Analysis of precipitation data in Rio de Janeiro city using Extreme Value Theory

Augusto J. M. da Fonseca[1], **Fabio Porto**[2], **Mariza Ferro**[2,3], **Eduardo Ogasawara**[1], **Eduardo Bezerra**[1]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ) Rio de Janeiro, Brazil.

[2]Laboratório Nacional de Computação Científica (LNCC) Getúlio Vargas, 333 – Petrópolis – RJ – Brazil

[3]Instituto de Computação Universidade Federal Fluminense (UFF) – Niterói, RJ – Brazil

augusto.fonseca@eic.cefet-rj.br, {fporto,mariza}@lncc.br,

eogasawara@ieee.org, ebezerra@cefet-rj.br

Abstract. *Every year the city of Rio de Janeiro is hit by heavy rains that cause several disasters. To be able to reduce potential damages that may arise due to extreme rainfall, it is important to make accurate estimates of such events. In this paper, we present preliminary work on using Extreme Value Theory (EVT) to model rainfall data in Rio de Janeiro municipality. We use twenty-five years of historical data coming from rain gauges spread throughout the city. We investigate the behavior of both approaches in EVT to identify extreme values, namely, Block Maxima and Peeks-over-Threshold. After determining the best sampling distributions given the available data, we present an analysis of their goodness of fit and of their corresponding return period plots.*

1. Introduction

The city of Rio de Janeiro is located in the southeastern region of Brazil and is annually hit by intense rains that cause disasters such as landslides on slopes, floods and flooding. Given this scenario, forecast models become fundamental elements in preventing and mitigating the social impacts caused by heavy rains. To this end, the Rio de Janeiro municipality works in collaboration with `Alerta Rio`[1], a monitoring, forecasting and warning system for the possibility of heavy rains, landslides, floods, among other geotechnical accidents. The Alerta Rio system has a network of telemetry stations (rain gauges and meteorological) and three weather radars. Using observations generated by these sensors, along with some numerical weather forecasting models, the team of Alerta Rio's meteorologists generates daily weather alert bulletins to the population.

In this paper, we describe a work in progress regarding the application of Extreme Value Theory to model extreme precipitation events in Rio de Janeiro. We use data coming from Alerta Rio's automatic rain gauge stations. For each station, we fit a distribution model to the detected extreme events. After finding a maximum likelihood estimate for the appropriate distribution, we present an aggregated analysis at various return periods.

[1]`http://alertario.rio.rj.gov.br/institucional/quem-somos/`

We organize this paper as follows. We provide some background on Extreme Value Theory in Section 2. We describe related work in Section 3. Then in Section 4 we describe our approach to applying EVT to the available data. In Section 5 we present the results so far and a corresponding analysis. In Section 6, we conclude by providing some final remarks and pointing to future work.

2. Background

By definition, an extreme (i.e., rare) value is one that has a low probability of occurrence, but with a high impact. Extreme values can be found in either the left or right tail of a probability distribution. EVT is a statistical technique that studies the limiting distributions for the minimum or the maximum of a large collection of independent and identically distributed random variables (Davison and Huser, 2015). The goal is to produce an smoothing and extrapolation model of the tails of a distribution that can be used to estimate the probabilities of extreme events. Given an original times series from which a set $\{X_1, \ldots, X_n\}$ of extreme values are identified, the purpose of EVT is to study the behavior of the statistic $M_n = \max\{X_1, \ldots, X_n\}$ as $n \to \infty$ (the $\max$ operator is replaced by $\min$ when the extreme values are in the lower side of the observation spectrum).

The pipeline to apply EVT is comprised of two steps. The first one has to do with identifying the sample of extreme values. The second one involves fitting an appropriate probability distribution to that sample. In the first step, the set $\{X_1, \ldots, X_n\}$ must be identified. EVT defines two approaches to identify extreme values, namely *Peaks over Threshold* (PoT) and *Block Maxima* (BM). PoT identifies as extreme all the observations greater than a predefined threshold τ. BM first construct a sequence of the maxima obtained by segregating the original series into blocks of a predefined size (e.g., daily, monthly, annually). The choice of the identification method is domain dependent. In the second step, we fit a probability distribution to the ordered set of random variables containing the extreme values. One result of EVT states that, as $n \to \infty$, M_n will follow the so called *Generalized Extreme Value* distribution: $M_n \sim \mathrm{GEV}(\mu, \sigma, \xi)$. GEV parameters are μ (location), σ (scale), and ξ (shape) (Davison and Huser, 2015). When the BM approach is used to select the sample of extreme values, three distributions are subsumed by the GEV distribution. These are the Gumbel ($\xi = 0$), Frechet ($\xi > 0$) and Weibull ($\xi < 0$) distributions. The choice of which one to use on the fitting procedure depends on the shape of the parent distribution being heavy tailed ($\xi > 0$), light tailed ($\xi = 0$), or very light tailed ($\xi < 0$). When the PoT approach is used, GEV is subsumed by the Generalized Pareto and Exponential distributions. The values of these parameters can be estimated from the sample of extreme values by using some estimation technique (e.g., Maximum Likelihood Estimation or Method of Moments). After the fitting procedure is finished, the resulting sampling distribution can be used to answer queries concerning the probabilities of extreme values.

3. Related work

The Mann–Kendall nonparametric test and the Sen's Curvature were employed to evaluate the significance and magnitude of precipitation in next two works. Luiz-Silva (2022) presents a synthesis of the main characteristics of precipitation in the State of Rio de Janeiro based on extreme rainfall indicators. They show that there is a growth of accumulated extreme precipitation in various stations near the ocean and display an increase in the extreme rainfall in 24 h in most Rio de Janeiro (+ 1 to + 5 mm/decade). Silva and Dereczynski (2014) elaborated an analysis of trends in climate extremes of air temperature and precipitation for the State of Rio de Janeiro, using observational data between 1961 and 2012. Results showed a significant increase in rainfall totals of more heavy rains in the year in the Baixadas Litorâneas and in part of the Metropolitan Region, with magnitudes (between +2.0 and +20 mm/year). Lyra et al. (2018) evaluated different methods of spatial interpolation methods of monthly rainfall over the state of Rio de Janeiro. This work is not focused only on extreme precipitations but the evaluated methods are also compared using frequency and extreme values analysis showing that all methods underestimated heavy rainfall.

Regarding approaches based on extreme value theory, in Lima et al. (2021) seven probability distribution functions were fitted to annual maximum daily rainfall series in Rio de Janeiro State. Time series data were obtained from 110 stations with a data coverage of at least 20 years, from 1960 to 2010. The best-fit distributions, Gumbel and GEV, were used to estimate extreme rainfall with different recurrence intervals. Results showed that the regions with higher values of extreme rainfall in all scenarios were the ones close to the coast, within 40 km, and south of Serra dos Órgãos mountain range, located in the middle of the state. The mountain range separates the state in two halves, concentrating higher values of extreme rainfall in the lower part, where the city of Rio de Janeiro is located. Scenarios for GEV and Gumbel indicated daily rainfall events up to 200 mm, with recurrence intervals of 50 to 100 years. dos Reis et al. (2022) aims to find probabilities of extreme values of the air temperature for the Cerrado, Pantanal and Atlantic Forest biomes in Mato Grosso do Sul in Brazil. Using the EVT approach this work estimates three probability distributions (GEV, Gumbel and the Log-Normal). The distributions have been fitted with monthly data used to estimate extreme levels of maximum temperatures. Álvaro José Back and Bonfante (2021) compares GEV and Gumbel distributions for estimating maximum daily rainfall in Santa Catarina (Brazil). They analyzed data of 224 series of maximum annual rainfall data ranging from 12 to 90 years. Results allow conclusions on parameter adjustment and selection of probability distributions to estimate maximum extreme rainfall. Ban et al. (2020) present an analysis of extreme precipitation events in convection-resolving climate simulations of Alpine region. Generalized extreme value theory was applied to address projections of 5-day, daily and hourly extreme precipitation events in all seasons.

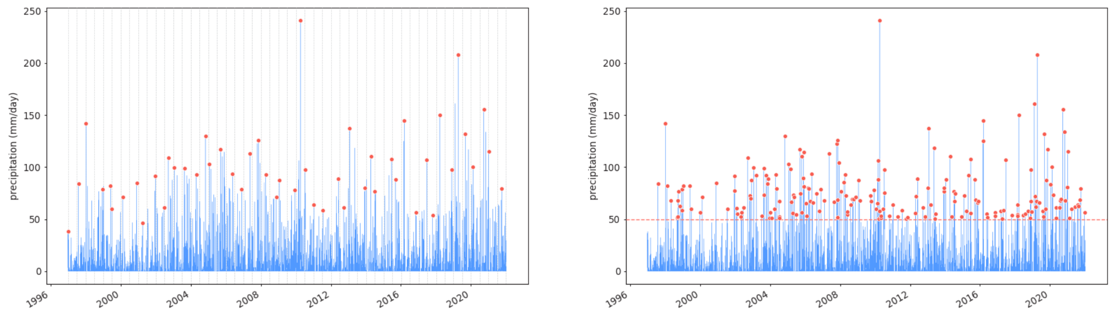
To the best of our knowledge, there is no work regarding the application of Extreme Value Theory to model extreme precipitation events specially for Rio de Janeiro city.

4. Method

Alerta Rio makes the observations of their thirty-three rain gauge stations publicly available[2]. From these time series, we consider twenty-seven, since six of them did not provide observations for the whole period (1997-2021). As a first preprocessing step, we aggregated (i.e., summed) the time series using a daily temporal resolution (the original time series have a fifteen-minutes temporal resolution).

We then applied the EVT pipeline separately to each time series generated in the

[2] http://alertario.rio.rj.gov.br/download/



(a) BM extremes **(b) PoT extremes**

Figure 1. Identified extreme precipitation values for Rocinha station.

preprocessing activity. In the first step of the pipeline, we decided to investigate the use of both approaches, BM and PoT. For the Block Maxima approach, we chose to use a semester as the block size, instead of usual size of a year, to produce a greater sample of extreme values for each station. As a result, we obtained a sample of 50 (25 years $\times$ 2 semesters) extreme values for each station. For PoT, this approach needs the specification of a threshold τ that is neither too high (to get enough observations) nor too low (not to take into account non-extreme values). In this paper, following the recommendation of Alerta Rio's meteorologists, we set $\tau = 50$ mm/h, since this value is the threshold they defined for rainstorms[3]. The resulting number of identified extreme values in this case was 196. Figure 1 displays the extreme values identified by each approach. In the second step (fitting the parameters of the sampling distribution), we used Maximum Likelihood Estimation as the estimation technique.

5. Results and discussion

To produce the results presented in this section, we used pyextremes[4], a Python library that provides tools necessary to perform several tasks required to perform an EVT pipeline. For lack of space, we present results only for the Rocinha station. (Corresponding plots for the other stations can be found at `https://github.com/MLRG-CEFET-RJ/dexea22-evt`). The rain gauge situated at the studied station recorded an average annual rainfall of 1730.37 mm, slightly 46% more than the city average of 1187.75 mm.

By running MLE, we selected the exponential distribution for the PoT extremes, with scale parameter 26.80. For the BM extremes, the Gumbel distribution was selected, with location and scale parameters equal to 82.62 and 27.79, respectively. The maximized log-likelihoods for PoT and BM were -840.560 and -245.473, respectively. Figure 2 and Figure 3 summarize the obtained results for PoT and BM cases, respectively. In the right upper corner of these figures is the plot of the corresponding empirical sampling distribution fitted from the data.

The lower half of both figures show the corresponding QQ-plot and PP-plot, that provide a diagnostic of the quality of fit in each case. The visual inspection of these plots give us confidence about the quality of the fitted models, since the points (show in black)

[3] The complete classification: Light ($[0, 5)$), Moderate ($[5, 25)$), Heavy ($[25, 50)$), Rainstorm ($[50, \infty)$).
[4] `https://georgebv.github.io/pyextremes/`

appear to be nearly the straight line (shown in blue). The goodness of fit is also confirmed by the values of the coefficient of determination R^2 and the p-value in both cases.

The upper-left corner in both figures shows a return value plot. This plot can be interpreted in the following way: for a given value t in the horizontal axis, there is a probability of $1/t$ that, in any given semester, an extreme precipitation such as the one corresponding to t will happen. These plots also show 95% confidence intervals (in shaded blue).

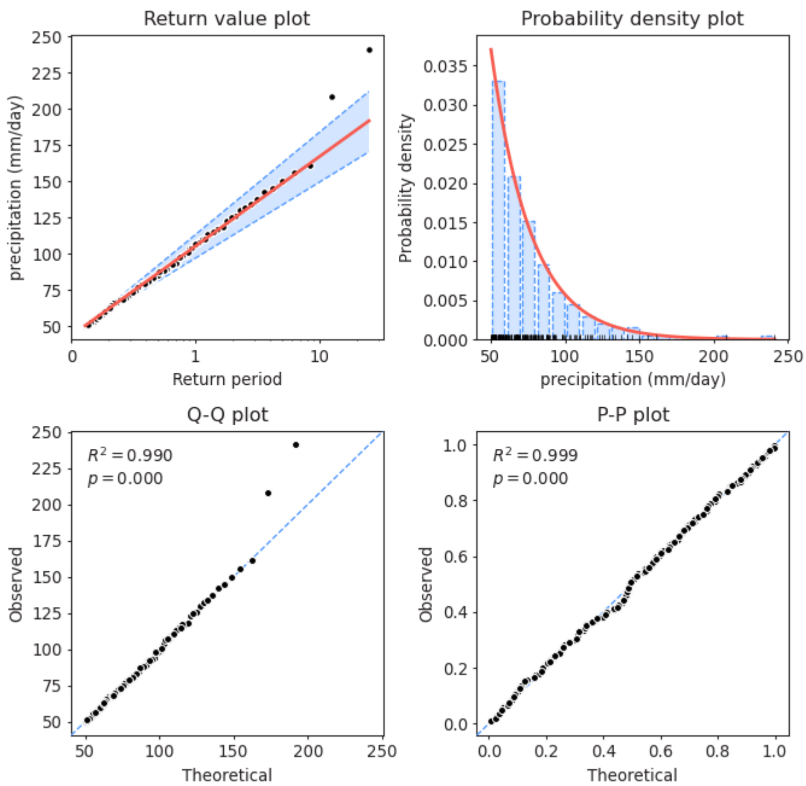


Figure 2. Diagnostics plots for the Rocinha station (PoT version).

6. Final remarks

Assessing the probability of extreme events is an important issue in meteorology. The aim of this work was to study extreme precipitation events in Rio de Janeiro City by using EVT. The preliminary results we present in this paper confirm that Extreme value Theory provides the needed framework for statistical modeling of such events. As future work, we plan to expand our analysis to consider other perspectives of extreme events, such as duration in time. We also plan to investigate data-driven ways to define the threshold τ in the PoT method, such as use statistical tests. Another subject we plan to investigate is modeling extreme values observed in a spatiotemporal setting with multiples data sources (e.g., radar, satellite), a common case in weather models.

References

Ban, N., Rajczak, J., Schmidli, J., and Schär, C. (2020). Analysis of alpine precipitation extremes using generalized extreme value theory in convection-resolving climate simulations. *Climate Dynamics*, 55:61–75.

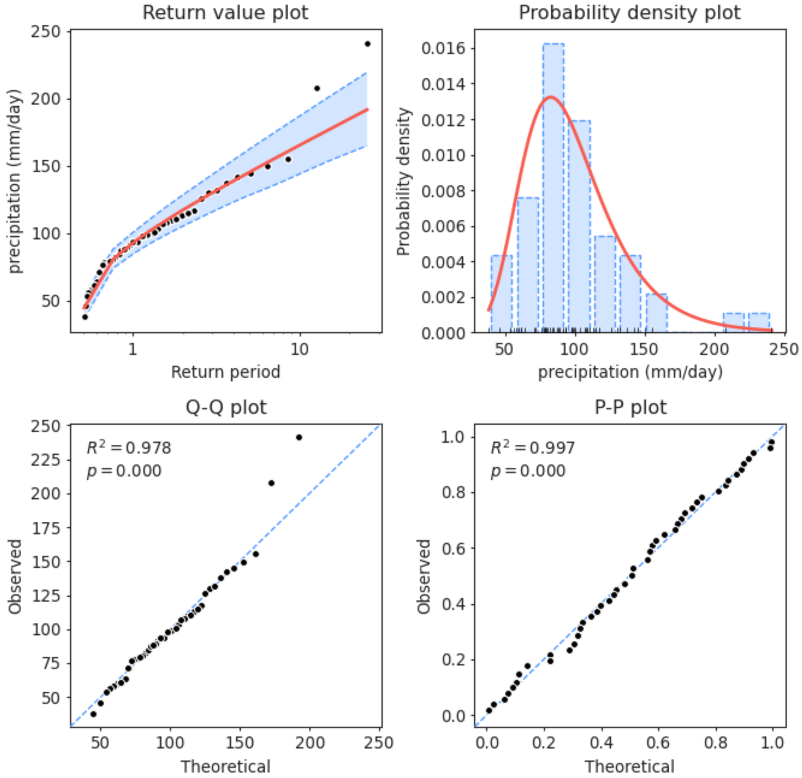


Figure 3. Diagnostics plots for the Rocinha station (BM version).

Davison, A. and Huser, R. (2015). Statistics of extremes. *Annual Review of Statistics and Its Application*, 2(1):203–235.

dos Reis, C. J., Souza, A., Graf, R., Kossowski, T. M., Abreu, M. C., de Oliveira-Júnior, J. F., and Fernandes, W. A. (2022). Modeling of the air temperature using the extreme value theory for selected biomes in mato grosso do sul (brazil). *Stochastic Environmental Research and Risk Assessment*.

Lima, A. O., Lyra, G. B., Abreu, M. C., Oliveira-Júnior, J. F., Zeri, M., and Cunha-Zeri, G. (2021). Extreme rainfall events over Rio de Janeiro State, Brazil: Characterization using probability distribution functions and clustering analysis. *Atmospheric Research*, 247:105221.

Luiz-Silva, W.and Oscar-Júnior, A. (2022). Climate extremes related with rainfall in the state of rio de janeiro, brazil: a review of climatological characteristics and recorded trends. *Nat Hazards*.

Lyra, G., Correia, T., Oliveira-Júnior, J., and Zeri, M. (2018). Evaluation of methods of spatial interpolation for monthly rainfall data over the state of rio de janeiro, brazil. *Theoretical and Applied Climatology*, 134.

Silva, W. and Dereczynski, C. (2014). Caracterização climatológica e tendências observadas em extremos climáticos no estado do rio de janeiro. *Anuário do Instituto de Geociências UFRJ*, 2(37):123–138.

Álvaro José Back and Bonfante, F. M. (2021). Evaluation of generalized extreme value and gumbel distributions for estimating maximum daily rainfall. *Brazilian Journal of Environmental Sciences*, 56(4):654–664.