Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Curso de Estatística

Douglas de Oliveira Matos Braga

Aplicação da Teoria de Valores Extremos para índice pluviométrico da cidade de Juiz de Fora - MG

Juiz de Fora

2015

Douglas de Oliveira Matos Braga

Aplicação da Teoria de Valores Extremos para índice pluviométrico da cidade de Juiz de Fora - MG

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do título de Bacharel em Estatística .

Orientador: Clécio da Silva Ferreira

Juiz de Fora

2015

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

de Oliveira Matos Braga, Douglas.
Aplicação da Teoria de Valores Extremos para índice pluviométrico da cidade de Juiz de Fora - MG / Douglas de Oliveira Matos Braga. – 2015.
50 f. : il.

Orientador: Clécio da Silva Ferreira
Trabalho de Conclusão de Curso – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Curso de Estatística, 2015.

1. Teoria de Valores Extremos. I. da Silva Ferreira, Clécio, orient. II. Título.

Douglas de Oliveira Matos Braga

Aplicação da Teoria de Valores Extremos para índice pluviométrico da cidade de Juiz de Fora - MG

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do título de Bacharel em Estatística .

Aprovada em: 15/06/2015

BANCA EXAMINADORA

Professor Dr. Clécio da Silva Ferreira - Orientador
Universidade Federal de Juiz de Fora

M.Sc. Márcio Antônio Deotti Ibrahim
Subsecretário de Defesa Civil
Prefeitura de Juiz de Fora

Professor Dr. Ronaldo Rocha Bastos
Universidade Federal de Juiz de Fora

AGRADECIMENTOS

À minha família, em especial à minha mãe Marina, pelo apoio constante e confiança.

Aos meus amigos, especialmente à minha turma do curso de Estatística, Bárbara, Bethânia, Isabela e Gabriely, pela amizade, companheirismo e solidariedade.

À Marcela, por todo o carinho e pelas revisões desse trabalho.

Aos professores do Departamento de Estatística da UFJF pelos ensinamentos, em especial ao meu orientador Clécio.

Aos professores Zempléni András e László Márkus da Eötvos Lorand University, por me apresentarem a Teoria de Valores Extremos.

A todos que tornaram esse trabalho possível, meu muito obrigado.

"Essencialmente, todos os modelos estão errados, mas alguns são úteis."
George E. P. Box

In cauda venenum.

RESUMO

A análise de eventos, muitas vezes críticos, que se encontram nas caudas das distribuições é dificultada pelo fato de haver pouca informação sobre eles, devido à sua raridade. A teoria de valores extremos apresenta metodologias para lidar com estes eventos através de distribuições limite, possibilitando a inferência sobre os mesmos. Este trabalho oferece uma introdução à esta teoria com a utilização de seus modelos univariados mais difundidos, a distribuição de valor extremo generalizada (GEV) e a distribuição Pareto generalizada (GPD). É feita uma revisão bibliográfica sobre suas propriedades, técnicas de modelagem, estimação e de avaliação da qualidade do ajuste. Ao final, é utilizada para análise de eventos extremos na precipitação pluvial diária da cidade de Juiz de Fora – MG. Tais eventos são responsáveis por situações de inundações, soterramentos e desabamentos na cidade e o conhecimento sobre o comportamento dos extremos da precipitação deve ser utilizado para a minimização de seu impacto e prevenção de tais tragédias. Para este estudo, foram utilizados índices da precipitação pluvial diária entre 01 de janeiro de 1961 e 31 de dezembro de 2014, cedidos pelo Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) do Instituto Nacional de Meteorologia. Através destes dados, concluiu-se que chuvas com alto potencial danoso são esperadas ocorrerem uma vez entre 0,99 ano e 2,4 anos (modelo GEV) e entre 1,17 ano e 2,24 anos (modelo GPD). Estes resultados ressaltam a importância de ações preventivas que devem ser exercidas de forma conjunta pelo Poder Público e população.

Palavras-chave: Teoria de Valores Extremos. Precipitação Pluvial. Distribuição de Valor Extremo Generalizada. Distribuição Pareto Generalizada.

ABSTRACT

The analysis of events, often critical, which are at the tails of the distributions is difficult because there is little information about them, due to their rarity. The extreme value theory presents methods for dealing with these events through limit distributions, allowing inference about them. This paper provides an introduction to this theory with the use of their most widespread univariate models, the generalized extreme value distribution (GEV) and the generalized Pareto distribution (GPD). A literature review on their properties, modeling techniques, estimation and evaluation of goodness-of-fit is shown. At the end, it is used for analysis of extreme events in the daily rainfall in the city of Juiz de Fora - MG. Such events are responsible for situations of floods, landslides and burials in the city and the knowledge about the behavior of precipitation extremes must be used to minimize their impact and prevent such tragedies. For this study, we used indices of daily rainfall between January 1, 1961 and December 31, 2014, provided by the Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) of the Instituto Nacional de Meteorologia. Through these data, it was concluded that potential harmful rains are expected to occur once between 0.99 and 2.4 years (GEV model) and between 1.17 and 2.24 years (GPD model). These results highlight the importance of preventive actions to be performed jointly by the government and community.

Key-words: Extreme Value Theory, Rainfall, Generalized Extreme Value Distribution, Generalized Pareto Distribution

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O conhecimento sobre eventos extremos é de grande importância para o sociedade pois, muitas vezes, estão associados a situações de risco como, por exemplo, grandes enchentes, terremotos de grandes magnitudes, crises financeiras ou vazamentos em uma usina nuclear. Por definição, estes eventos são escassos, havendo a necessidade de inferir sobre valores além dos observados.

Como um exemplo, o governo da Holanda, cuja grande parte do território - cerca de 40% - encontra-se abaixo do nível do mar e protegido da água por diques, determinou que a altura dos diques deve ser suficiente para que a probabilidade de que, em um dado ano, o nível do mar exceda o topo do dique e cause uma enchente seja de 10^{-4} . Ou seja, é necessário estimar a altura dos diques para que se espere ocorrer uma enchente apenas uma vez a cada 10000 anos. Mas existem dados disponíveis sobre o nível do mar para pouco mais de 100 anos, nos quais não houve nenhuma enchente (HAAN; FERREIRA, 2006).

Para a resolução desta questão, a teoria de valores extremos proporciona classes de modelos assintóticos que tornam possíveis a extrapolação para níveis não observados nas caudas das distribuições. A modelagem através desta teoria consiste em utilizar as observações extremas para estudar o comportamento assintótico da cauda, quando a distribuição populacional é desconhecida. A teoria probabilística dos valores extremos foi desenvolvida por Frechét (1927), Fisher e Tippett (1928) e Mises (1936) e resultaram no trabalho de Gnedenko (1943). Já a teoria estatística foi iniciada por Pickands (1975).

A teoria de valores extremos possui aplicação em diversas áreas, como hidrologia para análise de níveis de precipitação (KATZ, 1999) e enchentes (MORRISON; SMITH, 2002); em finanças para a modelagem do risco de portifólios de investimentos (GILLI et al., 2006); em atuária para análise de perdas extremas causadas por desastres naturais (PFEIFER, 2001) e para estimação de prêmios de resseguros (VANDEWALLE; BEIRLANT, 2006); em engenharia para predição de resistência de materiais (TRYON; CRUSE, 2000) e confiabilidade de sistemas (CHEN; LI, 2007).

O objetivo desse trabalho é introduzir os dois principais modelos univariados da teoria de valores extremos, quais sejam, a distribuição de valores extremos generalizada, apresentada no Capítulo 2, e a distribuição de Pareto generalizada, apresentada no Capítulo 3. O Capítulo 4 traz um estudo de simulação com o objetivo de apresentadar as propriedades de ambos os modelos, métodos de estimação e de avaliação do ajuste. Além disto, o Capítulo 5 mostra uma aplicação da modelagem univariada pela teoria de valores extremos através de um estudo sobre os extremos da precipitação pluvial diária na cidade de Juiz de Fora - MG.

2 DISTRIBUIÇÃO VALOR EXTREMO GENERALIZADA (GEV)

A distribuição de valor extremo generalizada é um modelo assintótico que busca modelar probabilisticamente a parte extrema da cauda da distribuição de uma variável de interesse X a partir da distribuição do máximo dessa variável. A formulação dessa distribuição é possível devido ao Teorema 2.1.1, que especifica a forma da distribuição do máximo centralizado e padronizado de uma sequência de variáveis aleatórias. Este teorema, na teoria de valores extremos, é uma versão análoga ao teorema central do limite para a soma de variáveis aleatórias.

2.1 FORMULAÇÃO DO MODELO

Seja $X_1, \ldots, X_n$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas com função de distribuição comum F_X e denotemos M_n o máximo das n primeiras variáveis aleatórias, ou seja, $M_n = max(X_1, \ldots, X_n)$. A variável aleatória M_n terá sua função de distribuição F_{M_n} dada por:

$$F_{M_n}(z) = P(M_n \leq z) = P(X_1 \leq z, \ldots, X_n \leq z) = F_X^n(z).$$

Se existe uma sequência de números positivos $\{a_n, n \geq 1\}$ e uma sequência de números $\{b_n, n \geq 1\}$ tal que para todos valores de z

$$P\left(M_n^* = \frac{M_n - b_n}{a_n} \leq z\right) = F_X^n(a_n z + b_n) \to G(z)$$

quando $n \to \infty$ e G(z) é um limite não-degenerado, então F_X pertence ao domínio de atração de G, denotado como $F \in \mathcal{D}(G)$ e, além disso, G é uma distribuição max-estável, também chamada de distribuição valor extremo.

Teorema 2.1.1 *(Teorema de Fisher-Tippett-Gnedenko)(FISHER; TIPPETT, 1928) Seja* (X_n) *uma sequência de variáveis aleatórias i.i.d. Se existem constantes padronizadoras* $a_n > 0, b_n \in \mathbb{R}$ *e alguma função distribuição não degenerada* G *tal que* $M_n^* = \frac{M_n - b_n}{a_n}$ *converge em distribuição para* G*, então* G *é uma das três seguintes funções de distribuição:*

$$\begin{aligned} Gumbel : G_0(z) &= exp(-\mathrm{e}^{-z}), \quad z \in \mathbb{R}, \\ Fréchet : G_{1,\alpha}(z) &= exp(-z^{-\alpha}), \quad z \geq 0, \alpha > 0, \\ Weibull\ Reversa : G_{2,\alpha}(z) &= exp(-(-z)^{\alpha}), \quad z \leq 0, \alpha < 0. \end{aligned}$$

As três funções de distribuição dadas no Teorema 2.1.1 podem ser escritas como uma única família de distribuições, tendo um novo parâmetro $\gamma = 1/\alpha$ com função de distribuição

$$G_\gamma(z) = exp(-(1 + \gamma z)^{-\frac{1}{\gamma}}), \quad 1 + \gamma z > 0. \tag{2.1}$$

A função distribuição em (2.1) pode ser generalizada, tomando $G(z) = G_\gamma(\frac{z-\mu}{\sigma})$. A nova função de distribuição será

$$G(z) = exp\left\{ -\left(1+\gamma\frac{z-\mu}{\sigma}\right)^{-\frac{1}{\gamma}}\right\}, \quad 1+\gamma(z-\mu)/\sigma > 0, \tag{2.2}$$

onde $\mu \in \mathbb{R}$ é o parâmetro de locação, $\sigma > 0$ é o parâmetro de escala e $\gamma \in \mathbb{R}$ é o índice de valor extremo. G então é dita distribuição valor extremo generalizada (GEV). O Teorema 2.1.1 pode ser reescrito então, por conveniência, em uma forma modificada

Teorema 2.1.2 *Se existem sequências de constantes $a_n > 0$ e b_n tais que*

$$P(M_n^* = M_n - b_n/a_n \leq z) \to G(z) \quad \textit{quando } n \to \infty \tag{2.3}$$

para uma função de distribuição não degenerada G, então G é da família de distribuições GEV

$$G(z) = exp\left\{ -\left(1+\gamma\frac{z-\mu}{\sigma}\right)^{-\frac{1}{\gamma}}\right\},$$

definida em $z : 1+\gamma(z-\mu)/\sigma > 0$, onde $-\infty < \mu < \infty, \sigma > 0$ e $-\infty < \gamma < \infty$.

2.2 PROPRIEDADES DA DISTRIBUIÇÃO DE VALOR EXTREMO

A partir da função distribuição em (2.3) é possível encontrar a função densidade de probabilidade da família de distribuições GEV:

$$g(z) = \frac{1}{\sigma}\left(1+\gamma\frac{z-\mu}{\sigma}\right)^{-\frac{1}{\gamma}-1} exp\left\{ -\left(1+\gamma\frac{z-\mu}{\sigma}\right)^{-\frac{1}{\gamma}}\right\}, \quad 1+\gamma(z-\mu)/\alpha > 0. \tag{2.4}$$

O suporte da função é a reta para $\gamma = 0$. Para o caso de $\gamma < 0$, o suporte da função é limitado superiormente em $\mu - \sigma/\gamma$ e limitado inferiormente em $\mu - \sigma/\gamma$ para $\gamma > 0$, como pode ser visto na Figura 1 e o gráfico da função distribuição (2.3) na Figura 2.

Para $\gamma > 0$, a distribuição aparenta ter a cauda direita pesada. De fato, se γ é positivo, todos os momentos de ordem maior ou igual a $1/\gamma$ serão infinitos. Contudo, para $\gamma \leq 0$ a distribuição possui todos os α-ésimos - $\alpha > 0$ - momentos finitos. Ou seja, dado que Y segue uma distribuição da família GEV com parâmetros γ, μ e σ, se $\gamma \leq 0$, temos que $E|Y|^\alpha < \infty$, $\forall \alpha > 0$ e para o caso de $\gamma > 0$, então $E|Y|^\alpha < \infty$, $\forall \alpha > 0$ e infinito para $\alpha > 1/\gamma$. A Tabela 1 mostra um resumo de algumas importantes medidas da GEV.

2.2.1 Domínios de Atração

Pelo Teorema 2.1.1, são possíveis 3 distribuições para as quais o máximo de uma amostra possa convergir: Gumbel, Fréchet e Weibull Reversa. Cada uma dessas distribuições representa um domínio de atração do máximo, e depende apenas do índice de

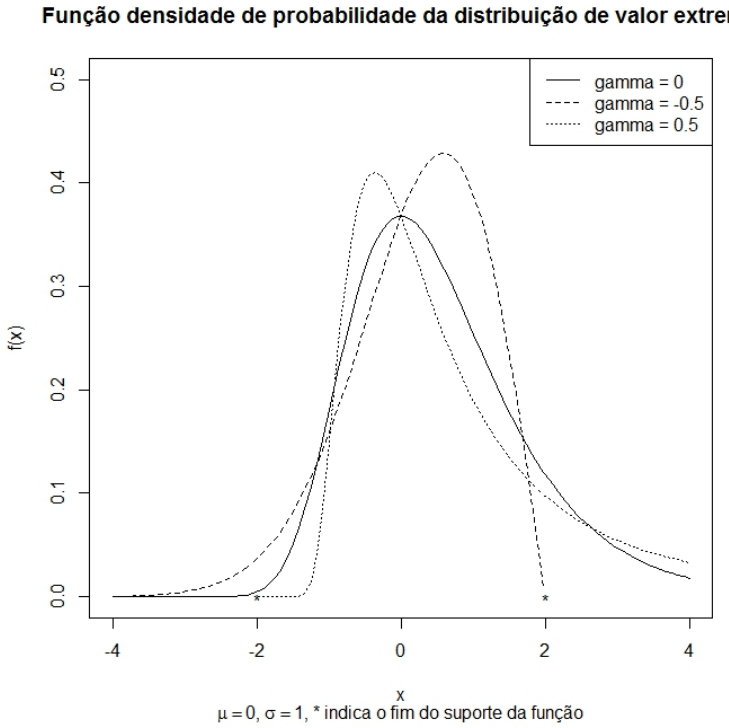

Figura 1 – Função densidade de probabilidade GEV

Tabela 1 – Medidas de resumo GEV

Medida	Valor
Média	$\mu + \sigma\frac{\Gamma(1-\gamma)-1}{\gamma}$, se $\gamma < 1$ ∞, se $\gamma \geq 1$
Mediana	$\mu + \sigma\frac{(log2)^{-\gamma}-1}{\gamma}$, se $\gamma \neq 0$ $\mu - \sigma log(log2)$, se $\gamma = 0$
Moda	$\mu + \sigma\frac{(1+\gamma)^{-\gamma}-1}{\gamma}$, se $\gamma \neq 0$ μ, se $\gamma = 0$
Variância	$\sigma^2(\Gamma(1-2\gamma) - \Gamma(1-\gamma)^2)/\gamma^2$, se $\gamma \neq 0, \gamma < \frac{1}{2}$ $\sigma^2\frac{\pi^2}{6}$, se $\gamma = 0$ ∞, se $\gamma \geq \frac{1}{2}$

valor extremo γ. Gnedenko (1943) obteve as condições para que, dada uma distribuição F_x, ela pertença a um dos três domínios de atração.

Teorema 2.2.1 *Seja X_q tal que $P(X \leq X_q) = q$, ou seja, $F_X(X_q) = q$. Para que F_x pertença a um determinado domínio de atração, é necessário e suficiente que*

- *Tipo I - Domínio de atração Gumbel*

$$\lim_{n\to\infty} n[1 - F_X(X_{1-\frac{1}{n}} + a(X_{1-\frac{1}{na}} - X_{1-\frac{1}{n}}))] = e^{-a} \tag{2.5}$$

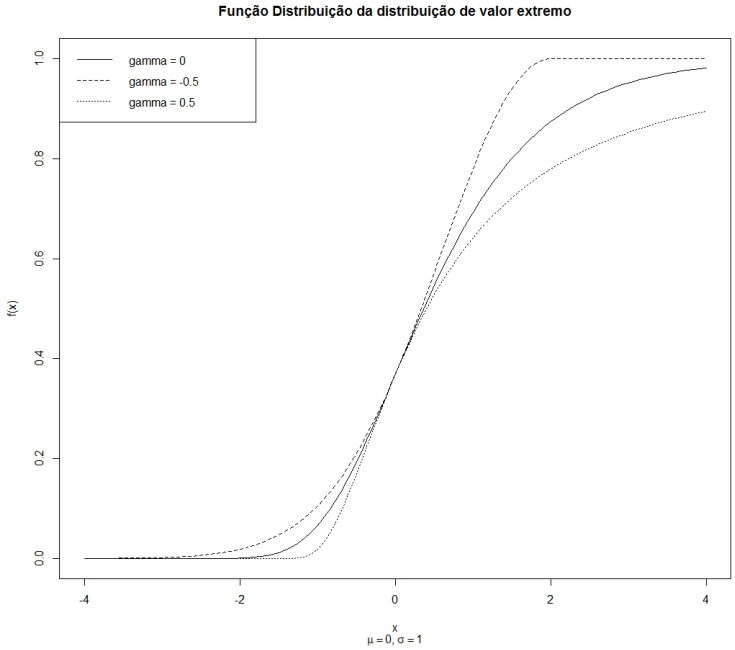

Figura 2 – Função distribuição GEV

- *Tipo II - Domínio de atração Fréchet*

$$\lim_{x\to\infty} \frac{1 - F_X(x)}{1 - F_X(cx)} = c^k, \quad c > 0, k > 0 \tag{2.6}$$

- *Tipo III - Domínio de atração Weibull Reversa*

$$\lim_{x\to o^-} \frac{1 - F_X(cx + x_{F_X})}{1 - F_X(x + x_{F_X})} = c^k, \quad c > 0, k > 0 \tag{2.7}$$

onde x_{F_x} é o limite superior da distribuição de X.

Para algumas distribuições de probabilidade, a partir dos resultados do Teorema 2.2.1 é conhecido o seu domínio de atração e ainda o índice de valor extremo, como mostrado por Charras-Garrido e Lezaud (2013) nas Tabelas 2, 3 e 4.

Tabela 2 – Uma lista de distribuições no domínio de atração tipo I - Gumbel ($\gamma = 0$)

Distribuição	1-F(x)
Weibull	$\exp(-\lambda x^\tau), \quad x > 0; \lambda, \tau > 0$
Exponencial	$\exp(-\lambda x), \quad x > 0; \lambda > 0$
Gamma	$\frac{\lambda^m}{\Gamma(m)} \int_x^\infty u^{m-1} exp(-\lambda u) du, \quad z > 0; \lambda, m > 0$
Logística	$1/(1 + exp(x)), \quad x \in \mathbb{R}$
Normal	$\int_x^\infty \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(u-\mu)^2}{2\sigma^2}) du, \quad x \in \mathbb{R}; \sigma > 0, \mu \in \mathbb{R}$
Log-Normal	$\int_x^\infty \frac{1}{\sqrt{2\pi\sigma^2 u}} exp(-\frac{(\log u-\mu)^2}{2\sigma^2}) du, \quad x > 0; \mu \in \mathbb{R}, \sigma > 0$

Tabela 3 – Uma lista de distribuições no domínio de atração tipo II - Fréchet ($\gamma > 0$)

Distribuição	1-F(x)	Índice de valor extremo
Pareto	$Kx^{-\alpha}, K, \alpha > 0$	$\frac{1}{\alpha}$
$F(m, n,)$	$\int_x^\infty \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\omega^{\frac{m}{2}-1}(1+\frac{m}{n}\omega)^{-\frac{m+n}{2}}d\omega$ $x > 0; m, n > 0$	$\frac{2}{n}$
Fréchet	1-exp(-$x^{-\alpha}$), $x > 0; \alpha > 0$	$\frac{1}{\alpha}$
T_ν	$\int_x^\infty \frac{2\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1+\frac{\omega^2}{\nu})^{-\frac{\nu+1}{2}}d\omega$ $x > 0; \nu > 0$	$\frac{1}{\nu}$

Tabela 4 – Uma lista de distribuições no domínio de atração tipo III - Weibull Reversa ($\gamma < 0$)

Distribuição	$1 - F(\omega(F) - \frac{1}{x})$	Índice de valor extremo
Uniforme	$\frac{1}{x}, \quad x > 1$	-1
Beta(p, q)	$\int_{1-\frac{1}{x}}^1 \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}u^{p-1}(1-u)^{q-1}du$ $x > 1; p, q > 0$	$-\frac{1}{q}$
Weibull Reversa	$1 - exp(-x^{-\alpha}), \quad x > 0; \alpha > 0$	$-\frac{1}{\alpha}$

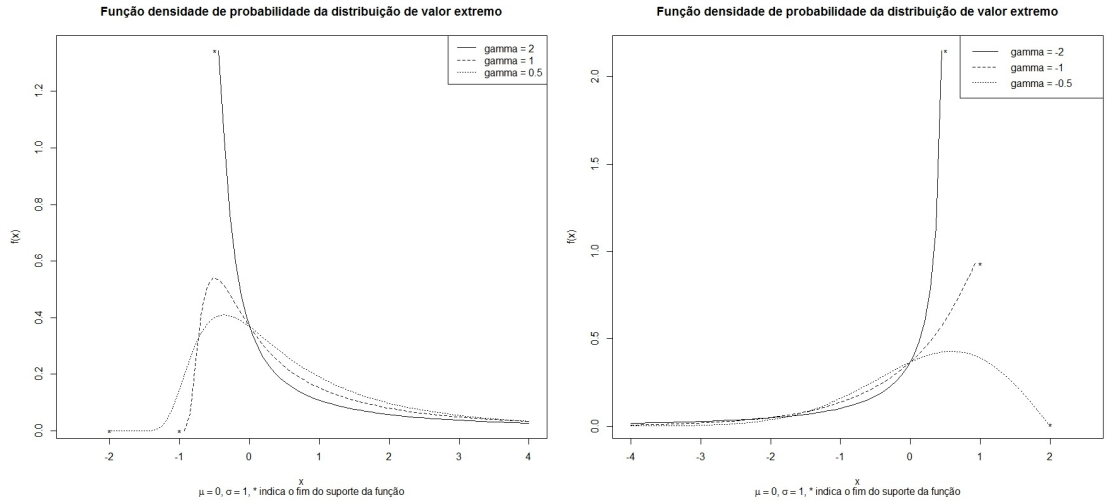

Figura 3 – Função densidade GEV para alguns $\gamma > 0$ (esquerda) e $\gamma < 0$ (direita)

A Figura 3 mostra como o parâmetro γ influencia a função densidade de probabilidade da distribuição GEV para as distribuições que estiverem no domínio de atração Fréchet e Weibull Reversa.

Algumas funções de distribuição não satisfazem nenhuma das condições do Teorema 2.2.1 e então o máximo não converge para uma distribuição não-degenerada, pois as três distribuições são as únicas max-estáveis. Esse é o caso, por exemplo, da distribuição de Poisson, cuja distribuição do máximo não pode ser aproximada pela família de distribuições

de valor extremo.

2.3 INFERÊNCIA PARA DISTRIBUIÇÕES DE VALOR EXTREMO

Pickands (1975) interpreta o Teorema 2.1.2 como uma aproximação para valores grandes de n, o uso da família de distribuições GEV parece sugestivo para a modelagem da distribuição do máximo de longas sequências de dados. O fato de as constantes padronizadoras a_n e b_n serem desconhecidas, a primeiro momento, parece ser uma dificuldade para a estimação, contudo esta questão pode ser facilmente resolvida. Assumindo (2.3) como uma aproximação

$$P((M_n - b_n)/a_n \leq z) \approx G(z)$$

para n suficientemente grande. Equivalentemente,

$$\begin{aligned} P(M_n < z) &\approx G((z - b_n)/a_n) \\ &= G^*(z), \end{aligned}$$

sendo G^* um outro membro da família de distribuições GEV. Ou seja, se o Teorema 2.1.2 permite a aproximação de M_n^* por uma distribuição da família GEV para n suficientemente grande, então o próprio M_n pode ser aproximado por um outro membro da família de distribuições GEV.

A partir deste argumento, pode-se realizar a seguinte abordagem para a modelagem de uma série de observações independentes $x_1, x_2, ...$ Os dados são divididos em blocos de sequência de observações de algum tamanho suficientemente grande j, gerando uma série de máximos de blocos

$$\begin{aligned} m_1 &= max\{x_1, ...x_j\} \\ m_2 &= max\{x_{j+1}, ...x_{2j}\} \\ &\vdots \\ m_k &= max\{x_{(k-1)j+1}, ...x_{kj}\}. \end{aligned}$$

O valor de j pode ser um número arbitrário ou escolhido de acordo com a natureza dos dados. Por exemplo, para dados climáticos muitas vezes pode ser interessante utilizar a máxima anual a fim de retirar os efeitos causados pela sazonalidade, já para dados do mercado financeiro, a máxima mensal ou trimestral é comumente utilizada. Através desta amostra dos máximos de blocos $m_1, m_2, ..., m_k$ podem ser estimados os parâmetros para a distribuição G da família GEV. A escolha de j muitas vezes resulta em um *trade-off*, pois um valor muito pequeno de j pode resultar na não convergência do modelo ou um viés maior nas estimativas. Já um valor muito grande de j ocasiona um número menor de valores de máximos de blocos utilizados na estimação do modelo, o que resulta em uma

maior variância do estimador, uma vez que o número k de máximo de blocos é o tamanho efetivo da amostra utilizada para estimar os parâmetros do modelo GEV.

Muitas vezes, em aplicações práticas, a suposição de independência das observações $x_1, ..., x_n$ no Teorema 2.1.1 não é realística. Se a sequência de observações $x_1, ...x_n$ for estritamente estacionária e possuir dependência temporal fraca (ver em Leadbetter (1983)), é possível escolher um tamanho dos blocos j que possa assumir que os máximos de bloco $m_1, ..., m_k$ sejam independentes e identicamente distribuídos. Tal propriedade pode ser verificada através do gráfico da função de autocorrelação empírica, dada por

$$\hat{R}(t) = \frac{1}{(k-t)S^2} \sum_{i=1}^{k-t} (m_i - \bar{m})(m_{i+t} - \bar{m}), \tag{2.8}$$

onde $\bar{m}$ é a média amostral dos máximos de bloco e S^2 é a variância amostral. Caso haja evidências de autocorrelação, é necessário aumentar o tamanho dos blocos.

2.3.1 Gráfico do Nível de Retorno

Estimativas dos quantis extremos da distribuição do máximo são obtidas invertendo (2.3):

$$z_p = \begin{cases} \mu - \frac{\sigma}{\gamma}[1 - \{-\log(\text{1-p})\}^{-\gamma}], & \gamma \neq 0 \\ \mu - \sigma \log\{-\log(1-p)\}, & \gamma = 0 \end{cases} \tag{2.9}$$

aonde $G(z_p) = (1-p)$. z_p é dito o Nível de Retorno associado com o Período de Retorno $1/p$, pois a probabilidade de o máximo de j observações exceder z_p será p. Logo, a cada $1/p$ blocos de n observações, é esperado que em uma, o máximo exceda z_p. Como os quantis permitem aos modelos probabilísticos serem expressos na escala dos dados, a relação do modelo GEV com seus parâmetros é mais facilmente interpretada utilizando os quantis em (2.9). Especialmente, definindo $w_p = -log(1-p)$, tal que

$$z_p = \begin{cases} \mu - \frac{\sigma}{\gamma}[1 - w_p^{-\gamma}], & \gamma \neq 0, \\ \mu - \sigma \log w_p, & \gamma = 0, \end{cases}$$

segue que um gráfico de z_p contra w_p em escala logarítmica, ou seja z_p contra $\log w_p$, será linear caso $\gamma = 0$. Se $\gamma < 0$, o gráfico é convexo com limite assintótico $p \to 0$ em $\mu - \sigma/\gamma$. Se $\gamma > 0$ o gráfico é côncavo sem limite finito, como ilustado pela Figura 4. Esse gráfico é chamado gráfico do nível de retorno, sendo a principal ferramenta utilizada para interpretar uma modelagem de valores extremos devido a sua simplicidade de entendimento e à escala que comprime a cauda à esquerda, destacando o efeito da extrapolação (COLES, 2001).

2.4 ESTIMAÇÃO

Existe uma grande variedade de métodos de estimação para os parâmetros das distribuições da família GEV, visto que não há um estimador que possua as propriedades

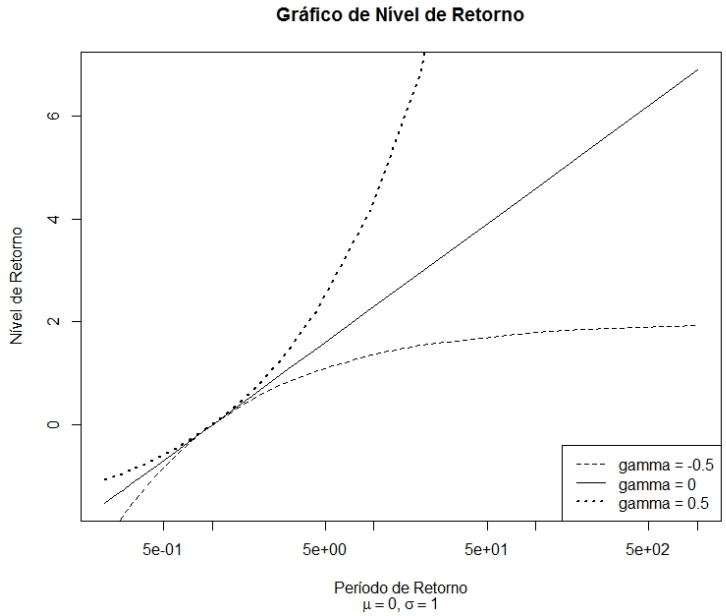

Figura 4 – Gráfico do Nível de Retorno da distribuição GEV para diferentes valores de γ

satisfatórias para todos os valores de γ. O Estimador de Máxima Verossilhança e o Estimador de Momentos Ponderados são os mais utilizados devido a sua simplicidade de cálculo e serão apresentados nessa seção.

2.4.1 Estimador de Máxima Verossimilhança (EMV)

Devido ao suporte da função distribuição depender dos parâmetros desconhecidos (2.3), o modelo não satisfaz às condições de regularidade que implicam nas propriedades assintóticas dos estimadores de máxima verossimilhança. Ainda assim, foi mostrado por Smith (1985) que para $\gamma > -1/2$ as propriedades usuais de consistência, eficiência assintótica e normalidade assintótica são válidas.

A partir da função densidade de probabilidade (2.4), podemos encontrar a função de log-verossimilhança para uma amostra de máximo de blocos $m_1, ..., m_k$

$$l(\sigma, \mu, \gamma; m_1, ..., m_k) = -k \log \sigma - \left(\frac{1}{\gamma}+1\right) \sum_{i=1}^{k} \log\left(1+\gamma\frac{m_i-\mu}{\sigma}\right) - \sum_{i=1}^{k} \log\left(1+\gamma\frac{m_i-\mu}{\sigma}\right)^{-\frac{1}{\gamma}},$$

se $1 + \gamma(m_i - \mu)/\sigma > 0 \ \forall i$.

Mesmo não existindo expressão analítica para os estimadores de máxima verossimilhança, através de métodos numéricos de otimização, pode-se encontrar a localização do máximo da função de log-verossimilhança. Para construção de intervalos de confiança, é possível utilizar a propriedade de normalidade assintótica do estimador, provada por Zhou (2009), para $\gamma > -1$. Apesar de ser um estimador simples de computar e possuir propriedades assintóticas desejáveis, essas propriedades são válidas apenas no caso de

$\gamma > -1$. Na prática, contudo, o valor real do parâmetro é desconhecido, fazendo com que a acurácia da estimação possa não ser bem avaliada. Para lidar com este problema, estimadores alternativos foram propostos.

2.4.2 Método de Momentos Ponderados por Probabilidade

Introduzido por Greenwood et al. (1979), os momentos ponderados por probabilidades são as quantidades $M_{p,r,s} = E(X^p F^r(X)(1 - F(X))^s)$, para p, r e s reais, sendo que, com $r = s = 0$ são obtidos os momentos convencionais. Para $\gamma < 1$ obtemos para a família de distribuições GEV, com $p = 1$ e $s = 0$,

$$M_{1,r,0} = \frac{1}{r+1}\left(b - \frac{a}{\gamma}[1 - (r+1)^{\gamma}\Gamma(1-\gamma)]\right).$$

Estes momentos podem ser estimados através de uma amostra de máximo de blocos para assim estimar os parâmetros a, b, e γ. Como não existe expressão analítica para o estimador de γ, este tem que ser computado numericamente. Para $\gamma \geq 1$ a distribuição de valor extremo generalizada não possui momentos ou momentos ponderados por probabilidade.

Essa abordagem é largamente usada devido à sua simplicidade conceitual e fácil implementação, além de apresentar melhores resultados para pequenas amostras do que o EMV. Ainda sim, por ser limitado por $\gamma < 1$, ele não é aplicável em situações de caudas realmente pesadas. Além disto, a propriedade de normalidade assintótica do estimador é válida apenas para $\gamma \in (-1, 1/2)$.

Uma modificação deste método, o método de L-momentos (HOSKING; WALLIS, 1990), possui vício e variância um pouco menores e é mais robusto em relação a presença de *outliers*.

Outros métodos que merecem destaque são o estimador de Hill (HILL, 1975) para $\gamma > 0$ e o estimador baseado em estatísticas de ordem, proposto por Jansen e Vries (1991), que é consistente para todo γ real e robusto à não independência dos máximos de bloco amostrais.

2.5 QUALIDADE DO AJUSTE

Quando se ajusta um modelo para os dados, busca-se uma estimação da função de distribuição desconhecida G da qual as amostras são suas realizações. Ajustando um modelo GEV aos dados, obtém-se uma estimativa $\hat{G}$ para G. A fim de verificar se a amostra dos máximos parciais $m_1, ..., m_k$ pode ser uma amostra aleatória da distribuição estimada $\hat{G}$, pode-se compará-la com uma estimativa de G independente de modelos, obtida empiricamente através dos dados.

Definição 2.5.1 *Dada uma amostra de observações independentes*

$$m_{(1)} \leq m_{(2)} \leq ... \leq m_{(k)}$$

de uma população com função de distribuição G, a ***função distribuição empírica*** *é definida por*

$$\tilde{G}(m) = \frac{i}{k+1} \quad para\ m_{(i)} \leq m \leq m_{(i+1)}, i = 1, ..., k.$$

Como $\tilde{G}$ é uma estimativa para a verdadeira probabilidade de distribuição G, ela deve ser razoavelmente condizente com $\hat{G}$, se $\hat{G}$ for uma estimativa adequada para G. Vários procedimentos de qualidade de ajuste são baseados na comparação entre $\tilde{G}$ e $\hat{G}$, particularmente duas técnicas gráficas que são comumente utilizadas.

Definição 2.5.2 *Dada uma amostra de observações independentes*

$$m_{(1)} \leq m_{(2)} \leq ... \leq m_{(k)}$$

de uma população com função de distribuição G, um gráfico PP plot consiste dos pontos

$$\left\{ \left(\hat{G}(m_{(i)}), \frac{i}{k+1} \right) \quad ;\ i = 1, ..., k \right\}.$$

Sendo $\hat{G}$ um modelo razoável para a distribuição populacional, os pontos do gráfico devem estar perto da diagonal unitária. Desvios substanciais da linearidade são evidências da falha de $\hat{G}$ como modelo para os dados.

Definição 2.5.3 *Dada uma amostra de observações independentes*

$$m_{(1)} \leq m_{(2)} \leq ... \leq m_{(k)}$$

de uma população com função de distribuição G, um gráfico QQ plot consiste dos pontos

$$\left\{ \left(\hat{G}^{-1}\left(\frac{i}{k+1} \right), m_{(i)} \right) \quad ;\ i = 1, ..., k. \right\}.$$

O nome *QQ plot* vem do fato de que as quantidades $m_{(i)}$ e $\hat{G}^{-1}(\frac{i}{k+1})$ são estimativas para o quantil $i/(k+1)$ da distribuição G. Se $\hat{G}$ é uma estimativa razoável para G, então os pontos do *QQ plot* também devem estar perto da diagonal unitária.

Ambos os gráficos possuem a mesma informação, mas em diferentes escalas. A importância de se utilizar os dois é que o que pode parecer um ajuste razoável em uma escala pode parecer ruim em outra (REISS; THOMAS, 2007).

Além de métodos gráficos, existem também testes de hipóteses para verificar o ajuste do modelo baseados na comparação entre uma distribuição hipotética $\bar{G}$ com parâmetros

θ e a função distribuição empírica $\tilde{G}$. A discrepância entre estas duas distribuições pode ser medida através de estatísticas quadráticas

$$Q^2 = k \int_{\forall m} [\tilde{G}(m) - \bar{G}(m, \theta)]^2 \Psi(m) \mathrm{d}\bar{G}, \tag{2.10}$$

onde $\Psi(m)$ é uma função peso. Quando $\Psi(m) = 1$, Q^2 é a estatística de Cramer-von Mises, usualmente chamada de W^2, diferença quadrática média entre a distribuição empírica e a hipotética. Quando $\Psi(m) = [\bar{G}(m, \theta)(1 - \bar{G}(m, \theta))]^{-1}$, as caudas da distribuição possuem peso maior que a parte central e esta é a estatística de Anderson-Darling, denominada A^2. No caso em que o conjunto de parâmetros θ da distribuição a ser testada é desconhecido, os parâmetros devem ser substituídos por suas estimativas $\hat{\theta}$. Então sua distribuição hipotética $\bar{G}$ passa a ser sua distribuição estimada $\hat{G}$. Com a transformação $t = \hat{G}(m, \hat{\theta})$, a estatística quadrática geral em (2.10) se torna

$$Q_p^2 = k \int_0^1 [\tilde{F}(m) - t]^2 \Psi(m) \mathrm{d}t.$$

Nesse caso, a distribuição de Q_p^2 não é independente de $\hat{F}$ e por isso sua distribuição se diferencia do caso em que o conjunto de parâmetros θ é conhecido. Uma transformação para Q_p^2 que a torna independente de $\hat{F}$ para as distribuições da família GEV e tabelas para os testes de Cramer-von Mises e Anderson-Darling são apresentadas por Laio (2004) para o caso de estimadores assintoticamente eficientes para o conjunto de parâmetros $\theta = \mu, \sigma$ e γ, permitindo a realização de ambos os testes para verificar a adequação do modelo GEV estimado aos dados.

Outra forma também importante de analisar o ajuste do modelo é através da análise dos resíduos. Sob a hipótese nula de que a amostra de máximos de blocos é uma realização de uma distribuição GEV, os resíduos

$$w_i = \left(1 + \hat{\gamma} \frac{m_i - \hat{\mu}}{\hat{\sigma}}\right)^{-\frac{1}{\hat{\gamma}}}, \quad i = 1, ..., n,$$

seguem uma distribuição Exponencial(1)(HOSKING; WALLIS, 1997). Uma técnica gráfica de análise de resíduos é um gráfico *PP plot* para w_i. Além disso, pode-se testar o ajuste desses resíduos à distribuição Exponencial(1) através do teste de Kolmogorov-Smirnov. A rejeição do teste evidencia que o modelo não representa bem os dados.

3 DISTRIBUIÇÃO PARETO GENERALIZADA (GPD)

A modelagem através do máximo de blocos pode ser uma abordagem para análise de valores extremos que gera um desperdício de dados se outros dados nos extremos estão disponíveis, ou seja, se há mais de um valor extremo no mesmo bloco. Isto ocorre principalmente quando há necessidade de utilizar blocos de muitas observações. Como as observações nos extremos são escassas, o ideal é utilizar todas disponíveis. É natural que sejam chamadas de observações extremas observações cujos valores excedam um limiar alto u. O Teorema 3.1.1 relaciona a distribuição dessas observações com a distribuição de valor extremo generalizada e permite uma abordagem que aproveita melhor a informação disponível sobre os extremos.

3.1 FORMULAÇÃO DO MODELO

Teorema 3.1.1 *Seja $X_1, \ldots, X_n$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas com função de distribuição comum F_X, denotando M_n o máximo das n variáveis aleatórias, ou seja,*

$$M_n = max(X_1, \ldots, X_n).$$

Supondo que F_X satisfaça as condições do Teorema 2.2.1, então para n suficientemente grande,

$$P(M_n \leq z) \approx G(z)$$

onde

$$G(z) = exp\left\{ -\left(1 + \gamma\frac{z-\mu}{\sigma}\right)^{-\frac{1}{\gamma}} \right\}$$

para algum $\mu, \sigma > 0$ e γ. Então, para u suficientemente grande, a distribuição de $Y = (X - u)$, condicionada em $X > u$, é aproximadamente

$$H(y) = 1 - \left(1 + \frac{\gamma y}{\tilde{\sigma}}\right)^{-1/\gamma}, \tag{3.1}$$

definida em $\{y{:}y{>}0$ e $(1 + \gamma y/\tilde{\sigma}) > 0\}$, onde

$$\tilde{\sigma} = \sigma + \gamma(u - \mu).$$

A prova desse teorema é apresentada por Leadbetter, Lindgren e Rootzen (1983).

A família de distribuições definida em (3.1) é denominada Pareto generalizada (GPD), primeiramente definida por Pickands (1975). O Teorema 3.1.1 implica que se os máximos de blocos possuem distribuição aproximada G, então os excessos sob limiar possuem uma distribuição aproximada dentro da família Pareto generalizada correspondente. Além disto, os parâmetros da distribuição Pareto generalizada são unicamente

determinados por aqueles da distribuição GEV associada. Em particular, o parâmetro γ em (3.1) é igual ao da distribuição GEV correspondente.

Denota-se por N_u o número de observações que excedem o limiar u, isto é, $N_u = \sum_{i=1}^{n} 1_{(X_i>u)}$, onde $1_{(X_i>u)} = 1$ se $X_i > u$ e 0 caso contrário. Os excessos além do limiar u, denotados por $Y_1, ..., Y_{N_u}$ são os valores $X_i - u \geq 0$. Para um limiar u fixo, N_u também é uma variável aleatória. Como $X_1, X_2, ...X_n$ possuem distribuição comum F_x, então

$$P(N_u > t) = \binom{n}{t}(1 - F_X(u))^t F_X(u)^{n-t}, \quad t = 0, 1, ..., n.$$

Ou seja, N_u segue uma distribuição Binomial$(n, 1 - F_X(u))$ e o número médio de excessos além de u será seu valor esperado

$$E[N_u] = n(1 - F_X(u)),$$

que é uma função decrescente em u. N_u será o tamanho de amostra efetivo utilizado para estimar os parâmetros do modelo GPD.

3.2 PROPRIEDADES DA DISTRIBUIÇÃO PARETO GENERALIZADA

Com a função distribuição em (3.1) pode-se encontrar a função densidade de probabilidade $h(y)$ da família de distribuições GPD

$$h(y) = \frac{1}{\tilde{\sigma}}\left(1 + \frac{\gamma y}{\tilde{\sigma}}\right)^{(-\frac{1}{\gamma}-1)}, \tag{3.2}$$

para $y > 0$ quando $\gamma \geq 0$ e $0 < y \leq -\tilde{\sigma}/\gamma$ quando $\gamma < 0$. A Figura 5 mostra o gráfico da função densidade e a Figura 6 mostra o gráfico da função distribuição (3.1).

Há também uma relação analítica entre a função distribuição H (3.1) e a função distribuição G da GEV correspondente (2.3), dada por

$$H_\gamma(x) = 1 + log(G_\gamma(x)), \quad log(G_\gamma(x)) > -1.$$

Tal relação explica o fato da densidade da GPD possuir cauda extrema assintoticamente equivalente à da GEV correspondente, como ilustrado na Figura 7.

A Tabela 5 apresenta um resumo de algumas importantes medidas da GPD.

3.2.1 Domínios de atração

Assim como no caso GEV, existem 3 casos possíveis para as distribuições limites das excedências de um limiar. Para domínio tipo I ($\gamma = 0$), a distribuição se torna

$$H(y) = 1 - e^{-\frac{y}{\tilde{\sigma}}}, y > 0,$$

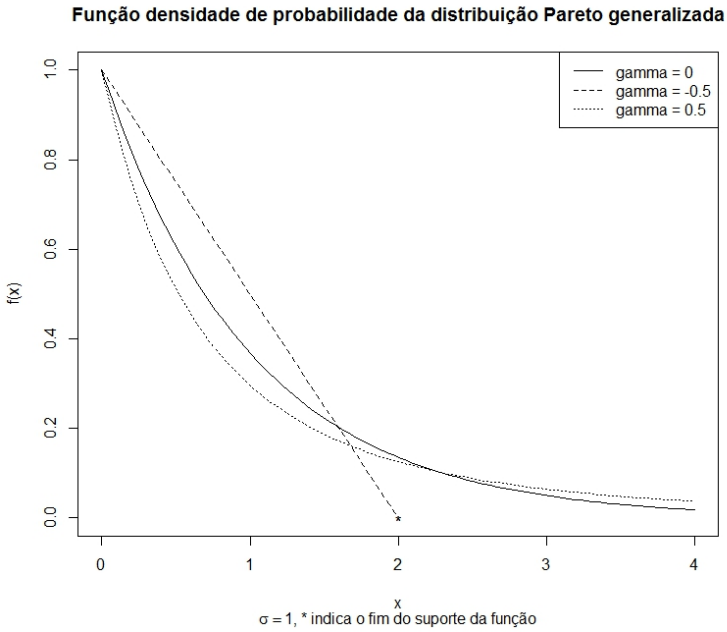

Figura 5 – Função densidade de probabilidade da GPD

sendo assim, o domínio, uma distribuição Exponencial com parâmetro $\frac{1}{\tilde{\sigma}}$. Para o domínio tipo II ($\gamma > 0$), a distribuição limite será a distribuição de Pareto. Já para o domínio tipo III ($\gamma < 0$), quando $\tilde{\sigma} = -\frac{1}{\gamma}$, a distribuição limite será uma Beta e quando $\tilde{\sigma} \neq -\frac{1}{\gamma}$, a distribuição limite será uma Beta reescalada com suporte em $(0, \tilde{\sigma}/\gamma)$.

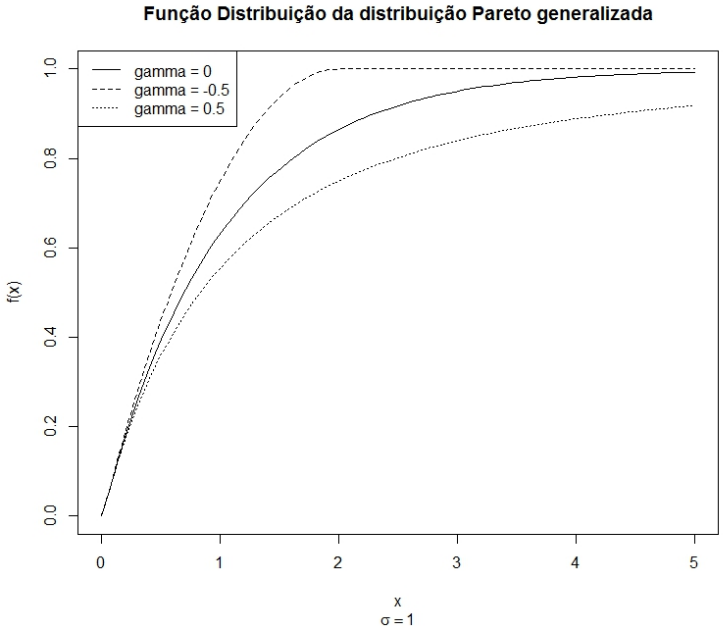

Figura 6 – Função distribuição da GPD

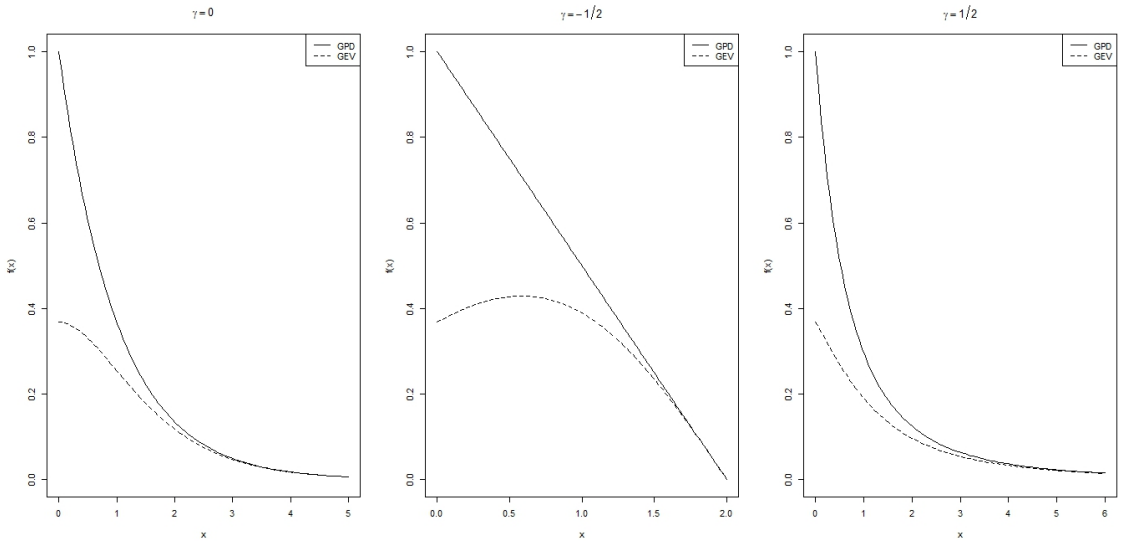

Figura 7 – Comparação das densidades de GPD e GEV equivalentes com $\mu = 0$ e $\sigma = 1$

Tabela 5 – Medidas de resumo GPD

Medida	Valor
Média	$\frac{\tilde{\sigma}}{1-\gamma}, \quad \gamma < 1$
Mediana	$\frac{\tilde{\sigma}(2^{\gamma}-1)}{\gamma}$
Variância	$\frac{\tilde{\sigma}}{(1-\gamma)^2(1-2\gamma)}$

Como o parâmetro γ da distribuição GPD equivale ao da GEV, pelo Teorema 3.1.1 -que estabelece a relação entre as 2 distribuições-, tem-se que as condições do Teorema 2.2.1 são válidas de maneira análoga para a convergência para os domínios de atração da GPD e as distribuições apresentadas nas Tabelas 2, 3 e 4 terão a distribuição dos excedentes de um limiar convergindo para o domínio equivalente da família de distribuições GPD.

3.3 ESCOLHA DO LIMIAR u

Assim como a escolha dos tamanho dos blocos j para o modelo GEV, a escolha do limiar u também é uma parte crucial na modelagem dos excessos por uma distribuição GPD. Um valor de u pequeno pode não garantir a convergência dos excessos Y para a família de distribuições Pareto generalizada, levando a um vício alto. Entretanto, à medida que se aumenta o valor de u, a amostra efetiva N_u diminui, causando um aumento na variância das estimativas. Novamente, a escolha de u leva a um *trade-off* entre vício e variância.

Definindo

$$e(u) = E[Y] = E[X - u|X > u]$$

como a função Média dos Excessos, ou seja, a esperança condicional dos excessos além de um limiar u, temos que se Y segue uma distribuição da família GPD com $\gamma < 1$, então

$$e(u) = \frac{\tilde{\sigma} + \gamma u}{1 - \gamma}, \quad \tilde{\sigma} + \gamma u > 0,$$

que é uma função linear em u. É possível então utilizar o gráfico da função Média dos Excessos empírica, $e_{N_u}(u)$, ou sua versão robusta a função Mediana dos Excessos empírica, $e^*_{N_u}(u)$, versus o limiar u. Para uma amostra $X_1, ...X_n$, essas funções são definidas por

$$e_{N_u}(u) = \frac{1}{N_u} \sum_{i}^{N_u} Y_i(u),$$

$$e^*_{N_u}(u) = Mediana\{Y_i(u), \quad i = 1, ..., N_u\}.$$

Estes gráficos servem como uma técnica exploratória para a definição do limiar u, pois para o intervalo em que o gráfico aparentar comportamento linear, há indício de que o modelo GPD ajusta bem para aqueles valores de u. Além disso, existe uma técnica complementar que consiste em ajustar modelos GPD para um intervalo de valores para u e procurar estabilidade nas estimativas dos parâmetros, já que se uma distribuição da família GPD é um modelo razoável para um u_0, ela também será para $u > u_0$ com mesmo parâmetro γ. Já o parâmetro de escala será

$$\tilde{\sigma}_u = \tilde{\sigma}_{u_0} + \gamma(u - u_0),$$

ou seja, função linear de u se $\gamma \neq 0$. Então se u_0 é um limiar válido para os excessos, as estimativas de γ devem ser constantes e as de $\tilde{\sigma_u}$ apresentarem comportamento linear para valores de u maiores que u_0. Isto sugere que gráficos de $\hat{\sigma_u}$ e $\hat{\gamma}$ contra u, juntamente com os respectivos intervalos de confiança, são uma técnica para encontrar o valor u_0. Métodos para essas estimações serão apresentados na Seção 3.4.

Muitas vezes, os dados que se deseja modelar provém de uma série temporal e possuem estrutura de autocorrelação, não sendo independentes. Nesse caso, para utilizar os resultados do Teorema 3.1.1 é necessário averiguar se os excessos de limiar não possuem estrutura de autocorrelação, o que pode ser feito pelo gráfico da função de autocorrelação dado em (2.8). Caso esse gráfico indique que há uma estrutura de autocorrelação entre os excessos de limiar, é necessário amostrar os excessos com um espaçamento entre eles para o qual não haja mais indícios de autocorrelação. Tal técnica é chamada desclusterização (DAVISON; SMITH, 1990).

3.3.1 Gráfico do Nível de Retorno

A interpretação dos modelos para valores extremos é mais conveniente quando feita em termos de quantis ou níveis de retorno, como discutido na Seção 2.3.1. Então supondo

que uma distribuição GPD com parâmetros γ e $\tilde{\sigma}$ é um modelo adequado para os excessos de um limiar u para uma variável X, então para $x > u$,

$$P(X > x|X > u) = \left[1 + \gamma\left(\frac{x-u}{\tilde{\sigma}}\right)\right]^{-1/\gamma}.$$

Segue que

$$P(X > x) = \zeta_u\left[1 + \gamma\left(\frac{x-u}{\tilde{\sigma}}\right)\right]^{-1/\gamma},$$

onde $\zeta_u = P(X > u)$. Dessa forma, o nível x_v que é excedido em média a cada v observações é a solução de

$$\zeta_u\left[1 + \gamma\left(\frac{x_v-u}{\tilde{\sigma}}\right)\right]^{-1/\gamma} = \frac{1}{v}.$$

Reorganizando,

$$x_v = \begin{cases} u + \frac{\tilde{\sigma}}{\gamma}[(v\zeta_u)^\gamma - 1] & \gamma \neq 0, \\ u + \tilde{\sigma}\log(v\zeta_u) & \gamma = 0, \end{cases} \tag{3.3}$$

dado que v seja suficientemente grande para assegurar que $x_v > u$.

O gráfico do nível de retorno é composto dos valores de x_v contra v em escala logarítmica e possui as mesmas características dos gráficos do nível de retorno baseados nos modelos GEV: linearidade se $\gamma = 0$, concavidade se $\gamma > 0$ e convexidade se $\gamma < 0$.

3.4 ESTIMAÇÃO

Vários estimadores foram propostos para os parâmetros da distribuição GPD, sempre tentando superar as limitações dos métodos existentes. Neste capítulo serão apresentados os dois métodos mais utilizados, devido à sua simplicidade: o Estimador de Máxima Verossimilhança e o Método de Momentos Ponderados por Probabilidade.

3.4.1 Estimador de Máxima Verossimilhança

O estimador de máxima verossimilhança é altamente utilizado para os parâmetros da distribuição GPD devido à sua facilidade de cálculo e boas propriedades para $\gamma > -\frac{1}{2}$ (SMITH, 1987). A partir da equação da função densidade de probabilidade da família de distribuições Pareto generalizada (3.2), é possível encontrar a função de log-verossimilhança para uma amostra de excessos de um limiar u, $y_1, y_2, ..., y_{N_u}$

$$l(\tilde{\sigma}, \gamma; y_1, ..., y_{N_u}) = -N_u log\tilde{\sigma} - \left(\frac{1}{\gamma} + 1\right)\sum_{i=1}^{N_u} log\left(1 + \frac{\gamma}{\tilde{\sigma}}y_i\right), \tag{3.4}$$

dado que $(1 + \gamma y_i/\tilde{\sigma}) > 0$ para $i = 1, ..., N_u$. As estimativas serão então os valores de γ e $\tilde{\sigma}$ que maximizam (3.4). Novamente, não é possível encontrar esses valores analiticamente, mas métodos numéricos conseguem resolver esse problema com facilidade.

Intervalos de confiança são construídos para $\gamma > -\frac{1}{2}$ através das propriedades assintóticas de eficiência e normalidade. Baseado no estimador de máxima verossimilhança, Smith (1987) propõe um estimador que é altamente eficiente caso a escolha do limiar u seja ótima.

3.4.2 Método de Momentos Ponderados por Probabilidade

Assim como mostrado na Seção 2.4.2 para as distribuições da família GEV, também é possível encontrar os momentos ponderados por probabilidade $M_{p,r,s} = E(X^p F^r(X)(1 - F(X))^s)$ para distribuições da família GPD. Assim como no caso GEV, esses momentos só existem para $\gamma < 1$. Para uma distribuição GDP, com $p = 1$ e $r = 0$,

$$M_{1,0,s} = \frac{\tilde{\sigma}}{(s+1)(s+1-\gamma)}.$$

Estimando esses momentos a partir de amostras de excessos de um limiar u, obtém-se estimativas para os parâmetros $\tilde{\sigma}$ e γ. Para o caso GPD, diferentemente do caso GEV, existe uma expressão analítica dada por Hosking e Wallis (1987)

$$\hat{\gamma}_{PWM}(N_u) = 2 - \frac{\hat{M}_{1,0,0}}{\hat{M}_{1,0,0} - 2\hat{M}_{1,0,1}}, \text{ com } \hat{M}_{1,0,s} = \frac{1}{k}\sum_{i=1}^{N_u}\left(1 - \frac{i}{N_u+1}\right)^s Y_{i,n}.$$

Esse método é popular devido ao seu conceito simples, fácil implementação e boa performance para amostras pequenas. Apesar disso, ele não se aplica para distribuições com caudas muito pesadas, devido à limitação de $\gamma < 1$ e possui normalidade assintótica apenas no intervalo $\gamma \in (-1, 1/2)$. Para a solução desses problemas, Diebolt, Guillou e Rached (2007) propuseram estimadores de momentos ponderados por probabilidade generalizados, que existem para distribuições da família GPD com $\gamma < 2$ e são assintoticamente normais para $\gamma \in (-1, 3/2)$.

Outros métodos que merecem destaque são: o estimador de Pickands (PICKANDS, 1975), que consiste em um procedimento simultâneo para a escolha do limiar u e estimação dos parâmetros γ e $\tilde{\sigma}$, não sendo muito utilizado por possuir uma variância assintótica grande e os estimadores bayesianos quasi-conjugados para distribuições GPD com caudas pesadas, $\gamma > 0$, propostos por Diebolt et al. (2005).

3.5 QUALIDADE DO AJUSTE

Para verificar o ajuste dos parâmetros estimados $\hat{\sigma}$ e $\hat{\gamma}$ de um modelo GPD ajustado a uma amostra $y_1, y_2, ..., y_{N_u}$ de excessos de um limiar u são utilizados métodos gráficos como na Seção 2.5 para o modelo GEV. Um gráfico *PP-plot* consistirá dos pontos

$$\{(i/(N_u+1), \hat{H}(y_{(i)})); i = 1, ..., N_u\},$$

onde

$$\hat{H}(y) = 1 - \left(1 + \frac{\hat{\gamma}y}{\hat{\sigma}}\right)^{-1/\hat{\gamma}}.$$

Um gráfico *QQ-plot* consistirá dos pares

$$\{(\hat{H}^{-1}(i/(N_u+1)), y_{(i)}), i = 1, ..., N_u\},$$

onde

$$\hat{H}^{-1} = u + \frac{\hat{\sigma}}{\hat{\gamma}}\Big[y^{-\hat{\gamma}} - 1\Big].$$

Em ambos so casos, os pontos devem estar próximos a uma reta com inclinação 1. Desvios substanciais dessa reta são indícios de mal ajuste, ou seja, o modelo não explica bem os dados. Ambos os gráficos são importantes, pois as diferentes escalas tornam visíveis diferenças que não seriam observáveis com a utilização de apenas uma. Além desses dois gráficos, também é possível verificar o ajuste com um gráfico sobrepondo o histograma dos excessos de limiar y_i com a densidade estimada $\hat{h}(y)$ e observando visualmente a diferença entre eles. A densidade estimada será

$$\hat{h}(y) = \frac{1}{\hat{\sigma}}\left(1 + \frac{\hat{\gamma}y}{\hat{\sigma}}\right)^{(-\frac{1}{\hat{\gamma}}-1)}.$$

Os métodos gráficos mencionados fornecem pistas do ajuste, mas não proporcionam uma definição objetiva do que seria um bom ajuste. Para isso existem testes de hipóteses assim como no caso da distribuição GEV. Choulakian e Stephens (2001) apresentam os testes de Cramer von-Mises e Anderson-Darling baseados na estatística 2.10 para distribuições da família GPD com parâmetros desconhecidos. A estatística de teste para Anderson-Darling, dada uma amostra $y_1, y_2, ..., y_{N_u}$ de excessos de limiar com uma distribuição ajustada $\hat{H}(y)$, será

$$A^2 = -n - (1/n)\sum_{i=i}^{N_u}(2i-1)[log(\hat{H}(y_{(i)} + log(1 - \hat{H}(y_{(n+1-i)})].$$

A Tabela 6 mostra os valores assintóticos ($N_u \geq 25$) z para os quais $P(A^2 > z) = \alpha$ para alguns valores de α e $\hat{\gamma}$ selecionados. A estatística A^2 então deve ser comparada com os valores de z da Tabela e caso seja maior para um nível de significância α, a hipótese nula de que a amostra $y_1, y_2, ..., y_{N_u}$ segue uma distribuição da família GPD deve ser rejeitada. Para valores intermediários de $\hat{\gamma}$, os valores z podem ser obtidos por interpolação linear com um erro máximo para α entre 0,0011 e 0,0003 e para $\hat{\gamma} > 0.5$ devem ser usados os mesmos valores para $\hat{\gamma} = 0.5$.

Tabela 6 – Pontos de porcentagem assintóticos para cauda superior de A^2

$\hat{\gamma}\backslash\alpha$	0,10	0,05	0,01
-0,90	0,641	0,771	1,086
-0,50	0,685	0,830	1,180
-0,20	0,741	0,903	1,296
-0,10	0,766	0,935	1,348
0,00	0,796	0,974	1,409
0,10	0,831	1,020	1,481
0,20	0,873	1,074	1,567
0,30	0,924	1,140	1,672
0,40	0,985	1,221	1,799
0,50	1,061	1,321	1,958

Fonte: Adaptado de Choulakian e Stephens (2001)

4 ESTUDO DE SIMULAÇÃO

Neste capítulo, serão apresentadas simulações realizadas para demonstração e ilustração das propriedades apresentadas nos dois capítulos anteriores sobre as famílias de distribuições de valor extremo generalizada e Pareto generalizada. Serão utilizados os estimadores de máxima verossimilhança, mas o mesmo estudo pode ser refeito utilizando os estimadores de Momentos Ponderados por Probabilidade. Para essas simulações foi utilizado o software R 3.1.2 (R Core Team, 2014).

4.1 TAMANHO DOS BLOCOS

Para avaliar a influência do tamanho dos blocos na estimação do índice de valor extremo para a distribuição GEV e o *trade-off* que essa escolha causa entre o vício do estimador e sua variância, foram simuladas 500 amostras de tamanho 4000 a partir de um distribuição normal padrão ($\mu = 0$ e $\sigma = 1$) e outras 500 amostras de tamanho 4000 a partir de uma distribuição exponencial com média 3. Então foram selecionados tamanho de blocos que fossem divisores de 4000 para não ocorrer o viés causado por bloco de tamanho menor, que seria o resto da divisão. Os gráficos na Figura 8 mostram estimativas do vício, da variância do estimador e do Erro Quadrático Médio para possíveis valores de tamanho de bloco tanto para os dados simulados a partir da distribuição normal quanto da distribuição exponencial.

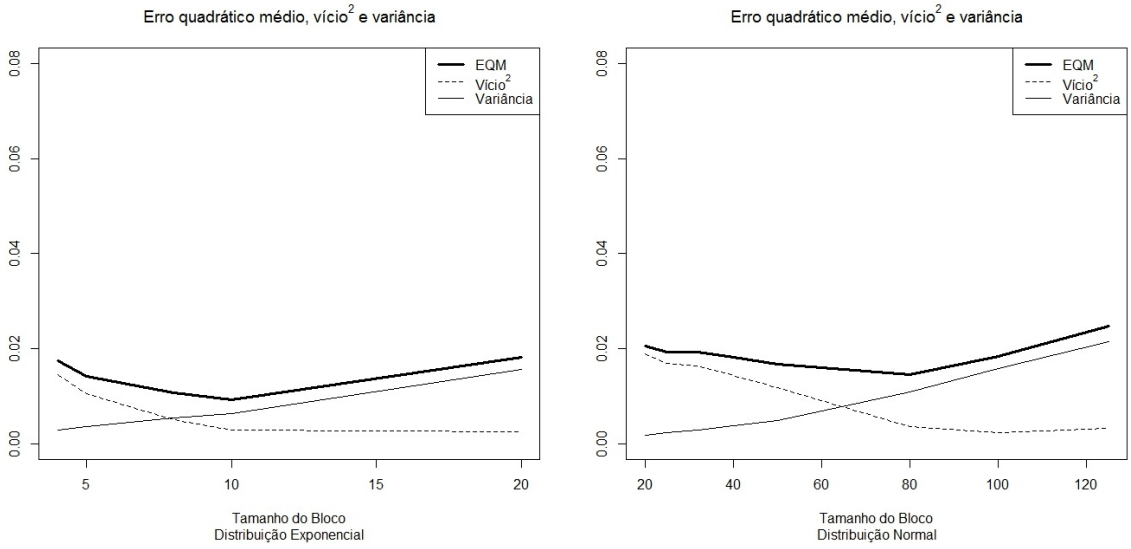

Figura 8 – Estimativas de EQM, vício^2 e variância para $\hat{\gamma}$

Neste estudo de simulação o EQM atinge o mínimo para blocos de tamanho 80, no caso dos dados possuírem distribuição normal, e 10 para distribuição exponencial, sendo que após estes pontos o vício decai mais lentamente. Isso é indício de que o máximo de

uma amostra normalmente distribuída converge mais lentamente para uma distribuição da família GEV que o máximo de uma amostra distribuída exponencialmente, reforçando o afirmado na Seção 2.2.1.

4.2 DOMÍNIO DE ATRAÇÃO E ÍNDICE DE VALOR EXTREMO

A partir da simulação de amostras de algumas das distribuições listadas nas Tabelas 2, 3 e 4 será possível observar o comportamento das distribuições em relação aos domínios de atração e o quanto seus parâmetros influenciam na estimativa do índice de valor extremo γ. Para isso, foram simuladas várias amostras de 4000 observações das diferentes distribuições e com diferentes parâmetros. De cada uma dessas amostras foram retirados os máximos de blocos de tamanho 80 e a eles ajustadas distribuições da família GEV, resultando em uma amostra efetiva de tamanho 50. Também foram obtidos os 50 maiores valores dessas amostras e ajustadas a elas uma distribuição da família GPD.

Tabela 7 – Estimativas para distribuições com domínio de atração tipo I ($\gamma = 0$)

Distribuição Exponencial(λ)

λ	GEV			GPD		
	$\hat{\gamma}$	$IC_{\alpha=0.05}(\hat{\gamma})$		$\hat{\gamma}$	$IC_{\alpha=0.05}(\hat{\gamma})$	
5	-0,0314	-0,2085	0,1457	-0,0351	-0,2977	0,2275
2	-0,0315	-0,2085	0,1455	-0,0351	-0,2977	0,2276
1	-0,0316	-0,2086	0,1455	-0,0349	-0,2977	0,2278
0,5	-0,0316	-0,2085	0,1454	-0,0349	-0,2977	0,2279
0,2	-0,0315	-0,2085	0,1456	-0,0350	-0,2977	0,2278

Distribuição Normal($0, \sigma^2$)

σ	GEV			GPD		
	$\hat{\gamma}$	$IC_{\alpha=0.05}(\hat{\gamma})$		$\hat{\gamma}$	$IC_{\alpha=0.05}(\hat{\gamma})$	
5	0,0277	-0,1633	0,2187	0,0656	-0,2167	0,3479
1	0,0275	-0,1633	0,2184	0,0658	-0,2166	0,3483
0,5	0,0275	-0,1633	0,2184	0,0656	-0,2167	0,3478
0,1	0,0276	-0,1633	0,2185	0,0661	-0,2165	0,3487
0,01	0,0276	-0,1617	0,2169	0,0656	-0,2012	0,3324

Como pode ser observado na Tabela 7, o Intervalo de Confiança de ambos os modelos sempre está contendo o verdadeiro valor de γ (0) e os parâmetros das distribuições simuladas têm pouca ou nenhuma influência na estimativa de γ.

Na Tabela 8 é perceptível que, para a distribuição T_ν, a medida que ν cresce o Intervalo de Confiança passa a incluir o 0, sugerindo que o máximo amostral pode estar no domínio de atração tipo I. Isso ocorre devido ao fato de que a distribuição T_ν, quando $\nu \to \infty$, converge para a distribuição Normal. Já para a distribuição $F(m, n)$, apesar de o

Tabela 8 – Estimativas para distribuições com domínio de atração tipo II ($\gamma > 0$)

Distribuição T_ν							
ν	$\gamma = \frac{1}{\nu}$	GEV			GPD		
		$\hat{\gamma}$	$IC_{\alpha=0.05}(\hat{\gamma})$		$\hat{\gamma}$	$IC_{\alpha=0.05}(\hat{\gamma})$	
2	0,5000	0,5828	0,2533	0,9123	0,5531	0,1420	0,9642
3	0,3333	0,3546	0,0795	0,6297	0,3786	-0,0503	0,8075
5	0,2000	0,1881	-0,0528	0,4291	0,1212	-0,1801	0,4225
7	0,1429	0,0092	-0,2065	0,2248	-0,0837	-0,3587	0,1913
10	0,1000	-0,1224	-0,3461	0,1013	-0,1454	-0,4537	0,1629
Distribuição $F(m, n)$							
m	n	$\gamma = \frac{2}{n}$	GEV			GPD	
			$\hat{\gamma}$	$IC_{\alpha=0.05}(\hat{\gamma})$		$\hat{\gamma}$	$IC_{\alpha=0.05}(\hat{\gamma})$	
5	2	1,0000	0,7923	0,4565	1,1282	0,8677	0,3414	1,3940
4	5	0,4000	0,6017	0,3176	0,8858	0,5318	0,1255	0,9382
5	5	0,4000	0,3615	0,0994	0,6236	0,3870	0,0098	0,7642
7	5	0,4000	0,5510	0,2068	0,8952	0,3149	-0,0618	0,6915
5	7	0,2857	0,1854	-0,0446	0,4155	0,1356	-0,1747	0,4458

valor verdadeiro de γ depender apenas do parâmetro n, o parâmetro m acaba influenciando na estimativa $\hat{\gamma}$ para ambos os modelos.

Tabela 9 – Estimativas para distribuições com domínio de atração tipo III ($\gamma < 0$)

Distribuição Beta(α, β)								
α	β	$\gamma = \frac{1}{n}$	GEV			GPD		
			$\hat{\gamma}$	$IC_{\alpha=0.05}(\hat{\gamma})$		$\hat{\gamma}$	$IC_{\alpha=0.05}(\hat{\gamma})$	
2	2	-0,5000	-0,5227	-0,5776	-0,4678	-0,4975	—	—
1	3	-0,3333	-0,3251	-0,3873	-0,2628	-0,3115	-0,4031	-0,2199
2	3	-0,3333	-0,3488	-0,4001	-0,2974	-0,3492	-0,4165	-0,2818
3	3	-0,3333	-0,3819	-0,4306	-0,3332	-0,3417	-0,4205	-0,2629
2	5	-0,2000	-0,2061	-0,2630	-0,1493	-0,2384	-0,3022	-0,1745
Distribuição Uniforme(a, b)								
a	b	$\gamma = -1$	GEV			GPD		
			$\hat{\gamma}$	$IC_{\alpha=0.05}(\hat{\gamma})$		$\hat{\gamma}$	$IC_{\alpha=0.05}(\hat{\gamma})$	
0	1	-1,0000	-0,9416	—	—	-1,0115	—	—
-3	3	-1,0000	-1,0003	—	—	-1,0193	—	—
-5	0	-1,0000	-0,9382	—	—	-1,0067	—	—

A partir dos resultados da Tabela 9 pode-se observar que o parâmetro α da distribuição $Beta(\alpha, \beta)$ influencia pouco na estimativa pontual $\hat{\gamma}$, já que γ depende apenas de β, mas tem uma maior influência na largura dos Intervalos de Confiança. Para o caso da Uniforme(a, b), é perceptível que a estimava de γ se aproxima do valor real -1, mas não é possível encontrar o intervalo de confiança de $\hat{\gamma}$ a partir do Estimador de Máxima Verossimilhança, pois ele possui as propriedades desejadas mencionadas na Seção 2.4.1

apenas para $\gamma > -1$ para o modelo GEV e para $\gamma > -1/2$ para o modelo GPD. Por esse motivo também não foi calculado intervalo de confiança para a distribuição simulada $Beta(2, 2)$ para as estimativas do modelo GPD, pois o valor verdadeiro de γ é -1/2.

4.3 CONVERGÊNCIA E QUALIDADE DO AJUSTE

Além disso, para verificar que os máximos das distribuições apresentadas converge para a família de distribuições GEV, foram simuladas 500 amostras de 4000 observações e ao máximo dos blocos de tamanho 80 de cada uma dessas amostras foi ajustado um modelo GEV com os parâmetros estimados através do método de máxima verossimilhança, gerando 500 amostras de tamanho efetivo $k = 50$. Para efeito de comparação foram simuladas o mesmo número de amostras a partir da própria distribuição de valor extremo e da distribuição de Poisson, cuja função distribuição não satisfaz as condições do Teorema 2.2.1, e por isso seu máximo não converge para a família de distribuições GEV. O ajuste foi avaliado pelo teste de Anderson-Darling modificado para distribuições de valores extremos com parâmetros desconhecidos (LAIO, 2004) com o nível de significância $\alpha = 0.05$. Utilizando as mesmas amostras também foram selecionados os excessos de um limiar u, utilizando u como o quantil 0.95 das amotras, gerando 500 amostras de excessos de limar de tamanho $N_u = 200$ para cada distribuição. A essas amostras foram ajustadas distribuições da família GPD, além de 500 amostras de tamanho 200 geradas a partir da própria distribuição GPD. Utilizando o teste de Anderson-Darling para distribuições da família Pareto generalizada (CHOULAKIAN; STEPHENS, 2001), foram computados os percentuais de rejeição do teste para cada distribuição utilizada. Os resultados dessa simulação encontram-se na Tabela 10.

Tabela 10 – Percentual de rejeições do teste de Anderson-Darling

Disitribuição	Percentual de Rejeição ($\alpha = 0.05$)	
	GEV	GPD
GEV(0,0,1)	3,00%	—
GPD(0,1)	—	2,60%
Normal(0,1)	4,20%	5,80%
Exponencial(3)	3,60%	6,00%
T_3	2,20%	2,60%
$F(5, 5)$	1,00%	1,00%
Beta(3,3)	6,00%	12,00%
Poisson(5)	99,40%	100,00%

Os resultados encontrados corroboram a afirmação de que os extremos de uma população que segue uma distribuição de Poisson não convergem pras distribuições da Teoria de Valor Extremo, dado o alto índice de rejeição do teste de Anderson-Darling para ambas distribuições. Além disto, não indica não-convergência das demais distribuições

utilizadas. A distribuição Uniforme não foi utilizada nesse estudo pois o estimador de máxima verossimilhança não é assintóticamente eficiente para $\gamma < -1/2$, condição necessária para o teste realizado.

5 APLICAÇÃO

A cidade Juiz de Fora tem a maior parte de sua área de ocupação localizada nos fundos do vale do Rio Paraibuna e dos vales secundários de seus afluentes e, por isto, sua população sofre frequentemente com alagamentos e inundações (SILVA; MACHADO, 2012). Além disto, devido à ocupação de encostas com declividade acentuada, é comum a ocorrência de deslizamentos de terra e soterramentos (VARGAS, 2011). A situaçã é particularmente agravada quando ocorre um grande volume de chuva em um curto espaço de tempo (SOUZA; SANTOS et al., 2012). Exemplos foram as fortes chuvas ocorridas em 9 de janeiro de 2012 e em 26 de dezembro de 2013, onde houve precipitação de 82,8 mm e 85,2 mm, respectivamente. Segundo a Defesa Civil, a primeira chuva foi responsável pelo soterramento de uma casa, causando uma morte e 123 pessoas ficaram desalojadas; já a segunda chuva causou o desabamento de um prédio e duas casas, causando uma morte e vários pontos de alagamento (Figura 9).

Fonte: Jornal Tribuna de Minas

Figura 9 – Estragos causados pelas chuvas dos dias 26 de dezembro de 2013 (esquerda) e 9 de janeiro de 2012 (direita)

Devido a estes fatos, o conhecimento sobre os extremos da precipitação pluvial em Juiz de Fora pode ajudar o planejamento urbano e a Defesa Civil a se prevenirem contra os possíveis desastres. Neste capítulo serão utilizadas as distribuições apresentadas nos Capítulos 2 e 3 para modelagem dos valores extremos da precipitação diária.

Os dados utilizados nesse estudo sobre a precipitação diária (em mm) de Juiz de Fora foram cedidos pelo Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) do Instituto Nacional de Meteorologia. O banco de dados contém a precipitação diária medida de 01 de janeiro de 1961 à 31 de dezembro de 2014. Os anos de 1972, 1979, 1986, 1988, 1990 e 1992 não foram utilizados neste estudo por conterem um número muito alto de observações faltantes, sendo assim, obteve-se uma amostra de 16921 precipitações diárias, cuja série histórica é apresentada na Figura 10.

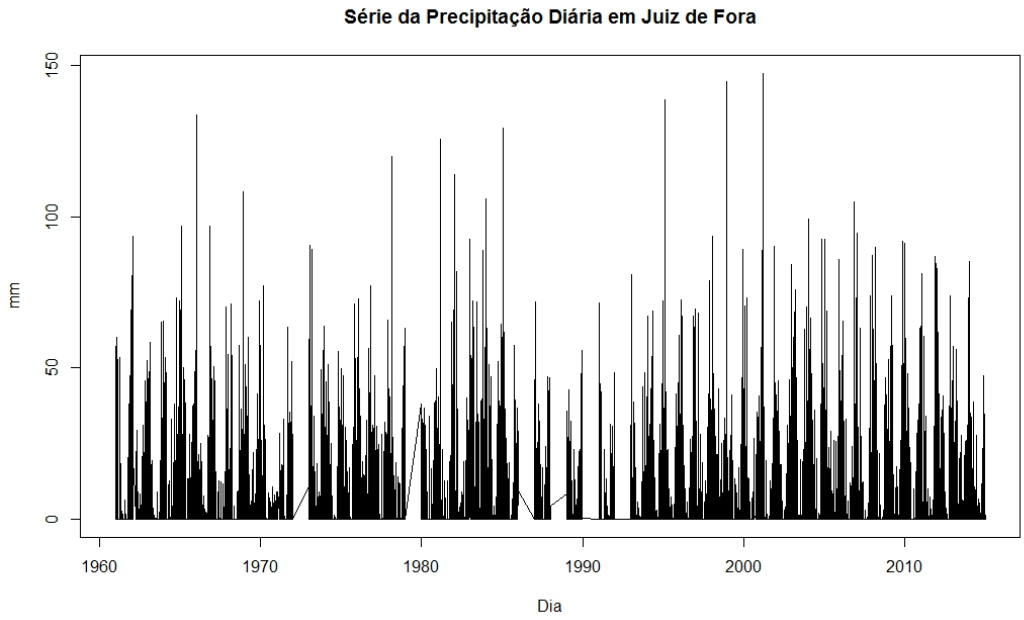

Figura 10 – Série da precipitação diária em Juiz de Fora de 01/01/1961 à 31/12/2014

A distribuição da precipitação diária fica concentrada perto de 0mm, mas também há um número substancial de observações com valores de precipitação elevado na cauda direita, como pode ser observado na Figura 11. O objetivo do estudo será justamente modelar o comportamento da cauda direita da precipitação diária, podendo assim conhecer melhor a distribuição de seus valores extremos.

5.0.1 Modelagem dos máximos anuais

Para a modelagem pela distribuição de valores extremos generalizada, foram selecionados os máximos anuais da precipitação diária de 1961 à 2014 - excluídos os 6 anos em que os dados não estão disponíveis -, resultando em uma amostra efetiva de k = 48 máximos anuais. A média dos máximos anuais m é 87,83 mm e o desvio padrão 24,47 mm. O gráfico da função de autocorrelação para os máximos anuais (Figura 12) parece não indicar que há estrutra de autocorrelação nas observações.

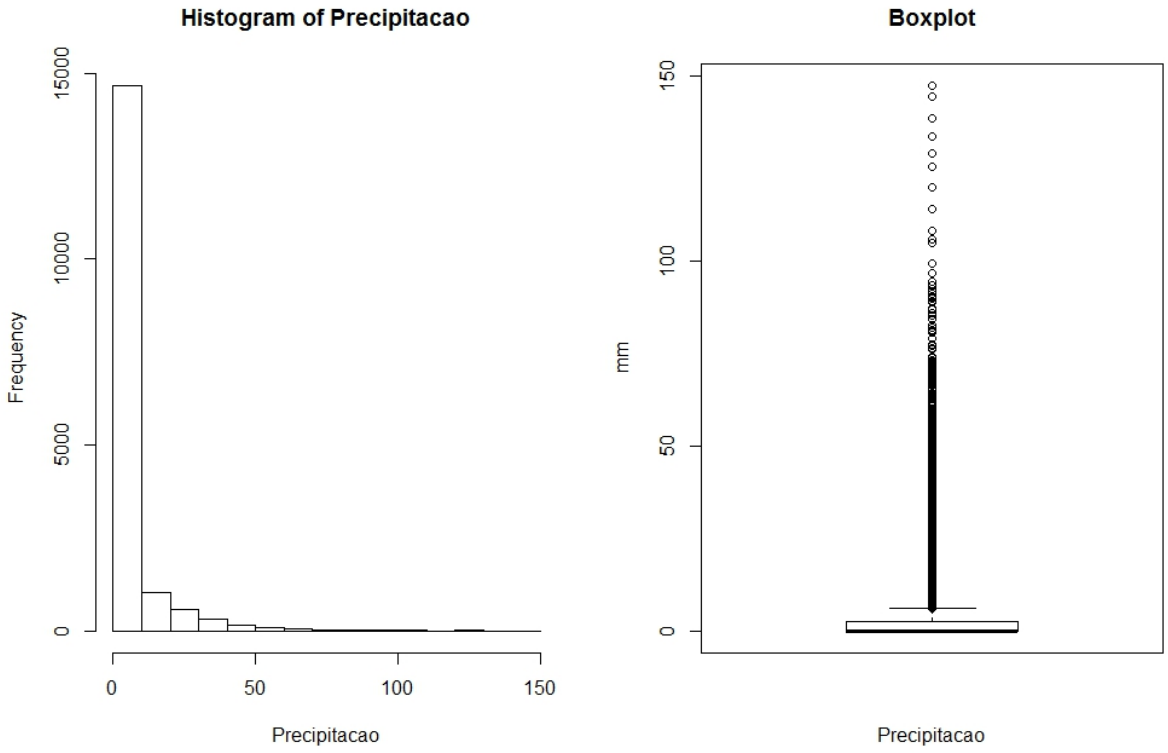

Figura 11 – Histograma e Boxplot da precipitação diária em Juiz de Fora

Para a amostra $m_1, ..., m_{48}$ dos máximos anuais da precipitação diária, ajustou-se um modelo GEV utilizando o método de máxima verossimilhança para estimação dos parâmetros, como descrito na Seção 2.4. Os parâmetros estimados com seus respectivos intervalos de confiança, com nível de significância $\alpha = 0,05$, estão apresentados na Tabela 11

Tabela 11 – Estimativas de Máxima Verossimilhança dos parâmetros de um modelo GEV para os máximos anuais da precipitação diária em Juiz de Fora

Parâmetro	Estimativa	$IC_{\alpha=0.05}$	
γ	-0,0088	-0,2331	0,2156
μ	76,8169	70,6560	82,9778
σ	19,2028	14,7244	23,6812

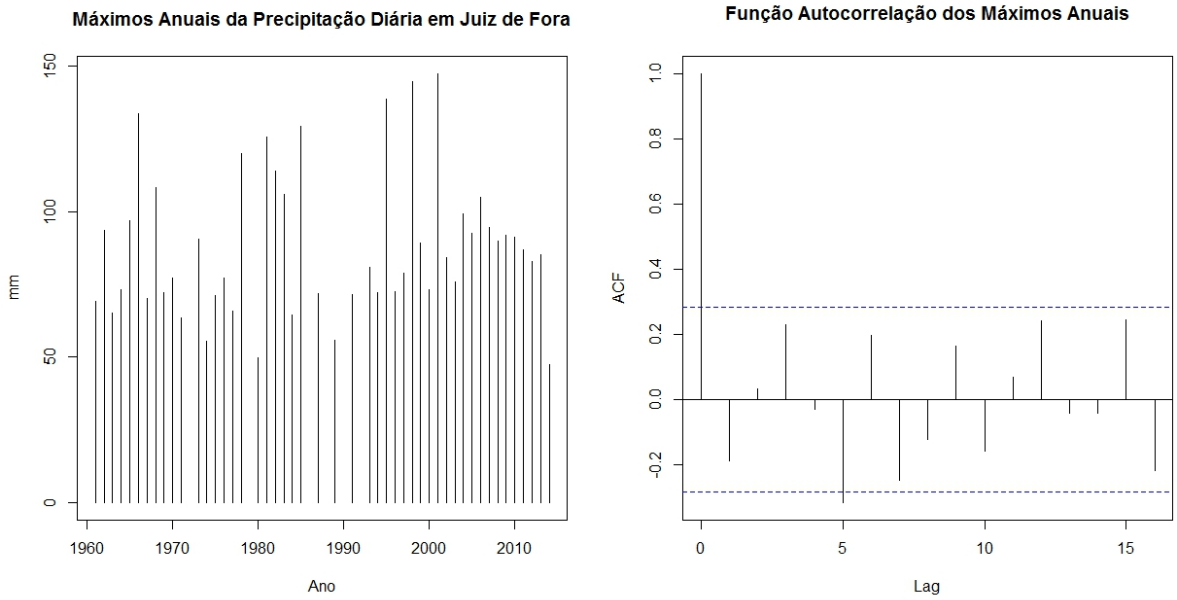

Figura 12 – Máximos anuais da precipitação diária em Juiz de Fora (esquerda) e Função de Autocorrelação estimada para os máximos anuais da precipitação diária em Juiz de Fora (direita).

A Figura 13 traz os gráficos de qualidade de ajuste explicados na Seção 2.5, que dão indícios de que a distribuição GEV com os parâmetros estimados se ajusta bem aos máximos anuais da precipitação diária. Além disto, foi realizado o teste de Anderson-Darling para distribuições de valor extremo com parâmetros desconhecidos. A estatística de teste ω obtida foi 0,1342 e seu p-valor correspondente é 0,4416. Logo, a hipótese nula de que a amostra de máximos anuais da precipitação diária segue uma distribuição GEV não é rejeitada ao nível de significância $\alpha = 0.05$.

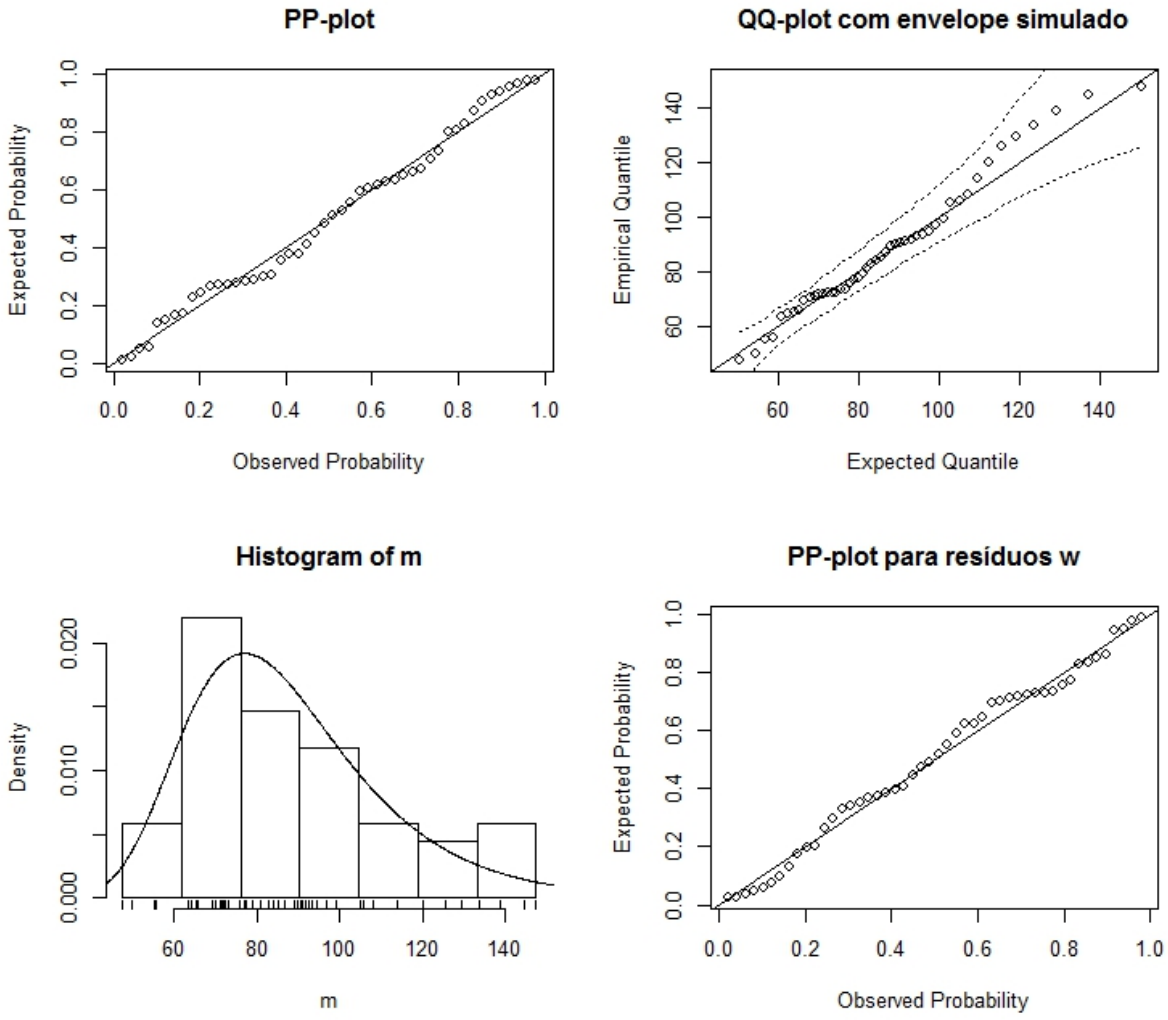

Figura 13 – Gráficos para análise da qualidade do ajuste dos máximos anuais a uma distribuição GEV com parâmetros estimados $\hat{\gamma} = -0,0088, \hat{\mu} = 76,8169,$ e $\hat{\sigma} = 19,2028$

A Figura 14 mostra o gráfico do nível de retorno para a distribuição GEV estimada. Os pontos são os máximos anuais da amostra; a linha central representa o nível de retorno; e as linhas em seu entorno, o intervalo de confiança para suas estimativas. Como a estimativa para o parâmetro γ é próxima a 0, a linha central se assemelha a uma reta. Através deste gráfico pode-se interpretar que se espera que ocorra uma vez a cada dez anos um dia em que a precipitação seja de 120 mm, com intervalo de confiança de 106,75 mm à 134,40 mm, ao nível de 5% de significância. Já um dia em que a precipitação seja de 150 mm (IC[1]: 119,86 mm - 181,41 mm) é esperado uma vez a cada 50 anos.

[1] Intervalo de Confiança para nível de significância $\alpha = 0.05$

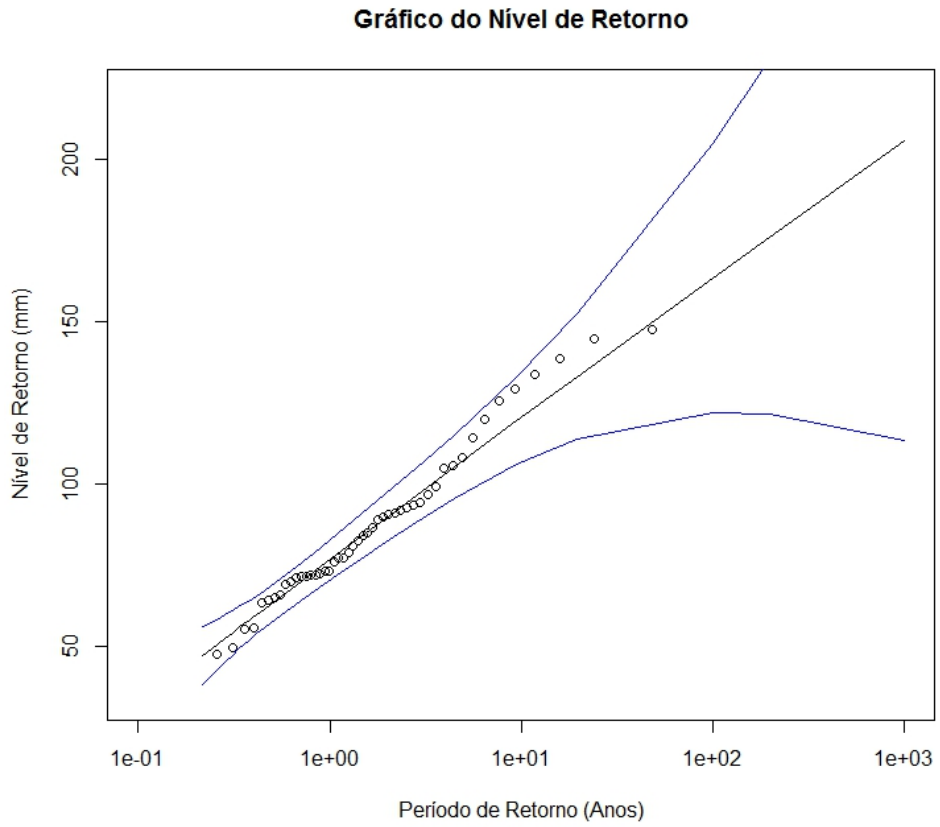

Figura 14 – Gráfico do Nível de Retorno anual do máximo da precipitação diária em Juiz de Fora a partir da distribuição GEV estimada

5.0.2 Modelagem dos excessos de um limiar

Para a modelagem através da família de distribuições Pareto generalizada, o primeiro passo é escolher o limiar u, tal que serão selecionadas apenas as amostras cujo valor ultrapassá-lo. Para essa seleção foram utilizados os gráficos da média e mediana empíricas dos excessos (Figura 15) e as estimativas de máxima verossimilhança dos parâmetros $\tilde{\sigma}$ e γ (Figura 16) para cada valor de u selecionado. Analisando esses gráficos, o objetivo é encontrar o menor valor de u para os quais eles sejam constantes (ou lineares no caso do gráfico das estimativas de $\tilde{\sigma}$).

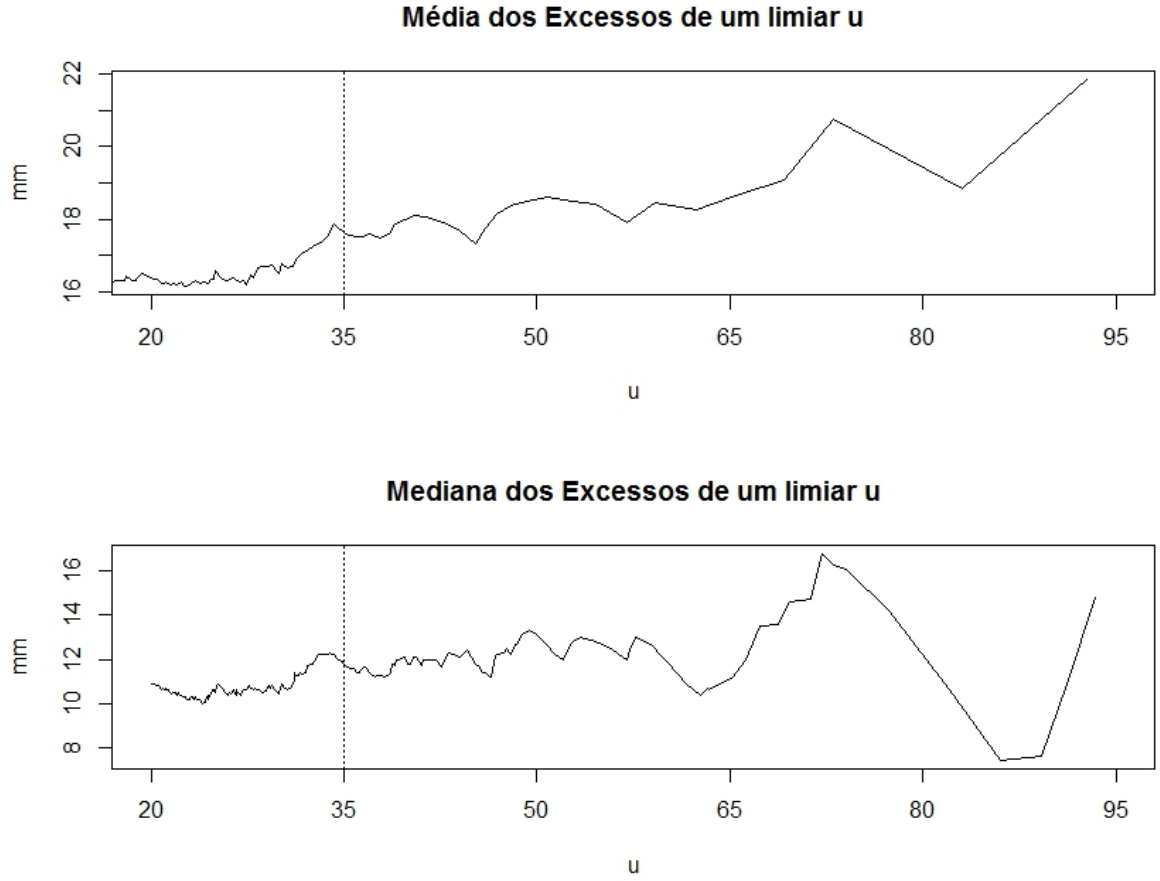

Figura 15 – Gráficos da Média Empírica (*superior*) e Mediana Empírica (*inferior*) dos excessos de um limiar u da precipitação diária em Juiz de Fora

O valor escolhido para u a partir da análise dos gráficos foi 35, correspondente ao quantil 97,2% das precipitações diárias em Juiz de Fora. A amostra efetiva de excessos de limiar obtida foi $N_u = 475$, os quais são apresentados na Figura 17. Na mesma figura se encontra o gráfico da função de autocorrelação dos excessos de limiar, que não apresenta evidências de autocorrelação dos excessos, não havendo então necessidade de desclusterização.

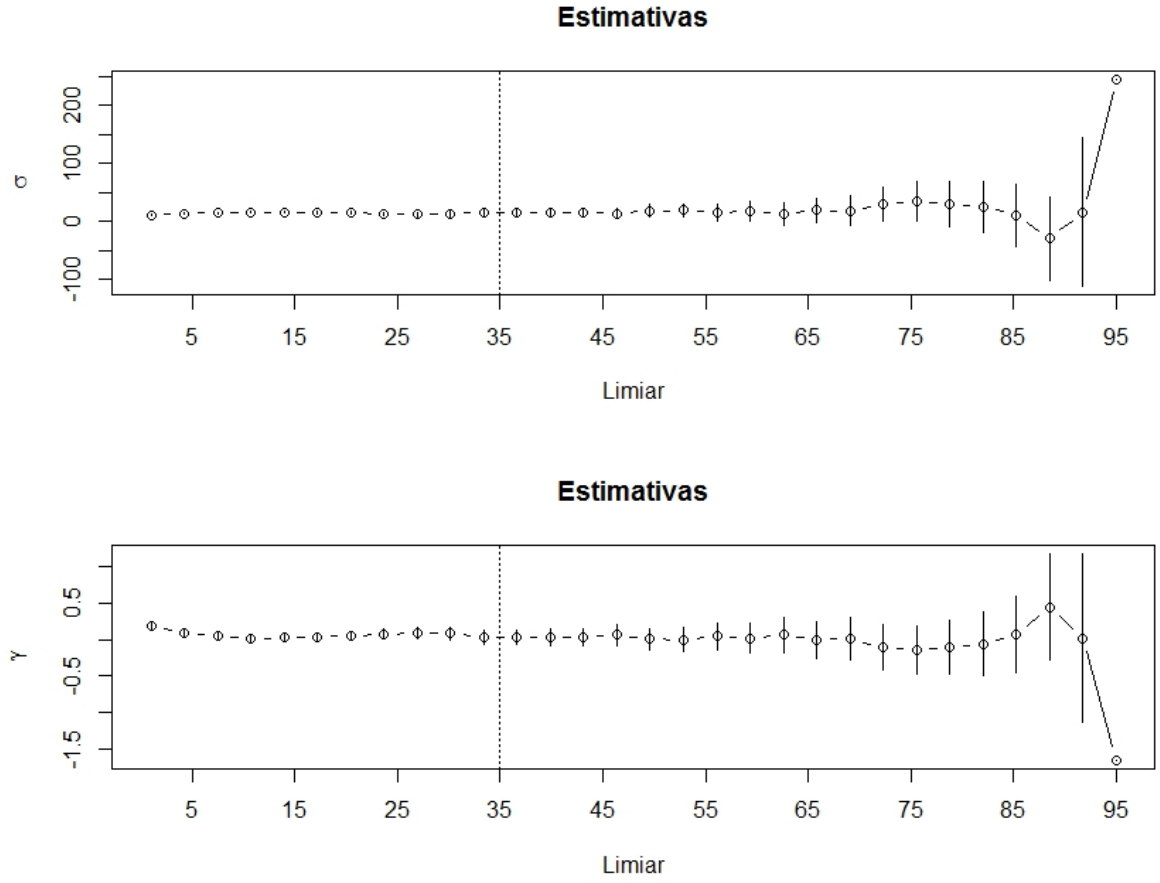

Figura 16 – Gráfico das estimativas dos parâmetros $\tilde{\sigma}$ (*superior*) e γ (*inferior*) de um modelo GPD para diferentes valores do limiar u

À essa amostra de 475 excessos do limiar $u = 35$mm da precipitação diária de Juiz de Fora foram estimados os parâmetros de um modelo GPD, pelo método de máxima verossimilhança. A Tabela 12 apresenta as estimativas desse parâmetros com os respectivos intervalos de confiança para um nível de significância $\alpha = 0,05$.

Tabela 12 – Estimativas de Máxima Verossimilhança dos parâmetros de um modelo GPD para os Excessos do limiar $u = 35$ da precipitação diária em Juiz de Fora

Parâmetro	Estimativa	$IC_{\alpha=0.05}$	
γ	0,0319	-0,0634	0,1271
$\tilde{\sigma}$	17,1112	14,8700	19,3524

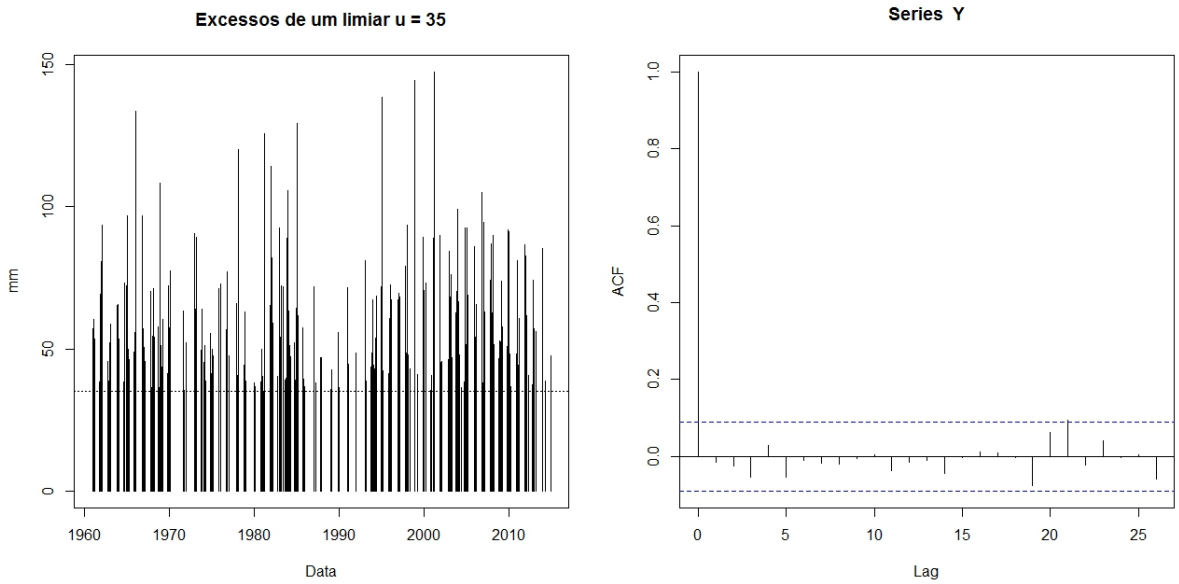

Figura 17 – Excessos do limiar $u = 35$ da precipitação diária em Juiz de Fora (esquerda) e Função de Autocorrelação estimada para os excessos do limiar $u = 35$ da precipitação diária em Juiz de Fora (direita)

A Figura 18 apresenta os gráficos para verificar a qualidade do ajuste de um modelo GPD, discutidos na Seção 3.5, que indicam que a distribuição Pareto generalizada estimada se ajusta bem à amostra de excessos do limiar $u = 35$ da precipitação diária. Além disso, para verificar o ajuste, foi realizado o teste de Anderson-Darling para distribuições da família GPD com parâmetros desconhecidos (CHOULAKIAN; STEPHENS, 2001). A estatística de teste A^2 encontrada foi 0,3282 e o valor crítico ao nível de 5% de significância é $z = 0,9887$, ou seja, a hipótese nula de que os excessos de limiar seguem uma distribuição Pareto generalizada não é rejeitada.

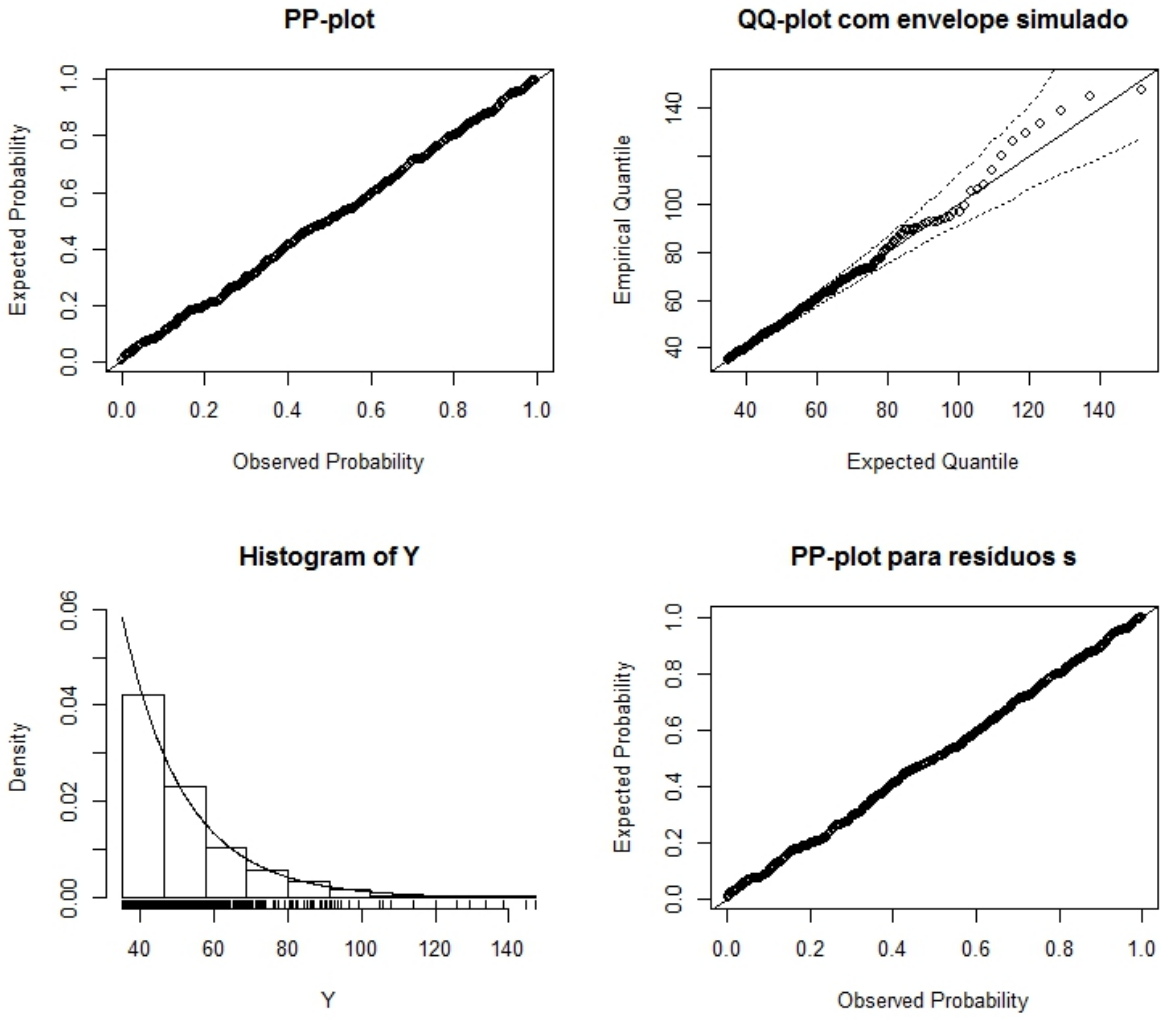

Figura 18 – Gráficos para análise da qualidade do ajuste dos excessos do limiar $u = 35$ à uma distribuição GPD com parâmetros estimados $\hat{\gamma} = 0.0319$ e $\hat{\sigma} = 17,1112$

O gráfico do nível de retorno, utilizado para interpretar os resultados da modelagem dos excessos de limiar da precipitação diária em Juiz de Fora através da distribuição Pareto generalizada, é apresentado na Figura 19. A interpretação desse gráfico é feita da mesma forma que para a distribuição GEV, ou seja, espera-se que ocorra uma vez a cada dez anos um dia em que a precipitação seja de 120,36 mm, com intervalo de confiança de 106,35 mm à 134,28 mm. Já um dia no qual a precipitação seja de 153 mm (IC[2]: 125,42 mm - 180,68 mm) é esperado uma vez a cada 50 anos.

[2] Intervalo de Confiança para nível de significância $\alpha = 0.05$

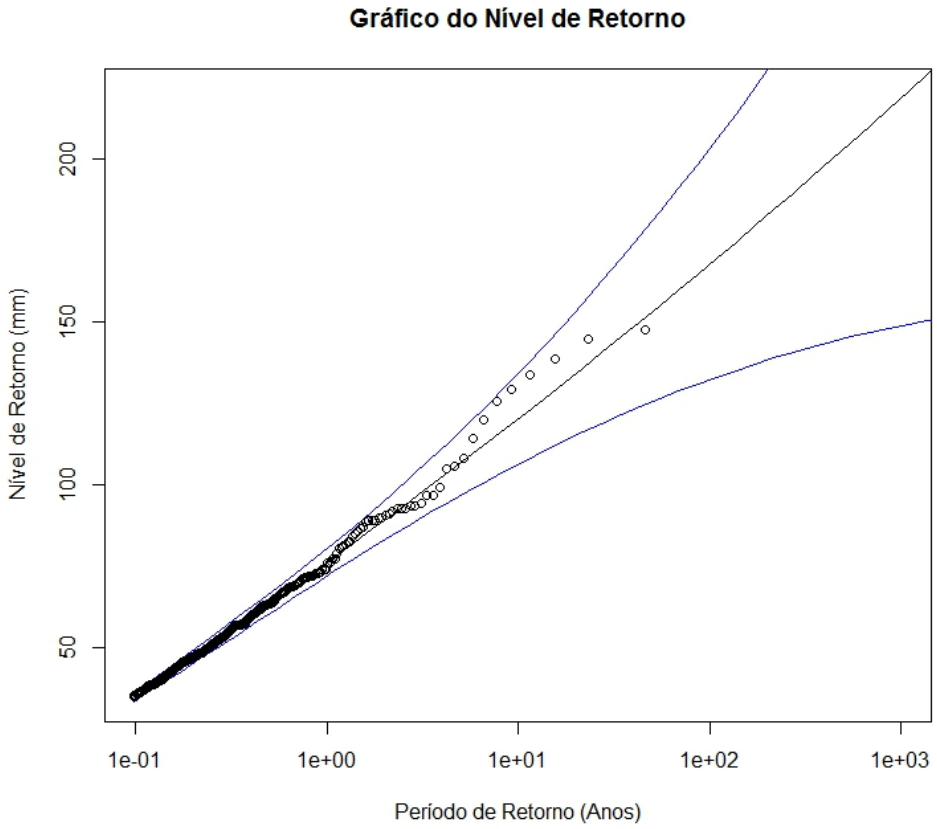

Figura 19 – Gráfico do Nível de Retorno anual da precipitação diária em Juiz de Fora a partir da distribuição GPD estimada

6 CONCLUSÕES

Nesta monografia foram discutidas metodologias para modelagem de eventos extremos, a modelagem de máximos parciais, através da distribuição de valor extremo generalizada, e a modelagem de excessos de um limiar, através da distribuição Pareto generalizada. O estudo de simulação realizado Capítulo 4 foi importante para o conhecimento prático de algumas propriedades dos modelos utilizados nesse estudo. Foi ilustrado o efeito de tamanho de blocos para a convergência para a distribuição GEV. Também foi possível verificar a influência dos parâmetros da distribuição populacional na estimação do índice de valor extremo γ para ambas distribuições. E ainda verificou-se a acurácia dos testes de Anderson-Darling modificados para o ajuste às distribuições, mostrando também a convergência ou não dos extremos de certas distribuições populacionais.

No estudo realizado para a precipitação pluvial em Juiz de Fora foi encontrado que dias com quantidade de chuva maior ou igual aos que causaram grandes estragos em 9 de janeiro de 2012 e 16 de dezembro de 2013 são esperados ocorrerem uma vez a cada 1,54 (IC[1]: 1,10 - 2,40) anos e 1,36 (IC: 0,99 1,99) anos, utilizando o modelo GEV, e 1.48 (IC: 1,26 - 2,24) anos e 1,42 (IC: 1,17 - 2.11), utilizando o modelo GPD. Estes resultados mostram que é alta a probabilidade de ocorrência de chuvas potencialmente danosas e, portanto, deve ser uma preocupação do poder público e da população planejar para a prevenção e minimização do impacto causado.

Como sugestão para estudos futuros, poderia ser considerada a modelagem conjunta da precipitação pluvial e do nível do rio Paraibuna, utilizando, por exemplo, distribuições bivariadas (distribuições de valor extremo generalizada e Pareto generalizada) ou cópulas (Gumbel e Clayton). Esta modelagem auxiliaria na avaliação, simultânea, do risco de enchentes e inundações nas áreas à margem do rio. Além disto, seria interessante haver um acompanhamento anual dos parâmetros estimados nesse estudo, para detecção de possíveis alterações no comportamento dos extremos da precipitação pluvial em Juiz de Fora devido às mudanças climáticas.

[1] Intervalo de Confiança para nível de significância $\alpha = 0.05$

REFERÊNCIAS

CHARRAS-GARRIDO, M.; LEZAUD, P. Extreme value analysis: an introduction. *Journal de la Societe Francaise de Statistique*, v. 154, n. 2, p. 66–97, 2013.

CHEN, J.-B.; LI, J. The extreme value distribution and dynamic reliability analysis of nonlinear structures with uncertain parameters. *Structural Safety*, Elsevier, v. 29, n. 2, p. 77–93, 2007.

CHOULAKIAN, V.; STEPHENS, M. A. Goodness-of-fit tests for the generalized pareto distribution. *Technometrics*, v. 43, n. 4, p. 478–484, 2001.

COLES, S. *An Introduction to Statistical Modeling of Extreme Values*. London: Springer, 2001.

DAVISON, A. C.; SMITH, R. L. Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society*, B 52, p. 393–442, 1990.

DIEBOLT, J. et al. Quasi-conjugate bayes estimates for gpd parameters and application to heavy tails modelling. *Extremes*, v. 8, n. 12, p. 57–78, 2005.

DIEBOLT, J.; GUILLOU, A.; RACHED, I. Approximation of the distribution of excesses through a generalized probability-weighted moments method. *Journal of Statistical Planning and Inference*, v. 137, n. 3, p. 841–857, 2007.

FISHER, R. A.; TIPPETT, L. H. C. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, v. 24, n. 2, p. 180–190, 1928.

FRECHÉT, M. Sur la loi de probabilité de lécart maximum. *Annales de la Société Polonaise de Mathématique*, v. 6, p. 93–116, 1927.

GILLI, M. et al. An application of extreme value theory for measuring financial risk. *Computational Economics*, Springer, v. 27, n. 2-3, p. 207–228, 2006.

GNEDENKO, B. V. Sur la distribution limite du terme maximum d'une serie aleatoire. *Ann. Math.*, v. 44, p. 423–453, 1943.

GREENWOOD, J. et al. Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. *Water Ressources Research*, v. 15, n. 5, p. 1049–1054, 1979.

HAAN, L. D.; FERREIRA, A. *Extreme Value Theory: An Introduction*. Boston: Springer, 2006.

HILL, B. A simple general approach to inference about the tail of a distribution. *The annals of statistics*, v. 3, n. 5, p. 1163–1174, 1975.

HOSKING, J.; WALLIS, J. Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, v. 29, n. 3, p. 339–349, 1987.

HOSKING, J. R. M.; WALLIS, J. R. Analysis and estimation of distributions using extreme order statistics. *Journal of the Royal Statistic al Society*, v. 52, p. 105–124, 1990.

HOSKING, J. R. M.; WALLIS, J. R. *Regional Frequency Analysis*. [S.l.]: Cambridge University Press, 1997.

JANSEN, D.; VRIES, C. de. On the frequency of large stock returns: Putting booms and busts into perspective. *Review of Economics and Statistics*, v. 73, p. 18–24, 1991.

KATZ, R. Extreme value theory for precipitation: sensitivity analysis for climate change. *Advances in Water Resources*, v. 23, n. 2, p. 133 – 139, 1999.

LAIO, F. Cramer von mises and anderson-darling goodness of fit tests for extreme value distributions with. *Water Resources Research*, v. 40, 2004.

LEADBETTER, M.; LINDGREN, G.; ROOTZEN, H. *Extremes and Related Properties of Random Sequences and Processe*. Berlim: Springer, 1983.

LEADBETTER, M. R. Extremes and local dependence in stationary sequences. *Zeitschrift fur Wahrscheinlichkeitstheorie und Werwandtl Gebiete*, v. 65, p. 291–306, 1983.

MISES, R. von. La distribution de la plus grande de n valeurs. *Rev. Math. Union Interbalcanique*, v. 1, p. 141–160, 1936.

MORRISON, J. E.; SMITH, J. A. Stochastic modeling of flood peaks using the generalized extreme value distribution. *Water Resources Research*, Wiley Online Library, v. 38, n. 12, p. 41–1, 2002.

PFEIFER, D. Extreme value theory in actuarial consulting: windstorm losses in central europe. *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields. Basel: Birkhäuser Verlag*, p. 373–378, 2001.

PICKANDS, J. Statistical inference using extreme order statistics. *Annals of Statistics*, v. 3, n. 1, p. 119–131, 1975.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2014. Disponível em: <http://www.R-project.org/>.

REISS, R.-D.; THOMAS, M. *Statistical analysis of extreme values: with applications to insurance, finance, hydrology and other fields*. [S.l.]: Birkhauser, 2007.

SILVA, R. S.; MACHADO, P. J. O. Inundações urbanas: o caso da micro-bacia hidrogáfica do córrego ipiranga–juiz de fora/mg. *Periódico Eletrônico Fórum Ambiental da Alta Paulista*, v. 7, n. 2, 2012.

SMITH, R. Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, v. 72, n. 1, p. 67–90, 1985.

SMITH, R. L. Estimating tails of probability distributions. *Annals of Statistics*, v. 15, p. 1174–1207, 1987.

SOUZA, L.; SANTOS, C. B. d. et al. O crescimento urbano e a ocupação de áreas de sob riscos de escorregamentos na região noroeste da área urbana de juiz de fora-mg. *Revista Interface (Porto Nacional)*, n. 02, 2012.

TRYON, R. G.; CRUSE, T. A. Probabilistic mesomechanics for high cycle fatigue life prediction. *Journal of engineering materials and technology*, American Society of Mechanical Engineers, v. 122, n. 2, p. 209–214, 2000.

VANDEWALLE, B.; BEIRLANT, J. On univariate extreme value statistics and the estimation of reinsurance premiums. *Insurance: Mathematics and Economics*, Elsevier, v. 38, n. 3, p. 441–459, 2006.

VARGAS, M. A. R. Construção social da moradia de risco: a experiência de juiz de fora (mg). *Revista Brasileira de Estudos Urbanos e Regionais*, v. 8, n. 1, p. 59–78, 2011.

ZHOU, C. Existence and consistency of the maximum likelihood estimator for the extreme value index. *Journal of Multivariate Analysis*, v. 100, n. 4, p. 794–815, 2009.