VALORES EXTREMOS: CONCEITOS, ABORDAGEM CLÁSSICA E MÉTODOS DE MODELAGEM

Pedro Pelacani Berger
Programa de Pós-Graduação em Engenharia Civil -UFES/ES
Universidade Federal do Espírito Santo, Av. Fernando Ferrari, s/n, 29060-900, Vitória-ES
pedropberger@gmail.com

Gutemberg Hespanha Brasil
Departamento de Estatística/PPGECO/PPGGP -UFES/ES
Universidade Federal do Espírito Santo, Av. Fernando Ferrari, s/n, 29060-900, Vitória-ES
ghbrasil@terra.com.br

RESUMO

Este trabalho objetiva apresentar as premissas da modelagem de valores extremos, abordar uma ideia geral quanto a sua aplicação e sumariar conceitos básicos. São apresentadas as características gerais da distribuição de valores extremos generalizada por meio das três famílias de distribuições: Gumbel, Weibull e Fréchet. Também são abordados métodos de estimação da cauda pesada dessas distribuições e duas aplicações práticas da teoria: série de chuvas em Fortaleza e série de manchas solares, ambas em base mensal.

PALAVARAS CHAVE. Valores extremos, Metódo de Gumbel, Peaks over Threshold.

EST - Estatística.

ABSTRACT

This work attempts to bring the premises of extreme value modelling, addressing a general idea as to its application and summarize the basic concepts. General characteristics of the extreme value distribution are presented, generalized by three families of distributions: Gumbel, Fréchet and Weibull. Also discussed are methods for estimating the heavy tail of these distributions and two practical applications of the theory: series of rainfall in Fortaleza and sunspot numbers, both on a monthly basis.

KEYWORDS. Extreme Values. Gumbel Method. Peaks over Threshold.

EST - Statistics.

1. Introdução

Valores extremos estão associados a baixas probabilidades. A teoria dos valores extremos (TEV) trata das questões probabilísticas e estatísticas relacionadas a valores muito baixos ou muito altos em sequências da variáveis aleatórias e em processos estocásticos. Assim, dados considerados como eventos raros não podem ser tratados simplesmente como "outliers", pois, podem ser inseridos em uma modelagem de extremos.

Mais diretamente, o interesse pode ser modelar as caudas de uma distribuição de alguma variável aleatória; esta pode representar dados sobre climatologia ou hidrologia ou outra área, como seguros ou finanças. Busca-se eventualmente precaver-se contra algum evento extremo (evento raro) cujas consequências podem ser indesejáveis ou desastrosas. Modelando essa cauda, pode-se ter ciência sobre as probabilidades de ocorrências de valores altos ou baixos). Entretanto, extremos não são usualmente bem modelados via distribuições clássicas, como, por exemplo, a distribuição normal.

O texto clássico reunindo a teoria de valores extremos é Gumbel (1958), onde já se encontram as distribuições de Fréchet, Weibull e Gumbel. Desenvolvimentos posteriores incluem a distribuição dos Valores Extremos Generalizados (GEV ou Generalized Extreme Value) e a Distribuição de Pareto Generalizada (GPD); Coles (2001), Smith (1989, 2003). Poderosos recursos computacionais, incluindo softwares, têm possibilitado um uso mais eficiente dessas distribuições (ou mesmo tornado possível seu uso).

Alguns problemas envolvendo extremos são: (i) quer-se prevenir de algum evento extremo, atribuindo ao evento um tempo médio de espera até nova ocorrência; (ii) estabelecer um certo valor extremo, o qual se denomina limiar, e obter o tempo de espera até a ocorrência desse evento; e, (iii) fixar um tempo de espera e calcular a probabilidade de se observar uma ocorrência além de um limiar pré-fixado. No decorrer do artigo aborda-se a teoria dos valores extremos apropriada para tratar esses problemas.

Na seção 2 comenta-se sucintamente sobre conceitos fundamentais para a modelagem de extremos. Na seção 3, descreve-se sobre distribuições Limitantes, as condições gerais de valores extremos, e Estimação. Na seção 4 conceitua-se o problema da taxa de retorno. As seções 5 e 6 apresentam os métodos de Gumbel e "Peak over Threshould" (POT), respectivamente. Na seção 7 são feitas aplicações da teoria de valores extremos a duas séries reais. Finalmente na seção 8 são feitos comentários à guisa de conclusões.

2. Conceituações Fundamentais para a Modelagem de Extremos

A teoria de valores extremos atua explicitamente nos resultados/valores extremos e provê uma série de "modelos naturais" para lidar com esses extremos. Toda a teoria está baseada em argumentos assintóticos onde as distribuições obtidas são casos limite de estatísticas de ordem; Gumbel (1958), Coles (2001), Smith (2003). "Essa abordagem pode ser denominada paradigma do valor extremo, visto que compreende um princípio para extrapolação de modelo baseada na implementação de limites matemáticos como aproximações de nível-finito"; Coles (2001, p. 2). Isso pressupõe que o mecanismo estocástico subjacente ao processo sendo modelado é suficientemente suave para possibilitar a extrapolação do valores não observados.

Os dados observados do modelo podem originalmente pertencer a essas distribuições de valores extremos, mas na grande maioria das vezes é necessária uma intervenção do pesquisador para a seleção correta dos extremos, além do tratamento prévio da série temporal quando vem ao caso. Além disso, supõe-se que os eventos são amostrados de uma população estacionária, o que significa que as propriedades estatísticas (média, variância, assimetria, etc) não variam no tempo.

Extremos apresentam características próprias, e não convém serem modelados como a maioria dos conjuntos de dados, onde geralmente o interesse é concentrado nas características centrais da população. Possuem uma teoria específica; e suas aplicações se diversificam em uma grande variedade de áreas e métodos de estimação, Finkenstädt e Rootzén (2004).

Existem abordagens clássicas e bayesianas para a confecção dos modelos, tanto no caso univariado como multivariado. Este trabalho trata apenas da visão clássica, mas ao mesmo tempo fornece a base e a noção das distribuições e métodos necessários para uma abordagem Bayesiana.

3. Distribuições Limitantes e Estimação

A teoria assintótica dos extremos amostrais tem seu desenvolvimento em paralelo com o teorema central do limite, e de fato as duas teorias têm alguma semelhança, Hann e Ferreira (2006). Ver também Coles (2001).

Seguindo o teorema de Fisher-Tippett-Gnedenko (ou teorema de Fisher-Tippett ou teorema dos valores extremos), sejam $X_1, X_2, X_3, \dots, X_n$, variáveis aleatórias independentes e identicamente distribuídas. O teorema central do limite está mais preocupado com o comportamento das somas parciais $X_1 + X_2 + \cdots + X_n$ quando $n \to \infty$, enquanto que a teoria dos extremos amostrais está mais preocupada com o comportamento dos extremos das amostras que são o máximo $\max(X_1, X_2, \dots, X_n)$ ou o mínimo $\min(X_1, X_2, \dots, X_n)$ quando $n \to \infty$.

Desejamos então encontrar as possíveis distribuições limite para o máximo amostral de variáveis aleatórias independentes e identicamente distribuídas. Seja F uma função de distribuição subjacente e x^* seu ultimo ponto a direita, ou seja:

$$x^* := \sup\{x : F(x) < 1\};$$

podendo este ser infinito, então:

$$\max(X_1, X_2, \dots, X_n) \xrightarrow{P} x^*, \qquad n \to \infty,$$

onde $\xrightarrow{P}$ significa convergência em probabilidade, desde que

$$P[max(X_1, X_2, \dots, X_n) \le x] = P[X_1 \le x, X_2 \le x, \dots, X_n \le x] = F^n(x),$$

Que converge para zero quando $x < x^*$e para 1 quando $x \ge x^*$. Assim, a fim de obter uma distribuição limite não degenerativa, uma normalização é necessária, Hann e Ferreira (2006).

Suponhamos que exista uma sequencia de constantes $a_n > 0$, e b_n reais onde $n = 1,2, \dots$, de tal forma que

$$\frac{\max(X_1, X_2, \dots, X_n) - b_n}{a_n}$$

tem uma distribuição limite não degenerativa quando $n \to \infty$; então

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G(x)$$

Onde $G(x)$ é uma função degenerativa para cada ponto de continuidade x. Todas as funções de distribuição que são obtidas a partir do limite acima são chamadas de distribuições de valores extremos.

De acordo com o teorema definido por Fisher-Tippet (1928), dado que exista uma sequência de constantes $a_n > 0$, e b_n reais ($n = 1,2, \dots$), e G uma função não degenerada então:

$$\frac{\max(X_1, X_2, \dots, X_n) - b_n}{a_n} \xrightarrow{d} G$$

Onde $\xrightarrow{d}$ representa convergência em distribuição. Nesse caso G pertence a uma das 3 famílias de distribuições de valores extremos, que são:

Quadro 1: Distribuições de Valores Extremos

Distribuição	*Função de Distribuição Acumulada*	*Função de Densidade de Probabilidade*
Tipo I	$\Pr[X \le x] = \exp\left\{-e^{-\frac{x-\xi}{\theta}}\right\}, \quad -\infty < \xi < \infty$	$p_x(x) = \theta^{-1} e^{-(x-\xi)/\theta} exp\left[-e^{-\frac{x-\xi}{\theta}}\right]$
Tipo II	$\Pr[X \le x] = \begin{cases} 0, & x < \xi \\ \exp\left\{-\left(\frac{x-\xi}{\theta}\right)^{-k}\right\}, & x \ge \xi \end{cases}$	$p_x(x) = \frac{k}{\theta}\left(\frac{x-\xi}{\theta}\right)^{-1-k} e^{-\left(\frac{x-\xi}{\theta}\right)^{-k}}$
Tipo III	$\Pr[X \le x] = \begin{cases} \exp\left\{-\left(\frac{x-\xi}{\theta}\right)^{k}\right\}, & x \le \xi \\ 1, & x > \xi \end{cases}$	$p_x(x) = \frac{k}{\theta}\left(\frac{(x-\xi)}{\theta}\right)^{k-1} e^{-\left(\frac{x-\xi}{\theta}\right)^{k}}$

Onde $(-\infty < \xi < \infty)$, $(\theta > 0)$, e $(k > 0)$ são parâmetros das distribuições. A distribuição correspondente de $(-X)$ é também chamada de distribuição dos valores extremos. As distribuições de valores extremos tipo I, II e III também são conhecidas como as famílias Gumbel, Fréchet e Weibull respectivamente. Dentre essas famílias de distribuições, a de uso mais comum é Gumbel, que também é conhecida como Exponencial Dupla ou log-Weibull.

Como exposto em Kotz e Nadarajah (2000), tem-se uma convergência similar quanto à modelagem dos máximos amostrais, a qual leva à distribuição de valores extremos generalizada. A partir da qual também podem ser obtidas as famílias de distribuições de valores extremos tipo I, II e III, a função de distribuição da GEV (Generalized Extreme Value) é dada por

$$P(X \geq x) = \exp\left\{-\left[1 + k\left(\frac{x-\xi}{\theta}\right)\right]^{-\frac{1}{k}}\right\}$$

onde

$$\left(1 + \frac{k(x-\xi)}{\theta}\right) > 0.$$

Várias técnicas foram propostas para a estimação de parâmetros de modelos de valores extremos. Em relação a estas, existe uma dificuldade clara quanto ao uso de métodos baseados na verossimilhança dos dados. Algumas condições necessárias para a simples aplicação desses métodos não são satisfeitas pelo modelos GEV porque os pontos máximos da distribuição são funções dos valores dos parâmetros. No entanto, Coles (2001, pp. 48-56) apresenta uma solução adequada para esses problemas. Mais detalhes sobre as distribuições de valores extremos e seus parâmetros podem ser encontrados em Berger (2011).

4. Taxa de Retorno

Após a obtenção dos estimadores de máxima verossimilhança dos parâmetros da GEV, o estimador de máxima verossimilhança para o quantil da distribuição de valores extremos generalizada x_p para $0 < p < 1$, a taxa de retorno $1/p$, é dado por:

$$\hat{x}_p = \begin{cases} \hat{\xi} - \frac{\hat{\theta}}{\hat{k}}\left[1 - y_p^{-\hat{k}}\right], & para\ \hat{k} \neq 0 \\ \hat{\xi} - \theta\ log y_p, & para\ \hat{k} = 0 \end{cases}$$

Onde $y_p = -\log(1-p)$.

Períodos de retorno longos são geralmente correspondentes a valores de p pequenos, que são de maior interesse, Coles (2001), Smith (2003).

O valor de p representa a probabilidade de retorno, que é útil quando desejamos encontrar a magnitude de um evento com p chance de ocorrer. A taxa de retorno é o retorno esperado em um dado período de tempo $1/p$. E x_p é o quantil p da GEV. (Em uma terminologia mais frequente, x_p é o **nível de retorno** associado com o **período de retorno** $1/p$).

É necessária cautela na interpretação das inferências para o período de retorno, especialmente para taxas de retorno que correspondem a longos períodos, pois a aproximação à normalidade da distribuição do estimador de máxima verossimilhança pode ser fraca. Melhores aproximações são obtidas através de funções do perfil da verossimilhança; Coles (2001).

Estimativas numéricas do perfil da verossimilhança (o perfil da verossimilhança é o comportamento da função de verossimilhança a partir de parâmetros anteriormente definidos) para qualquer um dos parâmetros ξ, θ ou k podem ser obtidas. Por exemplo, para se obter o perfil da verossimilhança de k, fixamos $k = k_0$, e maximizamos o logaritmo da verossimilhança, sem modificar os outros parâmetros, ξ e θ. Repete-se isso para uma sequência de valores de k_0. Os valores maximizados correspondentes constituem o perfil do logaritmo da verossimilhança, a partir dos quais podemos obter os intervalos de confiança aproximados.

Essa metodologia pode ser aplicada quando precisamos usar a inferência em alguma combinação de parâmetros. Em particular, podemos obter intervalos de confiança para qualquer taxa de retorno especificada. Isso requer uma reparametrização do modelo GEV, onde x_p passa a

ser um dos parâmetros do modelo; após isso o perfil da log-verossimilhança pode ser obtido pela maximização, mantendo os demais parâmetros, da forma que foi descrita acima. A reparametização ficaria:

$$\xi = x_p + \frac{\theta}{k}[1 - \{-\log(1-p)\}^{-k}]$$

De modo que a substituição de ξ no logaritmo da função de verossimilhança da GEV terá o efeito desejado de expressar o modelo de valores extremos generalizados em termos dos parâmetros (x_p, θ, k), Coles (2001). Embora impossível checar a validade de uma extrapolação baseada num modelo GEV, avaliações do ajuste são feitas com base nos dados observados. O que não é suficiente para justificar a extrapolação, mas é um bom pré-requisito. Utilizamos para isso ferramentas como os usuais P-P plot e o Q-Q plot.

O gráfico da taxa de retorno, dada pelo gráfico de

$$x_p = \xi - \frac{\theta}{k}[1 - \{-\log(1-p)\}^{-k}],$$

onde $x_p = -\log(1-p)$ numa escala logarítmica, é particularmente conveniente para a interpretação de um modelo de valores extremos. A cauda da distribuição é comprimida, de tal forma que as estimativas da taxa de retorno para um longo período pode ser apresentada, onde a linearidade do gráfico no caso em que $k = 0$ daria uma base para julgar o efeito da estimativa do parâmetro da curva.

5. Método de Gumbel

O método de Gumbel é um dos métodos baseados na abordagem clássica para a estimativa da taxa de retorno. Ele basicamente consiste na divisão de um ano de uma série temporal em n grupos (podendo ser definidos como meses, estações, ou a critério do pesquisador). Os máximos dos grupos são ajustados através da distribuição de Gumbel, e a partir daí busca-se estimar a taxa de retorno e o valor que pode retornar.

Fazendo $Max_{i1}, Max_{i2}, \ldots, Max_{ij}$ os máximos de cada um dos j grupos no ano i, se obtém uma estrutura de matriz da seguinte forma

$$\begin{matrix} & Grupos & & & \\ Anos & \begin{bmatrix} Max_{11} & Max_{12} & \ldots & Max_{1j} \\ Max_{21} & Max_{22} & \ldots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ Max_{i1} & Max_{i2} & \ldots & Max_{ij} \end{bmatrix} \end{matrix}$$

Note que cada linha contém o máximo de cada grupo de um ano, e cada coluna os máximos de cada grupo em sequência anual. Cada coluna dessa matriz é ajustada a uma distribuição de Gumbel $F_i(x)$. Usando essa distribuição, estima-se então $\hat{x}_p$ que é o valor máximo de retorno estipulado em um período, como já apresentado em sessão anterior; Ferreira, Souza e Brasil (1999), Coles (2001).

A grande vantagem desse método é que evita a necessidade de um estudo anterior quanto à sazonalidade da série, pois estima-se o grupo ano após ano, e em grande maioria das vezes, as séries apresentam dados de sazonalidade mensal, só havendo a necessidade específica de um tratamento quando se tratam de séries financeiras. A desvantagem é que o modelo fica suscetível a possíveis tendências ou alterações nos padrões sazonais.

6. Peak over Threshould (POT)

O principal objetivo do Peak Over Thershold (POT) é solucionar o problema da estimação da cauda, Belitsky e Moreira (2007). Diferente do método de Gumbel, onde selecionamos os valores máximos dos blocos da amostra, o método POT consiste em determinar um limiar ótimo u (threshold limit), o qual, acima desse, todos os valores serão tratados como valores extremos, ou seja, tem-se a certeza de sempre estar se tratando de eventos raros, o que não acontece quando separados por blocos.

Passa-se então a tratar também de $Y = (X - u)$ que é uma variável aleatória que representa a diferença entre o valor de X que ultrapassa o limiar u e é comumente modelado por uma distribuição exponencial. Essa função de distribuição designada por

$$F_u(x) = P[X - u | X > u]$$

E também é conhecida por função de distribuição de excessos da distribuição $F(\cdot)$ dos valores acima do limiar u. A partir desses valores acima do limiar se obtém a taxa de retorno, ou a probabilidade da ocorrência de eventos raros, medidas de risco entre outros elementos. É importante observar que não existe critério para a escolha da altura do limar ótimo u.

O método não trata explicitamente de componentes de sazonalidade e tendência, ou seja, é necessário que a série seja tratada corretamente antes da análise; outra peculiaridade do método é que, assim como no método de Gumbel, é necessário o pressuposto de independência, que não se verifica na prática, mas é comumente admitido em estudos, pois além de permitir uma maior facilidade dos cálculos, fornece resultados próximos aos que seriam derivados caso a verdadeira dependência fosse conhecida e levada em conta, Belitsky e Moreira (2007).

7. Aplicações

Inicialmente, foram realizados testes de aderência (goodness-of-fit) - Kolmogorov-Smirnov (Kolmogorov, 1933)., dos dados a doze tipos de distribuição (Qui-quadrado, GEV, Normal, Pareto, Gumbel, Fréchet, Weibull, T, e outras) para as duas séries da aplicação, considerando-se diversos padrões: (i) máximos anuais; (ii) toda a série em base mensal; e, (iii) Máximos (POT). Para a série de chuvas em Fortaleza, casos (i) e (ii), a melhor aderência foi para as distribuições GEV e Weibull. Para o caso (iii) foi a Pareto Generalizada. No caso da série de manchas solares, casos (i) e (ii), a melhor aderência foi para as distribuições de Pareto generalizada e Weibull.

7.1 Série de Chuvas em Fortaleza

A série de precipitação pluviométrica em Fortaleza, Ceará, Brasil é uma das séries climáticas mais longas, com registro desde 1849. A análise temporal desta série procura extrair suas características, principalmente a mensuração de ciclos, e a previsão, Brasil e Souza (2001). A análise realizada nesse trabalho se faz para a obtenção da taxa de retorno e do máximo pluviométrico que podem ser obtidos em diferentes períodos de tempo, utilizando os dados em base mensal, de Janeiro de 1849 a Dezembro de 1999. A estimação dos modelos e os cálculos das taxas de retorno foram realizados com o software "R"; R-Cran (2011).

Observados esses requerimentos básicos, a série foi dividida em blocos anuais, e os máximos de cada um dos blocos estimados; foram obtidos 151 máximos. Feito o teste de razão de verossimilhança a 95% de confiança para testar a hipótese de $k = 0$, a hipótese nula de que a distribuição pertencia ao tipo I não foi aceita, o p-valor do teste foi 0,00064. Em suma, pode-se assumir que se trata de uma $GEV(\xi, \theta, k)$. Os parâmetros então estimados para o modelo podem ser observados na Tabela 1. A qualidade do ajuste pode ser observada graficamente na Figura 1.

Tabela 1: Parâmetros estimados da GEV pelo método de Gumbel

	$\hat{\xi}$	$\hat{\theta}$	$\hat{k}$
Estimativas	361,9947	135,0446	-0,21401
Erro Padrão	12,13538	8,541517	0,051085

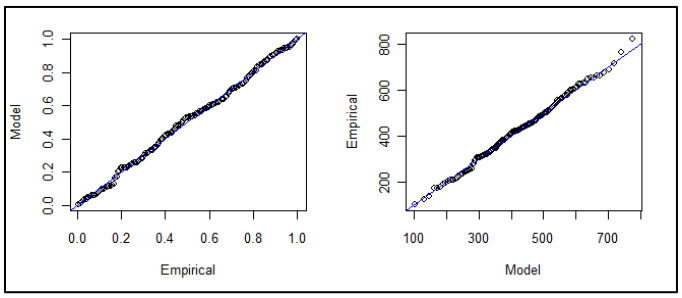

Figura 1: PP-plot e QQ-plot do modelo ajustado GEV(ξ, θ, k)- método de Gumbel

Estimou-se então a taxa de retorno dos para 10, 100, 200, 300, 500 e 1000 anos, a partir do fim da série (dezembro de 1999), com um intervalo de confiança de 95%. A tabela com as taxas estimadas e o gráfico podem ser observados a seguir, na Tabela 2 e Figura 2.

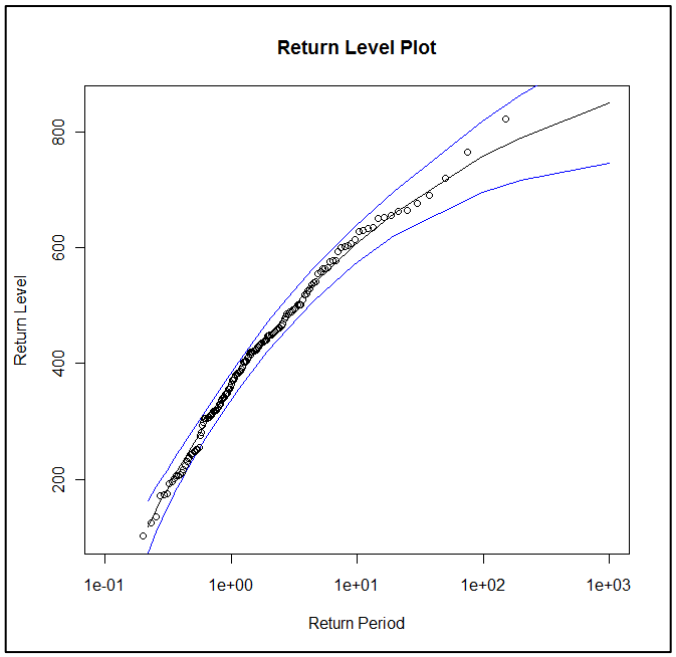

Figura 2: Gráfico das taxas de retorno estimadas em relação ao período: As linhas azuis representam os intervalos de confiança de 95% e a linha central o modelo estimado

Tabela 2: Taxas de retorno estimadas pelo método de Gumbel

Período de Retorno (Anos)	Retorno Esperado	IC95%	
		Inferior	*Superior*
10	603,17	571,11	635,23
100	757,24	695,3	819,19
200	789,86	715,66	864,05
300	806,77	725,19	888,36
500	826,09	735,12	917,05
1000	849,11	745,51	952,71

Foi então aplicado o método POT ao conjunto de dados, a fim de comparar sua eficiência com o método de Gumbel. O limiar selecionado foi o mesmo que define extremos em um gráfico Box-Plot, ou seja

$$u = 1{,}5 * (Q_3 - Q_1)$$
$$u \approx 263$$

Onde $(Q_3 - Q_1)$ representa o intervalo interquartil. A série dos picos acima do limiar possui ao todo 294 dados. Da mesma forma que no método de Gumbel, os parâmetros foram estimados a partir dos extremos que convergiam para uma GEV, estes podem ser observados na Tabela 3.

Tabela 3: Valores dos parâmetros estimados pelo método POT

	$\hat{\xi}$	$\hat{\theta}$	$\hat{k}$
Estimativas	343,9116	704,7682	0,226643
Erro Padrão	5,02668	4,14294959	0,068669

O ajuste através do método POT não se mostrou tão eficiente, como se pode ver no QQ-plot na Figura 3, apesar disso, também foram tabeladas as taxas de retorno para 10, 100, 200, 300, 500 e 1000 anos com um intervalo de confiança de 95%, as estimativas estão presentes da Tabela 4.

Tabela 4: Taxas de retorno estimadas pelo método POT

Período de Retorno (Anos)	*Retorno Esperado*	*IC95%*	
		Inferior	*Superior*
10	550,80	514,62	586,98
100	915,00	720,91	1109,10
200	1065,64	776,22	1355,05
300	1165,25	805,42	1525,07
500	1304,42	837,54	1771,31
1000	1520,88	869,75	2172,02

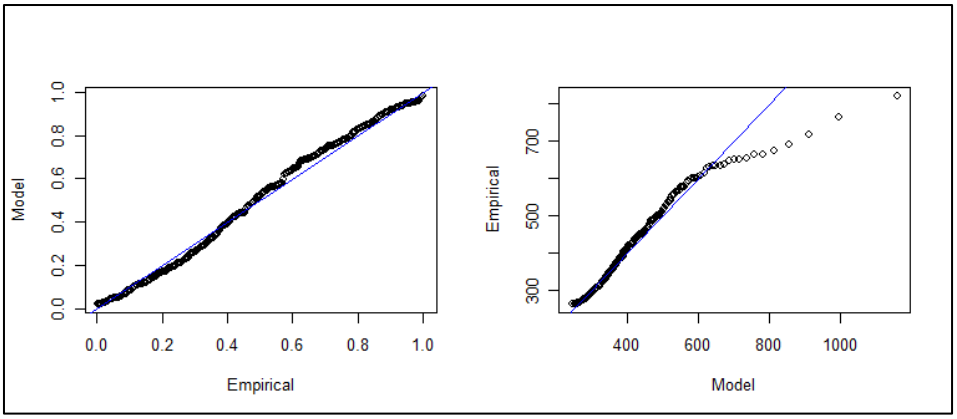

Figura 3: PP-plot e QQ-plot do modelo ajustado GEV(ξ, θ, k) - método POT

A taxa de retorno apresenta um comportamento diferente do quando estimada pelo método POT em relação ao método de Gumbel. O modelo estimado apresenta uma tendência de crescimento exponencial, não seguindo a tendência aparente dos dados, que em duas observações ultrapassam os limites de confiança. Possivelmente o limiar selecionado foi muito baixo, o que leva a uma má convergência para a distribuição GEV. O que poderia explicar que as taxas de retorno menores apresentam outro comportamento. Podemos observar isso na Figura 4.

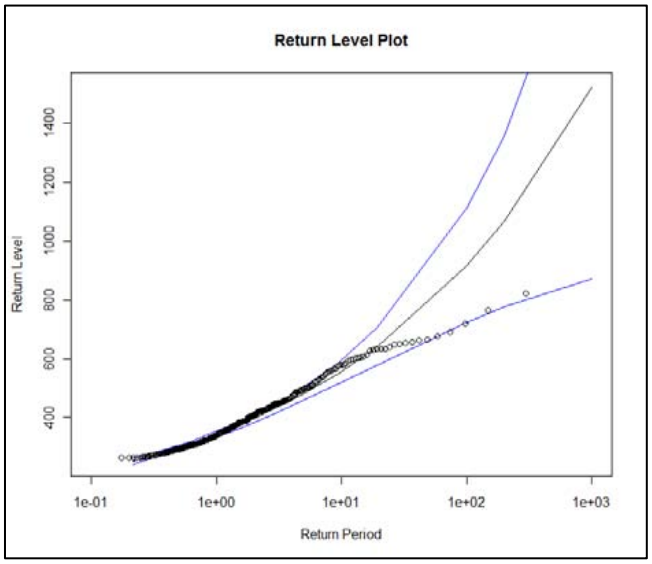

Figura 4: Gráfico das taxas de retorno estimadas em relação ao período: As linhas azuis representam os intervalos de confiança de 95% e a linha central o modelo estimado

7.2 Manchas Solares (sunspot numbers)

A atividade das manchas solares ocorre como parte de um ciclo de aproximadamente onze anos chamado ciclo solar ou ciclo de Schwabe, em que há períodos de atividade máxima e mínima. Quanto maior o número de manchas na superfície solar, maior será sua interferência na ionosfera do planeta Terra; Willson, Gulkis, Janssen, Hudson e Chapman (1981). A densidade do fluxo magnético das manchas solares é medida em Gauss. Os dados, constantemente atualizados, para a aplicação podem ser encontrados em NASA-GISS (2011). A série tem ao todo 3144 observações, que são as médias mensais de Janeiro de 1749 até Dezembro de 2010.

Diferente da série pluviométrica, os comportamentos mensais da série de manchas solares são similares (sazonalidade praticamente inexistente, i.e. doze fatores sazonais aproximadamente iguais a um), o que não acontece com os padrões anuais, que atendem a um ciclo de difícil estimação. O objetivo dessa aplicação é testar a convergência da distribuição GEV como uma boa forma de estimar a taxa de retorno, visto que a série temporal de sua aplicação tem um comportamento "complicado", e analisar qual dos dois métodos de estimação propostos trazem melhores resultados em termos de precisão e exatidão. Primeiramente os dados foram trabalhados pelo método de Gumbel, onde foi obtida a série com os 262 máximos anuais.

O teste de razão de verossimilhança que testa a hipótese nula $k = 0$ retornou um p-valor de 0,061252, que mostra que a GEV pode também ser uma Gumbel. Devido a isso, os parâmetros foram estimados também considerando essa hipótese, e considerando-se a distribuição dos máximos como uma $GEV(\xi, \theta, k)$ podem ser observados na Tabela 5. A qualidade do ajuste pode ser observada na Figura 5 através do QQ-plot e PP-plot.

Tabela 5: Valores dos parâmetros estimados pelo método de Gumbel

	Gumbel		*GEV*		
	$\hat{\xi}$	$\hat{\theta}$	$\hat{\xi}$	$\hat{\theta}$	$\hat{k}$
Estimativas	40,77234	53,16801	50,43854	38,40353	0,12761
Erro Padrão	2,052693	2,648241	2,912703	2,317825	0,072566

A partir desses parâmetros, buscou-se então estimar as taxas de retorno para 10, 100, 200, 300, 500 e 1000 anos. Os valores podem ser observados na Tabela 6. O gráfico com o modelo e seus intervalos de confiança podem ser observados na Figura 6.

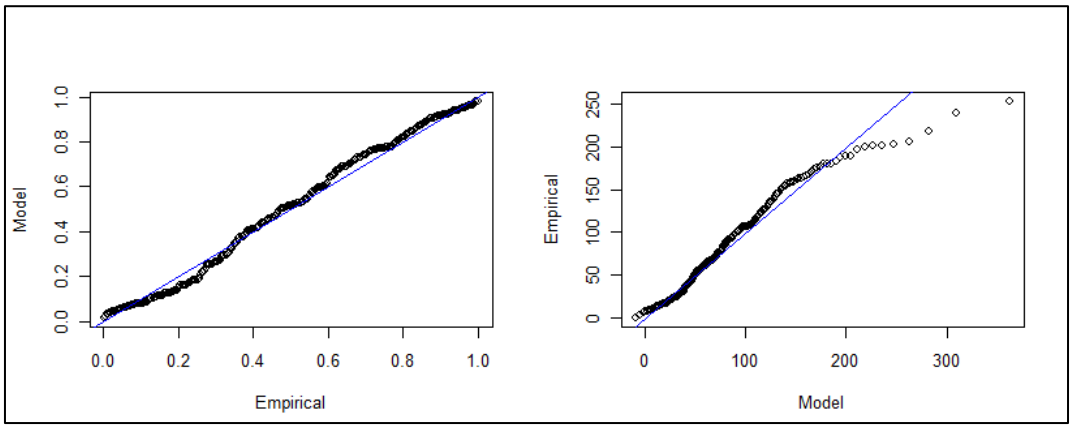

Figura 5: PP-plot e QQ-plot do modelo ajustado $\mathbf{GEV}(\boldsymbol{\xi}, \boldsymbol{\theta}, \mathbf{k})$ *- método de Gumbel*

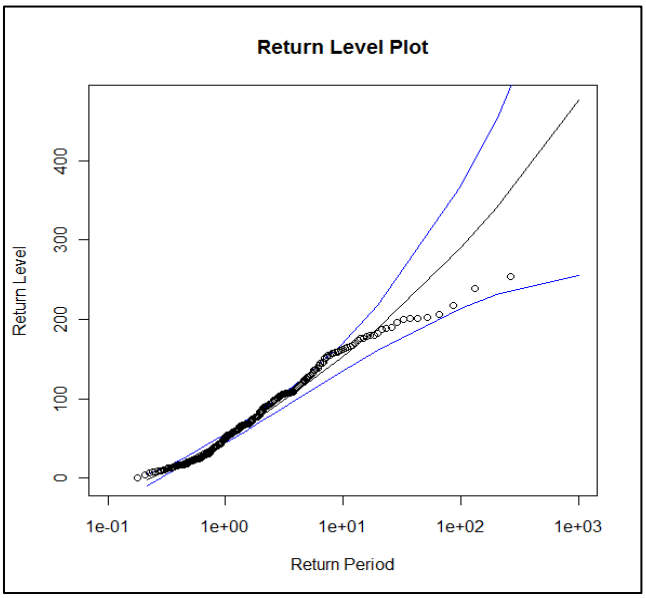

Figura 6: Gráfico das taxas de retorno estimadas em relação ao período: As linhas azuis representam os intervalos de confiança de 95% e a linha central o modelo estimado

Para o método POT, o limiar u selecionado foi 262, utilizando o mesmo critério da aplicação anterior. A convergência foi evidenciada, mais uma vez pelo teste Kolmogorov-Smirnov, o ajuste se mostrou melhor que pelo método de Gumbel, observando-se o QQ-plot e PP-plot dados na Figura 7.

Tabela 6: Taxas de retorno utilizando o método de Gumbel e POT para a série de manchas solares

Periodo de Retorno (Anos)	*Método de Gumbel*			*Método POT*		
	Retorno Esperado	*IC95%*		*Retorno Esperado*	*IC95%*	
		Inferior	*Superior*		*Inferior*	*Superior*
10	150,54774	133,9265	167,169	167,8463	160,5405	175,1522
100	290,78193	213,5932	367,9706	271,7755	229,3694	314,1817
200	341,03145	231,0394	451,0235	316,2934	252,0154	380,5715
300	372,5106	239,4713	505,5499	346,1297	265,4764	426,7829
500	414,53252	247,9444	581,1206	388,2918	282,4522	494,1314
1000	476,08336	255,0152	697,1516	454,8537	305,0059	604,7016

Os parâmetros estimados para o modelo e as taxas de retorno para 10, 100, 200, 300, 500 e 1000 anos juntamente com seus intervalos de confiança de 95% estão na Tabela 7 e Tabela 6, respectivamente.

Tabela 7: Parâmetros estimados da distribuição GEV obtida a partir do método POT

	$\hat{\xi}$	$\hat{\theta}$	$\hat{k}$
Estimativas	112,0694	18,43523	0,251263
Erro Padrão	0,967321	0,809876	0,053475

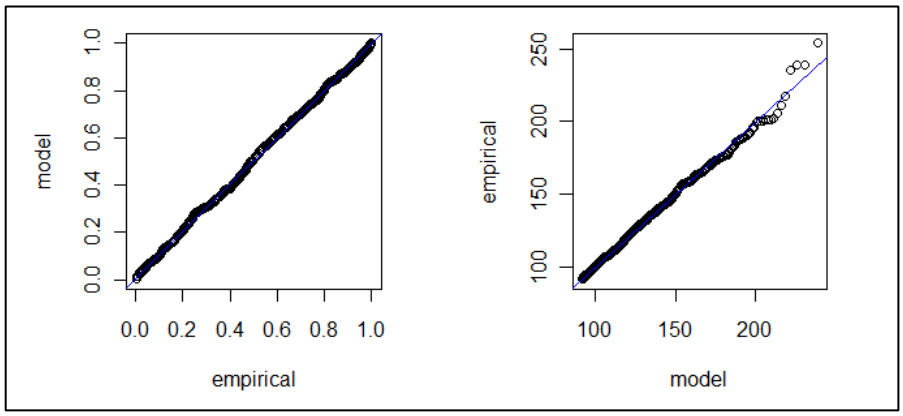

Figura 7: PP-plot e QQ-plot do modelo ajustado $\mathbf{GEV}(\boldsymbol{\xi}, \boldsymbol{\theta}, \mathbf{k})$ - método POT

Por fim, foi plotado o gráfico da taxa de retorno, na Figura 8. O comportamento dos dados em relação ao modelo ajustado foi diferente dos demais apresentados aqui, com um grande número de valores fora dos intervalos de confiança, o que entra em contraste com o bom ajuste observado no PP-plot e QQ-plot da Figura 7. Esse bom ajuste quanto à distribuição, mas com uma estimação ruim da taxa de retorno indica problemas no tratamento da série (possivelmente a escolha do limiar u), o que é evidente, visto que a mesma foi tratada a partir dos seus dados brutos.

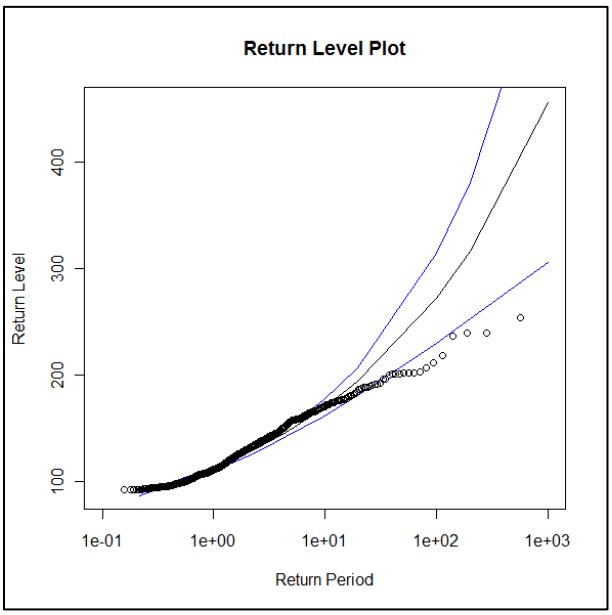

Figura 8: Gráfico das taxas de retorno estimadas em relação ao período: As linhas azuis representam os intervalos de confiança de 95% e a linha central o modelo estimado (POT)

8. Conclusão

A convergência dos máximos amostrais proposta pelo teorema dos valores extremos para a distribuição de valores extremos generalizada se mostrou verdadeira nos conjuntos de dados propostos independente do método utilizado.

A taxa de retorno obtida para os dados pluviométricos de Fortaleza mostra que uma chuva de grande magnitude esperada para os próximos 100 anos é quase tão intensa quanto uma esperada para os próximos 300 anos, 757mm e 806mm respectivamente. Apesar de que estimar uma taxa de retorno para os próximos 300 anos se mostra algo pouco relevante ou preciso, essa pequena diferença entre as duas taxas indica que medidas de prevenção tomadas para um retorno de 100 anos seriam praticamente suficientes por um tempo quase inestimável. Contudo, considerando-se o intervalo de predição, essa é uma conclusão que deve ser tomada com cautela.

O método POT não apresentou um ajuste tão bom quanto o esperado, isso se deve ao fato de que, considerando a temporada de chuvas no fim do verão, praticamente todos os pontos acima do limiar estipulado se encontraram nessa faixa, captando também dados que não são extremos, apenas valores altos devido a estação, prejudicando a convergência do teorema, que é a base para todas as técnicas apresentadas aqui, além do fato da série não ter sido previamente tratada, e os dados selecionados possivelmente apresentarem autocorrelação.

A série de manchas solares foi útil para mostrar qual dos dois métodos se mostram mais eficientes diante de uma série não ajustada, na presença ou não de sazonalidade e comportamentos desconhecidos. O esperado era de que o método de Gumbel se apresentasse superior por selecionar sempre os extremos nesse caso, independente da tendência da série, enquanto o método POT viria a selecionar muitos dados que não apresentam características de observações raras. O método de Gumbel por sua vez apresentou um parâmetro de locação menor (mais próximo de zero) enquanto o parâmetro de dispersão muito maior. Por sua vez, o método POT apesar de apresentar uma locação maior mostrou uma menor dispersão dos dados, mas as estimativas das taxas de retorno de ambos se mostraram próximas.

O próximo passo é o uso de modelos de séries temporais, em especial, modelos estruturais com a distribuição GEV.

Agradecimentos: Os autores agradecem às sugestões de dois revisores.

Referências

Belitsky, V.; Moreira F. M. (2007), Emprego do método "Peaks-Over-Threshold" na estimação de risco; uma exposição abrangente, detalhada mas simples, Instituto de Matemática e Estatística, Universidade de São Paulo, 2007.

Berger, P.P. (2011), Valores Extremos: Conceitos, Abordagem Clássica e Métodos de Modelagem, monografia, Departamento de Estatística da Universidade Federal do Espírito Santo.

Brasil, G. H.; Souza, R.C. (2002), Uma Reanálise da Série de Chuvas em Fortaleza, Brasil, XI CLAIO-Congreso Latino Iberoamericano de Investigacion de Operaciones, 27-31/10/2002, Concepción, Chile, Artigo completo na Acta de Trabajos, Cd-Rom (Trabajo A42n02, 13 p).

Coles, S. (2001), An introduction to statistical modeling of extremes values, Springer, 2001.

Cruciani, D.E. (1980), A drenagem na agricultura. São Paulo: Nobel, 1980.

Ferreira, M.J.S.; Souza, R.C.; Brasil, G.H. (1999), Modelos Dinâmicos Bayesianos para Extremos, Investigação Operacional, APDIO-CESUR-IST, Portugal, Volume 19, Número 1, 95-121, Junho/1999.

Finkenstädt, B.; Rootzén, H. (2004), Extreme Values in Finance, Telecommunications, and the Environment, Chapman & Hall/CRC Press, 2004.

Fisher, R. A.; Tippet, L. H. C. (1928), Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proceedings of the CambrigePhilophical Society, n.24, p.180-190, 192, 1928.

Gumbel, E.J. (1958), Statistics of Extremes, Columbia Univ. Press, New York, 1958

Hann, L.; Ferreira, A. (2006), Extreme value theory, Springer, 2006.

Kotz, S.; Nadarajah, S. (2000), Extreme value distributions: Theory and Aplications, Imperial College Press, 2000.

NASA-GISS (2011), NASA Goddard Institute for Space Studies, http://www.giss.nasa.gov/.

R-CRAN, (2011), The R Project for Statistical Computing, http://www.r-project.org/, 2011.

Smith, R. L. (1989). Extreme value analysis of environmental time series: an example based on ozone data (with discussion). Statistical Science, 4, 367-393.

Smith, R.L. (2003), Statistics of extremes, with applications in environment, insurance and finance, disponível em www.stat.unc.edu/postscript/rs/semstatrls.ps.

Willson, R. C.; Gulkis, S.; Janssen, M.; Hudson, H. S.; Chapman, G. A. (1981), Observations of solar irradiance variability, Science, vol. 211, Feb. 13, p. 700-702, 1981.