Kuwait J. Sci. **41** (1) pp. 1-19, 2014

Statistical modeling of extremes under linear and power normalizations with applications to air pollution

H. M. BARAKAT*, E. M. NIGM* AND O. M. KHALED**

**Department of Mathematics, Faculty of Science, Zagazig University, Zagazig-EGYPT, corresponding author e-mail: s_nigm@yahoo.com*

***Department of Mathematics, Faculty of Science, Port Said University - Port Said - EGYPT*

ABSTRACT

In this paper the Block Maxima (BM) and the Peak Over Threshold (POT) 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 (K-S) test. A new efficiency approach for modeling extreme values is suggested. This approach can convert any ordered data to enlarged block data by using sub-sample bootstrap. By using power normaliziation, for the first time in litrarure, the BM and sub-sample bootstrap methods are applied to model the air pollution. Although, this study is applied on three pollutants in two cities in Egypt, the suggested approaches may be applied on other pollutants in other regions in any country.

Keywords: Air pollution; bootstrap technique; generalized extreme value model; generalized Pareto distribution; Kolmogorov-Smirnov test.

AMS Subject Classification: 60G70; 62G09; 66P12.

INTRODUCTION

The traditional method of analyzing extreme values is based on the extreme value limiting distributions, which were derived by Gnedenko (1943) (Reiss & 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 BM and the POT methods 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 of the block is a year or a day and the observed quantities may be some environmental quantities 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., any change of the number of blocks 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 blocks number tends to infinity) to one of EVD. (ii) The distributions belonging to the GPD are the only stable ones, i.e., 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; Kim *et al.,* 2004). We choose in this study three pollutants: Sulphur Dioxide SO_2, Ozone O_3 and Particulate Matter $PM10$ in 10^{th} of Ramadan and Zagazig cities. The study of the Ozone pollutant was restricted to 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 during the twenty-four hours through out the year 2009 for the two cities, except Ozone was recorded every half hours. The detailed description of these pollutants and the collected data can be found in (Barakat *et al.,* 2010b). This study considered the BM and POT methods, which were used to evaluate the measurement of 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 K-S 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 & Fukuchi (1997), can convert any ordered data to enlarged block data by using sub-sample bootstraps. This method enables the engineers to analyse the rare events to construct dams for rivers, breakwater for sea defence and to design nuclear power plants against earthquakes, where the

number of available maxima about the relevant phenomena of these activities are often limited.

MATHEMATICAL MODELS

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

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

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 $\gamma > 0, \quad \gamma < 0$ 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 use the maximum likelihood method (ML) and improve 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 classical bootstrap approach uses Monte Carlo simulation to generate an empirical estimate for the sampling distribution of a given statistic by randomly drawing a large number of samples of the same size from the data. Therefore, the bootstrap is a way of finding the sampling distribution 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 use this method to get the preliminary result, to help simulate data with the same nature as the real data.

The POT method initiated by Pickand (1975) is an alternative approach to determine the type of asymptotic distribution for extremes. This approach is based on the concept of GPD and it is used to model data arising as independent threshold exceedances. Actually, the POT method is based on the fact that the conditional df $F^{[u]}(x+u) = P(X \le x+u | X > u)$ 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 $W_\gamma(x;\bar{\sigma}) = 1 - (1+\gamma x\bar{\sigma})^{-\frac{1}{\gamma}}$, where $\bar{\sigma} = \sigma - \gamma\mu$ and it is assumed that the df of BM weakly converges to $G_\gamma(x;\mu,\sigma)$ (Reiss & Thomas, 2003). In this case, we have $W_\gamma(x;\sigma) = 1 + \log G_\gamma(x;0,\sigma)$, $\log G_\gamma(x;0,\sigma) > -1$, and the left truncated GPD yields again a GPD, i.e.,

$$W_\gamma^{[c]}(x;\sigma^*) = W_\gamma(x;\bar{\sigma}), \quad \text{where} \quad \sigma^* = \bar{\sigma} + \gamma c. \tag{2}$$

The GPD family 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. Actually, the threshold should be high enough to justify the assumptions of the model but low enough to 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 use the simulation technique to choose a suitable threshold value. Namely, let γ_0, σ_0 and μ_0 be the preliminary estimates of the parameters γ, σ and μ, respectively (which are obtained by the BM method). Now, simulated data with the same size n as the realistic 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 the realistic data (this choice of c guarantees that the simulated and realistic data have nearly the same range) and $\sigma_0^* = \sigma_0 + \gamma_0(c - \mu_0)$. In view of the POT stability property of GPD, the simulated data will have the same nature of the realistic 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.

Remark 1 (the validity of approximation by iid variables). Although, the assumption that our variables are iid rv's is rarely correct in practice, we have many dependent models such as $m-$dependence and mixing models where the asymptotic results remain the same as for iid variables (Galambos, 1987). To be more specific, let X_j be the concentration of a given pollutant in the j th time interval (in our study, hour or half hour). It is reasonable to assume that the X_j are identically distributed but successive X_j values are dependent. However, the dependence weakens as the time passes. As a first approximation, $m-$dependence model is reasonable. More cautious researchers would incline toward mixing model (Galambos, 1987). In any case, the approximation by iid variables is reasonable, if asymptotic EVD are of interest.

GENERALIZED EXTREME VALUE DISTRIBUTION UNDER POWER NORMALIZATION (GEVP)

During the last two decades, E. Pancheva and her collaborators used a wider class of normalizing mappings than the linear ones, to get a wider class of limit laws, which can be used in solving approximation problems. Another reason for using nonlinear normalization concerns the problem of refining the accuracy of approximation in the limit theorems using relatively non difficult monotone mappings in certain cases that can achieve a better rate of convergence (Barakat *et al.*, 2010a). Pancheva (1985) considered the power normalization $C_n(x) = d_n|x|^{c_n}S(x),\ d_n, c_n > 0,$where $S(x) = -1, 0, 1,$ if $x < 0, x = 0, x > 0,$ respectively. Pancheva (1985) determined all the possible limit types of H, for which

$$H_n(x) = P[M_n \leq C_n(x)] = F^n(d_n|x|^{c_n}S(x)) \to H(x), \quad as\, n \to \infty. \tag{3}$$

These limit types are usually called the power max stable df's (P-max stable df's). Mohan & Ravi (1992) showed that the P-max stable df's (six P-types of df's) attract more than linear stable df's. Therefore, using the power normalization, we get a wider class of limit df's, which can be used in solving approximation problems. As in the case of linear normalization, Nasri-Roudsari (1999) has summarized these types by the following von Mises type representations

$$H_{1;\gamma}(x) = \exp[-(1 + \gamma(\log \alpha x^{\beta}))^{-\frac{1}{\gamma}}], x > 0, 1 + \gamma(\log \alpha x^{\beta}) > 0, \tag{4}$$

and

$$H_{2;\gamma}(x) = \exp[-(1 - \gamma(\log(\alpha(-x)^{\beta})))^{-\frac{1}{\gamma}}], x < 0, 1 - \gamma(\log(\alpha(-x)^{\beta}) > 0. \tag{5}$$

Each of the families (4) and (5) is called generalized extreme value distribution under power normalization (GEVP). Moreover, both GEVP (4) and (5) satisfy the P-max stable property, i.e., for every n there exist power normalizing constants $c_n, d_n > 0$, for which we have $H^n_{i,\gamma}(C_n(x)) = H_{i,\gamma}(x),\ i = 1, 2$. Clearly, the two parametric models (4) and (5) enable us to apply the BM method under power normalization. For these models, the parametric approach to modeling extremes is based on the assumption that the data in hand form an iid sample from an exact GEVP(γ, α, β) df in (4) and (5). Christoph & Falk (1996) showed that the upper tail behaviour of F, might determine whether F belongs to the domain of attraction of $H_{1;\gamma}(x; \alpha, \beta)$ or of $H_{2;\gamma}(x; \alpha, \beta)$. In the first case, the right end point of F should be positive, while in the second it should be negative. Therefore, the modeling under power normalization by using the BM method can only be applied if all data of maximums have the same sign. More specifically, if all the maximum obsrvations are positive, such as the case of our study, we have to select the model (4) and if all these observations are negative we select the model (5). In this paper we apply the mathematical modeling of extreme under power normalization by using (4). To the best of the authors knowledge, till now, there is not any published work related to this topic.

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

$$l(\underline{x}; \mu, \sigma, \gamma) = -n \log \sigma + \sum_{i=1}^{n} \left(-[1 + \gamma(x_i - \mu\sigma)]^{-\frac{1}{\gamma}} - (1 + \frac{1}{\gamma}) \log [1 + \gamma(\frac{x_i - \mu}{\sigma})]\right), \quad (6)$$

provided $1 + \gamma(x_i - \mu)/\sigma > 0$, for each i, otherwise (6) is undefined. For the maximization of $l(\underline{x}; \mu, \sigma, \gamma)$, 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. Moreover, the log likelihood functions for GPD and the model (4) are respictively given by

$$l^*(\underline{x}, \bar{\sigma}, \gamma) = -n \log \bar{\sigma} - (1 + \frac{1}{\gamma}) \sum_{i=1}^{k} \log\ (1 + \frac{\gamma x_i}{\bar{\sigma}}), \quad (7)$$

where k is the number of POT, and

$$l^{**}(\underline{x}; \alpha, \beta, \gamma) = -n \log \beta - \sum_{i=1}^{n} x_i - (1 + \frac{1}{\gamma}) \sum_{i=1}^{n} \log\ (1 + \gamma(\log \alpha + \beta \log x_i))$$

$$- \sum_{i=1}^{n} [1 + \gamma(\log \alpha + \beta \log x_i)]^{-\frac{1}{\gamma}}. \quad (8)$$

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 $\gamma > -0.5$ (Smith, 1985). Cases with $\gamma \leq -0.5$ 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 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 K-S statistic quantifies a distance between the empirical df of the sample and the reference df. In this study, all computations are achieved by the 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 the null hypothesis H_0; if $H = 0$; $KSSTAT \leq CV$ and $P >$ level of significant. Otherwise, we reject H_0. Although the K-S test is not the most adequate to test quality of the fit in the tail, but this fact does not bother us whenever we use it for comparison purposes, since this test provides us a digital indication (i.e., $KSSTAT$).

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 $X_j^*, j = 1, 2, ..., m, \quad m = m(n) \to \infty$, as $n \to \infty$, are conditionally iid rv's with $P(X_1^* = X_j | \underline{X}_n) = \frac{1}{n}, j = 1, 2, ..., n$, where $\underline{X}_n = (X_1, X_2, ..., X_n)$ is a random sample of size n from the unknown df F. Hence $X_1^*, ..., X_m^*$ is a re-sample of size m from the empirical df $F_n(x) = \frac{1}{n}\sum_{i=1}^{n}(X_i)$, where $I_A(x)$ is the indicator function. Furthermore, let $H_{n,m} = P(M_m \leq a_m x + b_m | \underline{X}_n) = F_n^m(a_m x + b_m)$ and $H_{n,m}^* = P(M_m \leq C_m(x) | \underline{X}_n) = F_n^m(C_m(x))$, where $H_{n,m}$ and $H_{n,m}^*$ are called the bootstrap distributions of $(M_m - b_m)/a_m$ and $(M_m/d_m)^{1/c_m}$, respectively. 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; Athreya & 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, if $m = n$, $H_{n,m}$ has a random limit and thus the naive bootstrap fails to approximate H_n. In other words the naive bootstrap of

the maximum order statistics, when $m = n$, fails to be consistent estimator for the limit df 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. 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. 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 $1 \leq a < 10$, $0 \leq c \leq (10)^{b-1}$. 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 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 the simulated data, a large number of bootstrap replicates (each of size m) and determine the maximum of each of them. Then, we used these maxima, as a sample drawn from the df G_{γ_0} to estimate the shape parameter γ_0 by using the ML method.

Example 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 question is: How can we to choose the sub-sample m?

To answer the first question, we use the observed maximum values over 365 blocks (daily maximum through one year) for each pollutant and estimate the shape, scale and location parameters of G_γ in (1) (Table 1). Applying the full-bootstrap 50000 times for the data (maximum values) and again estimate the

same parameters for each pollutant. The after bootstrap estimate of each parameter is the arithmatic mean of the obtained 50000 estimates of this parameter (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 (Table 3). In the case of without bootstrap, Table 3 shows that we have no goodness of fit for SO_2 and $PM10$ in Zagazig and 10^{th} of Ramadan cities, respectively. 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 (Figures 1-5). Figures 1-5 compare between the empirical GEVD and $G_{\gamma_0}(.;\mu_0,\sigma_0)$ curves (y-axis), for all pollutants after bootstrap (x-axis). Noting that, since in engineering desisgn, only the right tail is relevent, in Figures 1-5 we include only the right tails which are nearly larger than 0.9. 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 $W_{\gamma_0}^{[c]}(.;\sigma_0^*)$ (as we have shown in Section 2, Table 4a and 4b). It is noted 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, which is due to the inactivation and maintenance of the monitoring 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,(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 $k^* = 2132$. In this case, the corresponding threshold u is the upper quantile of order $[\lambda n] = [0.7500586n] = 6397$ (note that $\lambda \approx n - k^* n = \frac{8530 - 2132}{8530}$). Now, by using the determined suitable threshold values, from Tables 4a and 4b, we can apply the POT method on the realistic

data for each pollutant to determine its extreme value model, (Table 6). Finally, apply the full bootstrap technique (50000 times) to improve the obtained estimates, (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 $G_{\gamma_0}(.;\mu_0,\sigma_0)$, (Table 5). Determine, for each pollutant the value $\hat{m} = a(10)^{b-1}$ (see 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 suggested 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, (Table 8).

Table 1. Zagazig and 10^{th} of Ramadan for GEVD

City	ML parameters estimation								
	SO_2			$PM10$			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	0.106	81.24	39.49	0.22	249.75	67	-0.087	54.9	9.6

Table 2. Zagazig and 10^{th} of Ramadan for GEVD, after bootstrap

City	ML parameters estimation								
	SO_2			$PM10$			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. K-S 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 4a. Simulation study for choosing a suitable number of POT (k). Note that k^* is the best value

SO_2 in Zagazig: GPD with $\gamma_0 = 0.15$, $\sigma_0^* = 8.48$, $c = 0.226$, $n = 8633$							
k	431	1033	1549	1721	1979	2056^{Σ}	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^{th} of Ramadan: GPD with $\gamma_0 = 0.1$, $\sigma_0^* = 31.5$, $c = 2.5$, $n = 8530$							
k	432	1027	1549	1707	1962	2047	2132^{Σ}
$\hat{\gamma}_0$	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

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

***PM*10 in Zagazig: GPD with $\gamma_0 = 0.094$, $\sigma_0^* = 49.2$, $c = 2$, $n = 8540$**							
k	460	970	1480	1735	1990	2075	2160^{Σ}
$\hat{\gamma}_0$	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
***PM*10 in 10^{th} of Ramadan: GPD with$\gamma_0 = 0.21, \sigma_0^* = 14.8, c = 3.6$, $n = 8720$**							
k	440	962	1484	1745	2006^{Σ}	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 $\gamma_0 = -0.1$, $\sigma_0^* = 14.25$, $c = 7.46$, $n = 17000$**							
k	850	2040	3060	3400	3910	4080	4250^{Σ}
$\hat{\gamma}_0$	- 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 choosing m sub-sample bootstrap.
Note that m^* is the best value

***SO*$_2$ in Zagazig: GEVD with $\gamma_0 = 0.15, \sigma_0 = 11.69, \mu_0 = 21.6, n = 8633$**		
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 $\gamma_0 = 0.1, \sigma_0 = 39.4, \mu_0 = 81.3$, $n = 8530$		
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
***PM*10 in Zagazig: GEVD with $\gamma_0 = 0.094$, $\sigma_0 = 67.5$, $\mu_0 = 197$, $n = 8640$**		
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

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

$PM10$ in 10^{th} of Ramadan: GEVD with $\gamma_0 = 0.21$, $\sigma_0 = 65.9$, $\mu_0 = 249.8$, $n = 8720$		
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 $\gamma_0 = -0.1$, $\sigma_0 = 9.5$, $\mu_0 = 54.98$, $n = 17000$		
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

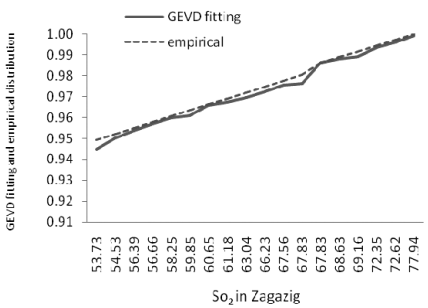

Fig. 1. SO_2, in Zagazig

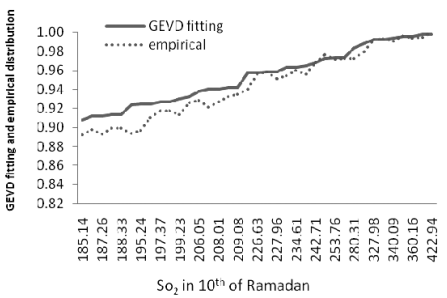

Fig. 2. SO_2, in 10^{th} of Ramadan

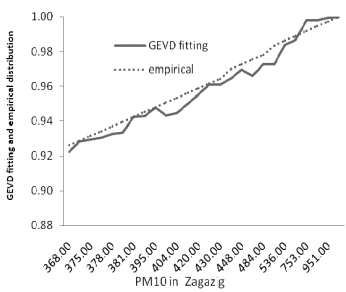

Fig. 3. $PM10$ in Zagazig

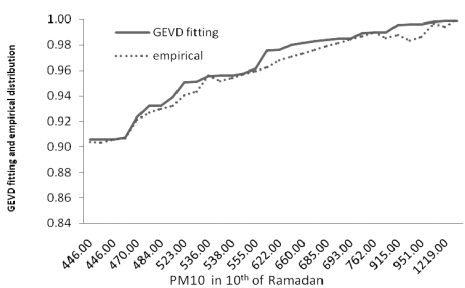

Fig. 4. $PM10$ in 10^{th} of Ramadan

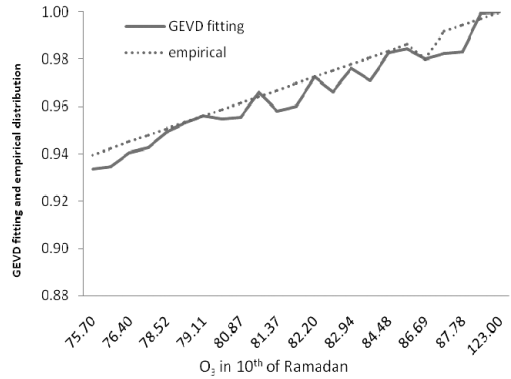

Fig. 5. O_3 in 10^{th} of Ramadan after bootstrap

Table 6. Zagazig and 10^{th} of Ramadan for GPD

City	ML parameters estimation					
	SO_2		PM10		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	-0.08	38.8

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

City	ML parameters estimation					
	SO_2		PM10		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 10^{th} of Ramadan for GEVD

City	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

Cont. Table 8. Zagazig and 10^{th} of Ramadan for GEVD

City	ML parameters estimation by sub-sample					
	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

Remark 2. We note that all data of the considered pollutants with the exception of Ozone lead to positive values of the shape parameter γ, which implies a Frechet domain of attraction, that is, unlimited range. This discrimination is due to the fact that the three pollutants differ radically in their chemical and physical properties so we do not expect to obtain the same models for them. Probably, obtaining negative values of the estimated shape parameter for the Ozone is due to that the monitored data were fallen in narrow range comparing with the other pollutants. On the other hand, we can find many works in which the shape parameter γ has negative value for the Ozone, e.g., Example 5.1.2, in Reiss & Thomas (2003). Even more, the difference in any pollutant's place can lead to a sharp change in its estimated shape parameter, where the shape parameter's sign may be changed. For example the different values -0.06; $+0.06$; 0.00; 0.12 and 0.11 were obtained as the estimates for the parameter γ for the Ozone for the years 1983-1987 for different five monitoring stations in San Francisco, c.f. Example 12.3.2, in Reiss & Thomas (2003).

APPLICATION OF BM AND SUB-SAMPLE BOOTSTRAP METHODS FOR GEVP TO AIR POLLUTION

We use the same technique of the preceeding subsection for applying the BM and sub-sample bootstrap methods for GEVP. Table 9 gives the ML parameters estimation of model (4). Applying the full-bootstrap 50000 times for the data, we again estimate the same parameters for each pollutant (Table 10). For fitting the real data, concerning $SO_2, PM10$ we use the K-S test and calculate its functions $H, P, KSSTAT$ and CV, with and without bootstrap (Table 11). In the case of without bootstrap, Table 11 shows that, we have no goodness of fit for $PM10$ in 10^{th} of Ramadan citiy. On the other hand, in the case of with bootstrap we have goodness fit for both the 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'. Note that the model (4) fail to fitting O_3, in 10^{th} of Ramadan citiy. Finally, Table 12 presents the estimate parameter by using sub-sample bootstrap as in Example 1.

Table 9. Zagazig and 10^{th} of Ramadan for GEDP

City	SO_2			$PM10$		
	γ	α	β	γ	α	β
Zagazig	-0.415	$4.6 * 10^{-3}$	1.759	-0.1448	$2.29 * 10^{-7}$	2.9
$10^{th}$of Ramadan	-0.277	$2.8 * 10^{-4}$	1.864	-0.0253	$1.59 * 10^{-9}$	3.67

Table 10. Zagazig and 10^{th} of Ramadan for GEVP, after bootstrap

City	SO_2			$PM10$		
	γ	α	β	γ	α	β
Zagazig	-0.418	$4.5 * 10^{-3}$	1.764	-0.152	$2.23 * 10^{-7}$	2.906
$10^{th}$of Ramadan	-0.277	$2.8 * 10^{-4}$	1.873	-0.0248	$1.69 * 10^{-9}$	3.68

Table 11. Kolmogorov-Smirnov test for the data with and without bootstrap

Data of SO_2 in Zagazig					
	H	P	$KSSTAT$	CV	Decision
Without bootstrap	0	0.2861	0.0435	0.0644	accept the null hypothesis
With bootstrap	0	0.2687	0.0425	0.0644	accept the null hypothesis
Data of SO_2 in 10^{th} of Ramadan					
	H	P	KSSTAT	CV	Decision
Without bootstrap	0	0.1030	0.0553	0.0636	accept the null hypothesis
with bootstrap	0	0.1130	0.0544	0.0636	accept the null hypothesis
Data of $PM10$ in Zagazig					
	H	P	KSSTAT	CV	Decision
Without bootstrap	0	0.2211	0.0450	0.0636	accept the null hypothesis
with bootstrap	0	0.2727	0.0417	0.0636	accept the null hypothesis
Data of $PM10$ in 10^{th} of Ramadan					
	H	P	KSSTAT	CV	Decision
Without bootstrap	1	0.0063	0.0828	0.0636	reject the null hypothesis
with bootstrap	0	0.0509	0.0634	0.0636	accept the null hypothesis

Table 12. Zagazig and $10^{th}$of Ramadan for GEVP

	ML parameters estimation by sub-sample					
	SO_2					
City	γ	$C.V$	α	$C.V$	β	$C.V$
Zagazig	-0.284	0.0035	$1.92 * 10^{-2}$	$1.2 * 10^{-3}$	1.6	$8.6 * 10^{-4}$
10^{th} of Ramadan	-0.185	$8.9 * 10^{-4}$	$5.9 * 10^{-6}$	$3.9 * 10^{-4}$	2.7 1	$2.12 * 10^{-4}$
	$PM10$					
	γ	C.V	α	$C.V$	β	$C.V$
Zagazig	-0.155	0.0021	$-1.46 * 10^{-8}$	$2.02 * 10^{-4}$	3.449	$2.57 * 10^{-4}$
10^{th} of Ramadan	-0.029	$1.0 * 10^{-3}$	$1.4 * 10^{-11}$	$1.27 * 10^{-4}$	4.339	$2.3 * 10^{-4}$

ACKNOWLEDGEMENT

The authors's work is supported by Zagazig University, jointly with National Center of Nuclear Safety and Radiation Control -Atomic Energy Authority-Egypt/2008-2009.

REFERENCES

Athreya, K. B. & Fukuchi, J. 1997. Confidence interval for end point of a c.d.f, via bootstrap, J. Statist. Plann. Inference **58**: 299-320.

Barakat, H. M., Nigm, E. M. & El-Adll, M. E. 2010a. Comparison between the rates of convergence of extremes under linear and under power normalization. Statistical Papers **51**(1): 149-164.

Barakat, H. M., Nigm, E. M., Ramadan. A. A. & Khaled, O. M. 2010b. A study of the air polluants by extreme value models, J. Appl. Statist. Sci. **18**(2): 199-209.

Barakat, H. M., Nigm, E. M. & El-Adll, M. E. 2011. Bootstrapping generalized extreme order statistics. Arab. J. Sci. Eng.(AJSE) **36**: 1083-1090.

Christoph, G. & Falk, M. 1996. A note on domains of attraction of p-max stable laws Statist. Probab. Lett. **28**: 279-284.

Coles, S. G. 2001. An introduction to statistical modeling of extreme values. Springer- Verlag, London.

Cox, D. R. & Hinkley, D. V. 1974. Theoretical statistics. Chapman and Hall, London.

Efron, B. 1979. Bootstrap methods: Another look at the Jackknife. Ann. Stat. **7**(1): 1-26.

Galambos, J. 1987. The asymptotic theory of extreme order statistics, Wiley, New York,

Goldberg, M. S., Burnett, R. T., Brook, J., Bailor, J. C., Valois, M. F. & Vincent. R. 2001. Associations between daily cause-specific mortality and concentrations of ground- level ozone in Montreal, Quebec. American J. of Epidemiology 154- 817.

Gnedenko, B.V. 1943. Sur la distribution limite du terme maximum d'une série aléatoire. Ann. Math. **44**: 423-453.

Kim, S. Y. Lee, J. T., Hong, Y. C., Ahn, K. J. & Kim, H. 2004. Determining the threshold effect of Ozone on daily mortality: an analysis of Ozone and mortality in Seoul, Korea, Environmental 1995-1999.

Pancheva, E. 1985. Limit theorems for extreme order statistics under nonlinear normalization. Lecture Notes in Mathematics **1155**: 284-309.

Pickands, J. 1975. Statistical inference using extreme order statistics. Ann. Statist. **3**: 119- 131.

Mohan, N.R. & Ravi, S. 1992. Max domains of attraction of univariate and multivariate p-max stable laws. Theory Probab. Appl. **37**: 632-643.

Nasri-Roudsari, D. 1999. Limit distributions of generalized order statistics under power normalization.Commun Stat Theory Methods **28**(6):1379--1389.

Nigm, E. M. 2006. Bootstrapping extremes of random variables under power normalization. Test **15**(1): 257-269.

Reiss, R. D. & Thomas, M. 2003. Statistical analysis of extreme values from insurance, finance, Hydrology and other fields. Berlen: Birkhäuser Verlag.

Smith, R. L. 1985 Maximum likelihood estimation in a class of nonregular cases. Biometrika **72**: 67-90.

Submitted : *11/12/2011*
Revised : *05/08/2013*
Accepted : *07/08/2013*

الإحصاءات المرتبة في دراسة النمذجة الإحصائية للقيم المتطرفة تحت ثوابت اتزان خطية وثوابت اتزان القوى مع التطبيق على تلوث الهواء

[*]هـ. م. بركات و [**]اي. م. نجم و أو. م. خالد

[*]قسم الرياضيات – كلية العلوم – جامعة الزقازيق – الزقازيق – مصر
[**]قسم الرياضيات – كلية العلوم – جامعة بورسعيد – بورسعيد – مصر

خلاصة

في هذا البحث تم استخدام طريقتي كتل القيم العظمى (BM) وحد عتبة الذروة (POT) كطريقتين للنمذجة الإحصائية لتلوث الهواء في مدينتين في جمهورية مصر العربية. وتم استخدام أسلوب المحاكاة لاختيار قيمة مناسبة للعتبة، كما تم تقييم أسلوب البوتستراب التام لتحسين معالم النماذج المقدرة عن طريق اختبار كولموغوروف – سميرنوف. أيضاً تم اتباع نهج جديد كفء لنمذجة القيم المتطرفة. وباستخدام هذا النهج يمكن تحويل أية بيانات مرتبة إلى بيانات موسعة على شكل كتل للقيم العظمى، وذلك باستخدام عينات البوتستراب الجزئية. وفي هذا البحث تم تطبيق طريقتي كتل القيم العظمى وعينات البوسترات الجزئية باستخدام ثوابت اتزان القوى لدراسة تلوث الهواء، علماً بأن ذلك الأسلوب لم يطبق من قبل. وأخيراً ننوه بأنه على الرغم من إجراء تلك الدراسة على ثلاثة ملوثات في مدينتين في مصر، إلا أن معالجة البيانات والنماذج المستخدمة في تلك الدراسة يمكن تطبيقها على ملوثات أخرى وفي مناطق ودول مختلفة.

المجلة العربية للعلوم الإدارية

Arab Journal of Administrative Sciences

- صدر العدد الأول في نوفمبر ١٩٩٣ .
- First issue, November 1993.
- علمية محكمة تعنى بنشر البحوث الأصيلة في مجال العلوم الإدارية.
- Refereed journal publishing original research in Administrative Sciences.
- تصدر عن مجلس النشر العلمي في جامعة الكويت ثلاثة إصدارات سنوياً (يناير - مايو - سبتمبر).
- Published by Academic Publication Council, Kuwait University, 3 issues a year (January, May, September).
- تسهم في تطوير الفكر الإداري واختبار الممارسات الإدارية وإثرائها.
- Contributes to developing and enriching administrative thinking and practices.
- مسجلة في قواعد البيانات العالمية.
- Listed in several international databases.
- تخضع للتقييم الأكاديمي الخارجي.
- Reviewed periodically by international referees for high academic standards.

الاشتراكات

الكويت: 3 دنانير للأفراد - 15 ديناراً للمؤسسات - الدول العربية: 4 دنانير للأفراد - 15 ديناراً للمؤسسات
الدول الأجنبية: 15 دولاراً للأفراد - 60 دولاراً للمؤسسات

توجه المراسلات إلى رئيس التحرير على العنوان الآتي:

المجلة العربية للعلوم الإدارية - جامعة الكويت ص.ب: 28558 الصفاة 13146 - دولة الكويت
هاتف: Tel: (965)24827317 - أو 4734 / 4416 /(965)24984415 - فاكس: Fax: (965)24817028
E-mail: ajas@ku.edu.kw - Web Site: http://www.pubcouncil.kuniv.edu.kw/ajas