CHAPTER 6

EXTREME – VALUE ANALYSIS

6.1 Introduction

As we have already mentioned in section 1.3, one of the areas where extreme-value theory has recently gained ground is teletraffic engineering. Indeed, as the World-Wide Web system becomes more and more popular, the need of evaluating its performance becomes more and more compelling. In order to achieve that and go on to possible modifications, one needs to know the behaviour of the users' 'demands' from the system. This can be expressed either in files lengths, CPU time to complete a job, call holding times and so on. All these quantities, essentially, constitute random variables and consequently their distribution summarizes any information of their behaviour. It's not difficult to realize that, in order the system to function adequately, its capacity should be adjusted so as to handle even the largest 'demands'. Hence, the study of the extremal 'demands' (e.g. longest file lengths, longest call holding times etc) turns out to be an important tool of teletraffic engineers.

In this thesis we apply the notion of extreme-value theory to teletraffic data-set obtained from the Internet Traffic Archive (ITA) (http://ita.ee.lbl.gov/index.html). The Internet Traffic Archive is a moderated repository to support widespread access to traces of Internet network traffic, sponsored by ACM SIGCOMM. The traces can be used to study network dynamics, usage characteristics, and growth patterns, as well as providing the grist for trace-driven simulations. The archive is also open to programs for reducing raw trace data to more manageable forms, for generating synthetic traces, and for analyzing traces. The Internet Traffic Archive was put together by Peter Danzig (University of Southern California), Jeff Mogul (Digital's Western Research Lab), Vern Paxson (Lawrence Berkeley National Lab), and Mike Schwartz (University of Colorado at Boulder). It was made possible by Carl Malamud and the Internet Multicasting Service

giving it its original home. The archive is sited at the Lawrence Berkeley National Laboratory.

More particularly, we are going to analyze data from the 'EPA-HTTP' trace. This trace contains a day's worth of all HTTP requests to the EPA WWW server located at Research Triangle Park, NC. The logs were collected from 23:53:25 EDT on Tuesday, August 29 1995 through 23:53:07 on Wednesday, August 30 1995, a total of 24 hours. There were 47,748 total requests, 46,014 GET requests, 1,622 POST requests, 107 HEAD requests, and 6 invalid requests. Timestamps have one-second precision. The WWW server software used was not recorded. The logs were collected by Laura Bottomley (laurab@ee.duke.edu) of Duke University. The dataset contains the following pieces of information:

- host making the request
- date and time of the request
- type of request
- HTTP reply code
- bytes in the reply.

In the present extreme-value analysis we are going to concentrate on the analysis of the files lengths requested (i.e. on the bytes in the reply).

6.2 Exploratory Data Analysis

6.2.1 Description of the Data

The original data-set contained 47,748 cases of requests. Still, in 5,331 of these cases the file length (in bytes) requested was not recorded, thus leaving as with 42,417 valid observations to be analyzed. However, in 5,718 of these valid cases no file was actually requested (i.e. the file length in bytes is zero). These observations were also removed from the data-set to be analyzed. So, finally, the data-set that we analyzed contained 36,699 file lengths in bytes. In the table that follows we present the main descriptive statistics of the variable under investigation, so as to get a first idea of the form of the data. A more intuitive description of the data is provided by the histogram that follows in

figure 6.1. However, the fact that we have, as it seems, some very large observations, spoils the original histogram of our data-set. For this reason, we split and display the data in two parts. Figure 6.2.a is the histogram of the smaller (shorter) values of 'File lengths', while in figure 6.2.b the histogram of large 'File lengths' is depicted. Now, we can see more clearly the 'behaviour' of 'File lengths'. It is obvious that we are dealing with a possible heavy-tailed underlying d.f. Still, a more thorough discussion on this issue is postponed until a later section of this chapter.

Table 6.1. Descriptives Statistics of 'File lengths' (in bytes)

Measure		**Statistic**	**Std. Error**
Mean		8497.8880	369.1512
95% Confidence Interval for Mean	*Lower Bound*	7774.3412	
	Upper Bound	9221.4347	
5% Trimmed Mean		3046.7888	
Median		1897.0000	
Variance		5001069499.289	
Std. Deviation		70718.2402	
Minimum		33.00	
Maximum		4816896	
Range		4816863	
Interquartile Range		3182.0000	
Skewness		30.318	0.013
Kurtosis		1267.888	0.026

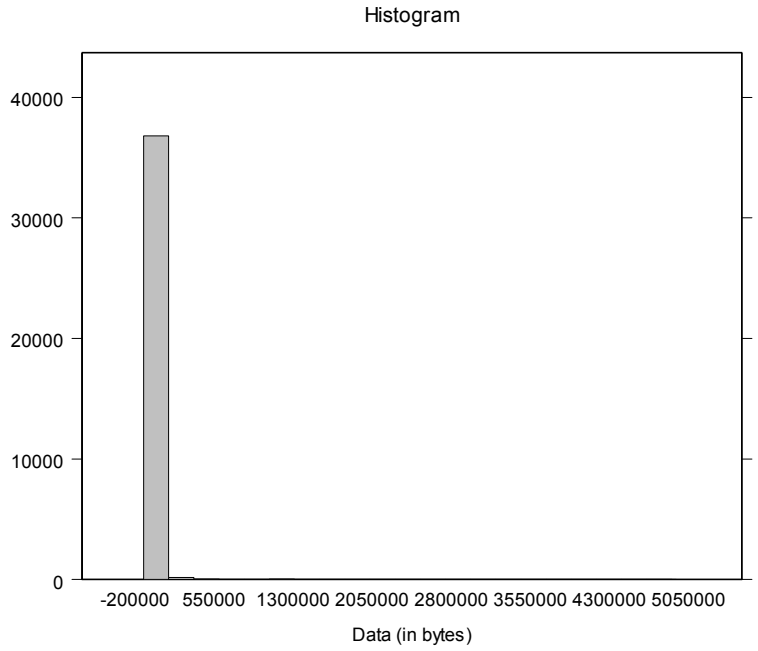
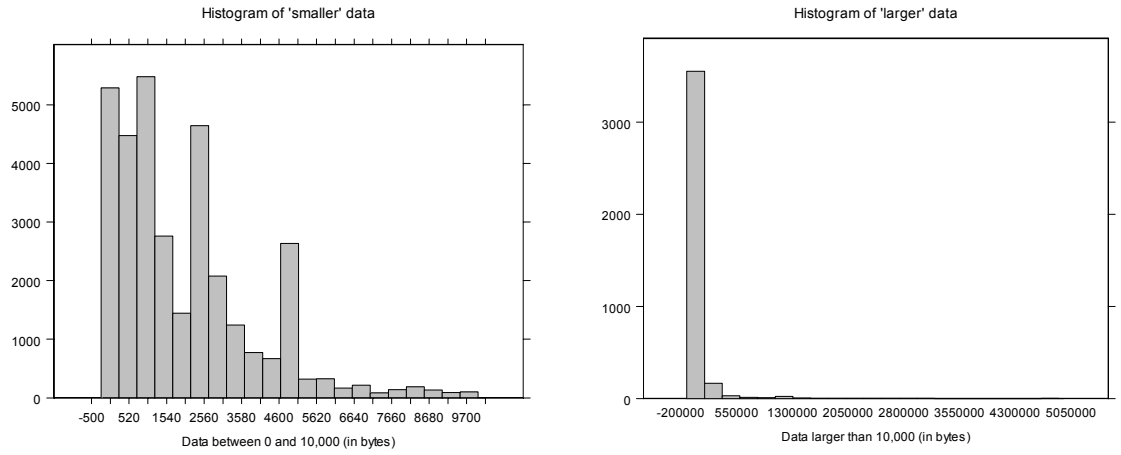


Figure 6.1. Histogram of 'File lengths'



(a) (b)

Figure 6.2. Histograms of separate parts of data ('File lengths')

A raw histogram may, however, be misleading as an indicator of how frequently high levels occur, since it fails to capture phenomena such as seasonality of data or the tendency of extreme values to occur in clusters. These are better revealed by a sequence (time-series) plot. The corresponding plot for our data is given in figure 6.3. A first examination of it reveals no problems of seasonality or clustering. Still, this is a very

rough check, there are more elaborate methods to check for the existence of the previous mentioned problems.

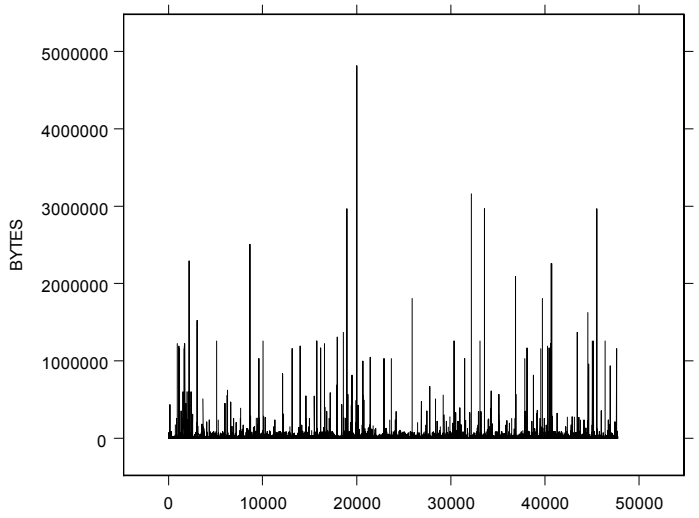


Figure 6.3. Sequence plot of 'File lengths' (in bytes)

Other common graphical illustrations of a data-set are the plots of the empirical distribution function and its inverse, the quantile function. For the 'File lengths' these plots are given in figures 6.4 and 6.5. Both of these plots, as well as the previous graphs provide a strong indication that our data are highly skewed to the right, i.e. they are long-tailed.

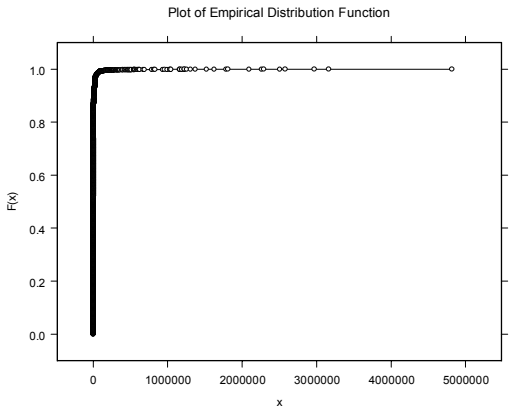


Figure 6.4. Plot of empirical distribution function of 'File lengths' in bytes.

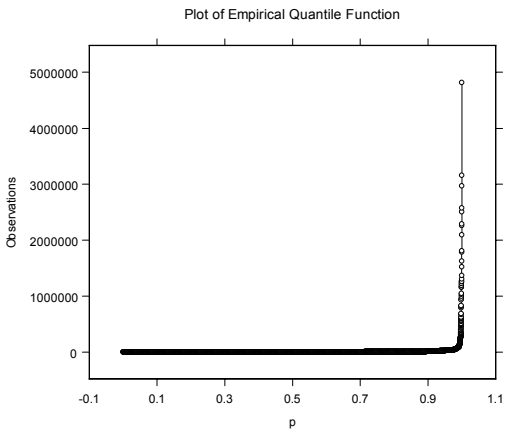


Figure 6.5. Plot of empirical quantile function of 'File lengths' in bytes.

6.2.2 Investigation of Independence

One of the assumptions widely used and required for almost all the results in extreme-value theory is that of independence of the data. Of course, for the case of Hill estimator, there are results that prove that the good properties of the estimator hold under quite general conditions of non-independence. Still, these results are not verified (at least yet) for other estimators. In general, in cases that some kind of dependence of data is detected, adjustments in the statistical methods used are needed. So, before proceeding with any analysis of the data, it is useful to check whether independence holds. Since our data are essentially time-series, we apply the standard tests of randomness (i.e. independence). Moreover, the classical autocorrelation function (acf) was calculated, as well as a test of independence based on records.

❖ Standard Tests of Randomness

The most popular standard (non-parametric) time-series tests of randomness (Resnick, 1997b) are the following:

- The turning point test

Let T be the number of turning points among $\{X_1, X_2, ..., X_n\}$. Under the null hypothesis of i.i.d. data we have that $T \sim Normal(2(n-2)/3, (16n-29)/90)$.

- The difference-sign test

Let S be the number of indices i such that $(X_i - X_{i-1}) > 0$. Under the null hypothesis of i.i.d. data we have that $S \sim Normal((n-1)/2, (n+1)/12)$.

We applied these tests to our data (using S-Plus) and the results were

- The turning point test :

According to this test, our data does not constitute an i.i.d. sample (the null hypothesis is rejected with almost zero observed level of significance)

- The difference-sign test

According to it, the hypothesis of independence cannot be rejected (the observed level of significance is 0.08)

❖ A Test of Randomness Based on Records

For any sequence of observations $\{X_1, X_2, ..., X_n\}$, there is a corresponding sequence of record values $\{R_1, R_2, ..., R_{N_n}\}$. The number of records N_n constitutes the so-called record counting process, defined as :

$N_1=1$, $N_n = 1 + \sum_{j=2}^{n} I_{\{X_j > M_{j-1}\}}$, $n \geq 2$ and M_j is the maximum of $\{X_1, ..., X_j\}$.

As Embrechts et al. (1997) prove, if the underlying data $\{X_1, X_2, ..., X_n\}$ are i.i.d., then the first two moments of the r.v.'s N_i are given by the formulae:

$$E(N_n) = \sum_{j=1}^{n} \frac{1}{j}, \text{ and } Var(N_n) = \sum_{j=1}^{n} \left(\frac{1}{j} - \frac{1}{j^2} \right).$$

The above result combined with the fact that the standardized number of records of an i.i.d. sample converges to a distribution closely related to the Normal (Kinnison, 1985) leads to another (rough) non-parametric test of randomness. Such a test is also applied in Embrechts et al. (1997).

For the data-set under investigation, with the sample size equal to 36699, the expected number of records is 11.1 with variance 9.4, while the actual, observed number of records is 12. These values seem to support the hypothesis of independence (the observed level of significance under the approximate assumption of normality is 0.77).

❖ A Method Based on Sample Auto-Correlation Function

An exploratory, informal method for testing for independence can be based on the sample autocorrelation function, defined as

$$\hat{\rho}(h) = \frac{\sum_{i=1}^{n-h} (X_i - \overline{X})(X_{i+h} - \overline{X})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}, \; h \in \aleph_+ .$$

Large values of the autocorrelation function or, to be more precise, values exceeding the constructed 95% confidence bounds, indicate lack of independence. The pattern of the values of the autocorrelation function can be better viewed by the so-called autocorrelation plot of $(h, \hat{\rho}(h))$. For our data-set, the autocorrelation plot, for lag h up to

100, is given in figure 6.6. In the plot, 95% confidence bounds are also given (automatically calculated in SPSS). Though some values of the autocorrelation function exceed the bound, still since these are too few (6 exceedances in 100 points, lags) they can be even regarded as included in the common statistical error. Moreover, the fact that the autocorrelation function doesn't display any particular pattern with respect to the lag h, is reassuring that no statistical autocorrelation exists in the data.

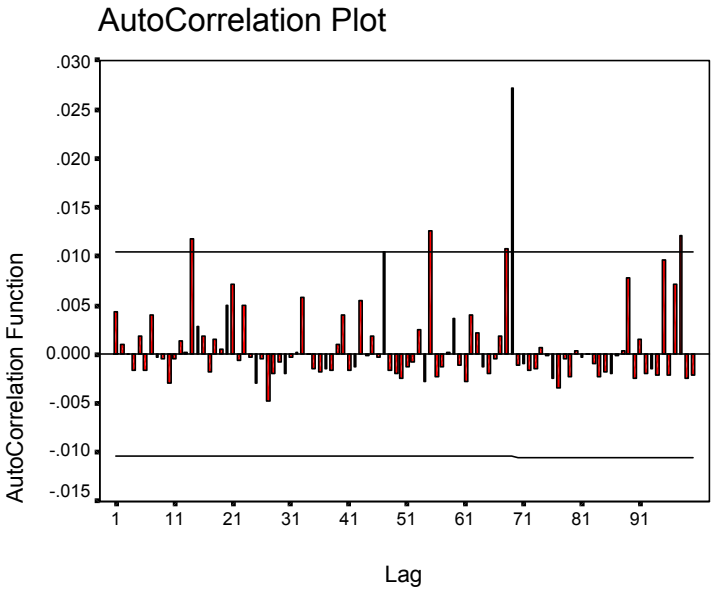


Figure 6.6. Autocorrelation plot of 'File lengths' for lags h=1,..,100.

However, it is important to realize that the 95% confidence bounds drawn by a typical statistical package (like S-Plus or SPSS) are drawn using Bartlett's formula, under the assumption that the data are normally distributed or, at best, that they have finite fourth moment. But, this assumption is totally inappropriate for heavy tailed-data, as is probably our case. So, not much faith should be put on these bounds. Moreover, in many cases of heavy-tailed data the centering by the sample mean is omitted, since if the mathematical expectation does not exist, it is totally meaningless to center by the sample mean. In such cases, the following heavy-tailed modification of autocorrelation function is more appropriate :

$$\hat{\rho}_H(h) = \frac{\sum_{i=1}^{n-h} X_i X_{i+h}}{\sum_{i=1}^{n} X_i^2}, \; h \in \aleph_+ .$$

As Resnick (1998) mentions, for the case of moving-average sequences, though there are cases where the mathematical correlation does not even exist ($\gamma > \frac{1}{2}$), $\hat{\rho}_H(h)$ still converges to a limiting constant. However, the limit law for $\hat{\rho}_H(h)$ is very complex. Many and strong assumptions have to be imposed in order to derive the limiting d.f. of $\hat{\rho}_H(h)$. Nonetheless, the function $\hat{\rho}_H$ and its corresponding plot can be used as an exploratory tool to make preliminary investigations of dependence. If on graphing the sample heavy-tailed autocorrelation function, one finds only small values, then it may be possible to model the data as i.i.d. Similarly, if the sample heavy-tailed autocorrelation function is small beyond lag g, then there is some evidence that MA(q) (moving average model of order q) may be an appropriate model. Of course, without firm knowledge of the quantiles of the limit distribution of $\hat{\rho}_H(h)$, it is impossible to say with precision what 'small' means. Still, in our case, we will be restricted to that exploratory view, since, as we will see in the sequel, our data don't even seem to follow any of the known and usually used long-tailed d.f.'s. The heavy-tailed autocorrelation plot for our data is given in figure 6.7. The majority of values does not exceed the limit 0.015, while the largest autocorrelation is observed for lag 57 and is approximately 0.04. Generally speaking, one could judge these values to be "small", indicating lack of autocorrelation in data.

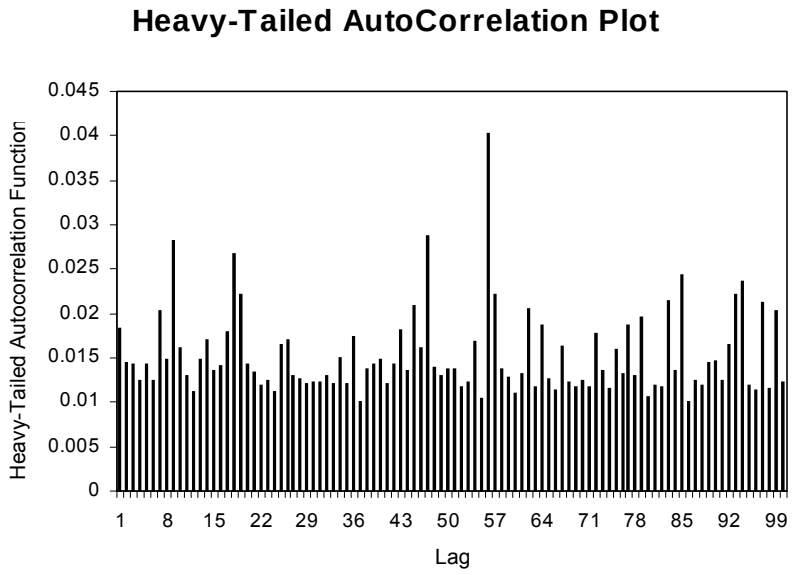


Figure 6.7. Heavy-tailed autocorrelation plot of 'File lengths' for lags h=1,..,100.

Since all the previous non-parametric and graphical checks of dependence do not give us a clear indication that dependence in our data exists, we are going to move on to other analyses assuming that our data are indeed independently (and identically) distributed. Moreover, even the nature of our data do not suggest that any form of dependence or correlation should exist. For example, if at some point during the day a large file is requested, there is no reason to believe that the next requested file should also be large or should be small.

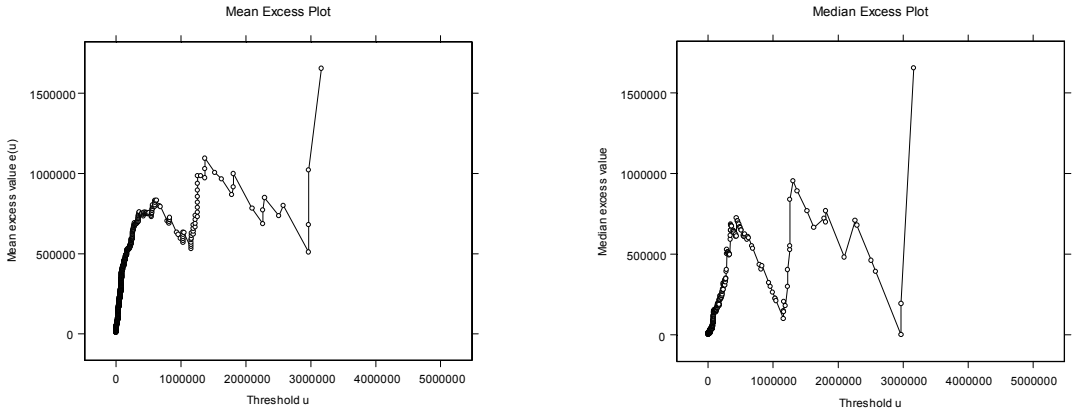
6.2.3 Investigation of Heavy Tails

Before proceeding to the formal study of extremes of the data in hand (i.e. to the formal estimation of extreme value index γ), there are several exploratory methods that can be used to give us a first insight into the behaviour of the extremes of a data-set. Such methods are the mean (or median) excess plots and the QQ plots based on exponential or other long-tailed d.f.'s. The usefulness of these tools is mainly that they provide us with an indication of whether our data are long-tailed ($\gamma \geq 0$) or short tailed ($\gamma < 0$). Knowledge, even rough, of the sign of γ can direct to the choice of more preferable extreme-value index estimators. Moreover, in the former case our interest should be

focused on the estimation of large quantiles, while in the latter case the estimation of upper end-point is more meaningful.

❖ Excess Plots

The definition and properties of mean and median excess functions have been given in section 5.3.1. As we have mentioned there, if the empirical mean excess function (MEF) of a data-set ultimately increases for the larger thresholds u, then this implies that the underlying d.f. is sub-exponential, while if it decreases, the corresponding d.f. is super-exponential. Ultimately constant behaviour of empirical MEF implies that the d.f has tails equivalent to exponential tails. The mean excess plot, accompanied with the corresponding median excess for the 'file lengths' we are examining are given below. If we ignore the last values of the empirical MEF (which is based on very few observations), an increase seems to exist, implying that we are dealing with a distribution with heavier than exponential tails.

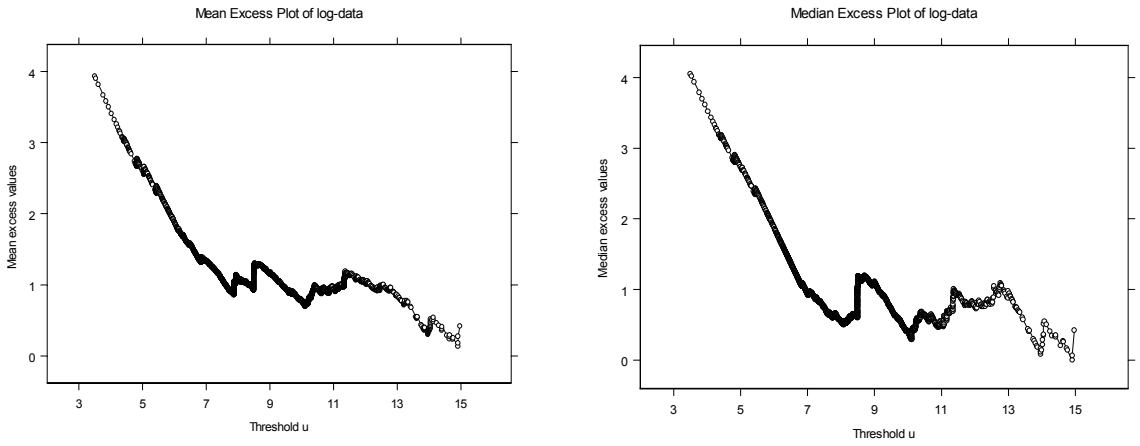


(a) (b)

Figure 6.8. Empirical mean excess plot (a) and median excess plot (b) of 'File lengths'.

However, we are interested in checking whether the underlying d.f. of our data is long-tailed in the sense that it belongs to $MDA(H_\gamma)$, $\gamma>0$. So, we should look at the mean excess plot not of the original data but of their logarithmic transformations. Indeed, we have already showed that if the empirical MEF of the logarithmic-transformed data is ultimately increasing, then the d.f. belongs to $MDA(H_\gamma)$ with $\gamma>0$, and the values of the

MEF converge to the true value of γ. For the 'file lengths' we are analyzing the plots are given below. These plots indicate a decrease implying that our data-set is not actually long-tailed. Still this result may be misleading. Indeed, one of the main assumptions of mean excess plot is that the underlying distribution has a finite first moment. However, long-tailed distributions with γ>1 do not satisfy such a condition, and the corresponding mean excess plot of such distributions is totally misleading. For this reason, we proceed to QQ plots, which do not have such restrictive assumptions.



(a) (b)

Figure 6.9. Empirical mean excess plot (a) and median excess plot (b) of log-transformed data.

❖ Quantile (QQ) Plots

The use of quantile plots (or QQ plots) as exploratory tools in extreme-value analysis is described in detail in Beirlant et al. (1996). The usefulness of these plots lies in the fact that for important classes of distributions the quantiles Q(p) are linearly related with the corresponding quantiles of a standard member from this class of distributions. As linearity in a graph can be easily checked by eye, this tool can be used in order to check goodness-of-fit of a data-set to a particular d.f.

In the sequel we present the QQ plot of our data-set, with respect to the Exponential (figure 6.10) and the Pareto (figure 6.11) distribution. These are distributions medium and long tailed, respectively, commonly used in practice. In each case the straight line, that indicates the 'perfect fit', has been estimated (using Least Square method) based on the whole data-set, i.e. these QQ plots evaluate the overall fit of our data-set. However,

since our main interest is focused on the 'upper part' of the data (largest values) it would be more interesting to look only on this part of the plots (right part).

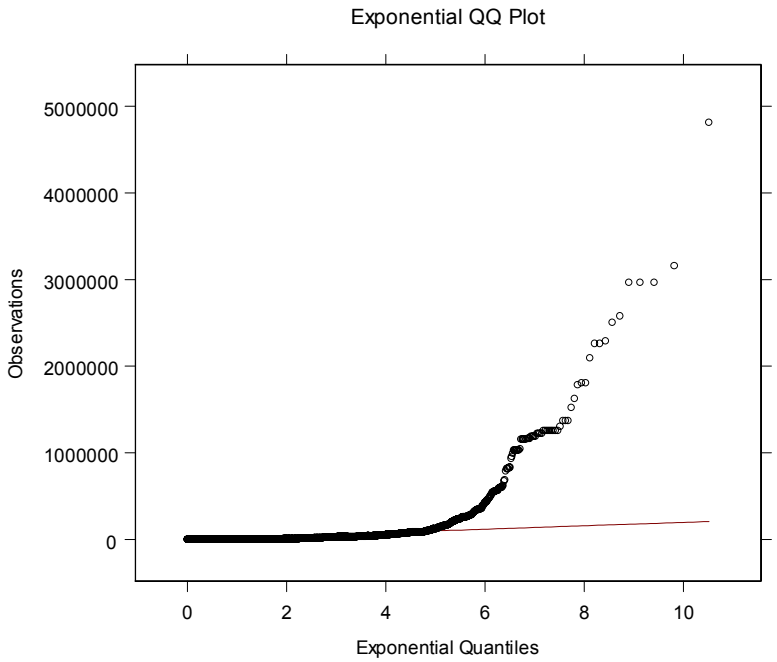
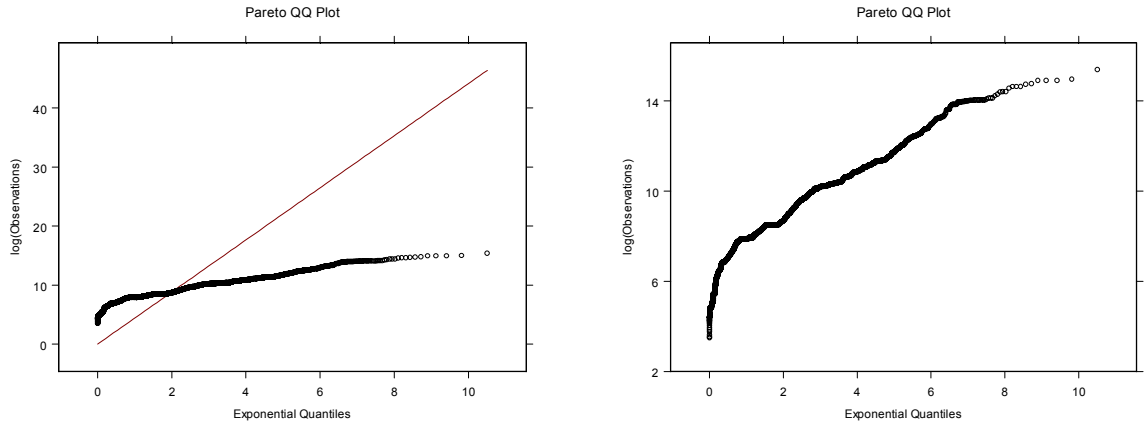


Figure 6.10. Exponential QQ (quantile) plot of the 'File Lengths' data-set.



(a) (b)

Figure 6.11. Pareto QQ (quantile) plot of the 'File Lengths' data-set, with the fitted LS line (a) and without fitted line (b).

The exponential QQ plot (figure 6.10) indicates that the main part of the data fits well the exponential d.f., but the right tail deviates strongly from the perfect fit. In fact, the observed values are much larger than the expected ones, implying subexponentiality of the data (same conclusion as the one drawn from the mean excess plot).

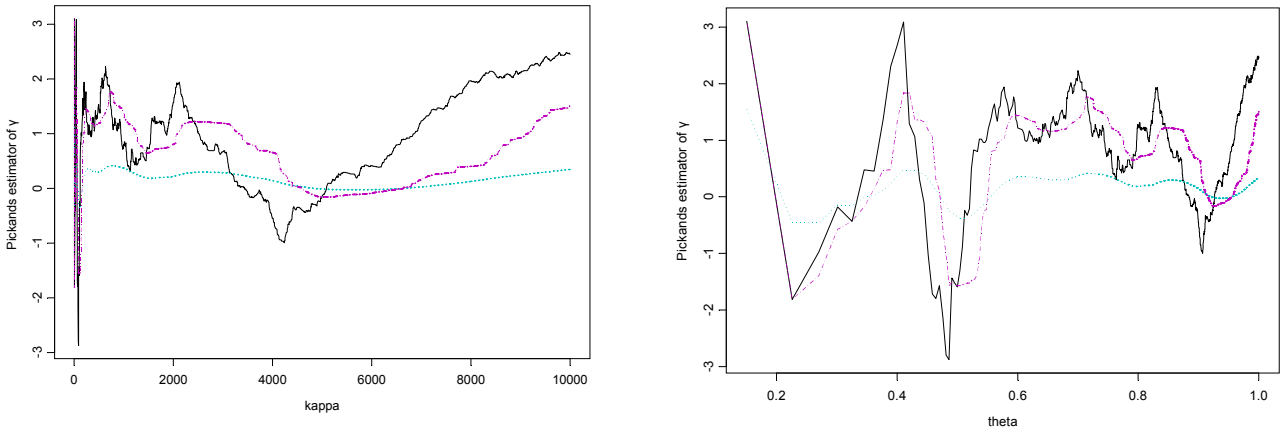
A first look at the Pareto QQ plot (figure 6.11.a) makes clear that by no means do our data fit the Pareto d.f. Still, if we do not try to fit Pareto globally , but look only on the behaviour of the right tail of the data (right part of figure 6.11.b), we can see a linear pattern for these large values, i.e. ultimately the data do seem to follow a Pareto d.f. This remark suggests that, probably, we are dealing with a long-tailed underlying d.f. F, i.e. $F \in MDA\left(H_\gamma\right),\ \gamma > 0$. Still, the formal investigation of the extremal behaviour of the data-set under study, comes in the section that follows.

6.3 Extreme-Value Analysis

6.3.1 Estimation of Extreme-Value Index γ

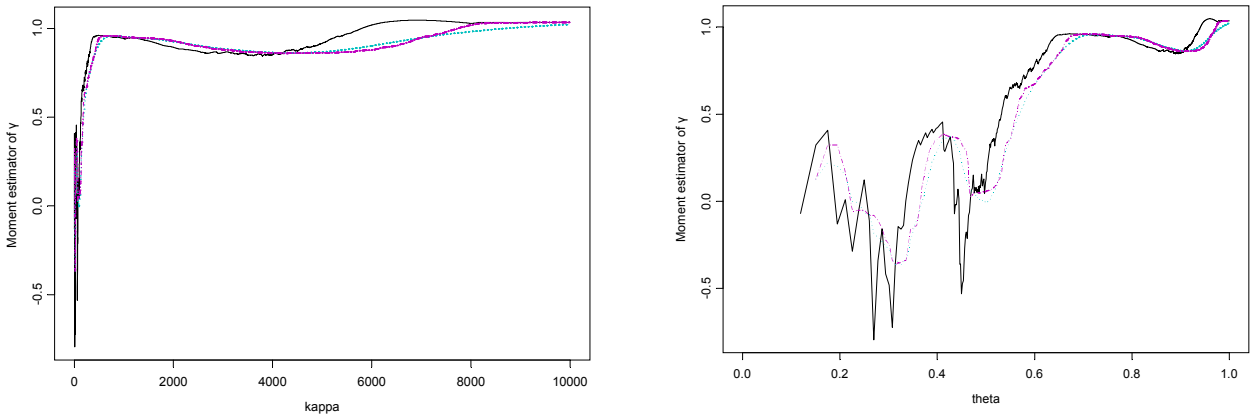
Here, we deal with the main scope of the current analysis, which is the investigation of the extremal behaviour of 'File Lengths' transported via the site of EPA. This is partly achieved through the estimation of extreme-value index γ. From the previous exploratory analysis we believe that γ is positive. For this reason, apart from extreme-value index estimators applicable to $\gamma \in \Re$, we are also going to use extreme-value index estimators restricted to the case $\gamma > 0$. In chapters 4 and 5 we have presented several semi-parametric extreme-value index estimators, while in section 5.4 we have evaluated their performance via simulation. According to these simulations, not a uniformly best estimator exists. So, we are going to estimate γ with more that one estimator. In particular, we are going to apply Pickands, Moment, Peng's and W estimators (applicable for any $\gamma \in \Re$), as well as Hill and Moment-Ratio (only for $\gamma > 0$). Each of these estimation techniques provides us with a sequence of estimated values of γ (one for each k, number of upper order statistics used in the estimation). So, a vital step before deciding on the estimated value of extreme-value index γ is the choice of k. On of the methods to

accomplish that, is a graphical one, already discussed. In the sequel we provide the plots $\{k, \gamma(k)\}$ of the estimators used, as well as the corresponding 'alternative plots' (see section 4.3.2), which are more useful and reliable in the case that our data do not follow closely a Pareto d.f. (as is probably the case here). Moreover, in each plot, apart from the standard estimators, the mean and median averaged estimators are depicted. Note that in the graphs to follow we display the estimated values of γ that correspond to k up to 10,000 (27% of the whole dataset). The purpose of this is to focus on the part of data that essentially concerns us and since, in any case we are looking for the proper number k of upper order statistics used and observations smaller than the upper 20% or 25% cannot actually be regarded to be "large". So, we can get a better view of the part of the graph that we are actually interested in.



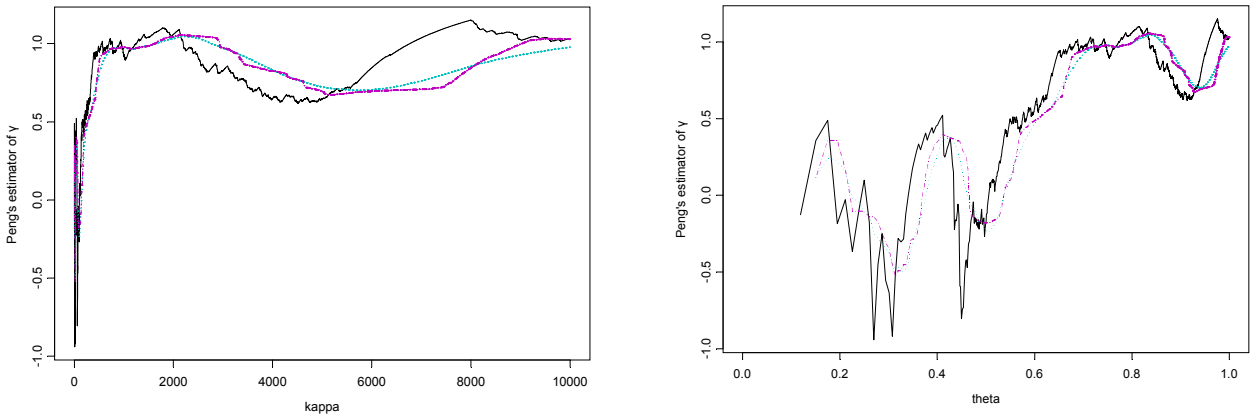
(a) (b)

Figure 6.12. Plot (a) and alternative plot (b) of Pickands estimator of γ (—), and the corresponding mean-averaged (⋯), and median-averaged (– · –) estimators.



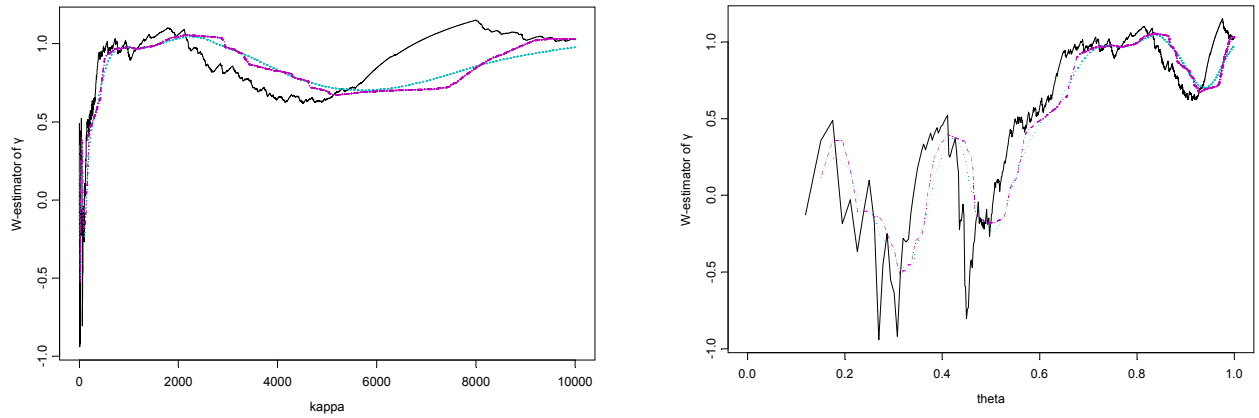
(a) (b)

Figure 6.13. Plot (a) and alternative plot (b) of Moment estimator of γ (—), and the corresponding mean-averaged (⋯), and median-averaged (– · –) estimators.



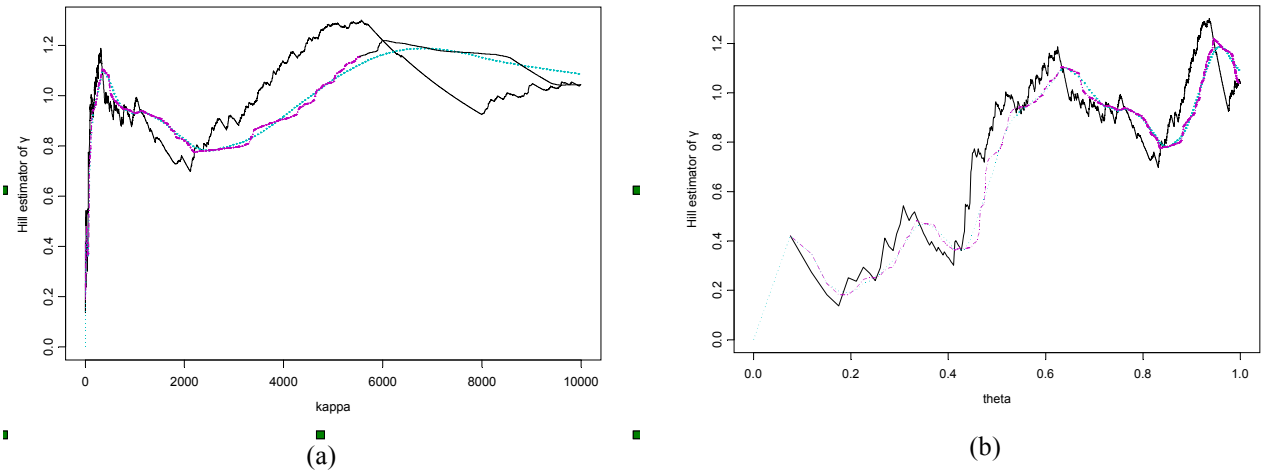
(a) (b)

Figure 6.14. Plot (a) and alternative plot (b) of Peng's estimator of γ (—), and the corresponding mean-averaged (⋯), and median-averaged (– · –) estimators.



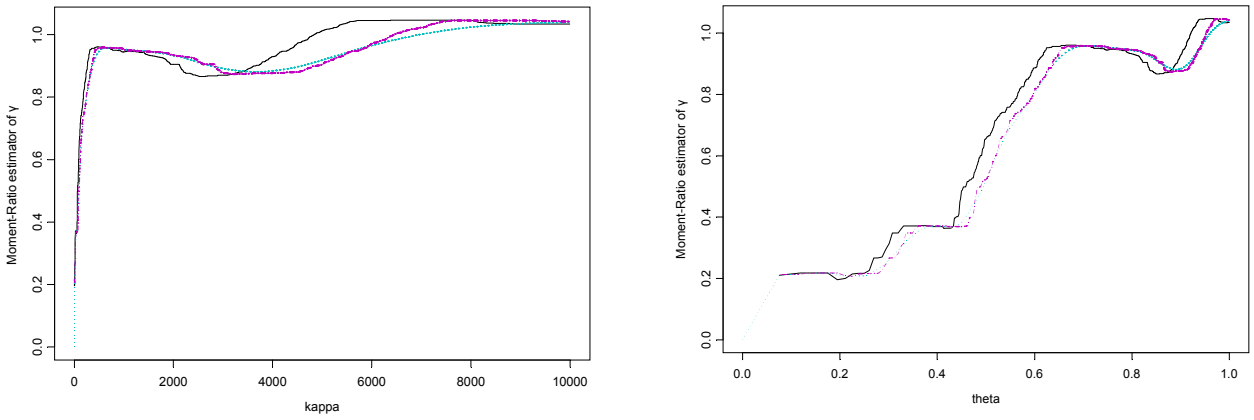
(a) (b)

Figure 6.15. Plot (a) and alternative plot (b) of W estimator of γ (—), and the corresponding mean-averaged (···), and median-averaged (– · –) estimators.



(a) (b)

Figure 6.16. Plot (a) and alternative plot (b) of Hill estimator of γ (—), and the corresponding mean-averaged (···), and median-averaged (– · –) estimators.



(a) (b)

Figure 6.17. Plot (a) and alternative plot (b) of Moment-Ratio estimator of γ (—), and the corresponding mean-averaged (···), and median-averaged (– · –) estimators.

It is a fortunate event that in our case all the estimators tend to almost the same value of γ and, more precisely, the value 1. Especially, Moment-Ratio and Moment estimators which have, according to the previous simulation results, the best performance for positive γ, display almost a straight line to 1. So, we can deduce that the value of γ that best describes the sizes of requested files from the size of EPA is approximately 1. This implies that the underlying distribution of the data under study belongs to the maximum domain of attraction of the Frechet(1) distribution, i.e. it a Pareto-type d.f. asymptotically decaying like Pareto(1) ($\overline{F}(x) \xrightarrow[x\to\infty]{} x^{-1}$).

6.1.2 Estimation of Large Quantiles

Though the value of extreme-value index estimator is indicative of the tail-heaviness of the underlying distribution of our data, a quantity that is more useful for practical purposes is large quantiles. That is, in practice what is desirable to know is what is the 'File Length' that is exceeded only 1 in x times/transactions (x should be large, such as 100, 1000 or even larger). Each extreme-value index estimator leads to a different estimation formula for large quantiles, too, (see chapter 4) that is, also, dependent on the number k of upper-order statistics used in the estimation. Here, we use the generic formula proposed by Dekkers et al. (1989) :

$$\hat{x}_p = \frac{a_n^{\hat{\gamma}_M} - 1}{\hat{\gamma}_M} \cdot \frac{X_{(k+1):n} M_1}{\rho_1(\hat{\gamma}_M)} + X_{(k+1):n},$$

where $a_n = \frac{k}{n(1-p)}$, $\rho_1(\gamma) = \begin{cases} 1, & \gamma \geq 0 \\ (1-\gamma)^{-1}, & \gamma < 0 \end{cases}$ (see also section 4.5),

substituting each different estimator.

In the figures that follow we present the estimators of 99-quantile based on the extreme-value index estimators previously used vs k.

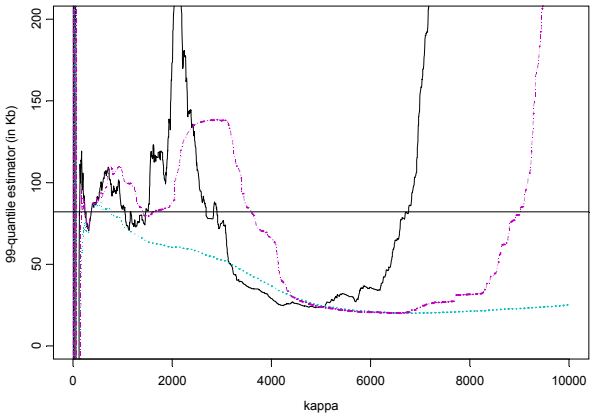


Figure 6.18. Plot of 99-quantile based on Pickands estimator (—), and the corresponding mean-averaged (···), and median-averaged (– · –) estimators.

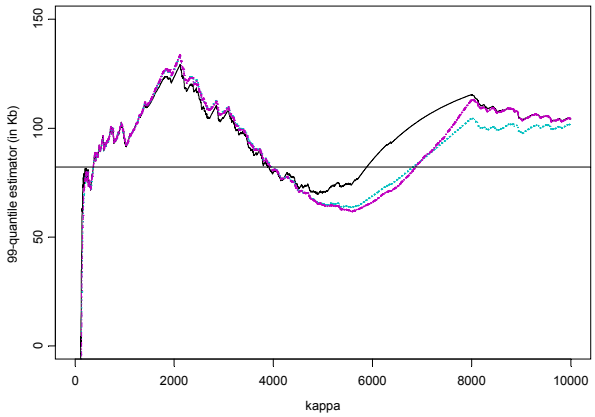


Figure 6.19. Plot of 99-quantile based on Moment estimator (—), and the corresponding mean-averaged (···), and median-averaged (– · –) estimators.

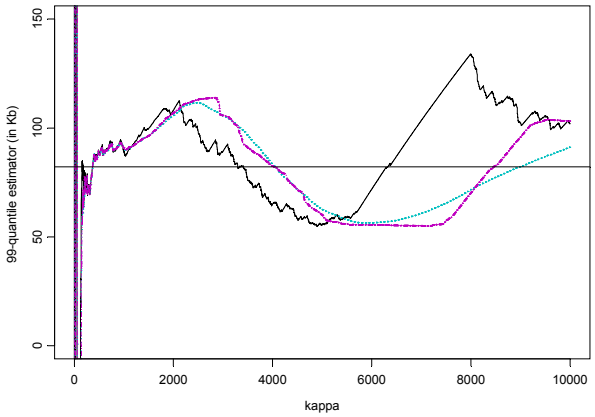


Figure 6.20. Plot of 99-quantile based on Peng's estimator (—), and the corresponding mean-averaged (···), and median-averaged (– · –) estimators.

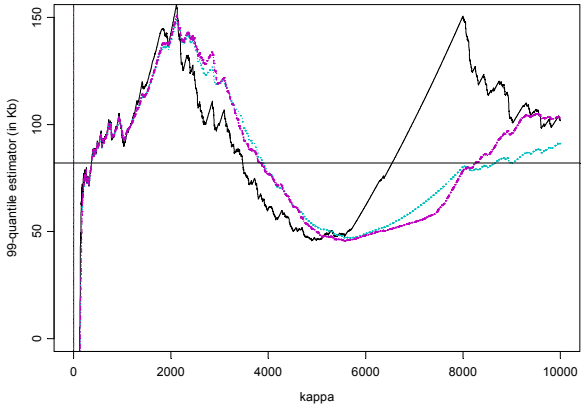


Figure 6.21. Plot of 99-quantile based on W estimator (—), and the corresponding mean-averaged (···), and median-averaged (– · –) estimators.

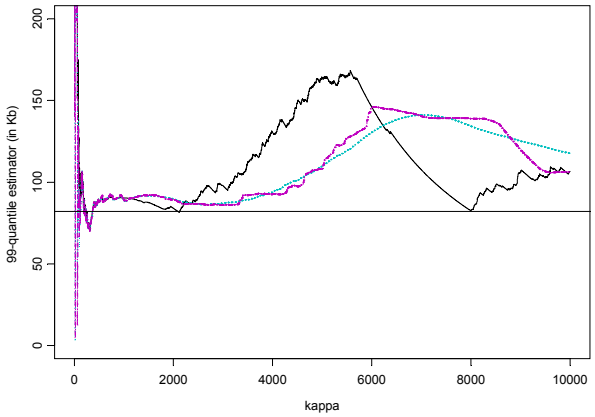


Figure 6.22. Plot of 99-quantile based on Hill estimator (—), and the corresponding mean-averaged (···), and median-averaged (– · –) estimators.

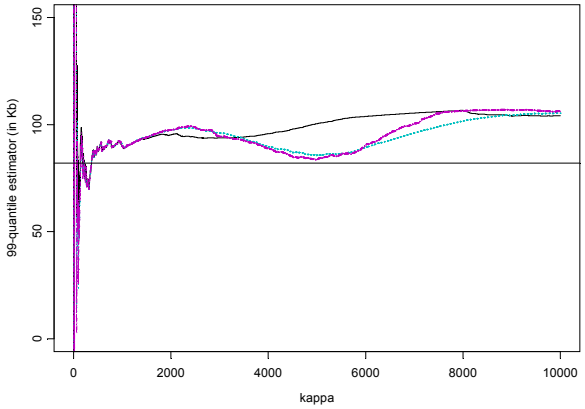


Figure 6.23. Plot of 99-quantile based on Moment-Ratio estimator (—), and the corresponding mean-averaged (···), and median-averaged (– · –) estimators.

Generally speaking, we could say that the behaviour of quantile estimators does not display the nice stability (with respect to k) as was the case for the extreme-value index estimators. However, the 99-quantile estimator based on the Moment-Ratio estimator displays the most stable behaviour indicating a value of 99-quantile approximately 100Kb (though constantly larger than the corresponding empirical estimate which is 82Kb).

In the table below we provide the estimators of 95-, 99-, and 99.9- quantiles, based on the Moment-Ratio estimator of γ, for several different values of k.

Table 6.2. Estimation of large quantiles (in Kb) using Moment-Ratio estimator of γ.

Estimation Methods	**Moment-Ratio Estimation of γ**	**Quantiles**		
		95%	**99%**	**99.9%**
Empirical Estimate	-	*25.846*	*82.054*	*1136.764*
k used				
1000	0.945	19.133	89.760	795.025
2000	0.905	25.579	95.276	734.338
3000	0.871	22.631	93.820	700.947
4000	0.930	19.974	95.793	828.905
5000	1.012	18.467	100.483	1047.887
6000	1.045	18.653	103.873	1161.415
7000	1.047	19.526	105.445	1174.495
8000	1.047	20.248	106.693	1182.169
9000	1.034	19.680	104.126	1127.136
10000	1.035	19.680	104.182	1128.756

To sum up it can be concluded that, we may assume that the size of files requested from the particular site of EPA follows a long-tailed distribution (which decays similar to a Pareto(1) distribution). This property may be further exploited in order to derive other useful outcomes. As far as large quantiles are concerned, we could say that, based on the extreme-value approach, the 95-quantile is roughly 20Kbs, the 99-quantile reaches 100Kbs, while a file larger than 1Mb is requested only one in a thousands times.