Analysis of extreme data values using the distribution of extreme generalized (GEV)

Wyara Vanesa Moura e Silva
ICV: Voluntary Scientific Initiation (UFPI); Date: August 29, 2013

1. Introduction

The world is undergoing environmental changes increasingly stringent in recent years. Such alterations cause disasters, which cause great losses, both material and financial, leading to concern from the authorities, to seek ways to minimize the effects of these events. Events of this size, because they are of low frequency and high severity are difficult to predict, there appeared studies related to extreme data to try to predict such disasters, in order to avoid or reduce losses.

2. Objectives

Propose a new model for estimating various types of Piaui state of extreme data, such as rivers quotas. In order to propose the model, we use the knowledge that the observations of the maxima of this type of data can be well approximated by the distribution of generalized extreme values (GEV). We will find probability of occurring natural disasters in the studied region, besides estimation of parameters for series with behavior change after modifications in the course of the rivers analyzed.

The analyzed data represent maximum values of fluviometric quotas of the river Parnaíba, located in the northeastern region of Brazil, in the state of Piauí. And data of maxima of the Paraná river located between the states of So Paulo, Minas Gerais and Mato Grosso do Sul. The analysis is based on hydrological data of water levels (fluviometric level) from the hydro meteorological network of the National Water Agency (ANA), using the Hydrological Information System (Hidro Web).

3. Methodology

3.1 Extreme value theory

The Extreme Values Theory (EVT) is an area of probability that studies the stochastic behavior of extremes associated with a set of random variables (or random vectors) with common distribution $\mathcal{F}$. Included in the general name of extremes aggregates are the maximum and minimum order statistics and extreme excesses above (or below) the threshold high (or low).

The theorem proposed by Fisher-Tippett (1928) has the characteristic of providing for maximum distribution limit in which these data are collected in block size n. Such distribution of extreme values are known as Type I (Gumbel), type II (Frchet) and Type III (Weibull). These three types of distributions (I, II, III) can be seen as members of a single family of distributions: a generalized extreme value distribution (Generalized Extreme Value, GEV) standard, which is the representation of the three extreme value distributions . Its density ($\mu = 0$) is given by:

$$h(y|\xi,\sigma) = \begin{cases} \exp\left\{-\left(1+\xi\frac{y}{\sigma}\right)^{-\frac{1}{\xi}}\right\}\frac{1}{\sigma}\left(1+\xi\frac{y}{\sigma}\right)^{-\frac{1}{\xi}-1} : \\ \text{if } \xi < 0 \ e \ -\infty < y < (-\sigma/\xi)\,; \\ \text{or } \xi > 0 \ e \ y \geq (-\sigma/\xi) \\ \exp\left\{-\exp\left\{-\frac{y}{\sigma}\right\}\right\}\frac{1}{\sigma}\exp\left\{-y\right\} : \\ \text{if } \xi = 0 \\ \text{and } y \in \mathbb{R}. \end{cases}$$

The shape parameter ξ can be used to model a large number of behaviors tails.

4. Change Point Model

Csörgõ and Horvth (1997) have performed studies concerning parameter changes in a distribution, based on a sample X_1, X_2, ..., X_{m*}, X_{m*+1}, ..., X_n, with density function f(x;θ_i,η), in order to test if θ_i was altered in some point $m*$.

In this work, the proposed model aims to find changes that occurred in the parameters of a serie that have two types of GEV distribution .

A simple formulation of the point change problem in a series is given as follows.

Let us assume that the densities $h_1(x)$ and $h_2(x)$ belong to a known parametric class $\mathbf{H} = h(x|\boldsymbol{\theta})$; $\boldsymbol{\theta} \in \Theta$ indexed by an unknown parameter $\boldsymbol{\theta}$ such that for a sequence of n independent random variables $\mathbf{X} = (x_1,\, x_2,\, ...,\, x_n)$, we have

$$\begin{array}{rclr} X_i \sim h_1(x) & = & h(x_i|\boldsymbol{\theta_1}), \quad i = 1, ..., \tau_1 & \\ X_i \sim h_2(x) & = & h(x_i|\boldsymbol{\theta_2}), \quad i = \tau_1 + 1, ..., n & \end{array} \tag{1}$$

where $\boldsymbol{\theta_1} \neq \boldsymbol{\theta_2}$ and τ_1 = 1, 2, ..., n - 1 is the unknown parameter, called the point of change (Perreault, 2000).

That is, the first and second part of the sequence of random variables belong to the same family of statistical distributions, but with different unknown parameters $\boldsymbol{\theta}$.

In which $\boldsymbol{\theta}$, specifically in this work, are the location (μ), scale (σ) and tail (ξ) parameters of the GEV distribution .

The resulting likelihood function of n observations $\mathbf{X} = (x_1,\, x_2,\, ...,\, x_n)$ generated by model in (1) becomes

$$L(\boldsymbol{X}|\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \tau_1) \;=\; \prod_{i=1}^{\tau_1} h(x_i|\boldsymbol{\theta_1}) \prod_{i=\tau_1+1}^{n} h(x_i|\boldsymbol{\theta_2})$$

in which, for example, the maximum likelihood estimator of $\boldsymbol{\theta_1}$, $\boldsymbol{\theta_2}$, τ_1 can be obtained.

Regarding the parameters of the GEV distribution of which it is part of the series under analysis, in which they will be two GEV, due to the existence of the point of change, estimation is based on the Bayesian approach, estimated by MCMC (Gamerman and Lopes, 2006).

The prior joint distributions of the $\boldsymbol{\theta_i} = (\mu_i, \sigma_i, \xi_i)$ parameters are given by:

$$\begin{aligned} p(\mu_1, \sigma_1, \xi_1) \;&\propto\; \sigma_1^{a-1} \exp(-b\sigma_1) \exp(-\frac{(\mu_1 - \mu_{1,0}^2)^2}{2\sigma_{1,\mu_1}^2} \\ &\quad \times \exp(-\frac{(\xi_1 - \xi_{1,0}^2)^2}{2\sigma_1^{2,\xi_1}} \\ p(\mu_2, \sigma_2, \xi_2) \;&\propto\; \sigma_2^{a-1} \exp(-b\sigma_2) \exp(-\frac{(\mu_2 - \mu_{2,0}^2)^2}{2\sigma_{2,\mu_2}^2} \\ &\quad \times \exp(-\frac{(\xi_2 - \xi_{2,0}^2)^2}{2\sigma_2^{2,\xi_2}} \end{aligned}$$

Then we obtain the distribution proportional to a posteriori distribution, taking

$$\begin{aligned} \pi(\mu_1, \sigma_1, \xi_1) \;&\propto\; p(\mu_1, \sigma_1, \xi_1) l(\mu_1, \sigma_1, \xi_1) \\ \pi(\mu_2, \sigma_2, \xi_2) \;&\propto\; p(\mu_2, \sigma_2, \xi_2) l(\mu_2, \sigma_2, \xi_2) \end{aligned}$$

After algebraic manipulations, it can be seen that the posterior distribution has neither a known form nor its complete conditional distribution. Hence the possibility of using the Metropolis-Hastings algorithm to sample points of the posterior distribution in order to find simple solutions for this type of problem.

For the parameter referring to the point of the change in distributions, given by τ_1, will be obtained by means of the probability:

$$p(\tau|\boldsymbol{X}, \boldsymbol{\theta_i}) \;=\; \frac{L(\boldsymbol{X}; \tau, \boldsymbol{\theta_i})}{\sum_{j=1}^{n} L(\boldsymbol{X}; j, \boldsymbol{\theta_i})}, \qquad i = 1, 2$$

in which due to limitations of the software used to build the algorithm such probability became

$$\begin{aligned} & p(\tau|\boldsymbol{X}, \boldsymbol{\theta_i}) \\ & \quad = \quad \frac{\exp\{L(\boldsymbol{X}; \tau, \boldsymbol{\theta_i})\}}{\displaystyle\sum_{j=1}^{n} \exp\{\min(L(\boldsymbol{X}; j, \boldsymbol{\theta_i})) - L(\boldsymbol{X}; j, \boldsymbol{\theta_i})\}} \end{aligned} \tag{2}$$

where $i = 1, 2$, and instead of $L(\boldsymbol{X}; j, \boldsymbol{\theta_i})$, we apply the logarithm in the likelihood function, making $l(\boldsymbol{X}; j, \boldsymbol{\theta_i})$, making it easier to process the algorithm.

In that $l(\boldsymbol{X}; j, \boldsymbol{\theta_i})$ will be given by

$$\begin{aligned} l(\tau|\boldsymbol{X}, \boldsymbol{\theta_i}) \;&=\; \sum_{j=1}^{\tau} \log\left(\prod_{i=1}^{\tau} h(x_i|\theta_i)\right) \\ &\quad + \sum_{j=\tau+1}^{n-1} \log\left(\prod_{i=\tau+1}^{n-1} h(x_i|\theta_i)\right) \end{aligned}$$

The sampling was done using Metropolis-Hastings to estimate the parameters of the GEV functions, and the probability condition to estimate the point of change. The parameters of each GEV were estimated separately, $\boldsymbol{\theta_i} = (\mu_i, \sigma_i, \xi_i)$.

Details of the MCMC sampling scheme are given below. In iteration s, the parameters are updated as follows:

Sampling $\boldsymbol{\theta_i}$ $(= \mu_i, \sigma_i, \xi_i)$: where i, indicates the amount of GEV, according to how many change points in the series. To estimate the parameters an initial kick will be given initially for $\boldsymbol{\theta_i^0}$ $(= \mu^0, \sigma^0, \xi^0)$, then a new value σ' of the Gamma distribution $(\sigma^{(t)2}/V_\sigma, \sigma^{(t)}/V_\sigma)$, μ' the Normal distribution $(\mu^{(t)}, V_\mu)$ and ξ' e ξ' the Normal distribution $(\xi^{(t)}, V_\xi)$.

Thus, the probability of acceptance $\alpha(\boldsymbol{\theta^{(t)}}, \boldsymbol{\theta'})$ will be calculated as in

$$\alpha(\boldsymbol{\theta^{(t)}}, \boldsymbol{\theta'}) \quad = \quad \min\left\{1, \frac{\pi\left(\boldsymbol{\theta'}\right) f_G\left(\sigma^{(t)}|\dfrac{\sigma^{'2}}{V_\sigma}, \dfrac{\sigma'}{V_\sigma}\right)}{\pi(\boldsymbol{\theta^{(t)}}) f_G\left(\sigma'|\dfrac{\sigma^{(t)2}}{V_\sigma}, \dfrac{\sigma^{(t)}}{V_\sigma}\right)}\right\}$$

and generate $u \sim U(0,1)$. If $u \leq \alpha(\boldsymbol{\theta^{(t)}}, \boldsymbol{\theta'})$ then accept the new value and make $\mu^{(t+1)} = \mu'$, $\sigma^{(t+1)} = \sigma'$ and $\xi^{(t+1)} = \xi'$. Otherwise reject and make $\mu^{(t+1)} = \mu^{(t)}$, $\sigma^{(t+1)} = \sigma^{(t)}$ and $\xi^{(t+1)} = \xi^{(t)}$.

Sampling τ: Then, the first kick will be given to the point of change of the series. Then they will be updated according to the probability, $p(\tau|\boldsymbol{X}, \boldsymbol{\theta_i})$ data in (2).

Thus, we will sample a value for the change point with probability as in (2).

5. Results

We analyzed historical series corresponding to maximum levels of quotas (in cm), of the rivers: Parnaíba and Paraná.

The Table 1 below shows the information about these historical series.

Table 1: Descriptive informations about the time series.

River	City	Period	n
Parnaíba	Teresina	07/01/1963 to 08/01/2012	413
Paraná	Guaíra	06/01/1920 to 10/01/2012	1100

5.1 Parnaíba River

The temporal series of maximum data for the Parnaíba River is shown below in Figure 5. It can be observed in the time series graph an increase in the maximum level of quotas, the period in which this increase occurred since 1970, the period from 1970 to 1980 not all the quota data has been recorded.

In 1970, it was the inauguration year of the Boa Esperança Dam, in the municipality of Guardalupe, which was built with the objective of promoting the hydraulic exploitation of the Parnaíba River. What is observed after its inauguration was that there was an increase in the maximum quotas, but we do not have the true knowledge of the real reasons for such a change of maximum levels.

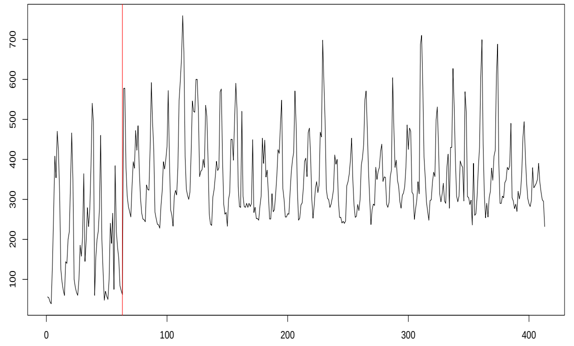


Figure 1: Time Series of Parnaíba River

The proposed point change model, using Bayesian approach, initial kick $k = 100$ resulted in plot in Figure 2 of the probabilities of each number to be the point that the change occur.

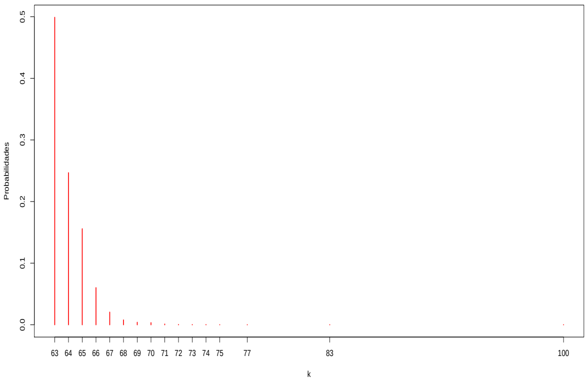


Figure 2: Change-Point probabilities, Parnaíba River

In that in point 63 that obtained the highest probability of occurrence of the change, dated 09/01/1969. From the proposed model, we also estimate the parameters of the GEV distributions taking into account the point of change. The estimated parameters are shown in Table 2.

Table 2: Estimates of the parameters, for points $\tau \leq 63$ e $\tau > 63$.

	$\tau \leq 63$	
Parameters	Estimates	95% Credible Interval
Location (μ)	104.7935	[84.44076 ; 126.2104]
Scale (σ)	76.4508	[55.6550 ; 102.4004]
Shape (ξ)	0.4828	[0.1101 ; 0.9018]
	$\tau > 63$	
Paramters	Estimates	95% Credible Interval
Location (μ)	307.0897	[300.1058 ; 314.4478]
Scale (σ)	63.1190	[57.4499 ; 69.2535]
Shape (ξ)	0.2577	[0.1480 ; 0.3688]

In this way, we can see that there was an increase in the lease parameter μ, since the σ scale and form ξ there was a decrease.

Analyzing the return graphs shown in Figure 3, after the change point, it can be seen that for $\boldsymbol{\tau > 63}$, it is observed that in a period corresponding to every 1 year and 10 years , the maximum level of dimensions is expected to be 515.5729 cm and 691.5289 cm respectively.

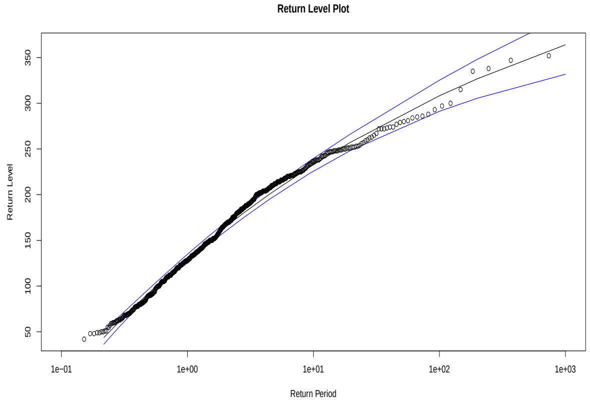
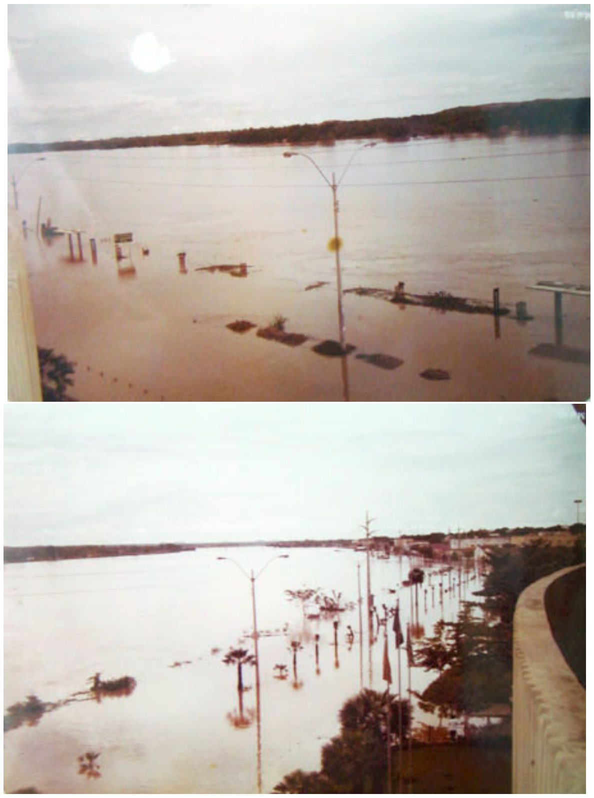


Figure 3: Returns Levels, Parnaíba River

An example of such returns for 10 years occurred in May 2009 when CHESF registered a maximum of 688cm at the monthly quota level in the Parnaíba River, which caused severe flooding and impacts to society and the region's ecosystem.

More severe extremes are likely to occur in a period of 100 years, with maximum heights between 770.7334 cm and 975.0603 cm. Such an occurrence of these extremes was witnessed by the capital of Piauí - Teresina, in 1985, in which one of the busiest avenues of the cities that stood by the river flooded. As shown in Figure 4:

Figure 4: Flood, Maranhão Avenue - Teresina - Piauí



The photos above are from Valmira Cabral, who runs the Vaqueiro Museum in the Alto Longá municipality. In 1985, she worked at Companhia Energtica do Piauí - Cepisa, and recorded the damage in the Teresina Center from the top of the building on Maranhão Avenue.

Floods have always been considered problems of great public concern, since they cause great damages, which compromise the quality of life of the population of Piau, especially in the riverside regions.

5.2 Paraná River

We can observe the graph of the temporal series of maximums of the Paraná River, which is found in Figure 6.5, the occurrence of a point of change just as there was in the Rio Parnaíba. This point of change occurred approximately after the construction of the Itaipu Hydroelectric Power Plant in 1982.

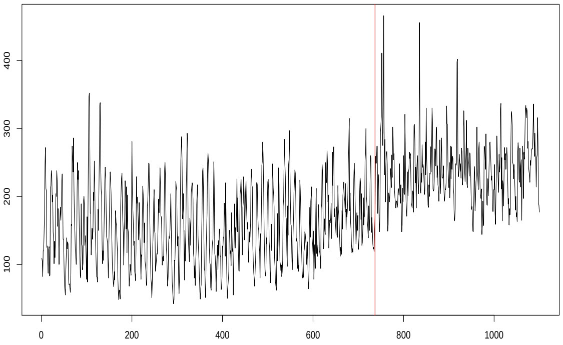


Figure 5: Time Series of Paran River

Thus, the modeling was done by applying the proposed scheme change point $\boldsymbol{\tau}$, using Bayesian approach, initial kick $\mathbf{k = 800}$ had as a result:

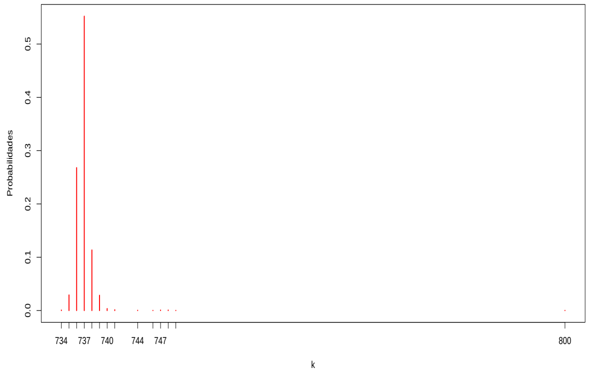


Figure 6: Change-Point probabilities, Paraná River

The most likely period of occurrence to change point was 737, dated 11/11/1981.

Table 3: Estimates of the parameters, for points $\tau \leq 737$ e $\tau > 737$.

	$\boldsymbol{\tau \leq 737}$	
Parameters	Estimates	95% Credible Interval
Location (μ)	129.9135	[125.8823 ; 133.7578]
Scale (σ)	51.4345	[48.4643 ; 55.0175]
Shape (ξ)	-0.1248	[-0.1765 ; -0.0646]
	$\boldsymbol{\tau > 737}$	
Parameters	Estimates	95% Credible Interval
Location (μ)	216.1418	[211.7028 ; 220.6419]
Scale (σ)	40.9981	[37.9853 ; 44.2026]
Shape (ξ)	-0.0576	[-0.1142 ; 0.0040]

The estimation of the parameters of the GEV distributions taking into account the point of change can also be found using the proposed model. The estimated parameters are shown in Table 3

The observed parameters are similar to those in the Parnaíba River, where there was an increase in the lease parameter (μ), the scale (σ) and form (ξ) there was a decrease.

Starting from the fact of the existence of such a point of change in the distribution regime, a comparison will be made between the returns before and after the point, where we will have:

Looking at the returns to $\tau \leq 737$, it is observed that in a period corresponding to every 1 year and 10 years, the maximum level of quotas is expected to be 237.9595 cm and 313.1509 cm respectively. Such returns can be seen in the return graph in Figure 3.

The returns to $\tau > 737$, it is noted that in a period corresponding to every 1 year and 10 years, the maximum level of quotas is expected to be 309.4345 cm and 384.0969 cm respectively. These returns can be seen in the return graph in Figure 7.

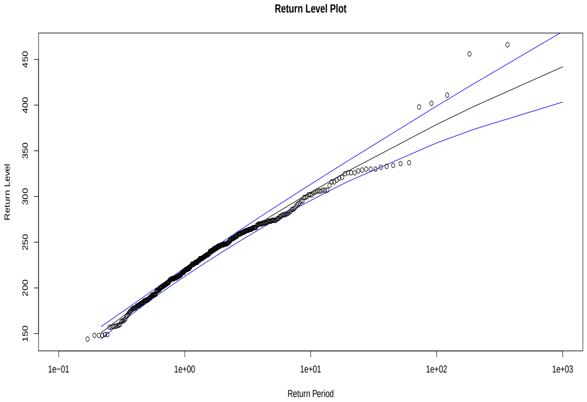


Figure 7: Returns Levels, Paraná River

Now showing through data, what is observed graphically is that after the point of change, there was an increase in the maximum values expected for returns in the time series.

6. Remarks

The proposed model for determining change points in time series was considerably effective. We verified the effectiveness of the model, from proposed simulations, where the change points were found, or accurately, or within a 95 % confidence interval. In addition, we observed that the model has a fast identification when the parameter that changes in the GEV distribution is the lease (μ).

Applications were made using fluviometric data, a future application would be to use such a model in other types of data, such as financial data. In addition to trying to create a model in which could be found more than one point regime in a time series.

Such applications in river basin data have made us realize that through human interference the course of a river can be significantly altered, causing predictions expected for height of quotas before the intervention to become larger after human action. In the case under analysis, the intervention was the construction of hydroelectric plants.

REFERENCES

CSÖRG$\tilde{O}$, M e HORVTH, L. (1997) **Limit Theorems in Change Point Analysis. Wiley: Chichester**.

FISHER, R. A. e TIPPET, L. H. C. (1928) **On the estimation of the frequency distributions of the largest and smallest sumber of a sample**, Proceedings of the Cambridge Philosophycal Society, 24, 180-190.

GAMERMAN, D. e LOPES, H. F. (2006) **Markov chain Monte Carlo: Stochastic Simulation for Bayesian Inference**, 2a. edio, Chapman & Hall.

Perreault, L. Berniera, J. Bobe, B. Parenta, E. (2000) **Bayesian change-point analysis in hydrometeorological time series. Part 1. The normal model revisited**. Journal of Hydrology 235, 221-241.