UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

SÔNIA ELIANE GONÇALVES DOS SANTOS

TEORIA DE VALORES EXTREMOS APLICADOS A DADOS DE TEMPERATURA MÁXIMA EM CAMPINA GRANDE

CAMPINA GRANDE - PB

DEZEMBRO 2017

SÔNIA ELIANE GONÇALVES DOS SANTOS

TEORIA DE VALORES EXTREMOS APLICADOS A DADOS DE TEMPERATURA MÁXIMA EM CAMPINA GRANDE

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. Ricardo Alves Olinda

CAMPINA GRANDE - PB

DEZEMBRO 2017

S237t Santos, Sonia Eliane Goncalves dos.
Teoria de valores extremos aplicados a dados de temperatura máxima em Campina Grande [manuscrito] : / Sonia Eliane Goncalves dos Santos. - 2017.
30 p.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2017.
"Orientação : Prof. Dr. Ricardo Alves de Olinda, Coordenação do Curso de Estatística - CCT."

1. Distribuição assintótica. 2. Valores extremos. 3. Probabilidade. 4. Climatologia.

21. ed. CDD 519.232

SÔNIA ELIANE GONÇALVES DOS SANTOS

TEORIA DE VALORES EXTREMOS APLICADOS A DADOS DE TEMPERATURA MÁXIMA EM CAMPINA GRANDE

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Trabalho aprovado em 11 de dezembro de 2017.

BANCA EXAMINADORA

Ricardo Alves de Olinda

Dr.Ricardo Alves de Olinda
Universidade Estadual da Paraíba

Tiago Almeida de Oliveira

Dr.Tiago Almeida de Oliveira
Universidade Estadual da Paraíba

Gustavo

Dr. Gustavo Henrique Esteves
Universidade Estadual da Paraíba

Agradecimentos

A Deus primeiramente, pois sem a vontade e misericórdia dele não estaria aqui presente.

Ao meu amado pai Edimundo Evaristo dos Santos que sempre esteve ao meu lado nas horas mais difíceis dessa vida, aos meus amados filhos Malili Malheiros Gomes, Allana Beatriz Gonçalves dos Santos Flor e Allan Marcos Gonçalves dos Santos Silva, meu irmão Elenilson Gonçalves dos Santos e minha mãe Severina Gonçalves dos Santos. Ao Professor Ricardo Alves de Olinda, pela orientação e incentivo, colaborando de maneira excelente para minha vida acadêmica.

Aos docentes do Curso de Bacharelado em Estatística da Universidade Estadual da Paraíba, que auxiliaram na minha formação acadêmica citarei apenas alguns Ana Patrícia B. Peixoto, Tiago Almeida, Gustavo Esteves, Diana Maia, Juarez Oliveira, Gil de Luna, Giselly Oliveira, Tanyse Kely, Maria Joseane Cruz, Nathielly Lima, Vitória Serafim, Kleber Barros, Sílvio Fernandes e outros professores que não são do curso porém dão aula no mesmo Edson Vasconcelos, Vandik, José Elias, Dalva Lobão, Maurício Tavares, Ana Patricia Sampaio, enfim vários outros que não estão aqui citados, porém todos contribuíram muito em minha formação acadêmica.

Aos meus amigos e colegas de curso que são muitos e cito alguns e peço perdão para não cometer injustiça com nenhum deles Márcia de Lourdes, Cláudio Cruz, Wanessa Isthéwany amiga não só de curso mas também companheira de trabalho, Aline Porto, Analu Cabral, Diego Alves, Damião Flávio, Arnete Campos, Klaini Clemente, Vitória, Fátima Pereira, Angela Ismenia, Filipe Sousa, Abraão, Leomir Sousa, Rayane Santos, Carlos Camilo, José Ednaldo, Shirley Oliveira, Deyse Pereira, Itamara Júnior, Nayara Lima e há alguns que desistiram no meio do caminho porém a amizade permanece Waleska S.Tavares, Eduarda Monteiro, Cassandra Gonzaga, Rodolfo Rodrigo, Liedson, Daliane Oliveira, Cristiano Valério, Valfredo Mendes, Regina Coeli, Jonatha Lopes, Patrícia Xavier e assim aos demais que não citei mais que permanece em meu coração.

Aos amigos de infância que sempre acreditaram na minha capacidade e sempre me deram força para continuar essa jornada Rosângela Silva, Fabiana Santos, Isabele Melo, Lindalva, Socorro Silva, Sirleide Silva, Silene Silva, Erivaldo Silva, Estefânia Carla, Tereza, Neidjane, Nadja, aos amigos de agora Eduardo Lisboa, Jackeline Costa, Renally Lisboa, Charles Silveira, Keila Almeida, Adenise Duarte aos meus tios Maria das Neves e José Nilton de Araújo, ao meu avô seu Manoel Evaristo, enfim a todos que torceram para meu exito pessoal e profissional.

Deus é bom o tempo todo, o tempo todo Deus é bom!!!

Resumo

Pesquisas revelam que as temperaturas no agreste e no semiárido paraibano estão ficando mais altas com o passar do tempo e as chuvas têm ocorrido com menor intensidade. Neste cenário, a escassez de água constitui um forte entrave ao desenvolvimento socioeconômico e, até mesmo, à "subsistência" da população. Sendo assim, será de suma importância ajustar aos dados, uma distribuição de probabilidade capaz de representar e extrapolar medições para ocorrências futuras de um fenômeno relacionado a uma série histórica, como por exemplo, temperatura máxima mensal de uma região. Uma metodologia capaz de modelar estes eventos é a Teoria de Valores Extremos (TVE). Diante do exposto este trabalho tem por objetivo ajustar a distribuição Generalizada de Valores Extremos (GEV), que inclui como casos particulares, as distribuições Gumbel, Fréchet e Weibull, aos dados de temperatura máxima mensal do município de Campina Grande; implementar o método de máxima verossimilhança para obter as estimativas dos parâmetros da GEV, seguido do teste de Kolmogorov-Sminorv; gráficos de probabilidade-probabilidade e o quantil-quantil, aplicados para verificar o ajuste do modelo aos dados. Verificou-se que as distribuições Gumbel e Weibull são adequadas para representar os dados de temperatura máxima dos meses em estudo.

Palavras-chave: Distribuição assintótica. Valores extremos. Probabilidade. Climatologia

Abstract

Research has shown that temperatures in the agreste and semi-arid Paraíba are getting higher over time and rains have occurred less intensely. In this scenario, water scarcity constitutes a serious obstacle to socioeconomic development and even to the subsistence of the population. Thus, it will be extremely important to adjust to the data a probability distribution capable of representing and extrapolating measurements for future occurrences of a phenomenon related to a historical series, such as the maximum monthly temperature of a region. One methodology capable of modeling these events is the Extreme Values Theory (TVE). In view of the above, this work aims to adjust the Generalized Distribution of Extreme Values (GEV), which includes, as particular cases, the Gumbel, Fréchet and Weibull distributions, to the data of maximum monthly temperature of the municipality of Campina Grande; to implement the maximum likelihood method to obtain the estimates of the GEV parameters, followed by the Kolmogorov-Sminorv test; probability-to-quantile and quantile-quantile charts applied to verify model fit to data. It was found that the Gumbel and Weibull distributions are adequate to represent the maximum temperature data of the months under study.

Key-words: Asymptotic distribution. Extreme values. Probability. Climatologia

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Pesquisas revelam que as temperaturas no agreste e no semiárido paraibano estão ficando mais altas com o passar do tempo e as chuvas têm ocorrido com menor intensidade. Com o aumento das temperaturas, a tendência é que a chuva se torne cada vez mais escassa, consequência do aquecimento global associado ao aquecimento local, que é provocado, principalmente, pelo desmatamento e o processo de urbanização das cidades. Segundo o Instituto Nacional de Pesquisas Espaciais - *INPE*[1], enquanto em alguns lugares do semiárido nordestino a temperatura máxima diária aumentou até 3ºC nos últimos 40 anos, a média do aumento da temperatura no mundo no mesmo período foi de 0,4ºC. Para se ter ideia, da última era glacial para os dias atuais, a temperatura aumentou em torno de 3,5ºC, em um período de 15 mil anos.

Segundo Beltrão (2005), o polígono das secas apresenta um regime pluviométrico marcado por extrema irregularidade de chuvas, no tempo e no espaço. Neste cenário, a escassez de água constitui um forte entrave ao desenvolvimento socioeconômico e, até mesmo, à "subsistência" da população. A ocorrência cíclica das secas e seus efeitos catastróficos são por demais conhecidos e remontam aos primórdios da história do Brasil. Por meio de uma série meteorológica, pode-se determinar uma distribuição de probabilidade capaz de representar e extrapolar medições para ocorrências futuras de um fenômeno relacionado com esta série, como por exemplo, temperatura máxima mensal de uma região.

Uma das perguntas mais importantes relacionadas aos eventos extremos é se as suas ocorrências estão aumentando ou diminuindo com o tempo (FARIA et al., 2010). Uma maneira de modelar estes eventos é por meio da teoria de valores extremos (TVE). A TVE originou-se da necessidade dos astrônomos de utilizar ou rejeitar observações discrepantes; os primeiros artigos sobre o assunto são atribuídos a Fuller em 1914 e a Griffith em 1920, nos quais discutiram as utilidades dessa teoria, tanto no campo das aplicações como no dos métodos de análise matemática.

Diante do exposto este trabalho teve por objetivo ajustar a distribuição generalizada de valores extremos (GEV), que inclui como casos particulares, as distribuições Gumbel, Fréchet e Weibull, aos dados de temperatura máxima mensal do município de Campina Grande; aplicar o método da máxima verossimilhança para obter as estimativas dos parâmetros da GEV; verificar o ajuste do modelo aos dados por meio do teste de Kolmogorov-Sminorv; gráficos de probabilidade-probabilidade e o quantil-quantil e verificar qual das três distribuições assintóticas (Weibull, Fréchet e Gumbel), de valores extremos se ajustam melhor aos dados de temperatura máxima.

[1] http://www.inpe.br/

2 Fundamentação Teórica

Nesta sessão estão descritos os principais aspectos teóricos da teoria de valores extremos, bem como a distribuição generalizada de valores extremos, que por sua vez, servirão de base para nossa aplicação.

2.1 Teoria de valores extremos para variáveis aleatórias univariadas

Suponha que $Y_1,\ Y_2,\ ...,\ Y_n$ sejam variáveis aleatórias independentes e identicamente distribuídas (*i.i.d.*) com função de distribuição $F_Y(y)$, considera-se M_n como sendo o máximo das n variáveis aleatórias, isto é,

$$M_n \ = \ max\left(Y_1,\ Y_2,\ ...,\ Y_n\right). \tag{2.1}$$

A função de distribuição da variável aleatória M_n, $F_{M_n}(y)$ é definida por

$$F_{M_n}(y) = P\left\{M_n \leq y\right\} = P\left\{Y_1 \leq y,\ ...,\ Y_n \leq y\right\} = \left(F_Y(y)\right)^n. \tag{2.2}$$

Mas, para n tendendo ao infinito, a função de distribuição de M_n pode ser degenerada[2] (COLES, 2001).

Conforme Olinda (2012), o máximo de uma amostra simplesmente tende para extremidade direita da distribuição quase certamente $(q.c.)$[3], não importando se a variável é finita ou infinita. Seja Y_F o ponto final à direita, então

$$\begin{aligned}\sum_{n=1}^{\infty} P\left\{|M_n - Y_F| > \varepsilon\right\} \ &= \ \sum_{n=1}^{\infty} P\left\{M_n < Y_F - \varepsilon\right\} = \\ \sum_{n=1}^{\infty} P\left\{Y_1 < Y_F - \varepsilon\right\}^n \ &= \ \frac{P\left\{Y_1 < Y_F - \varepsilon\right\}}{1 - P\left\{Y_1 < Y_F - \varepsilon\right\}} < \infty,\end{aligned}$$

e isso mostra que $M_n \xrightarrow{q.c.} Y_F$. Em forma de limite, a TVE assegura a existência de uma distribuição assintótica não degenerada, Z, para uma transformação linear de M_n, isto é, para constantes apropriadas $a_n > 0$ e $b_n \in \mathcal{R}$, tem-se que

[2] Observe que no Teorema Central do Limite este denominador é da ordem de $\sqrt{n}$, pois $Var\left[S_n\right] = nVar\left[Y_1\right]$ devido ao fato que $Y_1,\ Y_2,\ ...$ são variáveis aleatórias independentes e identicamente distribuídas. A substituição de $\sqrt{n}$ por n diminui a variância de tal forma que o limite é uma variável aleatória degenerada, ou seja, a massa de probabilidade é centrada em um único valor da variável aleatória.

[3] Convergência Quase-Certa: **Definição** Seja Y_n uma sequência de variáveis aleatórias. Diz que Y_n converge quase certamente para Y, e escreve-se $Y_n \xrightarrow{q.c.} Y$, se $P[w \in \Omega : Y_n(w) \xrightarrow{n} Y(w)] = 1$. A convergência quase certa é um tipo de convergência mais forte do que a em probabilidade e por isso implica a convergência em probabilidade (mas o inverso não é verdadeiro).

$$\begin{aligned} \lim_{n\to\infty} (F_Y(y))^n &= \lim_{n\to\infty} P\left[\frac{M_n - b_n}{a_n} \leq y\right] = \lim_{n\to\infty} P\left(M_n \leq a_n y + b_n\right) = \\ &= \lim_{n\to\infty} \left(F_Y\left(a_n y + b_n\right)\right)^n = Z\left(y\right), \end{aligned} \tag{2.3}$$

para $y \in \mathcal{R}$ já que as possíveis funções de distribuições de Z (que serão apresentadas a seguir) são funções contínuas em $\mathcal{R}$. As características e propriedades da distribuição *GEV* são determinadas pelas caudas extremas (inferior e superior) da distribuição dos dados (KOTZ; NADARAJAH, 2000).

Teorema de Fisher-Tippett: *Seja* $M_n = max(Y_1,\ Y_2,\ ...,\ Y_n)$ *em que* $Y_i, i = 1,\ 2,\ ...,\ n$, *são variáveis aleatórias i.i.d. Se para algumas sequências numéricas (constantes em n)* $a_n > 0$ *e* $b_n \in \mathcal{R}$, *tem-se*

$$P\left(a_n^{-1}\left(M_n - b_n\right) \leq y\right) \xrightarrow{d} Z(y), \tag{2.4}$$

para alguma Z *não degenerada. Reciprocamente, cada função de distribuição* Z *do tipo valor extremo aparece como um dos limites em (2.4) e de fato, aparece quando* Z *é, ela mesma, a função de distribuição de cada* Y_i.

O Teorema de Fisher-Tippett (1928), fornece a distribuição limite para o máximo coletado em blocos de tamanho n. Sendo $y_1,\ ...,\ y_k$ as realizações da variável aleatória Y, consideram-se blocos de tamanho k desta amostra, isto é, os dados amostrais são particionados em n blocos tais que $nk \leq m$. Define-se, então

$$\begin{aligned} M_1 &= max\left\{y_1,\ ...,\ y_k\right\} \\ M_2 &= max\left\{y_{k+1},\ ...,\ y_{2k}\right\} \\ &\vdots \\ M_n &= max\left\{y_{nk-k+1},\ ...,\ y_{nk}\right\}. \end{aligned}$$

Por meio dessa nova amostra, $M_1,\ ...,\ M_n$, podem-se obter os estimadores de $\hat{\mu}$, $\hat{\sigma}$ e $\hat{\xi}$ e , assim, determinar a distribuição extremal de Z.

Pelo Teorema pode-se, então, estimar a distribuição assintótica de $\frac{M_n - b_n}{a_n}$ diretamente da família Z sem fazer nenhuma referência à distribuição de Y pois tem-se que a distribuição Z (que corresponde à distribuição dos máximos $M_n = max(Y_1,\ ...,\ Y_n)$ em que $Y_i,\ i = 1,\ ...,\ n$, são variáveis aleatórias *i.i.d.*) é uma das três distribuições descritas anteriormente. A expressão seguinte incorpora os três tipos de distribuições de valores extremos:

$$Z\left(y\right) = \begin{cases} \exp\left[-\left(1+\gamma y\right)^{-\frac{1}{\gamma}}\right], & se\ \ \gamma \neq 0 \\ \exp\left[-\exp\left(-y\right)\right], & se\ \ \gamma = 0 \end{cases}$$

em que $(1+\gamma y) > 0$.

Pode-se, também, utilizar a reparametrização $\xi = -\gamma$ e, assim, obter:

$$Z(y) = \begin{cases} \exp\left[-(1-\xi y)^{\frac{1}{\xi}}\right], & se \;\; \xi \neq 0 \\ \exp\left[-\exp(-y)\right], & se \;\; \xi = 0 \end{cases}$$

em que $(1-\xi y) > 0$.

Pode-se, ainda, incorporar parâmetros de locação e escala na distribuição de valores extremos substituindo-se y por $\frac{y-\mu}{\sigma}$ em que $\mu \in \mathcal{R}$ e $\sigma > 0$. Assim, a família locação-escala e forma pode ser escrita em uma distribuição GEV a partir de

$$Z(y;\mu,\sigma,\xi) = \exp\left\{-\left[1-\frac{\xi(y-\mu)}{\sigma}\right]^{1/\xi}\right\}, \tag{2.5}$$

em que $1-\xi(y-\mu)/\sigma > 0,\ \sigma > 0$ e μ e σ arbitrário. O caso $\xi = 0$ é interpretado $\xi \to 0$, que é

$$Z(y;\mu,\sigma,0) \;\; = \;\; \exp\left\{-\exp\left[-\frac{(y-\mu)}{\sigma}\right]\right\}. \tag{2.6}$$

A condição $1-\xi(y-\mu)/\sigma > 0,\ \sigma > 0$ garante que y seja limitado superiormente e inferiormente por $\frac{\mu+\sigma}{\xi}$, respectivamente, se $\xi > 0$ e $\xi < 0$.

2.2 Teste de aleatoriedade

Conforme Medeiros (2011), na primeira etapa da análise verifica-se a hipótese de independência dos dados observados por meio do teste de chorrilho ("run tes") descrito por Zar (1999). De acordo com Silva (2008) esse teste inicialmente consiste em definir uma sequência dicotômica de tamanho n, a partir dessa amostra aleatória Y_1, ..., Y_n aplica-se a cada Y_i a função indicadora $A(y_i) = I_{y_i > M_d}(Y_i)$ e omitindo-se os casos em que $y_i = M_d$ sendo M_d a mediana dos dados.

O valor da mediana, M_d, é definido a seguir

$$M_d = \begin{cases} y_{\left(\frac{n+1}{2}\right)}, & se \;\; n \;\; for \;\; ímpar, \\ \\ \frac{y_{\left(\frac{n}{2}\right)} + y_{\left(\frac{n}{2}+1\right)}}{2}, & se \;\; n \;\; for \;\; par. \end{cases}$$

Os valores da variável indicadora A_i , com (i=1,2,. .., n), é definida por

$$A_i = \begin{cases} 1, & se \quad y_i > M_d \\ 0, & se \quad y_i < M_d. \end{cases}$$

A variável aleatória tem um total de sequências de zeros e uns ao longo da amostra R, e seu valor observado é definido por uma sequência de variáveis aleatórias N_1 como sendo o número total de ocorrências de $Y_1 > M_d$ e N_2 como sendo o número total de ocorrências de $Y_i < M_d$, cujos valores observados são, respectivamente, n_1 e n_2.

Os procedimentos adotados por Silva (2008), para $n_1 < 30$ e $n_2 < 30$, e Zar (1999), apresentam pares de valores críticos exatos $(r_{1,\,\alpha,\,n_1,\,n_2};\ r_{2,\,\alpha,\,n_1,\,n_2})$ ao nível de significância α. Sendo assim, rejeita-se a hipótese nula se $r \leq [r_{1,\,\alpha,\,n_1,\,n_2};\ \alpha]$ ou se $r \leq [r_{2,\,\alpha,\,n_1,\,n_2}; \alpha]$. Caso $n_1 \geq 30$ ou $n_2 \geq 30$, sob hipótese nula H_0 de independência tem-se que, assintoticamente R segue uma distribuição normal com esperança definida por

$$E\left(R\right) = \frac{2N_1N_2}{N} + 1,$$

com variância descrita na forma

$$Var\left(R\right) = \frac{2N_1N_2(2N_1N_2 - n)}{n^2(n-1)}$$

e suas estimativas definidas por

$$\hat{E}\left(R\right) = \frac{2n_1n_2}{n} + 1$$

$$\hat{Var}\left(R\right) = \frac{2n_1n_2\left(2n_1n_2 - n\right)}{n^2(n-1)},$$

em que n_1 e n_2 são valores observados de N_1 e N_2.

Na sequência, efetuando-se o teste deve-se calcular a mediana da amostra observada da temperatura máxima $y_1, ..., y_n$, obtendo-se uma sequência dicotômica dessa mesma amostra $A\left(y_1\right),\ ...,\ A\left(y_n\right), n_1,\ n_2,\ r,\ \hat{E}\left(R\right)$ e $\hat{Var}\left(R\right)$ e por fim, calcula-se o p-valor da seguinte forma

$$P\left(\frac{|r - E\left(R\right)| - 0,5}{\sqrt{Var\left(R\right)}} \geq q_r\right)$$

em que, q_r é o quantil de ordem $\frac{\alpha}{2}$ da normal padrão e, α é o nível de significância adotado para o teste. Uma vez testada a independência das observações segui-se à estimação dos parâmetros.

2.3 Método de estimação

Métodos de estimação é o meio pelos quais os parâmetros desconhecidos do modelo são inferidos com base nos dados amostrais. Diferentes abordagens tem sido propostas para estimar modelos de valores extremos, adotando-se uma visão singular que restringe e observar as técnicas baseadas na função de verossimilhança. Neste trabalho adota-se o método de máxima verossimilhança, que tem todo um aspecto e propriedades convenientes na inferência estatística.

2.3.1 Método da máxima verossimilhança

Segundo Ferrari (2011), algumas técnicas estão sendo apresentadas para fazer inferências sobre os parâmetros da distribuição *GEV*. As mesmas incluem técnicas gráficas que são baseadas nos gráficos de probabilidade, estimadores baseados no método de momentos, métodos de regressão, métodos de L-momentos e o método de máxima verossimilhança. Em alguns casos normais, os estimadores de máxima de verossimilhança são consistentes, assintoticamente normais e eficientes, os casos não regulares acontecem quando a distribuição em estudo depende de parâmetros desconhecidos. Smith (1985) realizou um estudo cuidadoso sobre o comportamento assintótico dos estimadores de máxima verossimilhança para a distribuição *GEV* e obteve os seguintes resultados no que se refere ao parâmetro de forma:

1. quando $\xi > -0,5$, os estimadores de máxima verossimilhança são regulares no sentido de ter as propriedades assintóticas habituais;

2. quando $-1 < \xi < -0,5$, os estimadores de máxima verossimilhança existem mas não são regulares;

3. quando $\xi < -1$, esses estimadores provavelmente não existem.

O caso em que $\xi < -0,5$ corresponde a uma distribuição com uma cauda superior muito curta e leve e, segundo Smith (1985), essa situação raramente é encontrada em aplicações de modelagem de valores extremos sendo que, as limitações teóricas da abordagem de máxima de verossimilhança geralmente não são obstáculos na prática.

Supondo que $Y_1, Y_2, ..., Y_n$ são variáveis aleatórias *i.i.d.*, os estimadores são apresentados da seguinte forma:

$$L\left(\mu,\sigma,\xi|y\right) = \prod_{i=1}^{n} f_Y\left(y_i|\mu,\sigma,\xi\right) =$$

$$= \sigma^{-n} \prod_{i=1}^{n} \left[1+\xi\left(\frac{y_i-\mu}{\sigma}\right)\right]^{-\left(\frac{1+\xi}{\xi}\right)} exp\left\{-\sum_{i=1}^{n}\left[1+\xi\left(\frac{y_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}. \tag{2.7}$$

Calculando-se o logaritmo da função de verossimilhança da Equação (2.7) , tem-se que

$$(\boldsymbol{\theta}|y) = ln\left[L\left(\mu,\sigma,\xi|y\right)\right] =$$

$$= -n\ln\left((\sigma)\right) - \left(\frac{1+\xi}{\xi}\right)\sum_{i=1}^{n}\ln\left[1+\xi\left(\frac{y_i-\mu}{\sigma}\right)\right] - \sum_{i=1}^{n}\left[1+\xi\left(\frac{y_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}$$

$$= \sum_{i=1}^{n}\left\{-\ln\left(\sigma\right) - \left(\frac{1+\xi}{\xi}\right)\ln\left[1+\xi\left(\frac{y_i-\mu}{\sigma}\right)\right] - \left[1+\xi\left(\frac{y_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\},$$

para $\xi < 0$ e $y_i < \mu - \frac{\sigma}{\xi}$ (ou seja, $\mu - \frac{\sigma}{\xi} > M_{(n)}$), ou para $\xi > 0$ $y_i > \mu - \frac{\sigma}{\xi}$ (ou seja, $\mu - \frac{\sigma}{\xi} > M_{(1)}$). Maximizando-se a Equação (2.7), com relação ao vetor de parâmetros $\boldsymbol{\theta}^{'} = (\mu, \sigma, \xi)^{'}$, obtém-se a estimativa da função de máxima verossimilhança para a família *GEV*, conduzindo-se ao sistema de equações não lineares, definido por:

$$\frac{l\left(\boldsymbol{\theta}|y\right)}{\partial\mu} = \frac{1}{\hat{\sigma}}\sum_{i=1}^{n}\left(\frac{1+\hat{\xi}-\omega_i^{-\frac{1}{\hat{\xi}}}}{\omega_i}\right) = 0.$$

$$\frac{\partial l\left(\boldsymbol{\theta}|y\right)}{\partial\sigma} = -\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}\left\{\frac{\left(1+\hat{\xi}-\omega_i^{-\frac{1}{\hat{\xi}}}\right)(y_i-\hat{\mu})}{\omega_i}\right\} = 0. \tag{2.8}$$

$$\frac{\partial l\left(\boldsymbol{\theta}|y\right)}{\partial\xi} = \sum_{i=1}^{n}\left\{\left(1-\omega_i^{-\frac{1}{\hat{\xi}}}\right) - \left[\frac{\ln\left(\omega_i\right)}{\hat{\xi}} - \left(\frac{y_i-\hat{\mu}}{\hat{\xi}\hat{\sigma}\omega_i}\right)\right] - \frac{y_i-\hat{\mu}}{\hat{\sigma}\omega_i}\right\} = 0,$$

em que $\omega_i = 1 + \hat{\xi}\left(\frac{y_i-\hat{\mu}}{\hat{\sigma}}\right)$. No caso particular da distribuição Gumbel a função de verossimilhança, definida pela a Equação (2.7), em que $\xi = 0$, conduz a seguinte equação

$$L\left(\mu,\sigma|y\right) = \prod_{i=1}^{n} f_Y\left(y_i|\mu,\sigma\right) = \sigma^{-n}\exp\left\{\sum_{i=1}^{n}\left(-\frac{y_i-\mu}{\sigma}\right)\right\}\exp\left\{\sum_{i=1}^{n}\exp\left(\frac{y_i-\mu}{\sigma}\right)\right\},$$

$$l\left(\boldsymbol{\theta}|y\right) = ln\left(L\left(\mu,\sigma|y\right)\right) = \sum_{i=1}^{n}\left\{-ln\left(\sigma\right) - \left(\frac{y_i-\mu}{\sigma}\right) - \exp\left(-\frac{y_i-\mu}{\sigma}\right)\right\}. \tag{2.9}$$

derivando-se a Equação (2.9), em que $\hat{\mu}$ e $\hat{\sigma}$ são os parâmetros de máxima verossimilhança, obtidos pelo sistema de equações não lineares, isto é,

$$\frac{\partial l\left(\boldsymbol{\theta}|y\right)}{\partial\mu} = \frac{1}{\hat{\sigma}}\left\{\left[\sum_{i=1}^{n}\exp\left(-\frac{y_i-\hat{\mu}}{\hat{\sigma}}\right)\right] - n\right\} = 0. \tag{2.10}$$

$$\frac{\partial l\left(\boldsymbol{\theta}|y\right)}{\partial\sigma} = -\frac{n}{\hat{\sigma}} + \sum_{i=1}^{n}\left(\frac{y_i-\hat{\mu}}{\hat{\sigma}^2}\right)\left[1-\exp\left(-\frac{y_i-\hat{\mu}}{\hat{\sigma}}\right)\right] = 0. \tag{2.11}$$

Os sistemas de Equações (2.10) e (2.11), em geral, não possuem soluções exatas pois são equações não lineares. Uma solução aproximada é calculada pelo método iterativo de quasi-Newton que, para iniciar o algoritmo, especifica uma estimativa inicial para μ, σ e ξ. Neste trabalho, o software *R* (*R* CORE TEAM, 2017) é utilizado para calcular as estimativas de máxima verossimilhança com auxílio do pacote *evd* satisfazendo-se o critério de convergência do método.

2.4 Seleção da distribuição de valores extremos

Conforme Hoskink (1984), um dos procedimentos para verificar se as observações seguem a distribuição de valores extremos tipo I(Gumbel), II(Fréchet) ou III(Weibull), basta verificar se ξ é igual a zero na distribuição GEV, o que pode ser feito através do teste da razão de verossimilhança modificado, exibido a seguir.

Toma-se uma série de n observações ($y_1,\ y_2,\ ...,\ y_n$), $l(\hat{\boldsymbol{\theta}}_{GVE})$ e $l(\hat{\boldsymbol{\theta}}_{Gumbel})$ do máximo do logaritmo das funções de verossimilhança das distribuições GVE e Gumbel em que, $\hat{\boldsymbol{\theta}}'_{GVE} = (\hat{\mu}, \hat{\sigma}, \hat{\xi})'$ e $\hat{\boldsymbol{\theta}}'_{Gumbel} = (\hat{\mu}, \hat{\theta})'$, são vetores das estimativas da função de máxima verossimilhança. A estatística de razão de verossimilhança (T_{LR}) é definido por

$$T_{LR} = -2\left[l\left(\hat{\boldsymbol{\theta}}_G\right) - l\left(\hat{\boldsymbol{\theta}}_{GEV}\right)\right] = 2\left[l\left(\hat{\boldsymbol{\theta}}_{GEV}\right) - l\left(\hat{\boldsymbol{\theta}}_G\right)\right] \tag{2.12}$$

que tem distribuição assintótica χ^2 com 1 grau de liberdade.

Hosking (1984) sugere a utilização da estatística modificada cujo objetivo é de atender uma melhor aproximação à distribuição assintótica para Equação (2.12)

$$T^*_{LR} = \left(1 - \frac{2,8}{n}\right) \times T_{LR}, \tag{2.13}$$

sendo n o tamanho da amostra.

Para testar a hipótese $H_0 : \xi = 0$ versus $H_1 : \xi \neq 0$, basta comparar o valor da estatística do teste T^*_{LR} com o valor tabelado da distribuição χ^2 com grau 1 de liberdade e um certo nível de significância (α), $\chi^2_{[\alpha,1]}$. Se $T^*_{LR} \geq \chi^2_{[\alpha,1]}$, rejeita-se H_0, ou seja, há fortes evidências de que as observações não são de uma distribuição do tipo I(Gumbel).

2.5 Diagnóstico do ajuste da *GEV*

Em geral, quando se tem interesse em verificar o ajuste da distribuição GEV aos dados amostrais utiliza-se alguns testes de aderência, como por exemplo, o teste de Kolmogorov-Smirnov (OLINDA, 2012). A utilização desta metodologia pode ser encontrada em Bautista (2002), Freire e Beijo (2010), Sansigolo (2008) e Ferrari (2011).

De acordo com Ferrari (2011), ao se ajustar uma distribuição de probabilidade a um conjunto de dados, trabalha-se com a hipótese de que a distribuição representa adequadamente aquele conjunto de informações. Seja $y_{(1)},\ y_{(2)},\ ...,\ y_{(n)}$, uma série de dados observados ordenados de forma crescente, a função de distribuição acumulada empírica da variável aleatória Y poderá ser descrita da seguinte forma,

$$\hat{F}\left(y_{(i)}\right) = \frac{i}{n+1}, i = 1, 2, ..., n. \tag{2.14}$$

Para se testar a suposição de que os dados seguem uma distribuição *GEV* selecionada, pode-se recorrer à estatística D do teste de Kolmogorov-Smirnov, definida da seguinte forma

$$D = \max \left| F\left(y_{(i)}\right) - \hat{F}\left(y_{(i)}\right) \right|, \quad i = 1, 2, ..., n, \tag{2.15}$$

sendo $F(y_{(i)})$ a distribuição teórica da distribuição *GEV* com suas estimativas obtidas e $\hat{F}(y_{(i)})$ é a distribuição empírica definida pela Equação (2.14). Rejeita-se a hipótese H_0 de que os dados seguem uma distribuição *GEV* se a estatística de teste for $D \geq D_{[n,\,\alpha]}$, em que o valor crítico é $D_{[n,\,\alpha]}$ para os valores de n com um nível de significância predeterminado. Além do teste descrito anteriormente, o ajuste da distribuição poderá ser avaliado graficamente por meio da construção do gráfico *qq*-plot (gráfico quantil-quantil) e *pp*-plot (gráfico de probabilidade-probabilidade).

O gráfico *pp*-plot é construído com os pontos dados pelas coordenadas,

$$\left[\hat{F}\left(y_{(i)}\right), F\left(y_{(i)}\right)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right], i = 1, 2, ..., n,$$

em que $\hat{\boldsymbol{\theta}}$ são as estimativas de $\hat{\boldsymbol{\theta}}' = (\mu,\ \sigma,\ \xi)'$, $F(y_{(i)})$ é a função de distribuição acumulada da *GEV* e $\hat{F}(y_{(i)})$ é uma distribuição empírica definida pela Equação 2.14, se a função da distribuição *GEV* é um modelo razoável para a distribuição dos dados amostrais, os pontos estarão aproximadamente alinhados em uma reta pelos pontos (0,0) e (1,1). Assim, uma forma de interpretar o gráfico é observar o quão distante esses pontos estão da reta. Quanto mais distante, menos adequada é a distribuição de probabilidade.

Conforme Bautista (2002), outra forma de avaliar graficamente o ajuste da distribuição é a utilização do gráfico *qq*-plot, formado pelos pontos de coordenadas

$$\left[\hat{F}^{(-1)}\left(\frac{i}{n+1}\right), y_i\right], i = 1, 2,\ ...,\ n,$$

em que $\hat{F}^{(-1)}(.)$ é a função inversa da Equação 2.14. Neste caso, também sob a hipótese de que os dados apresentam distribuição *GEV*, os pontos do gráfico estarão aproximadamente alinhados em uma reta. Quanto mais afastados de uma reta, menos adequada é a distribuição de probabilidade.

2.6 Período de retorno e nível de retorno

Conforme Coles (2001), o período de retorno (τ) é o intervalo de tempo estimado para ocorrência de um determinado evento e é definido como o inverso da probabilidade de um evento a ser igualado ou superado, ou seja,

$$\tau = \frac{1}{p},$$

em que p é a probabilidade do evento ser igualado ou ultrapassado $[(P(Y \geq y))]$.

No caso em estudo, o período de retorno é o intervalo de tempo estimado para a ocorrência de temperaturas máximas em é definido por

$$\tau = \frac{1}{1 - F(y)},$$

O nível de retorno (y_p), está associado ao período do retorno τ e a sua função é obtida por meio da solução da equação abaixo

$$\int_{i=\infty}^{y_p} f(\boldsymbol{\theta}) dy = 1 - p, \tag{2.16}$$

em que $p = \frac{1}{\tau}$, ou seja,

$$F(y_p) = (1 - p),$$

ao inverter a Equação (2.14), tem-se a solução

$$y_p = F^{-1}(1 - p) = \mu - \frac{\sigma}{\xi}\left\{1 - [-ln(1 - p)]^{-\xi}\right\}.$$

Para $\xi \neq 0$, do qual o limite $\xi \longrightarrow 0$ é definido a seguir

$$y_p = F^{-1}(1 - p) = \mu - \sigma\{ln[-ln(1 - p)]\}.$$

De acordo com Medeiros (2011), o nível y_p deverá ser excedido em média uma vez a cada $\frac{1}{p}$ anos. Mais precisamente, y_p é excedido pelo máximo anual em algum ano particular com probabilidade p. A estimativa de $\hat{y}_p$ do nível de retorno y_p para períodos de retorno τ é obtida pela substituição das estimativas de máxima verossimilhança de μ, σ e ξ em (2.16).

2.7 Obtenção dos intervalos de confiança

Os Intervalos de Confiança $(I.C.)$ com nível de $(1-\alpha)100\%$ para os níveis de retorno Y_p foram construídos e baseados no método delta, logo depois, no método estatístico de razão de verossimilhança. O intervalo de confiança para Y_p com $(1 - \alpha)100\%$ de confiança é descrito a seguir:

$$[I.C.(y_p)] = \left[\hat{y}_p \pm z_{\alpha/2}\sqrt{Var(\hat{y}_p)}\right],$$

em que α é o nível de significância, $z_{\alpha/2}$ o valor tal que $P(\mid Z \mid < z_{\alpha/2}) = 1 - \alpha$, Z uma variável com distribuição normal padronizada e $Var(\hat{y}_p)$ é a variância associada ao nível de

retorno $\hat{y}_p$ calculada através do método delta. Esse método é baseado no fato de que uma distribuição de $\hat{\boldsymbol{\theta}}' = (\hat{\mu}, \hat{\sigma}, \hat{\xi})'$ ser assintoticamente normal com média $\hat{\boldsymbol{\theta}}' = (\mu, \sigma, \xi)'$ e as matrizes de variâncias e covariância dado $\boldsymbol{I}(\boldsymbol{\theta})^{-1}$. Tendo em vista que a Equação (2.7) é uma função não linear em μ, σ e ξ, pode-se linearizá-la por meio da expansão da primeira ordem de Taylor em torno do ponto inicial correspondente ao vetor das estimativas dos parâmetros.

Conforme Ferrari (2011), o método delta descrito por Rao e Toutemburg (1999), é realizado da seguinte forma: calcula-se $Var(\hat{y}_p)$ por meio da matriz de variâncias e covariâncias de μ, σ e ξ, estimada pela inversa da matriz de segundas derivadas da função log-verossimilhança (matriz hessiana calculada em μ, σ, e ξ). Assim o método delta estima a variância de $\hat{y}_p$ por meio da expressão

$$Var\left(\hat{y}_p\right) \approx \nabla y_p' V \nabla y_p. \tag{2.17}$$

Sendo, para o caso em que $\xi \neq 0$, $\boldsymbol{J}$ é uma matriz de variâncias e covariâncias de $\hat{\boldsymbol{\theta}}' = (\hat{\mu}, \hat{\sigma}, \hat{\xi})'$ obtidos por meio da inversa da matriz de informação a seguir

$$\boldsymbol{J} = \left[\begin{array}{ccc} \frac{\partial^2}{\partial\mu\partial\mu} l\left(\boldsymbol{\theta}\right) & \frac{\partial^2}{\partial\mu\partial\sigma} l\left(\boldsymbol{\theta}\right) & \frac{\partial^2}{\partial\mu\partial\xi} l\left(\boldsymbol{\theta}\right) \\ \frac{\partial^2}{\partial\sigma\partial\mu} l\left(\boldsymbol{\theta}\right) & \frac{\partial^2}{\partial\sigma\partial\sigma} l\left(\boldsymbol{\theta}\right) & \frac{\partial^2}{\partial\sigma\partial\xi} l\left(\boldsymbol{\theta}\right) \\ \frac{\partial^2}{\partial\xi\partial\mu} l\left(\boldsymbol{\theta}\right) & \frac{\partial^2}{\partial\xi\partial\sigma} l\left(\boldsymbol{\theta}\right) & \frac{\partial^2}{\partial\xi\partial\xi} l\left(\boldsymbol{\theta}\right) \end{array} \right]^{-1}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \left[\begin{array}{ccc} Var\left(\hat{\mu}\right) & Cov\left(\hat{\mu}, \hat{\sigma}\right) & Cov\left(\hat{\mu}, \hat{\xi}\right) \\ Cov\left(\hat{\mu}, \hat{\sigma}\right) & Var\left(\hat{\sigma}\right) & Cov\left(\hat{\sigma}, \hat{\xi}\right) \\ Cov\left(\hat{\mu}, \hat{\xi}\right) & Cov\left(\hat{\sigma}, \hat{\xi}\right) & Var\left(\hat{\xi}\right) \end{array} \right],$$

e

$$\nabla y_p^T = \left[\frac{\partial y_p}{\partial u}, \frac{\partial y_p}{\partial \sigma}, \frac{\partial y_p}{\partial \xi}\right],$$

a matriz de derivadas parciais de y_p avaliada em μ, σ, e ξ.

Portanto, a variância do nível de retorno estimado $y_{\hat{p}}$,para $\xi \neq 0$, é calculada da seguinte forma

$$Var\left(\hat{y}_p\right) = \left(\frac{\partial \hat{y}_p}{\partial \mu}\right) Var\left(\hat{\mu}\right) + \left(\frac{\partial \hat{y}_p}{\partial \sigma}\right)^2 Var\left(\hat{\sigma}\right) + \left(\frac{\partial \hat{y}_p}{\partial \xi}\right) Var\left(\hat{\xi}\right)$$

$$+2\frac{\partial \hat{y}_p}{\partial \mu}\frac{\partial \hat{y}_p}{\partial \sigma} Cov\left(\hat{\mu}, \hat{\sigma}\right) + 2\frac{\partial \hat{y}_p}{\partial \mu}\frac{\partial \hat{y}_p}{\partial \xi} Cov\left(\hat{\mu}, \hat{\xi}\right) + 2\frac{\partial \hat{y}_p}{\partial \sigma}\frac{\partial \hat{y}_p}{\partial \xi} Cov\left(\hat{\sigma}, \hat{\xi}\right),$$

em que,

$$\frac{\partial \hat{y}_p}{\partial \mu} = 1$$

$$\frac{\partial \hat{y}_p}{\partial \sigma} = -\frac{1}{\hat{\xi}}\left\{1 - \left[-\ln\left(1-p\right)\right]^{-\hat{\xi}}\right\},$$

$$\frac{\partial \hat{y}_p}{\partial \xi} = \frac{\hat{\sigma}}{\hat{\xi}^2}\left\{1 - \left[-\ln\left(1-p\right)\right]^{-\hat{\xi}}\right\} - \frac{\hat{\sigma}}{\xi}\left[-\ln\left(1-p\right)\right]^{-\hat{\xi}} \ln\left[\ln\left(1-p\right)\right].$$

Para o caso em que $\xi = 0$, tem-se a matriz de variâncias e covariâncias de $\hat{\boldsymbol{\theta}}' = (\hat{\mu}, \hat{\sigma})'$ obtidos da inversa da matriz de informação definida por

$$\boldsymbol{V} = \left[\begin{array}{cc} \frac{\partial^2}{\partial\mu\partial\mu} l\left(\boldsymbol{\theta}\right) & \frac{\partial^2}{\partial\mu\partial\sigma} l\left(\boldsymbol{\theta}\right) \\ \frac{\partial^2}{\partial\sigma\partial\mu} l\left(\boldsymbol{\theta}\right) & \frac{\partial^2}{\partial\sigma\partial\sigma} l\left(\boldsymbol{\theta}\right) \end{array}\right]^{-1}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \left[\begin{array}{cc} Var\left(\hat{\mu}\right) & Cov\left(\hat{\mu}, \hat{\sigma}\right) \\ Cov\left(\hat{\mu}, \hat{\sigma}\right) & Var\left(\hat{\sigma}\right) \end{array}\right],$$

em que $\hat{\boldsymbol{\theta}}' = (\hat{\mu}, \hat{\sigma})'$ são as estimativas de máxima verossimilhança.

3 Resultados e Discussão

Os dados de temperatura máxima mensais (expressos em graus Celsius (°C)) foram obtidos no período de 1970 à 2010 entre janeiro a dezembro, fornecidos pela EMBRAPA algodão, localizada no município de Campina Grande, estado da Paraíba, os dados foram obtidos brutos e nós calculamos as máximas. Realizou-se um estudo descritivo para cada um dos meses, cujos resultados encontram-se na Tabela 1. Conforme pode-se observar, ocorreu uma assimetria entre os meses de janeiro, março, abril, maio, junho, julho, agosto e outubro e os meses de fevereiro, setembro, novembro e dezembro não se encaixaram nessa assimetria.

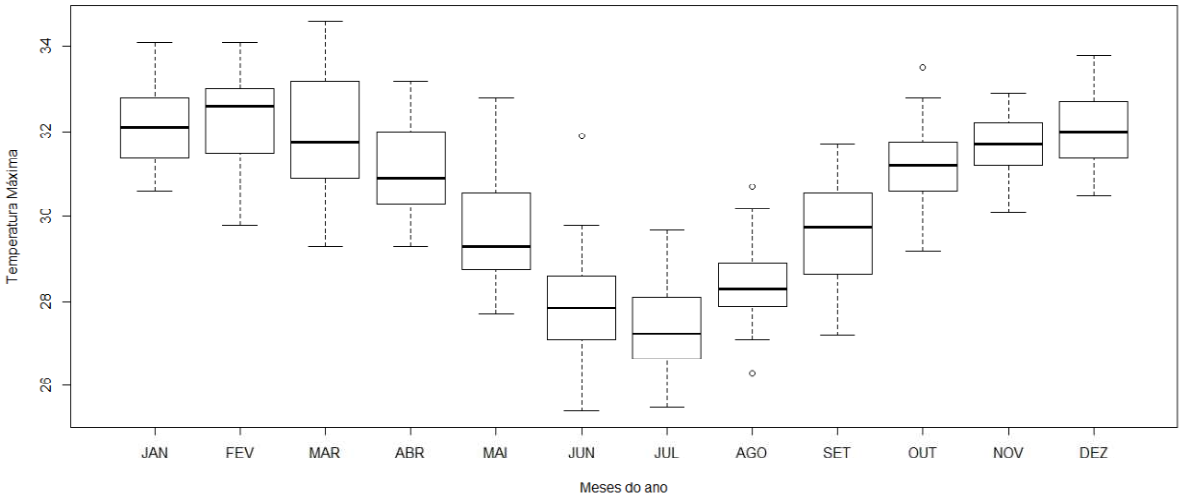
Tabela 1 – Estatísticas descritivas da variável aleatória temperatura máxima (°C) mensal no período entre 1970 a 2010, do município de Campina Grande-PB, em que DP é o desvio padrão, C.V. é o coeficiente de variação e C.A. é o coeficiente de assimetria

Meses	Média	Mediana	Variância	*D.P.*	*C.V.*(%)	*C.A.*	Curtose
janeiro	32,24	32,10	0,6968	0,8347	2,6005	0,2584	-0,6476
fevereiro	32,32	32,60	1,1563	1,0753	3,2985	-0,4265	-0,7429
março	31,94	31,75	2,2218	1,4905	4,6947	0,1580	-1,1463
abril	31,02	30,90	0,9509	0,9751	3,1559	0,3106	-0,7926
maio	29,63	29,30	1,7823	1,3350	4,5565	0,4766	-0,6584
junho	27,93	27,85	1,6601	1,2884	4,6264	0,7990	1,3095
julho	27,38	27,25	0,9765	0,9882	3,6264	0,2375	-0,3763
agosto	28,38	28,30	1,1312	1,0635	3,7582	0,0684	-0,3561
setembro	29,66	29,75	1,2852	1,1336	3,8106	-0,2386	-0,8852
outubro	31,26	31,20	0,8795	0,9378	3,0058	0,2540	-0,1213
novembro	31,64	31,70	0,5125	0,7159	2,2585	-0,2394	-0,5843
dezembro	32,06	32,00	0,8637	0,9293	2,9043	-0,0158	-0,9506

Pode-se observar, por meio da Tabela 1, que os meses de janeiro, fevereiro, março e dezembro, apresentam em média os valores mais altos de temperatura máxima, no qual o mês de fevereiro apresenta-se, em média, o maior valor. Para os meses de abril, outubro e novembro, a média e a mediana estão mais próximas e os demais meses, isto é, maio, junho, julho, agosto e setembro a mediana e a média são bem menores, podendo-se levar a uma possível assimetria.

Observa-se por meio da Figura 1, o gráfico box-plot, que por sua vez, refere-se aos dados da variável temperatura máxima de cada mês analisado, no qual pode-se perceber presença de alguns valores discrepantes (atípicos) representado pelo símbolo (∘), nos respectivos meses de junho, agosto e outubro, os demais não apresentam valores atípicos. Percebe-se que os meses de março, maio e setembro possui uma considerável

Figura 1 – Gráfico de caixa (Box-Plot) referente a temperatura máxima no período de 1970 a 2010 do município de Campina Grande.



assimetria em relação aos demais meses analisados. Este fato corrobora Bezerra (2013) ao ajustar a distribuição *GEV* aos dados de temperatura máxima mensal no município de Campina Grande, estado da Paraíba. A referida autora observou também uma considerável assimetria entre os meses de março e maio.

Na sequência (Tabela 2) observa-se o teste de chorrilho, realizado para verificar a pressuposição de independência entre os dados de temperatura máxima, ao nível de significância de 5%, sendo o mesmo comparado com seu respectivo valor-p. Pode-se observar também que apenas no mês de outubro o nível descritivo é menor que o nível de significância adotado, concluindo-se que a hipótese de independência entre os dados, para o mês de outubro, foi rejeitada. Nos demais meses não houve rejeição da hipótese, ou seja, a hipótese de independência não foi rejeitada ao nível de significância de 5%. De acordo com Bautista (2004), o cumprimento dessa pressuposição é importante no que se refere a obtenção de inferências estatísticas satisfatórias para o ajuste da distribuição *GEV*.

Após a obtenção desses resultados calcula-se as estimativas dos parâmetros μ, σ e ξ, isto é, locação, escala e forma respectivamente da distribuição *GEV*, juntamente com suas respectivas variâncias e covariâncias estimadas, obtidos por meio da função de máxima verossimilhança, apresentadas na Tabela 3. Dando sequência as análises, pode-se observar por meio da Tabela 3, que as estimativas pontuais para o parâmetro de forma (ξ) com valores menores do que zero, que correspondem à distribuição de Weibull, ocorrem em todos os meses analisados.

De acordo com Holmes e Moriarty (1999), esta distribuição é a mais adequada para representar fenômenos ambientais, como por exemplo, temperatura máxima mensal, devido

Tabela 2 – Teste de chorrilho sob o pressuposto de independência aplicado aos dados de temperatura máxima do município de Campina Grande.

Meses	p-valor
janeiro	0,0504
fevereiro	0,0640
março	0,4573
abril	0,8550
maio	0,4573
junho	0,4411
julho	0,7001
agosto	0,8550
setembro	0,0570
outubro	0,0005
novembro	0,0740
dezembro	0,0932

ao fato de possuir uma cauda superior com limite finito. Um fato que chama atenção neste trabalho é que não se observou nenhum mês com a estimativa pontual do parâmetro de forma maior que zero, que por sua vez, levaria à distribuição de Fréchet. De acordo com Coles (2001), a distribuição de Fréchet não é adequada para estudar o comportamento de alguns fenômenos ambientais, pois apresenta uma cauda superior com limite infinito. Entretanto vale ressaltar que, este fato só será consumado por meio do teste da razão de verossimilhança modificado T^*_{LR} com seus respectivos intervalos de confiança.

Tabela 3 – Estimativas dos parâmetros da GEV e suas respectivas variâncias e covariâncias estimadas, pelo método de máxima verossimilhança, para os dados de temperatura máxima do município de Campina Grande-PB.

Meses	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	Vâr($\hat{\mu}$)	Vâr($\hat{\sigma}$)	Vâr($\hat{\xi}$)	Côv ($\hat{\mu}, \hat{\sigma}$)	Côv($\hat{\mu}, \hat{\xi}$)	Côv($\hat{\sigma}, \hat{\xi}$)
janeiro	31,9308	0,7710	-0,2045	0,0605	0,0300	0,0113	-0,0093	-0,0095	-0,0114
fevereiro	32,0775	1,1476	-0,5163	0,1369	0,0826	0,0369	0,0065	-0,0342	-0,0385
março	31,4317	1,4268	-0,2893	0,0549	0,0357	0,0477	0,0272	-0,0234	-0,0139
abril	30,6383	0,8788	-0,1763	0,0879	0,0446	0,0213	0,0112	-0,0175	-0,0158
maio	29,0572	1,1122	-0,0724	0,0413	0,0390	0,0795	0,0320	-0,0261	-0,0105
junho	27,4064	1,1094	-0,0996	0,0970	0,0561	0,0404	0,0243	-0,0315	-0,0297
julho	27,0160	0,9241	-0,2092	0,0870	0,0405	0,0090	0,0144	-0,0085	-0,0057
agosto	28,0125	1,0370	-0,2703	0,0397	0,0203	0,0135	-0,0056	-0,0086	-0,0102
setembro	29,3560	1,1779	-0,4315	0,0275	0,0157	0,0342	0,0092	-0,0136	-0,0106
outubro	30,9095	0,8835	-0,2132	0,0188	0,0099	0,0168	-0,0033	-0,0069	-0,0080
novembro	31,4509	0,7453	-0,4294	0,0276	0,0167	0,0195	-0,0057	-0,0092	-0,0119
dezembro	31,7741	0,9276	-0,3467	0,0280	0,0142	0,0155	-0,0031	-0,0078	-0,0091

Levando-se em consideração que o parâmetro de forma (ξ) define qual tipo de distribuição de valores extremos deve-se ajustar aos dados amostrais, apresenta-se por meio da Tabela 4 os valores de estatística da razão de verossimilhança modificada e seus respectivos intervalos de confiança. Analisando-se o intervalo de confiança apresentado na Tabela 4, para o parâmetro de forma, pode-se observar que o valor zero está contido na maioria dos intervalos analisados. Dessa forma pode-se observar que a maioria dos meses em estudo (em 8 dos 12 meses) obteve-se um bom ajuste para a distribuição Gumbel. Este resultado

é corroborado pela estatística da razão de verossimilhança modificada T^*_{LR}, comparando-se o valor que se encontra na Tabela 4, com o valor tabelado de ($\chi^2_{1;0,05} = 3,84$).

Diante do exposto, pode-se concluir que para os meses de fevereiro, setembro, novembro e dezembro os resultados parecem seguir uma distribuição Weibull, observando-se para estes meses que o intervalo de confiança, com 95% significância, contém todos os valores negativos para o parâmetro de forma, contemplando-se a confiabilidade desta conclusão acerca da distribuição ajustada aos dados de temperatura máxima do município de Campina Grande. Medeiros (2011), ajustou a distribuição *GEV* a dados de precipitação máxima mensal no município de Moreilândia, estado de Pernambuco. O referido autor observou também um bom ajuste para a distribuição Gumbel, fato este corroborado por Coles (2001), ao afirmar que a distribuição Gumbel, apesar de apresentar cauda superior com limite infinito, levam a predições de níveis de retorno inferiores aos obtidos quando se utiliza a distribuição de Fréchet. Sendo assim, obtém-se uma considerável representação no ajuste de fenômenos ambientais, como por exemplo, precipitação e temperatura máxima mensal.

Tabela 4 – Intervalo de 95% de confiança para o parâmetro de forma (ξ) e valores da estatística de razão de verossimilhança modificado T^*_{LR}.

Meses	Limites de 95% de confiança para $\hat{\xi}$		T^*_{LR}
	Superior	Inferior	
janeiro	-0,4854	0,0761	1,8137
fevereiro	-0,7732	-0,2594	10,4944
março	-0,7011	0,1225	2,4073
abril	-0,4837	0,1310	1,2288
maio	-0,4299	0,2848	0,1791
junho	-0,4299	0,2848	0,7702
julho	-0,4570	0,0484	2,1389
agosto	-0,5073	0,0332	3,7949
setembro	-0,6955	-0,1675	6,9145
outubro	-0,4393	0,0128	2,7027
novembro	-0,6563	-0,0373	7,5695
dezembro	-0,6563	-0,0373	3,9853

Mediante resultados da Tabela 5 observa-se as informações do teste Kolmogorov-Smirnov ao nível de 5% de significância, encontra-se nesta Tabela 4 as diferenças máximas absolutas observadas entre os valores probabilísticos das funções de distribuição empírica e teórica para cada mês observado, seguido dos níveis descritivos. Conforme o teste, as distribuições ajustam-se bem aos dados de temperatura máxima, tendo em vista que, $D \leq D_{n,\alpha} = 0,20$ para todos os meses analisados. Na Figura 3 observa-se o esboço do teste de Kolmogorov-Sminorv, que visualmente proporciona as mesmas conclusões citadas anteriormente.

Tabela 5 – Resultados do teste de Kolmogorov-Smirnov para verificar a adequabilidade do ajuste da distribuição aos dados de temperatura máxima do município de Campina Grande-PB.

Meses	Diferença máxima absoluta (D)	p-valor
janeiro	0,1491	0,5392
fevereiro	0,1817	0,2939
março	0,0906	0,9664
abril	0,1274	0,7343
maio	0,0820	0,9918
junho	0,0797	0,9943
julho	0,0980	0,9508
agosto	0,1389	0,6307
setembro	0,1412	0,6323
outubro	0,0839	0,9891
novembro	0,1142	0,8410
dezembro	0,1329	0,7060

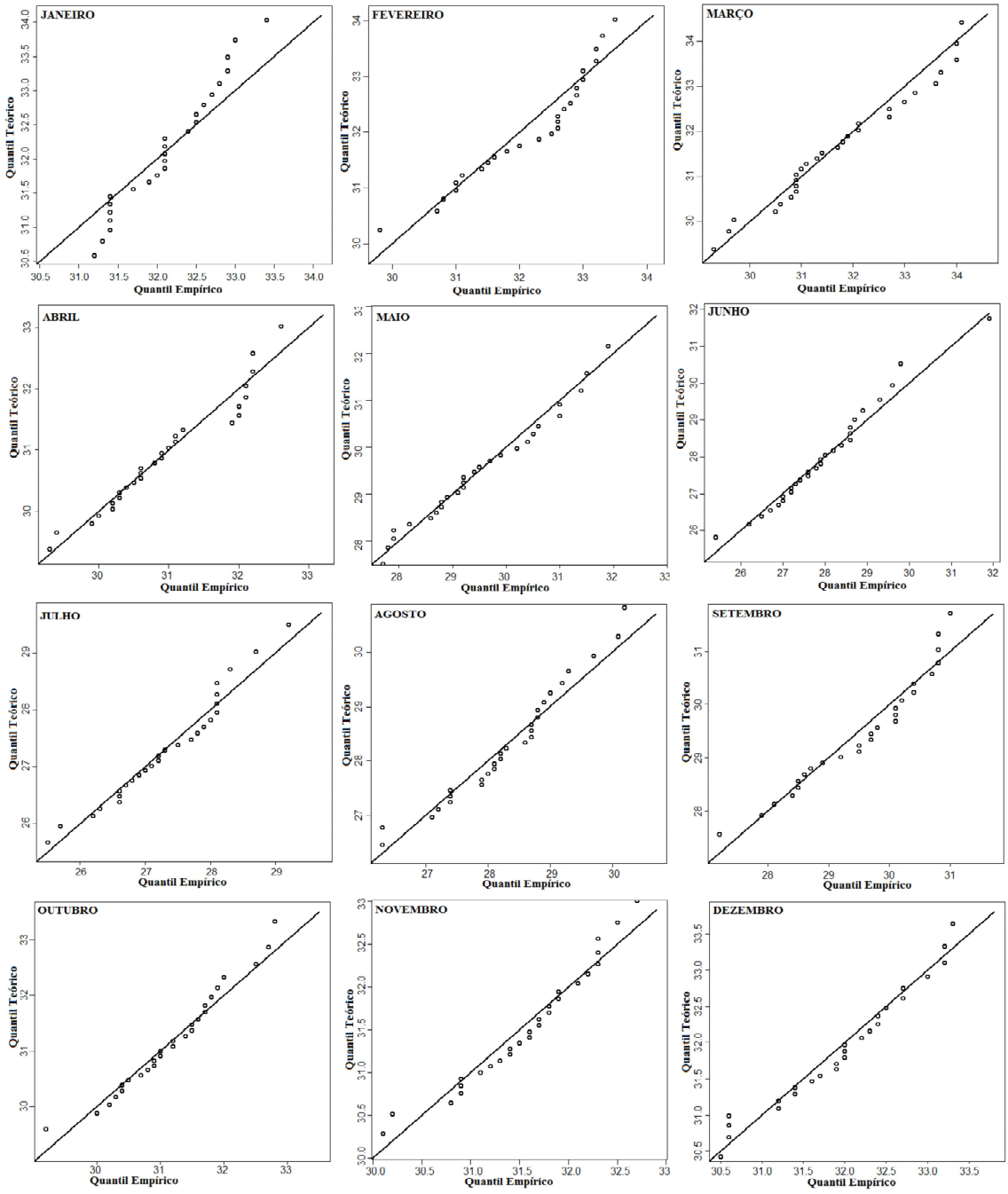
Dando sequência as análises, pode-se observar por meio da Tabela 6 as respectivas probabilidades de ocorrência de temperatura máxima acima de 30, 31 e 32ºC entre os meses de janeiro a dezembro. Observa-se que temperaturas acima de 32ºC os meses de janeiro, fevereiro, setembro, outubro, novembro e dezembro foram bastante expressivas. Para o restante dos meses (março a julho), a probabilidade de ocorrência de temperaturas máximas acima de 32ºC não são tão expressivas, estes resultados eram esperados tendo em vista que de abril a julho é o período mais chuvoso no município de Campina Grande.

Tabela 6 – Probabilidade de ocorrência de temperaturas máximas acima de 30, 31 e 32ºC, para os 12 meses do ano, no município de Campina Grande-PB.

Meses	> 30	> 31	> 32
janeiro	0,65905	0,85684	0,94837
fevereiro	0,13797	0,52082	0,84898
março	0,08952	0,26276	0,51944
abril	0,02759	0,11646	0,34346
maio	0,00052	0,05278	0,40111
junho	0,01308	0,12477	0,46051
julho	0,01631	0,18045	0,66200
agosto	0,07902	0,44709	0,77467
setembro	0,58503	0,88861	0,99967
outubro	0,88684	0,95650	0,98366
novembro	0,93252	0,96479	0,98569
dezembro	0,96911	0,98983	0,99667

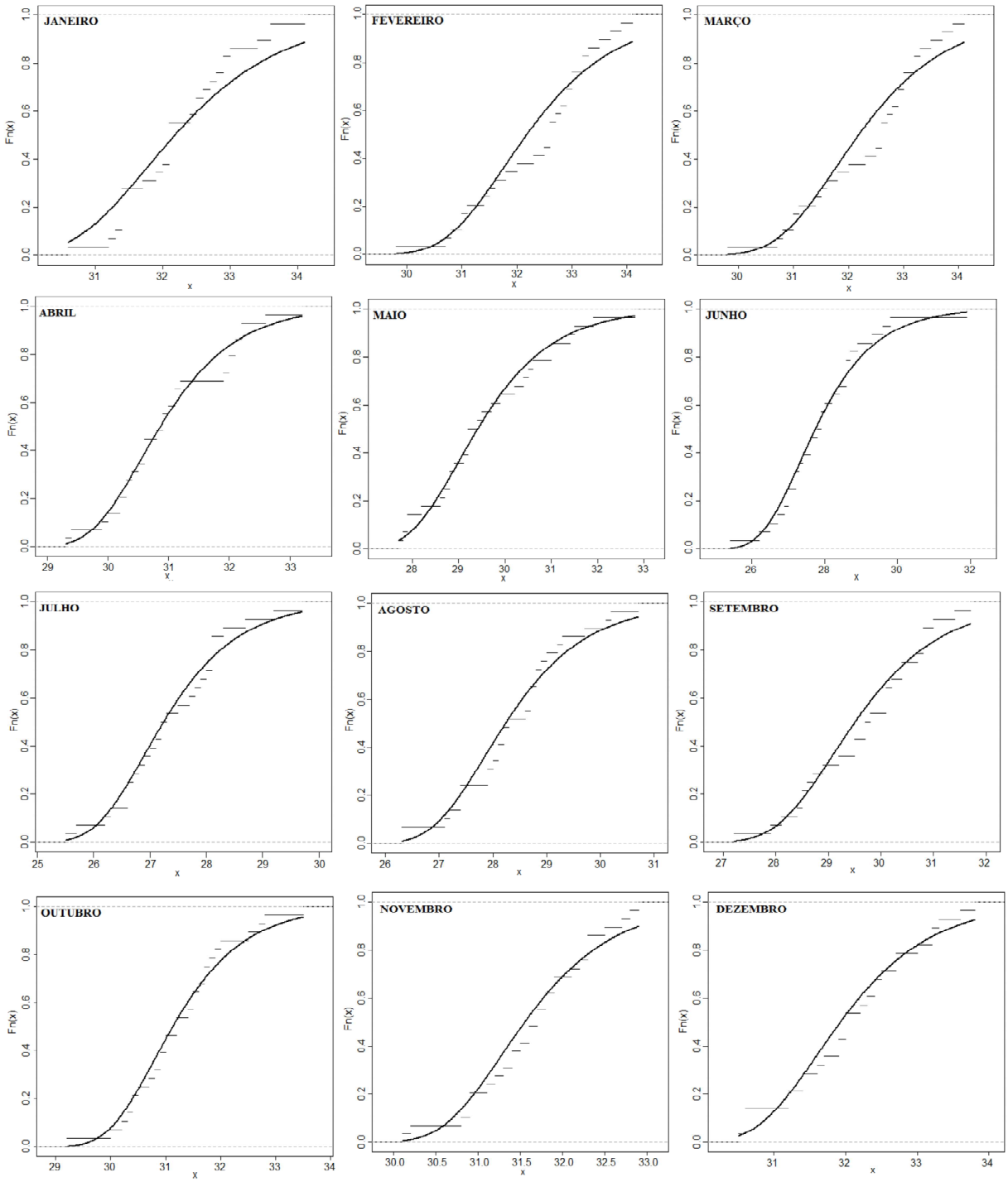
Dando sequência as análises, pode-se observar por meio da Figura 2 os gráficos do quantil empírico versus o quantil teórico para o diagnóstico da distribuição *GEV*. Pode-se observar que no mês de janeiro ocorreu um leve desvio nas caudas da distribuição ajustada.

Figura 2 – Gráficos de quantil-quantil, diagnóstico das distribuições para os dados de temperatura máxima para todos os meses do ano.



O gráfico do quantil empírico versus o quantil teórico compara o ajuste da distribuição teórica com a verdadeira distribuição dos dados amostrais em estudo. De acordo com Sansigolo(2008), tem-se uma maior precisão no ajuste da distribuição *GEV* quando se utiliza uma série histórica igual ou superior a 30 anos, em nosso estudo utilizou-se uma série histórica de 40 anos. Consequentemente, estes ajustes podem ser utilizados com bastante confiabilidade para extrapolação em períodos mais longos, desde que não ocorram futuras variações climáticas significativas nesta região.

Figura 3 – Gráficos do teste de Kolmogorov-Smirnov da função de distribuição acumulada empírica (linhas tracejadas) e teórica (representada pela curva) para diagnóstico dos modelos ajustados aos dados de temperatura máxima mensal.



4 Conclusão

A distribuição generalizada de valores extremos com parâmetro $\xi = 0$, que corresponde à distribuição de valores extremos tipo I e II ou de Gumbel e Weibull é adequada para estudar o comportamento da temperatura máxima nos meses que foram analisados, no município de Campina Grande-PB. Esta distribuição apresentou bom ajuste aos dados, fato comprovado através do gráfico pp-plot. A partir dessas distribuições pode-se determinar diversas aplicações como tempo de retorno e probabilidade de ocorrência para determinado limiar de temperatura máxima.

Um fato que chama atenção neste trabalho é que não se observou nenhum mês com a estimativa pontual do parâmetro de forma maior que zero, que por sua vez, levaria à distribuição de Fréchet, este fato foi confirmado pelo teste da razão de verossimilhança modificado. Vale ressaltar, conforme afirma vários autores, que a distribuição de Fréchet não é adequada para estudar o comportamento de alguns fenômenos ambientais pois apresenta uma cauda superior com limite infinito.

De acordo com os gráficos que comparam o ajuste da distribuição teórica com a verdadeira distribuição dos dados amostrais, pode-se concluir que estes ajustes podem ser utilizados com bastante confiabilidade para extrapolação em períodos de retorno mais longos, desde que não ocorram futuras variações climáticas significativas nesta região.

5 Referências Bibliográficas

BAUTISTA, E.L.B. A distribuição generalizada de valores extremos no estudo da velocidade máxima do vento em Piracicaba - SP. *Dissertação de Mestrado*, ESALQ, USP, Piracicaba estado de São Paulo. 2002. 61p.

BAUTISTA,E. A. L.; ZOCCHI, S. S.; ANGELOCCI, L. R. A Distribuição generalizada de valores extremos aplicada ao ajuste dos dados de velocidade máxima de vento em Piracicaba - SP. *Revista de Matemática e Estatística*, Marília, v.22, p95-111, 2004.

BELTRÃO, B. A. Diagnóstico do município de Morelândia-PE. Programa de desenvolvimento energético dos Estados e Municípios - PRODEEM. *Serviço Geológico do Brasil.* 2005. 150p.

BEZERRA, S. R. de S. *Modelagem estatística de valores extremos aplicados a dados de temperatura máxima em São Gonçalo - PB.* Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Campina Grande, 2013. 35p.

COLES, S. *An Introduction to Statistical Modeling of Extreme Values.* Berlin: Springer. 2001. 208p.

FARIA P. N; OLINDA, R. A.; OZAKI, V. A.; CAMPOS, R. C. *Análise da probabilidade de perda no rendimento da soja: Implicações para o seguro agrícola.* Simpósio Nacional de Probabilidade e Estatística-SINAPE, 2010, São Paulo. Disponível em: http://www.ime.unicamp.br/sinape/19sinape/node/243, 2010.

FERRARI, G. T. Imputação de dados pluviométricos e sua aplicação na modelagem de eventos extremos de seca agrícola. *Dissertação de Mestrado*, ESALQ, Piracicaba estado de São Paulo. 2011. 70p.

FISHER, R. A; TIPPETT, L. H. C. Limiting forms of the frenquency distribution of the largest or smallest member of a sample, *Procs. Cambridge Philos. SOC.* v.24, p.180-190, 1928.

FREIRE, F. R.; BEIJO, L. A. Análise dos métodos de estimação para os parâmetros da distribuição Gumbel na precipitação de chuvas máximas para a cidade de Piracicaba - SP outubro (2010).

HOSKING, J. R. M. Terting whether the shape parameter is zero in the generalized extreme-value distribution. *Biometrika*, v. 71. p.367-374, 1984.

HOLMES, J. D. MORIARTY, W. W. Aplication of the generalized Pareto distribution to extreme value analysis in wind engineering. *Journal of Wind Enginering and Industrial Aerodynamics*, v.83, p.1-10, 1999.

KOTZ, S.; NADARAJAH, S. *Extreme Value Distributions; Theory and Applications*. London: Imperial College Press. 2000. 195p.

MEDEIROS, E. S. *Distribuição generalizada de valores extremos aplicada a dados de precipitação máxima na região de Moreilândia - PE*. Trabalho de Conclusão de Curso de Bacharelado em Estatística, UEPB, Campina Grande-Paraíba. 2011. 41p.

OLINDA, R. A. Modelagem estatística de extremos espaciais com base em processos max-stable aplicados a dados meteorológicos no estado do Paraná. *Tese de Doutorado*, ESALQ, Piracicaba estado de São Paulo. 2012. 163p.

RAO, C.R. TOUTENBURG, H. *Linear models*. 2nd. ed. New York: Springer-Verlag, 1999. 443p.

R Core Team (2017). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

SANSIGOLO, C. A. A distribuição de extremos e precipitação diária, temperatura máxima e mínima e velocidade do vento em Piracicaba - SP. *Revista Brasileira de Meteorologia*, v.23, p.341-346, 2008.

SILVA, R. R. A distribuição generalizada do Pareto e mistura da distribuição de Gumbel no estudo da vazão e da velocidade máxima do vento em Piracicaba - SP. *Dissertação de Mestrado*, ESALQ, Piracicaba estado de São Paulo. 2008. 70p.

SMITH, R. L. Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, London, v. 72, p. 67-92, 1985.

ZAR , J. H. *Biostatistical analysis*. 4.ed. New Jersey: Prentice Hall. 1999. 911p.