International Journal of Wildland Fire **2009**, *18*, 983–991. doi:10.1071/WF07044_AC

Accessory publication

Spatial and temporal extremes of wildfire sizes in Portugal (1984–2004)

P. de Zea Bermudez[A], *J. Mendes*[B], *J. M. C. Pereira*[C], *K. F. Turkman*[A,E] *and M. J. P. Vasconcelos*[D]

[A]Departamento de Estatística e Investigação Operacional (DEIO) and Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL), University of Lisbon, PT-1749-016 Lisbon, Portugal.

[B]Instituto Superior de Estatística e Gestão de Informação (ISEGI), New University of Lisbon, PT-1070-124 Lisbon, Portugal.

[C]Deparment of Forestry and Center for Forest Studies, Instituto Superior de Agronomia (ISA), Technical University of Lisbon, PT-1349-017 Lisbon, Portugal.

[D]Tropical Research Institute, PT-1300-344 Lisbon, Portugal.

[E]Corresponding author. Email: kfturkman@fc.ul.pt

This manuscript is a complementary publication to supplement the article referred above. The statistical techniques used are standard in extreme value theory (see Embrechts *et al.* (1997) or the book by Coles (2001)).

Global Analysis

We start by doing a preliminary data analysis in order to capture the main features of the Portuguese wildfire data. The strong spatial and temporal variations shown by the data will be addressed at later stages.

Figure 1 shows the behavior of all the data set, as well as the observations ≥ 100 hectares.

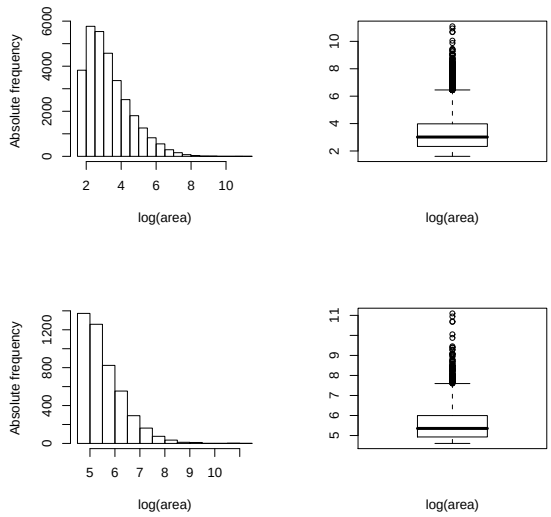

Figure 1: Histograms and boxplots - global sample (top row) and observations ≥ 100 (bottom row)

The box-plot of the complete data clearly displays a considerable number of large values. It is quite evident from this plot that any model fitted to the complete data set, no matter how good it is, would result in the underestimation of large fires. This suggests that large fires should be modeled separately. Asymptotic models suggested by extreme value theory are good candidates. Moreover, the two histograms and the two boxplots presented in Figure 1 look the same. This suggests that the data is heavy-tailed.

The first step is to find out the extent of the tail heaviness.A common way to assess the heaviness of a tail is by means of the QQ-plots. Let $x_{i:n}$ be the ith observation of an ascending ordered sample $(x_{1:n} \leq x_{2:n} \leq ... \leq x_{n:n})$. In an **Exponential QQ-plot**, the exponential quantiles, given as $-\ln(1-p_i)$, where $p_i = i/(n+1)$, are plotted as a function of the ordered sample, $x_{i:n}$, $i = 1, 2, ..., n$. In a **Pareto QQ-plot**, $-\log(1-p_i)$ are plotted as a function of $\log(x_i)$. As is known, the Pareto distribution is heavier tailed than the exponential distribution. Exponential distribution is known to be in the domain of attraction of the Gumbel distribution, whereas, the Pareto distribution is in the domain of attraction of the Fréchet distribution. Therefore, both distributions are commonly used as benchmark to assess the thickness of the distribution's tail and consequently for choosing the appropriate model for large values. The linearity of a QQ-plot indicates that the underlying model might be adequate to the data. The exponential and the Pareto tails are given by $\bar{F}(x) = \exp(-\lambda x)$, $x > 0$, $\lambda > 0$ and $\bar{F}(x) = x^{-\lambda}$, $x > 1$, $\lambda > 0$, respectively.

Figure 2 shows the exponential and Pareto QQ-plots of the global sample and of the observations equal or larger that 100 hectares.

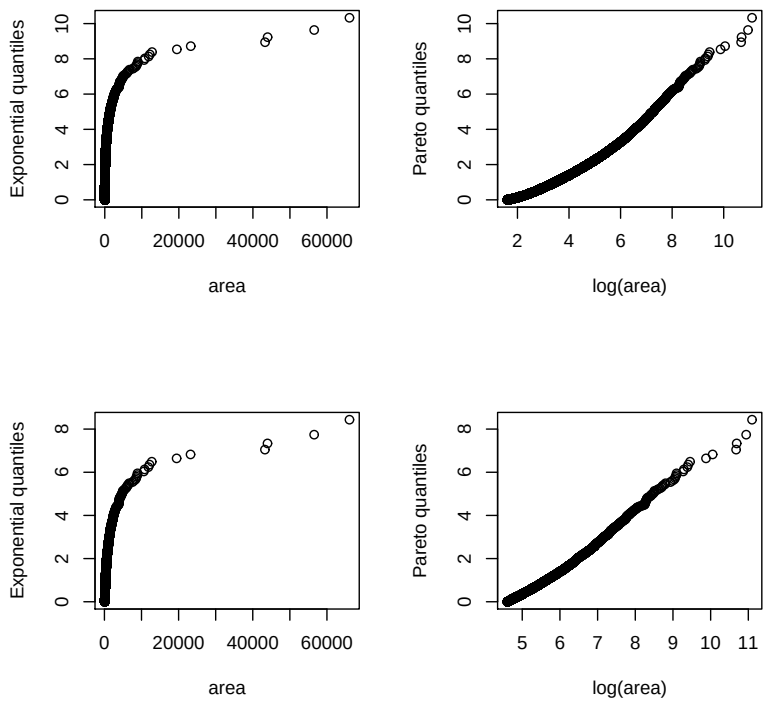

Figure 2: Exponential (left) and Pareto (right) QQ-plots - global sample (top row) and observations ≥ 100 (bottom row)

The linearity of the plot is evident for the Pareto QQ-plots, specially for the "$\geq$ 100 hectares" sample. The overall conclusion of this preliminary data analysis is that the wildfires with scars above 100 hectares satisfy the threshold stability criteria and that the data are consistent with a heavy-tailed distribution, belonging to the domain of attraction of a Fréchet distribution. The statistical methods of inference on large observations are based on the assumption that the data are independent and identically distributed. However, by nature, the data clearly show spatial and temporal variations which invalidate this assumption. Therefore, a careful analysis of the temporal and the spatial variation inherent in the large values is needed.

Temporal Analysis

In order to assess the temporal variation that is expected to exist in large observations, QQ-plots for each of the 21 years of data are represented in Figures 3 - 7. The Pareto QQ-plots of the 21 years look very similar. Because of their evident linearity, we conclude the tails of the underlying distributions (for all the years) are heavy-tailed.

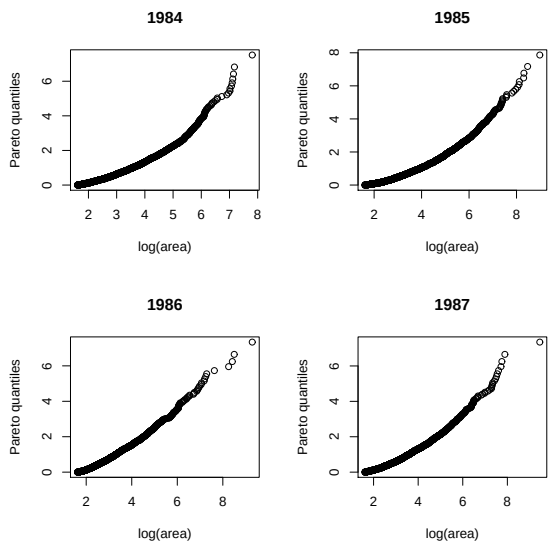

Figure 3: Pareto QQ-plots of the wildfire sizes (1984-1987)

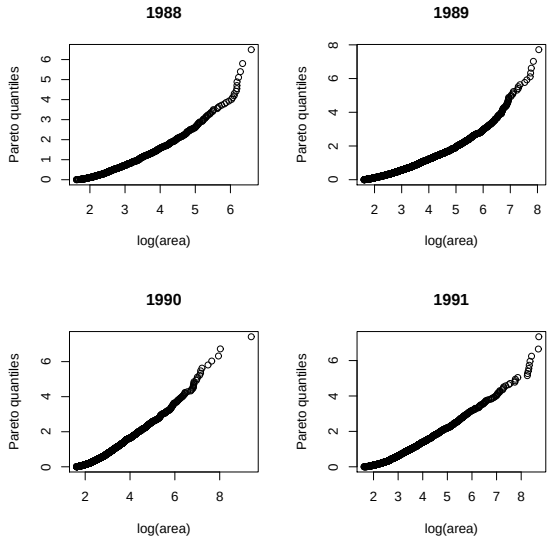

Figure 4: Pareto QQ-plots of the wildfire sizes (1988-1991)

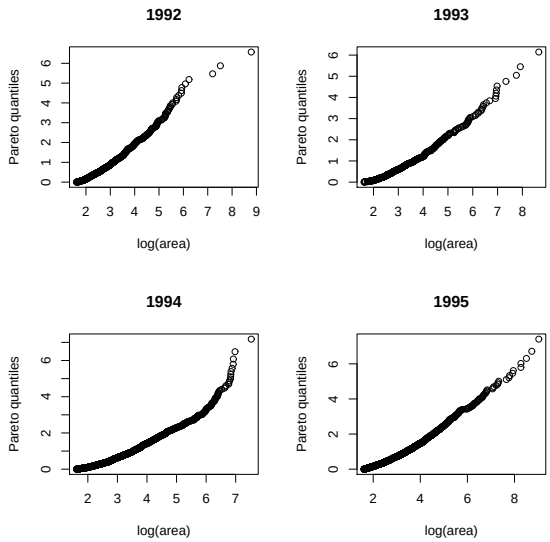

Figure 5: Pareto QQ-plots of the wildfire sizes (1992-1995)

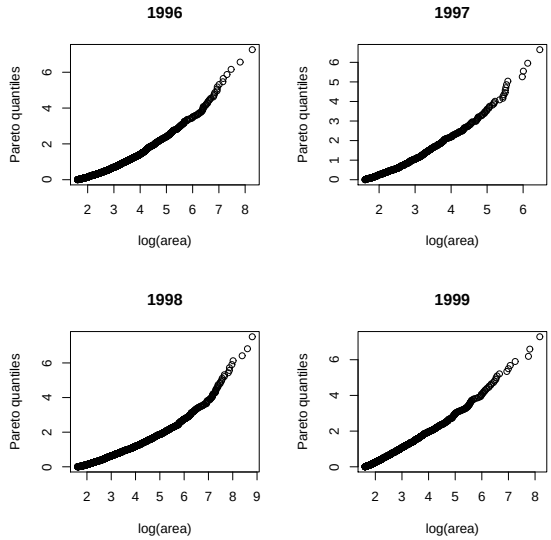

Figure 6: Pareto QQ-plots of the wildfire sizes (1996-1999)

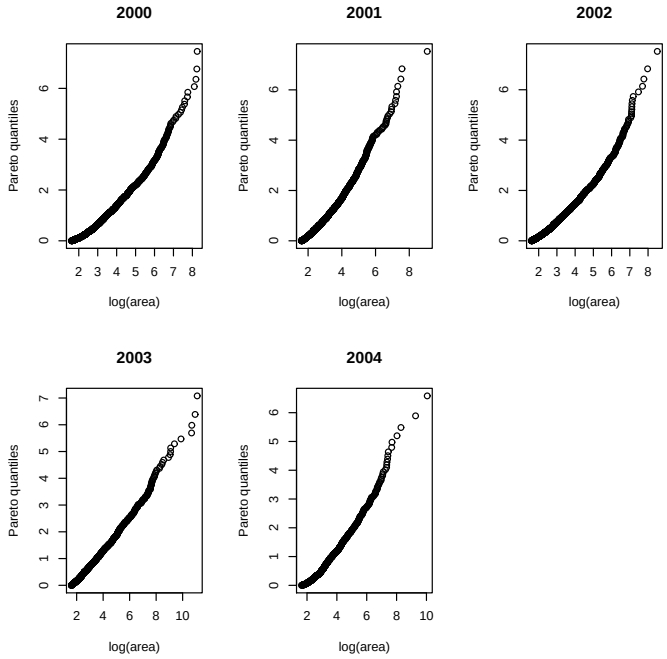

Figure 7: Pareto QQ-plots of the wildfire sizes (2000-2004)

Spatial analysis

The regions of Portugal (Figure 8) that are considered in this study, as well as the notation used, are:

1 - Minho and Douro Litoral; 2 - Trás-os-Montes and Alto Douro; 3 - Porto; 4- Beira Interior; 5 - Beira Litoral, Estremadura and Ribatejo; 6 - Lisboa; 7 - Alentejo; 8 - Algarve

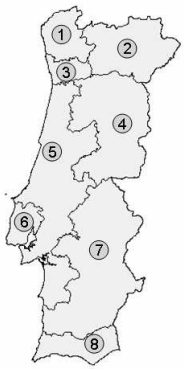

Figure 8: Regions of Portugal

The distribution of the number o fires that occurred in terms of the regional distribution is presented in Figure 9, both for the original sample and for the sample of all the observations $\geq$ than 100 hectares. The data is presented in Table 1. The plots show that region 4 has (both) the largest number of wildfires and also of large wildfires.

Region	1	2	3	4	5	6	7	8
% of wildfires	17	25	7	30	10	7	3	1
Number of wildfires	5060	7738	2198	9313	2998	2110	819	380

Table 1: Percentage of wildfires per region

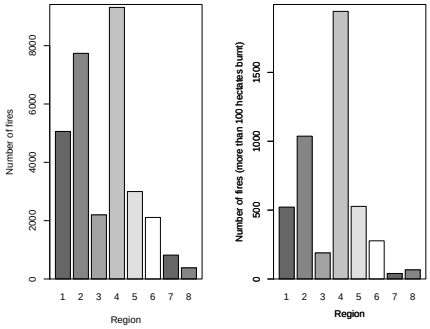

Figure 9: Distribution of wildfires according to the eight regions - global sample (left) and observations ≥ 100 (right)

The histograms of the eight regions are presented in Figure 10.

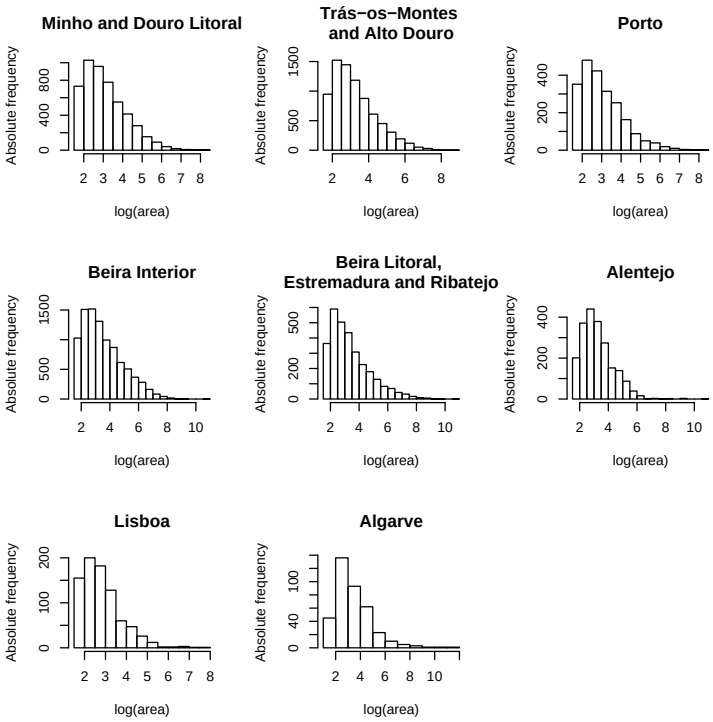

Figure 10: Histograms of the eight regions

The Pareto QQ-plots of the eight regions are presented in Figure 11. These plots clearly show that, although there is some variation from one region to another, all the underlying distributions of the eight regions are heavy-tailed.

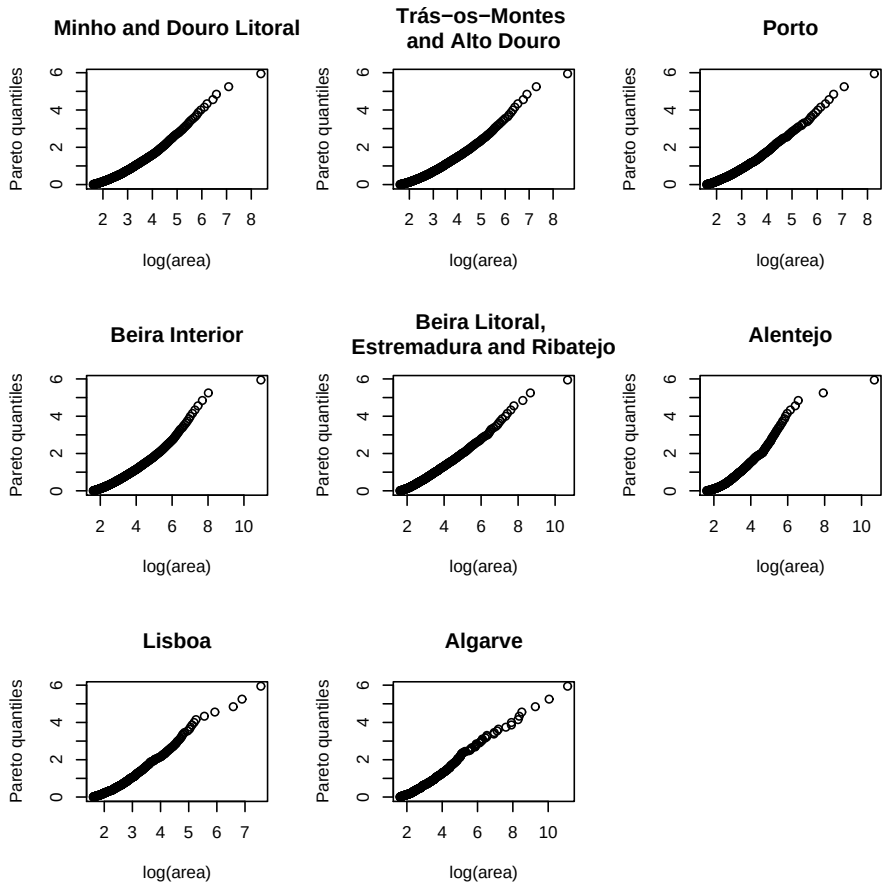

Figure 11: Pareto QQ-plots for the eight regions

The boxplots of the eight regions are presented in Figure 12.

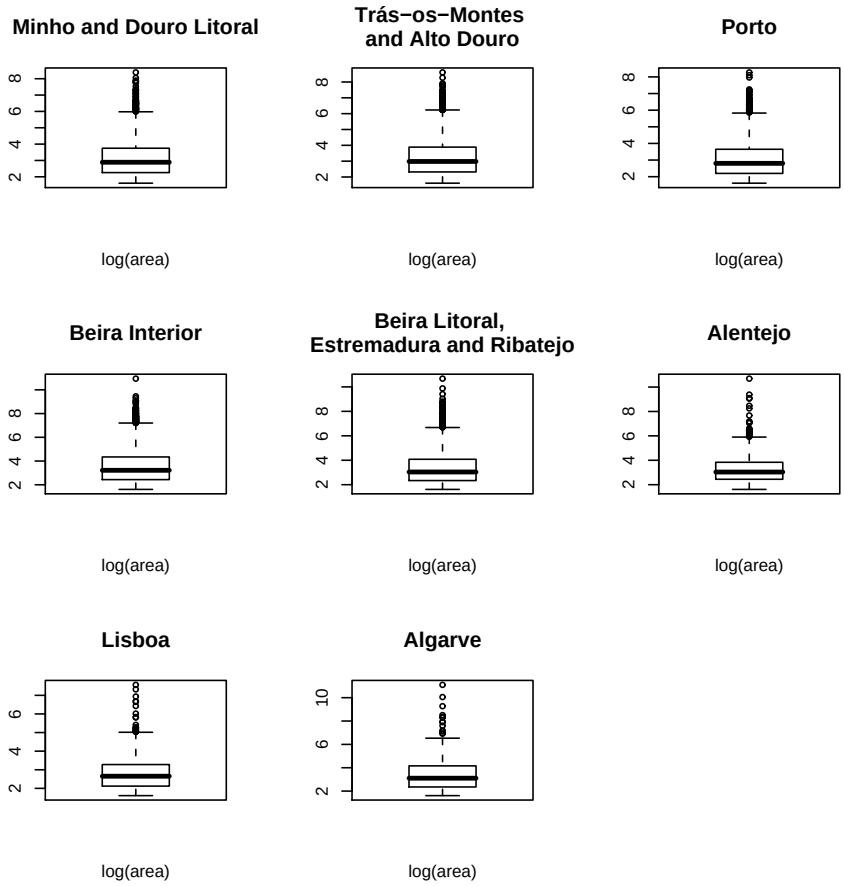

Figure 12: Box-plots of the eight regions

So far, the preliminary data analysis has indicated us that:

1. The data is consistent with a heavy-tailed wildfire size distribution, belonging to the domain of attraction of the Fréchet distribution.

2. There are strong annual, as well as regional, variations in the wildfire sizes. However, large wildfires sizes, over different year and regions, are all consistent with a heavy-tailed distributions, all belonging to the domain of attraction of (possibly different) Fréchet distributions. Hence, different Fréchet extreme value distributions could be fitted to each of the annual and regional data sets to assess the regional and temporal variations in the behavior of extreme wildfires.

However, rather that making inference on the extreme values by fitting Fréchet distributions to the annual and to the regional data, we will make in-

ference on excess wildfire sizes over a fixed threshold, using the generalized Pareto distribution (GPD). This inferential method is often called *Peaks Over Threshold (POT)*. The duality between the two inferential techniques is well established. Inference based on excesses is often more data-efficient, and is preferred to direct inference on extreme value distributions. The estimated model parameters of the extreme value model can be recovered from the estimated model parameters of the corresponding GPD. We refer the reader to any book on extreme value theory, such as Embrechts *et al.* (1991) or Coles (2001).

Hence, we assume that the excesses (or exceedances) above a sufficiently high threshold (u) have, in the limit, a GDP. The distribution function of the GPD(k,σ), $k \in \Re$ and $\sigma > 0$, is given by:

$$F(x \mid k, \sigma) = \begin{cases} 1 - \left(1 + \frac{kx}{\sigma}\right)^{-1/k} & , k \neq 0 \\ 1 - \exp(-\frac{x}{\sigma}) & , k = 0 \end{cases}$$

where $\sigma > 0$ and $k \geq 0$, $x > 0$, while for $k < 0$ the range is $0 < x < -\sigma/k$. The parameters k and σ correspond, respectively, to the shape and scale parameters of the distribution.

We note that the exponential and the Pareto distributions belong to generalized Pareto class of distributions.

The crucial issue of the POT method is the choice of the threshold u. There must be a "trade-off" between bias and variance of the estimators of the parameters. In fact, a too low u results directly in an increase of the number of large observations used for the inference, which increases the bias while decreasing the variance of the estimators. On the other hand, a high u reduces the portion of the sample used for the inference and therefore decreases the bias of the estimators, while increasing their variances.

In order to choose an adequate u, we usually plot the estimates of the parameters, as a function of either the threshold or the number of upper order statistics (r):

- the sample mean excess function (MEF) given as

$$e_n(u) = \frac{\sum_{i=1}^{n}(X_i - u)^+}{\sum_{i=1}^{n} 1_{\{X_i > u\}}}, \text{where } \mathbf{1} \text{ is the indicator function.}$$

- the Pickands' [Pickands (1975)] and the Moments [Dekkers and de Haan (1989)] parameter estimates,
- the estimates provided by a bias-corrected version of Hill's estimator [Beirlant et al.(2004)],
- the Maximum Likelihood (ML) estimates.

The aspect of the MEF gives valuable information regarding the tail of the distribution. Again, the exponential and the Pareto are used as benchmarks. It is known that the theoretical MEF of an exponential distribution is constant, regardless of the value of u considered. The theoretical Pareto MEF is linear with a positive slope, for every threshold u considered (see *e.g.* Embrechts *et al.* (1991) - Figure 4.2.4. page 295).

It may be interesting, at this stage, to recall the fact that, if $X \sim GPD(k, \sigma)$, then the MEF is given as,

$$e(u) = \frac{\sigma + ku}{1 - k}$$

for $k < 1$, $u > 0$ and $\sigma + ku > 0$. This means that for the GPD the plot of the MEF should be linear with slope and intercept equal to $k/(1-k)$ and $\sigma/(1-k)$, respectively.

The MEFs of the data classified by region are presented in Figure 13. The linearity and the fact that all the lines have positive slopes above some point definitely supports the Pareto tail.

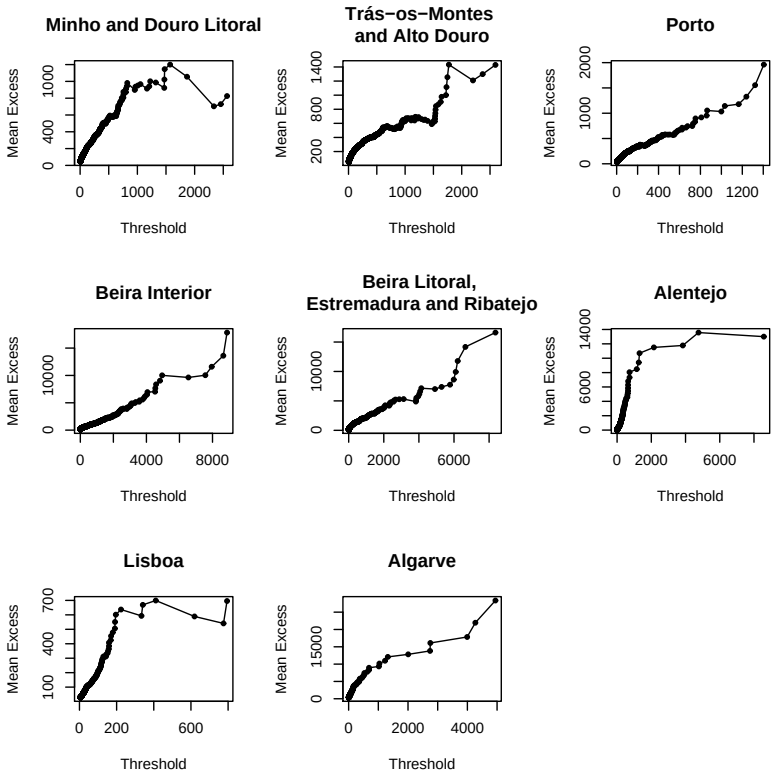

Figure 13: MEF of the eight regions of Portugal

The moments' estimator

These plots, as well as the ones that follow, may be difficult to analyze. As such, they should all be considered, simultaneously, to produce an adequate value of u for each region. By a careful analysis of the plots presented in Figure 14, it can be deduced that the number of upper order statistics r which should be selected, or the corresponding threshold $u = x_{n-r:n}$, for each of the region can be chosen as:

$$\begin{array}{l} \text{Regions 1 - } r \approx 350 \ (u = 130) \\ \text{Regions 2 - } r \approx 500 \ (u = 200) \\ \text{Regions 3 - } r \approx 150 \ (u = 120) \end{array}$$

The plots of regions 4-8 are a bit messy and as such we will not, at this stage, propose any value for r.

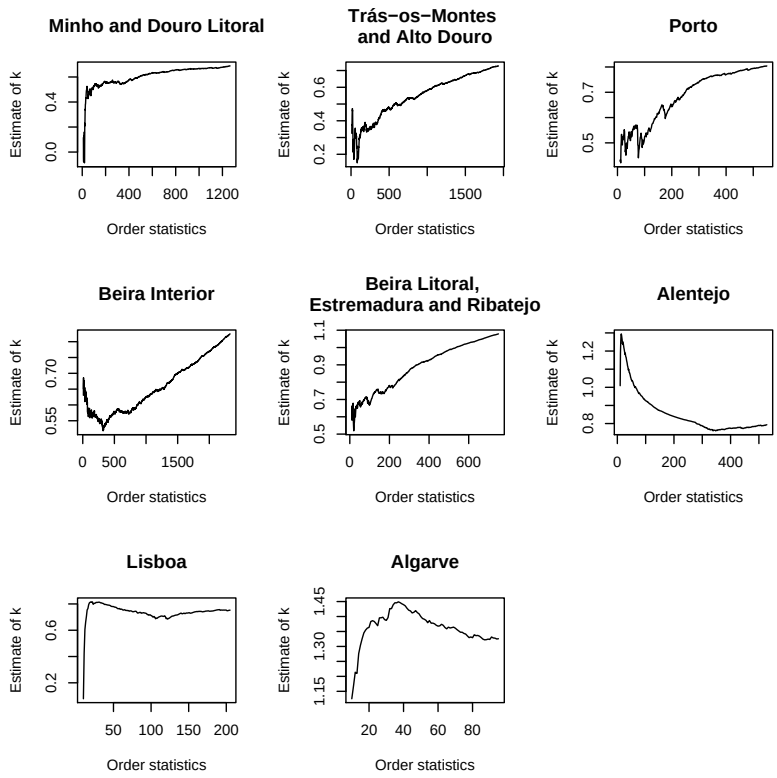

Figure 14: Estimates of k for the eight regions of Portugal obtained by the moments' estimator

The bias-corrected Hill's estimator

Regions 1 - $r \approx 119$ ($u = 300$); Regions 2 - $r \approx 68$ ($u = 800$)
Regions 3 - $r \approx 565$ ($u = 30$); Regions 4 - $r \approx 214$ ($u = 900$)
Regions 5 - $r \approx 137$ ($u = 600$); Regions 6 - $r \approx 275$ ($u = 100$)
Regions 7 - $r \approx 272$ ($u = 20$); Regions 8 - $r \approx 148$ ($u = 30$)

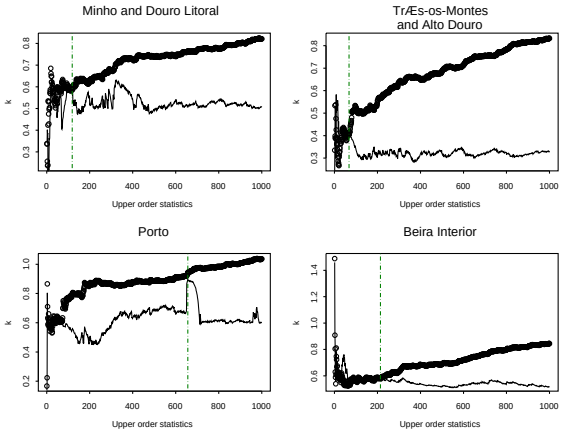

Figure 15: Hill's estimates (circle) and Bias corrected estimates (solid line) for regions 1 (top left), 2 (top right), 3 (bottom left) and 4 (bottom right)

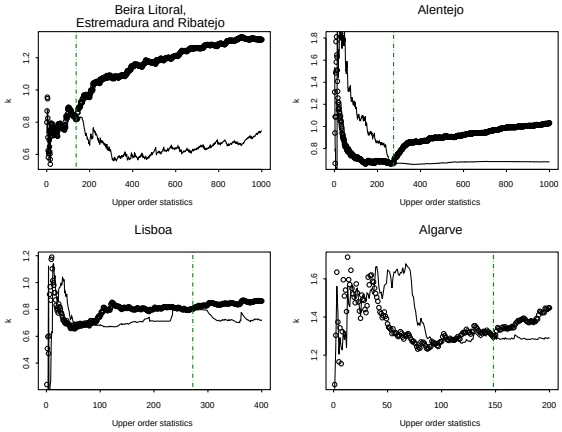

Figure 16: Hill's estimates (circle) and Bias corrected estimates (solid line) for regions 5 (top left), 6 (top right), 7 (bottom left) and 8 (bottom right)

The ML estimator

The ML estimates of the shape parameter of the GPD are presented (for the eight regions) as a function of the number of exceedances.

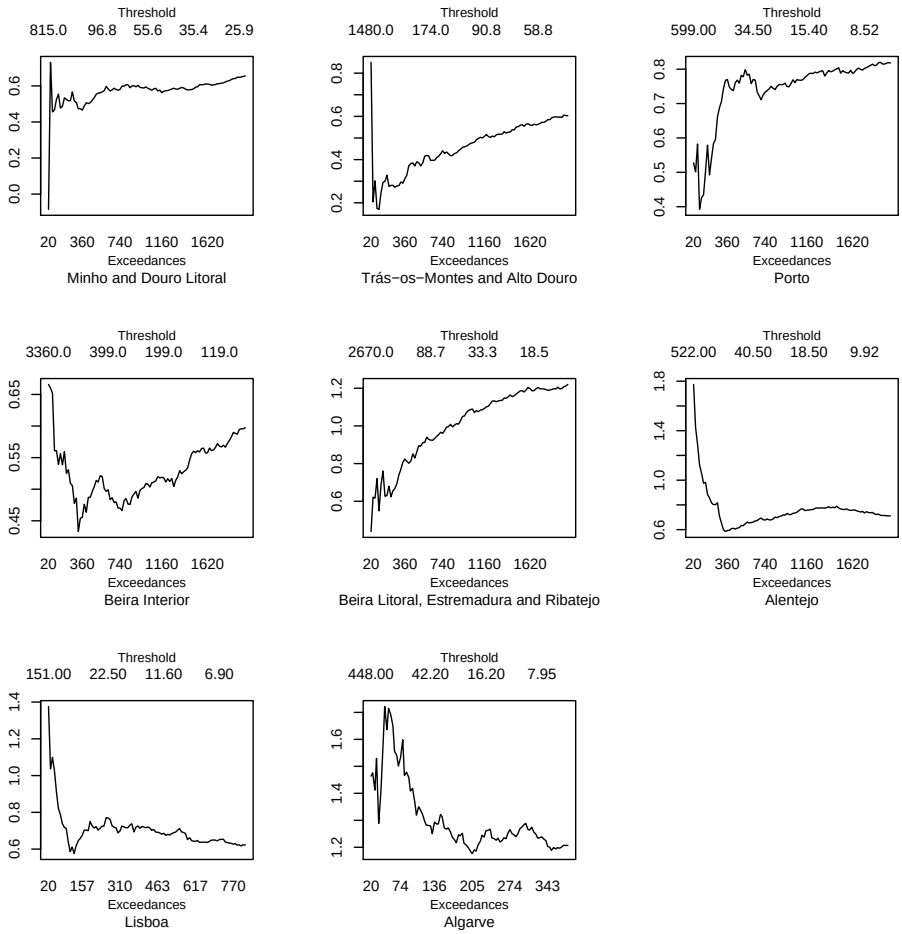

Figure 17: ML estimates for the shape parameter k ($r = 20, ..., 2000$)

Based on the above arguments, we have fitted different GPD models with several different thresholds. The best models for each of the eight regions are chosen based on bias-variance arguments.

GPD models - let M_{ij} be the jth GPD model of the ith region, $j = 1, 2, ..., mod_i$, being mod_i the number of models considered for region i and $i = 1, 2, ..., 8$.

Model	u	**# of excesses** (%)	$\hat{k}$	**Se** $\hat{k}$	**CI** k **(95%) Inf**	**Sup**	$\hat{\sigma}$	**Se** $\hat{\sigma}$	**CI** σ **(95%) Inf**	**Sup**
M_{11}	100	522 (10.3)	0.57	0.07	0.43	0.71	87.10	6.97	73.43	100.76
M_{12}	250	158 (3.1)	0.49	0.12	0.26	0.73	180.71	24.86	131.99	229.44
M_{13}	500	53 (1.0)	0.41	0.20	0.01	0.81	352.85	84.17	187.88	517.82
M_{14}	750	26 (0.5)	0.26	0.34	-0.41	0.92	595.57	232.27	140.32	1050.82
M_{21}	100	1037 (13.4)	0.48	0.05	0.38	0.57	117.54	6.54	104.72	130.37
M_{22}	250	379 (4.9)	0.33	0.07	0.19	0.47	225.64	19.48	187.46	263.82
M_{23}	500	149 (1.9)	0.29	0.11	0.07	0.50	318.03	43.28	233.20	402.86
M_{24}	750	76 (1.0)	0.22	0.13	-0.03	0.47	412.96	71.07	273.66	552.25
M_{25}	1000	41 (0.5)	0.22	0.18	-0.13	0.57	487.96	114.20	264.13	711.80
M_{31}	100	189 (8.6)	0.53	0.11	0.30	0.75	117.31	15.18	87.55	147.06
M_{32}	250	75 (3.4)	0.47	0.16	0.16	0.79	196.07	37.76	122.06	270.08
M_{33}	500	28 (1.3)	0.52	0.28	-0.03	1.07	298.08	97.07	107.83	488.34
M_{34}	750	13 (0.6)	0.51	0.49	(1)	(1)	468.66	255.38	(1)	(1)
M_{35}	1000	8 (0.4)	0.29	0.88	(1)	(1)	769.79	728.14	(1)	(1)
M_{41}	100	1943 (20.9)	0.59	0.04	0.52	0.67	166.68	6.98	153.01	180.36
M_{42}	250	946 (10.2)	0.50	0.05	0.40	0.60	284.77	16.14	253.13	316.40
M_{43}	500	460 (4.9)	0.51	0.07	0.37	0.64	402.92	32.80	338.63	467.21
M_{44}	750	272 (2.9)	0.48	0.09	0.31	0.66	536.23	54.00	430.39	642.07
M_{45}	1000	186 (2.0)	0.59	0.12	0.35	0.82	548.75	72.62	406.42	691.09
M_{46}	1250	125 (1.3)	0.56	0.14	0.29	0.83	711.64	110.85	494.38	928.91
M_{47}	1500	88 (0.9)	0.51	0.15	0.22	0.81	939.26	162.49	620.78	1257.74
M_{48}	1750	74 (0.8)	0.64	0.19	0.27	1.01	823.28	169.79	490.49	1156.07
M_{51}	100	527 (17.6)	0.90	0.09	0.73	1.07	157.25	13.96	129.89	184.62
M_{52}	250	259 (8.6)	0.68	0.11	0.46	0.89	383.59	44.32	296.72	470.46
M_{53}	500	153 (5.1)	0.65	0.13	0.40	0.91	557.70	81.84	397.30	718.11
M_{54}	750	98 (3.3)	0.57	0.15	0.28	0.87	854.78	147.07	566.52	1143.03
M_{55}	1000	79 (2.6)	0.74	0.21	0.32	1.15	755.88	166.42	429.69	1082.06
M_{56}	1250	57 (1.9)	0.61	0.21	0.20	1.03	1134.89	267.64	610.31	1659.46
M_{57}	1500	48 (1.6)	0.71	0.26	0.20	1.22	1097.36	305.92	497.76	1696.96
M_{58}	1750	38 (1.3)	0.64	0.27	0.11	1.16	1428.49	428.80	588.04	2268.94
M_{61}	100	277 (13.1)	0.76	0.10	0.57	0.96	62.62	6.59	49.70	75.53
M_{62}	150	153 (7.3)	0.91	0.14	0.64	1.19	78.89	11.54	56.27	101.51
M_{63}	200	95 (4.5)	1.08	0.20	0.69	1.47	97.38	18.96	60.22	134.54
M_{64}	250	64 (3.0)	1.19	0.25	0.70	1.68	127.19	30.27	67.87	186.52
M_{65}	300	50 (2.4)	1.52	0.36	0.81	2.22	107.36	34.13	40.47	174.26
M_{66}	350	37 (1.8)	1.67	0.46	0.77	2.57	137.39	55.06	29.48	245.31
M_{67}	400	28 (1.3)	1.52	0.46	0.62	2.42	257.93	103.28	55.50	460.36
M_{71}	50	100 (12.2)	0.65	0.15	0.36	0.94	41.34	6.92	27.78	54.91
M_{72}	100	40 (4.9)	1.05	0.32	0.42	1.67	46.08	14.66	17.35	74.81
M_{73}	150	21 (2.6)	1.52	0.69	(1)	(1)	49.80	27.67	(1)	(1)
M_{81}	100	66 (17.4)	1.64	0.33	1.00	2.29	85.24	24.39	37.44	133.04
M_{82}	250	31 (8.2)	1.51	0.46	0.61	2.41	325.59	133.67	63.59	587.58
M_{83}	500	18 (4.7)	1.35	0.61	(1)	(1)	895.43	507.44	(1)	(1)
M_{84}	1000	13 (3.4)	1.37	0.78	(1)	(1)	1291.41	950.16	(1)	(1)

Table 2: The various models fitted to the regional data

(1) - unreliable confidence interval

The best models selected for the eight regions are the following:

Region	Model	Threshold
1	M_{12}	250
2	M_{22}	250
3	M_{32}	250
4	M_{43}	500
5	M_{53}	500
6	M_{64}	250
7	M_{72}	100
8	M_{81}	100

Table 3: Final GPD models fitted to the data

The quantiles $Q_{0.995}$ and $Q_{0.999}$, estimated using the above fitted models and the corresponding confidence intervals (95%), are given in the following table.

Region	$\mathbf{Q_{0.995}}$	**L Inf**	**L Sup**	$\mathbf{Q_{0.999}}$	**L Inf**	**L Sup**
1	788.45	681.76	945.98	1887.07	1415.81	2970.56
2	1016.68	917.15	1148.35	2029.34	1675.29	2647.22
3	863.93	697.43	1164.00	2038.41	1385.04	4263.03
4	2240.95	1999.12	2564.26	5431.73	4332.37	7360.62
5	3544.96	2784.20	4939.07	10819.69	6765.85	22129.75
6	1059.79	733.89	1930.85	6399.20	2903.29	29603.92
7	534.60	327.52	1556.12	2637.72	861.13	2890.41
8	17668.15	4974.84	99105.93	247863.56	28189.84	99105.93

Table 4: Estimated quantiles using the final models fitted to the data

The estimated quantiles can be compared with the empirical quantiles (to be presented in the next table) to have an idea of how well the model fits to data.

Some empirical quantiles

Region	$Q_{0.995}$	$Q_{0.999}$
1	746.81	2308.29
2	1009.53	1734.52
3	809.38	2587.27
4	2213.13	4746.19
5	3855.85	8372.32
6	724.93	8873.61
7	761.23	1590.68
8	11950.35	49829.96

Table 5: Empirical quantiles

Fitting annual extremes

Similar inferential techniques are used to fit models for large values, separately for each of the 21 years. Therefore, a detailed information will not be provided here. One pertinent question that needs to be answered is: "Is if the probability structure of the extreme wildfires changing over time?". The fitted shape and scale parameters for each of the 21 years are given in Figure 18. The above question can then be formulated as a statistical test based on the time series of the estimated parameters. These tests, based on the estimated autocorrelation structures, are standard in time series analysis. Unfortunately, due to the short series (21 consecutive observations), they have very low power. Therefore, we avoid making formal statements about the temporal variation of the probability structure. However, we note that the estimated autocorrelation and partial autocorrelation functions given in Figure 18 do not seem to indicate any temporal structure for the estimated parameters.

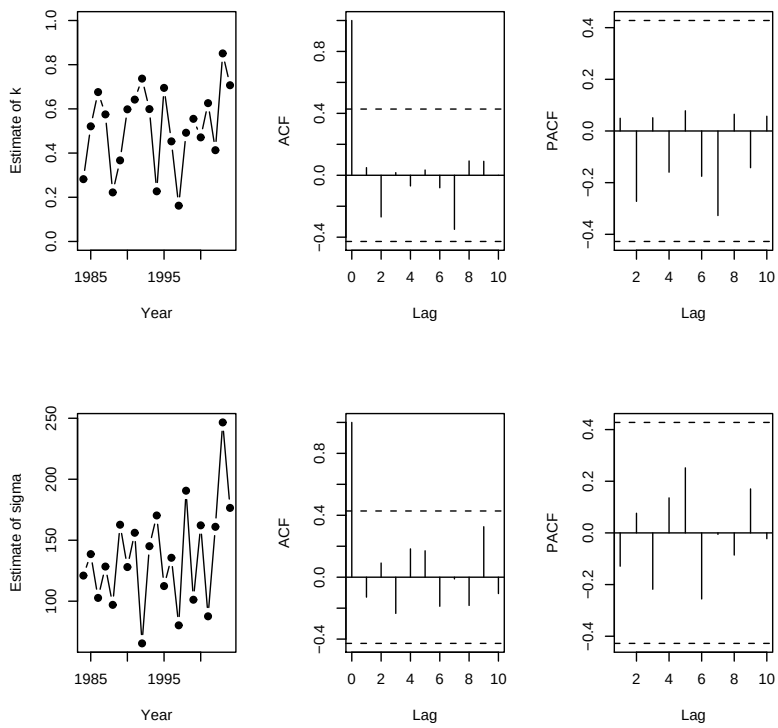

Figure 18: Estimated parameters and corresponding ACF and $PACF$ for the 21 years (top row - k; bottom row - σ), considering $u = 100$

An important issue is also to access the influence of the largest four observations recorded in 2003. Are they influential in terms of tail heaviness? Should they be rejected on the grounds that they are most likely outliers? We

proceeded as follows:

- retrieve the largest observation and fit a first model;
- retrieve the largest and 2nd largest observations and fit a second model;
- retrieve the largest, 2nd and 3rd largest observations and fit a third model;
- retrieve the largest, 2nd, 3rd and 4th largest observations and fit a fourth model.

The estimates of k and σ, as well as the observations that were left out from this analysis, are presented in Table 6.

Sample size	Observations retrieved	Estimate of k	Estimate of σ
1185	66070.63	0.83	239.25
1184	66070.63 and 56550.80	0.80	235.22
1183	66070.63, 56550.80 and 43970.25	0.76	233.53
1182	66070.63, 56550.80, 43970.25 and 43282.33	0.70	237.43

Table 6: Additional models for 2003 - $u = 100$

The values contained in Table 6 clearly show the influence of these large observations on the estimated parameters (see Table 2).

Software

Graphs and general data handling - R
Models - **evir 1.5** functions (Splus - A. McNeil; nowadays R - Alec Stephenson).
Hill's estimator and Bias-reduced estimator (Beirlant *et al.* (2004)) retrieved in http://ucs.kuleuven.be/Wiley/index.html.

References

Beirlant J, Goegebeurg Y, Segers J, Teugels J (2004) 'Statistics of Extremes - Theory and Applications.' (John Wiley and Sons:Chichester)
Coles S (2001) 'An Introduction to Statistical Modeling of Extreme Values.' (Springer-Verlag:London)
Dekkers A L M, de Haan L (1989) On the estimation of the extreme-value index and large quantile estimation. *Annals of Statistics* **17**, 1795-1832.
Embrechts P, Kluppelberg C, Mikosch T (1997) 'Modeling Extremal Events.' (Springer-Verlag:Berlin)
Hill B M (1975) A simple general approach to inference about the tail of a distribution. *Annals of Statistics* **3**, 1163-1174.
Pickands J(1975) Statistical Inference using extreme order statistics, *Annals of Statistics* **3**:119-131.