Universidade Federal
de São João del-Rei

Romário Pires

ANÁLISE DO DESEMPENHO PREDITIVO DA DISTRIBUIÇÃO GENERALIZADA DE VALOR EXTREMO PARA DADOS DE CONTAGEM: UM ESTUDO DE CASO

São João del-Rei - MG

2018

Romário Pires

ANÁLISE DO DESEMPENHO PREDITIVO DA DISTRIBUIÇÃO GENERALIZADA DE VALOR EXTREMO PARA DADOS DE CONTAGEM: UM ESTUDO DE CASO

Trabalho de Conclusão de Curso apresentado à Coordenadoria do Curso de Matemática, da Universidade Federal de São João del-Rei, como requisito parcial à obtenção do título de Licenciado em Matemática.

Orientador: Prof. Dr. Davi Butturi-Gomes

São João del-Rei, ____ de _____ de _____

Banca Examinadora

_____

Orientador: Prof. Dr. Davi Butturi-Gomes

_____

Prof. Dr. Marcos Santos de Oliveira

_____

Profa. Dra. Rejane Corrêa da Rocha

São João del-Rei - MG

Junho de 2018

Agradecimentos

Agradeço primeiramente a Deus pelo dom da vida, por ter me propiciado caminhos que me levam ao conhecimento e por estar sempre junto comigo me guiando e me livrando dos perigos.

Ao meu orientador, professor Dr. Davi Butturi-Gomes pela prodigiosa orientação, dedicação, preocupação com todo o trabalho e toda paciência e conhecimento partilhado.

Aos professores Dr. Marcos Santos de Oliveira e Dra. Rejane Corrêa da Rocha que aceitaram fazer parte da banca examinadora, a fim de avaliar meu trabalho e, cada vez mais, me preparando para a vida acadêmica.

À Universidade Federal de São João del-Rei pela oportunidade de cursar uma graduação.

Aos meus amigos de república, os quais estiveram sempre ao meu lado, vivendo junto a mim as dificuldades enfrentadas no ambiente acadêmico.

Aos amigos do Grupo de Oração Universitário - GOU que sempre estiveram comigo, estudando e me ajudando a fundamentar a fé que tenho.

Aos meus pais Gilson e Marta e ao meu irmão Rodrigo por todo apoio, incentivo e preocupação durante todo tempo da graduação.

À Letícia, por todo apoio e carinho, por toda ajuda, especialmente nessa fase final, e por tudo que me ensinou quando juntou seus passos aos meus.

Lista de Tabelas

Lista de Figuras

Sumário

Resumo

Análise do desempenho preditivo da distribuição generalizada de valor extremo para dados de contagem: um estudo de caso

A Teoria do Valor Extremo (TVE) é uma área de estudo utilizada na modelagem de eventos extremos. A TVE depende, de forma geral, de uma distribuição generalizada de valores extremos, que inclui três tipos de distribuição assintótica (Weibull, Gumbel e Fréchet). Existem vários métodos para estimar seus parâmetros, sendo o estimador de máxima verossimilhança o mais utilizado. Devido à importância do conhecimento da ocorrência de veranicos para o planejamento de atividades na agricultura, realizou-se, no presente trabalho, um estudo de caso utilizando um banco de dados de veranicos na Estação Chuvosa do Posto Pluviométrico da cidade de Floresta do Estado de Pernambuco, para o qual foi feito ajuste da generalizada do valor extremo a dados de contagem via método da máxima verossimilhança. Para isso, o conjunto de dados máximos foi dividido em dois: a primeira parte utilizada para estimação e a segunda parte para verificar o desempenho preditivo. Além disso, foi realizado o ajuste da distribuição generalizada do valor extremo para o conjunto completo de dados para realizar predições para os 20 anos seguintes da série de dados. Foi utilizado do teste de Ljung-Box para verificar a independência entre as observações e o teste de Kolmogorov-Smirnov para adequabilidade das distribuições, ambos a um nível de 5% de significância. Verificou-se que as distribuições Gumbel e Fréchet ajustaram-se aos dados de veranicos máximos anuais da primeira parte e a distribuição Fréchet ajustou-se a série total de dados de veranicos máximos anuais.

Palavras-chave: Dados de contagem; GEV; Método da Máxima Verossimilhança; Teoria do Valor Extremo.

Abstract

Predictive performance analysis of the generalized extreme value distribution applied to count data: a case study

The Extreme Value Theory is a field of study useful for modeling extreme phenomena. It depends generally on the generalized extreme value (GEV) disitribution, which encompasses three possible asymptotic results (Weibull, Gumbel and Fréchet). Among the several existing methods for estimation, the maximum likelihood method (MLM) is perhaps the mostly used one. Due to the importance of the occurences of dry spells for agricultural practices, the present study focused on modeling the maximum number of days without rain available of a dataset withdrawn from Floresta municipality, located in the State of Pernambuco, Brazil using the MLM for the GEV distribution. For such, the dataset was divided in two parts: the first part was used for model fitting and the second part was used to evaluate its prediction performance. Furtermore, a final model was fitted to the complete data and predictions were made for 20 years in the future of the last year of the dataset. The Ljung-Box test for independency and the Kolmogorov-Smirnov for goodness-of-fit were used at 5% level of significance. It was verified that both Gumbel and Fréchet distributions fitted to the first part of the dataset and that only the Fréchet distribution fitted complete series of maximum duration of anual dry spells.

Keywords: Count data; GEV; Maximum Likelihood Method; Extreme Value Theory.

1 Introdução

A chuva é um fenômeno meteorológico de extrema importância para o meio ambiente e sobrevivência humana. Ela é um dos fatores que mais influenciam no rendimento e sucesso de produção das culturas agrícolas, sobretudo, em áreas não irrigadas. E a comercialização dos produtos agrícolas é uma das fortes características observadas na economia brasileira, que é responsável por grande parte da produção mundial (REGO, PAULA, 2012).

O veranico, que pode ser entendido como um período de interrupção da precipitação durante a estação chuvosa, tem forte influência sobre a produtividade agrícola, principalmente quando coincide com a fase na qual o plantio é mais sensível à deficiência hídrica, isto é, a fase de florescimento. Assim, o conhecimento sobre possíveis riscos, como por exemplo, duração elevada de veranicos em um ano, é fundamental para tomadas de decisões nas atividades agrícolas (CARVALHO et al. 2000).

Pensando que agricultores tenham interesse na previsão da duração máxima de veranicos para um determinado período de tempo, faz-se necessário utilizar de técnicas estatísticas que, por ventura, ainda são pouco exploradas. Uma das formas de modelar máximos de veranicos, dadas n observações, é utilizando as distribuições de valores extremos, que serão introduzidos na subseção que trata da Teoria do Valor Extremo. Assim, este trabalho busca explicitar, a partir de um conjunto de dados discretos, a Teoria do Valor Extremo com sua distribuição generalizada, mostrando sua aplicação em dados veranicos máximos observados na cidade de Floresta (PE).

1.1 Objetivos

O presente trabalho teve por objetivo estudar o desempenho preditivo da distribuição generalizada do valor extremo para dados de extremos discretos, partindo de pontos gerais em teoria da estimação e utilizando um banco de dados de veranicos na Estação Chuvosa do Posto Pluviométrico da cidade de Floresta do Estado de Pernambuco.

1.1.1 Objetivos Específicos

- Realizar o ajuste da distribuição generalizada do valor extremo a dados de contagem;
- Verificar a qualidade do ajuste entre as três formas possíveis (Weibull, Gumbel e Fréchet);

- Comparar as predições teóricas fornecidas pelo modelo selecionado com as observações reais; e

- Realizar predições para os 20 anos seguintes da série de dados (2005 a 2024).

2 Revisão de Literatura

2.1 Teoria do Valor Extremo

A Teoria dos Valores Extremos é uma área da Estatística e Probabilidade que vem sendo comumente utilizada em eventos que ocorrem com pouca frequência, isto é, estudando valores que estão distantes das tendências centrais. O seu estudo é considerado como um dos mais importantes para as ciências aplicadas nos últimos 50 anos (COLES, 2001).

Sob o olhar de Nogueira (2016), a motivação principal para o desenvolvimento inicial se deu quando as barragens que protegem a Holanda do avanço do mar se romperam devido à rajadas de ventos e ondas gigantes, causando a inundação de boa parte do país, provocando a morte de dezenas de milhares de animais e de mais de 1800 pessoas, além do desabrigamento de outras 72 mil. Após o desastre, o governo da Holanda criou um comitê para estudar o aquecimento global e suas consequências, que utilizava o ferramental ligado à Teoria dos Valores Extremos.

Há indícios que já no século XVIII, em 1709, Nicolas Bernoulli estudou problemas envolvendo Valores Extremos (GUMBEL, 1958 apud NOGUEIRA, 2016). Contudo, a utilização da Teoria do Valor Extremo teve impulso por volta de 1920 a partir do trabalho de Fisher e Tippett (1928), que descreveram três tipos de distribuições assintóticas dos valores extremos. Porém, o primeiro a estudar e normalizar a aplicação destas distribuições foi Gumbel (ALMEIDA, 2018).

Segundo Coles (2001), os extremos dos bancos de dados podem ser coletados de duas maneiras diferentes, a saber, selecionando os máximos (ou mínimos) de cada período de uma série de dados, que é denotado como máximos (ou mínimos) em blocos; ou pontos que ultrapassam um limite pré-fixado (alto ou baixo), que é chamado de excessos sobre um limiar. Neste trabalho, foram considerados apenas os primeiros.

Seja uma sequência de variáveis aleatórias $X_1, X_2, \ldots, X_n$. Define-se M_n o máximo do bloco como sendo:

$$M_n = \max\{X_1, X_2, \ldots, X_n\}.$$

Além disso, se a amostra é independente e identicamente distribuída (iid), é possível encontrar a função de distribuição dos máximos (ou mínimos),

$$\begin{aligned} P(M_n \leq z) &= P(X_1 \leq z, \ldots, X_n \leq z) \\ &= P(X_1 \leq z) \times \cdots \times P(X_n \leq z) \\ &= [F(z)]^n, \end{aligned}$$

para $z \in \mathbb{R}, n \in \mathbb{N}$.

Porém, esse resultado não é imediatamente útil, visto que a distribuição permanece incógnita e, além disso, trata-se de uma função distribuição degenerada[1]. Contudo, isto pode ser contornado se utiliza uma renormalização linear da variável M_n, isto é:

$$M_n^* = \frac{M_n - b_n}{a_n},$$

para sequências de constantes apropriadas (PORTES, 2017, p. 24). Toda esta ideia foi mostrada por Fisher e Tippet (1928) e posteriormente por Gnedenko (1943) através de um teorema - O teorema Fisher-Tippet-Gnedenko (ou Teorema do Valor Extremo).

O Teorema Fisher-Tippet-Gnedenko diz que

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \xrightarrow[n \to \infty]{\mathcal{D}} G(z)$$

e, caso $G(\cdot)$ seja não degenerada, então a convergência se dará para uma das distribuições de valores extremos: Gumbel (I), ou para Weibull (II) ou para Fréchet (III),

$$\begin{aligned} \text{(I) } G(z) &= \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\}, \quad -\infty < z < \infty \\ \text{(II) } G(z) &= \exp\left\{-\left(\frac{z-\mu}{\sigma}\right)^{-\xi}\right\}, \qquad z > \mu \\ \text{(III) } G(z) &= \exp\left\{-\left[\left(\frac{z-\mu}{\sigma}\right)^{\xi}\right]\right\}, \qquad z < \mu. \end{aligned}$$

As três funções de distribuições possuem em comum dois parâmetros, sendo eles o de posição (μ) e o de escala (σ). Porém, as distribuições de Fréchet e Weibull apresentam o parâmetro de forma (ξ), que ajusta a cauda da distribuição (PORTES, 2017).

Estudando os três diferentes tipos de comportamento que as distribuições I, II e III caracterizam, Jenkinson (1955) apud Nogueira (2016) propôs uma distribuição generalizada que abrange essas três possíveis famílias de distribuições - a distribuição Generalizada de Valores Extremos (GVE). Essa distribuição possui função de distribuição acumulada que pode ser expressa como:

[1]Uma distribuição degenerada é definida como uma distribuição de probabilidade de uma variável aleatória cujo suporte consiste em somente um valor.

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

em que ξ, σ, μ são, respectivamente, os parâmetros de forma, de escala e de posição, com $\sigma \in \mathbb{R}_+$ e ξ e $\mu \in \mathbb{R}$ (NOGUEIRA, 2016).

As três funções de distribuições de valores extremos diferenciam-se no parâmetro de forma (ξ), pois ele que determina a forma assintótica de valores extremos, isto é, quando $\xi > 0$, a GVE representa a distribuição de Fréchet; se $\xi < 0$ a GVE representa a distribuição Weibull; e para o caso em que o limite de $G(z)$, com ξ tendendo a zero, a GVE corresponde à distribuição Gumbel (NOGUEIRA, 2016).

2.2 Aplicações práticas da Teoria do Valor Extremo

Dentre as aplicações práticas da teoria de valores extremos, pode ser citada a análise e previsão meteorológica; temperatura esperada em uma dada região (REIS; BEIJO; AVELAR, 2017); precipitações máximas (REIS et al., 2017); modelagem das ondas do oceano (DAWSON, 2000 apud PORTES, 2017); falha na memória celular (MCNULTY et al., 2000 apud PORTES, 2017); engenharia eólica (HARRIS, 2001 apud PORTES, 2017); estratégias administrativas (DAHAN; MENDOLSON, 2001 apud PORTES, 2017); processamentos de dados biológicos (ROBERTS, 2000 apud PORTES, 2017); avaliação da mudança climática; ciências Atuariais para cálculo de riscos que resultam em grandes perdas econômicas para as seguradoras, resseguradoras, instituições financeiras ou entidades de previdência, entre elas: risco financeiro nas bolsas de valores (ARRAES; ROCHA, 2006 apud PORTES, 2017) e perda máxima operacional (ERGASHEV et al., 2013 apud PORTES, 2017).

Reforçando alguns exemplos citados, os elevados regimes de precipitação podem ocasionar graves efeitos, como enchentes, deslizamentos de terras, atrasos em colheitas, entre outros. E, devido à intensidade desses fenômenos, setores importantes no Brasil, como o energético e o agrícola, têm sofrido grandes impactos ocasionados por eventos climáticos extremos (REIS; BEIJO; AVELAR, 2017). Nesse sentido, a Teoria do Valor Extremo foca em estudar e analisar o comportamento e a ocorrência dos máximos nessas ocasiões.

Levando-se em consideração o tamanho do território brasileiro, é possível observar diferentes regimes pluviométricos. De acordo com Menezes, Brito e Lima (2010), de norte a sul no país se encontra uma grande variedade de climas com distintas características

regionais.

As regiões Sudeste e Centro-Oeste sofrem influência tanto de sistemas tropicais quanto de latitudes médias, com estação seca bem definida no inverno (de Abril a Setembro) e estação chuvosa no verão (de Outubro a Março) (MENEZES; BRITO; LIMA, 2010).

Em algumas regiões no Brasil, principalmente nos cerrados, a precipitação total do período chuvoso é suficiente para o desenvolvimento da agricultura, porém é comum a ocorrência de sequência de dias secos durante a estação chuvosa. A esses dias secos durante a estação chuvosa, dá-se o nome de "veranicos". Todavia, na região sul mineira, veranico possui dois nomes populares, que são "varanico" ou "varanique" (informação pessoal).

Os veranicos, dependendo da duração e da época, podem afetar de forma acentuada o desenvolvimento das plantações e, consequentemente, a produtividade final (SOUZA; PERES, 1998). O desenvolvimento do semiárido do Nordeste do Brasil é dependente da precipitação pluviométrica e, em consequência, suas variações provocam prejuízos econômicos e sociais. A Paraíba, por exemplo, tem como características climáticas marcantes, as irregularidades, tanto espacial quanto temporal, do seu regime de chuvas (MENEZES; BRITO; LIMA, 2010).

2.3 Estimação de Parâmetros

O conhecimento de θ fornece uma descrição completa do comportamento probabilístico sobre uma determinada população descrita por uma função densidade de probabilidade (fdp) ou função de probabilidade $f(x|\theta)$. Em geral, não é possível ter acesso à toda a população, então retiram-se amostras e, a partir delas, procura-se aproximar os valores dos parâmetros. Neste contexto, a Teoria da Estimação se ocupa de buscar métodos para encontrar um bom estimador pontual de θ.

Definição 2.3.1. *Um estimador pontual é uma função da amostra, ou seja, uma estatística especial que contém informação sobre o parâmetro de interesse.*

Definição 2.3.2. *Estimativa é o valor observado de um estimador que é obtido quando uma amostra é efetivamente colhida.*

Exemplo 2.3.1. *No segundo bimestre de um ano letivo, Romário alcançou as seguintes notas: 10,0 em Matemática, 7,3 em Português, 7,0 em História, 8,5 em Geografia, 9,2 em*

Inglês, 8,4 em Espanhol, 9,0 em Física, 8,2 em Química, 8,0 em Biologia e 9,4 em Educação Física. Logo, a População é $X = \{10,0,\ 7,3,\ 7,0,\ 8,5,\ 9,2,\ 8,4,\ 9,0,\ 8,2,\ 8,0,\ 9,4\}$. *Então, é sabido que a média das notas obtidas por Romário naquele bimestre é* $\mu = 8,5$. *Suponha que seja desconhecido alguns valores das notas obtidas por Romário e se tem acesso a apenas às notas de Geografia, Física e Biologia. Então, para a amostra* $x = \{8,5,\ 9,4,\ 8,2\}$, *a média amostral é* $\hat{\mu} = 8,5$. *O que pode ser observado no exemplo é que a média da amostra é um estimador da média desconhecida da população (e uma boa estimativa, ao fato que* $\mu = \bar{X}$*) e o valor encontrado para a média é a estimativa. Além disso, pode ser notado que foi encontrado um valor único, isto é, uma estimativa pontual.*

Existem diversos métodos para encontrar estimadores de parâmetros. Neste trabalho foram estudados dois deles, o Método dos Momentos e o Método da Máxima Verossimilhança.

2.3.1 Método dos Momentos

O Método dos Momentos, segundo Casella e Berger (2010), é o método mais antigo para encontrar estimadores pontuais, datado no final do século XIX. Ele é bem simples de executar e trás um objetivo claro e fácil de compreensão.

Definição 2.3.3. *Seja* $X_1, X_2, \ldots, X_n$ *uma amostra de uma população com função de probabilidade* $f(x|\theta_1, \theta_2, \ldots, \theta_k)$ *com* θ_j, $j = 1, \ldots, k$, *sendo os parâmetros. Estimadores pelo Método dos Momentos são encontrados igualando-se os primeiros momentos amostrais aos momentos da população correspondentes a* k *e resolvendo o sistema resultante de equações simultâneas.*

Mais precisamente, é definido

$$\begin{cases} m_1 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^1, & \mu_1 = EX^1, \\ m_2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2, & \mu_2 = EX^2, \\ \quad \vdots & \\ m_k = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^k, & \mu_k = EX^k. \end{cases}$$

O momento da população μ_j geralmente será uma função de $\theta_1, \theta_2, \ldots, \theta_k$, isto é, $\mu_j(\theta_1, \theta_2, \ldots, \theta_k)$. O método de momentos é obtido resolvendo-se o seguinte sistema de equações para $\theta_1, \theta_2, \ldots, \theta_k$ em termos de $m_1, m_2, \ldots, m_k$:

$$\begin{cases} m_1 = \mu_1(\theta_1, \theta_2, \ldots, \theta_k), \\ m_2 = \mu_2(\theta_1, \theta_2, \ldots, \theta_k), \\ \quad \vdots \\ m_k = \mu_k(\theta_1, \theta_2, \ldots, \theta_k). \end{cases}$$

Exemplo 2.3.2. *Sejam $X_1, X_2, \ldots, X_n$ binomiais (k, p) independente e identicamente (iid), isto é, $P(X_i = x_i|k, p) = \binom{k}{x_i} p^{x_i}(1-p)^{k-x_i}$; $x_i = 0, 1, \ldots, k$. Assumindo que k e p são desconhecidos, devem ser obtidos estimadores $\hat{p}$ e $\hat{k}$ para p e k, respectivamente.*

Solução: Por definição, o método dos momentos propõe igualar os t-ésimos momentos populacionais ($\mu_t = EX^t$) aos respectivos t-ésimos momentos amostrais ($m_t = \frac{1}{n}\sum X_i^t$). Dessa maneira,

$$\begin{cases} m_1 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^1 \\ m_2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2 \end{cases}$$

e

$$\begin{cases} \mu_1 = EX = kp \\ \mu_2 = EX^2 = kp(1-p) + (kp)^2. \end{cases}$$

Igualando-se os dois primeiros momentos populacionais aos dois primeiros momentos amostrais, resulta-se:

$$\begin{cases} \hat{k}\hat{p} = \dfrac{1}{n}\sum X_i \\ \hat{k}\hat{p}(1-\hat{p}) + (\hat{k}\hat{p})^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2 \end{cases} \quad \Rightarrow \quad \begin{cases} \hat{k} = \dfrac{\bar{X}^2}{\bar{X} - \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \bar{X}^2)} \\ \hat{p} = \dfrac{\bar{X}}{\hat{k}}. \end{cases}$$

Estimando, então, de forma genérica, valores para $\hat{k}$ e $\hat{p}$ para a distribuição Binomial. ■

2.3.2 Estimadores de Máxima Verossimilhança

O Método de Máxima Verossimilhança constitui-se em adotar, como estimadores de parâmetros, as quantidades que maximizam a função de probabilidade de a amostra observada ter sido obtida. Para obter estimadores por esse método, é necessário conhecer a distribuição da variável em estudo. Segundo Casella e Berger (2010), esse método é o mais amplamente utilizado para obter estimadores.

Definição 2.3.4. *Para cada ponto amostral x, seja $\hat{\theta}(x)$ um valor de parâmetro no qual $L(\theta|\boldsymbol{x})$ atinge seu máximo como uma função de θ, com x mantido fixo. Um Estimador de Máxima Verossimilhança do parâmetro θ com base em uma amostra X é $\hat{\theta}(X)$.*

Como a Estimativa pela Máxima Verossimilhança visa maximizar a função para estimar parâmetros, se ela for diferenciável, ela pode ser executada por meio do cálculo diferencial, isto é, derivando a função e igualando a zero.

Os zeros da primeira derivada localizam apenas os extremos no interior da função trabalhada. Contudo, se os extremos ocorrerem no limite, a derivada primeira poderá não ser zero. Se o caso ocorrer, o limite deverá ser verificado separadamente para os extremos. Além disso, os pontos que zeram a primeira derivada podem ser máximos globais ou locais, ou mínimos globais ou locais. Nos casos regulares, a primeira derivada da função de verossimilhança fornece os pontos de máximo global.

Exemplo 2.3.3. *Seja um tetraedro[2] regular com suas faces pintadas de branco ou de azul, porém desconhecidos o número de faces de cada cor. Ao lançar o tetraedro, o resultado é considerado "sucesso" se a face que ficar em contato com a mesa for azul. Imagine que foram feitos cinco experimentos com o tetraedro e, em cada cada experimento, ele foi lançado quatro vezes ao ar. A pessoa que o lançou relatou que os cinco experimentos ocorreram com a seguinte ordem de números de faces azuis registradas: primeiro experimento não houve nenhuma face azul; no segundo experimento houve uma face azul; no terceiro, uma face azul; no quarto, duas faces azuis; e no quinto, nenhuma face azul. Com base nessas informações, obtenha a estimativa de máxima verossimilhança da probabilidade de sucesso em um lançamento qualquer desse tetraedro.*

Solução: Sabendo que o problema é descrito pela função de probabilidade de uma variável aleatória Binomial e que os eventos são todos iid, tem-se a seguinte função de

[2]Poliedro regular composto por quatro faces triangulares.

verossimilhança:

$$L(p|\boldsymbol{x}) = \prod_{i=1}^{5} \binom{4}{x_i} p^{x_i}(1-p)^{4-x_i} = \prod_{i=1}^{5} \binom{4}{x_i} \prod_{i=1}^{5} p^{x_i} \prod_{i=1}^{5} (1-p)^{4-x_i}.$$

Derivar esse produtório é uma tarefa muito difícil. Porém, pode-se usar de manipulações matemáticas e propriedades dos logaritmos naturais, que resultam em uma função fácil de diferenciação. Como

$$\begin{aligned} L(p|\boldsymbol{x}) &= \prod_{i=1}^{5} \binom{4}{x_i} \prod_{i=1}^{5} p^{x_i} \prod_{i=1}^{5} (1-p)^{4-x_i} \\ &= \left[\prod_{i=1}^{5} \binom{4}{x_i}\right] \left(p^{\sum_{i=1}^{5} x_i}\right) \left[(1-p)^{\sum_{i=1}^{5}(4-x_i)}\right], \end{aligned}$$

logo,

$$\begin{aligned} L^* = \ln[L(p|\boldsymbol{x})] &= \ln\left[\left(\prod_{i=1}^{5} \binom{4}{x_i}\right) \left(p^{\sum_{i=1}^{5} x_i}\right) \left((1-p)^{\sum_{i=1}^{5}(4-x_i)}\right)\right] \\ &= \ln(p^4) + \ln(1-p)^{16} + \ln\left[\prod_{i=1}^{5} \binom{4}{x_i}\right] \\ &= \ln(p^4) + 16\ln(1-p) + \ln\left[\prod_{i=1}^{5} \binom{4}{x_i}\right]. \end{aligned}$$

Então,

$$\frac{dL^*}{dp} = \frac{4\hat{p}^3}{\hat{p}^4} + \frac{16(-1)}{1-\hat{p}} + 0 = \frac{4}{\hat{p}} - \frac{16}{1-\hat{p}}.$$

Fazendo a derivada igual a zero, tem se:

$$\frac{dL^*}{dp} = 0 \Rightarrow \frac{4}{\hat{p}} - \frac{16}{1-\hat{p}} = 0 \Rightarrow \hat{p} = \frac{1}{5}.$$

Quando a derivada da função de probabilidade é igual a zero, $\hat{p}$ assume valor $\frac{1}{5}$. Essa é, então, a estimativa de máxima verossimilhança para a probabilidade de obter um sucesso (face azul). ■

Como observado, foi encontrado o parâmetro dado a amostra. É possível, por esse método, encontrar o estimador para o parâmetro p de um modo geral, isto é, como se fosse uma fórmula para se aplicar em qualquer amostra de uma população descrita pela função de distribuição Binomial.

Exemplo 2.3.4. *Sejam $X_1, X_2, \ldots, X_n$ variáveis aleatórias iid com distribuição Binomial (m,p). Dado que a função de verossimilhança é*

$$L(p|\boldsymbol{x}) = \prod_{i=1}^{n} \binom{m}{x_i} p^{x_i}(1-p)^{m-x_i},$$

encontre o estimador $\hat{p}$.

Solução: Assim como no caso anterior, a função é difícil de diferenciar. Então, será usado novamente das propriedades de logaritmo natural da verossimilhança. Como

$$L(p|\boldsymbol{x}) = \prod_{i=1}^{n} \binom{m}{x_i} p^{x_i}(1-p)^{m-x_i} = \prod_{i=1}^{n} \binom{m}{x_i} \prod_{i=1}^{n} p^{x_i} \prod_{i=1}^{n} (1-p)^{m-x_i},$$

logo,

$$\begin{aligned} L^* = \ln[L(p|\boldsymbol{x})] &= \ln\left[\left(\prod_{i=1}^{n}\binom{m}{x_i}\right)\left(p^{\sum_{i=1}^{n} x_i}\right)\left((1-p)^{\sum_{i=1}^{n}(m-x_i)}\right)\right] \\ &= \ln\prod_{i=1}^{n}\binom{m}{x_i} + \ln(p^{\sum_{i=1}^{n} x_i}) + \ln\left[(1-p)^{\sum_{i=1}^{n}(m-x_i)}\right] \\ &= \sum_{i=1}^{n} x_i \ln(p) + \sum_{i=1}^{n}(m-x_i)\ln(1-p) + \ln\prod_{i=1}^{n}\binom{m}{x_i}. \end{aligned}$$

Fazendo, então, a derivada com relação a p, tem-se:

$$\begin{aligned} \frac{dL^*}{dp} &= \frac{\sum_{i=1}^{n} x_i}{p} + \frac{\sum_{i=1}^{n}(m-x_i)}{-(1-p)} \\ &= \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\sum_{i=1}^{n}(m-x_i)}{(1-p)}. \end{aligned}$$

Igualando a derivada a 0, resulta-se:

$$\begin{aligned} \frac{\sum_{i=1}^{n} x_i}{\hat{p}} - \frac{\sum_{i=1}^{n}(m-x_i)}{(1-\hat{p})} &= 0 \\ \frac{\sum_{i=1}^{n} x_i}{\hat{p}} &= \frac{\sum_{i=1}^{n}(m-x_i)}{(1-\hat{p})} \\ (1-\hat{p})\sum_{i=1}^{n} x_i &= \hat{p}\sum_{i=1}^{n}(m-x_i) \\ \sum_{i=1}^{n} x_i - \hat{p}\sum_{i=1}^{n} x_i &= \hat{p}\sum_{i=1}^{n}(m-x_i) \\ \sum_{i=1}^{n} x_i &= \hat{p}\left[\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}(m-x_i)\right]. \end{aligned}$$

Logo,

$$\begin{aligned}\hat{p} &= \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}(m - x_i)} \\ &= \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} m} \\ &= \frac{\sum_{i=1}^{n} x_i}{nm} \\ &= \frac{\bar{x}}{m}.\end{aligned}$$

■

Então, encontrar o valor do parâmetro p pelo método de máxima verossimilhança de uma amostra descrita pela função de distribuição binomial, pode ser usada dessa generalização encontrada. Veja a verificação do Exemplo 2.3.3: $\hat{p} = \frac{\bar{X}}{m} = \frac{0,8}{4} = 0,2 = \frac{1}{5}$.

Se tratando de funções de probabilidade que possuem mais de um parâmetro, para estimar os seus parâmetros, dado uma amostra $X(X_1, X_2, \ldots, X_n)$, o processo é parecido, distinguindo apenas em derivar a função parcialmente e igualando a zero as derivadas parciais, obtendo um sistema de equação.

Assim, para este trabalho, é de interesse obter, genericamente, os estimadores dos parâmetros μ, σ e ξ da distribuição GVE via método de máxima verossimilhança.

2.4 Estimação dos Parâmetros da GVE via Máxima Verossimilhança

A GVE possui a seguinte função de distribuição:

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} 1_{Z(z)},$$

em que $\sigma > 0$ e $-\infty < \mu, \xi, z < +\infty$. Para obter a função densidade de probabilidade desta distribuição, basta fazer a primeira derivada em relação a z.

$$g(z) = \frac{1}{\sigma}\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\left(\frac{1+\xi}{\xi}\right)} \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\},$$

definida em $-\infty < z < +\infty$.

A Figura 1 apresenta os gráficos da função densidade de probabilidade da GVE para $\xi = -0,5$ (Weibull - linha verde), $\xi \longrightarrow 0$ (Gumbel - linha vermelha) e $\xi = 0,5$ (Fréchet - linha azul), com $\mu = 0$ e $\sigma = 1$, dos quais pode observar que o parâmetro de forma determina a natureza das caudas da distribuição. O ξ da Weibull gera uma cauda inferior e o ξ da Gumbel e Fréchet gera uma cauda superior. Contudo, a cauda da Gumbel é mais leve em relação a Fréchet.

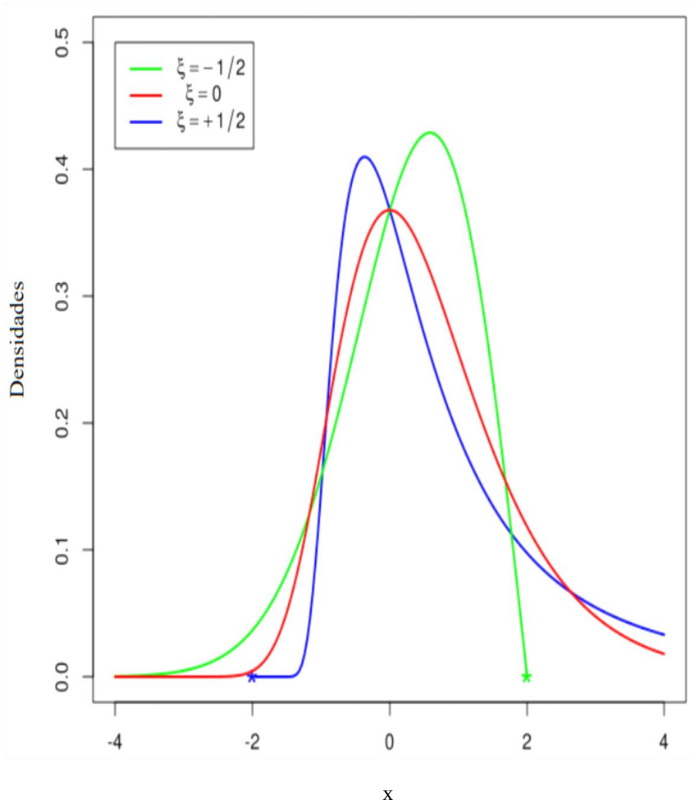

Figura 1: Função densidade de probabilidade da distribuição GVE para $\xi = -0,5$ (Weibull), $\xi \longrightarrow 0$ (Gumbel) e $\xi = 0,5$ (Fréchet), com $\mu = 0$ e $\sigma = 1$. Fonte: Adaptado de Wikipedia (2018).

Seja $X_1,\ X_2, \ldots,\ X_n$ uma sequência independente de máximos (ou mínimos) observados em uma amostra. Para estimar os parâmetros da função de probabilidade de GVE, deve-se derivar o logaritmo da função de verossimilhança em relação a μ, σ e ξ e igualar a zero. Fazendo isso, encontra-se o seguinte sistema de derivadas parciais:

$$\begin{cases} \dfrac{\partial L^*}{\partial \mu} = \dfrac{1}{\hat{\sigma}} \displaystyle\sum_{i=1}^{n} \left(\dfrac{1 + \hat{\xi} - w_i^{-\frac{1}{\hat{\xi}}}}{\hat{\sigma}} \right) = 0 \\ \dfrac{\partial L^*}{\partial \sigma} = -\dfrac{n}{\hat{\sigma}} + \dfrac{1}{\hat{\sigma}^2} \displaystyle\sum_{i=1}^{n} \left\{ \dfrac{(x_i - \hat{\mu}) \left[(1 + \hat{\xi}) - w_i^{-\frac{1}{\hat{\xi}}} \right]}{w_i} \right\} = 0 \\ \dfrac{\partial L^*}{\partial \xi} = \displaystyle\sum_{i=1}^{n} \left\{ \left(1 - w_i^{-\frac{1}{\hat{\sigma}}} \right) \left[\dfrac{1}{\hat{\xi}^2} \ln(w_i) - \dfrac{(x_i - \hat{\mu})}{\hat{\xi}\hat{\sigma} w_i} \right] - \dfrac{(x_i - \hat{\mu})}{\hat{\xi} w_i} \right\} = 0 \end{cases},$$

em que $w_i = 1 + \hat{\xi} \left(\dfrac{X_i - \hat{\mu}}{\hat{\sigma}} \right)$ e $\hat{\mu}$, $\hat{\sigma}$ e $\hat{\xi}$ são estimadores de máxima verossimilhança dos parâmetros da GVE. Entretanto, esse método para estimar ξ trás consigo alguns problemas:

(i) Se $-1 < \xi < -0,5$ ou $0,5 < \xi < 1$, o estimador de Máxima Verossimilhança não satisfaz as condições de regularidade, isto é, normalidade assintótica, consistência e eficiência; e

(ii) Se $\xi < -1$ ou $\xi > 1$, o estimador de Máxima Verossimilhança não existe.

Como o sistema de equações não possui solução analítica, pode-se usar os métodos iterativos para obtenção da solução numérica do sistema. No presente trabalho foi utilizado o método quase newton para a solução do mesmo.

2.5 Predição pela Distribuição Generalizada do Valor Extremo

Uma vez que foram obtidas as estimativas de máxima verossimilhança $\hat{\mu}, \hat{\sigma}$ e $\hat{\xi}$, pode-se buscar formas de realizar predições. Na Teoria do Valor Extremo tem-se dois métodos preditivos - por Período de Retorno e por Nível de Retorno.

2.5.1 Período de Retorno

Segundo Almeida (2018), pode-se definir o período de retorno como sendo o valor médio da variável T de tempo aleatório estimado entre ocorrências consecutivas de um fenômeno natural (A).

Exemplo 2.5.1. *Seja um evento qualquer como, por exemplo, um veranico de 45 dias. Suponha que ele seja igualado ou excedido em média a cada 50 anos. Então, espera-se*

que ele tenha um período de retorno $T = 50$ anos, isto é, que ocorra o mesmo veranico a cada 50 anos.

No presente estudo, o fenômeno natural é: "Veranico máximo excede um determinado valor z". Matematicamente, o período de retorno é o inverso da probabilidade de igualdade ou excedência, isto é

$$T = \frac{1}{P(A)} = \frac{1}{1 - F(z)},$$

sendo $F(z)$ no presente trabalho a função de distribuição GVE.

2.5.2 Nível de Retorno

De acordo com Bautista (2004) apud Nogueira (2016), o nível de retorno, denotado z_p, está associado ao período de retorno por T, e é obtido a partir da solução da equação

$$\int_{-\infty}^{z_p} f(z|\theta)dz = 1 - p$$

para $p = \frac{1}{T}$, isto é,

$$F(z_p|\theta) = (1 - p).$$

Contudo, o interesse está em fazer com que z_p dependa de p para um valor escolhido *a priori* de p. Para tal, deve-se aplicar a função inversa, de modo a se obter

$$z_p = F^{-1}(1 - p|\theta).$$

Quando $F(\cdot)$ é a função de distribuição da GVE, com um pouco de álgebra, pode-se demonstrar que

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

Em geral, os valores de μ, σ e ξ são desconhecidos e substituímos pelas estimativas de máxima verossimilhança. Assim,

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

3 Material e Métodos

3.1 Conjunto de Dados

Os dados analisados neste trabalho foram obtidos em Assis (2012), com uma adaptação reproduzida no Anexo A. Trata-se de uma série de durações de veranicos (em número de dias) na cidade de Floresta (PE) organizados anualmente no período de 1964 até 2004 (Figura 2). Para aplicar a Teoria do Valor Extremo, foram extraídas as durações máximas em cada ano, obtendo um conjunto de dados, e métodos de inferência estatística e computacional foram aplicados.

O conjunto total de dados é composto por 41 elementos, ou seja, os blocos de máximo em cada ano, e estão descritos a seguir: $X = \{16, 21, 31, 16, 33, 25, 45, 29, 21, 34, 61, 21, 120, 38, 38, 31, 19, 43, 30, 22, 51, 11, 19, 33, 26, 33, 34, 31, 25, 59, 16, 120, 121, 120, 30, 29, 18, 47, 21, 29, 44\}$.

Figura 2: Região da Cidade de Floresta - Pernambuco. Fonte: Wikipédia (2018).

3.2 Métodos de Inferência Estatística e Computacional

Primeiramente, o conjunto de dados foi dividido em dois: a primeira parte, que é formada por 21 elementos, foi utilizada para a estimação, e as 20 observações remanescentes foram utilizadas para verificar o desempenho preditivo.

Uma vez que a Teoria do Valor Extremo requer independência entre as observações,

foi aplicado o teste de Ljung-Box (LJUNG; BOX, 1978) a 5% de significância para as 21 observações considerando as hipóteses:

$$\begin{cases} H_0 : \text{A série de veranicos máximos anuais é independente.} \\ H_1 : \text{A série de veranicos máximos anuais não é independente.} \end{cases}$$

Na sequência, foi ajustada a GVE pelo método da máxima verossimilhança (COLES, 2001) e foram obtidos: (*i*) o viés de predição; (*ii*) a acurácia preditiva (inclusão ou não do máximo verdadeiro no intervalo de predição a 95% de confiança); (*iii*) a precisão preditiva (comprimento do intervalo de predição).

O terceiro passo foi realizar o ajuste da distribuição generalizada do valor extremo ao total de dados de contagem e realizar predições para os anos os 20 anos seguintes da série de dados (2005 a 2024).

3.2.1 Teste de Kolmogorov-Smirnov

De acordo com Zar (1998), para avaliar o bondade de ajuste de um modelo probabilístico a um conjunto de dados, para um dado nível de significância, deve-se utilizar de testes de aderência. Exemplos de testes de aderência são o teste de Qui-quadrado e o teste de Kolmogorov-Smirnov, sendo este último o utilizado neste trabalho.

O teste de Kolmogorov-Smirnov é um método não-paramétrico, que fornece apenas a conclusão de que a distribuição é adequada ou não. Caso a distribuição não seja boa, deve-se procurar o ajuste para outra distribuição. Para a realização do teste devem ser seguidos os seguintes passos:

(*i*) Colocar a sequência de dados em ordem crescente;

(*ii*) Obter os valores de probabilidade da distribuição teórica $F(z_{(i)})$ e os valores de probabilidade da distribuição empírica $\hat{F}(z_{(i)})$;

(*iii*) Calcular a estatística D através da seguinte expressão:

$$D = \sup|F(z_{(i)}) - \hat{F}(z_{(i)})|, i = 1, 2, \ldots, n;$$

(*iv*) A hipótese de que a distribuição empírica $\hat{F}(X)$ é suficientemente próxima de $F(X)$ teórica pode ser testada pela comparação do resultado da estatística D com um valor crítico para um dado nível de significância ou pela comparação do valor de p do teste com o nível de significância adotado.

Com isso, o teste de Kolmogorov-Smirnov ao nível de 5% de significância foi realizado após a modelagem, com o objetivo de verificar se a distribuição GVE se ajustou bem à série de veranicos máximos anuais da cidade de Floresta (PE), para as hipóteses:

$$\begin{cases} H_0 : \text{A série de veranicos máximos anuais segue a distribuição GVE.} \\ H_1 : \text{A série de veranicos máximos anuais não segue a distribuição GVE.} \end{cases}$$

Todas as análises estatísticas foram conduzidas em plataforma R (R CORE TEAM, 2017) e a rotina utilizada foi disponibilizada no Apêndice I. O teste de Ljung-Box foi realizado conforme o disponibilizado no pacote 'tseries' (TRAPLETTI, HORNIK, 2018) e as estimativas de Máxima Verossimilhança foram realizadas com auxílio do pacote 'evd' (STEPHENSON, 2002).

4 Resultados e Discussão

4.1 Avaliação do Desempenho Preditivo

Na Figura 3 pode-se observar a sequência das 21 durações máximas anuais dos veranicos (série de ajuste) no município de Floresta (PE), para a qual o resultado do teste de Ljung-Box indicou independência ($p = 0,5698$).

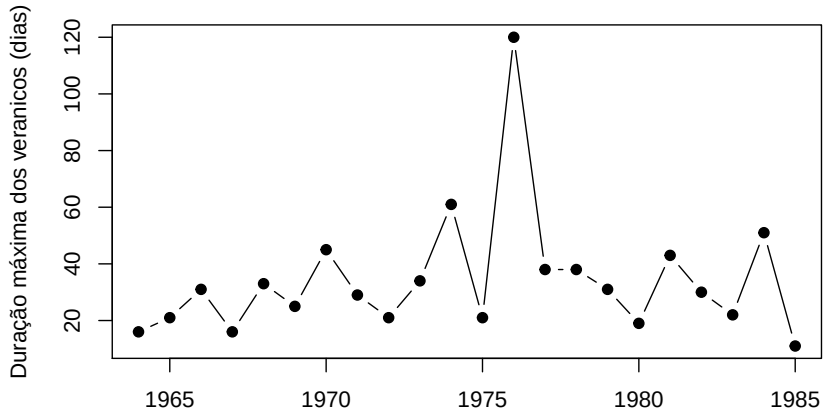

Figura 3: Máximos anuais de veranicos.

Com base nesse resultado, foi possível ajustar a distribuição GVE utilizando o método da máxima verossimilhança para estimação dos parâmetros. Na Tabela 1 é apresentado o resultado obtido da estimativa dos parâmetros da GVE.

Tabela 1: Estimativas pontuais e intervalares (IC 95% de confiança) dos parâmetros da distribuição GVE via método da máxima verossimilhança.

Parâmetro	Estimativa	IC 95%
μ	25,1374	[20,3993 ; 29,8755]
σ	9,3539	[5,1571 ; 13,5505]
ξ	0,3838	[-0,0795 ; 0,8472]

Observando o parâmetro ξ, pode-se concluir que a distribuição Gumbel é a que se ajusta na série de veranicos, visto que o zero está incluído no intervalo de confiança. Além disso, $\hat{\xi} = 0,3838$, o que significa que o estimador de máxima verossimilhança satisfaz as condições de regularidade.

O teste de Kolmogov-Smirnov indicou que $p = 0,8385$, concluindo que o teste não rejeitou H_0. Isso significa que a distribuição Gumbel via máxima verossimilhança ajustou-se a série de veranicos máximos.

Na Tabela 2 estão os valores máximos reais observados presentes no banco de dados. Os valores observados são acumulados, ou seja, o máximo de cada período dado (em anos). Estão presentes também os resultados das predições teóricas fornecidas pelo modelo, níveis de retorno com seus intervalos de predição com a respectiva amplitude e o viés (valor observado subtraído do valor predito). O nível de retorno necessita ser calculado para um tempo de pelo menos dois anos, uma vez que é necessário calcular os máximos acumulados até o período. A partir da tabela foi possível fazer a análise da acurácia e precisão dos níveis de retorno.

Tabela 2: Viés, intervalos de predição (IP) e respectivas amplitudes para os níveis de retorno com os períodos das observações reais, a partir da distribuição Gumbel.

Período	Valor Observado	Valor Predito	Viés	IP 95%	Amplitude
2	19	29	10	[26 ; 38]	12
3	33	34	1	[**31** ; **46**]	15
4	33	37	4	[34 ; 51]	17
5	33	39	6	[36 ; 55]	19
6	34	41	7	[38 ; 58]	20
7	34	43	9	[39 ; 61]	22
8	34	44	10	[40 ; 63]	23
9	59	45	-14	[**41** ; **65**]	24
10	59	46	-13	[**42** ; **67**]	25
11	120	47	-73	[43 ; 68]	25
12	121	48	-73	[44 ; 70]	26
13	121	49	-72	[44 ; 71]	27
14	121	49	-72	[45 ; 72]	27
15	121	50	-71	[45 ; 73]	28
16	121	51	-70	[46 ; 74]	28
17	121	51	-70	[46 ; 75]	29
18	121	52	-69	[47 ; 76]	29
19	121	52	-69	[47 ; 77]	30
20	121	53	-68	[48 ; 78]	30

Analisando os resultados apresentados da Tabela 2 foi possível observar que a distribuição Gumbel englobou, em seus intervalos de predição, apenas três de dezenove valores observados, para 3, 9 e 10 anos. Isso significa que a distribuição teve aproximadamente 16% de acerto (acurácia preditiva). Pode-se perceber que não houve um bom desempenho preditivo, que deve estar associado à pequena quantidade de dados e ao amplo tempo de predição. Como a GVE é uma distribuição assintótica, usar de 21 elementos para estimar e predizer os niveis de retorno pode trazer muitos erros, assim como ocorreu neste caso.

Além disso, é possível averiguar que os intervalos de confiança dos níveis de retorno calculados estão subestimando os valores reais a partir de um período de nove anos e que existem vários casos onde os valores observados não foram englobados pelos intervalos de predição. Conclui-se que isto pode trazer consigo um prejuízo ao produtor.

Os casos em que os valores observados foram menores do que o predito pelo ajuste (viés positivo) exigem do produtor gastos a mais do que o necessário no momento de sua preparação e planejamento na suas atividades agrícolas. Por outro lado, os casos em que os valores observados foram maiores que o predito pelo ajuste (viés negativo), implicam, por exemplo, que o produtor deixe de ter uma melhor preparação ou planejamento das suas atividades. Assim, o viés negativo trás consigo um prejuízo maior para o produtor do que o viés positivo, uma vez que planejar as atividades agrícolas para um veranico muito maior do que o ocorrido provoca, nos piores cenários, que ele compre produtos desnecessários. Já o viés negativo pode fazer com que o produtor planeje suas atividades para um determinado limite de veranico, sendo que se o veranico máximo for muito superior do que o que está preparado, pode comprometer todo o seu lucro e produção.

Considerando que a estimativa pontual para o parâmetro de forma é relativamente distante de zero, isto é, um intervalo de menor confiança facilmente não incluiria o zero, então pode ser de interesse avaliar o desempenho preditivo da distribuição Fréchet. Assim, foram realizadas as predições teóricas e verificadas a acurácia e precisão preditivas para essa distribuição, utilizando as estimativas dos parâmetros via método de máxima verossimilhança da Tabela 1. O teste de Kolmogov-Smirnov indicou que $p = 0,8989$, concluindo que novamente o teste não rejeitou H_0. Isso significa que a distribuição Fréchet também ajustou-se a série de veranicos máximos. Estes resultados podem ser visualizados na Na Tabela 3.

Analisando os resultados apresentados na Tabela 3, foi possível observar que a distribuição Fréchet englobou, em seus intervalos de predição, dez de dezenove valores observados. Isso significa que a distribuição teve aproximadamente 53% de acerto (acurácia preditiva). Assim como no caso anterior, foi possível verificar que os valores níveis de retorno preditos subestimam os valores reais a partir do período de nove anos, com exceção do período de dez anos, visto que a predição foi exata. Além disso, pode-se perceber que não houve um bom desempenho preditivo, que deve estar também associado à pequena quantidade de dados e ao amplo tempo de predição. Também foi concluído que isto pode

Tabela 3: Viés, intervalos de predição (IP) e respectivas amplitudes para os níveis de retorno com os períodos das observações reais, a partir da distribuição Fréchet.

Período	Valor Observado	Valor Predito	Viés	IP 95%	Amplitude
2	19	29	10	[23 ; 35]	12
3	33	35	2	[**27** ; **44**]	17
4	33	40	7	[**29** ; **51**]	22
5	33	44	11	[**31** ; **57**]	26
6	34	48	14	[**32** ; **63**]	31
7	34	51	17	[**33** ; **69**]	36
8	34	54	20	[**33** ; **74**]	41
9	59	56	-3	[**33** ; **79**]	46
10	59	59	0	[**33** ; **84**]	51
11	120	61	-59	[33 ; 89]	56
12	121	63	-60	[33 ; 93]	60
13	121	65	-58	[33 ; 97]	64
14	121	67	-56	[24 ; 120]	96
15	121	69	-52	[32 ; 106]	74
16	121	71	-50	[31 ; 111]	80
17	121	72	-49	[31 ; 114]	83
18	121	74	-47	[30 ; 118]	88
19	121	75	-46	[**29** ; **123**]	94
20	121	77	-44	[**29** ; **126**]	97

trazer consigo um prejuízo para o agricultor pela mesma razão do caso anterior.

Comparando os resultados das Tabelas 2 e 3, pode-se concluir que a distribuição Fréchet apresentou mais acurácia para os níveis de retorno do que a distribuição Gumbel. Em contrapartida, a distribuição Gumbel apresentou menores amplitudes do que a distribuição Fréchet, o que significa maior precisão.

4.2 Ajuste da Série Completa

Na Figura 4 pode-se observar a sequência total durações máximas dos veranicos em Floresta (PE), de 1964 a 2004, cujo resultado do teste de Ljung-Box indicou que a série é independente ($p = 0,05697$).

Com base nesse resultado do teste, foi possível ajustar a distribuição GVE utilizando o método da máxima verossimilhança para estimação dos parâmetros. Na Tabela 4 é apresentado o resultado obtido da estimativa dos parâmetros da GVE.

Observando o parâmetro ξ, pode-se concluir que a distribuição Fréchet é a que se ajusta na série total de veranicos, visto que o intervalo de confiança está todo no lado positivo.

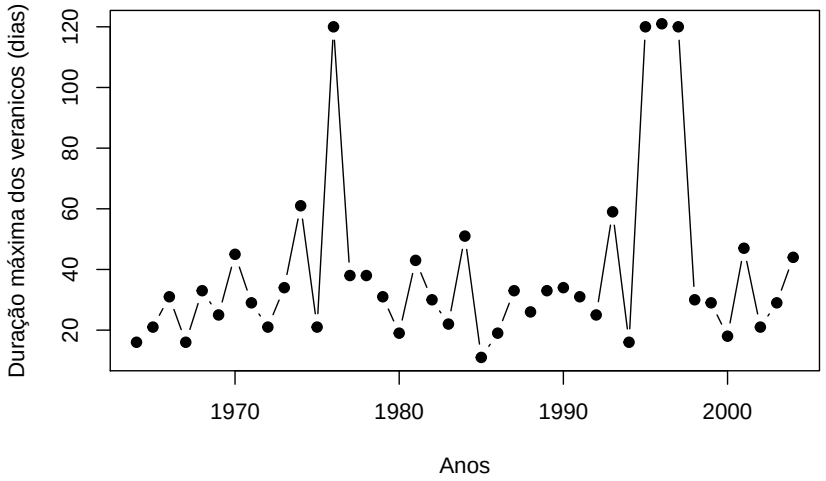

Figura 4: Máximos anuais de veranicos de 1964 a 2004.

Tabela 4: Estimativas dos parâmetros da distribuição GVE via método da máxima verossimilhança.

	Parâmetro	Estimativa	I.C. 95%
GVE	μ	25,6739	[21,6219 ; 29.7257]
	σ	11,8053	[8,2078 ; 15,4027]
	ξ	0,3828	[0,1231 ; 0,6424]

Além disso, $\hat{\xi} = 0,3828$, o que significa que o estimador de máxima verossimilhança satisfaz as condições de regularidade.

O teste de Kolmogov-Smirnov indicou que $p = 0,8129$, concluindo que o teste não rejeitou H_0. Isso significa que a distribuição GVE via máxima verossimilhança ajustou-se a série de veranicos máximos.

Na Tabela 5 estão os resultados do valor máximo predito de cada período (em anos) durante 20 anos (2005 a 2024), dos níveis de retorno com seus intervalos de predição com a respectiva amplitude.

Como exemplo, escolhendo-se um período de 3 anos, uma interpretação dos resultados apresentados na Tabela 5 pode ser feita da seguinte maneira: em um período de três anos após o fim das observações fornecidas pelo modelo (2007), espera-se que ocorra um veranico máximo em torno de 38 dias, podendo variar entre 31 e 46 dias. Se tratando do intervalo desse exemplo e do contexto inserido, pode-se imaginar que o produtor planejará suas atividades para suportar um veranico de 46 dias.

Para o planejamento das atividades, o intervalo de predição é fundamental. Contudo, quando o intervalo é muito amplo, o agricultor sentirá dificuldade e não dará o devido crédito para o modelo, visto que quanto maior o intervalo, maior a variação do veranico.

Tabela 5: Valor predito, intervalo de predição (IP) e respectivas amplitudes para os níveis de retorno, a partir da distribuição Fréchet.

Período	Valor Predito	IP 95%	Amplitude
2	30	[25 ; 35]	10
3	38	[31 ; 46]	15
4	45	[35 ; 54]	19
5	50	[38 ; 61]	23
6	54	[40 ; 68]	28
7	58	[42 ; 74]	32
8	61	[44 ; 80]	36
9	65	[45 ; 85]	40
10	68	[46 ; 90]	44
11	71	[47 ; 95]	48
12	73	[48 ; 99]	51
13	76	[49 ; 103]	54
14	78	[49 ; 108]	59
15	81	[50 ; 112]	62
16	83	[50 ; 117]	67
17	85	[50 ; 120]	70
18	87	[51 ; 124]	73
19	89	[51 ; 129]	78
20	91	[51 ; 132]	81

Desse modo, um intervalo de 15 dias, assim como no exemplo acima, pode ser considerado útil.

5 Conclusão

De acordo com os resultados obtidos, pode-se concluir que distribuição Gumbel ajustou-se à série de dados dos 21 veranicos máximos observados de Floresta (PE). Porém, levando em consideração que a estimativa pontual para o parâmetro de forma foi relativamente distante de zero, foi feito o teste novamente e as respectivas predições teóricas via distribuição Fréchet, verificou-se que ela também ajustou-se à serie de dados de 21 veranicos máximos. Comparando-as, pode-se perceber que a acurácia para distribuição Fréchet levou a um menor erro, enquanto a distribuição Gumbel apresentou intervalos de predição mais precisos, isto é, mais curtos.

Concluiu-se que não houve um bom desempenho preditivo das distribuições Gumbel e Fréchet, visto a Gumbel apresentou aproximadamente 16% de acurácia e a Fréchet apresentou aproximadamente 53% de acurácia preditiva. Como mencionado ao longo do trabalho, isto deve estar associado à pequena quantidade de dados e ao amplo tempo de predição, visto que a GVE é uma distribuição assintótica, portanto, utilizar dados com poucos elementos para estimar e predizer os niveis de retorno pode trazer muitos erros.

Se tratando do ajuste da série completa, e de acordo com os resultados obtidos, pode-se concluir que a distribuição Fréchet ajustou à série. Sendo assim, as predições obtidas no trabalho, mesmo que grande parte esteja desatualizada, pode ser utilizado por agricultores que tenham interesse em planejar as atividades agrícolas na região de Floresta (PE).

Referências

ALMEIDA, C. G. **Uma Abordagem Bayesiana para a modelagem dos ventos máximos de Sorocaba-SP e Bauru-SP.** 2018. 77 f. Dissertação (Mestrado) apresentada ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, Universidade Federal de Alfenas, Alfenas-MG, 2018.

ASSIS, J. M. O. **Análise de tendências de mudanças climáticas no semiárido de Pernambuco**. 166 p. Dissertação (Mestrado em Desenvolvimento e Meio Ambiente) – Centro de Filosofia e Ciências Humanas, Universidade Federal de Pernambuco, Recife-PE, 2012.

CARVALHO, D. F.; FARIA, R. A.; SOUSA, S. A. V.; BORGES, H. Q. Espacialização do período de veranico para diferentes níveis de perda de produção na cultura do milho, na bacia do rio verde grande, mg. **Revista Brasileira de Engenharia Agrícola e Ambiental**, Campina Grande, v. 4, n. 2, p. 172-176, 2000.

CASELLA, G.; BERGER, R. L. **Inferência Estatística.** 2. ed. São Paulo: Cengage Learning, 2010, 281-288p.

COLES, S. **An introduction to statistical modelling of extreme values**. London: Springer, 2001. 208p.

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, Cambridge**, v. 24, n. 2, p. 180-190, 1928.

GNEDENKO, B. Sur la distribution limite du terme maximum d'une serie aleatoire. **Annals of Mathematics**, Lawrenceville, v. 44, n. 3, p. 423-453, 1943.

LJUNG, G. M.; BOX, G. E. On a measure of lack of fit in time series models. **Biometrika**, Oxford University Press, v. 65, n. 2, p. 297-303, 1978.

MENEZES, H. E. A.; BRITO, J. I. B.; LIMA, R. A. F. A. Veranico e produção agrícola no Estado da Paraíba, brasil. **Revista Brasileira de Engenharia Agrícola e Ambiental**,

Campina Grande, v. 14, n. 2, p. 181-186, 2010.

NOGUEIRA, R. S. **Precisão e acurácia dos estimadores de máxima verossimilhança dos parâmetros da distribuição Gumbel não estacionária**. 99 p. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Instituto de Ciências Exatas, Universidade Federal de Alfenas, Alfenas-MG, 2016.

PORTES, P. C. **Modelagem Bayesiana dos níveis máximos do índice de preços ao consumidor**. 2017. 81 f. Dissertação (Mestrado) apresentada ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, Universidade Federal de Alfenas, Alfenas-MG, 2017.

R CORE TEAM. **R: A language and environment for statistical computing**. Vienna, R Foundation for Statistical Computing, 2017.

REGO, B. R.; PAULA, F. O. O mercado futuro e a comercialização do café: Influências, Riscos e Estratégias com o uso de *Hedge*. **Revista Gestão & Conhecimento**, Poços de Caldas, v. 7, n.11, mar./jun., 2012. Disponível em: <https://www.pucpcaldas.br/graduacao/administracao/revista/artigos/v7n1/v7n1a1.pdf>. Acesso em: 16 de maio de 2018.

REIS, J. R.; BEIJO, B. L.; AVELAR, F. G. Temperatura esperada para Piracicaba-SP via distribuições de valores extremos. **Revista Brasileira de Agricultura Irrigada**, Fortaleza, v. 11, n. 4, p. 1639-1650, 2017

SOUSA, S. A. V.; PERES, F. C. Programa computacional para simulação da Ocorrência de veranicos e queda de rendimento. **Pesquisa Agropecuária Brasileira**, Brasília, v. 33, n. 12, p. 1951-1956, 1998.

STEPHENSON, A. G. evd: Extreme Value Distributions. Vienna, **R News**, v. 2, n. 2, p. 31-32, 2002.

TRAPLETTI, A.; HORNIK, K. tseries: Time Series Analysis and Computational Finance. **R package** v. 0, p. 10-45, 2018.

WIKIPEDIA. (Wikipedia, the free encyclopedia). Generalized extreme value distribution.

Disponível em: <https://en.wikipedia.org/w/index.php?title=Generalized_extreme-_value_distribution>. Acesso em: 06 de agosto de 2018.

WIKIPÉDIA. (Wikipédia, a enciclopédia livre). Floresta (Pernambuco). Disponível em: <https://pt.wikipedia.org/w/in-dex.php?title=Floresta_(Pernambuco)>. Acesso em: 05 de junho de 2018.

ZAR, J. H. **Biostatistical Analysis.** 4. ed. United States of America: Prentice Hall, 1998. 916 p.

Apêndices

Apêndice A: Rotina no R utilizada para calcular as estimativas dos parâmetros e níveis de retorno via máxima verossimilhança.

```
rm(list=ls())
require(evd)
require(tseries)

#########################
###### Criar vetor dos máximos anuais de veranicos.
dados_veranicos_max<-as.numeric(c(16, 21, 31, 16, 33, 25, 45, 29, 21, 34, 61, 21,
                                  120, 38, 38, 31, 19, 43, 30, 22, 51, 11, 19, 33,
                                  26, 33, 34, 31, 25, 59, 16, 120, 121, 120, 30,
                                  29, 18, 47, 21, 29, 44))


### Separando os dados para ajuste e para predição.
dados_veranicos_max.1 <- dados_veranicos_max[1:21]
dados_veranicos_max.2 <- dados_veranicos_max[22:41]

###FAZENDO COM AS 21 OBSERVAÇÕES REAIS.
### Independencia da série.
Box.test(dados_veranicos_max.1, type = "Ljung-Box")

###### Encontra as estimativas e intervalos de confiança dos parâmetros da
###### distribuição GVE via MMV para.
modelo.pre<-fgev(dados_veranicos_max.1, corr=T)
modelo.pre
mu.pre  <-modelo.pre$estimate[1]
sig.pre <-modelo.pre$estimate[2]
xi.pre  <-modelo.pre$estimate[3]

confint(modelo.pre) # inclui o zero?

### Predição com a Gumbel.
round( qgev( (1/2:20), loc=mu.pre, scale=sig.pre, shape=0, lower.tail = F ))
round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/2)  ) )[1,] )


round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/3)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/4)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/5)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/6)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/7)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/8)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/9)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/10)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/11)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/12)  ) )[1,] )
```

```
round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/13)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/14)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/15)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/16)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/17)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/18)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/19)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, shape=0, prob=(1/20)  ) )[1,] )



# Predição com a Fréchet.
round( qgev( (1/2:20), loc=mu.pre, scale=sig.pre, shape=xi.pre, lower.tail = F ))
round( confint( fgev(dados_veranicos_max.1, prob=(1/2)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/3)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/4)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/5)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/6)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/7)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/8)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/8)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/9)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/10)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/11)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/12)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/13)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/14)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/15)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/16)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/17)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/18)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/19)  ) )[1,] )

round( confint( fgev(dados_veranicos_max.1, prob=(1/20)  ) )[1,] )
```

```
###PARA A SÉRIE TOTAL
###### Criar vetor dos máximos anuais de veranicos.
dados_veranicos_max<-as.numeric(c(16, 21, 31, 16, 33, 25, 45, 29, 21, 34, 61, 21,
                                  120, 38, 38, 31, 19, 43, 30, 22, 51, 11, 19, 33,
                                  26, 33, 34, 31, 25, 59, 16, 120, 121, 120, 30,
                                  29, 18, 47, 21, 29, 44))

##### Gera o gráfico dos máximos anuais de veranicos
par(mar=c(4.5,4.5,1,1))
plot(1964:2004,dados_veranicos_max, type = "b",pch=19,
     xlab='Anos',
     ylab='Duração máxima dos veranicos (dias)')

### Independencia da série.
Box.test(dados_veranicos_max, type = "Ljung-Box")

###### Encontra as estimativas e intervalos de confiança dos parâmetros da
###### distribuição GVE via MMV para.
modelo1<-fgev(dados_veranicos_max, corr=T)
modelo1
mu  <-modelo1$estimate[1]
sig <-modelo1$estimate[2]
xi  <-modelo1$estimate[3]

confint(modelo1) # inclui o zero?

round( qgev( (1/2:20), loc=mu, scale=sig, shape=xi, lower.tail = F ))
round( confint( fgev(dados_veranicos_max, prob=(1/2)  ) )[1,] )

round( confint( fgev(dados_veranicos_max, prob=(1/3)  ) )[1,] )

round( confint( fgev(dados_veranicos_max, prob=(1/4)  ) )[1,] )

round( confint( fgev(dados_veranicos_max, prob=(1/5)  ) ) [1,])

round( confint( fgev(dados_veranicos_max, prob=(1/6)  ) )[1,] )

round( confint( fgev(dados_veranicos_max, prob=(1/7)  ) )[1,] )

round( confint( fgev(dados_veranicos_max, prob=(1/8)  ) )[1,] )

round( confint( fgev(dados_veranicos_max, prob=(1/9)  ) )[1,] )

round( confint( fgev(dados_veranicos_max, prob=(1/10)  ) )[1,] )

round( confint( fgev(dados_veranicos_max, prob=(1/11)  ) )[1,] )

round( confint( fgev(dados_veranicos_max, prob=(1/12)  ) )[1,] )

round( confint( fgev(dados_veranicos_max, prob=(1/13)  ) )[1,] )

round( confint( fgev(dados_veranicos_max, prob=(1/14)  ) )[1,] )

round( confint( fgev(dados_veranicos_max, prob=(1/15)  ) )[1,] )

round( confint( fgev(dados_veranicos_max, prob=(1/16)  ) )[1,] )

round( confint( fgev(dados_veranicos_max, prob=(1/17)  ) )[1,] )

round( confint( fgev(dados_veranicos_max, prob=(1/18)  ) )[1,] )

round( confint( fgev(dados_veranicos_max, prob=(1/19)  ) )[1,] )

round( confint( fgev(dados_veranicos_max, prob=(1/20)  ) )[1,] )
```

Anexos

Anexo A: Veranicos na Estação Chuvosa do Posto Pluviométrico da Floresta (PE). Modificado de Assis (2012).

INSTITUTO DE TECNOLOGIA DE PERNAMBUCO

LABORATÓRIO DE METEREOLOGIA

			DURAÇÃO DO VERANICO									
ANO	VER. MÁX.	DATA MÁX.	1	2	3	4	5	6	7	8	9	10
1964	16	29/mar	6	11	8	5	9	6	16			
1965	21	24/jan	17	21	17	5	10	9	6			
1966	31	06/mar	9	18	5	13	31	16				
1967	16	01/jan	16	16	12	16	6	9	8			
1968	33	01/jan	33	16	11	23	5	11				
1969	25	19/mar	16	14	11	10	5	25	17			
1970	45	17/mar	18	14	17	8	45					
1971	29	01/jan	29	9	6	17	8	5				
1972	21	22/jan	5	11	21	5	8	6	5	16	12	12
1973	34	29/jan	6	5	10	34	6	5	6	5	9	
1974	61	01/mar	10	12	13	61						
1975	21	13/fev	6	13	5	21	7	7	14	10		
1976	120	01/jan	120									
1977	38	24/mar	5	9	26	19	38					
1978	36	01/jan	36	17	29	19						
1979	31	27/jan	9	8	31	11	15	23				
1980	19	12/abr	8	9	8	10	14	17	19			
1981	43	30/jan	6	12	5	43	5	14	11			
1982	30	23/jan	14	6	30	8	8	7	25	6		
1983	22	01/jan	22	6	7	14	5	7	9	13		
1984	51	09/jan	7	51	5	8	5	7	9			
1985	11	15/abr	5	6	8	9	11					
1986	19	11/jan	6	19	6	11	7	7	5	7	6	6
1987	33	08/jan	33	9	10	11	5	6	6	15		
1988	26	08/fev	8	16	8	26	7	12	5			
1989	33	11/fev	10	12	13	33	5	8	5	6		
1990	34	07/jan	5	34	6	7	14	8	18			
1991	31	21/fev	18	10	10	31	7	16	12			
1992	25	01/jan	25	12	11	7	23					
1993	59	02/fev	5	10	12	59	22					
1994	16	24/jan	5	6	16	5	11	8	13	10	10	
1995	120	01/jan	120									
1996	121	01/jan	121									
1997	120	01/jan	120									
1998	30	01/abr	9	7	27	9	9	5	6	30		
1999	29	28/jan	26	29	14	23	18					
2000	18	29/jan	12	8	18	5	11	16	11	9		
2001	47	15/mar	32	33	47							
2002	21	24/mar	5	9	5	11	19	13	21	16		
2003	29	02/abr	19	12	6	18	7	29				
2004	44	18/mar	11	17	14	5	44					