مجلة العلوم البحثة والتطبيقية
Journal of Pure & Applied Sciences
www.Suj.sebhau.edu.ly ISSN 2521-9200
Received 11/08/2018 Revised 14/01/2019 Published online 01/04/2019

Threshold Selection Using Graphical Methods

*Hafid Aboubakr Alaswed , Mohamed Amraja Mohamed
Statistica Department, Faculty of Science, Sebha University-Libya
*Corresponding author: haf.alaswed@sebhau.edu.ly

Abstract One of the major challenges in Peak over Threshold model (POT) the selection of the best threshold in fitting the Generalized Pareto Distribution (GPD) which is widely used in many applications. The choice of threshold must be a balance between bias and variation. In this paper we comparison between two graphical methods to determine the best threshold in the POT model and estimate the tail index. The results obtained from different estimators used to estimate the shape distribution of GPD by using maximum likelihood (ML). Finally, in this paper we use application on real data to compare the properties of different estimators for estimating tail index. The results show that GPD model with threshold of threshold choice plot (TCP) is a better choice basis on the Deviance and Akaike information test. For the calculations, we will use the R programming with packages POT and ismev for parameter estimation and diagnostic plots.
Keywords: Generalized Pareto Distribution (GPD), Goodness-of-Fit Test, Methods of Threshold Selection, Mean Residual Plot (MRLP), Threshold Choice Plot (TCP).

اختيار العتبة باستخدام الطرق البيانية

*حافظ ابوبكر محمد و محمد امراجع محمد
قسم الإحصاء– كلية العلوم– جامعة سبها، ليبيا
haf.alaswed@sebhau.edu.ly :*للمراسلة

الملخص إحدى التحديات الرئيسية في طريقة ذروة فوق العتبة هو كيفية اختيار أفضل عتبة عند نمذجة توزيع باريتو المعمم الذي يستخدم على نطاق واسع في العديد من التطبيقات. عند اختيار العتبة يجب أن يكون هناك توازنا بين التحيز والتباين. في هذه الورقة تم عرض طريقتين بيانيتين والمقارنة بينهما لتحديد أفضل عتبة لنمذجة توزيع باريتو وتقدير مؤشر الذيل. النتائج التي تم الحصول عليها من هاتين الطريقتين (مقدرات مختلفة) تستخدم لتقدير معلمة الشكل لتوزيع باريتو المعمم. وأخيرا، في هذه الورقة تم التطبيق على بيانات لمقارنة الخصائص المختلفة لهذين المقدرين المختلفين لتقدير مؤشر الذيل حيث أظهرت النتائج أن نمذجة توزيع باريتو باستخدام قيمة العتبة الخاصة برسمة اختيار العتبة هو الاختيار الأفضل بناء على اختبار ديفاينس واكاكيا. تم استخدام برنامج R لإجراء العمليات الحسابية والأحصائية لتقدير المعلمة والرسومات التشخيصية.
الكلمات المفتاحية: توزيع باريتو المعمم، اختبار جودة التوفيق، طرق اختيار العتبة، رسمة متوسطات البواقي ، رسمة اختيار العتبة.

1 Introduction

Extreme value theory (EVT) is one of major importance in many fields of applications where extreme values may appear and have detrimental effects explored by [15],[22]. In EVT, the problem of threshold selection to estimate the tail index of distributions is very important in many applications, see, [8]. The last decade has seen development methods of threshold selection in extreme value applications. In, [16], the classical asymptotically motivated model for excesses above a high threshold is the generalized Pareto distribution (GPD), and [14] for the original theoretical development and [7], for further developments and applications.

The paper is organized as follows: In Section 2 the general theoretical background of GPD is provided. In Section 3, method of threshold selection. In Section 4, some graphical methods for dealing with the issue of choosing the threshold value for the estimation of shape are introduced. Parameter estimation and model selection are introduced in Section 5. Section 6, describes case study on real data. Finally, concluding remarks are given in Section 7.

2 Generalized Pareto Distribution (GPD)

The generalized Pareto distribution (GPD) was introduced by [14] as a three parameter distribution and has been used widely by many scientists. In [19], the GPD is usually expressed as a three parameters distribution with d.f.

$$P_\gamma(x;\mu,\sigma_\mu)=\begin{cases}1-(1+\gamma\frac{x-\mu}{\sigma_\mu})^{-1/\gamma}, & if\ \gamma>0\\ 1-\exp(\frac{x}{\sigma_\mu}), & if\ \gamma=0\\ 1-(1+\gamma\frac{x-\mu}{\sigma_\mu})^{-1/\gamma}, & if\ \gamma<0\end{cases} \quad (1)$$

Where, $\sigma>0$ scale parameter, μ is location parameter and γ the shape parameter. The support is $x\geq\mu$ when $\gamma>0$ and $\mu\leq x\leq\mu-\frac{\sigma_\mu}{\gamma}$ when $\gamma<0$. The GPD subsumes three other distributions under its parameterization. When $\gamma>0$, we have a version of the usual Pareto

distribution; if $\gamma = 0$ we give the exponential distribution while $\gamma < 0$, we have a type II Pareto distribution. To achieve a good model fit of GPD, we need to choose a suitable value of threshold. We now outline our methods for threshold selection, see for example, [20].

3 Methods of Threshold Selection

Extremes can be defined in a number of ways. Here we adopt the excesses over a threshold approach. In [3],[20], the selection of an appropriate threshold is one of the important concerns of the POT approach and still an unsolved problem an area of ongoing research in the literature which can be of the critical importance. In [6],[9], it states that the selection of the threshold process always is a trade-off between the bias and variance. If a too high threshold is selected, the bias decreases while the variance increases as there is not enough data above this threshold. On the other hand, by taking a lower threshold, the variance decreases as the number of observations is larger and the bias increases. [9], outlines a graphical methods used for the threshold selection in [5], and also suggested by [22]. Next we will give some graphical methods of threshold selection.

4 Graphical methods of threshold selection

Several graphics have been proposed to assist in threshold selection. We will illustrate some graphics plot to select the suitability threshold of the fitted GPD. In the following, we will describe two diagnostic plot of threshold selection [13].

4.1 Mean Residual Life Plot (MRLP)

The mean residual life plot (MRLP) is a graphical tool widely used for assessing the behavior of a distribution function (d.f.). The MRLP was introduced by [5],[2], uses the expectation of the GPD excesses, see [16]. One tool for choosing suitable thresholds is the sample MRLP:

$$\left\{(\mu, e_n(\mu), x_{1:n} < \mu < x_{n:n}\right\} \qquad 2$$

where $x_{1:n}$ and $x_{n:n}$ are the minimum and maximum order statistics of the data sample, μ is the threshold and $e_n(\mu)$ is the sample mean excess function defined by

$$e_n(\mu) = \frac{\sum_{i=1}^{n} (x_i - \mu)}{\sum_{i=1}^{n} 1(x_i > \mu)} \qquad 3$$

i.e. the sum of the excesses over the threshold μ divided by the number of data points which exceed the threshold. In particular, if the empirical plot seems to follow a reasonably straight line with positive gradient above a certain value of μ, then this is an indication that the excesses over this threshold follow a GPD with positive shape parameter in [12]. In practice, if x represents excess over a threshold μ_0, and if the approximation by a GPD is good enough, we have:

$$E(x - \mu_0 / x > \mu_0) = \frac{\sigma_{\mu_0}}{1-\gamma} \qquad 4$$

For all new threshold μ_1 such as $\mu_1 > \mu_0$, excesses above the new threshold are also approximate by a GPD with updated parameters.

$$E(x - \mu_1 / x > \mu_1) = \frac{\sigma_{\mu_1}}{1-\gamma} = \frac{\sigma_{\mu_0} + \gamma\mu_1}{1-\gamma} \qquad 5$$

Another graphical tool can be performed to choose the threshold that is the stability plot or threshold choice plot (TCP).

4.2 Threshold Choice Plot (TCP)

Let X be a random variable (r.v) with $GPD(\mu_0, \gamma_0, \sigma_0)$. Let μ_1 be an another threshold as $\mu_1 > \mu_0$. The random variable $X \setminus X > \mu_1$ is also GPD with new parameters $\sigma_1 = \sigma_0 + \gamma_0(\mu_1 - \mu_0)$ and $\gamma_1 = \gamma_0$.

Let $\sigma^* = \sigma_1 - \gamma_1\mu_1$ with this new parameterization, σ^* is independent of μ_1. Thus, estimates of σ^* and γ_1 are constant for all $\mu_1 > \mu_0$ if μ_0 is a suitable threshold for the asymptotic approximation. Threshold choice plots represent the points defined by:

$$\left\{(\mu_1, \sigma^*) : \mu_1 \le x_n\right\} \text{ and } \left\{(\mu_1, \gamma_1) : \mu_1 \le x_n\right\} \qquad 6$$

where x_n is the maximum observations of the X. Such plots are found in [8], for GPDs fitted to wave height data by [10],[20].

5 Parameter Estimate and Model Selections

Traditionally, the threshold was chosen before fitting the GPD. Threshold choice involves balancing bias and variance. In [16], practical, the parameters of distribution must be estimated from the data. [21[. There are several methods to estimate parameters. We focus on maximum likelihood estimation (MLE) because of nice asymptotic. [16], it has described how a GPD can be fitted with MLE See [5], [21]. In order to make comparison between the estimates of shape parameters for different threshold choices provide different fitted models of GPD, we use goodness-of-fit tests using different statistics [4], namely the Deviance, Akaike information criterion (AIC) and results appear in Table 1 test as follows:

Deviance test: the deviance statistic is defined by:

$$D = 2\{L_0(M_0) - L_1(M_1)\} \sim \chi^2_k \qquad 7$$

Where $L_0(M_0)$ and $L_1(M_1)$ be the maximized values of the log-likelihood for models under null and alternative hypothesis respectively.

To reject models under null hypothesis if:

$$D > C_\alpha \quad 8$$

where C_α is the $(1-\alpha)$ quantile of the distribution. For more details see [8].

AIC test : we apply the Akaike information criterion (AIC),see, [18]. The AIC is calculated as

$$AIC = 2NLL(\theta_i) + 2i \quad 9$$

where NNL is the negative log likelihood and θ_i is parameter vector with i elements. The model with the smallest value for AIC is preferred [11].

6 Case Study

To illustrate the above method of threshold selection we used the sample data of large fire insurance claims in Denmark from 1980 to 1990. The data are contained 2167 observations in a numeric vector. The source data is taken from the evir package in R [1],[17]. We use this dataset as an example of graphical methods to select an appropriate threshold. Next, we can fit the GPD to those excesses by applying the MLE to estimate the shape parameter of the GPD. In Table 1, we summarize the estimation results for different choices of graphical tools for the threshold selection via the MRLP and TCP.

Table 1: Graphical methods results of threshold selection

Graphical methods		MRLP	TCP
True threshold		20	26
n_μ		36	22
Quantail		0.98	0.99
MLE of shape		0.68	0.84
Testing	Deviance	284.36	186.04
	Order	2	1
	AIC	288.36	190.04
	Order	2	1

Table 1 provides the results of graphical methods to select an appropriate value of threshold. The first row of Table 1, represent the value of threshold selection by using graphical method and these thresholds give exceedances and quantail (probability less than threshold) reported in the second and third row respectively. Shape parameters are estimated of GPD by MLE in the four row of Table 1. For comparison, results of Goodness of Fit tests for both Deviance and AIC are listed in the last row. GPD model with threshold of TCP is a better chosen depended on D and AIC test. The diagnostic plots for two threshold choices shown in Fig.1. On the top panel of Fig. 1 is the MRLP while middle and bottom panel is the TCP. In Fig. 1, the value of threshold is 20 in MRLP because the graphs are approximately linear and a lot of stability is the most an appropriate to fit the studied dataset. While, the threshold value of TCP is 26 would be more an appropriate to find the minimum value above which the estimations of both parameters remain constant. This indicates that, the fitted GPD model is satisfactory for the fire insurance claims in Denmark.

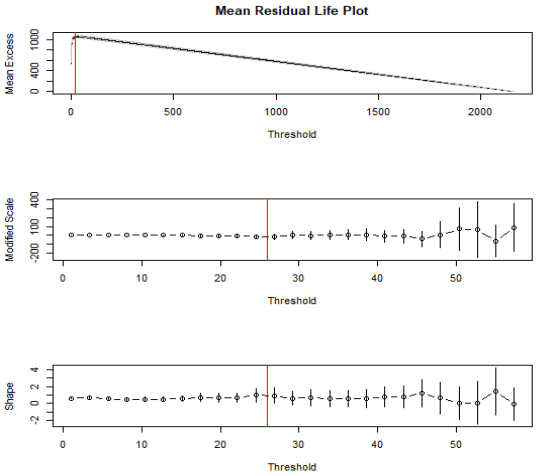

Fig. 1: Graphical tools of threshold selection and the vertical line indicates to the threshold selection via MRLP=20 and TCP=26

7 Conclusion

In fitting GPD, there is the issue of threshold selection. If the chosen threshold is too low, the GPD approximation may not hold and bias can occur. If the threshold is chosen too high, reduced sample size increases the variance of parameter estimates. In this paper, graphical methods are applied to select the best threshold and a suitable threshold should be specified to find the GPD. We have presented two graphical methods MRLP and TCP, which provide different threshold choices. We provide a data analysis to see how the two graphical works in practice. Application on set of real data showed that the TCP, suitable threshold can be chosen an appropriate value of threshold selection when the estimators of the shape parameter keep stable above the threshold. Goodness-of-fit such as the Deviance and AIC of the GPD for the exceedances, and select the lowest one, above which the data provides adequate fit to the GPD, and shows that the GPD model is a good choice. Further research will be conducted to compare graphical methods with numerical methods to select the best threshold. In addition, an automated graphical threshold selection procedure based on a sequence of goodness-of-fit tests, more work is needed in this direction.

References

[1]- Bader, B. and Yan, J. (2015). *Extreme Value Analysis with Goodness-of-Fit Testing*, R package version 0.1.2.

[2]- Beirlant, J., Broniatowski, M., Teugels, J.L. and Vynckier, P. (1995). The Mean Residual Life Function at Great Age: Applications to Tail Estimation. *Journal of Statistical Planning andInference,* 45, 21-48.

[3]- Caeiro, F. and Gomes, M. I. (2016), Threshold Selection in Extreme Value Analysis, *Extreme*

[4]- *Value Modeling and Risk Analysis: Methods and Applications*, 69-82.

[5]- Choulakian, V. and Stephens, M. A. (2001), Goodness-of-Fit Tests for the Generalized Pareto Distribution, *Technometrics*, 43, 478-484.

[6]- Coles, S., Dixon, M. (1999). Likelihood-based inference for extreme value models. Extremes.2(1):5–23.

[7]- Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, Springer, 1st ed.

[8]- Davison, A.C., Smith, R.L., (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series*. B 52 (3), 393–442.

[9]- de Sousa, B. and G. Michailidis (2004), A diagnostic plot for estimating the tail index of a distribution. Journal of Computational and Graphical Statistics 13(4), 974-995.

[10]- Dupuis, D.J., (1999). Exceedances over high thresholds: a guide to threshold selection. Extremes 1 (3), 251–261.

[11]- Guillou, A., Hall, P., (2001). A diagnostic for selecting the threshold in extreme value analysis. J. R. Stat. Soc. Ser. B 63, 293–305.

[12]- Kolbjørn Engeland, Hege H. and Arnoldo F. (2004). Practical extreme value modelling of hydrological floods and droughts: a case study. *Extremes*, 7(1): 5–30.

[13]- McNeil, A. J. and Saladin, T. (1997). The Peaks over Thresholds Method for Estimating High Quantiles of Loss Distributions. Proceedings of XXVIIth International Astin Colloquium,Cairns, Australia, 23–43.

[14]- Northrop, P. J. and Coleman, C. L. (2014), Improved threshold diagnostic plots for extreme value analyses, *Extremes*, 17, 289-303.

[15]- Pickands, III, J. (1975). Statistical Inference Using Extreme Order Statistics, *The Annals of Statistics*, 3, 119-131.

[16]- Reiss, R.-D. & Thomas, M. (2007). Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields, 3rd edition. Birkhäuser. Boston.

[17]- Scarrott, C. and MacDonald, A. (2012), A Review of Extreme Value Threshold Estimation and Uncertainty Quantification, *REVSTAT-Statistical Journal*, 10, 33-60.

[18]- Southworth, H. and Heffernan, J. E. (2013), *texmex: Statistical Modelling of Extreme Values*, R package version 2.1.

[19]- Strupczewski, W.G., Singh, V.P. and Feluch, W. (2001): Non-stationarity approach to at-site flood frequency modelling I. Maximum likelihood estimation. *J. Hydrol.* 248:123–142.

[20]- Tekin. ÖZTEKİN (2005). Comparison of parameter estimation methods for the three-parameter generalized Pareto distribution. *Turkish J. Agric. For.*,29(6): 419–428.

[21]- Thompson, P.; Cai, Y.; Reeve, D. & Stander, J. (2009). Automated threshold selection methods for extreme wave analysis, Coastal Engineering, 56, 1013–1021.

[22]- Xiangxian, Z. and Wenlei, G. (2009). A New Method to Choose the Threshold in the POT Model, ICISE(9), First International Conference on Information Science and Engineering.750-753

[23]- Zoi T. and John P. (2003). Extreme value index estimators and smoothing alternatives: A critical review.