The Journal of Open Source Software

thresholdmodeling: A Python package for modeling excesses over a threshold using the Peak-Over-Threshold Method and the Generalized Pareto Distribution

Iago Pereira Lemos[1, 2, 3], **Antônio Marcos Gonçalves Lima**[4, 2, 3], **and Marcus Antônio Viana Duarte**[4, 1, 2, 3]

1 Acoustics and Vibration Laboratory **2** School of Mechanical Engineering **3** Federal University of Uberlândia **4** Associate Professor

DOI: 10.21105/joss.02013

Software

- Review
- Repository
- Archive

Editor: Vincent Knight

Reviewers:

- @bahung
- @kellieotto

Submitted: 06 January 2020
Published: 10 February 2020

Summary

Extreme value analysis has emerged as one of the most important disciplines for the applied sciences when dealing with reduced datasets and when the main idea is to extrapolate the observations over a given time. By using a threshold model with an asymptotic characterization, it is posible to work with the Generalized Pareto Distribution (GPD) (Coles, 2001) and use it to model the stochastic behavior of a process at an unusual level, either a maximum or minimum. For example, consider a large dataset of wind velocity in Florida, USA, during a certain period of time. It is possible to model this process and to quantify extreme events' probability, for example hurricanes, which are maximum observations of wind velocity, in a time of interest using the return value tool.

In this context, this package provides a complete toolkit to conduct a threshold model analysis, from the beginning phase of selecting the threshold, going through the model fit, model checking, and return value analysis. Moreover, statistical moments functions are provided. In case of extremes of dependent sequences it is also possible to conduct a declustering analysis.

In a software context, it is possible to see a strong community working with `R` packages like `POT` (Ribatet & Dutang, 2019), `evd` (Stephenson, 2018), and `extRemes` (Gilleland, 2019) that are used for complete extreme value modeling. In `Python`, it is possible to find the `scikit-extremes` (Correoso, 2019), which does not contain threshold models yet. Another package is `scipy`, which has the `genpareto` (Scipy, 2019) functions, but this does not provide any Peak-Over-Threshold modeling functions since it is not possible to define a threshold using this package. Moreover, this package brings to the community of scientists, engineers, and any other interested person and programmer, the possibility to conduct an extreme value analysis, using a strong, consolidated and high-level programming language, given the importance of the extreme value theory approach for statistical analysis in corrosion engineering (see Scarf & Laycock (1994) and Tan (2017)), hydrology (see Katz, Parlange, & Naveau (2002)), enviromental data analysis (see Rydman (2018) and Bommier (2014)) and many other fields of natural sciences and engineering. (For a large number of additional applications, see Coles (2001) p. 1.)

Hence, the `thresholdmodeling` package presents numerous functions to model the stochastic behavior of an extreme process. For a complete introduction to the complete fifteen package functions, it is crucial to go to the Functions Documentation on the GitHub page.

Package Features

Threshold Selection

- **Mean Residual Life Plot**: It is possible to plot the Mean Residual Life function as it is defined in Coles (2001);
- **Parameter Stability Plot**: Also, it is possible to obtain the two parameter stability plots of the GPD: the Shape Parameter Stability Plot and the Modified Scale Parameter Stability Plot, which is defined from a reparametrization of the GPD scale parameter. (See Coles (2001) for a complete theoretical introduction about these two plots.)

Model Fit

- **Fit the GPD Model**: Fitting a given dataset to a GPD model using some fit methods (see **Model Fit**).

Model Checking

- **Probability Density Function, Cumulative Distribution Function, Quantile-Quantile and Probability-Probability Plots**: Plots the theoretical probability density function with the normalized empirical histograms for a given dataset, using some bin methods (see `gpdpdf`). Also, the theoretical CDF in comparison to the empirical one with the Dvoretzky–Kiefer–Wolfowitz confidence bands can be drawn. In addition, The QQ and PP plots, comparing the sample and the theoretical values can be obtained, where the first uses the Kolmogorov-Smirnov Two Sample Test for getting the confidence bands while the second uses the Dvoretzky–Kiefer–Wolfowitz method;
- **L-Moments Plots**: L-Skewness against L-Kurtosis plot for a given threshold values using the Generalized Pareto parametrization. Be warned, L-Moments plots are really difficult to interpret. See Ribatet & Dutang (2019) and Hosking & Wallis (1997) for more details.

Model Diagnostics and Return Level Analysis

- **Return Level Computation and Plot**: Computing a return value for a given return period is also possible, with a confidence interval obtained by the Delta Method (Coles, 2001). Furthermore, a return level plot is provided, using the Delta Method in order to obtain the confidence bands. In order to compare, the empirical return level plot is provided.

Declustering and Data Visualization

It is possible to visualize the data during the unit of a return period. In case of extreme dependences sequences, for a given empirical rule (number of days, for example), it is possible to cluster the dataset and, taking the maximum observation of each cluster, a declustering of maximums is done.

Further Functions

It is also possible to compute sample L-Moments, model L-Moments, non-central moments, differential entropy, and the survival function plot.

Installation

For installation instructions, see the README on the GitHub page.

Reproducibility and User's Guide

The repository on the GitHub page contains a link to the dataset: Daily Rainfall in the South-West of England from 1914 to 1962. It can be used to test the software in order to verify its results and compare it with the forseen ones in Coles (2001). For a more detailed tutorial of using of each function, go to the Test directory.

A minimal simple example on how to use the software and get some of the results presented by Coles (2001) is given below. For information about the functions employed, see the Functions Documentation and for more details of reproducibility, see the README.

```
from thresholdmodeling import thresh_modeling
import pandas as pd

url = 'https://raw.githubusercontent.com/iagolemos1
/thresholdmodeling/master/dataset/rain.csv'
df = pd.read_csv(url, error_bad_lines=False)
data = df.values

thresh_modeling.MRL(data, 0.05)
thresh_modeling.return_value(data, 30, 0.05, 365, 36500, 'mle')
```

Fig. 1: Mean Residual Life Plot for the daily rainfall dataset.

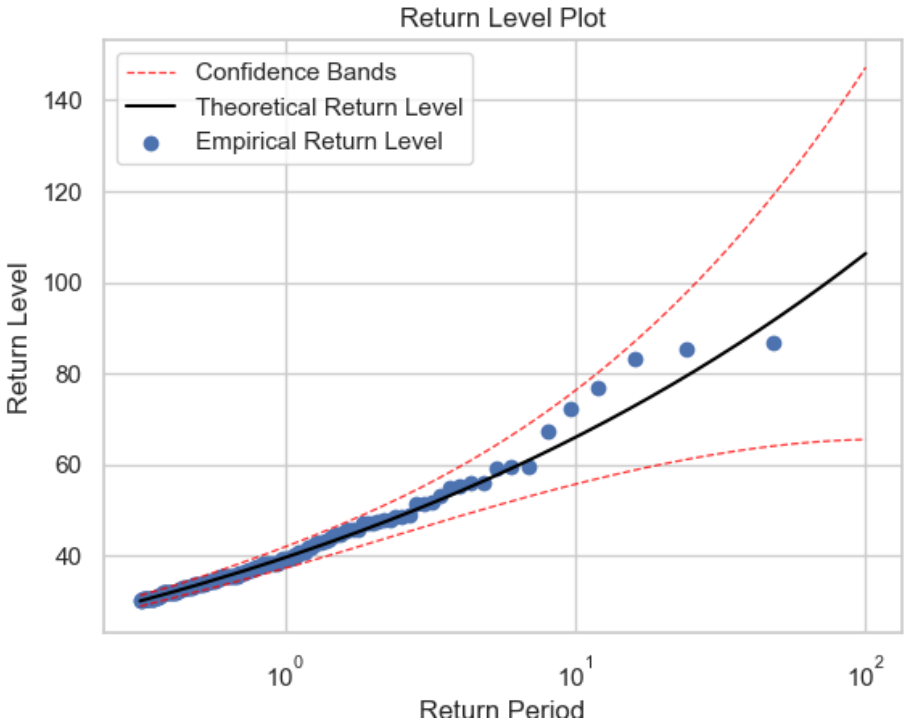

Fig. 2: Return level plot with the empirical estimatives of the return level and the confidence bands based on the Delta Method.

Also, for the given return period (100 years), the software presents the following results in the terminal:

```
The return value for the given return period is 106.3439 ± 40.8669
```

For more details, the documentation on the GitHub page is up-to-date.

Acknowledgements

The authors would like to thanks the School of Mechanical Engineering at Federal University of Uberlândia and CNPq and CAPES for the financial support to this research.

References

Bommier, E. (2014). *Peaks-Over-Threshold Modelling of Environmental Data*.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values* (1st ed.). London: Springer. doi:10.1007/978-1-4471-3675-0

Correoso, K. (2019). Scikit-extremes. Retrieved from https://github.com/kikocorreoso/scikit-extremes

Gilleland, E. (2019). *extRemes: Extreme Value Analysis*. Retrieved from https://cran.r-project.org/web/packages/extRemes/index.html

Hosking, J. R. M., & Wallis, J. R. (1997). *Regional Frequency Analysis: An Approach Based on L-Moments*. (1st ed.). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511529443

Katz, R. W., Parlange, M. B., & Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources*, *25*(8–12), 1287–1304. doi:10.1016/S0309-1708(02)00056-8

Ribatet, M., & Dutang, C. (2019). *POT: Generalized Pareto Distribution and Peaks Over Threshold*. Retrieved from https://cran.r-project.org/web/packages/POT/index.html

Rydman, M. (2018). *Application of the Peaks-Over-Threshold Method on Insurance Data*.

Scarf, P. A., & Laycock, P. J. (1994). Applications of Extreme Value Theory in Corrosion Engineering. *Journal of Research of the National Institute of Standards and Technology*, *99*(4), 313–320. doi:10.6028/jres.099.028

Scipy. (2019). Scipy.stats.genpareto. Retrieved from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.genpareto.html

Stephenson, A. (2018). *evd: Functions for Extreme Value Distributions*. Retrieved from https://cran.r-project.org/web/packages/evd/index.html

Tan, H.-Y. (2017). *Analysis of Corrosion Data for Integrity Assessments* (Thesis for the Degree of Doctor of Philosophy). Brunel University London.