, $\alpha = (0.9064, 0.0936)$, $\lambda^{-1} = 86288.5$

Lastly, the model fit can be assessed by comparing the fitted survival function of the spliced distribution with the empirical survival function by means of PP-plots and their logged versions (Fig. 5). It can be seen that the fitted distributions correspond very closely to their empirical counterparts. As such, choosing model 2 seems appropriate indeed, both w.r.t. the building and content losses alike.

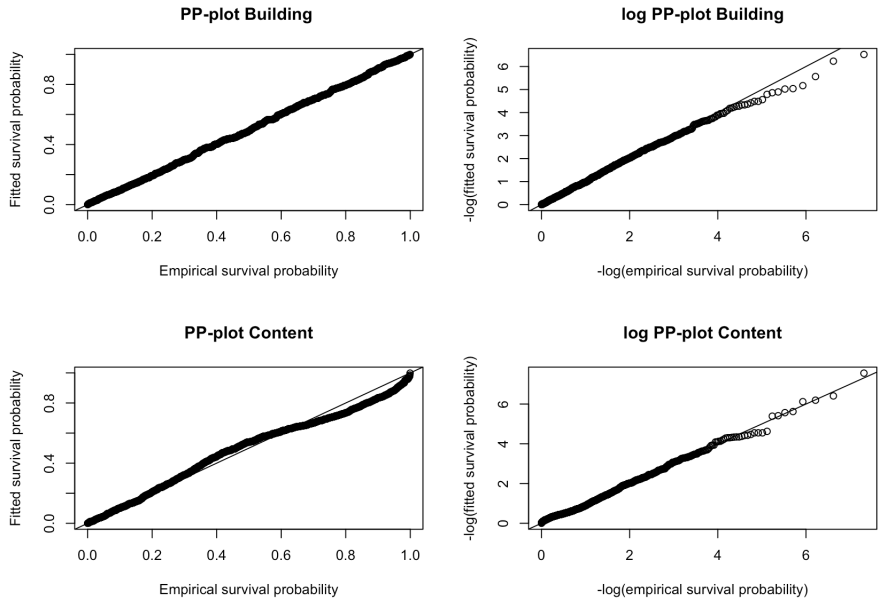


Figure 5: (log) PP-plots for building (top) and content (bottom)

# Question C

In this section, the coefficient of upper tail dependence is considered. I was unable to obtain it for the empirical copula. Therefore, I have considered standard copulas closely resembling the empirical one, namely the Student copula and the Gumbel Copula. As can be seen (Fig. 3), these copulas are indeed similar to the empirical. Note that the visualized Student copula is estimated using $df = 1$, whereas $\theta = 1.175$ is used for the Gumbel Copula. The corresponding upper tail coefficients are tabulated below (Table 3). Based on the AIC however, the Gumbel copula should clearly be preferred. As such, my suggested approximation for the empirical $\lambda_u$ is 0.1970.

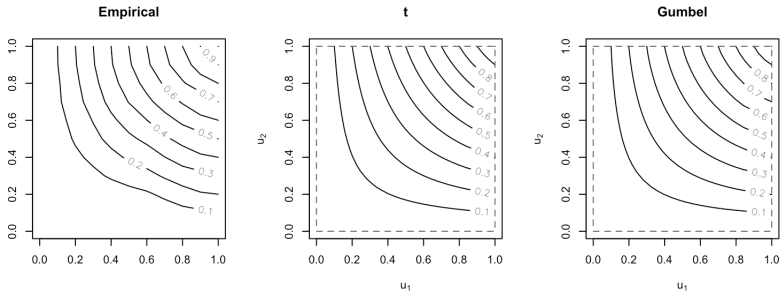


Figure 6: Empirical, Student and Gumbel copula

| | $\lambda_u$ | AIC |
|---|---|---|
| $C^{Student}_{rho,1}$ | 0.2655 | -47.6656 |
| $C^{Gumbel}$ | 0.1970 | -132.8761 |

Table 3: Coefficients of upper tail dependence & Copula fit

# Question D

In this last part of the analysis, a closer look is taken at the risk over time, without modeling of any kind. In common logic, the larger the claim sizes, the higher the associated risk should be. However, just looking at the values over time provides little insight on the time dynamic and its underlying risk (Fig. 7). However, there appear to be some more volatile clusters, such as during

1980 for building losses and 1990 for content losses. For purposes of visual appeal, the losses are log-transformed.

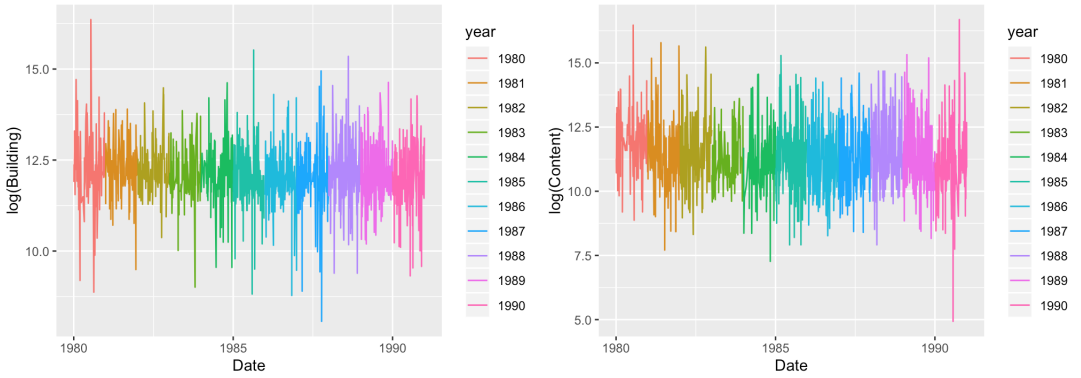


Figure 7: log time series plots

Next, it might be worthwhile to investigate monthly patterns. when visualizing the claims per month, an interesting pattern emerges (Fig. 8). It seems as though the most extreme building losses happen during summer and to a lesser extent also during fall. Moreover, content losses exert a similar pattern as represented by the color gradient. Still, Fig. 8 makes it hard to distinguish the smaller claims based on the season they appeared in. To that end, the distributional variation is more closely examined for the claim sizes once again on log-scale (Fig. 9). The distributions seem different for content losses particularly, where the fall season appears to be most risky.

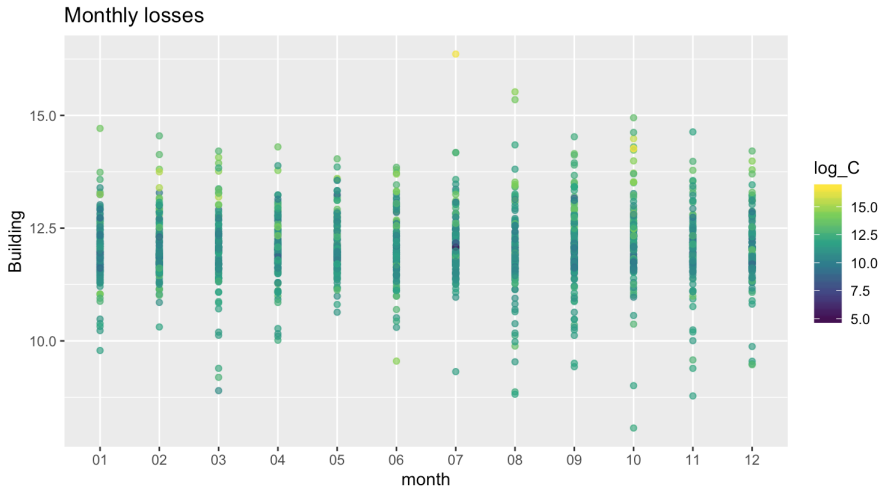


Figure 8: Losses over time

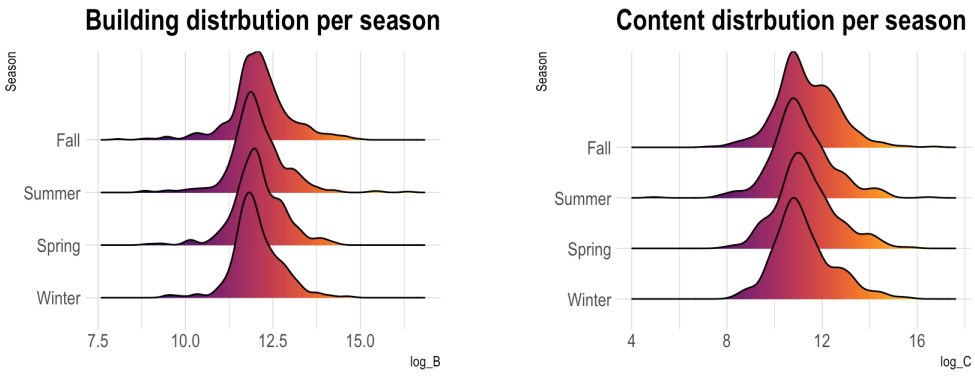


Figure 9: Distributional variation per season