UNIVERSITI TEKNOLOGI MARA

What's what PSPM

EISSN: 2756-7729

OKTOBER 2024/VOL 2

EXTREME EVENT ANALYSIS: ESTIMATING BOUNDARIES FOR DATA EXTREMITY

Zuraida Jaafar
Pengajian Sains Matematik , Kolej Pengajian Pengkomputeran, Informatik dan Matematik,
Universiti Teknologi MARA (UiTM), Cawangan Negeri Sembilan, Kampus Kuala Pilah, 72000,
Negeri Sembilan Darul Khusus, Malaysia.
zuraida@uitm.edu.my

1 INTRODUCTION

In practical applications of statistical modelling, the phenomena under investigation often involve many data points representing rare events with extremely high or low values compared to the typical range. These extreme events can significantly impact the data distribution, exhibiting long and heavy tails. The occurrence of extreme events can be observed across various disciplines, including climatology, earth sciences, ecology, engineering, hydrology, and social sciences. However, a critical question arises in extreme events analysis: How far can we reliably determine the extremity of data?

One of the most fundamental problems in the field of extreme value models is selecting a threshold value, a boundary or cutoff point used to determine the extremity of the data (McPhillips et al., 2018). The choice of the thresholds needs to be done properly, as a high threshold value will reduce the bias but increase the variance for the estimators while choosing a low value will give the opposite effect (Scarrott & MacDonald, 2012). Choosing the appropriate threshold value can help ensure that these extreme values are accurately identified and included in the analysis, leading to more accurate predictions and better decision-making.

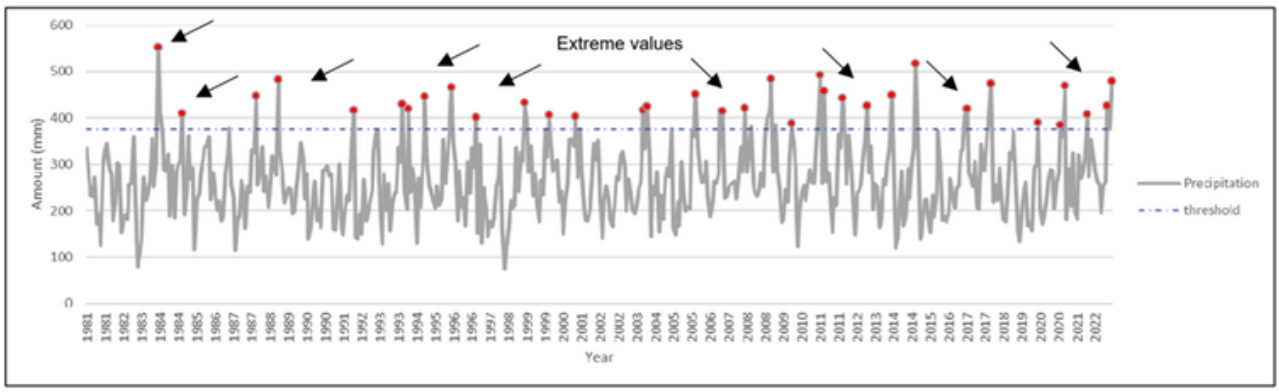

Figure 1:Extreme Values of Precipitation in Malaysia (1981-2022)

This article reviews the key historical approaches to estimating thresholds for extreme value analysis applications. It explores various methods developed over time to determine the appropriate threshold values for identifying and including extreme data points in statistical modelling and analysis.

2 SUMMARY OF ESTIMATION APPROACHES

2.1 GRAPHICAL DIAGNOSTICS

One basic concept in threshold selection is graphical diagnostic, which has been discussed deeply in Coles (2001), Kratz & Resnick (1996) and Drees et al. (2002). This approach focuses on analyzing the plot of the sample fraction, k, against the estimates of the tail index.

- Mean residual life plot;
  (A stable or approximately linear region in the plot suggests an appropriate threshold)
- Mean Excess plot;
  (Find a stable region in the plot, where the estimate becomes reliable)
- Q-Q plot;
  (A good threshold will produce a QQ plot that closely follows a straight line)
- Parameter threshold stability plot;
  (Look for a region where the shape and scale parameter estimates become relatively constant)

However, these methods are very subjective due to the necessarily personal interpretation of the plot. Hence, choosing each threshold manually makes the process burdensome.

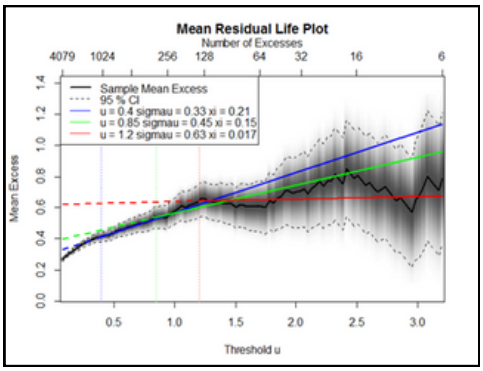

Figure 2a: Mean Residual Life plot

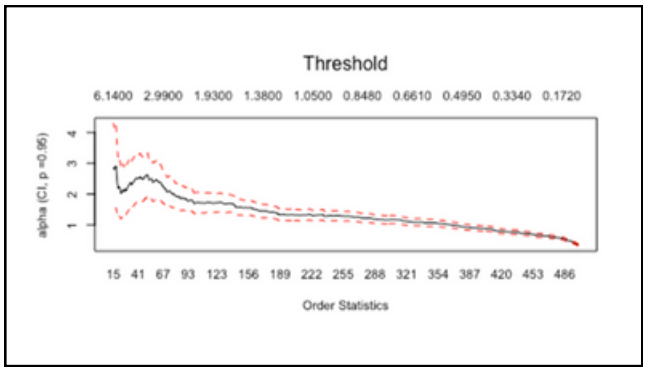

Figure 2b: Mean Excess plot

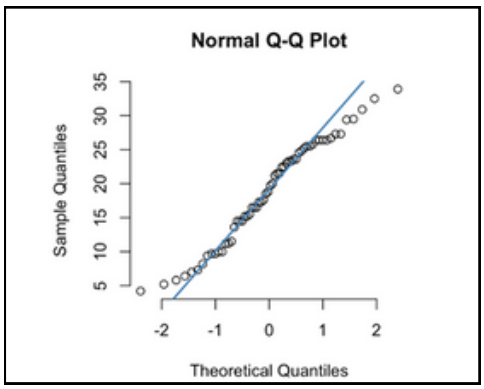

Figure 2c: Q-Q plot

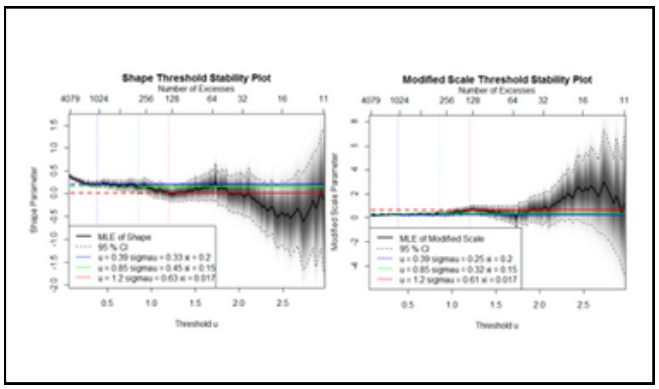

Figure 2d: Parameter Stability plot

2.2 RULES OF THUMB

The rule-of-thumb approach makes it easier to select the sample fraction of extreme events, which uses the upper 10% of the data (DuMouchel, 1983) or (Ferreira et al., 2003). Other than that, Ho & Wan (2002) work on the empirically driven rule of in their study of the stock returns.

2.3 PROBABILISTIC METHODS

Probabilistic methods which procedures aim for the optimal sample fraction for specific estimation, such as the Hill estimator (Hill, 1975). The Hill estimator is a classic tail index estimator for the Pareto-type distribution based on the upper-ordered statistics. Many researchers have found automated approaches to determine the tail fraction by, for example, minimizing the mean square error of estimators of properties of the tail distribution, such as the tail index (Beirlant et al., 1999), or the quantiles estimates (Ferreira et al., 2003). Besides, some study which compares the empirical distribution to the fitted Generalized Pareto distribution (GPD) via the goodness-of-fit test (Bader et al., 2018; Wadsworth, 2016) or by minimizing the distance between them (Clauset et al., 2009), where the latter approach is theoretically analysed in Drees et al. (2020).

2.4 COMPUTATION METHODS

Furthermore, several computation approaches aim to estimate the optimal sample fraction, k, which minimizes the asymptotic mean squared error of the Hill estimator. Hall (990) first proposed a resampling-based method for estimating k by minimizing the mean squared error. Drees & Kaufmann (998) utilize the Lepskii method and an upper bound on the maximum random fluctuation of $\hat{\gamma}k$ around γ. Danielsson et al. (2001) extended Hall's methodology to a double bootstrap method to identify optimal sample fractions. On the other hand, most methods related to minimizing mean squared error do not perform well in finite samples. Thus, to overcome this weakness, the usage of Kolmogorov-Smirnov measures has been used. Bickel & Sakov (2008) utilized the Kolmogorov-Smirnov distance metric in their bootstrap procedure to determine the difference between subsequently smaller subsample bootstrap distribution. Hence, this method inspired Danielsson et al. (2016) to use the Kolmogorov-Smirnov distance metric in their bootstrap procedure, which minimizes the distance between the tail of the empirical distribution and the fitted Pareto distribution. In 2021, Schneider et al. introduced a method that does not require users to manually choose tuning parameters in their computation and is easier. This method measures the fit of the exponential approximation above the threshold using integrated square error known as the Inverse Hill statistic.

2.5 MIXTURE MODELS

Another method that can be used to estimate the threshold values is using mixture models. A threshold value can be chosen automatically based on the separation between the bulk distribution and tail distribution (GPD). This approach considers all the observations regardless of whether it is extreme or not. Mixture models can be categorized by the type of bulk distribution models: parametric, semiparametric or nonparametric (Scarrott & MacDonald, 2012). The tail distribution over the threshold and the bulk distribution below the threshold can be simultaneously captured by extreme value mixture models. Without wasting any data, it takes into account every detail that are available. One of the main objectives of extreme value mixture model is to choose a flexible bulk model and tail model that simultaneously fits the non-extreme and extreme data.

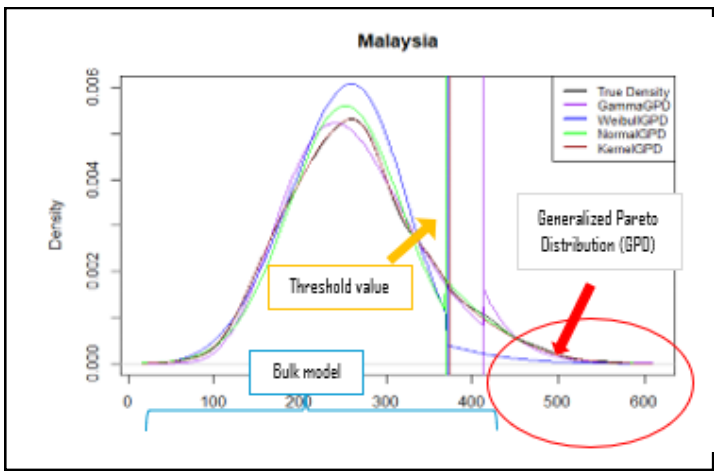

Figure 3: Extreme Value Mixture Model

3 CONCLUSIONS

Threshold selection is an important component in extreme event analysis. This paper provides several methods for determining appropriate thresholds, such as graphical diagnostics, rules of thumb, probabilistic methods, computational methods, and mixture models. Every approach has strengths and weaknesses, for example, graphical diagnostics provide quick intuition, but it can be subjective to interpret, and noisy data may make the plot harder to analyse. While probabilistic and computational methods offer more automated and objective solutions, they may be sensitive to a number of sample sizes and need complex computation. The optimal choice of threshold method requires a balance of accuracy and practicality in identifying extreme values and flexibility in applying the approach across different applications.

REFERENCES

1. Bader, B., Yan, J., & Zhang, X. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. The Annals of Applied Statistics, 12(1), 310–329. https://doi.org/10.1214/17-AOAS1092
2. Beirlant, J., Dierckx, G., Goegebeur, Y., & Matthys, G. (1999). Tail Index Estimation and an Exponential Regression Model. Extremes, 2(2), 177–200. https://doi.org/10.1023/A:1009975020370
3. Bickel, P. J., & Sakov, A. (2008). On The Choice of m In The m Out of n Bootstrap And Confidence Bounds For Extrema. Statistica Sinica, 18, 967–985. https://api.semanticscholar.org/CorpusID:12825582
4. Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-Law Distributions in Empirical Data. SIAM Review, 51(4), 661–703. https://doi.org/10.1137/070710111
5. Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. In Technometrics (Vol. 44, Issue 4). https://doi.org/10.1198/tech.2002.s73
6. Danielsson, J., de Haan, L., Peng, L., & de Vries, C. G. (2001). Using a Bootstrap Method to Choose the Sample Fraction in Tail Index Estimation. Journal of Multivariate Analysis, 76(2), 226–248. https://doi.org/https://doi.org/10.1006/jmva.2000.1903
7. Danielsson, J., ergun, lerby murat, de Haan, L., & de Vries, C. G. (2016). Tail Index Estimation: Quantile Driven Threshold Selection. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2717478
8. Drees, H., Jan\ssen, , Anja, Resnick, S. I., & Wang, T. (2020). On a Minimum Distance Procedure for Threshold Selection in Tail Analysis. SIAM Journal on Mathematics of Data Science, 2(1), 75–102. https://doi.org/10.1137/19M1260463
9. Drees, H., & Kaufmann, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation. Stochastic Processes and Their Applications, 75(2), 149–172. https://doi.org/https://doi.org/10.1016/S0304-4149(98)00017-9
10. Drees, H., Resnick, S., & de Haan, L. (2002). How to make a Hill plot. The Annals of Statistics, 28(1). https://doi.org/10.1214/aos/1016120372
11. Ferreira, A., Haan, L. de, & Peng, L. (2003). On optimising the estimation of high quantiles of a probability distribution. 37(5), 401–434.
12. Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. Journal of Multivariate Analysis, 32(2), 177–203. https://doi.org/https://doi.org/10.1016/0047-259X(90)90080-2
13. Hill B.M. (1975). A simple general approach to inference about the tail of a distribution. Annals of Statistics, 19(3), 1403–1433.
14. Ho, A. K. F., & Wan, A. T. K. (2002). Testing for covariance stationarity of stock returns in the presence of structural breaks: an intervention analysis. Applied Economics Letters, 9(7), 441–447. https://doi.org/10.1080/13504850110090210
15. Kratz, M., & Resnick, S. I. (1996). The qq-estimator and heavy tails. Communications in Statistics. Part C: Stochastic Models, 12(4), 699–724. https://doi.org/10.1080/15326349608807407
16. McPhillips, L. E., Chang, H., Chester, M. V., Depietri, Y., Friedman, E., Grimm, N. B., Kominoski, J. S., McPhearson, T., Méndez-Lázaro, P., Rosi, E. J., & Shafiei Shiva, J. (2018). Defining Extreme Events: A Cross-Disciplinary Review. Earth's Future, 6(3), 441–455. https://doi.org/10.1002/2017EF000686
17. Scarrott, C., & MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. Revstat Statistical Journal, 10(1), 33–60.
18. Schneider, L. F., Krajina, A., & Krivobokova, T. (2021). Threshold selection in univariate extreme value analysis. Extremes, 24(4), 881–913. https://doi.org/10.1007/s10687-021-00405-7
19. Wadsworth, J. L. (2016). Exploiting Structure of Maximum Likelihood Estimators for Extreme Value Threshold Selection. Technometrics, 58(1), 116–126. https://doi.org/10.1080/00401706.2014.998345
20. William H. DuMouchel. (1983). Estimating the Stable Index α in Order to Measure Tail Thickness: A Critique. The Annals of Statistics, 11(4), 1019–1031.