Extreme Value Analysis of Teletraffic Data

Tsourti, Zoi and Panaretos, John

2004

Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®
Computational Statistics & Data Analysis 45 (2004) 85–103
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Extreme-value analysis of teletraffic data

Zoi Tsourti, John Panaretos*

Department of Statistics, Athens University of Economics and Business, 76 Patission Str, 10434 Athens, Greece

Received 10 August 2001; received in revised form 1 July 2002

Abstract

An empirically verified characteristic of the expanding area of Internet is the long tailness of phenomena such as cpu time to complete a job, call holding times, files lengths requested, inter-arrival times and so on. Extreme values of the above quantities are liable to cause problems to the efficient operation of the network and call for effective design and management. Extreme-value analysis is an area of statistical analysis particularly concerned with the systematic study of extremes, providing useful insight to fields where extreme values are probable to occur and have detrimental effects, as is the case of teletraffics. In this paper, we illustrate the main elements of this analysis and proceed to a detailed application of extreme-value analysis concepts to a specific teletraffic data-set. This analysis verifies, too, the existence of long tails in the data.

Keywords: Teletraffic engineering; Long tails; Extreme-value index; Smoothing procedures

1. Introduction

In the last decade, a rapid evolution in telecommunications has taken place. The Internet, as the most representative of the telecommunication means, has also experienced an exponential growth, which, however, was accompanied by a high cost represented by network and servers delays. Accordingly, an impelling need has emerged for performance evaluation and subsequent network design.

* Corresponding author.
E-mail address: jpan@aueb.gr (J. Panaretos).

0167-9473/$ - see front matter
doi:10.1016/S0167-9473(03)00116-6

A useful step is the adequate statistical modelling of quantities such as cpu time to complete a job, call holding times, file length requested, inter-arrival times and so on. It is fortunate that the expansion of Internet also led to a proliferation of enormous-sized data-sets of high-quality network traffic measurements. However, the investigation of such data-sets provided researchers with strong indications of deviations from the usual assumptions of classical statistical theory. In the literature, one may find many articles reporting the long tailness of teletraffic data (see, for example, Naldi, 1999; Crovella et al., 1998; Fang et al., 1997; Resnick, 1997; Willinger et al., 1997; Kratz and Resnick, 1996; Resnick and Stărică, 1995; Duffy et al., 1994). In such cases, the classical queuing and network stochastic models, with their simplifying assumptions (assuming, at least, finite moments) are inappropriate and may lead to unreliable results. In the present paper, we present a statistical analysis that can successfully deal with long-tailed teletraffic data, providing researchers with useful insight about the extreme behaviour of the phenomena under study and enhancing, thus, the efficient design and management of the network.

Extreme-value analysis is the field of statistics particularly concerned with the systematic study of extreme values. The cornerstone of extreme-value theory is Fisher–Tippet's theorem for limit laws for maxima (Fisher and Tippet, 1928). According to this theorem, if the maximum value of a distribution function (d.f.) tends (in distribution) to a non-degenerate d.f. then this limiting d.f. can only be the generalized extreme-value (GEV) distribution, with d.f. $H_\gamma(x) = \exp\{-(1+\gamma x)^{-1/\gamma}\}$, where $1+\gamma x > 0$, and $\gamma \in \Re$.

A comprehensive sketch of the proof can be found in Embrechts et al. (1997). The random variable (r.v.) X (or equivalently, the d.f. F of X) is said to belong to the maximum domain of attraction of the extreme-value distribution H_γ if there exist constants $c_n > 0$, $d_n \in \Re$ such that $c_n^{-1}(M_n - d_n) \xrightarrow{d} H_\gamma$. We write $X \in \mathrm{MDA}(H_\gamma)$.

The parameter γ, called extreme-value index, determines, essentially, the behaviour of extremes. Generally speaking, $\gamma < 0$ corresponds to upper-bounded distributions (i.e. with no problem of extreme values), $\gamma > 0$ describes long-tailed distributions (with many extremes), while $\gamma = 0$ refers to moderately increasing to infinity distributions (here most usual distributions are included such as the Normal, the Gamma, the exponential, and so on).

Extreme-value analysis is a topic of major importance in many fields of application where extreme values may appear and have detrimental effects. Such fields range from hydrology (Smith, 1989; Davison and Smith, 1990; Coles and Tawn, 1996; Barão and Tawn, 1999) to insurance (Beirlant et al., 1994; Mikosch, 1997; McNeil, 1997; Rootzén and Tajvidi, 1997) and finance (Danielsson and de Vries, 1997; McNeil, 1998, 1999; Embrechts et al., 1998, 1999; Embrechts, 1999). Its usefulness in teletraffics, as previously mentioned, is recently proving its value.

In this paper, we deal with the extreme-value analysis (ranging from exploratory analysis up to estimation of γ) of a data-set from the field of teletraffic engineering. This issue is discussed in Section 3. Before that, in Section 2, the main ideas of extreme-value theory and corresponding analysis are briefly reviewed. In Section 4, some concluding remarks are provided.

2. Estimation issues of extreme-value analysis

2.1. Extreme-value index estimators

The most popular estimation approach in the context of extreme-value analysis is the so-called 'maximum domain of attraction approach' (Embrechts et al., 1997). In this framework, we are interested in the distribution of the maximum value. According to the Fisher–Tippet theorem, the limiting d.f. of the (normalized) maximum value (if it exists) is the GEV d.f. H_γ. The procedure followed in practice is that we assume that the asymptotic approximation is achieved for the largest k observations (where k is large but not as large as the sample size n), which we subsequently use for the estimation of the extreme-value index γ. However, the choice of k is a rather controversial issue and is further elaborated in Section 2.2. In the sequel, we present the most prominent answers to the issue of parameter estimation.

Pickands estimator $\hat{\gamma}_{\mathrm{P}}$ (Pickands, 1975) is the first suggested estimator for the parameter γ and is given by the formula

$$\hat{\gamma}_{\mathrm{P}} = \frac{1}{\ln 2} \ln\left(\frac{X_{(k/4):n} - X_{(k/2):n}}{X_{(k/2):n} - X_{k:n}}\right),$$

where $X_{1:n} \geqslant X_{2:n} \geqslant \cdots \geqslant X_{n:n}$ are the descending-order statistics of the corresponding sample of observations. A particular characteristic of Pickands estimator is the fact that the largest observation is not explicitly used in the estimation. The properties of Pickands estimator were mainly explored by Dekkers and de Haan (1989), who proved, under certain conditions, weak and strong consistency, as well as asymptotic normality.

However, the most popular tail index estimator is the Hill's estimator (Hill, 1975), which, though, is restricted to the case $\gamma > 0$. Hill's estimator is provided by the formula $\hat{\gamma}_{\mathrm{H}} = (1/k)\sum_{i=1}^{k} \ln X_{i:n} - \ln X_{k+1:n}$. Weak and strong consistency, as well as asymptotic normality of Hill's estimator hold under the assumption of independent and identically distributed data (Embrechts et al., 1997). Though the Hill's estimator has the apparent disadvantage that is restricted to the case $\gamma > 0$, it has been widely used in practice and extensively studied by statisticians. Its popularity is partly due to its simplicity and to the fact that in most of the cases where extreme-value analysis is called for, we have long-tailed d.f.'s (i.e. $\gamma > 0$).

The popularity of Hill's estimator made the problem of trying to extend this estimator to the general case $\gamma \in \Re$ a tempting one. Such an estimator has been proposed by Dekkers et al. (1989). This is the moment estimator, given by $\hat{\gamma}_{\mathrm{M}} = M_1 + 1 - 0.5(1 - (M_1)^2 M_2^{-1})^{-1}$, where $M_j \equiv (1/k)\sum_{i=1}^{k} (\ln X_{i:n} - \ln X_{(k+1):n})^j$, $j = 1, 2$. Weak and strong consistency, as well as asymptotic normality of the moment estimator have been proven by Dekkers et al. (1989).

Concentrating on cases where $\gamma > 0$, the main disadvantage of Hill's estimator is that it can be severely biased, depending on the second-order behaviour of the underlying d.f. F. Based on the behaviour of an asymptotic second-order expansion of the d.f. F, Danielsson et al. (1996) proposed the moments-ratio estimator: $\hat{\gamma}_{\mathrm{MR}} = 0.5 \cdot M_1^{-1} M_2$. They proved that $\hat{\gamma}_{\mathrm{MR}}$ has lower asymptotic square bias than the Hill's estimator (when

evaluated at the same threshold, i.e. for the same k), though the convergence rates are the same.

Apart from the above estimators, many others can be found in the literature, either for $\gamma \in \Re$ or for more specific ranges of γ. The aforementioned estimators share some common desirable properties, such as weak consistency and asymptotic normality. On the other hand, simulation studies or applications on real data can end up in large differences among these estimators. In any case, there is no 'uniformly best' estimator. Of course, Hill, Pickands and moment estimators are the most popular ones. This could be partly due to the fact that they are the oldest ones. Actually, most of the rest have been introduced as alternatives to the above estimators and some of them have been proven to be superior in some cases only.

2.2. Smoothing modifications of extreme-value index estimators

One of the most serious objections one could raise against the aforementioned estimators is their sensitivity towards the choice of k (number of upper-order statistics used in the estimation).

An exploratory way to subjectively choose the number k is based on the plot of the estimator $\hat{\gamma}(k)$ versus k (or $\hat{\gamma}(n^{\theta})$ versus θ in the so-called alternative plot). A stable region of the plot indicates a valid value for the estimator. The need for a stable region results from adapting theoretical limit theorems which are proved subject to the conditions that $k(n) \to \infty$ but also $k(n)/n \to 0$. However, since extreme events by definition are rare, there is only little data (few observations) that can be utilized and this inevitably involves an added amount of statistical uncertainty. A possible solution would be to smooth 'somehow' the estimates with respect to the choice of k (i.e. make the plot more insensitive to the choice of k), leading to a more stable plot and a more reliable estimate of γ. Such a method was proposed by Resnick and Stărică (1997,1999) for smoothing Hill and moment estimators, respectively.

The authors proposed simple averaging techniques for reducing the volatility of the corresponding Hill and moment plots. In both cases, the smoothing procedure consists of averaging the estimator's values corresponding to different values of order statistics p. The generic formula of the proposed averaged estimators is

$$\operatorname{av}\hat{\gamma}_{(\cdot)}(k) = \frac{1}{k - [ku]} \sum_{p=[ku]+1}^{k} \hat{\gamma}_{(\cdot)}(p),$$

where $\hat{\gamma}_{(\cdot)}$ is Hill's or moment estimator, $u < 1$, and $[x]$ the smallest integer greater than or equal to x.

The authors (Resnick and Stărică, 1997) derived the adequacy (consistency and asymptotic normality) of the averaged-Hill's estimator, as well as its improvement over Hill's estimator (smaller asymptotic variance).

In the case of averaged-moment estimator (Resnick and Stărică, 1999), the consequent reduction in asymptotic variance is not so profound. The authors actually showed that through averaging (using the above formula), the variance of the moment estimator can be considerably reduced only in the case $\gamma < 0$. For $\gamma > 0$, the simple moment

estimator turns out to be superior, while for $\gamma \approx 0$ the two moment estimators are almost equivalent.

In Tsourti and Panaretos (2001), the idea of the above smoothing procedure has been applied to other standard extreme-value index estimators and their adequacy has been evaluated via a simulation study.

3. Application of extreme-value analysis to teletraffic data

3.1. Introduction

One of the areas where extreme-value theory has recently gained ground is teletraffic engineering. Indeed, as the Internet becomes more and more popular, the need for evaluating its performance becomes 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 file length, cpu time to complete a job, call holding times and so on. In order for 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 length, longest call holding times, etc.) turns out to be very relevant to teletraffic engineers.

In this paper, 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*). In particular, we analyse data from the 'EPA-HTTP' trace. This trace contains a day's worth (August 30, 1995) of all HTTP requests to the EPA WWW server. In the present extreme-value analysis, we concentrate on the analysis of the file length requested (i.e. on the bytes in the reply). In Section 3.2, an exploratory data analysis is provided, while in Section 3.3 the core extreme-value analysis is presented consisting of the estimation of extreme-value index γ as well as some large quantiles.

3.2. Exploratory data analysis

3.2.1. Description of the data

The original data-set contained 47,748 cases of requests. Still, in 5331 of these the file length requested was not recorded, while in 5718 additional observations no file was actually requested (i.e. the file length is zero). These 11,049 observations, in total, were removed. So the final data-set, on which the analysis that follows has been based, considers 36,699 cases of requests (expressed as file lengths in bytes). In Table 1 that follows, we present the main descriptive statistics of the variable under investigation, while a more intuitive description of the data is provided by the histograms that follow in Fig. 1. Fig. 1a is the histogram of the smaller values of 'File length', while in Fig. 1b the histogram of large 'File length' is depicted. It is obvious that we are dealing with a possible heavy-tailed underlying distribution. Still, a more thorough discussion on this issue is postponed until the next section.

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

Table 1
Descriptives statistics of 'File length' (in bytes)

Measure	Statistic
Mean	8497.89
5% Trimmed mean	3046.79
Median	1897.00
Std. deviation	70718.24

the tendency of extreme values to occur in clusters. These are better revealed by a sequence plot (see Fig. 2). A first examination of it reveals no problems of seasonality or clustering.

3.2.2. Investigation of independence

One of the assumptions required for almost all the results in extreme-value theory is that of independence of the data. For the case of Hill's estimator, there are results that prove the good properties of the estimator 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. This issue is especially crucial in the context of World Wide Web, where correlations may occur, for example, in servers providing patches. In such cases, large communities might begin downloading patches as soon as they contact the server and, thus, extremal events may occur in clusters. When such phenomena take place, the good properties of the aforementioned estimators do not necessarily hold and special techniques need to be developed. In Embrechts et al. (1997), this issue is dealt with in depth, where the extremal index is introduced in order to characterize the dependence structure of the data and its connection to its extremal behaviour.

So, in order to examine whether further actions are needed in our case, we apply several tests of randomness and visually examine the autocorrelation plot.

In our case, we used the standard difference-sign test as well as a test of randomness based in records (Embrechts et al., 1997). In this context, records referring to 'temporary maximum values' (i.e., a value X_n is characterized as a record if $X_n > \max\{X_1, \ldots, X_{n-1}\}$), are indicative of the extreme-value behaviour of the phenomenon under study, too. Embrechts et al. (1997) explored their distribution under the assumption of independence, leading, thus, to a (rough) non-parametric test of randomness.

According to the difference-sign test, the hypothesis of independence cannot be rejected (the observed level of significance is 0.08). Moreover, for the data-set under investigation the expected number of records is 11.1 with variance 9.4, while the observed number of records equals 12. These values seem to support the hypothesis of independence (the observed level of significance under the approximate assumption of normality is 0.77).

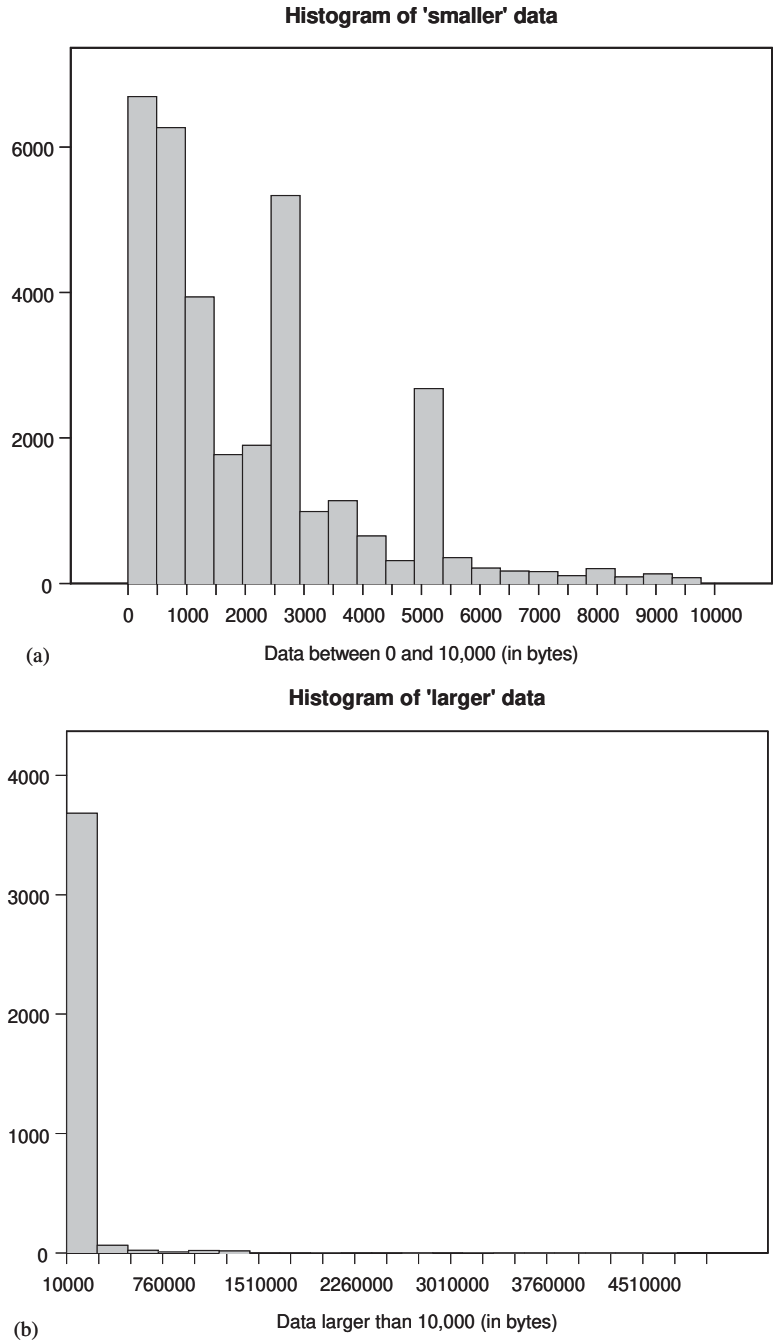


Fig. 1. Histograms of separate parts of data (File length).

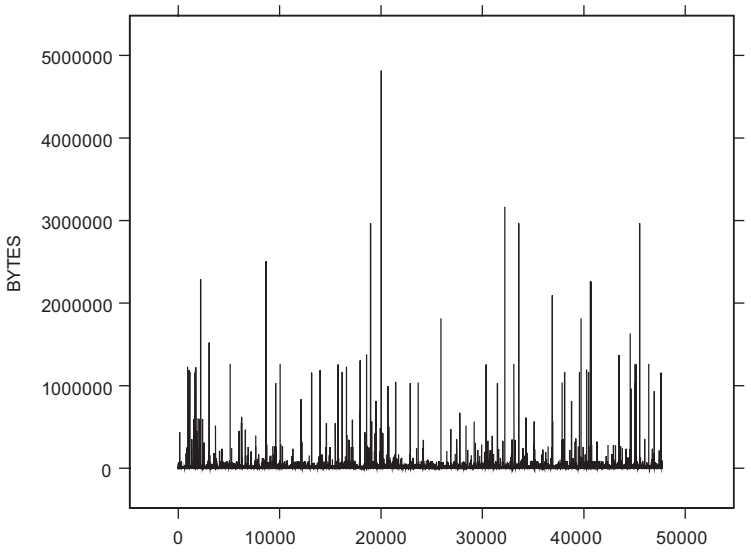


Fig. 2. Sequence plot of 'File length' (in bytes).

Another exploratory, informal method for testing for independence can be based on the sample autocorrelation function. However, 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:

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

If, on graphing the sample heavy-tailed autocorrelation function, one finds only small values, then it may be possible to model the data as independent and identically distributed. The heavy-tailed autocorrelation plot for our data is given in Fig. 3. The majority of values do 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 the data. Moreover, the fact that the autocorrelation function does not display any particular pattern with respect to the lag h, is reassuring that no statistical autocorrelation exists in the data.

Since all the previous non-parametric and graphical checks of independence do not give us an indication that dependence in our data exists, we proceed 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.

3.2.3. Investigation of heavy tails

Before proceeding to the formal study of extremes of the data in hand, there are several exploratory methods that can be used to give us a first insight into the

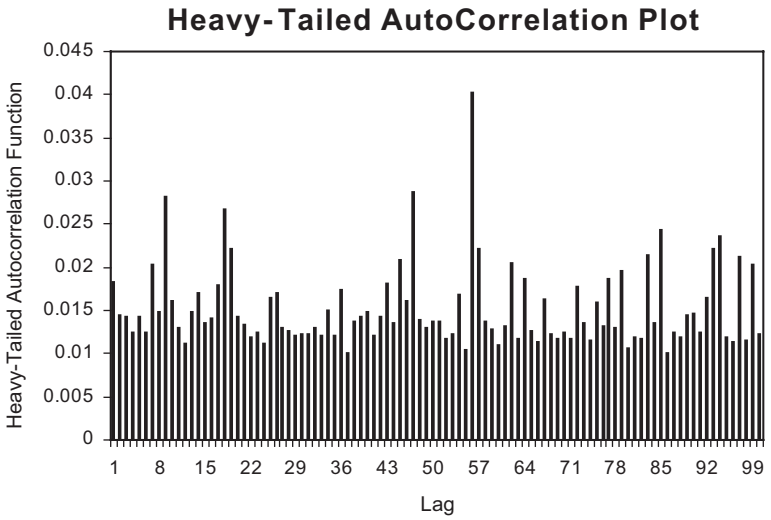


Fig. 3. Heavy-tailed autocorrelation plot of 'File length' for lags $h = 1, \ldots, 100$.

behaviour of the extremes of a data-set. Such methods include the mean excess plot and 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 ($\gamma \geqslant 0$), or short tailed ($\gamma < 0$). Knowledge, even rough, of the sign of γ can direct us to the choice of more appropriate 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.

The definition and properties of mean excess functions (MEF) and the corresponding mean excess plots (MEP) are given in Beirlant et al. (1996). If the MEF of the logarithmic-transformed data is ultimately increasing, then the d.f. belongs to MDA(H_γ) with $\gamma > 0$, and the values of the MEF converge to the true value of γ. However, one of the main assumptions of MEP, in order to be reliable, is that the underlying distribution has a finite first moment, which makes them inappropriate for long-tailed distributions with $\gamma > 1$. For this reason, we proceed to QQ plots, which do not have such restrictive assumptions.

The use of QQ plots as exploratory tools in extreme-value analysis is described in detail in Beirlant et al. (1996). The idea is that by constructing QQ plots of standard distributions (medium or heavy tailed) and by focusing on the upper right part of these plots (i.e. largest values), the evaluation of the fit of the data to the tails of the standard distribution may be insightful about the tails of the data themselves. In the sequel, we present the QQ plots of our data-set, with respect to the exponential (Fig. 4) and the Pareto (Fig. 5) distributions. These are distributions medium and long tailed, respectively, commonly used in practice.

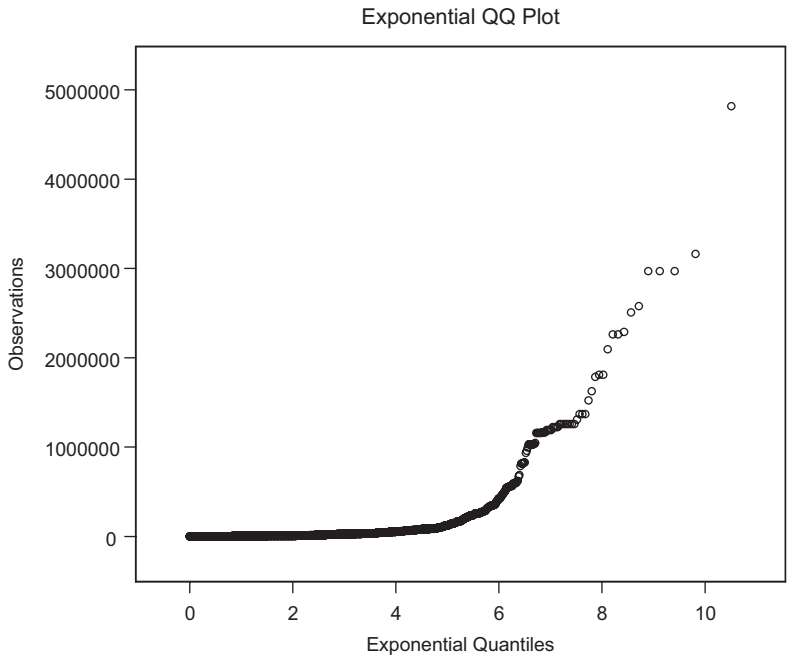


Fig. 4. Exponential QQ plot of the 'File length' data-set.

As we have previously mentioned, any interpretation of the above plots is going to be focused on the upper right part of those. In the exponential QQ plot, even if we ignore the very few extreme values which display great variability, not a straight pattern can be detected in the right part of the plot and, thus, gives us no indication that the tails of the teletraffic data in hand have exponential tails. On the other hand, in the right part of the Pareto QQ plot a linear pattern is made apparent, implying that 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, that is, $F \in \mathrm{MDA}(H_\gamma)$, $\gamma > 0$. Still, the formal investigation of the extremal behaviour of the data-set under study, comes in the section that follows.

3.3. Extreme-value analysis

3.3.1. Estimation of extreme-value index γ

We now deal with the main scope of the current analysis, which is the investigation of the extremal behaviour of 'File length' transported via the site of EPA. This is, essentially, 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 $\gamma > 0$. Based on simulation

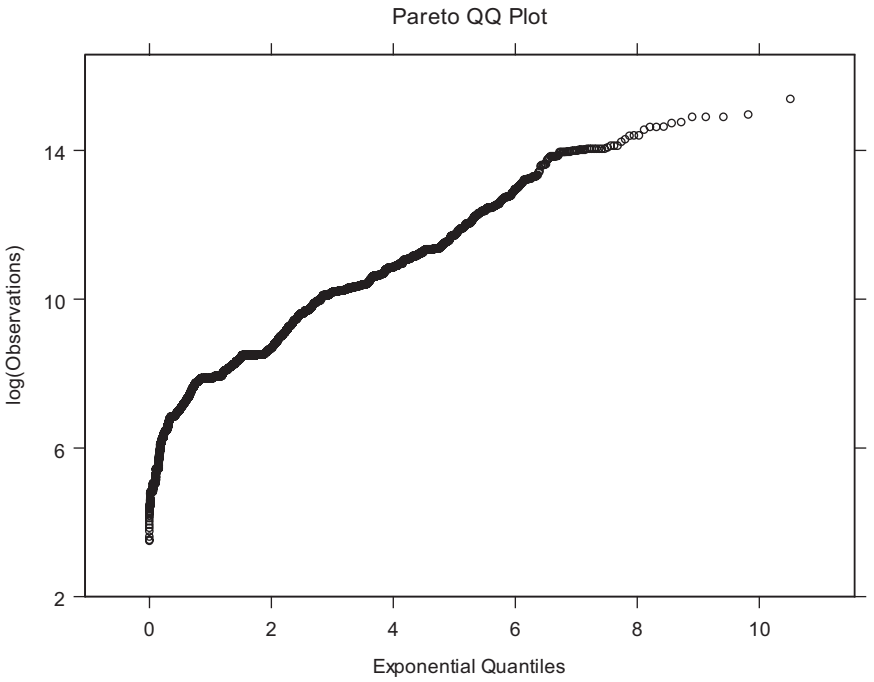
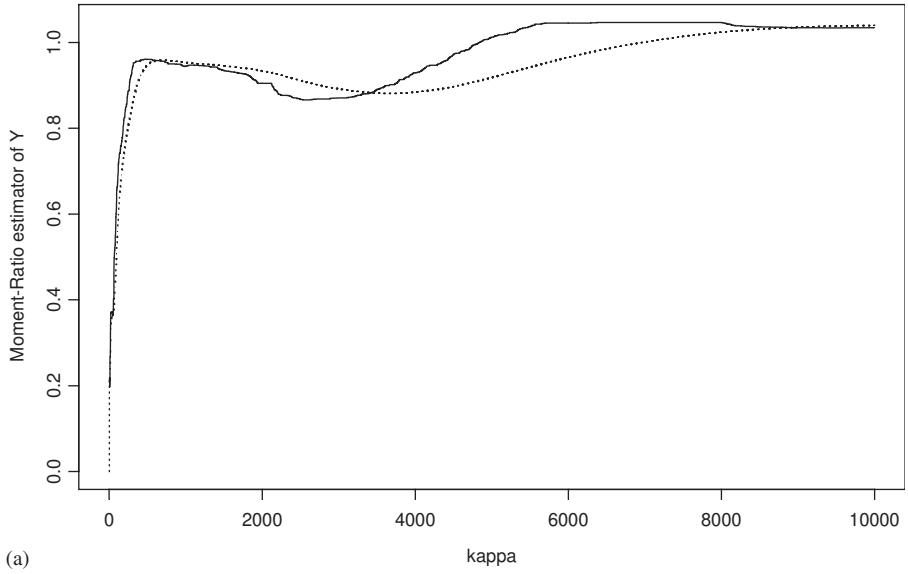


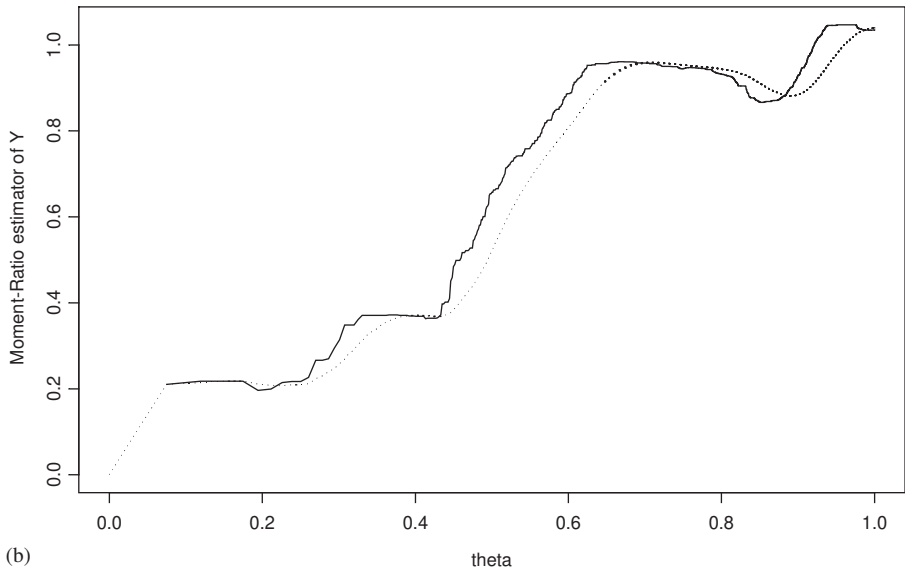
Fig. 5. Pareto QQ plot of the 'File length' data-set.

studies that exist in the literature (e.g. Deheuvels et al., 1997; Rosen and Weissman, 1996; Tsourti and Panaretos, 2001), we are going to estimate γ using moments-ratio, moment and Hill's estimators, which are judged to be more efficient for the case $\gamma > 0$. 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). In the sequel we provide the plots $(k, \gamma(k))$ of the estimators used, as well as the corresponding 'alternative plots', 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-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 data-set). The purpose of this is to focus only on the part of data that essentially concern us (large values). In this way, we can get a better view of the part of the graph that we are actually interested in (Figs. 6–8).

It is fortunate that in our case all the estimators tend to approximately the same value of γ, the value 1. Especially, moments-ratio and moment estimators which have (according to Tsourti and Panaretos, 2001) 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 close to 1. This implies that



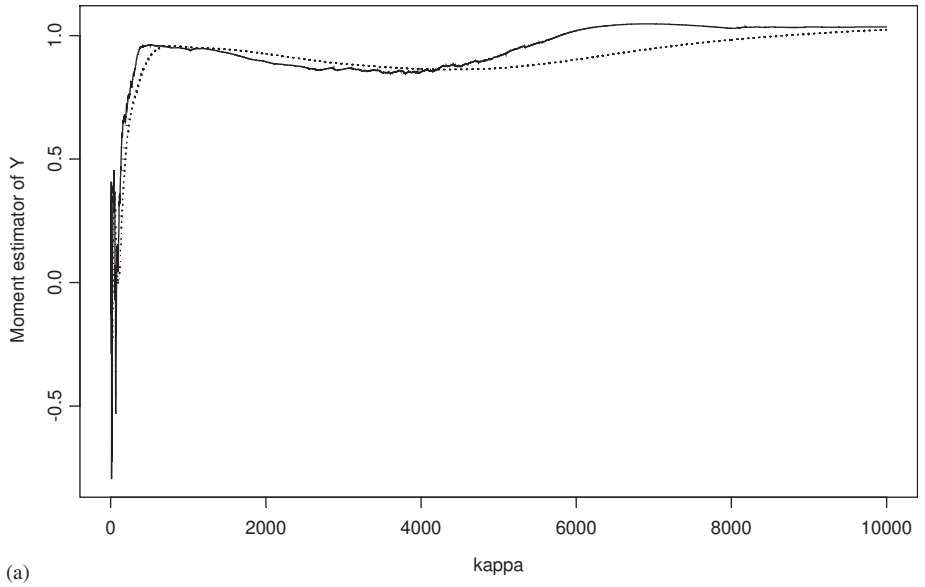
(a)



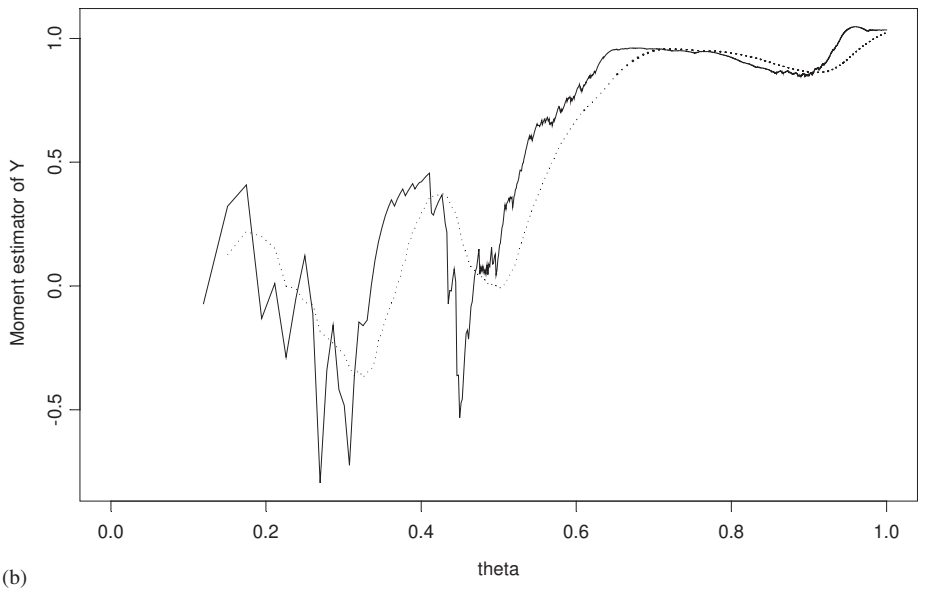
(b)

Fig. 6. Plot (a) and alternative plot (b) of moments-ratio estimator of γ (—) and the corresponding mean-averaged $(\cdots)$ estimator.

the underlying distribution of the data under study belongs to the maximum domain of attraction of the GEV(1) distribution, i.e. it is a Pareto-type d.f. asymptotically decaying like a Pareto (1) ($\bar{F}(x) \xrightarrow[x\to\infty]{} x^{-1}$).



(a)



(b)

Fig. 7. Plot (a) and alternative plot (b) of moment estimator of γ (—) and the corresponding mean-averaged $(\cdots)$ estimator.

3.3.2. 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

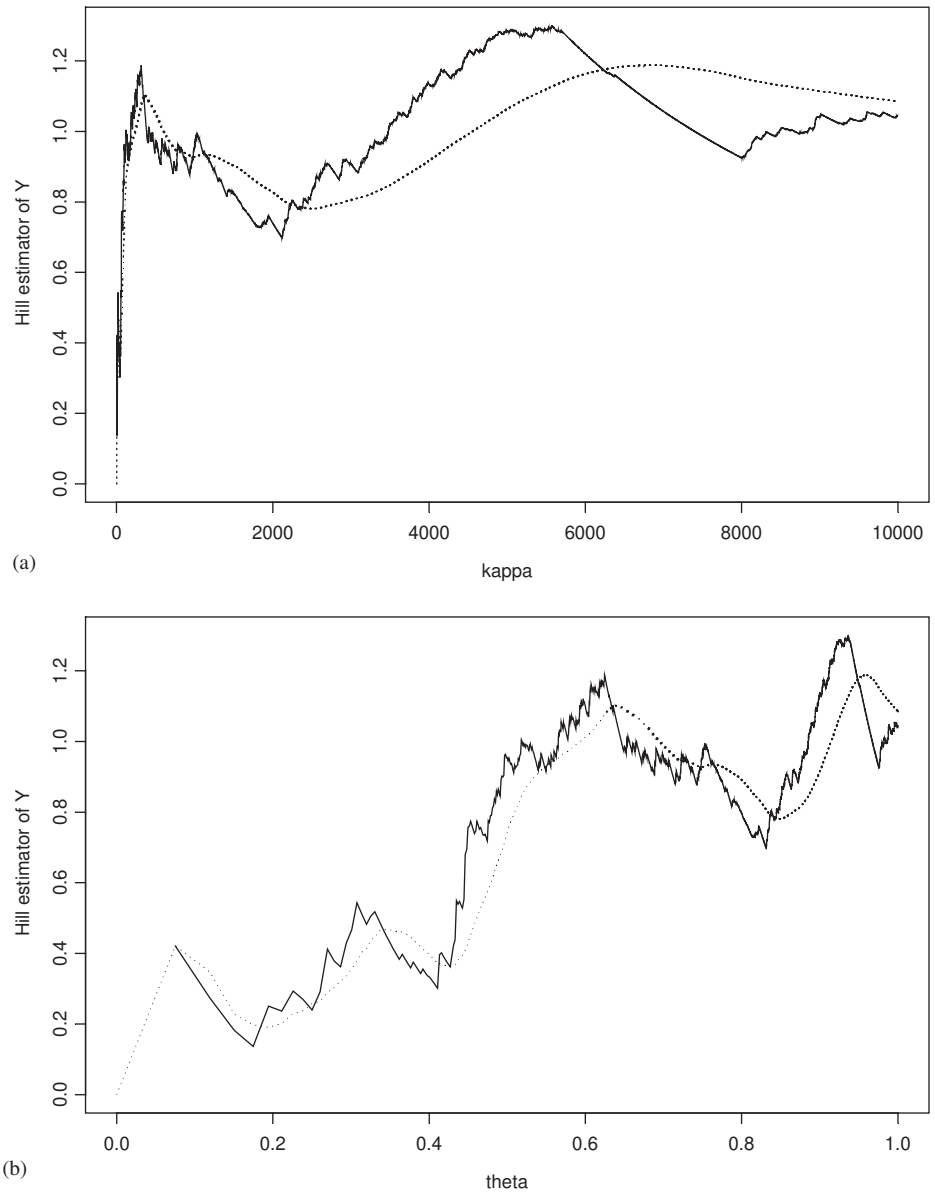


Fig. 8. Plot (a) and alternative plot (b) of Hill's estimator of γ (—) and the corresponding mean-averaged $(\cdots)$ estimator.

practical purposes is large quantiles. That is, in practice what is desirable to know is the 'File Length' that has exceeded only 1 in x times/transactions (x large). Each extreme-value index estimator leads to a different estimation formula for large quantiles, which is, also, dependent on k. Here, we use the generic formula proposed by

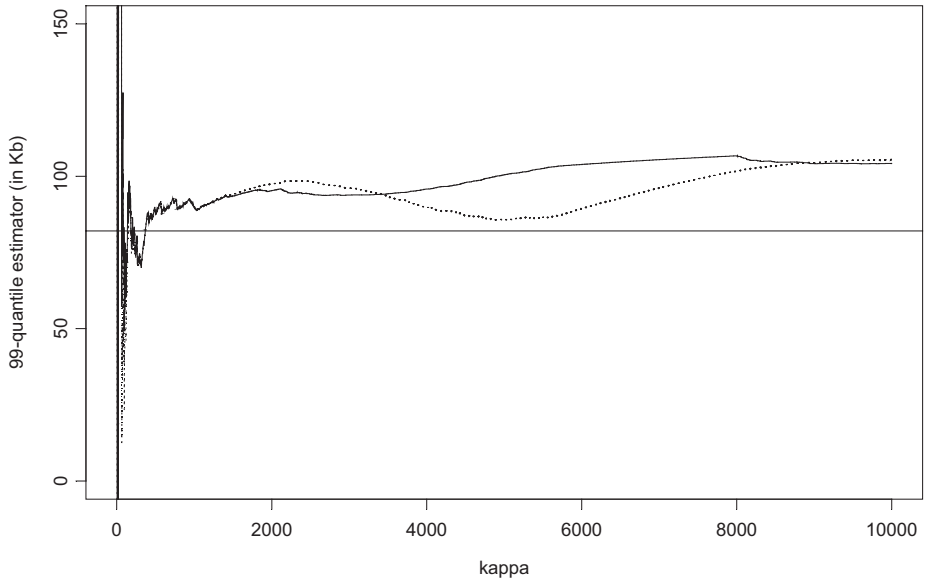


Fig. 9. Plot of 99-quantile based on moments-ratio estimator (—) and the corresponding mean-averaged (···) estimator.

Dekkers et al. (1989):

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

$$\text{where } a_n = \frac{k}{n(1-p)}, \quad \rho_1(\gamma) = \begin{cases} 1, & \gamma \geqslant 0, \\ (1-\gamma)^{-1}, & \gamma < 0, \end{cases}$$

substituting each different estimator $\gamma_{(\cdot)}$.

In Figs. 9–11 we present the estimators of 99% quantile based on the extreme-value index estimators previously used, versus k (the straight line indicates the empirical 99% quantile).

An interesting finding of these figures is that the behaviour of quantile estimators does not seem to display the nice stability (with respect to k) as was the case for the extreme-value index estimators themselves. In any case, the 99% quantile estimator based on the moments-ratio estimator displays the most stable behaviour indicating a value of 99-quantile approximately 100 Kb (though constantly larger than the corresponding empirical estimate which is 82 Kb).

In Table 2, we provide the estimators of 95%, 99%, and 99.9% quantiles, based on the moments-ratio estimator of γ, for several values of k.

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 similarly to a Pareto (1) distribution). This property may be further exploited in order

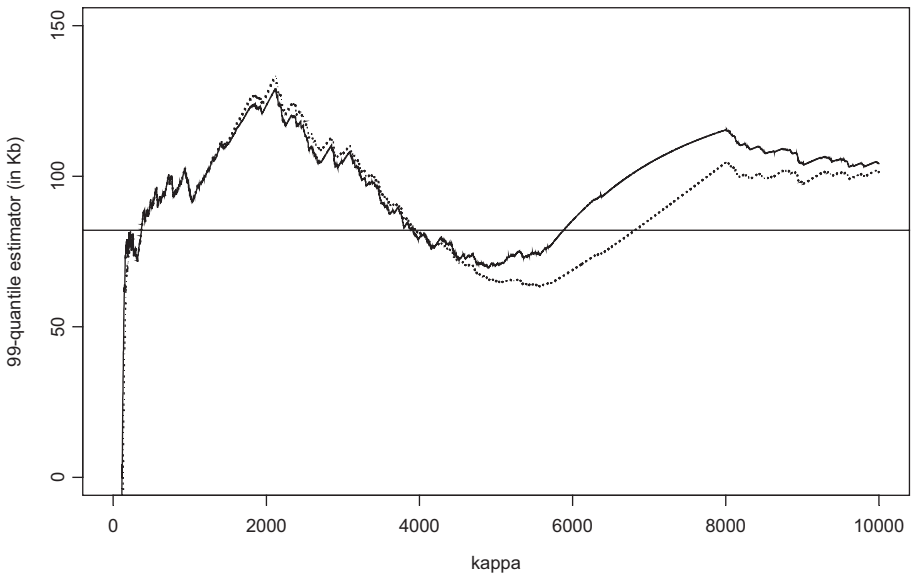


Fig. 10. Plot of 99-quantile based on moment estimator (—), and the corresponding mean-averaged (···) estimator.

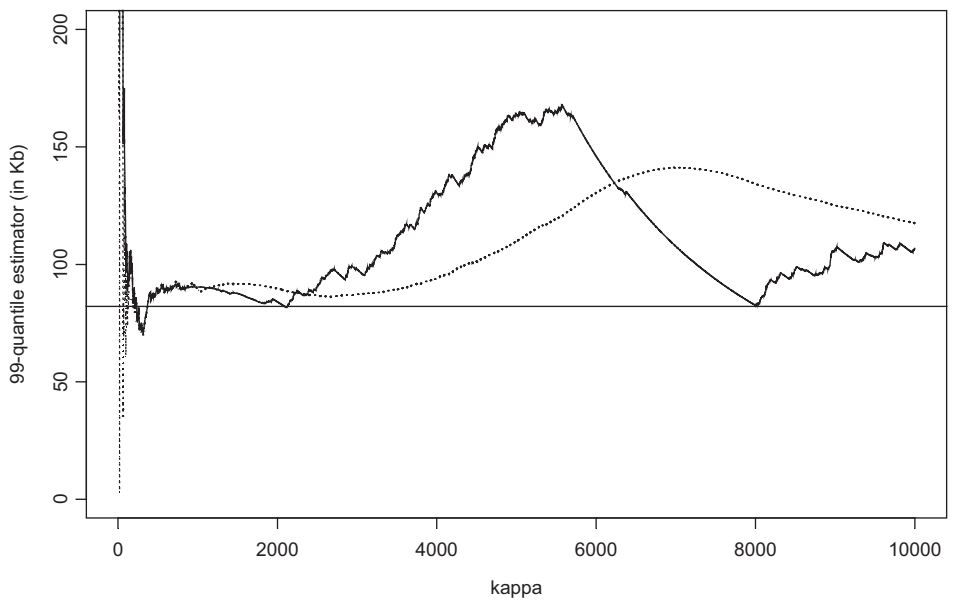


Fig. 11. Plot of 99-quantile based on Hill's estimator (—), and the corresponding mean-averaged (···) estimator.

Table 2
Estimation of large quantiles (in Kb) using moment-ratio estimator of γ

	Moments-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 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 20 Kbs, the 99-quantile reaches 100 Kbs, while a file larger than 1 Mb is requested only one in thousand times.

4. Conclusions

The sharp increase of Internet's popularity, as this can be expressed by the expansion of the number of servers and the number of users, led the Web to performance problems and generated the need for efficient design and administration. In order to achieve this, teletraffic engineers proceed to measurement and study of Internet behaviour's, through phenomena such as cpu time to complete a job, call holding times, file length requested and inter-arrival times. Increasing instrumentation of teletraffic networks has made possible the acquisition of large amounts of data, indicating, though, that the usual assumptions of classical queuing models are not valid, since long-tailed behaviours are observed. Extreme-value analysis can prove to be particularly helpful in such circumstances, providing systematic information on the extremes observed in a network.

In this paper, we have presented an application of extreme-value analysis on data from the field of teletraffic engineering. Particularly, the right tail of distribution of the size of files requested from EPA site has been examined. Graphical methods as well as the estimation of extreme-value index indicate the 'heavy-tailness' of the phenomenon under study. Based on our analysis, we concluded that a file of size 100 Kbs is requested once in hundred times, while a 1 Mb file only once in a thousands requests.

References

Barão, M.I., Tawn, J.A., 1999. Extremal analysis of short series with outliers: sea-levels and athletic records. Appl. Statist. 48 (4), 469–487.

Beirlant, J., Teugels, J.L., Vynckier, P., 1994. Extremes in non-life insurance. In: Galambos, J., Lenchner, J., Simiu, E. (Eds.), Extremal Value Theory and Applications. Kluwer, Dordrecht, pp. 489–510.

Beirlant, J., Teugels, J.L., Vynckier, P., 1996. Practical Analysis of Extreme Values. Leuven University Press, Leuven.

Coles, S.G., Tawn, J.A., 1996. Modelling extremes of the areal rainfall process. J. Roy. Statist. Soc. Ser. B 58 (2), 329–347.

Crovella, M.E., Taqqu, M.S., Bestavros, A., 1998. Heavy-tailed probability distributions in the world wide web. Chapter 1 in the book: A Practical Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy Tailed Distributions, Vol. 3-25. Chapman & Hall, New York.

Danielsson, J., de Vries, C.G., 1997. Beyond the sample: extreme quantile and probability estimation. Erasmus University, Rotterdam, Preprint.

Danielsson, J., Jansen, D.W., de Vries, C.G., 1996. The method of moment ratio estimator for the tail shape distribution. Comm. Statist.—Theory Methods 25 (4), 711–720.

Davison, A.C., Smith, R.L., 1990. Models for exceedances over high thresholds. J. Roy. Statist. Soc. Ser. B 52 (3), 393–442.

Deheuvels, P., de Haan, L., Peng, L., Pereira, T.T., 1997. NEPTUNE T400:EUR-09, Comparison of extreme-value index estimators.

Dekkers, A.L.M., de Haan, L., 1989. On the estimation of the extreme-value index and large quantile estimation. Ann. Statist. 17 (4), 1795–1832.

Dekkers, A.L.M., Einmahl, J.H.J., de Haan, L., 1989. A moment estimator for the index of an extreme-value distribution. Ann. Statist. 17 (4), 1833–1855.

Duffy, D., McIntosh, A., Rosenstein, M., Willinger, W., 1994. Statistical analysis of CCSN/SS7 traffic data from working CCS subnetworks. IEEE J. Selected Areas Commun. 12, 544–551.

Embrechts, P., 1999. Extreme Value Theory in Finance and Insurance. Manuscript, Dept. of Math., ETH, Swiss Federal Technical University.

Embrechts, P., Klüppelberg, C., Mikosch, T., 1997. Modelling Extremal Events for Insurance and Finance. Springer, Berlin.

Embrechts, P., Resnick, S., Samorodnitsky, G., 1998. Living on the edge. RISK Mag. 11 (1), 96–100.

Embrechts, P., Resnick, S., Samorodnitsky, G., 1999. Extreme value theory as a risk management tool. North Amer. Actuar. J. 26, 30–41.

Fang, Y., Chlamtac, I., Lin, Y.-B., 1997. Call performance for a PCS network. IEEE J. Selected Areas Commun. 15 (8), 1568–1581.

Fisher, R.A., Tippet, L.H.C., 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proc. Cambridge Philos. Soc. 24 (2), 163–190 (in Embrechts et al., 1997).

Hill, B.M., 1975. A simple general approach to inference about the tail of a distribution. Ann. Statist. 3 (5), 1163–1174.

Kratz, M., Resnick, S.I., 1996. The QQ estimator and heavy tails. Commun. Statist.—Stochast. Models 12 (4), 699–724.

McNeil, A.J., 1997. Estimating the tails of loss severity distributions using extreme value theory. ASTIN Bull. 27, 117–137.

McNeil, A.J., 1998. On extremes and crashes. A short-non-technical article. RISK, January 1998, p. 99.

McNeil, A.J., 1999. Extreme Value Theory for Risk Managers. Internal Modelling and CAD II, London: RISK Books, pp. 93–113.

Mikosch, T., 1997. Heavy-tailed modelling in insurance. Commun. Statist.—Stochast. Models 13 (4), 799–815.

Naldi, M., 1999. Measurement-based modelling of internet dial-up access connections. Comput. Networks 31 (22), 2381–2390.

Pickands, J., 1975. Statistical inference using extreme order statistics. Ann. Statist. 3 (1), 119–131.

Resnick, S., 1997. Heavy tailed modelling and teletraffic data. Ann. Statist. 25 (5), 1805–1869.

Resnick, S., Stărică, C., 1995. Consistency of Hill's estimator for dependent data. J. Appl. Probab. 32, 139–167.

Resnick, S., Stărică, C., 1997. Smoothing the Hill's estimator. Adv. Appl. Probab. 29, 271–293.

Resnick, S., Stărică, C., 1999. Smoothing the Moment estimator of the extreme value parameter. Extremes 1 (3), 263–293.

Rootzén, H., Tajvidi, N., 1997. Extreme value statistics and wind storm losses: a case study. Scand. Actuar. J. 70–94.

Rosen, O., Weissman, I., 1996. Comparison of estimation methods in extreme value theory. Comm. Statist.—Theory Methods 25 (4), 759–773.

Smith, R.L., 1989. Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. Statist. Sci. 4 (4), 367–393.

Tsourti, Z., Panaretos, J., 2001. Extreme-value index estimators and smoothing alternatives: review and simulation comparison. Technical Report No 149, Athens University of Economics and Business, Greece.

Willinger, W., Taqqu, M., Sherman, R., Wilson, D., 1997. Self-similarity through high variability: statistical analysis of Ethernet LAN traffic at the source level. IEEE/ACM Trans. Networking 5, 71–96.