



Graphical Diagnostics for Threshold Selection in Fitting the Generalized Pareto Distribution

H. A. Alaswed

Statistics Department, Faculty of Science, Sebha University, Libya

Keywords:

Akaike information test
Deviance test
Generalized Pareto Distribution
Pareto Distribution
Graphical diagnostics
Maximum likelihood estimation

ABSTRACT

When fitting the Generalized Pareto distribution (GPD), selecting an appropriate threshold value is important for achieving an effective fit. The main objective of this study is to give five graphical diagnostics for selecting GPD thresholds. The other objective of this study is to examine different graphical methods based on the goodness-of-fit test. Maximum likelihood method was used to estimate the shape of parameter. Finally, use flood data to compare five graphical diagnostics of threshold selection for shape parameter estimate. The results show that, the four graphical diagnostics (threshold choice plot, mean excesses plot, dispersion index plot and quantile quantile plot) yield the same threshold range, with the exception of the Hill plot. On the other hand, threshold choice plot is simple to identify the range of thresholds that should be stable to fit. When compared to other graphical diagnostic, the GP distribution becomes valid because they demand too much subjectivity and make it difficult to define a range threshold from the plots. In other words, graphical diagnostics of the higher are an acceptable option for fitting the GPD model based on the goodness-of-fit test. All statically analyses for the study are performed using R- statistical program.

الرسومات التشخيصية لاختيار العتبة لنموذج توزيع باريتو العام

حافظ ابوبكر الأسود

قسم الاحصاء ، كلية العلوم ، جامعة سبها ، ليبيا

الكلمات المفتاحية:

اختبار المعلومات للأكايكيو
اختبار نسبة الإمكان
الرسومات التشخيصية
تقدير معالم توزيع باريتو
توزيع باريتو
طريقة الامكان الأعظم

الملخص

عند نمذجة توزيع باريتو المعمم نحتاج إلى اختبار قيمة عتبة مناسبة لتحقيق الملاءمة المناسبة للنموذج. الهدف الرئيسي في هذه الدراسة هو تقديم خمس رسومات تشخيصية لاختيار عتبة مناسبة لتوزيع باريتو العام. الهدف الآخر من هذه الدراسة هو اختبار الطرق البيانية المختلفة بناء على اختبار جودة التوافق... تم استخدام طريقة الامكان الأعظم لتقدير معالم توزيع باريتو. أخيراً، تم التطبيق على البيانات الحقيقية لمقارنة طرق الرسومات البيانية لاختيار العتبة لتقدير معلمة الشكل. أظهرت النتائج أن الرسومات التشخيصية الأربعة (رسم اختيار العتبة ورسم المتوسطات المتجاوزة ورسم التجزئات) لها نفس نطاق العتبة بينما اختلفت طريقة (رسم هيل). نموذج باريتو مع الرسم البياني لطريقة (رسم اختيار العتبة) لاختيار العتبة يعد أفضل طرق الرسم في اختيار العتبة بناء على الاختبارين. من ناحية أخرى، من السهل الحصول على نطاق العتبة لمخطط (رسم اختيار العتبة) التي تكون مستقرة أو ثابتة فوق العتبة التي يصبح عندها نموذج باريتو هو التوزيع المناسب وذلك عند مقارنته بالرسومات التشخيصية الأخرى التي تتطلب الكثير من الخبرة الذاتية لتحديد مدي العتبة لهذه الرسومات. كما أظهرت النتائج أن نموذج باريتو مع الرسومات التشخيصية لاختيار العتبة العليا يعد اختيار مناسب بالاعتماد على الاختبارين. تم اجراء جميع الحسابات في هذا الورقة بواسطة برنامج آر لتقدير المعالم واجراء الرسومات التشخيصية والاختبارين.

*Corresponding author.

E-mail addresses: haf.alaswed@sebhau.edu.ly

Article History: Received 31 January 2024 - Received in revised form 12 April 2024 - Accepted 23 April 2024

1. Introduction

The probability of rare flood and drought events is an important in hydrology and other branches of high flow problem to estimate parameters of distribution, a review of flood frequency analysis is found in [1], see,[2] for floods, and [3,4] for droughts. For the original theoretical development Generalized Pareto Distribution (GPD), see [5,6] and for further developments and applications, see[7]. The method of Peaks Over Thresholds (POT) has been used in many fields, for example, applied to environmental data such as rainfall, sea levels and modeling high flows by GPD distribution, among others and a review is found in,[8,9]. The properties of the GP stated by [10,11] to make the distribution a logical candidate for the analysis of extreme events, and such practical problems are addressed in [12]. To fit the GPD, there is the problem of threshold choice. Several procedures for selection threshold value in the case of POT modeling are given in the literature (e.g., [13]) suggest the use of the Mean Excess Function (MEF). In, [14], outlined two diagnostics for the choice of threshold: Mean Excesses Plot (MEP) and Threshold Choice Plot (TCP). The shape and scale parameters of the GPD can be estimated by several methods such as the ML method, see,[13] and methods for estimating the GPD parameters have been reviewed by [10], and established the asymptotic normality and consistency of the MLE estimators. Various approaches of numerical methods have been suggested and applied by authors, for example, Square Error Method (SEM), by [15] and automatic choice using shape parameter, by [16] and Automated Threshold Selection Method (ATSM), by [17]. On the other hand, graphical method, by [18]. Various literature discusses the problem of threshold selection, and the best method is still to be found. The outline of the present study report as follows: In section 2, the background theoretical of GPD is described. Fitting the GPD is introduced in Section 3. While Section 4, model adequacy is provided. In section 5 is concerned with diagnostics plot with the problem of choosing the best value of threshold, five graphical diagnostics is introduced. Section 6, describes application on flood data. Our conclusions and future work are given in section 7.

2 Generalized Pareto Distribution ($GPD(\gamma, \sigma)$)

The 2-parameter generalized Pareto (GP) distribution with scale and the shape parameter (denoted, $GP(\sigma, \gamma)$) is the distribution of the $X = \sigma(1 - e^{-Y}) / \gamma$ where Y is a random variable with the standard exponential distribution. $GPD(\sigma, \gamma)$ has the distribution function (p.d.f.) is:

$$f(x; \sigma, \gamma) = \begin{cases} \frac{1}{\sigma} \left(1 - \frac{\gamma x}{\sigma}\right)^{\frac{1}{\gamma}-1}, & (1 - \frac{\gamma x}{\sigma}) \geq 0, \gamma \neq 0, \sigma > 0 \\ \frac{1}{\sigma} e^{-\frac{x}{\sigma}}, & x \geq 0, \gamma = 0, \sigma > 0 \end{cases} \quad 1$$

Where, σ, γ are the scale and shape parameter respectively. For,

$\gamma > 0$, $\gamma = 0$ and $\gamma < 0$, we have the usual GPD; a type II GPD; and exponential distribution respectively. The GPD can be extended by adding a location parameter μ . The $GPD(x; \sigma, \gamma)$ is defined to be $GPD(x - \mu; \sigma, \gamma)$. And the Cumulative Distribution Function (cdf) is:

$$F(x; \sigma, \gamma) = \begin{cases} 1 - \left(1 - \frac{\gamma x}{\sigma}\right)^{\frac{1}{\gamma}}, & \gamma \neq 0, \sigma > 0 \\ 1 - e^{-\frac{x}{\sigma}}, & \gamma = 0, \sigma > 0. \end{cases} \quad 2$$

The p-quantile of the $GPD(x; \gamma, \sigma)$ is:

$$x_p = \begin{cases} \sigma[1 - (1 - p)^\gamma] / \gamma, & \text{if } \gamma \neq 0 \\ -\sigma \log(1 - p), & \text{if } \gamma = 0 \end{cases} \quad 3$$

3 Fitting the GPD

After choice the value of threshold to estimate the GPD parameters by MLE. Considering that the data $x_{1:n}, x_{2:n}, \dots, x_{n:n}$ are known sorted random sample from $GPD(x; \gamma, \sigma)$ and the parameters for the distribution are unknown, then the likelihood function of GPD can be expressed as follows:

$$L = \prod_{i=1}^n f(x_i, \sigma, \gamma) = \prod_{i=1}^n \frac{1}{\sigma} \left(1 - \frac{\gamma x_i}{\sigma}\right)^{\frac{1}{\gamma}-1} \quad 4$$

The method of MLE works by finding the value for the parameters that maximize the value of likelihood function. Equivalently, the log-likelihood may be maximized on taking the logarithms of (4),

$$l(\sigma, \gamma; x) = -n \log \sigma + (1 - 1/\gamma) \sum_{i=1}^n \log\left(1 - \frac{\gamma x_i}{\sigma}\right) \quad 5$$

The log-likelihood is derived from (5) as:

$$\frac{\partial \ln l}{\partial \gamma} = \frac{\sum_{i=1}^n \log\left(1 - \frac{\gamma x_i}{\sigma}\right)}{\gamma^2} + \frac{(1 - \gamma) \sum_{i=1}^n \frac{\gamma}{\sigma(1 - \frac{\gamma x_i}{\sigma})}}{\gamma} = 0 \quad 6$$

$$\frac{\partial \ln l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{(1 - \gamma) \sum_{i=1}^n \frac{\gamma}{\sigma(1 - \frac{\gamma x_i}{\sigma})}}{\gamma} = 0 \quad 7$$

The MLE solution is found via numerical routines or numerical optimization method (i.e., Newton-Raphson method) are used to find the MLE of γ and σ For computational, see for more details, [19,20].

4 Model adequacy

We can use hypothesis testing to select the best fit model of the data over threshold. Therefore, we now implement two criteria to test if the GPD is a reasonable for our flood dataset or not.

4.1 Deviance Test (DT)

Deviance test is a statistical test provides one objective criterion for selecting among possible models. The null and the alternative hypotheses, written as:

$$H_0 : \gamma = 0 \text{ versus } H_1 : \gamma > 0$$

The first test presented can be found in [14]. The deviance statistic to be used is given by:

$$DT = 2\{L_0(M_0) - L_1(M_1)\} \approx \chi_k^2 \quad 8$$

Where $L_0(M_0)$ and $L_1(M_1)$ be the maximized values of the log-likelihood for distribution under H_0 . To accept exponential distribution as H_0 if $DT < \chi_{\alpha, k}^2$. For details see [21]. Another test that can be performed comes from Akaike information criterion (AIC).

4.2 Akaike information criterion (AIC test)

In order to check a number of models, we perform tests with Akaike information criterion (AIC), see [22,23]. The AIC test is given by:

$$AIC = 2nll(\hat{\phi}_i) + 2i \quad 9$$

Where nll is the negative log likelihood and $\hat{\phi}_i$ is parameter distribution with i events. The goodness of fit is a first term of AIC and the second term is penalizing model complexity. The model with the smallest value for AIC is preferred. The AIC, DT can be performed with package of POT and evd. in [30,31].

5 Graphical Diagnostics for Threshold Choice

Some graphics diagnostics should be introduced to select the value of threshold. In this section, we will be introduced five diagnostics plot to select the suitability threshold value to fit GPD. In,[14], outlined the diagnostics for the choice of threshold. There are several popular methods for choosing the suitable threshold, some of these are,

numerical procedure, plotting methods and combinations between them. Five diagnostic plot of threshold choice as follows:

5.1 Threshold Choice Plot (TCP)

Stability plot (SP) or Threshold Choice plot (TCP) is procedure for a range of thresholds selection to fit the GPD. If the GPD is a suitable fit for high a threshold μ_0 , then $\mu > \mu_0$ will be also follow a

GPD. The plotting (μ_1, σ^*) versus $(\mu_1, \hat{\xi})$ together with confidence interval for each of these plotting (see [14,30,34]), and to select μ_0 as the smallest value when the estimating remains stable region. The stability plot shows the points given by:

$$\{(\mu_1, \sigma^*) : \mu_1 \leq X_{\max}\} \text{ and } \{(\mu_1, \hat{\xi}) : \mu_1 \leq X_{\max}\} \tag{10}$$

Where X_{\max} is the maximum of the observations X .

5.2 Mean Excesses Plot (MEP)

Another graphical method which is widely used to determine the threshold is the mean excess plot (or shortly MEP), is a graphical method for determining a d.f. F . The MEP was described by [15,28,34] and used the mean excesses of the GP to select the best threshold. The mean excesses plot is defined by:

$$e(\mu) = E(X - \mu | X > \mu) = \frac{\sigma_\mu + \xi\mu}{1 - \xi} \text{ for } \xi < 1. \tag{11}$$

If the GPD is a suitable fit, the plot of $e(\mu)$ against μ , called MEP should follow a straight line with intercept $\sigma_\mu / 1 - \gamma$ and slope $\gamma / 1 - \gamma$ over threshold μ , it is mean that the data support the GPD with a long-tailed ($\gamma > 0$). Furthermore, a horizontal MEP (medium tailed, γ is near 0) mean that exponentially model while thin-tailed ($\gamma < 0$), see,[16]. for a discussion of the properties of this function. In practice, based on a sample size, n ,

X_1, X_2, \dots, X_n , $e(\mu)$ is estimated by its empirical mean. The sample ME function is given by:

$$e_n(\mu) = \frac{\sum_{i=1}^n (x_i - \mu)}{\sum_{i=1}^n I\{x_i > \mu\}} \tag{12}$$

Where I is a function of indicator $I\{x_i - \mu\} = 1$ if $\mu < X_i$, and 0 otherwise. To show the sample ME-plot, we plot the points as follows:

$$(X_{n-k:n}, \hat{e}_n(X_{n-k:n}), 1 < k < n-1) \tag{13}$$

Where $X_{n-k:n}$ denotes the $(k + 1)^{th}$ largest observation and where $\hat{e}_n(X_{n-k:n})$ may be rewritten as:

$$\hat{e}_n(X_{n-k:n}) = \frac{1}{k} \sum_{j=1}^k x_{n-j+1:n} - x_{n-k:n} \tag{14}$$

If the data follows at GPD greater than some high value of threshold, we will be expected the MEP to see linear in view of Eq-11. One quite hard with the MEP, is very much dispersion, especially at above thresholds, it will be hard to choose whether an observed departure from linearity. Another graphical plot should be used to select the value of threshold is the Hill plot (HP).

5.3 Hill plot (HP)

The most popular tail index estimator is the Hill plot (HP) by [25]. which, however, is restricted to the Pareto case $\gamma > 0$. The HP is another approach to estimate the positive shape (heavy-tailed, $\alpha = 1/\gamma$) distributions, the HP will be employed for the following reasons, easy implementation, asymptotic unbiasedness under large samples and the most efficient estimator of shape parameter γ . Let

$X_1 \leq X_2 \leq \dots \leq X_n$ be the ordered statistics of random variables iid. The HP is defined by the set of points $\{(k, H_{k,n}), 1 < k < n-1\}$ where the HP estimator is provided by the formula:

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^k \log X_{n-i+1:n} - \log X_{n-k:n} \tag{15}$$

For every choice of k , we obtain another estimator of γ . The threshold μ is choice from the plot of areas remain constant of the tail index (TI). However, this select is not always clear. We hope of finding a stationary area when estimates of TI do not change above a different value of threshold obtained. The results from the Hill estimator (HE) are based on the TI selection. The HP proved by [26], at another plot diagnostic to establish the TI, which HP for a range value of k versus the TI, or k against the corresponding threshold. Many authors were studied the statistical behavior and properties of the HE to extend the HE to the general case $\gamma \in \mathfrak{R}$. Recent generalizations of the HP for shape ($\gamma \in \mathfrak{R}$) are introduced by [26,27]. As [29], extend the idea of Hill to derive a plug-in estimator by applying the hypothesis test on an accumulation of the log spacing's further considered a kernel-based goodness of fit statistic of the tail fit in the Pareto type tail case. Another approach which has been considered for the threshold selection is the dispersion index plot.

5.4 Dispersion Index Plot (DIP)

According to [30], is presented another plot to determine the threshold, namely the Dispersion Index Plot (DIP). The DIP is special useful when dealing with over high a threshold will be asymptotic by a generalize pareto distribution. Let X be a random variable as a Poisson distribution with parameter λ That is:

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k}, k \in \mathcal{N} \tag{16}$$

where λ is the average of events. Moreover, a confidence interval can be computed by using a χ^2 test

$$I_\alpha = \left[\frac{\chi_{(1-\alpha)/2, M-1}^2}{M-1}, \frac{\chi_{1-(1-\alpha)/2, M-1}^2}{M-1} \right] \tag{17}$$

Where $P(DI \in I_\alpha) = \alpha$

The following is the last graphical plot based on the estimation of the model at a range of thresholds is call Quantail Quantail Plot (QQP).

5.5 Quantail Quantail plot (QQ-Plot)

The QQ-plot introduced by [30], as an alternative to the Hill plot. And, the HE is convergence to the slope of the line fitted to the upper TI of GPD, this estimator, will be defined by [15,32], can be represented by formula of $\hat{\gamma}$:

$$\hat{\gamma}_{qq} = \frac{\sum_{i=1}^k \ln \frac{i}{k+1} (\ln X_{i:n} - \ln X_{k+1:n})}{k \sum_{i=1}^k \left(\ln \frac{i}{k+1} \right)^2 - \left(\sum_{i=1}^k \ln \frac{i}{k+1} \right)^2} \tag{18}$$

Some features were established by the authors, such as the asymptotic variance of the q-q estimator, weak consistency, and asymptotic normality. However, when the convergence is not completely perfect, the main advantage of using the q-q-estimator over the HE is that the residuals of the GPD plot provide information to figure out the bias in the estimates. The estimator is represented as being reasonable in the case of the bias of the standard estimator.

6 Application on real data

In this study, we use the ardières data frame containing flood discharges over a period of 33 years of the Ardières river at Beaujeau. The source data is taken from the POT package in R by [30,33] and [35]. This data was used because it was used in some applications of extreme values, in addition to not obtaining real data for extreme value phenomena. First, we have to extract extreme events while preserving independent between events from the time series and select a suitable threshold such that asymptotic approximation in equation (1) is good enough.

6.1 Descriptive Statistic

Firstly, descriptive statistics summarizes the characteristics of a dataset to see the behavior of our data. Table-1, shows the results of some important descriptive statistic (min, max, mean, median and skewness). The results in **Table-1** shows that the sample mean is

(4.19) and the median is (3.03) of the two statistics, the mean is greater than the median, this situation is also called positive skewness, The skewness of the data have a positive skewness ($SK=4.78 > 0$), which suggests heavy tails. To make the comparison between distributions, first we performed tests to verify the symmetry (normality) or asymmetric and to ensure that on real data. For such verification we performed the Shapiro-Wilk test and ks.test to confirm the data does not follow a normal distribution. A positive value of skewness signifies a distribution with an asymmetric (not normal) tail extending out towards more positive and If your data are not symmetric, the mean and median are not equal (or similar). If the distribution of data is skewed to the right (or asymmetric), the median is often less than the mean, therefore, there is a strong reason to suggest the model of these data belong to heavy tail of GPD (long tail, $\gamma > 0$).

Table 1. Summary statistics for ardieres data

Statistics	Values
Min	1.51
Max	44.20
Median	3.03
Mean	4.19
Skewness	4.78
ks.test	0.24 (0.00**)
shapiro.test	0.59 (0.00**)

1% (P-value is in parentheses)

6.2 Graphical Diagnostics (GD)

To fit GPD of our data, it is first important to choose a threshold. Different graphical diagnostics have been suggested for choosing the threshold. All the different graphical diagnostic presented in section 5 have been used, and it is then possible to compare the different values of thresholds obtained. Fig-1, show five graphical diagnostics for selection the value of threshold, and help in choosing where to begin looking at thresholds. From Fig-1, TCP is on the top-left and right panel, medium left panel is MEP, while medium right panel is Hill plot, bottom-left panel is the DIP and bottom-right panel is QQ-plot. It is not easy to show by eye, and the interpretation of these plots often requires a good deal of subjective judgment.

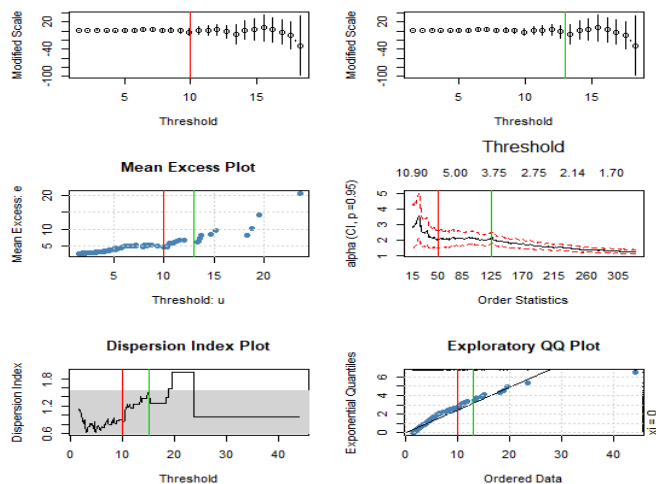


Fig-1: Graphic diagnostic for the threshold choice to fit GPD and the vertical line indicates to the graphical threshold selection.

In Fig-1, the results of graphical diagnostic are shown. According to Fig-1, TCP is shown on the top-left and right panel, we look to construct a horizontal line, that cuts through all the confidence bars and both modified scale and shape estimates seem to be constant or stability on the range ($u = 10$ to $u = 13$). TCP for lower shape ($u = 10$) and upper modified scale ($u = 13$) respectively for the data. Two thresholds ($u = 10$ & $u = 13$) are shown by vertical dashed lines. The MEP in Fig-1 (Medium left panel), is linear on the range for thresholds between lower and upper straight lines as ($u = 10$ to $u = 13$) respectively. We look for approximate linearity whilst keeping in between the the confidence bounds. Indeed,

the MEP are vertical dashed lines mark these thresholds, since there is some indicate that the plot slightly about this threshold, and represented a threshold at about the 97th percentile of the data. From the Fig-1, we see that the MEP ultimately increases, so the TI of our data is positive. It seemingly indicates linearity from threshold $u = 10$ to $u = 13$, but it is not easy to detect which point is an appropriate threshold. Based on linear property, all points between (10,13) may be good for threshold. From Fig-1, can be observed in the medium right panel is Hill plot. The vertical dashed lines, in combination with the connection HE, indicate that the HP is rather stable about the range threshold of (shape, $\gamma = .0.30$ and $\gamma = 0.16$ respectively). These two thresholds produce quite different tail extrapolations (shape, $\gamma = .0.30$ and $\gamma = 0.16$ respectively), but they are both in keeping with the overall recommendations for threshold selection using the HP. The HP, is inconsistencies are observed between the estimated shape parameter at this range and other thresholds of TCP and MEP. On the other hand, the DI plot is shown in Fig-1 (on bottom-left panel) does not quite stabilize in any particular region. From Fig-1, we realize that the DIP unstable in any especially area, making it extremely difficult to select a specific value for threshold value at this plot. A stable region in the DIP for threshold is (11,13) and the high dispersion present in the region of the high order statistics in the DIP. The Q-Q-plots show in Fig-1 (on bottom-right panel) for empirical versus GPD model quantiles is seen. The qq-plots shows a rapid increase up to around the threshold $u = 10$ to $u = 13$ should be chosen, then levels off. A vertical line marks the location of our threshold.

6.3 Fitting the GPD

Based on Fig-1, five different thresholds are chosen for fitting the shape parameter estimation of GPD by using MLE, we get the estimates of shape parameters. Fitting distributions was made in different levels of threshold values of u by MLE, and the goodness of fit tests are presented these results in the table-2 below.

Table 2. Estimated parameter of GPD by MLE for different graphical

Type of Graphical methods	TCP,MEP DIP,QQP $u=10$	TCP,MEP DIP, QQP $u=13$	HP $u=3.6$	HP $u=6$
Threshold	10	13	133	49
Proportion Above	0.07	0.03	0.40	0.14
Quantail	0.93	0.97	0.60	0.86
Shape	0.50	0.57	0.30	0.16
D-test	112.29 (0.00*)	58.82 (0.01*)	572.9 (0.00*)	251.81 (0.00*)
Order	2	1	4	3
AIC-test	116.26 (0.00*)	62.82 (0.01*)	576.9 (0.00*)	255.81 (0.00*)
Order	2	1	4	3

*(is P-value)

Table-2, summarizes some estimation results of the shape parameter γ via the MLE estimator at each four choices of threshold are summarized in the Table-2, and the results are stable for TCP, MEP and qq-plot except for HP. for comparison between diagnostic plots, to achieve a good model fit, traditionally, the threshold was chosen a suitable value of threshold before fitting GPD. The shape parameters estimated for both ($\gamma = 10$ to $\gamma = 13$ at threshold ($u = 10$ to $u = 13$)) are very close, there is not a great difference between them and the shape parameter was always positive.

6.4 Goodness-of-Fit Criteria (DT & AIC)

According to goodness-of-fit test via DT and AIC. Both DT and AIC information criteria test are used to obtain whether or not the observed differences are consistent to find the most an appropriate threshold choice to fit the data by GPD. Note that the values of AIC for TCP, MEP are less than the values for the HP, QQP in all diagnostic plots. Therefore, TCP best fitted in all cases where the stability of TCP was higher. The estimated shape parameters are significantly different from threshold. In this case, the data for all the considered thresholds pass both goodness of fit tests and the tests are significant. Another way of checking the adequacy of GPD fit estimates via diagnosis plot.

6.4 Graphical Tools for Data Analysis

After choosing our threshold, we create graphical diagnostic plots where we desire linearity amongst the fulfillment of other criteria to check the threshold's suitability. In Fig 2-5. some graphical method to check the suitable threshold of the GPD distribution are shown. Fig-2 and 3, show the fitted model at threshold 10 and 13, while Fig 4 and 5 at threshold 3.6 and 6.

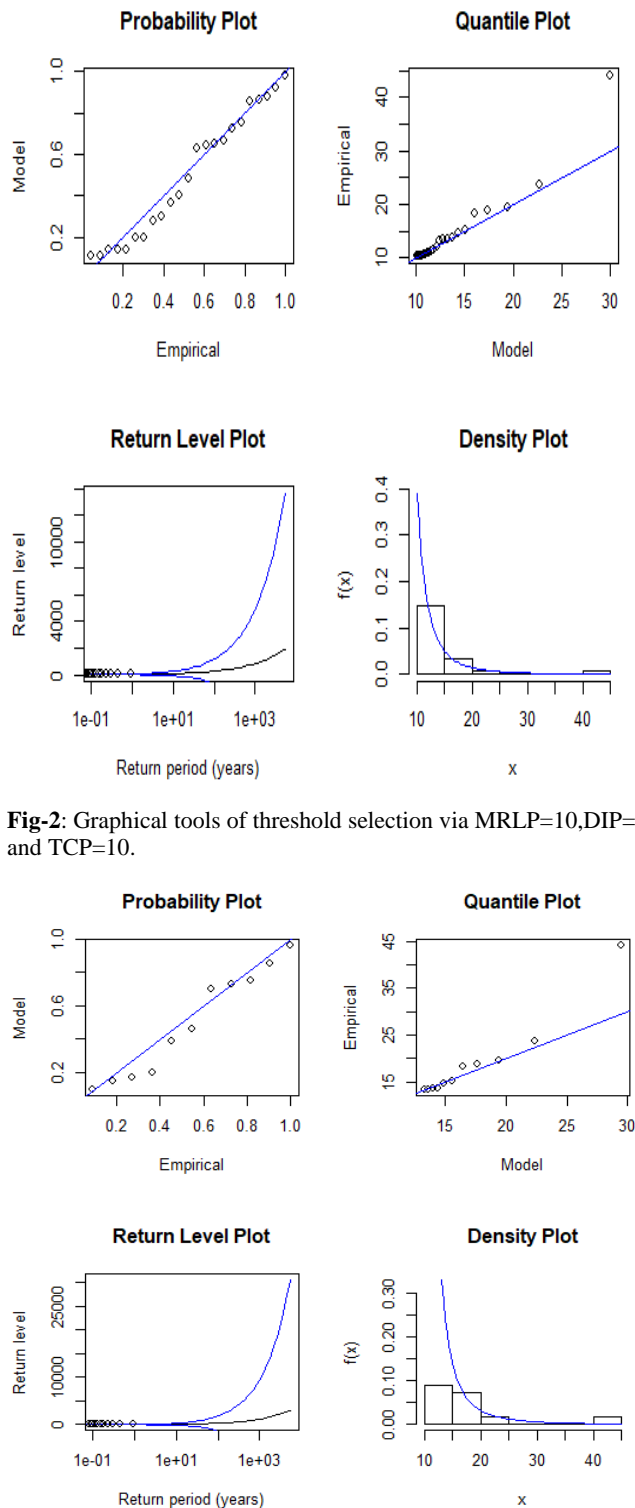


Fig-2: Graphical tools of threshold selection via MRLP=10,DIP=10 and TCP=10.

Fig-3: Graphical tools of threshold selection via MEP=13, DIP=13 and TCP=13

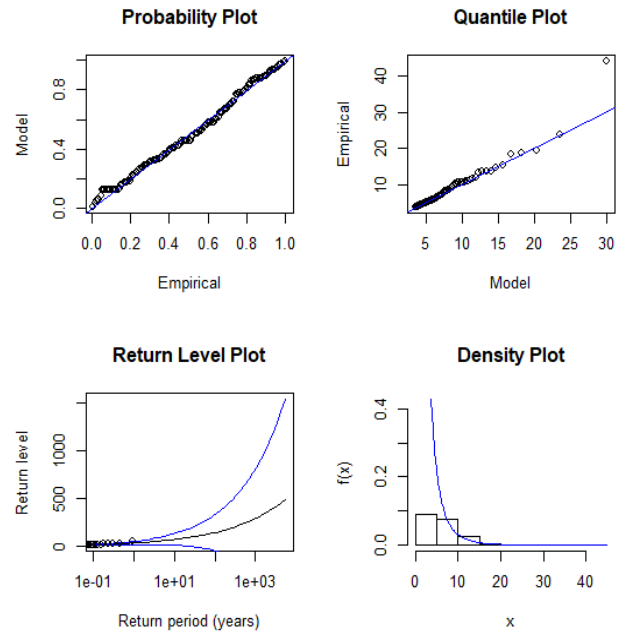


Fig-4: Graphical tools of threshold selection via Hill=3.6.

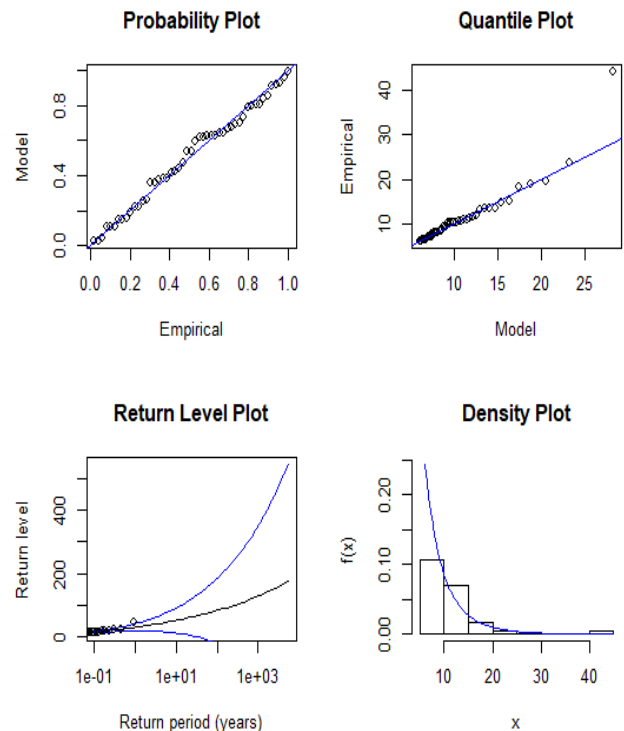


Fig-5: Graphical diagnostic plots of threshold selection via Hill=6.

Depended on Fig-2, Both the probability plot (top left) and the q-q plot (top right) show the reasonability of the GPD fit. The RL curve (bottom left) asymptotes as a consequence of the positive shape. Finally, the correspond density estimate (bottom right) seems consistent with the histogram of the data. Consequently, all four diagnostics plots reduce to GPD model. Fig-2, shows diagnostic plots at GP distribution is best fit. On other hand, all diagnostics plot in Fig 2-5, seem to indicate a reasonable fit of the GPD to our data.

7 Conclusion

The problem of threshold selection estimating of long-tailed model is very important in many practical applied. In this study, we have presented a five graphical threshold selection method. Various diagnostics plots to evaluate the GPD fit are commonly used for selecting threshold, such five-plot including, Stability plot (SP) or Threshold Choice plot (TCP), Mean Excesses Plot (MEP), Hill plot (HP), Dispersion Index Plot (DIP) and Q-Q plot (QQP) respectively. These graphical methods are a diagnostic plot drawn before fitting

any model (GPD) to choose an appropriate value of threshold. In addition, we can use diagnostic plots, P-P plots, q-q plots and RL plot to select the best fit of GPD. All diagnostics seem to indicate a reasonable fit of the GPD to our data. Visually we can see a slight difference in the estimation of the threshold four of these graphical diagnostics (TCP, MEP, DI and QQP) share with same selecting threshold range (10,13). On the other hand, Hill graphical diagnostics differences among these graphical diagnostics. All depend on visual inspection to identify the threshold. The level over which linearity is evident might be used as a threshold level for various levels. If the model provides significantly different outcomes for different values. An important step in any strategy is selecting a threshold depending on the ME plot so that the plot is nearly linear above it. Threshold selection can be difficult, and parameter estimations might be affected by the threshold selection, particularly when real data is analyzed. It's hard to determine the best threshold through using this method. The position of a stable zone in the HP and the significant dispersion present in the region of the upper-order statistics in the DI are quite difficult. The problems can occur when transfers from the GPD occur. The q-q-plot is an alternative to the HP. Although the HP seems to be less smoother than the QQ-plot, problems can still occur when departures from the Pareto distribution occur. There are some issues that would be taken into consideration, the first problem is how to deal with choosing the threshold value or the number of upper-order statistics required for the GPD by using an automated threshold selection instead of graphical diagnostics that require prior experience (or subjective) of their interpretation of plots. The second problem is how to find some modifications of these proposed in the literature based on smoothing and robustifying procedures and other diagnostic tests in order to select the best threshold of fit GPD model. Finally, all results of graphical diagnostics belonging to GPD under linear can be done nonlinear, so it would be recommended to have other practical applications to GPD under power.

References

- [1]- Bobée, B., & Rasmussen, P. F. (1995). Recent advances in flood frequency analysis. *Reviews of Geophysics*, 33(S2), 1111-1116.
- [2]- Ewemoje, T. A., & Ewemoje, O. S. (2011). Best distribution and plotting positions of daily maximum flood estimation at Ona River in Ogun-Oshun River basin, Nigeria. *Agricultural Engineering International: CIGR Journal*, 13(3).
- [3]- Clausen, B., & Pearson, C. P. (1995). Regional frequency analysis of annual maximum streamflow drought. *Journal of Hydrology*, 173(1-4), 111-130.
- [4]- Madsen, H., Mikkelsen, P. S., Rosbjerg, D., & Harremoës, P. (1998). Estimation of regional intensity-duration-frequency curves for extreme precipitation. *Water Science and Technology*, 37(11), 29-36.
- [5]- Balkema, A. A., & De Haan, L. (1974). Limit laws for order statistics (No. 2099-2018-3082).
- [6]- Pickands III, J. (1975). Statistical inference using extreme order statistics. *the Annals of Statistics*, 119-131.
- [7]- Zhang, X., & Ge, W. (2009, December). A new method to choose the threshold in the POT model. In 2009 First International Conference on Information Science and Engineering (pp. 750-753). IEEE.
- [8]- Rosbjerg, D., Madsen, H., & Rasmussen, P. F. (1992). Prediction in partial duration series with generalized Pareto-distributed exceedances. *Water Resources Research*, 28(11), 3001-3010.
- [9]- Abild, J., Andersen, E. Y., & Rosbjerg, D. D. (1992). The climate of extreme winds at the Great Belt, Denmark. *Journal of Wind Engineering and Industrial Aerodynamics*, 41(1-3), 521-532.
- [10]- Hosking, J. R., & Wallis, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, 29(3), 339-349.
- [11]- Singh, V. P., & Guo, H. (1995). Parameter estimation for 3-parameter generalized Pareto distribution by the principle of maximum entropy (POME). *Hydrological sciences journal*, 40(2), 165-181.
- [12]- McNeil, A. J. (1997). Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin: The Journal of the IAA*, 27(1), 117-137.
- [13]- Davison, A. C., & Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 52(3), 393-425.
- [14]- Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). An introduction to statistical modeling of extreme values (Vol. 208, p. 208). London: Springer.
- [15]- Beirlant, J., Dierckx, G., & Guillou, A. (2005). Estimation of the extreme-value index and generalized quantile plots. *Bernoulli*, 11(6), 949-970.
- [16]- Reiss, R. D., Thomas, M., & Reiss, R. D. (1997). *Statistical analysis of extreme values (Vol. 2)*. Basel: Birkhäuser.
- [17]- Silva-González, F., Heredia-Zavoni, E., & Inda-Sarmiento, G. (2017). Square Error Method for threshold estimation in extreme value analysis of wave heights. *Ocean Engineering*, 137, 138-150.
- [18]- Wuertz, Diethelm, et al. "Package 'fExtremes'." (2009).
- [19]- Gilleland, Eric, Mathieu Ribatet, and Alec G. Stephenson. "A software review for extreme value analysis." *Extremes* 16 (2013): 103-119.
- [20]- Gilleland, Maintainer Eric. "Package 'ismev'." (2018).
- [21]- Embrechts, P., Kluppelberg, C., & Mikosch, T. (1999). *Modelling extremal events*. *British actuarial journal*, 5(2), 465-465.
- [22]- Guillou, A., & Hall, P. (2001). A diagnostic for selecting the threshold in extreme value analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 293-305.
- [23]- Goegebeur, Y., & Guillou, A. (2010). Goodness-of-fit testing for Weibull-type behavior. *Journal of Statistical Planning and Inference*, 140(6), 1417-1436.
- [24]- Towler, E., Rajagopalan, B., Gilleland, E., Summers, R. S., Yates, D., & Katz, R. W. (2010). Modeling hydrologic and water quality extremes in a changing climate: A statistical approach based on extreme value theory. *Water Resources Research*, 46(11).
- [25]- Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 1163-1174.
- [26]- Drees, H., Resnick, S., & de Haan, L. (2000). How to make a Hill plot. *The Annals of Statistics*, 28(1), 254-274.
- [27]- Gomes, M. I., & Martins, M. J. (2001). Generalizations of the Hill estimator—asymptotic versus finite sample behaviour. *Journal of statistical planning and inference*, 93(1-2), 161-180.
- [28]- Beirlant, J., Vynckier, P., & Teugels, J. L. (1996). Excess functions and estimation of the extreme-value index. *Bernoulli*, 293-318.
- [29]- Goegebeur, Y., Beirlant, J., & de Wet, T. (2010). Kernel estimators for the second order parameter in extreme value statistics. *Journal of statistical Planning and Inference*, 140(9), 2632-2652.
- [30]- Ribatet, M. (2011). *A user's guide to the PoT package (version 1.4)*. University of Montpellier II.
- [31]- Gilleland, E., Ribatet, M., & Stephenson, A. G. (2013). A software review for extreme value analysis. *Extremes*, 16, 103-119.
- [32]- Kratz, M., & Resnick, S. I. (1996). The QQ-estimator and heavy tails. *Stochastic Models*, 12(4), 699-724.
- [33]- Gouy, V., Liger, L., Ahrouch, S., Bonnineau, C., Carluet, N., Chaumot, A., ... & Pesce, S. (2021). Ardières Morcille in the Beaujolais, France: A research catchment dedicated to study of the transport and impacts of diffuse agricultural pollution in rivers. *Hydrological Processes*, 35(10), e14384.
- [34]- Hafid Aboubakr Alaswed & Mohamed Amraja Mohamed, (2019). Threshold Selection Using Graphical Methods. *Journal of Pure & Applied Sciences*. 17(1): 42-45.
- [35]- Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J. (2019). *Extremes in R: Package "extRemes"*. R package version 2.0-10. Retrieved from <https://CRAN.R-project.org/package=extRemes>.