Statistical Modeling of Extreme Values with Applications to Air Pollution

H. M. Barakat[1], E. M. Nigm[1] and O. M. Khaled[2]

[1]Department of Mathematics Faculty of Science Zagazig University, Zagazig, Egypt
[2]Department of Basic Science Faculty of Engineering, Sinai University, El-Arish, Egypt Email
hbarakat2@hotmail.com s_nigm@yahoo.com osam87@yahoo.com

Abstract: In this paper the Block Maxima and the Peak Over Threshold methods are used to model the air pollution in two cities in Egypt. A simulation technique is suggested to choose a suitable threshold value. The validity of full bootstrapping technique for improving the estimation parameters in extreme value models has been checked by Kolmogorov-Smirnov test. A new efficiency approach for modeling extreme values is suggested. This approach can convert any ordered data to enlarged block data by using sup-sample bootstrap. Although, this study is applied on three pollutants in two cities in Egypt, but the suggested approaches may be applied on other pollutants in other regions in any country.

[H. M. Barakat, E. M. Nigm and O. M. Khaled **Statistical Modeling of Extreme Values with Applications to Air Pollution**] Life Science Journal, 2012; 9(1): 124-132] (ISSN: 1097-8135). http://www.lifesciencesite.com.

Key words: Air pollution; Generalized extreme value model; Generalized Pareto distribution; Kolmogorov-Smirnov test; Bootstrap technique.

1. Introduction

The traditional method of analyzing extreme values is based on the extreme value limiting distributions, which were derived by Gnedenko (1943) and Reiss and Thomas, (2003). These limits are known as Extreme Value Distributions (EVD) and they arise as limiting for distribution of maximum sample of independent and identically distributed (iid) random variables (rv's). EVD are often used to model natural phenomena such as sea levels, river heights, rainfall and air pollution. Two main methods for modeling, the Block Maxima (BM) method and the Peak Over Thresholds (POT) method, have been developed (Coles 2001).

In the BM method it is supposed to have observed maxima values of some quantities over a number of blocks. A typical example is a block is year or day and the observed quantities may be some environmental quantity such as the wind speed or air pollutant at a specific location. In this method, the block maxima is modeled by EVD. The choice of EVD is motivated by the facts: (i) The EVD are the only ones which can appear as the limit of linearly normalized maxima. (ii) They are the only ones which are max-stable, i.e., such that a change of block size only leads to a change of location and scale parameters in the distribution.

In the POT method it is supposed to have all observed values, which are larger than some suitable threshold. These values are then assumed to follow the Generalized Pareto Family of Distributions (GPD). The choice of GPD is motivated by two characterizations: (i) The distribution of scale normalized exceedance over threshold asymptotically converges to a limit belonging to GPD if and only if the distribution of BM converges (as the block length tend to infinity) to one of EVD. (ii) The distributions belonging to the GPD are the only stable ones, i.e., the only ones for which, the conditional distribution of an exceedance is scale transformation of the original distribution.

A number of studies have shown a positive association between air pollution and human health effects (Goldberg et al., 2001 and Kim et al., 2004). We choose in this study three pollutants: Sulphur Dioxide SO_3, Ozone O_3 and Particulate Matter $PM10$ in 10^{th} of Ramadan and Zagazig cities. The study of the Ozone pollutant was restricted on 10^{th} of Ramadan city. The first city is one of the largest industrial cites in Egypt and the second is one of the most populous. Devices have been installed to monitor these pollutants in different places in these two cities. The places of these devices have been selected by experts in environmental measurements. The measurement units of the pollutants is $\mu gm/m^3$. The data for these pollutants were recorded every hour on the twenty-four hours through year 2008 for the two cities, except Ozone was recorded every half hours. The detail description of these pollutants and the collected data can be founded in (Barakat et al., 2011). This study considered the BM and POT methods, which are used to evaluate the measurement O_3, SO_2 and $PM10$ in two cities in Egypt. Bootstrapping technique for improving the estimation parameters in extreme value model is used and its validity is checked by the

Kolmogorov − Smirnov test. A simulated technique is suggested to choose a suitable value of threshold in the POT method. Moreover, a new efficiency method for modeling extreme values is suggested. This method, based on the work of Athreya and Fukuchi (1997), can convert any ordered data to enlarged block data by using sup-sample bootstraps. This method enables the engineers to analys the rare events to construct dam for rivers, breakwater for sea defence, and to design nuclear power plant against earthquakes, where the number of available maxima about the relevant phenomena of these activities are often limited.

2. Mathematical Models

Let $X_1, X_2, \cdots, X_n$ be iid rv's with common df $F(x) = P(X \le x)$. Suppose that $M_n = \max\{X_1, X_2, \ldots, X_n\}$. The cornerstone of extreme value theory is the Extremal Type Theorem (ETT) (see, Reiss and Thomas, 2003), which states that: If there exist sequences of constants $a_n > 0$ and b_n , such that $P(\frac{M_n - b_n}{a_n} \le x) = F^n(a_n x + b_n)$ weakly converges to a nondegenerate df $G(x)$, then G should be of the same type of the Generalized Extreme Value Distribution (GEVD)

$$G_\gamma(x; \mu, \sigma) = \exp[-[1 + \gamma(\frac{x-\mu}{\sigma})]^{-\frac{1}{\gamma}}], \quad 2.1$$

which is a unified model for the EVD. Apart from a change of origin (the location parameter μ) and a change in the unit on the $x-$axis (the scale parameter $\sigma > 0$) the GEVD yields the three EVD, according as $\gamma > 0$, $\gamma < 0$ and $\gamma = 0\,(\gamma \to 0)$, which are known as Frechet, Weibull and Gumbel families of df's, respectively. In this case, any suitable standard statistical methodology from parametric estimation theory can be utilized in order to derive estimate of the parameters μ, σ and γ. In this paper, we used the maximum likelihood method (ML) and improved the obtained estimates by the bootstrap technique. The bootstrap is a data-driven method that has a very wide range of applications in statistics. This technique is initiated by Efron (1979). The classic bootstrap approach uses Monte Carlo simulation to generate an empirical estimate for the sampling distribution of the statistic by randomly drawing a large number of samples of the same size n from the data, where n is the size of the sample under consideration. Therefore, the bootstrap is a way of finding the sampling distribution, at least approximately, from just one sample. Here is the procedure:

Step 1: Re-sampling.

A sampling distribution is based on many random samples from the population. In place of many samples from the population, create many re-samples by repeated sampling with replacement from this one random sample. Each re-sample is of the same size as the original random sample.

Step 2: Bootstrap distribution.

The sampling distribution of a statistic collects the values of the statistics from many samples. The bootstrap distribution of a statistic collects its values from re-samples.

The BM approach is adopted whenever the data set consists of maxima of independent samples. In practice, some blocks may contain several among the largest observations, while other blocks may contain none. Therefore, the important information may be lost. Moreover, in the case that we have a few number of data, block maxima can not be actually implemented. For all these reasons, the BM method may be seen restrictive and not very realistic. In our study, we used this method to get the preliminary result, to help simulate data with the same nature as the real data.

An alternative approach, POT method, to determine the type of asymptotic distribution for extremes is based on the concept of GPD. This approach, which was initiated by Pickands (1975), is used to model data arising as independent threshold exceedances. Actually, the POT method is based on the fact that the conditional df $F^{[u]}(x+u) = P(X < x+u \mid x > u)$ may be approximated for large u (i.e., the threshold u is close to the right endpoint $w(F) = \sup\{x : F(x) < 1\}$) by the family $W_\gamma(x; \overline{\sigma}) = 1 - (1 + \gamma \frac{x}{\overline{\sigma}})^{-\frac{1}{\gamma}}$, provided that the df of BM weakly converges to the limit G_γ, which is defined by (2.1). In this case we have $\overline{\sigma} = \sigma - \gamma\mu$ (Reiss and Thomas, 2003). This family is connected by the GEVD by the simple relationship

$$W_\gamma(x;\sigma) = 1 + \log G_\gamma(x;0,\sigma), \log G_\gamma(x;0,\sigma) > -1.$$

It is worth to mention that the left truncated GPD yields again a GPD, namely:

$$W_\gamma^{[c]}(x;\sigma^*) = W_\gamma(x;\overline{\sigma}), \text{where } \sigma^* = \overline{\sigma} + \gamma c. \quad 2.2$$

Notice that the GPD nests the Pareto,

uniform and exponential distributions. Evidently, in the statistical modeling of threshold exceedance data, the whole data are used, in opposite of the case of the BM method. Possibly, the most important issue in statistical modeling of threshold exceedances data is the choice of threshold u. Did we choose a high enough threshold? the threshold should be hight enough to justify the assumptions of the model but low enough to a capture a reasonable number of observations. A threshold choice based on the observed sample is required to balance these two opposing demands. In this paper we used a simulation technique to choose a suitable threshold value. Namely, we first note that the GPD are the only continuous df's W such that for a certain choice of constants b_u and a_u,

$$W^{[u]}(b_u + a_u x) = W(x)$$

is again the exceedance df at u (Reiss and Thomas, 2003). This property is the (POT)-stability of GPD. Now, let γ_0, σ_0 and μ_0 be the preliminary estimates of the parameters γ, σ and μ, respectively (which is obtained by the BM method). Then, simulate data with the same size n as the real collected data from the GPD $W_{\gamma_0}^{[c]}(x; \sigma_0^*)$, with $c = \min\{x_1, x_2, ..., x_n\}$, where $x_1, x_2, ..., x_n$ is are the real data (this choice of c grantees that the simulated and realistic data have nearly the same range) and $\sigma_0^* = \sigma_0 + \gamma_0(c - \mu_0)$ (in view of (2.2)). In view of the POT stability property of GPD, the simulated data will have the same nature as the real collected data. Moreover, any POT u from the simulated data follows the GPD with the same shape parameter. Therefore, we choose the value of u which makes the estimate of the known shape parameter as best as we can. Finally, we take this value of u as a suitable threshold for our real data.

All the described models so far can be fitted by the method of ML, Cox and Hinkley (1974). Actually, the log likelihood function of the GEVD is given by

$$l(\underset{-}{x};\mu,\sigma,\gamma) = -n\log\sigma + \sum_{i=1}^{n}\left(-\left[1+\gamma\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\gamma}} - \left(1+\frac{1}{\gamma}\right)\log\left[1+\gamma\left(\frac{x_i-\mu}{\sigma}\right)\right]\right) \qquad 2.3$$

provided $1 + \gamma(x_i - \mu)/\sigma > 0$, for each i, otherwise (2.3) is undefined. For the maximization of $l(\underset{-}{x}; \mu, \sigma, \gamma)$ for a general model indexed by parameters μ, σ, γ, this may be performed using a packaged nonlinear optimization subroutine, of which several excellent versions are available. Also the log likelihood function for GPD is given by

$$l^*(\underset{-}{x};\bar{\sigma},\gamma) = -n\log\bar{\sigma} - \left(1+\frac{1}{\gamma}\right)\sum_{i=1}^{k}\log\left(1+\frac{\gamma x_i}{\bar{\sigma}}\right), \qquad 2.4$$

where k is the number of POT. Finally, we should say something about the theoretical status of the approximations involved. The asymptotic theory of ML for the GEV model is valid provided $\gamma > -0.5$ (see Smith, 1985). Cases with $\gamma \le -0.5$ correspond to an extremely short upper tail and hardly ever occurs in environmental applications. A more serious problem is that even when, $\gamma > -0.5$, the asymptotic theory may give rather poor results with small sample sizes.

The Kolmogorov − Smirnov test (K-S test) is a nonparametric test for the equality of continuous one-dimensional df that can be used to compare a sample with a reference df (one-sample K-S test). The Kolmogorov − Smirnov statistic quantifies a distance between the empirical df of the sample and the reference df. Assume we have the hypothesis-testing situation $H_0 : F = \hat{F}$ for all x, where $\hat{F}$ is a completely specified continuous df. The differences between F and $\hat{F}$ should be small for all x, except for sampling variation, if the null hypothesis is true. For the usual two-sided goodness-of-fit alternative $H_1 : F \ne \hat{F}$, for some x. Large absolute values of these deviations tend to discredit the hypothesis. All computations are achieve by Matlab package, where we have four functions $[H, P, KSSTAT, CV]$. Namely, H is equal to 0 or 1, P is the $p-$value, $KSSTAT$ is the maximum difference between the data and fitting curve and CV is a critical value. Therefore

- We accept H_0, if $H = 0$, $KSSTAT \le CV$ and $P >$ level of significant,
- We reject H_0, if $H = 1$, $KSSTAT > CV$ and $P \le$ level of significant.

Sub-sample bootstrap technique

Although the bootstrap has been widely used in many areas, the method has its limitation in extremes. It was shown in some cases that a full-sample bootstrap does not work for extremes. Namely, assume $X_j^*, j = 1,2,...,m,$ where $m = m(n) \to \infty,$ as $n \to \infty,$ are conditionally iid rv's with

$$P(X_1^* = X_j \mid \underset{-}{X}_n) = \frac{1}{n}, \qquad j = 1,2,...,n,$$

where $\underline{X}_n=(X_1,X_2,...,X_n)$ is a random sample of size n from the unknown df F. Hence $X_1^*,...,X^*$ is a re-sample of size m from the empirical df $F_n(x)=\frac{1}{n}\sum_{i=1}^{n}\#\{X_i\le x\}$. Furthermore, let

$$H_{n,m}(x)=P(\frac{M_m-b_m}{a_m}\le x\mid \underline{X}_n)=F_n^m(a_mx+b_m)$$

be the bootstrap df of $\frac{M_m-b_m}{a_m}$. A full-sample bootstrap is the case when $m=n$. In contrast, a sub-sample bootstrap is the case when $m<n$. If the df of BM converges to the limit G_γ, which is defined in (2.1), Athreya and Fukuchi (1997) showed that the bootstrap df $H_{n,m}$ is weakly consistent estimate for G_γ, if $m=\circ(n)$ and it is strongly consistent, if $m=o(\frac{n}{\log n})$. Otherwise, $H_{n,m}$ fails to approximate G_γ. For the maximum order statistics under power normalization, this result is extended by Nigm (2006). More recently, Barakat et al. (2011) extended the same result to the generalized order statistics. Actually, this result suggests an efficiency estimate for the GEVD by using the BM method, even if the data do not consist of blocks (in this case the bootstrap replicates of size m, from F_n, are treated as blocks). For applying the suggested technique, we have to choose a suitable value of m (i.e., the size of bootstrap replicates or the blocks size). Actually, the suitable choice of the value m is the cornerstone of this technique. However, this value should be small enough to satisfy the stipulation $m=o(\frac{n}{\log n})$ and in the same time should be large enough to satisfy the stipulation $m\to\infty$, as $n\to\infty$. To determine a suitable value of m, we first simulate data with the same size as the realistic data, from the known GEVD $G_{\gamma_0}(.;\mu_0,\sigma_0)$. Then put $\frac{n}{\log n}$ in the form $a(10)^b+c$, where a,b and c are integers such that $1\le a<10$, $0\le c\le(10)^{b-1}$. Thus in view of the above two stipulations, we can take $m\approx\hat{m}=a(10)^{b-1}$. Consequently, to choose such suitable value of $\hat{m}$, we select a value from an appropriate discrete neighborhood of $\hat{m}$ (see Example 2.1) that gives the best estimate $\hat{\gamma}_0$ for the shape parameter γ_0. The estimate $\hat{\gamma}_0$ is obtained by withdrawing, from each of the originals samples, a large number of bootstrap replicates (each of size m) and determined the corresponding maxima. Then, we used these maxima, as a sample drawn from the parametric G_{γ_0}, to estimate the shape parameter γ_0, by using the ML method.

Example 2.1. Suppose we have $n=20000$, then $a=2,b=3$ and $c=19.490588$. Consequently, $\hat{m}=200$. In this case we can select a suitable value of m from the discrete neighborhood $\{100,150,200,250,300\}$ that gives the best estimate $\hat{\gamma}_0$ comparing the other values in the neighborhood, provided that this value does not equal 100 or 300. Otherwise, we should enlarge this neighborhood.

Data Treatments And Simulation Study

This section aims to answer the three questions. The first question is: Did the bootstrap improve the estimation of the parameters of the extreme models? The second question is: How can we choose a suitable POT number for every pollutant? The third equation is: How can we to choose the sub-sample m?

To answer the first question, we use the observed maxima values over 365 blocks (daily maximum through one year) for each pollutant and estimate the shape, scale and location parameters of G_γ in (2.1) (see, Table 1). Applying the full-bootstrap 50000 times for the data and again estimate the same parameters for each pollutant (see, Table 2). For fitting the real data, concerning $SO_2, PM10$ and O_3, we use the K-S test and calculate its functions $H, P, KSSTAT$ and CV, with and without bootstrap (see, Table 3). In the case of without bootstrap Table 3 shows that, we have not goodness of fit for SO_2 and $PM10$ in Zagazig and 10^{th} of Ramadan cities, respectively, where $H=1$, $KSSTAT>CV$ and $P\le$ level of significant. On the other hand, in the case of with bootstrap we have goodness fit for the both pollutants in the two cities. Moreover, the maximum distances between fitting curve and the data ($KSSTAT$) in the case of with bootstrap are less than those distances in the case of without bootstrap, see Figures 1-5 (Figures 1-5 compare between the empirical GEVD and $G_{\gamma_0}(.;\mu_0,\sigma_0)$ curves, for all pollutants after bootstrap). Therefore, the bootstrap works to improve the parameters estimation.

To answer the second question, we generate 2000 random samples, each of them has the same size n (say) as the realistic data of the pollutant under consideration, from the GPD $W^{C}_{\gamma_0}(.;\sigma^{*}_{0})$, see Table 4a and 4b (as we have shown previously in Section 2). Note that the size of the generated samples actually is less than $365\times 24=8760$, for SO_2 and $MP10$, or $365\times 48=17520$, for O_3, this is due to the inactivation and maintenance of the monitor devices in some hours at some days. In view of the imposed stipulations on the threshold u (and consequently on the number of POT k) in Section 2, we vary the number of POT k over the values $[\frac{n}{20}],[\frac{n}{19}],...,[\frac{n}{4}]$, where $[\theta]$ is the integer part of θ, see Table 4. Actually, we only wrote 7 values of k in Table 4a and 4b, including $[\frac{n}{20}]$ and the best value. Then, we look for the value of k (or u), which gives the best estimate $\hat{\gamma}_0$ of the shape parameter (its true value γ_0 is known), where the estimate $\hat{\gamma}_0$ here is the mean value of 2000 estimates, which are calculated as we have shown in Section 2. When two values of k give the same best mean estimate, we favor between them by the coefficient of variations (C.V). For example, in the case of SO_2, in 10^{th} of Ramadan in Table 4a and 4b, we see that the values $k=2047$ and $k=2132$ give the same best estimate $\hat{\gamma}_0=0.0987$ (the true value is $\gamma_0=0.1$). Since, the second value corresponds the C.V=1.389, which is less than the C.V=1.4044 concerning the first value, we then choose the second value, i.e., the suitable number of POT is $k^{\Sigma}=2132$. In this case, the corresponding threshold u is the upper quantile of order $[\lambda n]=[0.7500586n]=6397$ (note that $\lambda;\frac{n-k^{å}}{n}=\frac{8530-2132}{8530}$). Now, by using the determined suitable threshold values, from Table 4, we can apply the POT method on the realistic data for each pollutant to determine its extreme value model, see Table 6. Finally, apply the full bootstrap technique (50000 times) to improve the obtained estimates, see Table 7. To answer the third question, we generate 2000 random samples, each of them has the same size n as the realistic data of the pollutant under consideration, from the GEVD $G_{\gamma_0}(.;\mu_0,\sigma_0)$, see Table 5. Determine, for each pollutant the value $\hat{m}=a(10)^{b-1}$ (as we have shown in Section 2). We can see that $\hat{m}=90$, for the SO_2 and $PM10$, i.e., for the first four rows of Table 5, while $\hat{m}=170$, for the O_3, i.e., for the last row of Table 5. Thus, for the first four rows, by checking the discrete neighborhood $\{60,70,80,90,100,110,120\}$, we find that the best value of m (according to the given method in Section 2) is the lower value 60. Thus, we consider a new discrete neighborhood, $\{20,30,50,60\}$, which yields the value $m=30$. In Similar way, for the last row of Table 5, we checked the the discrete neighborhoods $\{110,130,150,170,190,210\}$, $\{60,70,80,90,100,110\}$ and $\{20,30,50,60\}$. The last neighborhood gives the value $m=30$. Therefore, for all pollutants the value 30 is more suitable value of m. Take this value and apply the sub-sample bootstrap technique on the realistic data to get a more suitable extreme value models for these pollutants (as we have shown in Section 2), see Table 8.

Table 1: Zagazig and 10th of Ramadan for GEVD

ML parameters estimation									
	SO_2			PM_{10}			O_3		
	γ_0	μ_0	σ_0	γ_0	μ_0	σ_0	γ_0	μ_0	σ_0
Zagazig	0.16	21.9	11.72	0.099	196.78	66.01			
10^{th} of Ramadan	0.11	81.24	39.49	0.22	249.75	67	-0.087	54.9	9.6

Table 2: Zagazig and 10th of Ramadan for GEVD, after bootstrap

ML parameters estimation									
	SO_2			PM_{10}			O_3		
	γ_0	μ_0	σ_0	γ_0	μ_0	σ_0	γ_0	μ_0	σ_0
Zagazig	0.15	21.6	11.6	0.094	197	67.5			
10^{th} of Ramadan	0.1	81.3	39.4	0.21	249.8	65.9	-0.1	54.98	9.5

Table 3: Kolmogorov-Smirnov test for the data with and without bootstrap

Data of SO_2 in Zagazig					
	H	P	$KSSTAT$	CV	Decision
without bootstrap	1	0.0446	0.0656	0.0644	reject the null hypothesis
with bootstrap	0	0.0709	0.0605	0.0644	accept the null hypothesis
Data of SO_2 in 10^{th} of Ramadan					
	H	P	$KSSTAT$	CV	Decision
without bootstrap	0	0.2962	0.0507	0.0706	accept the null hypothesis
with bootstrap	0	0.3065	0.0502	0.0706	accept the null hypothesis
Data of $PM10$ in Zagazig					
	H	P	$KSSTAT$	CV	Decision
without bootstrap	0	0.4389	0.0450	0.0706	accept the null hypothesis
with bootstrap	0	0.4614	0.0442	0.0706	accept the null hypothesis
Data of $PM10$ in 10^{th} of Ramadan					
	H	P	$KSSTAT$	CV	Decision
without bootstrap	1	0.0305	0.0752	0.0706	reject the null hypothesis
with bootstrap	0	0.0548	0.0697	0.0706	accept the null hypothesis
Data of O_3 in 10^{th} of Ramadan					
	H	P	$KSSTAT$	CV	Decision
without bootstrap	0	0.1845	0.0565	0.0707	accept the null hypothesis
with bootstrap	0	0.2537	0.0528	0.0707	accept the null hypothesis

Table 4: Simulation study for choosing a suitable number of POT (k). Note that k*is the best value

SO_2 in Zagazig: GPD with $\gamma_0 = 0.15, \quad \sigma_0^* = 8.48, \quad c = 0.226, \quad n = 8633$							
k	431	1033	1549	1721	1979	2056*	2151
$\hat{\gamma}_0$	0.144	0.1504	0.1506	0.1505	0.1504	0.1502	0.1505
C.V	0.624	0.538	0.738	0.565	0.544	0.4144	0.336
$\hat{\sigma}_0^*$	13.45	11.67	10.99	10.8	10.59	10.5	10.45
SO_2 in $10^t h$ of Ramadan: GPD with $\gamma_0 = 0.1, \quad \sigma_0^* = 31.5, \quad c = 2.5, \quad n = 8530$							
k	432	1027	1549	1707	1962	2047	2132*
$\hat{\gamma}_0$	0.0934	0.098	0.0982	0.0985	0.0986	0.0987	0.0987
C.V	4.69	2.322	2.708	1.585	1.4156	1.4044	1.389
$\hat{\sigma}_0^*$	42.7	38.68	37.67	37.02	36.5	36.35	36.21
$PM10$ in Zagazig: GPD with $\gamma_0 = 0.094, \quad \sigma_0^* = 49.2, \quad c = 2, \quad n = 8540$							
k	460	970	1480	1735	1990	2075	2160*
$\hat{\gamma}_0$	0.0857	0.0891	0.0901	0.0911	0.0913	0.0914	0.0914
C.V	4.94	3.84	4.12	3.77	3.3	2.95	2.93

$\hat{\sigma}_0^*$	65.22	60.66	58.63	57.36	56.6	56.39	56.187
$PM10$ in $10^t h$ of Ramadan: GPD with $\gamma_0 = 0.21, \quad \sigma_0^* = 14.8, \quad c = 3.6, \quad n = 8720$							
k	440	962	1484	1745	2006^*	2093	2180
$\hat{\gamma}_0$	0.2047	0.2092	0.2092	0.2097	0.2098	0.2096	0.2097
C.V	1.33	0.5372	0.4247	0.3832	0.3727	0.3736	0.3239
$\hat{\sigma}_0^*$	27.92	23.52	21.48	20.74	20.33	20.14	19.8
O_3 : GPD with $\gamma_0 = -0.1, \quad \sigma_0^* = 14.25, \quad c = 7.46, \quad n = 17000$							
k	850	2040	3060	3400	3910	4080	4250^*
$\hat{\gamma}_0$	- 0.1053	-0.1026	-0.102	-0.1018	-0.1018	-0.1018	-0.1017
C.V	0.68	0.52	0.36	0.23	0.2003	0.2333	0.2427
$\hat{\sigma}_0^*$	10.6	11.56	12.03	12.266	12.32	12.38	12.43

Table 5: Simulation study for chosen m sub-sample bootstrap. Note that m^* is the best value

SO_2 in Zagazig: GEVD with $\gamma_0 = 0.15, \quad \sigma_0 = 11.69, \quad \mu_0 = 21.6, \quad n = 8633$		
m	$\hat{\gamma}_0$	$C.V$
20	0.147	0.352
30^*	0.152	0.374
50	0.1402	0.421
60	0.1355	0.507
SO_2 in 10^{th} of Ramadan: GEVD with $\gamma_0 = 0.1, \quad \sigma_0 = 39.4, \quad \mu_0 = 81.3, \quad n = 8530$		
m	$\hat{\gamma}_0$	$C.V$
20	0.0844	0.742
30^*	0.0994	0.517
50	0.0925	0.622
60	0.087	0.76
$MP10$ in Zagazig: GEVD with $\gamma_0 = 0.094, \quad \sigma_0 = 67.5, \quad \mu_0 = 197, \quad n = 8640$		
m	$\hat{\gamma}_0$	$C.V$
20	0.0854	0.5911
30^*	0.0987	0.7977
50	0.0782	1.741
60	0.074	0.941
$MP10$ in 10^{th} of Ramadan: GEVD with $\gamma_0 = 0.21, \quad \sigma_0 = 65.9, \quad \mu_0 = 249.8, \quad n = 8720$		
m	$\hat{\gamma}_0$	$C.V$
20	0.2017	0.2987
30^*	0.2064	0.2890
50	0.1906	0.3552
60	0.1909	0.3652
O_3 : GEVD with $\gamma_0 = -0.1, \quad \sigma_0 = 9.5, \quad \mu_0 = 54.98, \quad n = 17000$		
m	$\hat{\gamma}_0$	$C.V$
20	-0.1122	0.4165
30^*	-0.1077	0.4033
50	-0.1168	0.3807
60	-0.1178	0.4212

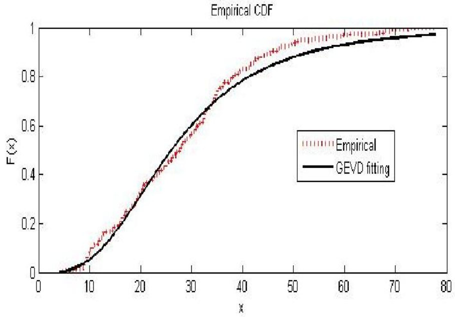

Figure 1: SO2 in Zagazig

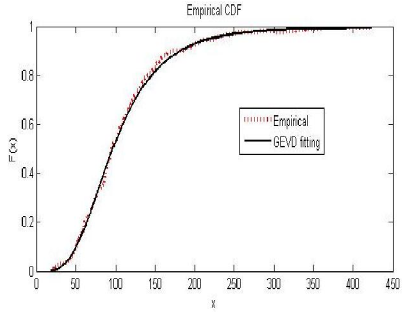

Figure 2: SO2 in 10th of Ramadan

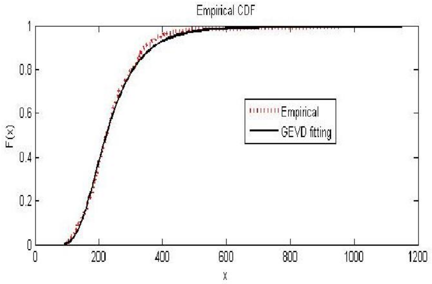

Figure 3: PM10 in Zagazig

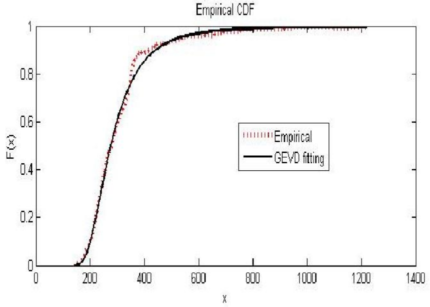

Figure 4: PM10 in 10th of Ramadan

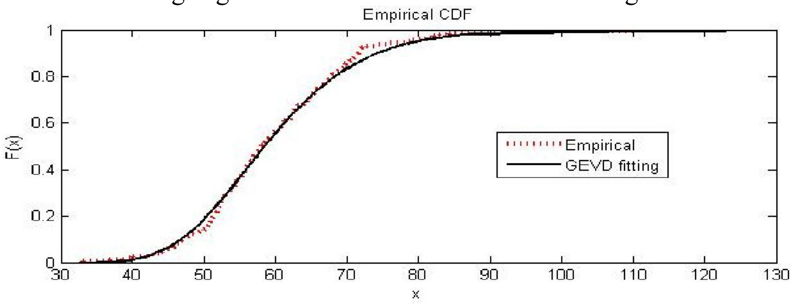

Figure 5: O3 in 10th of Ramadan after bootstrap

Table 6: Zagazig and 10th of Ramadan for GPD

ML parameters estimation						
	SO_2		PM_{10}		O_3	
	γ	σ	γ	σ	γ	σ
Zagazig	0.164	7.16	0.047	57.64		
10^{th} of Ramadan	0.046	33.44	0.13	68.27	-.08	38.8

Table 7: Zagazig and 10^{th} of Ramadan for GPD after bootstrap

ML parameters estimation						
	SO_2		PM_{10}		O_3	
	γ	σ	γ	σ	γ	σ
Zagazig	0.157	7.13	0.052	57.3		
10^{th} of Ramadan	0.062	32.4	0.14	67.9	-0.087	8.89

Table 8: Zagazig and 10th of Ramadan for GEVD

ML parameters estimation by sub-sample						
SO_2						
	γ	$C.V$	μ	$C.V$	σ	$C.V$
Zagazig	0.176	0.253	26.39	0.0134	7.34	0.0463
10^{th} of Ramadan	0.119	0.258	108.9	0.187	32.02	0.0489
$PM10$						
	γ	$C.V$	μ	$C.V$	σ	$C.V$
Zagazig	0.117	0.3728	264.41	0.0121	55.05	0.043
10^{th} of Ramadan	0.26	0.17	340.67	0.0124	70.587	0.088
O_3						
	γ	$C.V$	μ	$C.V$	σ	$C.V$
10^{th} of Ramadan	-0.08	0.739	64.36	0.0056	6.8	0.044

Corresponding author
H. M. Barakat[1]
Department of Mathematics Faculty of Science
Zagazig University, Zagazig, Egypt
hbarakat2@hotmail.com
s_nigm@yahoo.com
osam87@yahoo.com

References

1. Athreya, K.B and Fukuchi, j. (1997). Confidence interval for end point of a c.d.f, via bootstrap, J. Statist. Plann. Inference. 58, 299-320.
2. Barakat. H. M., Nigm, E. M., Ramadan. A. A., Khaled. O. M. (2011). a study of the air population by extreme value models, J. Appl. Statist. Sci. Vol. 18, Issue
3. Barakat, H. M., Nigm, E. M. and El-Adll, M. E. (to appear, 2011). Bootstrapping generalized extreme order statistics. Arabian Journal for Science and Engineering (AJSE). The paper is also has been presented in The Pyrenees International Workshop and summer School on Statistics, Probability and Operations Research. SPO 2009, September 15-18, 2009, Jaca-Huesca-Spain.
4. Coles, S.G. (2001). An introduction to statistical modeling of extreme values. Springer- Verlag, London
5. Cox, D. R. and Hinkley , D. V. (1974). Theoretical statistics. Chapman and Hall, London
6. Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. Ann. Stat. Vol 7, No. 1, 1-26.
7. Goldberg, M. S., Burnett, R. T., Brook, J., Bailor, J. C., Valois, M. F., and Vincent. R. (2001). Associations between daily cause-specific mortality and concentrations of ground-level ozone in Montreal, Quebec. American J. of Epidemiology, 154: 817
8. Gnedenko, B. V. (1943). Sur la distribution limite du terme maximum d'une s_erie al eatoire. Ann. Math. 44, 423-
9. Kim, S. Y. Lee, J. T., Hong, Y. C., Ahn, K. j. and Kim, H.(2004). Determining the threshold effect of Ozone on daily mortality: an analysis of Ozone and mortality in Seoul, Korea, Environmental, 1995 – 1999.
10. Pickands, J. (1975). Statistical inference using extreme order statistics. Ann. Statist. 3, 119-131.
11. Nigm, E. M. (2006). Bootstrapping extremes of random variables under power normalization. Test, Vol. 15, No. 1, 257-269.
12. Reiss, R. D. and Thomas, M. (2003). Statistical analysis of extreme values from insurance, finance, Hydrology and other fields. Berlen: Birkhäuser Verlag
13. Smith, R.L. (1985). Maximum likelihood estimation in a class of nonregular cases. Biometrika 72, 67-90.

12/30/2011