**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

A distribuição generalizada de Pareto e mistura de distribuições de Gumbel no estudo da vazão e da velocidade máxima do vento em Piracicaba, SP

Renato Rodrigues Silva

Dissertação apresentada para obtenção do título de Mestre em Agronomia. Área de concentração: Estatística e Experimentação Agronômica

**Piracicaba
2008**

Renato Rodrigues Silva
Engenheiro Agrônomo

A distribuição generalizada de Pareto e mistura de distribuições de Gumbel no estudo da vazão e da velocidade máxima do vento em Piracicaba, SP

Orientador:
Prof. Dr. **SILVIO SANDOVAL ZOCCHI**

Dissertação apresentada para obtenção do título de Mestre em Agronomia. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2008

Dados Internacionais de Catalogação na Publicação (CIP)
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - ESALQ/USP

Silva, Renato Rodrigues

A distribuição generalizada de Pareto e mistura de distribuições de Gumbel no estudo da vazão e da velocidade máxima do vento em Piracicaba, SP / Renato Rodrigues Silva. - - Piracicaba, 2008.

100 p.

Dissertação (Mestrado) - - Escola Superior de Agricultura Luiz de Queiroz, 2008.
Bibliografia.

1. Distribuições (Probabilidade) 2. Estatística aplicada I. Título

CDD 551.55
S586d

“Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor”

Agradecimentos

À Deus.

À minha família Derci Rodrigues Brito Silva, Osmar Marinho Silva e Rafael Rodrigues Silva pelo carinho e apoio.

Ao meu orientador Prof. Dr. Silvio Sandoval Zocchi pela orientação, amizade e ensinamentos.

Ao professor Prof. Dr. Luiz Roberto Angelocci pelos ensinamentos na área de meteorologia.

Ao professor Prof. Dr. Victor Hugo Lachos d'Avila por ter fornecido os artigos necessários para o estudo do algoritmo ECM.

À todos os professores do Curso de Estatística e Experimentação Agronômica da ESALQ-USP pela transmissão dos conhecimentos.

Ao pessoal da empresa Bayes Forecast Ltda. por ter incentivado a elaboração deste trabalho.

Aos meus amigos Fernando, Guarda e a meus colegas de mestrado Erick, Simone, Luziane, Renata, Elton, Edjane, Mirian, Eduardo, Marina, Jalmar, Juliana, enfim a todos os colegas de mestrado e doutorado do Curso de Estatística e Experimentação Agronômica pela amizade e bom convívio.

SUMÁRIO

RESUMO	5
ABSTRACT	6
1 INTRODUÇÃO	7
Referências	10
2 AJUSTE DA DISTRIBUIÇÃO GENERALIZADA DE PARETO PARA ANÁLISE DE DADOS DE VAZÃO MÉDIA DIÁRIA EM PIRACICABA, SP, BRASIL.	14
Resumo	14
Abstract	14
2.1 Introdução	15
2.2 Desenvolvimento	15
2.3 Resultados e Discussão	26
2.4 Conclusões	33
Referências	34
3 MISTURAS DE DUAS DISTRIBUIÇÕES DE GUMBEL PARA ANÁLISE DOS DADOS DE VE- LOCIDADE MÁXIMA DO VENTO EM PIRACICABA, SÃO PAULO, BRASIL.	37
Resumo	37
Abstract	37
3.1 Introdução	38
3.2 Desenvolvimento	40
3.2.1 Distribuição de Gumbel	40
3.2.2 Misturas de distribuições de Gumbel	40
3.2.3 Teste de aleatoriedade	41
3.2.4 Estimação dos parâmetros da distribuição de Gumbel e seus respectivos intervalos de confiança	42
3.2.5 Estimação dos parâmetros das misturas de distribuições de Gumbel e seus respectivos intervalos de confiança	45
3.2.6 Determinação do número de componentes das misturas de distribuições de Gumbel	53
3.2.7 Diagnóstico do ajuste	54
3.2.8 Obtenção das probabilidades de ocorrência	57
3.2.9 Estimação de níveis de retorno e seus respectivos intervalos de confiança	57
3.3 Resultados e Discussão	59
3.4 Conclusões	67
Referências	67
APÊNDICES	71

RESUMO

A distribuição generalizada de Pareto e mistura de distribuições de Gumbel no estudo da vazão e da velocidade máxima do vento em Piracicaba, SP

A teoria dos valores extremos é um tópico da probabilidade que descreve a distribuição assintótica das estatísticas de ordem, tais como máximos ou mínimos, de uma sequência de variáveis aleatórias que seguem uma função de distribuição F normalmente desconhecida. Descreve, ainda, a distribuição assintótica dos excessos acima de um valor limiar de um ou mais termos dessa sequência. Dessa forma, as metodologias padrões utilizada neste contexto consistem no ajuste da distribuição generalizada dos valores extremos a uma série de máximos anuais ou no ajuste da distribuição generalizada de Pareto a uma série de dados compostas somente de observações excedentes de um valor limiar. No entanto, segundo Coles et al. (2003), há uma crescente insatisfação com o desempenho destes modelos padrões para predição de eventos extremos causada, possivelmente, por pressuposições não atendidas como a de independência das observações ou pelo fato de que os mesmos não sejam recomendados para serem utilizados em algumas situações específicas como por exemplo e quando observações de máximos anuais compostas por duas ou mais populações independentes de eventos extremos sendo que a primeira descreve eventos menos frequentes e de maior magnitude e a segunda descreve eventos mais frequentes e de menor magnitude. Então, os dois artigos que compõem este trabalho tem como objetivo apresentar alternativas de análise de valores extremos para estas situações em que o ajuste dos modelos padrões não são adequados. No primeiro, foram ajustadas as distribuições generalizada de Pareto e exponencial, caso particular da GP, aos dados de vazão média diária do Posto de Artemis, Piracicaba, SP, Brasil, conjuntamente com a técnica do desagrupamento, (“declustering”), e comparadas as estimativas dos níveis de retorno para períodos de 5, 10, 50 e 100 anos. Conclui-se que as estimativas intervalares dos níveis de retorno obtidas por meio do ajuste da distribuição exponencial são mais precisas do que as obtidas com o ajuste da distribuição generalizada de Pareto. No segundo artigo, por sua vez, foi apresentada uma metodologia para o ajuste da distribuição de Gumbel e de misturas de duas distribuições de Gumbel aos dados de velocidades de ventos mensais de Piracicaba, SP. Selecionou-se a distribuição que melhor ajustou-se aos dados por meio de testes de hipóteses “bootstrap” paramétrico e critérios de seleção AIC e BIC. E concluiu-se que a mistura de duas distribuições de Gumbel é a distribuição que melhor se ajustou-se aos dados de velocidades máxima de ventos dos meses de abril e maio, enquanto que o ajuste da distribuição de Gumbel foi o melhor para os meses de agosto e setembro.

Palavras-chave: Distribuição generalizada de Pareto; Mistura de distribuições de Gumbel; Desagrupamento; Vazão; Velocidade do vento

ABSTRACT

The generalized Pareto distribution and Gumbel mixture to study flow and maximum wind speed in Piracicaba, SP

The extreme value theory is a probability topics that describes the asymptotic distribution of order statistics such as maximum or minimum of random variables sequence that follow a distribution function F normally unknown. Describes still, the excess asymptotic distribution over threshold of this sequence. So, the standard methodologies of extremes values analysis are the fitting of generalized extreme value distribution to yearly maximum series or the fitting of generalized Pareto distribution to partial duration series. However, according to Coles et al. (2003), there is a growing dissatisfaction with the use this standard models for the prediction of extremes events and one of possible causes this fact may be a false assumptions about a sequence of observed data as a independence assumptions or because the standards models must not used in some specific situations like for example when maximum sample arise from two or more independents populations, where the first population describes more frequents and low intense events and the second population describes less frequents and more intense events. In this way, the two articles this work has a objective show alternatives about extreme values analysis for this situations that the standards models doesn't recommended. In the first article, the generalized distribution Pareto and exponential distribution, particular case of GP, together with to declustering methods was applied to mean daily flow of the Piracicaba river, Artemis station, Piracicaba, SP, and the estimates the return levels of 5, 10, 50 and 100 years were compared. We conclude that the interval estimates of the 50 and 100 year return levels obtained using the fitting the exponential distribution are more precise than those obtained using the generalized Pareto distribution. In the second article, we propose the fit of Gumbel distribution and the Gumbel mixture to data maximum speed wind in Piracicaba, SP. We select the best model using bootstrap test of hypotheses and the AIC and BIC selection criteria We conclude that the mixture Gumbel is the best model to analyze the maximum wind speed data for months of april e may and otherside the fit of Gumbel distributions was the best fit to months of august e september.

Keywords: Generalized Pareto distribution; Gumbel mixture; Declustering; Flow; Wind speed

1 INTRODUÇÃO

A teoria dos valores extremos é um tópico da probabilidade que descreve a distribuição assintótica das estatísticas de ordem, tais como máximos ou mínimos, de uma seqüência de variáveis aleatórias que seguem uma função de distribuição F normalmente desconhecida. Descreve, ainda, a distribuição assintótica dos excessos acima de um valor limiar de um ou mais termos dessa seqüência.

Embora atualmente a teoria dos valores extremos seja tema de vários artigos científicos, os primeiros trabalhos sobre o assunto foram publicados há mais de 70 anos atrás.

Segundo Bautista (2002), o marco inicial da teoria dos valores extremos é o trabalho de Fisher e Tippett (1928). Nesse artigo, os autores provaram que a distribuição dos máximos ou dos mínimos de uma amostra aleatória convergem ou para uma distribuição degenerada ou convergem para uma dos três tipos de distribuições assintótica dos valores extremos, Gumbel ou tipo 1, Fréchet ou tipo 2 e Weibull ou tipo 3.

Posteriormente, Gnedenko (1943) demonstrou as condições necessárias e suficientes para conhecer-se a qual das três distribuições limites convergem a distribuição dos máximos ou dos mínimos de uma amostra aleatória, caso estes não convirjam a uma distribuição degenerada.

Uma outra contribuição importante no desenvolvimento inicial da teoria dos valores extremos foi o trabalho de Jenkinson (1955). Neste, o autor mostrou que as funções de distribuição acumulada de Gumbel, Fréchet, Weibull, apresentadas por Fisher e Tippett (1928) são casos particulares da distribuição generalizada dos valores extremos (GVE).

No entanto, foi Gumbel (1941) o primeiro a propor uma metodologia de análise estatística envolvendo a teoria dos valores extremos. Em seu artigo, foram estimados períodos de retornos dos dados de vazão máxima anual do rio Missipi em Vicksburg, Missipi, EUA e do rio Rhône em Lyon, França, por meio da distribuição de Gumbel. Desde então, diversos trabalhos foram e estão sendo feitos para o desenvolvimento de métodos de análise estatística no contexto da teoria dos valores extremos. Por exemplo, utilizando a inferência clássica, Prescott e Walden (1980) discutiram aspectos computacionais sobre a estimação dos parâmetros por meio do método da máxima verossimilhança e Smith (1985) estudou as condições de regularidades dos estimadores de máxima verossimilhança da distribuição generalizada dos valores extremos. Por outro lado, usando a abordagem Bayesiana, Coles e Powell (1996) ajustaram a distribuição generalizada dos valores extremos a dados meteorológicos usando como distribuição a priori dos parâmetros, uma distribuição normal multivariada com valores dos hiperparâmetros baseados em informações de outras localidades. Além disso, Coles e Tawn (1996) elucidaram a distribuição a priori baseando-

se no conhecimento de um especialista da área.

É importante dizer, que até o momento, somente foi mencionado o desenvolvimento da teoria dos valores extremos sob o enfoque da distribuição dos máximos ou mínimos de variáveis aleatórias. Porém, um outro enfoque bastante importante é o ajuste de modelos probabilísticos baseados em séries históricas de dados denominadas séries de duração parcial, que, em síntese, são constituídas somente pelas observações excedentes acima de um valor limiar.

Os primeiros trabalhos neste contexto foram desenvolvidos por Todorovic e Zelenhasic (1970) que ajustaram o modelo exponencial Poisson à série parcial de vazão do rio Susquehanna em Wilkes Barre, Pensylvania, EUA. Nessa época, essa metodologia foi denominada como picos acima de um valor limiar (“peaks over threshold”).

Contudo, Pickands (1975) demonstrou que a distribuição assintótica dos excessos de uma variável aleatória acima de um valor limiar pode ser aproximada por meio da distribuição generalizada de Pareto, GP, que pode ser interpretada como uma família de distribuições que, dependendo do valor da parâmetro da forma, menor, maior ou tendendo a zero, inclui como casos particulares a distribuição exponencial, Weibull ou Pareto, respectivamente.

Desde então, diversos trabalhos sobre a implementação do método dos picos acima de um valor limiar tem sido feitos também utilizando o ajuste da distribuição generalizada de Pareto, como, por exemplo, os trabalhos de Hosking e Wallis (1987) e de Martins e Stedinger (2001).

Convém ressaltar que um aspecto crítico dessa metodologia é a escolha do valor limiar. Para solucionar este problema, Davison e Smith (1990) propuseram a escolha do valor limiar por meio da análise do gráfico das médias dos excessos proposto pelos mesmos.

Coles (2001), por sua vez, propôs um método que consiste em se estabelecer uma gama de valores limiares e para cada um desses valores, estimar os parâmetros da GP e observar a estabilidade do parâmetro da forma, uma vez que este não depende do valor limiar.

Todavia, Behrens et al. (2004) afirmaram que esses métodos não garantem uma escolha apropriada do valor limiar e propuseram o ajuste de um modelo de mistura Gama - GP sob a abordagem Bayesiana, considerando que o valor limiar é mais um parâmetro do modelo. Entretanto, pressupõe-se, nesse caso, que a função de distribuição de cada termo da sequência de variáveis aleatórias segue uma distribuição Gama, o que nem sempre pode ser adequado.

Além das metodologias envolvendo as distribuições GVE e GP, existem ainda a modelagem de dados de valores extremos baseadas na distribuição conjunta de r , estatística de ordem proposta por Smith (1986) e a análise de valores extremos por meio de modelos baseados em processos pontuais proposta por Smith (1989).

No entanto, independente do método de estimação e modelo adotado, modelos mal especificados ou com pressuposições não atendidas com o tipo de dados analisados, podem

trazer sérios prejuízos à predição de eventos extremos.

Uma pressuposição básica bastante questionável é a de que as observações das variáveis estudadas sejam independentes e identicamente distribuídas. Segundo Coles (2001), uma pressuposição mais realista é assumir que a sequência de variáveis aleatórias sejam estacionárias.

Há no entanto, muitas formas de dependência para uma sequência estacionária e então, o que se faz é impor algumas restrições tais como assumir que dois ou mais eventos extremos são aproximadamente independentes, desde que sejam bastante distanciados no tempo. Essas restrições são definidas matematicamente por Leadbetter et al. (1983).

Para o caso específico do ajuste da distribuição generalizada de Pareto a séries parciais, existem diversas formas na literatura de como incluir a dependência na modelagem, como, por exemplo, Smith et al. (1997), que modelaram as séries parciais considerando que a mesma segue uma cadeia de Markov e McNeil e Frey (2000), que por sua vez, ajustaram a distribuição generalizada de Pareto aos resíduos de um modelo GARCH.

Há, ainda, um método mais geral e simples de tratar o problema de séries parciais estacionária que é o método do desagrupamento (“declustering”) proposto por Leadbetter et al. (1989), baseado na pressuposição que eventos extremos persistem durante algumas observações consecutivas formando grupos. Nesse caso, a idéia principal é identificar esses grupos, determinar os máximos de cada grupo e ajustar a distribuição generalizada de Pareto a esses máximos.

Um outro problema relevante na análise de valores extremos é que os modelos tradicionais baseados no contexto da teoria dos valores extremos não contemplam situações em que as observações são provenientes de duas ou mais populações de eventos extremos independentes. Nesse caso, uma alternativa é modelar os dados por meio de uma mistura de duas distribuições generalizadas de valores extremos que, segundo Coles et al. (2003), descrevem o fenômeno físico analisado de uma maneira mais eficiente do que os modelos padrões.

Como exemplo de utilização do modelos de misturas de duas ou mais distribuições generalizadas de valores extremos, Walshaw (2000) ajustou um modelo de mistura de duas distribuições de valores extremos sob enfoque Bayesiano aos dados de velocidades máximas de ventos da cidade Boston, Massachusetts e de Key west, Florida, EUA. Como premissa, considerou que os ventos máximos podem ser oriundos ou de tempestades tropicais, que geralmente são de maior magnitude e menos freqüentes, ou de tempestades frontais geralmente mais amenos porém mais freqüentes.

Por sua vez, Tartaglia et al. (2006) ajustaram mistura de duas distribuições de Gumbel para os dados de precipitação pluvial em Toscana, Itália com a justificativa de que a série de precipitação pluvial era composta de precipitações de menor magnitude e mais frequentes e por precipitações de maior magnitude porém menos frequentes. No entanto, na literatura pouco se discute sobre a questão da seleção entre modelos de mistura e as distribuições padrões da teoria dos

valores extremos.

Sendo assim, os dois artigos que compõem este trabalho tem como objetivo apresentar alternativas de análise de valores extremos para estas situações em que o ajuste dos modelos padrões não são adequados.

O primeiro, apresenta o ajuste das distribuições generalizada de Pareto (GP) e exponencial, caso particular da GP, aliada a técnica desagrupamento (“declustering”) para a análise de dados de vazão média diária do Posto de Artemis, curso d’água rio Piracicaba, município de Piracicaba, São Paulo, Brasil cujo objetivo principal é estimar os níveis de retorno para os períodos de 5, 10, 50 e 100 anos.

O segundo, apresenta os ajustes da distribuição de Gumbel e das misturas de duas distribuições de Gumbel aos dados de velocidades máximas de vento de Piracicaba, dos meses de abril, maio, agosto e setembro. Esses meses foram escolhidos pois segundo Silva e Zocchi (2005), há uma aparente bimodalidade, o que pode sugerir uma mistura de duas distribuições. Quanto à seleção entre esses modelos probabilísticos foram utilizados testes de hipóteses baseados na técnica “bootstrap” paramétrico e os critérios de seleção AIC e BIC.

Referências

BAUTISTA, E.A.L. **A distribuição generalizada de valores extremos no estudo da velocidade máxima de vento em Piracicaba, SP**. 2002. 49 p. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2002.

BEHRENS, C.N.; LOPES, H.F.; GAMERMAN, D. Bayesian analysis of extreme events with threshold estimation. **Statistical Modelling**, London, v. 4, n. 3, p. 227-244, 2004.

COLES, S.G. **An introduction to statistical modeling of extreme values**. London: Springer. 2001. 226 p.

COLES, S.G.; POWELL. Bayesian methods in extreme value modelling: a review and new developments. **International Statistical Review**, Hague, n. 1, v. 64, p. 119-136, 1996.

COLES, S.G.; TAWN, J.A. A Bayesian analysis of extreme rainfall data. **Applied Statistics**, London, n. 4, v. 45, p. 463-478, 1996.

COLES, S.G.; PERICCHI, L.R.; SISSON, S. A. A fully probabilistic approach to extreme value modelling. **Journal of Hydrology**, Amsterdam, v. 273, n. 1, p. 35-50, 2003.

DAVISON, A.C.; SMITH, R.L. Models for exceedances over high thresholds. **Journal of the Royal Statistical Society, B**, London, v.520, n. 3, p. 393-442, 1990.

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

GUMBEL, E.J. The return period of flood flows. **Annals of Mathematical Statistics**, Ann Arbor, v. 12, n. 2, p. 163- 190, 1941.

GNEDENKO, B.V. Sur la distribution limite du terme maximum d'une sèrie aléatoire. **Annals of Mathematics**, Princeton, v. 44, n. 3, 423 - 453, 1943.

HOSKING, J.R.M.; WALLIS, J.R. Parameters and quantile estimation for the generalized Pareto distribution. **Technometrics**, Alexandria, v. 29, n. 3, p. 339-349, 1987.

JENKINSON, A.F. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. **Quartely Journal of the Royal Meteorological Society**, London, v. 81, n. 348, p. 158 - 171, 1955.

LEADBETTER, M.R.; LINDGREN, G.; ROOTZÉN, H. **Extremes and related properties of random sequences and series**. New York: Springer. 1983.

LEADBETTER, M.R.; WEISSMAN, I.; DE HAAN, L.; ROOTZÉN, H. **On clustering of high values in statistically stationary series**. Chapel Hill: Center for Stochastic Process, University of

North Carolina, 1989. p. 253.

MARTINS, E.S.; STEDINGER. J. Generalized maximum likelihood Pareto - Poisson estimators for partial duration series. **Water Resources Research**, Washington, v. 37, n. 10, p. 2551 - 2557, 2001.

McNEIL, A.J.; FREY, R. estimation of tail-related risk measures of heterocedastic financial time series: an Extreme Value Approach. **Journal of Empirical Finance**, Orlando, v. 7, n. 3, p. 271-300, 2000.

PRESCOTT, P.; WALDEN. A.T. Maximum estimation of the parameters of the generalized extreme value distribution. **Biometrika**, Cambridge, v. 67, n. 3, p. 723-724, 1980.

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

SILVA, R.R.; ZOCCHI, S.S. A distribuição generalizada de Pareto-Poisson no estudo da precipitação pluvial total diária máxima em Piracicaba, SP. **Revista de Matemática e Estatística**, Marília, v. 24, n. 3, p. 77-94, 2006.

SILVA, R.R.; ZOCCHI, S.S. Densidades não paramétricas no estudo da velocidade máxima do vento em Piracicaba, SP. In: RBRAS, 50.; SEAGRO, 11., 2005, Londrina. **Resumos ...** Londrina: UEL, 2005, 1 CD-ROM.

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

SMITH, R.L. extreme value theory based on the r largest annual events. **Journal of Hydrology**,

Amsterdan , v. 86, n.1, p. 27-43, 1986.

SMITH, R.L. Extreme value analysis of environmental times series: an application to trend detection in ground-level ozone. **Statistical Science**, Hayward, v. 4, n. 4, p. 367-393, 1989.

SMITH, R.L., TAWN, J. A.; COLES, S.G. Markov chain models for threshold exceedances. **Biometrika**, Cambridge, v. 84, n. 2, p. 249-268, 1997.

TARTAGLIA, V.; CAPORALI, E.; CAVIGLI, E.; MORO, A. L - Moments based assessment of a mixture model for frequency analysis of rainfall extremes. **Advanced in Geosciences**, Katlenburg - Lindau, v.2, p. 331-334, Jan. 2006. Disponível em: <[http: // www.adv-geosci.net/11/index.html](http://www.adv-geosci.net/11/index.html)>. Acesso em: 01 maio 2007.

TODOROVIC, P.; ZELENHASIC, E. A stochastic model for flood analysis. **Water Resources Research**, Washington, v.6, n.6, p.1641-1648, 1970.

WALSHAW, D. Modelling extreme wind speeds in regions proper to hurricanes. **Applied Statistics**, London, v. 49, n.1, p. 51-62, 2000.

2 AJUSTE DA DISTRIBUIÇÃO GENERALIZADA DE PARETO PARA ANÁLISE DE DADOS DE VAZÃO MÉDIA DIÁRIA EM PIRACICABA, SP, BRASIL

Resumo

Uma das metodologias mais utilizadas no contexto da teoria dos valores extremos é o ajuste da distribuição generalizada de Pareto, GP, para observações excedentes de um determinado valor limiar. No entanto, para o uso da GP, pressupõe-se que as observações sejam independentes e identicamente distribuídas, o que em geral não se verifica na prática. Uma forma de contornar esse problema é a utilização da técnica do desagrupamento (“declustering”), proposto por Leadbetter et al. (1989), que em síntese, identifica grupos de ocorrência de vazão extrema para posteriormente ajustar-se a GP apenas para os máximos destes. Neste trabalho, foram ajustadas a distribuição generalizada de Pareto e a distribuição exponencial, caso particular da GP, aos dados de vazão média diária do rio Piracicaba, Posto de Artemis, Piracicaba, SP, Brasil, conjuntamente com a técnica do desagrupamento, (“declustering”), e estimados os níveis de retorno para períodos de 5, 10, 50 e 100 anos. Conclui-se que as estimativas intervalares dos níveis de retorno obtidas por meio do ajuste da distribuição exponencial são mais precisas do que as obtidas com o ajuste da distribuição generalizada de Pareto.

Palavras-chave: Teoria dos valores extremos; Distribuição generalizada de Pareto; Desagrupamento; Vazão

Abstract

One of the most used methodologies in the context of the extremes value theory is the fitting of the generalized distribution Pareto to data exceedances over a threshold and the fitting of generalized extreme value distribution. However, an assumption to fit the generalized Pareto distribution to a dataset is the that the data has to be sequence independent and having a common distribution, what it is usually unrealistic. One way of solving this problem is to use the declustering methods proposed by Leadbetter et al. (1989), to identify groups of occurrence of extreme flow and fit of the generalized distribution Pareto to maximum of the groups. In this work, the fit generalized distribution Pareto and exponential distribution, together with declustering methods was applied to mean daily flow of the Piracicaba river, Artemis station, Piracicaba, SP, Brazil and after the estimates the return levels of 5, 10, 50 and 100 years were compared. We conclude that the interval estimates of the 50 and 100 year return levels obtained using the exponential distribution are more precise than those obtained using the generalized Pareto distribution.

Keywords: Extremes value theory; Generalized distribution Pareto; Declustering; Flow

2.1 Introdução

Em estudos sobre bacias hidrográficas, um dos aspectos mais importantes a serem avaliados é a vazão de um curso d'água em determinados pontos dessas bacias. A importância de obter essas informações atribui-se à necessidade de se construírem estruturas hidráulicas de controle de águas naturais para atenuar os prejuízos causados por vazões extremas. Para o dimensionamento dessas estruturas devem-se, então, considerar vazões extremas cujas probabilidades de ocorrência sejam pequenas.

Assim, para o cálculo dessas probabilidades, geralmente é utilizada a teoria dos valores extremos, uma vez que, segundo Bautista et al. (2004), os modelos probabilísticos baseados nesta teoria visam prever, a partir de um conjunto de dados observados num período de tempo relativamente curto, 60 anos por exemplo, os valores extremos esperados num período de tempo maior, 100 anos ou mais.

Uma das metodologias mais utilizadas neste contexto e para observações que excedem a um determinado valor limiar é o ajuste da distribuição generalizada de Pareto, GP que inclui a distribuição exponencial como caso particular. No entanto, para o ajuste dessas distribuições, as observações tem que atender a premissa de serem independentes e identicamente distribuídas, o que, segundo Coles (2001), pode não ser adequada para os tipos de dados citados.

Nessa situação, Davison e Smith (1990) propuseram a utilização da distribuição generalizada de Pareto aliada a um processo duplamente estocástico de Poisson. Alternativamente McNeil e Frey (2000) estimaram níveis de retorno, por meio do ajuste da distribuição generalizada de Pareto aos resíduos ordinários do modelo GARCH estimado por meio do método da máxima pseudo-verossimilhança.

Por outro lado, Leadbetter et al. (1989) apresentaram um método mais pragmático que os já citados, em que não é necessário pressupor qual é o processo estacionário que os dados analisados seguem, denominado método do desagrupamento (“declustering”).

Sendo assim, neste trabalho, foram ajustadas as distribuições generalizada de Pareto e exponencial aos dados de vazão média diária do Posto de Artemis, Piracicaba, SP, Brasil, conjuntamente com a técnica do desagrupamento e comparadas as estimativas dos níveis de retorno para períodos de 5, 10, 50 e 100 anos obtidas por meio desses dois modelos.

2.2 Desenvolvimento

Os dados de vazões médias diárias do rio Piracicaba, em $m^3.s^{-1}$, usados neste trabalho são provenientes do posto de Artemis prefixo 4D-007, cuja localização é longitude $47^\circ 46' 31''$, latitude $22^\circ 40' 45''$, município de Piracicaba, SP, Brasil. Esses dados são relativos ao período de 1944 à 2003, pertencentes ao banco de dados fluviométrico do Estado de São Paulo e estão

disponíveis na página <http://www.sigrh.sp.gov.br/cgi-bin/bdhm.exe/flu>.

Como informação adicional, Pellegrino et al. (2001) observaram que quase a totalidade da bacia hidrográfica do rio Piracicaba apresenta clima subtropical Cwa, segundo a classificação de Koeppen, com temperaturas médias entre 18°C e 22°C e com precipitação média de 1400 mm.

A metodologia para a análise desses dados é apresentada a seguir

Considere X a variável aleatória vazão média diária e $Y = X - u$ a variável aleatória excesso em relação a um determinado valor limiar u , condicionada a $X > u$. Supondo que X tem uma função de distribuição acumulada F pertencente ao domínio de atração de uma das três distribuições generalizada dos valores extremos, Gumbel, Fréchet ou Weibull, a distribuição de Y dado u é dada por

$$\begin{aligned} F(y) = Pr(Y \leq y) &= Pr(X \leq u + y | X > u) = \frac{Pr(u < X \leq u + y)}{Pr(X > u)} \\ &= \frac{F(u + y) - F(u)}{1 - F(u)} \end{aligned}$$

que segundo Pickands (1975), para u suficientemente grande, é bem aproximada pela função de distribuição acumulada generalizada de Pareto $G(y)$ dada por

$$G(y) = 1 - \left(1 + \xi \frac{y}{\sigma}\right)^{-\frac{1}{\xi}} \quad (2.1)$$

cujas função densidade de probabilidade é dada por

$$g(y) = \frac{dG(y)}{dy} = \frac{1}{\sigma} \left(1 + \xi \frac{y}{\sigma}\right)^{-\left(\frac{1+\xi}{\xi}\right)},$$

em que $y > 0$ e $\left(1 + \xi \frac{y}{\sigma}\right) > 0$, sendo $-\infty < \xi < \infty$ o parâmetro de forma e $\sigma > 0$ o parâmetro de escala.

As funções de distribuição acumuladas de Pareto e de Weibull correspondem aos casos particulares de de (2.1) em que $\xi > 0$ e $\xi < 0$, respectivamente. Como limite de $G(y)$ com ξ tendendo a zero tem-se que

$$\lim_{\xi \rightarrow 0} G(y) = 1 - \exp\left(-\frac{y}{\sigma}\right),$$

para $y > 0$, que corresponde à função de distribuição acumulada exponencial com parâmetro de escala $\frac{1}{\sigma}$, cuja função densidade de probabilidade é

$$g(y) = \frac{1}{\sigma} \exp\left(-\frac{y}{\sigma}\right).$$

Como passo seguinte e antes de se fazer inferência a respeito dos parâmetros σ , ξ e dos níveis de retorno, deve-se escolher um valor limiar u apropriado.

Essa escolha geralmente é feita baseando-se no fato de que para u suficientemente grande e para $\xi < 1$, a esperança condicional de $Y - \delta$ dado $Y > \delta$ é dada por

$$\begin{aligned}
 E(Y - \delta | X > \delta) &= \frac{\int_0^\infty yg(y + \delta) dy}{1 - G(\delta)} \\
 &= \frac{\int_0^\infty y \frac{1}{\sigma} \left(1 + \frac{\xi}{\sigma}(y + \delta)\right)^{-\left(\frac{1+\xi}{\xi}\right)} dy}{\left(1 + \frac{\xi}{\sigma}\delta\right)^{-\frac{1}{\xi}}} \\
 &= \frac{1}{\sigma} \left(1 + \frac{\xi}{\sigma}\delta\right)^{\frac{1}{\xi}} \int_0^\infty y \left(1 + \frac{\xi}{\sigma}(y + \delta)\right)^{-\left(\frac{1+\xi}{\xi}\right)} dy. \tag{2.2}
 \end{aligned}$$

Fazendo a substituição

$$v = 1 + \frac{\xi}{\sigma}(y + \delta), \tag{2.3}$$

tem-se que

$$\frac{dv}{dy} = \frac{\xi}{\sigma} \Rightarrow dv = \frac{\sigma}{\xi} dy \tag{2.4}$$

e além disso, para $y \rightarrow \infty$, tem-se que $v \rightarrow \infty$ e para $y = 0$, $v = \left(1 + \frac{\xi}{\sigma}\delta\right) = \omega$ o que implica que, substituindo (2.3) e (2.4) em (2.2) obtém-se

$$\begin{aligned}
 E(Y - \delta | X > \delta) &= \frac{1}{\sigma} \left(1 + \frac{\xi}{\sigma}\delta\right)^{\frac{1}{\xi}} \lim_{a \rightarrow \infty} \int_\omega^a \left(\frac{\sigma}{\xi}(v - 1) - \delta\right) v^{-\frac{1+\xi}{\xi}} \frac{\sigma}{\xi} dv \\
 &= \frac{1}{\xi} \left(1 + \frac{\xi}{\sigma}\delta\right)^{\frac{1}{\xi}} \lim_{a \rightarrow \infty} \int_\omega^a \frac{\sigma}{\xi} v^{-\frac{1}{\xi}} - \frac{\sigma}{\xi} v^{-\left(\frac{1+\xi}{\xi}\right)} - \delta v^{-\left(\frac{1+\xi}{\xi}\right)} dv \\
 &= \frac{1}{\xi} \left(1 + \frac{\xi}{\sigma}\delta\right)^{\frac{1}{\xi}} \lim_{a \rightarrow \infty} \left[-\frac{\sigma}{1-\xi} v^{-\frac{1+\xi}{\xi}} + (\sigma + \xi\delta) v^{-\frac{1}{\xi}} \right] \Big|_\omega^a \\
 &= \frac{1}{\xi} \left(1 + \frac{\xi}{\sigma}\delta\right)^{\frac{1}{\xi}} \left[\frac{\sigma}{1-\xi} \left(1 + \frac{\xi}{\sigma}\delta\right)^{-\frac{1+\xi}{\xi}} - (\sigma + \xi\delta) \left(1 + \frac{\xi}{\sigma}\delta\right)^{-\frac{1}{\xi}} \right] \\
 &= \frac{\sigma + \xi\delta}{(1-\xi)\xi} - \frac{\sigma + \xi\delta}{\xi} \\
 &= \frac{\sigma + \xi\delta}{1-\xi}, \tag{2.5}
 \end{aligned}$$

válida para $\delta \in (0, -\frac{\sigma}{\xi})$ quando $\xi < 0$ e para $\delta \in (0, \infty)$ quando $0 < \xi < 1$.

Assim para um determinado valor limiar $w = u + \delta$ tem-se

$$\begin{aligned}
E(X - w|X > w) &= E(X - (u + \delta)|X > (u + \delta)) = E(Y - \delta|X > \delta) \\
&= \frac{\sigma + \xi(w - u)}{1 - \xi} = \frac{\sigma - \xi u}{1 - \xi} + \frac{\xi}{1 - \xi}w,
\end{aligned} \tag{2.6}$$

em que $w \in (u, u - \frac{\sigma}{\xi})$ quando $\xi < 0$ e $w \in (u, \infty)$ quando $0 < \xi < 1$.

Ou seja, dado um valor limiar u , suficientemente grande, a esperança dos excessos em relação a um limiar w maior que u é uma função de w afim com coeficiente angular $\frac{\xi}{1-\xi}$ e intercepto $\frac{\sigma - \xi u}{1 - \xi}$, conforme ilustra a figura 2.1.

Figura 2.1 – Gráficos da esperança de $X - w|X > w$ em função de $w \geq u$, para os casos em que $Y = X - u|X > u$ segue a distribuição Weibull, ($\xi < 0$), exponencial ($\xi \rightarrow 0$) e Pareto, ($\xi > 0$), respectivamente.

Na prática, dada uma série de n observações de vazões médias diárias, Davison e Smith (1990) propõem construir o gráfico das médias dos excessos em relação a um valor w , em função de w , ou seja, construir o gráfico da função

$$\mu(w) = \frac{1}{n_w} \sum_{h=1}^{n_w} (x_h - w) \tag{2.7}$$

para $x_{\min} < w < x_{\max}$, sendo $x_{\min}$ e $x_{\max}$, a mínima e a máxima vazão média diária observada, respectivamente, h o índice da h -ésima vazão excedente de w e n_w o número de excedentes sobre w .

Sugerem, então, como valor limiar u o menor valor de w a partir do qual a função das médias dos excessos amostral segue aproximadamente uma função afim.

No entanto, para tornar a interpretação do gráfico da média dos excessos e a escolha do valor limiar menos subjetiva, Smith (2004) propôs a construção de um envelope simulado, técnica proposta primeiramente por Atkinson (1985) no contexto de análise de resíduos de modelos de regressão, cujo procedimento é apresentado a seguir.

Considerando-se os dados da amostra original e um valor limiar fixo u :

1. Obter as estimativas de máxima verossimilhança, cujo método será descrito mais adiante, $\hat{\sigma}$ e $\hat{\xi}$ dos parâmetros σ e ξ da distribuição generalizada de Pareto e para cada $w \in [u; x_{max}]$, calcular $\mu(w)$, dado por (2.7);
2. Gerar 99 amostras aleatórias da distribuição generalizada de Pareto com parâmetros $\sigma = \hat{\sigma}$ e $\xi = \hat{\xi}$, de tamanhos iguais aos da amostra original;
3. Para cada amostra, obter as estimativas de máxima verossimilhança e a média amostral dos excessos de $w \in [u; x_{max}]$ para cada w , ou seja, obter $\hat{\sigma}^{(s)}$, $\hat{\xi}^{(s)}$ e $\hat{\mu}^{(s)}(w)$, para $s = 1, \dots, 99$;
4. Para cada $w \in [u; x_{max}]$, calcular os percentis de ordem $\frac{\alpha}{2}$ e $1 - \frac{\alpha}{2}$ de

$$\left\{ \mu(w) - \frac{\hat{\sigma}^{(1)} + \hat{\xi}^{(1)}(w - u)}{1 - \hat{\xi}^{(1)}} + \frac{\hat{\sigma} + \hat{\xi}(w - u)}{1 - \hat{\xi}}, \dots, \mu(w) - \frac{\hat{\sigma}^{(99)} + \hat{\xi}^{(99)}(w - u)}{1 - \hat{\xi}^{(99)}} + \frac{\hat{\sigma} + \hat{\xi}(w - u)}{1 - \hat{\xi}} \right\}$$

que formam, para cada w , os limites inferiores e superiores do envelope, de $100 \times (1 - \alpha)\%$ de confiança, respectivamente.

Para escolher o valor limiar u , Smith (2004) recomenda pressupor v valores limiares candidatos $\{c_1, \dots, c_v\}$, construir o gráfico das médias dos excessos para cada um desses valores, e escolher como sendo o valor limiar u , o menor valor limiar candidato cujo o gráfico tenha a média dos excessos amostral contida dentro do envelope simulado.

Uma observação importante é que além de ser útil para escolha do valor limiar o gráfico das médias dos excessos fornece informações a respeito do tipo da cauda da distribuição generalizada de Pareto, uma vez que se a partir do valor limiar u for observado que a função das médias dos excessos amostral for crescente, infere-se que $\xi > 0$, caso for decrescente pode-se dizer que $\xi < 0$ e caso contrário, $\xi = 0$, conforme ilustra a figura 2.1.

Uma vez definida a escolha do valor limiar u , seguindo as idéias de Rubem (2006), testa-se a hipótese nula de que a série de k observações de vazão média diária é estacionária utilizando o teste KPSS proposto por Kwiatkowski et al. (1992) e descrito a seguir.

Considere que cada elemento de $x_i^{(*)}$ possa ser decomposto pela soma de três componentes não observáveis, ou seja,

$$x_i^{(*)} = \kappa + \beta i + \lambda_i + \epsilon_i$$

sendo κ uma constante, β o efeito da tendência determinística, ϵ o erro estacionário com distribuição $N(0, \rho^2)$ e λ_i um passeio aleatório definido por

$$\lambda_i = \lambda_{(i-1)} + \gamma_i$$

em que γ o i -ésimo elemento de sequência aleatória independente e identicamente distribuída que segue $N(0, \psi^2)$.

Nesse caso, testar a hipótese nula de estacionariedade é equivalente a testar a hipótese $H_0 : \beta = \psi^2 = 0$, e sendo assim o teste KPSS pode ser implementado da seguinte forma.

Primeiramente, sob a hipótese H_0 a série de dados $\{x_1^{(*)}, \dots, x_k^{(*)}\}$, ou seja, ajusta-se o modelo

$$x_i^{(*)} = \kappa + \epsilon_i$$

cujas estimativas de mínimos quadrados de κ é dada por

$$\hat{\kappa} = \frac{1}{k} \sum_{i=1}^k x_i^{(*)} = \bar{x}_i^{(*)}.$$

Em seguida, obtém-se os resíduos ordinários para cada $x_i^{(*)}$ definidos por

$$\hat{\epsilon}_i = x_i^{(*)} - \hat{\kappa} = x_i^{(*)} - \bar{x}_i^{(*)},$$

sendo $\hat{x}_i^{(*)}$ a vazão média diária excedente de u predita pelo modelo, estima-se a soma de resíduos parciais, denotada por η_m , dada por

$$\eta_m = \sum_{m=1}^i \hat{\epsilon}_m$$

e obtém-se o estimador consistente de ρ^2 dado por

$$\hat{\rho}^2 = \frac{1}{k} \sum_{i=1}^k \hat{\epsilon}_i^2 + \frac{2}{k} \sum_{m=1}^{\phi} \frac{m}{\phi + 1} + \sum_{t=m+1}^k \hat{\epsilon}_i \hat{\epsilon}_{i-m}.$$

em que ϕ é o parâmetro de referência definido por meio de

$$\phi = \lfloor 3 \frac{\sqrt{k}}{13} \rfloor.$$

em que $\lfloor . \rfloor$ representa de maior número inteiro menor igual $3 \frac{\sqrt{k}}{13}$.

Por fim, obtém-se a estatística do teste KPSS, ζ , dada por

$$\zeta = \frac{\sum_{i=1}^k \eta_i}{k^2 \hat{\rho}^2}.$$

Então, para rejeitar H_0 a um nível de significância $100 \times \alpha$, deve-se verificar que o valor da estatística ζ seja maior que o valor crítico encontrado na tabela construída por Kwiatkowski et al. (1992, p. 166).

Neste momento é importante comentar que no caso de rejeição da hipótese de nulidade, a metodologia utilizada nesse trabalho não é recomendada, caso contrário, prossegue-se com o método do desagrupamento (“declustering”) descrito a seguir.

Sejam $\{x_1^{(*)}, \dots, x_k^{(*)}\}$ o conjunto de k vazões médias diárias excedentes de u observadas durante um período de N anos e $\{y_1, \dots, y_k\}$ os excessos em relação a u . Considere, ainda, que $\{d_i\} = \{l : x_l \geq u\}$, válido para todo $i = 2, \dots, k$ e $l = 1, \dots, n$, seja o conjunto das l -ésimas ocorrências de vazões médias diárias excedentes em relação a u durante o período de anos considerado e seja e_i o número de ocorrência de vazões médias diárias abaixo do valor limiar u entre as ocorrências $x_{(i-1)}^{(*)}$ e $x_i^{(*)}$ definido por

$$e_i = (d_i - d_{i-1}) - 1, \quad i = 2, \dots, k,$$

r um número fixo de ocorrência de excedentes abaixo de u , H_j , o j -ésimo grupo de excessos válidos para $j = 1, \dots, J$ em que J é o número de grupos formados e H_{j-1} , o grupo antecessor do j -ésimo grupo de excessos válido para $j = 2, \dots, J$. Uma regra empírica para identificar os grupos é dada por.

Para $i = 1$, $y_1 \in H_1$ e por sua vez para $i = 2, \dots, k$ tem-se

$$\begin{cases} \text{Se } e_i < r, & y_{i-1} \text{ e } y_i \in H_j \\ \text{caso contrário,} & y_{i-1} \in H_{j-1} \text{ e } y_i \in H_j. \end{cases}$$

válida para $i = 2, \dots, k$.

Convém ressaltar que como na literatura não há uma metodologia para escolha do valor de r , então, o critério proposto neste trabalho é fazer o gráfico de autocorrelação dos excedentes para alguns supostos valores de r e escolher o menor valor de r tal que verifique-se independência dos excedentes por meio de uma inspeção visual.

Uma vez escolhido o valor de r , seja Z a variável aleatória máximo dos excessos dentro dos grupos, em que $Z \sim GP(\sigma, \xi)$. Considere o conjunto de máximos observados $\{z_1, \dots, z_J\}$ dos J grupos e que esses valores sejam independentes e identicamente distribuídos. Pode-se definir, dessa forma, a função de verossimilhança da distribuição generalizada de Pareto como sendo

$$L(\sigma, \xi) = \prod_{j=1}^J \frac{1}{\sigma} \left(1 + \xi \frac{z_j}{\sigma}\right)^{-\left(\frac{1+\xi}{\xi}\right)}$$

cujo logaritmo é

$$l(\sigma, \xi) = -J \log \sigma - \frac{1+\xi}{\xi} \sum_{j=1}^J \log \left(1 + \xi \frac{z_j}{\sigma}\right). \quad (2.8)$$

Derivando - se (2.8) em relação a σ e ξ e igualando essas derivadas a zero obtém-se o seguinte sistema de equações

$$\begin{cases} \frac{\partial l(\sigma, \xi)}{\partial \sigma} = -\frac{J}{\sigma} + \left(\frac{1+\xi}{\xi}\right) \frac{\xi}{\sigma^2} \sum_{j=1}^J \frac{z_j}{\left(1 + \xi \frac{z_j}{\sigma}\right)} = 0 \\ \frac{\partial l(\sigma, \xi)}{\partial \xi} = \frac{1}{\xi^2} \sum_{j=1}^J \log\left(1 + \xi \frac{z_j}{\sigma}\right) - \left(\frac{1+\xi}{\xi}\right) \sum_{j=1}^J \frac{1}{\left(1 + \xi \frac{z_j}{\sigma}\right)} \frac{z_j}{\sigma} = 0 \end{cases}$$

cuja solução, $(\hat{\sigma}, \hat{\xi})$ são as estimativas de máxima verossimilhança para (σ, ξ) .

Uma vez que não há solução analítica para resolver esse sistema de equações utilizou-se o método quase Newton “BFGS” descrito por Nocedal e Wright (1999) e implementado no software R 2.5.1, R Development Core Team (2006).

Para o caso particular da distribuição GP com ξ tendendo a zero, ou seja, para a distribuição exponencial, no entanto, tem-se que o logaritmo da função de verossimilhança é dado por

$$l(\sigma) = -J \log(\sigma) - \frac{1}{\sigma} \sum_{j=1}^J z_i \quad (2.9)$$

e o estimador de máxima verossimilhança para σ é obtido solucionando a equação formada pela derivada primeira em relação a σ e igualando a zero, ou seja,

$$\frac{dl(\sigma)}{d\sigma} = -\frac{J}{\sigma} + \frac{\sum_{j=1}^J z_i}{\sigma^2} = 0.$$

cuja solução é dada por

$$\hat{\sigma} = \frac{\sum_{j=1}^J z_i}{J} = \bar{z},$$

ou seja, pela média dos máximos dos J grupos.

Uma vez obtidas as estimativas pontuais, pode-se construir os intervalos de $100 \times (1 - \alpha)\%$ de confiança para os parâmetros σ e ξ utilizando o método do perfil da verossimilhança.

Esse método consiste em estabelecer uma gama de valores do parâmetro para o qual deseja obter o intervalo e, para cada um desses valores fixados, substituí-lo em (2.9) e maximizá-la em relação ao outro parâmetro não fixado.

Dessa forma, o intervalo de $100 \times (1 - \alpha)\%$ para o parâmetro σ é definido da seguinte forma

$$I.C_{(100 \times (1 - \alpha)\%)}(\sigma) = \left\{ \sigma : 2 \left[l(\hat{\sigma}, \hat{\xi}) - l(\xi, \sigma_0) \right] \leq \chi_{1, (1 - \alpha)}^2 \right\}$$

e o intervalo de $100 \times (1 - \alpha)\%$ para o parâmetro ξ é dado por

$$I.C_{(100 \times (1 - \alpha)\%)}(\xi) = \left\{ \xi : 2 \left[l(\hat{\sigma}, \hat{\xi}) - l(\sigma, \xi_0) \right] \leq \chi_{1, (1 - \alpha)}^2 \right\}$$

em que σ_0 é um valor fixo de σ , ξ_0 é um valor fixo de ξ e $\chi_{1, (1 - \alpha)}^2$ é o quantil de ordem $100 \times (1 - \alpha)$ da distribuição qui-quadrado com 1 grau de liberdade.

Para o caso em que ξ tende a zero, seguindo as idéias propostas por Azzalini (1996), o intervalo de $100 \times (1 - \alpha)\%$ de confiança para o parâmetro pode ser obtido baseando-se na função “deviance” definida por

$$D(\sigma) = 2 [l(\hat{\sigma}) - l(\sigma)] = 2 \left[J \log \left(\frac{\sigma}{\hat{\sigma}} \right) + \sum_{j=1}^J \left(\frac{1}{\sigma} - \frac{1}{\hat{\sigma}} \right) \right] \quad (2.10)$$

e nese caso, o intervalo de $100 \times (1 - \alpha)\%$ de confiança para o parâmetro σ é obtido por meio de

$$I.C_{(100 \times (1 - \alpha)\%)}(\sigma) = \left\{ \sigma : D(\sigma) \leq \chi_{1, (1 - \alpha)}^2 \right\}. \quad (2.11)$$

Depois de estimar os parâmetros, geralmente há o interesse em estimar o nível de retorno, τ , que é definido por Chow (1964), como o nível que é excedido em média uma vez a cada T anos. Entretanto, deve-se levar em consideração que a série de vazões médias é observada diariamente e para obter o nível de retorno τ associado ao período de retorno de T anos, considera-se que t seja o número médio de dias esperados até a ocorrência de τ durante o período T pré especificado, ou seja,

$$t = T365, 25,$$

e que

$$Pr(X > \tau | X > u) = \left[1 + \xi \left(\frac{\tau - u}{\sigma} \right) \right]^{-\frac{1}{\xi}} = \frac{1}{t} = \frac{1}{T365, 25} \quad (2.12)$$

o que implica que

$$Pr(X > \tau) = p \left[1 + \xi \left(\frac{\tau - u}{\sigma} \right) \right]^{-\frac{1}{\xi}} = \frac{1}{T365, 25}. \quad (2.13)$$

Dessa forma, define-se o nível de retorno isolando τ em (2.13), ou seja,

$$\tau = u + \frac{\sigma}{\xi} \left[(T365, 25p)^{\xi} - 1 \right] \quad (2.14)$$

sendo $p = Pr(X > u)$ a probabilidade de ocorrer uma vazão média diária maior do que u .

Assim, dadas n observações independentes de vazões médias diárias e seja Q a variável aleatória número de ocorrência de vazões médias diárias excedentes de u que segue distribuição binomial com parâmetros p e n , pode-se estimar o parâmetro p por meio da máxima verossimilhança cujo estimador é definido como

$$\hat{p} = \frac{k}{n}.$$

Sendo assim, o nível de retorno τ pode ser estimado por meio de

$$\hat{\tau} = u + \frac{\hat{\sigma}}{\hat{\xi}} \left[(T365, 25\hat{p})^{\hat{\xi}} - 1 \right] \quad (2.15)$$

que é válida apenas quando considera-se $X_1, \dots, X_n$ independentes e identicamente distribuídas.

No entanto, está sendo considerado apenas que a série é estacionária e dessa forma, para levar-se em consideração a dependência das observações nas estimativas dos níveis de retorno, Coles (2001) propõe reescrever (2.14) da seguinte forma

$$\tau = u + \frac{\sigma}{\xi} \left[(T365, 25\Theta p)^{\xi} - 1 \right] \quad (2.16)$$

em que $\Theta \in (0, 1)$ é o parâmetro “extremal index” que pode ser interpretado como o tamanho médio dos grupos para $n \rightarrow \infty$, e estimado por meio de

$$\hat{\Theta} = \frac{J}{k}.$$

Sendo assim, a estimativa $\hat{\tau}$ é obtida substituindo-se σ , ξ , p e Θ por suas estimativas $\hat{\sigma}$, $\hat{\xi}$, $\hat{p}$ e $\hat{\Theta}$ respectivamente, ou seja,

$$\hat{x}_T = u + \frac{\hat{\sigma}}{\hat{\xi}} \left[(Tn_y \hat{\Theta} \hat{p})^{\hat{\xi}} - 1 \right] = u + \frac{\hat{\sigma}}{\hat{\xi}} \left[(T365, 25 \frac{J}{n} \hat{p})^{\hat{\xi}} - 1 \right]$$

Por sua vez, para o caso em que ξ tende a zero, tem-se que

$$\hat{\tau} = u + \hat{\sigma} \log \left(T365, 25 \frac{J}{n} \hat{p} \right).$$

Para a construção do intervalo de $100 \times (1 - \alpha)\%$ de confiança de τ utilizando a distribuição GP, primeiramente, define-se uma reparametrização da função densidade de probabilidade da distribuição generalizada de Pareto, isolando σ na expressão (2.16), ou seja,

$$\sigma = \frac{(\tau - u)\xi}{(T365, 25 \frac{J}{n} p)^{\xi} - 1}.$$

Ignorando a incerteza de Θ e p , a partir de então procede-se de maneira similar a construção de intervalo de $100 \times (1 - \alpha)\%$ de confiança para os parâmetros, ou seja, o intervalo dos níveis de retorno para um período de T anos fica definido como

$$I.C_{(100 \times (1 - \alpha)\%)}(\tau) = \left\{ \tau : 2 \left[l(\hat{\sigma}, \hat{\xi}) - l(\xi, \tau_0) \right] \leq \chi_{1, (1 - \alpha)}^2 \right\}$$

em que $l(\xi, \tau_0)$ é o logaritmo da função de verossimilhança dado um valor τ_0 fixo do nível de retorno.

Para o caso em que ξ tende a zero, a reparametrização de σ é dada por

$$\sigma = \frac{x - u}{\log(T365, 25\Theta p)}$$

e então o intervalo dos níveis de retorno para um período de T anos é definido por meio de

$$I.C_{(100 \times (1 - \alpha)\%)}(\tau) = \left\{ \tau : 2 [l(\hat{\sigma}) - l(\tau_0)] \chi_{1, (1 - \alpha)}^2 \right\}$$

Finalmente, para avaliar a qualidade do ajuste neste trabalho sugere-se a utilização gráficos probabilidade - probabilidade e quantil - quantil com envelope simulado. Ordenando de forma crescente os máximos dos excessos dos J grupos observados em relação a um valor limiar u , $\{z_{(1)}, \dots, z_{(J)}\}$, o gráfico probabilidade - probabilidade é definido pelos pontos de coordenadas $\left\{ \frac{j}{J+1}, \hat{G}(z_{(j)}) \right\}$, sendo

$$\hat{G}(z_{(j)}) = 1 - \left[1 + \hat{\xi} \left(\frac{y}{\hat{\sigma}} \right) \right]^{-\frac{1}{\hat{\xi}}}. \quad (2.17)$$

Para obter o envelope simulado de $100 \times (1 - \alpha)\%$ de confiança deve-se proceder da seguinte forma

- 1 - Gerar B amostras de tamanho J a partir da $GP(\hat{\sigma}, \hat{\xi})$ denotadas por $\left\{ z_1^{*(1)}, \dots, z_J^{*(1)} \right\}, \dots, \left\{ z_1^{*(B)}, \dots, z_J^{*(B)} \right\}$,
- 2 - Para cada uma das B amostras, obter os estimadores de máxima verossimilhança de σ e ξ , ordenar as amostras geradas em forma crescente e obter pontos de coordenadas $\left\{ \frac{j}{J+1}, \hat{G}(z_{(j)}) \right\}$.
- 3 - Para cada j obter os limites superior e inferior de confiança $100 \times (1 - \alpha)\%$ do envelope que são definidos como sendo o quantis de ordens $\frac{\alpha}{2}$ e $(1 - \frac{\alpha}{2})$ de $\left(z_j^{(1)}, \dots, z_j^{(B)} \right)$ respectivamente.

O gráfico quantil - quantil, por sua vez, é definido pelos pontos de coordenadas $\left(\hat{G}^{-1} \left(\frac{j}{J+1} \right), z_{(j)} \right)$, sendo

$$\hat{G}^{-1} \left(\frac{j}{J+1} \right) = u + \frac{\hat{\sigma}}{\hat{\xi}} \left[z^{-\hat{\xi}} - 1 \right] \quad (2.18)$$

e a obtenção do envelope é feita de forma similar ao gráfico probabilidade-probabilidade com a diferença que no passo 2, para cada uma das B amostras deve-se obter os pares de coordenadas $\left(\hat{G}^{-1}\left(\frac{j}{J+1}\right), z_{(j)}\right)$ ao invés dos pares de coordenadas $\left(\frac{j}{J+1}, \hat{G}(z_{(j)})\right)$.

A interpretação de ambos os gráficos é a seguinte. Se o modelo está bem ajustado aos dados, então ambos têm tendência linear e a amostra observada deverá estar contida dentro dos limites de confiança dos envelopes simulados.

2.3 Resultados e Discussão

A primeira etapa do ajuste da distribuição generalizada de Pareto é a escolha do valor limiar u e, para essa finalidade, primeiramente foi feito o gráfico das médias dos excessos proposto por Davison e Smith (1990), conforme mostra a figura 2.2. Analisando essa figura, observa-se que para $w > 730 \text{ m}^3 \cdot \text{s}^{-1}$, o gráfico da média dos excessos é aproximadamente linear e decrescente com um aspecto “serreado” evidente. No entanto, segundo Smith (2004), valores de w cuja a média dos excessos tem um aspecto serreado muito evidente devem ser considerados menos representativos para escolha do valor limiar. Então, adotando-se o critério de que o valor limiar “ótimo” é o menor valor de u tal que o gráfico da média dos excessos seja aproximadamente linear, escolheu-se $u = 400 \text{ m}^3 \cdot \text{s}^{-1}$. De modo a verificar empiricamente se esse limiar é adequado, foram construídos os gráficos da média dos excessos com envelope simulado, utilizando nível de significância de 5% e 99 simulações de Monte Carlo, para valores de w maiores que $u = 400 \text{ m}^3 \cdot \text{s}^{-1}$ e $u = 730 \text{ m}^3 \cdot \text{s}^{-1}$, apresentados na figura (2.3). Analisando essa figura, não se rejeita a hipótese de que os excessos sigam a distribuição generalizada de Pareto em nenhum dos dois gráficos apresentados, uma vez que a maioria dos pontos das médias dos excessos estão contidos entre os limites inferior e superior do envelope simulado nos respectivos gráficos. Portanto, adotando-se o critério de que deve-se escolher, como sendo o valor limiar u , o menor valor limiar candidato cujo os pontos do gráfico da média amostral dos excessos estão contidos dentro dos limites do envelope simulado, escolheu-se $u = 400 \text{ m}^3 \cdot \text{s}^{-1}$ confirmando a escolha anterior. Além disso, observa-se que o gráfico da estimativa da esperança da média dos excessos é uma reta decrescente, o que sugere que $\xi < 0$, ou seja, que se trata de uma distribuição de Weibull.

Uma vez escolhido o valor limiar u o próximo passo é verificar a hipótese nula de estacionariedade das vazões médias diárias observadas excedentes de $400 \text{ m}^3 \cdot \text{s}^{-1}$. Para esta finalidade, primeiramente foi feita uma inspeção visual do gráfico do logaritmo da vazão média diária excedente de $400 \text{ m}^3 \cdot \text{s}^{-1}$ conforme mostra a figura 2.4. Analisando-se o mesmo não se observa nenhum tipo de tendência ao longo dos anos. Logo em seguida, para verificar formalmente essa hipótese, foi feito o teste KPSS. Utilizando-se o parâmetro de referência igual a 6, foi obtida a estatística do teste igual a 0,0325 e valor-p aproximadamente igual a 1, o que implica que não há evidência estatística para rejeitar a hipótese nula de estacionariedade ao nível de significância de

Figura 2.2 – Gráfico das médias dos excessos de vazão média diária ($m^3.s^{-1}$) em função de w

Figura 2.3 – Gráfico das médias dos excessos para dois valores limiares candidatos: (a) $u = 400 \text{ m}^3 \cdot \text{s}^{-1}$ e (b) $u = 730 \text{ m}^3 \cdot \text{s}^{-1}$ para dados de vazões médias diárias com envelopes simulados de 95 % de confiança utilizando-se 99 simulações de Monte Carlo

5 %. Portanto, pode-se concluir que, embora Moraes et. al (1997) e Groppo et. al (2005) tenham relatado uma tendência linear decrescente da vazão do rio Piracicaba devido à implantação do sistema de abastecimento público de água Cantareira, essa tendência não é significativa para as observações de vazão acima de $400 \text{ m}^3.\text{s}^{-1}$.

Figura 2.4 – Gráfico da série temporal do logaritmo da vazão média diária excedentes de $400 \text{ m}^3.\text{s}^{-1}$

A figura 2.5, por sua vez, apresenta o gráfico de autocorrelação dos excedentes acima de $400 \text{ m}^3.\text{s}^{-1}$, e o gráficos de autocorrelação dos máximos dos grupos de excedentes acima $400 \text{ m}^3.\text{s}^{-1}$ para $r = 1$, $r = 2$ e $r = 4$, respectivamente. Analisando o gráfico de autocorrelação dos excedentes vê-se que as estimativas de autocorrelação para cada “lag”, em geral, ultrapassam os limites do intervalo de 95% de confiança calculado sob a hipótese de que as observações não sejam autocorrelacionadas o que implica que existem evidências estatística para rejeitar essa hipótese. Por outro lado, analisando os gráficos de autocorrelação dos máximos dos grupos dos excedentes para $r = 1$, $r = 2$ e $r = 4$, observa-se que as estimativas de autocorrelação dos máximos de grupos dos excedentes não apresentam nenhum tipo de padrão não ultrapassando os limites do intervalo de 95 % de confiança. Isto indica que não há evidências estatísticas para rejeitar a hipótese de que os máximos dos grupos de excedentes sejam aproximadamente independentes. Logo, como conclusão adotou-se $r = 1$.

Uma vez escolhido o valor limiar u e o valor de r , a distribuição generalizada de Pareto foi ajustada aos máximos dos grupos dos excessos de vazão média diária acima de 400

Figura 2.5 – Gráfico de autocorrelação dos excedentes acima do limiar $u = 400 \text{ m}^3 \cdot \text{s}^{-1}$ (a) e gráficos de autocorrelação dos máximos dos grupos de excedentes acima de $400 \text{ m}^3 \cdot \text{s}^{-1}$ dos dados de vazão média diária para (b) $r = 1$, (c) $r = 2$ e (d) $r = 4$

$m^3.s^{-1}$ para $r = 1$, cujas as estimativas de máxima verossimilhança dos parâmetros σ e ξ obtidas foram $\hat{\sigma} = 161,24$ e $\hat{\xi} = -0,04$, respectivamente. Além disso, a estimativa de máxima verossimilhança do parâmetro p e do “extremal index” foram de $\hat{p} = 0,037$ e $\hat{\Theta} = 0,223$, respectivamente. A figura (2.6) apresenta os perfis de verossimilhança para a obtenção dos intervalos de 95% de confiança para os parâmetros σ e ξ dados, respectivamente por (130,0; 199,6) e (-0,18; 0,13). Analisando-os vê-se que, com 95% de probabilidade o intervalo de confiança pode ter incluso o valor de $\xi = 0$, ou seja, não há evidência estatística para se rejeitar a hipótese de nulidade de $\xi = 0$, com um nível de significância de 5%. Então, nesse caso, para efeito de comparações, foi ajustada também a distribuição exponencial aos máximos dos grupos dos excessos cuja estimativa de máxima verossimilhança do parâmetro σ obtida foi de 154,85 e cujo intervalo de 95% de confiança (139,7; 170,7) é apresentado na figura (2.6).

Figura 2.6 – Perfis de verossimilhança e função deviance para a obtenção dos intervalos de 95% de confiança (LI; LS) dos parâmetros σ , (a), e ξ , (b), da distribuição GP e do parâmetro σ , (c) da distribuição exponencial

A tabela 2.1 apresenta as estimativas dos níveis de retorno e seus respectivos in-

tervalos de 95% de confiança utilizando a distribuição generalizada de Pareto para os dados de vazão média diária. Analisando os resultados apresentados nessa tabela, vê-se que por meio do ajuste da distribuição generalizada de Pareto, espera-se que, em média, ocorra uma vazão média diária equivalente a $1149 \text{ m}^3.\text{s}^{-1}$ uma vez a cada 60 anos, valor próximo da máxima vazão média diária registrada em 59 anos, $1141,5 \text{ m}^3.\text{s}^{-1}$, o que significa que essa estimativa é consistente com os dados observados. Por sua vez, analisando a tabela 2.2, observa-se que as estimativas de níveis de retorno utilizando a distribuição exponencial é maior do que as estimativas de níveis de retorno obtidas utilizando-se a distribuição generalizada de Pareto, porém, vê-se que as amplitudes dos intervalos de 95 % de confiança para os níveis de retorno obtidos por meio do ajuste da distribuição exponencial são menores do que os obtidos por meio do ajuste da distribuição generalizada de Pareto. Provavelmente isso é devido ao fato de que estimativa do parâmetro da forma da distribuição generalizada de Pareto ser menor que zero, que por consequência indica que a distribuição generalizada de Pareto tem a cauda direita finita, e, portanto, a medida que aumenta o número de anos associado a um nível de retorno esse nível de retorno fica cada vez mais próximo do limite superior da distribuição o que acaba acarretando em um aumento da imprecisão das estimativas.

Tabela 2.1 – Estimativas dos níveis de retorno e seus respectivos intervalos de 95% de confiança utilizando a distribuição generalizada de Pareto para os dados de vazão média diária

	5 anos	10 anos	50 anos	60 anos	100 anos
Estimativas	813	910	1126	1149	1214
$I.C_{95\%}(\tau)$	(760; 890)	(841; 1032)	(997; 1423)	(1012; 1472)	(1051; 1621)
Amplitude do I.C	130	191	427	460	570

Tabela 2.2 – Estimativas dos níveis de retorno e seus respectivos intervalos de 95% de confiança utilizando a distribuição Exponencial para os dados de vazão média diária

	5 anos	10 anos	50 anos	60 anos	100 anos
Estimativas	820	928	1177	1205	1284
$I.C_{95\%}(\tau)$	(779; 867)	(876; 986)	(1102; 1263)	(1126; 1294)	(1197; 1383)
Amplitude do I.C	88	110	161	168	186

Por fim, a figura 2.7 apresenta os gráficos probabilidade - probabilidade e quantil - quantil para o ajuste das distribuições exponencial e generalizada de Pareto com envelope simulado com 95 % de confiança e número de simulação igual a 1000. Analisando essa figura observa-se um

bom ajuste dos dois modelos probabilísticos aos dados sugerindo que se deve usar a distribuição exponencial para o ajuste dos excessos de vazão média diária uma vez que é um modelo mais parcimonioso e cuja as estimativas intervalares dos níveis de retornos são mais precisas.

Figura 2.7 – Gráfico probabilidade - probabilidade para o ajuste da distribuição generalizada de Pareto (a) e distribuição exponencial (b) com envelope simulado de 95% de confiança, gráfico quantil-quantil para o ajuste da distribuição generalizada de Pareto (c) e distribuição exponencial (d) com envelope simulado de 95% de confiança em escala logarítmica

2.4 Conclusões

Diante dos resultados obtidos pode concluir que:

- Recomenda-se o uso da distribuição exponencial para ajuste dos excessos de vazão média diária observada, uma vez que as estimativas intervalares dos níveis de retorno obtidas por meio desta são mais precisas do que as obtidas por meio do ajuste da distribuição generalizada de Pareto.
- Utilizando as estimativas de níveis de retorno obtidas por meio do ajuste da distribuição exponencial, espera-se que, em média, ocorra uma vazão média diária de $1205 \text{ m}^3.\text{s}^{-1}$ uma vez a cada 60 anos para o município de Piracicaba, SP.

Referências

- ATKINSON, A. C. **Plots, Transformations and Regressions**. Oxford: Statistical Science Series; 1985. 296 p.
- AZZALINI, A. **Statistical Inference Based on Likelihood**. London: Chapman-Hall, 1996. 352 p.
- 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**, São Paulo, v.22, n. 1, p.95-111, 2004.
- CHOW, V.T. Statistical and probability analysis of hidrologic data. In: ____ **Handbook of applied hydrology** : A compendium of water-resources technology, Part I. New York: McGraw-Hill; 1964. chap. 8, p. 8.1 - 8.42.
- COLES, S.G. **An introduction to statistical modeling of extreme values**. London: Springer; 2001. 226 p.
- DAVISON, A.C.; SMITH, R.L. Models for exceedances over high thresholds. **Journal of the Royal Statistical Society, B**, London, v.520, n. 3, p.393-442, 1990.
- GROPPO, J.D.; MORAES, J.M.; BEDUSCHI, C.E.; MARTINELLI, L. A.; Análise de séries temporais de vazão e precipitação em algumas bacias do estado de São Paulo com diferentes graus de intervenções antrópicas. **Geociências**, São Paulo, v. 24, n. 2, p. 181-193, 2005.

HOLMES, J.D.; MORIATY, W.W. Application of the generalized Pareto distribution to extreme value analysis in wind engineering. **Journal of wind engineering and industrial aerodynamics**, Amsterdam, v. 83, p.1-10, 1999.

LEADBETTER, M.R.; WEISSMAN, I.; DE HAAN, L.; ROOTZÉN, H. **On clustering of high values in statistically stationary series**. Chapel Hill: Center for Stochastic Process, University of North Carolina, 1989. p. 253.

KWIATKOWSKI, D.; PHILLIPS, P.C.B.; SCHMIDT P.; SHIN, Y. Testing the null hypothesis of stationarity against the alternative of a unit root. **Journal of Econometrics**, Amsterdam, v. 54, n. 3, p. 159-178, 1992.

McNEIL, A.J.; FREY, R. estimation of tail-related risk measures of heterocedastic financial time series: an Extreme Value Approach. **Journal of Empirical Finance**, Orlando, v. 7, n. 3, p. 271-300, 2000.

MORAES, J.M.; GENOVEZ, A.M.; MORTATTI, J.; BALLESTER, M.V.; KRUSCHE, A.V.; MARTINELLI, L.A.; VICTORIA, R.L. Análise de intervenção das séries temporais de vazão dos principais rios da bacia do Rio Piracicaba. **Revista Brasileira de Recursos Hídricos**, Rio de Janeiro, v. 2, n. 2, p. 65-79, 1997.

PELLEGRINO, G.Q.; MORAES, J.M.; GUANDIQUE, E.M.; BALLESTER, M.V.; MARTINELLI, L.A.; VICTORIA, R.L. Análise Espaço Temporal de componentes hidroclimáticos na bacia do Rio Piracicaba. **Revista Brasileira de Agrometeorologia**, Santa Maria, v. 9, n. 1, p. 125-135, 2001.

NOCEDAL, J.; WRIGHT, S.J. **Numerical Optimization** New York: Springer Verlag, 1999. 636 p.

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

RUBEM, A.P.S.; **Modelagem de extremos baseados nas r - maiores estatísticas de ordem: Uma aplicação no cálculo do valor em risco em mercados emergentes**. 2006. 114 p. Dissertação (Mestrado em Estatística) - Instituto de Matemática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2006.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, ISBN 3-900051-07-0, Vienna, Austria, 2006, disponível em: <<http://www.R-project.org>>. Acesso em: 01 out. 2007.

SMITH, R.L. Statistics of extremes, with applications in environment, insurance and finance. In: B. FINKENSTADT.; H, ROOTZEN. **Extreme values in finance, telecommunications and the environment**. London: Chapman and Hall/CRC Press, 2004, chap. 1, p. 1-78.

3 MISTURAS DE DUAS DISTRIBUIÇÕES DE GUMBEL PARA ANÁLISE DOS DADOS DE VELOCIDADE MÁXIMA DO VENTO EM PIRACICABA, SP

Resumo

Modelos de misturas finitas geralmente são utilizados em situações que se suspeita que a amostra observada seja proveniente de duas ou mais populações independentes em que não é possível identificar a qual das populações pertence cada elemento dessa amostra. No entanto, esses modelos são também utilizados quando o ajuste de um componente simples dessa mistura não é capaz de descrever o fenômeno estudado de uma forma adequada permitindo uma maior flexibilidade na modelagem de populações heterogêneas. No contexto da teoria dos valores extremos, exemplos de aplicações de misturas finitas são os trabalhos de Thom (1967), Fahmi (1991), Walshaw (2000), Bortollo et al. (2005), Tartaglia et al. (2006) entre outros. Neste trabalho, foram feitos os ajustes da distribuição de Gumbel e das misturas de duas distribuições de Gumbel com mesmo e distintos parâmetros de escala aos dados de velocidades máximas de vento dos meses de abril, maio, agosto e setembro em Piracicaba, SP. Esses meses foram escolhidos pois segundo Silva e Zocchi (2005), há uma aparente bimodalidade, o que pode sugerir uma mistura de duas distribuições. Selecionou-se a distribuição que melhor ajustou-se aos dados por meio de testes de hipóteses “bootstrap” paramétrico e dos critérios AIC e BIC e avaliou-se a qualidade do ajuste por meio dos gráficos probabilidade - probabilidade e quantil - quantil. Uma vez escolhida a distribuição mais adequada para cada um dos meses mencionados, foram estimadas as probabilidades de ocorrência acima de 40, 50, 60, 70, 80, 90 e 100 $km.h^{-1}$, os níveis de retornos para os períodos de 5, 10, 50 e 100 anos e seus respectivos intervalos de confiança. Conclui-se que o modelo de duas distribuições de Gumbel com mesmo parâmetro de escala foi o modelo que melhor ajustou-se aos dados de abril e maio enquanto que a distribuição de Gumbel foi o modelo de melhor performance para os meses de agosto e setembro.

Palavras-chave: Teoria dos valores extremos; Misturas de distribuições de Gumbel; Velocidade do vento

Abstract

Finite mixture models are normally used in situations where there is a suspect that a sample arises from two independent populations and where it is not possible to identify which of two populations belongs to each element of the sample. However, these models are also used in situations where a fit of a simple component to these mixtures is unable to describe the behavior of the studied phenomena because it allows for greater flexibility in modeling a heterogeneous population. Extreme value analysis using finite mixture models has also been studied by Thom (1967), Fahmi (1991), Walshaw (2000), Bortollo et al. (2005), Tartaglia et al. (2006). In this work, we fit the Gumbel distribution and Gumbel mixture to maximum speed wind of the months of April, May, August and September in Piracicaba, SP, because for these months, according to Silva and Zocchi (2005), the dataset of maximum speed wind is apparently bimodal. Furthermore, we choose the best model to data using tests of hypothe-

ses and the AIC and BIC selection criteria and we estimate occurrence probabilities of wind with speeds above 40 to 100 $km.h^{-1}$ and return levels with their respective 95 % confidence intervals. We conclude that the mixture models of Gumbel distributions is the best model to analyze the maximum wind speed data for months of april e may and otherside the fit of Gumbel distributions was the best fit to wind speed data for months of august e september.

Keywords: Extreme value theory; Gumbel mixture; Wind speed

3.1 Introdução

Modelos de misturas finitas geralmente são utilizados em situações que se suspeita que a amostra observada seja proveniente de duas ou mais populações independentes em que não é possível identificar a qual das populações pertence cada elemento dessa amostra. Esses modelos, no entanto, são também utilizados quando o ajuste de um componente simples dessa mistura não é capaz de descrever o fenômeno estudado de uma forma adequada permitindo uma maior flexibilidade na modelagem de populações heterogêneas (McLachlan e Peel, 2000). Exemplos de utilização desses modelos podem ser encontrados na literatura como os trabalhos de Rao (1948), Day (1969), Zocchi(1993), Titterton et al. (1985), Lin et al. (2007 a) entre outros.

No contexto da teoria dos valores extremos, que segundo Coles (2001) tem como objetivo descrever o comportamento aleatório de extremos (máximos ou mínimos) de seqüências de variáveis aleatórias, a aplicação de modelos de misturas finitas vem sendo utilizada desde o final da década de 60. Thom (1967) ajustou um modelo de duas distribuições de Fréchet para modelar dados de velocidade máxima do vento. Tartaglia et al. (2006) ajustaram uma mistura de duas distribuições de Gumbel aos dados de precipitação máxima pluvial em Toscana, Itália, com a justificativa de que esses dados previnham ou de uma população de precipitações pluviais de menor magnitude e mais freqüentes, oriundas de condições climáticas locais, ou de uma segunda uma população, de precipitações mais extremas e menos freqüentes, oriundas de fenômenos climáticos extremos, que abrangiam o país como um todo. Ainda nesse contexto, porém sob enfoque Bayesiano, Walshaw (2000) ajustou um modelo de mistura de duas distribuições generalizadas de valores extremos aos dados de velocidades máximas de ventos da cidade Boston, Massachusetts e de Key West, Flórida, EUA. Como premissa, considerou que os ventos máximos eram oriundos ou de tempestades tropicais, geralmente de maior magnitude e menos freqüentes, ou de tempestades frontais, geralmente mais amenos porém mais freqüentes. Outros trabalhos encontrados na literatura sobre modelos de misturas finitas para análise de valores extremos são os trabalhos de Fahmi e Abrasi (1987), Fahmi e Abrasi (1991), Bortollo et al. (2005).

Entretanto, independente de qual modelo de mistura finita for utilizado, um problema clássico é a determinação do número de componentes, principalmente quando esse número

for menor do que a do modelo proposto, pois nesses casos, os verdadeiros valores dos parâmetros da mistura proposta estão na borda do espaço paramétrico. Além disso, a hipótese nula de que a variável aleatória segue apenas um dos componentes é um subconjunto não identificável do espaço paramétrico impossibilitando o uso do teste de razão de verossimilhanças da forma tradicional.

Como alternativa para testar essa hipótese, McLachlan (1987) propôs a realização de testes de razão de verossimilhanças baseados na metodologia “bootstrap” paramétrico e utilizou esse teste para determinar o número de componentes de um modelo de mistura de distribuições normais. Em linhas gerais, segundo Efron e Tibshirani (1993), o “bootstrap” é um método de reamostragem computacionalmente intensivo que pode ser utilizado para estimar a distribuição de probabilidade de um estimador de interesse, quando a mesma é intratável por meio dos métodos tradicionais. Assim, essa metodologia pode ser uma alternativa adequada para a construção de intervalos de confiança e de testes de hipótese dos parâmetros de interesse quando os mesmos não podem ser construídos por meio de metodologias baseadas em teorias assintóticas.

Num estudo comparativo entre a metodologia de McLachlan (1987), e outras apresentadas na literatura, Feng e McCulloch (1994) concluíram que o teste de razão de verossimilhanças baseado na metodologia “bootstrap” paramétrico é o teste de melhor performance para testar o número de componentes de mistura de distribuições normais com distintas variâncias. Em contrapartida, Titterton (1990) questionou sobre a validade dos testes de razão de verossimilhanças “bootstrap” para a situação em que a hipótese nula de que dados seguem apenas um componente da mistura seja verdadeira, porque, neste caso, sob a hipótese alternativa, o estimador de máxima verossimilhança não é consistente.

Por outro lado, com o intuito de justificar o uso dos teste de razão de verossimilhanças “bootstrap” no contexto de ajuste de misturas finitas, Feng e McCulloch (1996) generalizaram o conceito de estimador consistente. Essa generalização consistiu em demonstrar que quando o verdadeiro parâmetro da mistura está contido em um subconjunto não identificável que pode estar na borda do espaço paramétrico, o estimador de máxima verossimilhança dos parâmetros das misturas convergem para um elemento que é função desse estimador e que também pertence a esse subconjunto não identificável.

Neste trabalho, foram feitos os ajustes da distribuição de Gumbel e das misturas de duas distribuições de Gumbel com mesmo parâmetro de escala e com parâmetros de escala distintos aos dados de velocidades máximas de vento de Piracicaba, dos meses de abril, maio, agosto e setembro. Esses meses foram escolhidos pois segundo Silva e Zocchi (2005), há uma aparente bimodalidade, o que pode sugerir uma mistura de duas distribuições. Quanto à seleção entre esses modelos probabilísticos propõe-se utilizar o teste de razão de verossimilhanças baseado na metodologia “bootstrap” paramétrico apresentado por Feng e McCulloch (1996).

3.2 Desenvolvimento

Os dados de velocidades máximas mensais de vento a 10 m acima do solo foram obtidos no período de 1956 a 1971 e de 1974 a 2007, por meio de registros de anemógrafo do tipo universal, marca Fuess, localizado no posto agrometeorológico da Escola Superior de Agricultura “Luiz de Queiroz”, em Piracicaba (latitude 22°42'30"S, longitude 47°30'00"W, e altitude 545 m), SP, Brasil. Esses dados estão disponíveis na página <http://www.lce.esalq.usp.br/postocon.html>.

Convém ressaltar que, para cada ano, a série de dados foi subdividida em 12 meses, obtendo-se, assim, uma série de dados para cada mês. Nesse trabalho, no entanto, foram analisadas apenas as séries de dados referentes aos meses de abril, maio, agosto e setembro.

A metodologia utilizada para a análise dos dados citados num dado mês é apresentada a seguir.

3.2.1 Distribuição de Gumbel

Seja X a variável aleatória velocidade máxima de vento. Define-se a função de distribuição acumulada de Gumbel por meio de

$$F(x) = \exp \left\{ - \exp \left[- \left(\frac{x - \mu}{\sigma} \right) \right] \right\},$$

em que μ e σ são, respectivamente, os parâmetros de localização e escala, sendo $-\infty < \mu < \infty$, $\sigma > 0$ e $-\infty < x < \infty$, cuja função densidade de probabilidade é dada por

$$f(x) = \frac{1}{\sigma} \exp \left[- \left(\frac{x - \mu}{\sigma} \right) \right] \exp \left\{ - \exp \left[- \left(\frac{x - \mu}{\sigma} \right) \right] \right\}, \quad (3.1)$$

com esperança matemática, $E(X) = \mu + \sigma\gamma$, e variância $Var(X) = \frac{\sigma^2\pi^2}{6}$, sendo γ a constante de Euler, ($\gamma \approx 0,577216$).

3.2.2 Misturas de distribuições de Gumbel

Considere, alternativamente, que a variável X de interesse tenha distribuição dada por uma mistura de duas distribuições de Gumbel. Tem-se nesse caso, que a função densidade de probabilidade é dada por

$$g_1(x) = pf_1(x) + (1 - p)f_2(x) \quad (3.2)$$

sendo $0 < p < 1$ e $f_1(x)$ e $f_2(x)$ as funções densidade de probabilidade de Gumbel com parâmetros μ_1 e σ_1 , e μ_2 e σ_2 , respectivamente.

O modelo (3.2) pode também ser explicado considerando que uma amostra aleatória $\mathbf{X} = \{X_1, \dots, X_n\}$ de X seja proveniente de duas populações, Ψ_1 e Ψ_2 em que não é possível identificar a qual das duas populações pertence cada elemento dessa amostra.

Nesse caso, para obter-se a distribuição de probabilidade da variável aleatória X , inicialmente, pressupõe-se a existência dos eventos $X_i \in \Psi_1$ e $X_i \in \Psi_2$ com probabilidades p e $(1 - p)$, respectivamente.

Define-se uma variável aleatória não observável $Z \in \{0, 1\} \sim \text{Bernoulli}(p)$ associada ao espaço amostral desses eventos tal que

$$\begin{cases} X|Z = 1 \sim \text{Gumbel}(\mu_1, \sigma_1) & \text{e} \\ X|Z = 0 \sim \text{Gumbel}(\mu_2, \sigma_2). \end{cases}$$

Logo em seguida, obtém-se a distribuição condicional de $X|Z = z$, $g_{1X|Z}(x)$, que é dada por

$$g_{1X|Z}(x) = f_1(x)^z f_2(x)^{(1-z)}$$

sendo $f_1(x)$ e $f_2(x)$ a função densidade de probabilidade de Gumbel com parâmetros de locação μ_1, μ_2 e de escala σ_1, σ_2 respectivamente. Na sequência, obtém-se a distribuição conjunta de X e Z por meio de

$$g_1(x, z) = [f_1(x)^z f_2(x)^{(1-z)}] p^z (1 - p)^{(1-z)}$$

e finalmente, define-se o modelo de mistura de duas distribuições de Gumbel como sendo a distribuição marginal de X , ou seja,

$$g_1(x) = \sum_{z=0}^1 [f_1(x)^z f_2(x)^{(1-z)}] p^z (1 - p)^{(1-z)} = p f_1(x) + (1 - p) f_2(x) \quad (3.3)$$

que é a mesma expressão dada em (3.2).

Ainda vale destacar que para $\sigma_1 = \sigma_2$ tem-se o caso particular modelo de mistura de duas distribuições de Gumbel com mesmo parâmetro de escala que será denotado por $g_2(x)$.

3.2.3 Teste de aleatoriedade

Seguindo o procedimento adotado também por Bautista et al. (2004), a primeira etapa da análise é verificar a hipótese de independência dos dados observados por meio do teste de chorrilho (“run test”) descrito por Zar (1999).

Em linhas gerais, esse teste consiste em, inicialmente, definir uma sequência dicotômica de tamanho n , a partir da amostra aleatória $\{X_1, \dots, X_n\}$ aplicando a cada X_i a função indicadora $A(x_i) = I_{X_i > M_d}(X_i)$ e omitindo os casos em que $X_i = M_d$ sendo M_d a mediana dos dados.

Uma vez feito isso, define-se a variável aleatória número total de subsequências de zeros e uns ao longo da amostra R , cujo valor observado é denotado por r e define-se as variáveis aleatórias N_1 como sendo o número total de ocorrências de $X_i > M_d$ e N_2 como sendo o número total de ocorrências de $X_i < M_d$, cujo valores observados são n_1 e n_2 , respectivamente.

Para $n_1 < 30$ e $n_2 < 30$, Zar (1999) apresenta os pares de valores críticos exatos $(r_{1,\alpha,n_1,n_2}, r_{2,\alpha,n_1,n_2})$ ao nível de significância de $100 \times \alpha\%$, usando o procedimento descrito por Eisenhart e Swed (1943) e Browlee (1965). Assim, rejeita-se a hipótese nula se $r \leq r_{1,\alpha,n_1,n_2}$, ou se $r \leq r_{2,\alpha,n_1,n_2}$. Caso $n_1 \geq 30$ ou $n_2 \geq 30$, sob a hipótese H_0 de independência tem-se que, assintoticamente R segue distribuição normal com esperança dada por

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

e variância dada por

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

cujas estimativas são dadas por

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

e

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

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

Então para efetuar-se o teste deve-se calcular a mediana da amostra observada de velocidade máxima de vento $\{x_1, \dots, x_n\}$, obter a sequência dicotômica da amostra observada $A(x_1), \dots, A(x_n)$, n_1 , n_2 , r , $\hat{E}(R)$ e $\hat{Var}(R)$ e, por fim, calcular o valor-p por meio de

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

em que ϕ_r é o quantil de ordem $100 \times \frac{\alpha}{2}\%$ da normal padrão e $100 \times \alpha\%$ é o nível de significância adotado para o teste.

Uma vez testada a independência das observações parte-se para a estimação dos parâmetros.

3.2.4 Estimação dos parâmetros da distribuição de Gumbel e seus respectivos intervalos de confiança

Considerando-se, inicialmente, a distribuição de Gumbel como modelo adotado, tem-se que a função de verossimilhança é dada por

$$L_1(\boldsymbol{\theta}) = L_1(\mu, \sigma) = \prod_{i=1}^n \frac{1}{\sigma} \exp \left[- \left(\frac{x_i - \mu}{\sigma} \right) \right] \exp \left\{ - \exp \left[- \left(\frac{x_i - \mu}{\sigma} \right) \right] \right\}$$

cujo logaritmo é

$$l_1(\boldsymbol{\theta}) = l_1(\mu, \sigma) = -n \log \sigma - \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right) - \sum_{i=1}^n \exp \left\{ - \left(\frac{x_i - \mu}{\sigma} \right) \right\}. \quad (3.4)$$

Dessa forma, para obter as estimativas de máxima verossimilhança de $\boldsymbol{\theta} = [\mu, \sigma]^\top$ deve-se achar o ponto de máximo de (3.4) em relação μ e σ . Para isso, soluciona-se o sistema de equações formado pelas derivadas primeira de (3.4) em relação a μ e σ , igualadas a zero, ou seja

$$\nabla l_1(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial l_1(\mu, \sigma)}{\partial \mu} \\ \frac{\partial l_1(\mu, \sigma)}{\partial \sigma} \end{bmatrix} = \begin{bmatrix} \frac{n}{\hat{\sigma}} - \frac{1}{\hat{\sigma}} \sum_{i=1}^n \exp \left\{ - \left(\frac{x_i - \hat{\mu}}{\hat{\sigma}} \right) \right\} \\ -\frac{n}{\hat{\sigma}} + \sum_{i=1}^n \frac{(x_i - \hat{\mu})}{\hat{\sigma}^2} - \sum_{i=1}^n \frac{(x_i - \hat{\mu})}{\hat{\sigma}^2} \exp \left\{ - \left(\frac{x_i - \hat{\mu}}{\hat{\sigma}} \right) \right\} \end{bmatrix} = \mathbf{0} \quad (3.5)$$

Uma vez que não há uma forma analítica para resolver o sistema de equações dado por (3.5), utilizou-se, para maximizar (3.4), o algoritmo “BFGS”, descrito por Nocedal e Wright (1999) e apresentado a seguir.

Dado um valor inicial $\boldsymbol{\theta}^{(0)}$ para $\boldsymbol{\theta}$ e, seguindo recomendações de Nocedal e Wright (1999), um valor inicial da inversa da matriz Hessiana $\mathbf{H}^{(0)} = \mathbf{I}$, em que $\mathbf{I}$ seja a matriz identidade, o algoritmo “BFGS” consiste em

1 Obter

$$\mathbf{q}^{(k)} = -\mathbf{H}^{(k)} \nabla l_1(\boldsymbol{\theta}^{(k)})$$

2 Obter

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \phi^{(k)} \mathbf{q}^{(k)},$$

sendo $\phi^{(k)}$ uma variável denominada tamanho do passo cujo o valor para $k = 0$ deve ser igual a 1 e, caso contrário, deve satisfazer as seguintes condições

$$\begin{aligned} l_1(\boldsymbol{\theta}^{(k)} + \phi^{(k)} \mathbf{q}^{(k)}) &\leq l_1(\boldsymbol{\theta}^{(k)}) + c_1 \alpha^{(k)} \nabla^\top l_1(\boldsymbol{\theta}^{(k)}) \mathbf{q}^{(k)}, \\ |l_1^\top(\boldsymbol{\theta}^{(k+1)}) \mathbf{q}^{(k)}| &\leq c_2 |\nabla^\top l_1(\boldsymbol{\theta}^{(k)}) \mathbf{q}^{(k)}|, \end{aligned} \quad (3.6)$$

em que $0 < c_1 < c_2 < 1$ são constantes pré especificadas que por recomendação de Nocedal e Wright (1999), utiliza-se $c_1 = 10^{-4}$, $c_2 = 10^{-9}$.

3 Obter $\mathbf{y}^{(k)} = \nabla l_1(\boldsymbol{\theta}^{(k+1)}) - \nabla l_1(\boldsymbol{\theta}^{(k)})$

4 Obter $\mathbf{e}^{(k)} = \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}$

5 Obter

$$\mathbf{H}^{(k+1)} = \left(\mathbf{I} - \rho^{(k)} \mathbf{e}^{(k)} \mathbf{y}^{(k)\top} \right) \mathbf{H}^{(k)} \left(\mathbf{I} - \rho^{(k)} \mathbf{y}^{(k)} \mathbf{e}^{(k)\top} \right) + \rho^{(k)} \mathbf{e}^{(k)} \mathbf{e}^{(k)\top}$$

$$\text{em que } \rho^{(k)} = \frac{1}{\mathbf{y}^{(k)\top} \mathbf{e}^{(k)}}$$

6 Verificar se $l_1(\boldsymbol{\theta}^{(k+1)}) - l_1(\boldsymbol{\theta}^{(k)}) \geq \epsilon$. Em caso afirmativo, parar, caso contrário, fazer $k = k + 1$ e voltar ao passo 1.

Além disso, para obtenção do vetor $\boldsymbol{\theta}^{(0)}$ utilizou-se o método dos momentos. A obtenção de estimadores de parâmetros por meio desse método consiste em resolver um sistema de equações formado igualando-se os momentos teóricos com os momentos amostrais. Para o caso da distribuição de Gumbel tem-se o seguinte sistema de equações

$$\begin{cases} E(X) = \mu + \sigma\gamma = \bar{x} \\ Var(X) = \frac{\sigma^2\pi^2}{6} = s^2 \end{cases}$$

cujas soluções são dadas por

$$\sigma_0 = \frac{\sqrt{6}}{\pi} s \approx 0,77970s \quad (3.7)$$

e por

$$\mu_0 = \bar{x} - \gamma \frac{\sigma}{\pi} \quad (3.8)$$

em que $\bar{x}$ e s^2 são a média e a variância amostrais, respectivamente.

Uma vez obtidas as estimativas pontuais, pode-se construir os intervalos de $100 \times (1 - \alpha)\%$ de confiança para os parâmetros μ e σ da distribuição de Gumbel, utilizando o método do perfil da verossimilhança.

Esse método consiste em estabelecer uma gama de valores do parâmetro para o qual deseja obter o intervalo e, para cada um desses valores fixados, substituí-lo em (3.4) e maximizá-la em relação ao outro parâmetro não fixado.

Dessa forma, o intervalo de $100 \times (1 - \alpha)\%$ para o parâmetro μ é definido da seguinte forma

$$I.C_{(100 \times (1 - \alpha)\%)}(\mu) = \left\{ \mu : 2 [l(\hat{\mu}, \hat{\sigma}) - l(\mu, \sigma)] \leq \chi_{1, \{1 - \alpha\}}^2 \right\}$$

e o intervalo de $100 \times (1 - \alpha)\%$ para o parâmetro σ é dado por

$$I.C_{(100 \times (1 - \alpha)\%)}(\sigma) = \left\{ \sigma : 2 [l(\hat{\mu}, \hat{\sigma}) - l(\mu, \sigma)] \leq \chi_{1, \{1 - \alpha\}}^2 \right\}$$

em que μ_0 é um valor fixo de μ , σ_0 é um valor fixo de σ , e $\chi^2_{\{1, (1-\alpha)\}}$ é o quantil de ordem $100 \times (1 - \alpha)\%$ da distribuição qui-quadrado com 1 grau de liberdade.

3.2.5 Estimação dos parâmetros das misturas de distribuições de Gumbel e seus respectivos intervalos de confiança

Para fazer inferência a respeito dos parâmetros do modelo de mistura de duas distribuições de Gumbel com distintos parâmetros de escala tem-se que a função de verossimilhança é dada por

$$L_2(\boldsymbol{\theta}) = L_2(\mu_1, \mu_2, \sigma_1, \sigma_2, p) = \prod_{i=1}^n \left[p \frac{1}{\sigma_1} \exp \left[- \left(\frac{x_i - \mu_1}{\sigma_1} \right) \right] \exp \left\{ - \exp \left[- \left(\frac{x_i - \mu_1}{\sigma_1} \right) \right] \right\} + (1 - p) \frac{1}{\sigma_2} \exp \left[- \left(\frac{x_i - \mu_2}{\sigma_2} \right) \right] \exp \left\{ - \exp \left[- \left(\frac{x_i - \mu_2}{\sigma_2} \right) \right] \right\} \right]$$

cujo o logaritmo é

$$l_2(\boldsymbol{\theta}) = l_2(\mu_1, \mu_2, \sigma_1, \sigma_2, p) = \sum_{i=1}^n \log \left[p \frac{1}{\sigma_1} \exp \left[- \left(\frac{x_i - \mu_1}{\sigma_1} \right) \right] \exp \left\{ - \exp \left[- \left(\frac{x_i - \mu_1}{\sigma_1} \right) \right] \right\} + (1 - p) \frac{1}{\sigma_2} \exp \left[- \left(\frac{x_i - \mu_2}{\sigma_2} \right) \right] \exp \left\{ - \exp \left[- \left(\frac{x_i - \mu_2}{\sigma_2} \right) \right] \right\} \right] \quad (3.9)$$

Sendo assim, para obtenção das estimativas de máxima verossimilhança dos parâmetros, assim como já foi visto anteriormente, deve-se solucionar o sistema de equações obtido pelas derivadas primeira de (3.9) em relação a $\mu_1, \sigma_1, \mu_2, \sigma_2$ e p igualadas a zero, ou seja,

$$\left\{ \begin{array}{l} \frac{\partial l_2(\boldsymbol{\theta})}{\partial \mu_1} = \sum_{i=1}^n \frac{\frac{\hat{p}}{\hat{\sigma}_1} \hat{f}_1(\hat{\mu}_1, \hat{\sigma}_1) \left[1 - \exp \left\{ - \left(\frac{x_i - \hat{\mu}_1}{\hat{\sigma}_1} \right) \right\} \right]}{\hat{g}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)} = 0 \\ \frac{\partial l_2(\boldsymbol{\theta})}{\partial \mu_2} = \sum_{i=1}^n \frac{\frac{(1 - \hat{p})}{\hat{\sigma}_2} \hat{f}_2(\hat{\mu}_2, \hat{\sigma}_2) \left[1 - \exp \left\{ - \left(\frac{x_i - \hat{\mu}_2}{\hat{\sigma}_2} \right) \right\} \right]}{\hat{g}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)} = 0 \\ \frac{\partial l_2(\boldsymbol{\theta})}{\partial \sigma_1} = \sum_{i=1}^n \frac{\frac{\hat{p}}{\hat{\sigma}_1} \hat{f}_1(\hat{\mu}_1, \hat{\sigma}_1) \left[-1 + \frac{(x_i - \hat{\mu}_1)}{\hat{\sigma}_1} - \exp \left\{ - \left(\frac{x_i - \hat{\mu}_1}{\hat{\sigma}_1} \right) \right\} \right] \frac{(x_i - \hat{\mu}_1)}{\hat{\sigma}_1}}{\hat{g}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)} = 0 \\ \frac{\partial l_2(\boldsymbol{\theta})}{\partial \sigma_2} = \sum_{i=1}^n \frac{\frac{(1 - \hat{p})}{\hat{\sigma}_2} \hat{f}_2(\hat{\mu}_2, \hat{\sigma}_2) \left[-1 + \frac{(x_i - \hat{\mu}_2)}{\hat{\sigma}_2} - \exp \left\{ - \left(\frac{x_i - \hat{\mu}_2}{\hat{\sigma}_2} \right) \right\} \right] \frac{(x_i - \hat{\mu}_2)}{\hat{\sigma}_2}}{\hat{g}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)} = 0 \\ \frac{\partial l_2(\boldsymbol{\theta})}{\partial p} = \sum_{i=1}^n \frac{\hat{f}_1(\hat{\mu}_1, \hat{\sigma}_1) - \hat{f}_2(\hat{\mu}_2, \hat{\sigma}_2)}{\hat{g}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)} = 0 \end{array} \right. \quad (3.10)$$

em que

$$\hat{f}_1(\hat{\mu}_1, \hat{\sigma}_1) = \frac{1}{\hat{\sigma}_1} \exp \left\{ - \left(\frac{x_i - \hat{\mu}_1}{\hat{\sigma}_1} \right) \right\} \exp \left\{ - \exp \left\{ - \left(\frac{x_i - \hat{\mu}_1}{\hat{\sigma}_1} \right) \right\} \right\},$$

$$\hat{f}_2(\hat{\mu}_2, \hat{\sigma}_2) = \frac{1}{\hat{\sigma}_2} \exp \left\{ - \left(\frac{x_i - \hat{\mu}_2}{\hat{\sigma}_2} \right) \right\} \exp \left\{ - \exp \left\{ - \left(\frac{x_i - \hat{\mu}_2}{\hat{\sigma}_2} \right) \right\} \right\}$$

e

$$\hat{g}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, p) = \hat{p} \hat{f}_1(\hat{\mu}_1, \hat{\sigma}_1) + (1 - \hat{p}) \hat{f}_2(\hat{\mu}_2, \hat{\sigma}_2).$$

Uma vez que também não há solução analítica para (3.10), então para a obtenção das estimativas do parâmetro da mistura de duas distribuições de Gumbel com distintos parâmetros de escala, neste trabalho foi utilizado o algoritmo ECM proposto primeiramente por Meng e Rubin (1993) que é uma variante do algoritmo EM proposto por Dempster et al. (1977).

Para a implementação do algoritmo ECM primeiramente define-se a função de verossimilhança da distribuição conjunta de X e Z, denominada função de verossimilhança de “dados completos”, dada por

$$\begin{aligned} L_{2c}(\boldsymbol{\theta}) &= L_{2c}(\mu_1, \mu_2, \sigma_1, \sigma_2, p) = p^{\sum_{i=1}^n z_i} \frac{1}{\sigma_1^{\sum_{i=1}^n z_i}} \exp \left\{ - \sum_{i=1}^n z_i \left(\frac{x_i - \mu_1}{\sigma_1} \right) \right\} \times \\ &\exp \left\{ - \sum_{i=1}^n z_i \exp \left\{ - \frac{x_i - \mu_1}{\sigma_1} \right\} \right\} (1 - p)^{(n - \sum_{i=1}^n z_i)} \frac{1}{\sigma_2^{(n - \sum_{i=1}^n z_i)}} \times \\ &\exp \left\{ - \sum_{i=1}^n (1 - z_i) \left(\frac{x_i - \mu_2}{\sigma_2} \right) \right\} \exp \left\{ - \sum_{i=1}^n (1 - z_i) \exp \left\{ - \frac{x_i - \mu_2}{\sigma_2} \right\} \right\} \end{aligned}$$

cujo logaritmo é

$$\begin{aligned} l_{2c}(\boldsymbol{\theta}) &= l_{2c}(\mu_1, \mu_2, \sigma_1, \sigma_2, p) = \sum_{i=1}^n z_i \log p - \sum_{i=1}^n z_i \log \sigma_1 - \sum_{i=1}^n z_i \left(\frac{x_i - \mu_1}{\sigma_1} \right) \\ &- \sum_{i=1}^n z_i \exp \left\{ - \left(\frac{x_i - \mu_1}{\sigma_1} \right) \right\} + (n - \sum_{i=1}^n z_i) \log(1 - p) - (n - \sum_{i=1}^n z_i) \log \sigma_2 \\ &- \sum_{i=1}^n (1 - z_i) \left(\frac{x_i - \mu_2}{\sigma_2} \right) - \sum_{i=1}^n (1 - z_i) \exp \left\{ - \left(\frac{x_i - \mu_2}{\sigma_2} \right) \right\} \end{aligned} \quad (3.11)$$

em que $z_{1i} = z_i$ e $z_{2i} = 1 - z_{1i}$.

O algoritmo ECM é composto por dois passos, o primeiro denominado passo E, consiste em dada uma iteração k e um vetor parâmetros $\boldsymbol{\theta}^{(k)} = [\mu_1^{(k)}, \mu_2^{(k)}, \sigma_1^{(k)}, \sigma_2^{(k)}, p^{(k)}]^\top$,

obter a esperança matemática do logaritmo da verossimilhança dos dados completos condicionada a $\mathbf{X} = \mathbf{x}$, ou seja,

$$E_Z [l_{2c}(\mu_1, \mu_2, \sigma_1, \sigma_2, p) | \mathbf{X} = \mathbf{x}] = \sum_{j=1}^2 \left[\log p_j \sum_{i=1}^n E_Z [Z_{ji} | \mathbf{X} = \mathbf{x}]^{(k)} - \log \sigma_j \sum_{i=1}^n E_Z [Z_{ji} | \mathbf{X} = \mathbf{x}]^{(k)} - \sum_{i=1}^n E_Z [Z_{ji} | \mathbf{X} = \mathbf{x}]^{(k)} \left(\frac{x_i - \mu_j}{\sigma_j} \right) - \sum_{i=1}^n E_Z [Z_{ji} | \mathbf{X} = \mathbf{x}]^{(k)} \exp \left\{ - \left(\frac{x_i - \mu_j}{\sigma_j} \right) \right\} \right] \quad (3.12)$$

Então para obter-se $E_Z [Z_{ji} | \mathbf{X} = \mathbf{x}]^{(k)}$, considera-se que $Z_{2i} = 1 - Z_{1i}$, $Z_{11}, \dots, Z_{1n}$ e $X_1, \dots, X_n$ são independentes e identicamente distribuídos. Assim, assumindo $Z_i \sim \text{Bernoulli}(p)$ tem-se que

$$\begin{aligned} E_Z [Z_{1i} | \mathbf{X} = \mathbf{x}]^{(k)} &= Pr(Z_{1i} = 1 | X_i = x_i) \\ &= \frac{Pr(Z_{1i} = 1, X_i = x_i)}{Pr(X_i = x_i)} = \frac{Pr(Z_{1i} = 1, x_i < X_i \leq x_i + \Delta x_i)}{Pr(x_i < X_i \leq x_i + \Delta x_i)} \\ &= \lim_{\Delta x_i \rightarrow 0} \frac{Pr(Z_{1i} = 1, x_i < X_i \leq x_i + \Delta x_i)}{Pr(x_i < X_i \leq x_i + \Delta x_i)} = \frac{p^{(k)} f_1(x_i)^{(k)}}{g_1(x_i)^{(k)}} \end{aligned}$$

em que

$$f_1(x_i)^{(k)} = \frac{1}{\sigma_1^{(k)}} \exp \left[- \left(\frac{x - \mu_1^{(k)}}{\sigma_1^{(k)}} \right) \right] \exp \left\{ - \exp \left[- \left(\frac{x - \mu_1^{(k)}}{\sigma_1^{(k)}} \right) \right] \right\}$$

e

$$g_1(x)^{(k)} = \sum_{j=1}^2 \frac{1}{\sigma_j^{(k)}} \exp \left[- \left(\frac{x - \mu_j^{(k)}}{\sigma_j^{(k)}} \right) \right] \exp \left\{ - \exp \left[- \left(\frac{x - \mu_j^{(k)}}{\sigma_j^{(k)}} \right) \right] \right\}.$$

A segunda etapa do algoritmo, denominado passo CM consiste em particionar o vetor de parâmetros $\boldsymbol{\theta} = [\mu_1, \mu_2, \sigma_1, \sigma_2, p]^T$ em três subvetores $[p]$, $[\mu_1, \sigma_1]^T$ e $[\mu_2, \sigma_2]^T$ e obter as estimativas dos parâmetros da seguinte forma

Primeiramente a obtenção da estimativa do parâmetro p na iteração $(k+1)$ é dada por

$$\frac{\partial E_Z [l_{2c}(\mu_1, \mu_2, \sigma_1, \sigma_2, p) | \mathbf{X} = \mathbf{x}]}{\partial p} = \frac{\sum_{i=1}^n E [Z_{1i} | \mathbf{X} = \mathbf{x}]^{(k)}}{p^{(k+1)}} - \frac{\sum_{i=1}^n E [Z_{2i} | \mathbf{X} = \mathbf{x}]^{(k)}}{(1 - p^{(k+1)})}. \quad (3.13)$$

Então igualando (3.13) a zero tem-se

$$\begin{aligned} \frac{\sum_{i=1}^n E [Z_{1i} | \mathbf{X} = \mathbf{x}]^{(k)}}{p^{(k+1)}} &= \frac{\sum_{i=1}^n 1 - E [Z_{1i} | \mathbf{X} = \mathbf{x}]^{(k)}}{(1 - p^{(k+1)})} \\ p^{(k+1)} &= \frac{1}{n} \sum_{i=1}^n E [Z_{1i} | \mathbf{X} = \mathbf{x}]^{(k)}. \end{aligned} \quad (3.14)$$

Logo em seguida, para estimar os outros parâmetros procede-se da seguinte forma

Fixando os valores de $p = p^{(k+1)}$, $\mu_2 = \mu_2^{(k)}$, $\sigma_2 = \sigma_2^{(k)}$ obtém-se os valores de $\mu_1^{(k+1)}$ e $\sigma_1^{(k+1)}$ por meio de

$$\left[\mu_1^{(k+1)}, \sigma_1^{(k+1)} \right] = \arg \max_{(\mu_1^{(k+1)}, \sigma_1^{(k+1)})} E_Z [l_{2c}(\mu_1, \mu_2, \sigma_1, \sigma_2, p) | \mathbf{X} = \mathbf{x}] \quad (3.15)$$

e de modo análogo, fixando os valores de $p = p^{(k+1)}$, $\mu_1 = \mu_1^{(k+1)}$, $\sigma_1 = \sigma_1^{(k+1)}$ obtém-se os valores de $\mu_2^{(k+1)}$ e $\sigma_2^{(k+1)}$ por meio de

$$\left[\mu_2^{(k+1)}, \sigma_2^{(k+1)} \right] = \arg \max_{(\mu_2^{(k+1)}, \sigma_2^{(k+1)})} E_Z [l_{2c}(\mu_1, \mu_2, \sigma_1, \sigma_2, p) | \mathbf{X} = \mathbf{x}] \quad (3.16)$$

vale comentar que para maximização de (3.15) e (3.16) foi utilizado o algoritmo “BFGS”.

E assim repete-se o passo E e o passo CM enquanto

$$|l_{2c}(\mu_1^{(k+1)}, \mu_2^{(k+1)}, \sigma_1^{(k+1)}, \sigma_2^{(k+1)}, p^{(k+1)}) - l_{2c}(\mu_1^{(k)}, \mu_2^{(k)}, \sigma_1^{(k)}, \sigma_2^{(k)}, p^{(k)})| \geq \epsilon$$

sendo ϵ uma tolerância admitida.

Para o caso particular em que $\sigma_1 = \sigma_2$ o logaritmo da verossimilhança é dado por

$$l_3(\boldsymbol{\theta}) = l_3(\mu_1, \mu_2, \sigma, p) = \sum_{i=1}^n \log \left[p \frac{1}{\sigma} \exp \left[- \left(\frac{x_i - \mu_1}{\sigma} \right) \right] \exp \left\{ - \exp \left[- \left(\frac{x_i - \mu_1}{\sigma} \right) \right] \right\} + \right. \\ \left. (1 - p) \frac{1}{\sigma} \exp \left[- \left(\frac{x_i - \mu_2}{\sigma} \right) \right] \exp \left\{ - \exp \left[- \left(\frac{x_i - \mu_2}{\sigma} \right) \right] \right\} \right]$$

cujo sistema de equações é dado por

$$\left\{ \begin{array}{l} \frac{\partial l_3(\mu_1, \mu_2, \sigma, p)}{\partial \mu_1} = \sum_{i=1}^n \frac{\frac{p}{\hat{\sigma}} \hat{f}_1(\hat{\mu}_1, \hat{\sigma}) \left[1 - \exp \left\{ - \left(\frac{x_i - \hat{\mu}_1}{\hat{\sigma}} \right) \right\} \right]}{\hat{g}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma})} = 0 \\ \frac{\partial l_3(\mu_1, \mu_2, \sigma, p)}{\partial \mu_2} = \sum_{i=1}^n \frac{\frac{(1-p)}{\hat{\sigma}} \hat{g}_j(\hat{\mu}_j, \hat{\sigma}) \left[1 - \exp \left\{ - \left(\frac{x_i - \hat{\mu}_j}{\hat{\sigma}} \right) \right\} \right]}{\hat{g}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma})} = 0 \\ \frac{\partial l_3(\mu_1, \mu_2, \sigma, p)}{\partial \sigma} = \sum_{i=1}^n \left[\frac{\frac{p}{\hat{\sigma}} \hat{f}_1(\mu_1, \sigma) \left[-1 + \frac{(x_i - \hat{\mu}_1)}{\hat{\sigma}} - \exp \left\{ - \frac{x_i - \hat{\mu}_1}{\hat{\sigma}} \right\} \frac{(x_i - \hat{\mu}_1)}{\hat{\sigma}} \right]}{\hat{g}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma})} \right] + \\ \sum_{i=1}^n \left[\frac{\frac{(1-p)}{\hat{\sigma}} \hat{f}_2(\mu_2, \sigma) \left[-1 + \frac{(x_i - \hat{\mu}_2)}{\hat{\sigma}} - \exp \left\{ - \frac{x_i - \hat{\mu}_2}{\hat{\sigma}} \right\} \frac{(x_i - \hat{\mu}_2)}{\hat{\sigma}} \right]}{\hat{g}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma})} \right] = 0 \\ \frac{\partial l_3(\mu_1, \mu_2, \sigma, p)}{\partial p} = \sum_{i=1}^n \frac{\hat{f}_1(\hat{\mu}_1, \hat{\sigma}) - \hat{f}_2(\hat{\mu}_2, \hat{\sigma})}{\hat{g}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma})} = 0 \end{array} \right.$$

em que

$$\hat{f}_1(\hat{\mu}_1, \hat{\sigma}) = \frac{1}{\hat{\sigma}} \exp \left\{ - \left(\frac{x_i - \hat{\mu}_1}{\hat{\sigma}} \right) \right\} \exp \left\{ - \exp \left\{ - \left(\frac{x_i - \hat{\mu}_1}{\hat{\sigma}} \right) \right\} \right\},$$

$$\hat{f}_2(\hat{\mu}_2, \hat{\sigma}_2) = \frac{1}{\hat{\sigma}} \exp \left\{ - \left(\frac{x_i - \hat{\mu}_2}{\hat{\sigma}} \right) \right\} \exp \left\{ - \exp \left\{ - \left(\frac{x_i - \hat{\mu}_2}{\hat{\sigma}} \right) \right\} \right\}$$

e

$$\hat{g}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}, \hat{p}) = \hat{p} \hat{f}_1(\hat{\mu}_1, \hat{\sigma}) + (1 - \hat{p}) \hat{f}_2(\hat{\mu}_2, \hat{\sigma}).$$

e a função de verossimilhança da distribuição conjunta de Z e X é dada por

$$\begin{aligned} L_{3c}(\boldsymbol{\theta}) = L_{3c}(\mu_1, \mu_2, \sigma, p) &= p^{\sum_{i=1}^n z_i} \frac{1}{\sigma^n} \exp \left\{ - \sum_{i=1}^n z_i \left(\frac{x_i - \mu_1}{\sigma} \right) \right\} \exp \left\{ - \sum_{i=1}^n z_i \exp \left\{ - \frac{x_i - \mu_1}{\sigma} \right\} \right\} \\ &\quad (1-p)^{(n-\sum_{i=1}^n z_i)} \exp \left\{ - \sum_{i=1}^n (1-z_i) \left(\frac{x_i - \mu_2}{\sigma} \right) \right\} \exp \left\{ - \sum_{i=1}^n (1-z_i) \exp \left\{ - \frac{x_i - \mu_2}{\sigma} \right\} \right\} \end{aligned}$$

cujos logaritmo é

$$\begin{aligned}
l_{3c}(\boldsymbol{\theta}) &= l_{3c}(\mu_1, \mu_2, \sigma, p) = \sum_{i=1}^n z_i \log p - n \log \sigma - \sum_{i=1}^n z_i \left(\frac{x_i - \mu_1}{\sigma} \right) \\
&\quad - \sum_{i=1}^n z_i \exp \left\{ - \left(\frac{x_i - \mu_1}{\sigma} \right) \right\} + (n - \sum_{i=1}^n z_i) \log(1 - p) - \sum_{i=1}^n (1 - z_i) \left(\frac{x_i - \mu_2}{\sigma} \right) \\
&\quad - \sum_{i=1}^n (1 - z_i) \exp \left\{ - \left(\frac{x_i - \mu_2}{\sigma} \right) \right\}
\end{aligned} \tag{3.17}$$

em que $z_{1i} = z_i$ e $z_{2i} = 1 - z_{1i}$.

Para implementação do algoritmo ECM, tem-se que o passo E é definido por

$$E_Z [Z_{1i} | \mathbf{X} = \mathbf{x}]^{(k)} = \frac{p^{(k)} f_1(x_i)^{(k)}}{g_2(x_i)^{(k)}}$$

em que

$$f_1(x_i)^{(k)} = \frac{1}{\sigma^{(k)}} \exp \left[- \left(\frac{x - \mu_1^{(k)}}{\sigma^{(k)}} \right) \right] \exp \left\{ - \exp \left[- \left(\frac{x - \mu_1^{(k)}}{\sigma^{(k)}} \right) \right] \right\}$$

e

$$g_1(x)^{(k)} = \sum_{j=1}^2 \frac{p^{(k)}}{\sigma^{(k)}} \exp \left[- \left(\frac{x - \mu_j^{(k)}}{\sigma^{(k)}} \right) \right] \exp \left\{ - \exp \left[- \left(\frac{x - \mu_j^{(k)}}{\sigma^{(k)}} \right) \right] \right\}.$$

Por sua vez, para a implementação do passo CM, tem-se que nesse caso o vetor de parâmetros $\boldsymbol{\theta} = [\mu_1, \mu_2, \sigma, p]^T$ é particionado nos subvetores $[p]$ e $[\mu_1, \sigma, \mu_2]$ e dessa forma a obtenção da estimativa do parâmetro p é dada por

$$p^{(k+1)} = \frac{1}{n} \sum_{i=1}^n E [Z_{1i} | \mathbf{X} = \mathbf{x}]^{(k)}$$

No entanto, para obter-se as estimativas dos parâmetros $[\mu_1, \mu_2, \sigma]$ maximiza o logaritmo da função de verossimilhança dado por (3.17), ou seja

$$\left[\mu_1^{(k+1)}, \mu_2^{(k+1)}, \sigma^{(k+1)} \right] = \arg \max_{(\mu_1^{(k+1)}, \mu_2^{(k+1)}, \sigma^{(k+1)})} l_{3c}(\mu_1, \mu_2, \sigma, p)$$

e novamente para maximizar (3.17) foi utilizado o algoritmo “BFGS”.

Entretanto, independente do método numérico utilizado algumas dificuldade podem surgir, tais como o problema de falta de identificabilidade. Uma função densidade de probabilidade qualquer denotada por $\psi(\cdot)$ é identificável se para distintos valores dos parâmetros dessa

função tem-se distintas funções de densidade de probabilidade, ou seja, se para $\Lambda \neq \Lambda' \in \Omega$ tem-se que $\psi(\mathbf{y}) \neq \psi(\mathbf{y})^\top$ em que Λ é um vetor de parâmetros e $\mathbf{y}$ é o suporte dessa distribuição.

Para o caso de misturas finitas de j distribuições de probabilidade, no entanto, tem-se que a mesma é invariante a j permutações dos índices dos parâmetros de cada um de seus componentes, e, nesse caso, pode-se ter que cada componente da mistura é identificável mas a mistura não, originando assim o problema de falta de identificabilidade.

Por exemplo, para o caso da mistura de duas distribuições de Gumbel com distintos parâmetros de escala tem-se que

$$g_1(x) = pf_1(x; (\mu_1, \sigma_1)) + (1 - p)f_2(x; (\mu_2, \sigma_2)) = (1 - p)f_1(x; (\mu_2, \sigma_2)) + pf_2(x; (\mu_1, \sigma_2)).$$

Além disso, para p igual a 0 ou igual a 1, vê-se que a função densidade dos modelos de mistura fica sendo igual a um dos seus componentes o que, por consequência, torna a mistura de distribuições não identificável.

Então para evitar problemas de falta de identificabilidade, seguindo as recomendações de McLachlan e Peel (2000), estimou-se os parâmetros dos modelos de misturas de duas distribuições de Gumbel com as restrições no espaço paramétrico, $\mu_1 < \mu_2$ uma vez que o passo E do algoritmo ECM já garante $0 < p < 1$. A restrição nos parâmetros de locação é normalmente utilizada para evitar a permutações dos parâmetros entre os componentes da mistura.

Uma outra questão importante para o ajuste de uma mistura finita é a escolha do valor inicial do algoritmo de estimação. Essa questão é importante porque, normalmente, a função de verossimilhança de modelos de misturas apresentam vários pontos de máximos locais, sendo que para distintos valores iniciais do algoritmo, pode-se ter distintos pontos máximos locais. O procedimento empregado neste trabalho para a escolha dos valores iniciais dos modelos de mistura de duas distribuições de Gumbel com distintos ou com mesmo parâmetro de escala foi baseado no trabalho de Lin et al. (2007 b) e é apresentado a seguir

- 1 - A partir da amostra original de tamanho n , retirar $S = 20$ amostras por meio da metodologia “bootstrap” não paramétrico, ou seja 20 amostras com reposição de tamanho n ,
- 2 - Para cada amostra “bootstrap”, particioná-la utilizando o algoritmo k - médias descrito por Rencher (2002) e cujo os detalhes são apresentados a seguir.
A idéia desse algoritmo é classificar os dados em j grupos, no caso $j = 2$, somente baseando-se na distância euclidiana dos valores dos dados da seguinte forma
 - 2.1 - Definir os centróides $\bar{d}_1, \bar{d}_2$, de cada um dos dois grupos. Para isso, retira-se uma amostra de tamanho 2 sem reposição dos valores 1 até n e define-se esses valores oriundos da amostra como sendo os valores dos dois centróides de cada um dos grupos.

2.2 - Gera-se uma matriz de distância entre cada ponto e os centróides, ou seja,

$$D = \begin{pmatrix} d_{1,1} & d_{1,2} \\ \vdots & \vdots \\ d_{n,1} & d_{n,2} \end{pmatrix}$$

em que $d_{i,k}$ é a distância do i -ésimo elemento ao k -ésimo centróide.

2.3 Alocar cada elemento da amostra observada em um dos dois grupos, de acordo com a distância euclidiana deste em relação ao centróide, ou seja, para cada determinado x_i se $d_{i,1} < d_{i,2}$ alocar x_i no grupo 1, caso contrário alocar x_i no grupo 2.

2.4 Calcular os novos centróides para cada grupo de acordo com seguinte expressão

$$\bar{d}_k = \frac{1}{n_{d_1} n_{d_2}} \sum_{i=1}^n \sum_{j=1}^2 d_{ij}$$

em que n_{d_1} número de elementos da amostra que pertence ao grupo 1 e n_{d_2} o número de elementos da amostra que pertence ao grupo 2.

2.5 Repetir o passos 2 a 4 até que $\bar{d}_j^{(k)} = \bar{d}_j^{(l+1)}$ sendo l a l -ésima iteração.

3 - Calcular $p^{(0)}$ por meio de

$$p^{(0)} = \frac{n_{d_1}}{n}$$

4 - Para cada subamostra obtida por meio do passo anterior, calcular $\mu_j^{(0)}, \sigma_j^{(0)}$ por meio de (3.8) e (3.7) respectivamente.

E então para escolher os valores das estimativas dos parâmetros, deve-se obter as estimativas de máxima verossimilhança para cada valor inicial dos parâmetros e escolher como estimativa de máxima verossimilhança, a que proporciona o maior valor logaritmo da função de verossimilhança.

Seguindo as sugestões de Lin et al. (2007 b), foram construídos os intervalos de $100 \times (1 - \alpha)\%$ de confiança dos parâmetros da mistura de duas distribuições de Gumbel com distintos parâmetros de escala utilizando a metodologia “bootstrap”.

O primeiro passo é simular B amostras “bootstrap” de tamanho n , $\mathbf{x}^{*(1)} = \{x_1^{*(1)}, \dots, x_n^{*(1)}\}, \dots, \mathbf{x}^{*(B)} = \{x_1^{*(B)}, \dots, x_n^{*(B)}\}$ do modelo de mistura de duas distribuições de Gumbel por meio do seguinte algoritmo

1 - Para i de 1 até n simular a variável aleatória latente $Z_i^{(*)}$ por meio de

$$Z_i^{(*)} = W^{-1}(\hat{p}),$$

em que $W^{-1}(\hat{p})$ é o inverso da distribuição de Bernoulli com parâmetro $\hat{p}$.

2 - Para cada i simular $x_i^{(*)}$ dado $Z_i^{(*)}$ por meio de

$$\begin{cases} \text{Se } Z_i^{(*)} = 1, & x_i^{(*)} = F^{-1}(U|\hat{\mu}_1, \sigma_1) \\ \text{caso contrário,} & x_i^{(*)} = F^{-1}(U|\hat{\mu}_2, \sigma_2). \end{cases}$$

Logo em seguida, para cada amostra “bootstrap” de tamanho n , $\mathbf{x}^{*(b)}$ obter as estimativas de máxima verossimilhança para os parâmetros $\mu_1, \mu_2, \sigma_1, \sigma_2$ e p , obtendo assim a distribuição empírica “bootstrap” das estimativas de cada um dos parâmetros $(\mu_j^{(*)1}, \dots, \mu_j^{(*)B})$, $(\sigma_j^{(*)1}, \dots, \sigma_j^{(*)B})$ e $(p^{(*)1}, \dots, p^{(*)B})$,

Por fim, para construir o intervalo $100 \times (1 - \alpha)\%$ de confiança de cada parâmetro da mistura de duas distribuições de Gumbel com distintos parâmetro de escala, basta tomar como limites inferiores e superiores desses intervalos os percentis de ordem $100 \times \frac{\alpha}{2}\%$ e $100 \times (1 - \frac{\alpha}{2})\%$ da distribuição empírica das estimativas “bootstrap” de cada um dos parâmetros.

3.2.6 Determinação do número de componentes das misturas de distribuições de Gumbel

Com a finalidade de selecionar qual é a distribuição que melhor ajusta-se aos dados, primeiramente testou-se a seguinte hipótese, $H_0 : X \sim \text{Gumbel}(\mu, \sigma)$ versus $H_a : X \sim p \text{ Gumbel}(\mu_1, \sigma) + (1 - p) \text{ Gumbel}(\mu_2, \sigma)$ por meio de um teste de razão de verossimilhanças baseando-se na distribuição empírica “bootstrap” da estatística 2λ , sendo $\lambda = l_3(\mu_1, \mu_2, \sigma, p) - l_1(\mu, \sigma)$, cujo algoritmo é descrito a seguir.

- 1 - Calcular $2\hat{\lambda}$ para amostra original, sendo $\hat{\lambda} = l_3(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}, \hat{p}) - l_1(\hat{\mu}, \hat{\sigma})$;
- 2 - Gerar B amostras “bootstrap” de tamanho n , sendo $B = 1000$, sob H_0 e para cada um dessas amostras calcular $2\lambda^* = 2(l_3(\hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\sigma}^*, \hat{p}^*) - l_1(\hat{\mu}^*, \hat{\sigma}^*))$. sendo $\hat{\mu}^*, \hat{\sigma}^*$, estimativas de máxima verossimilhança de cada amostra “bootstrap” para distribuição de Gumbel e $\hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\sigma}^*, \hat{p}^*$ estimativas de máxima verossimilhança da misturas de distribuição de Gumbel com mesmo parâmetro de escala para cada uma das amostras “bootstrap”.

Uma observação importante é que, seguindo as idéias sugeridas por Feng e McCulloch (1996), adotou-se, neste trabalho, a metodologia “bootstrap” paramétrico, em que a amostra “bootstrap” é definida por

$$\mathbf{x}^* : x_i^* = F^{-1}(U|\hat{\mu}, \hat{\sigma})$$

em que $U \sim \text{Uniforme}(0, 1)$ e $F^{-1}(U|\hat{\mu}, \hat{\sigma})$ o inverso da distribuição de Gumbel com $\mu = \hat{\mu}$ e $\sigma = \hat{\sigma}$

3 - Calcular o Valor-p definido por

$$\text{Valor- p} = \frac{1}{B} \sum_{b=1}^B I_{\{2\lambda > 2\lambda^*\}}(2\lambda^*)$$

Logo em seguida, faz-se um segundo teste de hipótese, denotado aqui por teste 2, que consiste em testar $H_0 : X \sim \text{Gumbel}(\mu, \sigma)$ versus $H_a : X \sim p \text{ Gumbel}(\mu_1, \sigma_1) + (1 - p) \text{ Gumbel}(\mu_2, \sigma_2)$ seguindo o mesmo procedimento descrito acima, exceto ao cálculo de λ^* , que foi calculado da seguinte forma $\lambda^* = l_2(\hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\sigma}_1^*, \hat{\sigma}_2^*, \hat{p}^*) - l_1(\hat{\mu}^*, \hat{\sigma}^*)$.

No caso de rejeição de H_0 para os dois testes anteriores, realiza-se um terceiro teste de hipótese, denotado aqui por teste 3, para testar $H_0 : X \sim p \text{ Gumbel}(\mu_1, \sigma) + (1 - p) \text{ Gumbel}(\mu_2, \sigma)$ versus $H_a : X \sim p \text{ Gumbel}(\mu_1, \sigma_1) + (1 - p) \text{ Gumbel}(\mu_2, \sigma_2)$ por meio do teste de razão de verossimilhança assintótico.

O procedimento do teste de razão de verossimilhanças assintótico consiste em calcular $2\lambda = 2(l_2(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, \hat{p}) - l_3(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}, \hat{p}))$, e supondo que $2\lambda \sim \chi^2_{1, (1-\alpha)}$, rejeita-se H_0 ao nível de significância de $100 \times \alpha\%$ caso $2\lambda > c_{(1-\alpha)}$, sendo $c_{(1-\alpha)}$ o quantil de ordem $100 \times (1 - \alpha)\%$ da distribuição χ^2 com, 1 grau de liberdade.

Além dos testes de hipóteses, para auxiliar a escolha da melhor distribuição candidata foram calculados o critério de informação de Akaike (1974) e o critério de informação Bayesiano proposto por Schwarz (1978), dados, respectivamente, por

$$AIC = -2l + 2k$$

e

$$BIC = -2l + kn$$

sendo k o número de parâmetros, l o logaritmo da função de verossimilhança da distribuição candidata considerada e n o tamanho da amostra.

3.2.7 Diagnóstico do ajuste

Para avaliar a qualidade do ajuste, foram construídos os gráficos probabilidade - probabilidade, (“PP - plot”) e quantil - quantil (“QQ - plot”) com envelopes simulados, técnica proposta primeiramente por Atkinson (1985) no contexto de análise de resíduos de modelos de regressão .

Para construí-los, inicialmente, obtém-se os dados observados de velocidade máxima mensal de vento em ordem crescente, $\{x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}\}$, e então, para o caso da

distribuição de Gumbel, o gráfico probabilidade - probabilidade é definido pelos pontos de coordenadas $\left(\frac{1}{i+1}, \hat{F}(x_{(i)})\right)$ em que $\hat{F}(x_{(i)})$ é definido por

$$\hat{F}(x_{(i)}) = \exp \left\{ - \exp \left[- \left(\frac{x_{(i)} - \hat{\mu}}{\hat{\sigma}} \right) \right] \right\},$$

Para obter o envelope simulado de $100 \times (1 - \alpha)\%$ de confiança deve-se proceder da seguinte forma

1 - Gerar B amostras de tamanho n a partir de $Gumbel(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ denotadas por

$$\left\{ x_1^{*(1)}, \dots, x_n^{*(1)} \right\}, \dots, \left\{ x_1^{*(B)}, \dots, x_n^{*(B)} \right\},$$

2 - Para cada uma das B amostras, obter os estimadores de máxima verossimilhança para os parâmetros que para a distribuição de Gumbel são μ e σ , para a mistura de duas distribuições de Gumbel com distintos parâmetros de escala são $(\mu_1, \mu_2, \sigma_1, \sigma_2, p)$ e para o caso particular $\sigma_1 = \sigma_2$ são $(\mu_1, \mu_2, \sigma, p)$.

3 - Ordenar as amostras geradas em forma crescente e obter pontos de coordenadas $\left\{ \frac{1}{i+1}, \hat{F}(x_i^*) \right\}$. No caso da distribuição de Gumbel, $\left\{ \frac{1}{i+1}, \hat{G}_1(x_i^*) \right\}, \left\{ \frac{1}{i+1}, \hat{G}_2(x_i^*) \right\}$ para a mistura de duas distribuições de Gumbel com distintos e com mesmo parâmetros de escala, respectivamente.

3 - Para cada i obter os limites superior e inferior de confiança $100 \times (1 - \alpha)\%$ do envelope que são definidos como sendo o quantis de ordens $100 \times \frac{\alpha}{2}\%$ e $100 \times \left(1 - \frac{\alpha}{2}\right)\%$ de $\hat{F}(x_i^{*(1)}), \dots, \hat{F}(x_i^{*(B)})$ respectivamente no caso da distribuição de Gumbel e quantis de ordens $100 \times \frac{\alpha}{2}\%$ e $100 \times \left(1 - \frac{\alpha}{2}\right)\%$ de $\hat{G}_1(x_i^{*(1)}), \dots, \hat{G}_1(x_i^{*(B)})$ e de $\hat{G}_2(x_i^{*(1)}), \dots, \hat{G}_2(x_i^{*(B)})$ para mistura de duas distribuições de Gumbel com distintos e com mesmo parâmetro de escala.

Por sua vez, para o caso dos modelos de mistura de duas distribuições de Gumbel o gráfico probabilidade - probabilidade é definido pelos pontos de coordenadas $\left(\frac{1}{i+1}, \hat{G}_1(x_{(i)})\right)$ sendo $\hat{G}_1(x_{(i)})$ dado por

$$\hat{G}(x_{(i)}) = \int_{-\infty}^{x_{(i)}} \sum_{j=1}^2 \hat{p}_j \frac{1}{\hat{\sigma}_j} \exp \left\{ - \left(\frac{x - \hat{\mu}_j}{\hat{\sigma}_j} \right) \right\} dx$$

E ainda, para o caso de modelos de mistura de duas distribuições de Gumbel com mesmo parâmetro de escala, tem-se que o gráfico de probabilidade - probabilidade é definido por $\left(\frac{1}{i+1}, \hat{G}_2(x_{(i)})\right)$ sendo

$$\hat{G}_2(x_{(i)}) = \int_{-\infty}^{x_{(i)}} \sum_{j=1}^2 \hat{p}_j \frac{1}{\hat{\sigma}} \exp \left\{ - \left(\frac{x - \hat{\mu}_j}{\hat{\sigma}} \right) \right\} dx.$$

A interpretação desse gráfico é a seguinte, sob a hipótese dos dados seguirem a distribuição proposta os pontos desse gráfico estarão alinhados em uma reta que passa pelos pontos (0,0) e (1,1), e dentro dos limites de confiança do envelope.

Outra forma utilizada para avaliar graficamente a qualidade do ajuste das distribuições candidatas foi a construção do gráfico quantil - quantil com envelope simulado que, para o caso da distribuição de Gumbel, é definido pelos pontos de coordenadas $\left(\hat{F}^{-1} \left(\frac{1}{i+1} \right), x_{(i)} \right)$ em que $\hat{F}^{-1} \left(\frac{1}{i+1} \right)$ é dado por

$$\hat{F}^{-1} \left(\frac{1}{i+1} \right) = \hat{\mu} - \hat{\sigma} \log \left[-\log \left(\frac{1}{i+1} \right) \right].$$

Ainda, para o caso dos modelos de mistura de duas distribuições de Gumbel o gráfico quantil - quantil é definido por meio dos pontos de coordenadas $\left(\hat{G}_1^{-1} \left(\frac{1}{i+1} \right), x_{(i)} \right)$ sendo

$$\hat{G}_1^{-1} \left(\frac{1}{i+1} \right) = \int_{-\infty}^{\hat{x}_{(i)}} \sum_{j=1}^2 \hat{p}_j \frac{1}{\hat{\sigma}_j} \exp \left\{ - \left(\frac{x - \hat{\mu}_j}{\hat{\sigma}_j} \right) \right\} dx = \frac{1}{i+1}$$

e em se tratando do modelo de mistura de duas distribuições de Gumbel com mesmo parâmetro de escala o gráfico quantil - quantil é definido pelos pontos de coordenadas $\left(\hat{G}_2^{-1} \left(\frac{1}{i+1} \right), x_{(i)} \right)$ em que

$$\hat{G}_2^{-1} \left(\frac{1}{i+1} \right) = \int_{-\infty}^{\hat{x}_{(i)}} \sum_{j=1}^2 \hat{p}_j \frac{1}{\hat{\sigma}} \exp \left\{ - \left(\frac{x - \hat{\mu}_j}{\hat{\sigma}} \right) \right\} dx = \frac{1}{i+1}.$$

O envelope simulado do gráfico quantil-quantil é feito de maneira análoga ao gráfico probabilidade - probabilidade, no entanto, deve-se calcular para cada uma das B amostras simuladas os pontos de coordenadas $\left(\hat{F}^{-1} \left(\frac{1}{i+1} \right), x_{(i)} \right)$, para a distribuição de Gumbel, $\left(\hat{G}_1^{-1} \left(\frac{1}{i+1} \right), x_{(i)} \right)$ e $\left(\hat{G}_2^{-1} \left(\frac{1}{i+1} \right), x_{(i)} \right)$ para a mistura de duas distribuições de Gumbel com distintos e com mesmo parâmetro de escala respectivamente.

A interpretação do gráfico quantil-quantil consiste em sob a hipótese dos dados seguirem a distribuição proposta os pontos desse gráfico estarão alinhados em uma reta e dentro dos limites do envelope. Convém ressaltar que o gráfico quantil-quantil tem uma vantagem adicional de poder-se detectar pontos atípicos.

3.2.8 Obtenção das probabilidades de ocorrência

Uma vez escolhida a distribuição candidata que melhor ajusta-se os dados, a próxima etapa é estimar as probabilidades de ocorrência de rajadas de ventos com velocidade acima de um determinado valor x , que no caso de se escolher a distribuição de Gumbel, podem ser estimadas por meio de

$$P(X > x) = 1 - \hat{F}(x) = 1 - \exp \left\{ \exp \left(-\frac{x - \hat{\mu}}{\hat{\sigma}} \right) \right\}.$$

Por outro lado, caso escolher a mistura de duas distribuições de Gumbel com distintos parâmetros de escala, as probabilidades de ocorrência podem ser estimada por meio de

$$P(X > x) = \int_x^\infty \sum_{j=1}^2 \hat{p}_j \frac{1}{\hat{\sigma}_j} \exp \left\{ -\left(\frac{x - \hat{\mu}_j}{\hat{\sigma}_j} \right) \right\} dx$$

e caso seja escolhido a mistura de duas distribuições de Gumbel com mesmo parâmetro de escala, as probabilidades de ocorrência podem ser estimadas por meio de

$$P(X > x) = \int_x^\infty \sum_{j=1}^2 \hat{p}_j \frac{1}{\hat{\sigma}} \exp \left\{ -\left(\frac{x - \hat{\mu}_j}{\hat{\sigma}} \right) \right\} dx.$$

3.2.9 Estimação de níveis de retorno e seus respectivos intervalos de confiança

Uma outra estimativa de interesse é a estimativa de nível de retorno, x_N , para um período pré fixado N que normalmente é mensurado em anos. O nível de retorno pode ser definido como o nível que pode ser excedido em qualquer ano com probabilidade $\frac{1}{N}$, que no caso da distribuição de Gumbel, é dado por

$$x_N = F^{-1} \left(1 - \frac{1}{N} \right) = \mu - \sigma \log \left[-\log \left(1 - \frac{1}{N} \right) \right]. \quad (3.18)$$

Em contrapartida, se for escolhida a misturas de duas distribuições de Gumbel, pode-se estimar x_N por meio de

$$x_N = G_1^{-1} \left(1 - \frac{1}{N} \right) = \int_{x_N}^\infty \sum_{j=1}^2 \hat{p}_j \frac{1}{\hat{\sigma}_j} \exp \left\{ -\left(\frac{x - \hat{\mu}_j}{\hat{\sigma}_j} \right) \right\} dx = \frac{1}{N} \quad (3.19)$$

e em se tratando de se escolher a mistura de duas distribuições de Gumbel com mesmo parâmetro de escala, pode-se estimar x_N por meio de

$$x_N = G_2^{-1} \left(1 - \frac{1}{N} \right) = \int_{x_N}^{\infty} \sum_{j=1}^2 \hat{p}_j \frac{1}{\hat{\sigma}} \exp \left\{ - \left(\frac{x - \hat{\mu}_j}{\hat{\sigma}} \right) \right\} dx = \frac{1}{N}. \quad (3.20)$$

Além das estimativas pontuais, foram contruídos intervalos de $100 \times (1 - \alpha)\%$ para os níveis de retorno da distribuição de Gumbel e para as misturas.

Para a distribuição de Gumbel os intervalos de confiança foram contruídos por meio do método do perfil da verossimilhança. Este método consiste em primeiramente fazer-se uma reparametrização na função densidade de probabilidade de Gumbel definida em (3.1) isolando μ na expressão (3.18), ou seja, o parâmetro de locação é definido agora por

$$\mu = x_N + \sigma \log \left[-\log \left(1 - \frac{1}{N} \right) \right]$$

e sendo assim, a função densidade de probabilidade de Gumbel fica com parâmetros (x_N, σ) . Em seguida, fixa-se o valor de x_N , isto é, $x_N = x_{0N}$, constrói-se a função de verossimilhança cujo o logaritmo é dado por

$$l(x_{0N}, \sigma) = -n \log \sigma - \sum_{i=1}^n \left(\frac{x_i - \mu_{x_{0N}}}{\sigma} \right) - \sum_{i=0}^n \exp \left[- \left(\frac{x_i - \mu_{x_{0N}}}{\sigma} \right) \right]$$

e maximiza-se a função de verossimilhança em relação a σ , sendo que

$$\mu_{x_{0N}} = x_{0N} + \sigma \log \left[-\log \left(1 - \frac{1}{N} \right) \right]$$

Assim a construção do intervalo de $100 \times (1 - \alpha)\%$ de confiança de x_N é dada, por

$$I.C_{(100 \times (1 - \alpha)\%)}(x_N) = \{x_N : 2(l(\hat{\mu}, \hat{\sigma}) - l(x_{0N}, \hat{\sigma})) \leq d_{(1 - \alpha)}\}$$

Por fim, para a construção do intervalo de $100 \times (1 - \alpha)\%$ de confiança de x_N para os modelos de misturas de duas distribuições de Gumbel neste trabalho é apresentado o seguinte procedimento.

- 1- Simular $B = 1000$ amostras “bootstrap” de tamanho n , $\mathbf{x}^{(s)} = \{x_1^{(s)}, \dots, x_n^{(s)}\}$ do modelo de mistura de duas distribuições de Gumbel conforme já foi descrito para a obtenção dos intervalos de confiança dos parâmetros.
- 2- Para cada amostra “bootstrap” de tamanho n , $\mathbf{x}^{(s)}$ obter os estimadores de máxima verossimilhança dos $\mu_1, \mu_2, \sigma_1, \sigma_2$ e p e obter os nível de retorno para o período de N anos por meio de (3.19).

- 3- Por fim, para a construção do intervalo $100 \times (1 - \alpha)\%$ de confiança de x_N para os modelos de misturas, basta tomar como limites inferiores e superiores do intervalo os percentis $100 \times \frac{\alpha}{2}\%$ e $100 \times (1 - \frac{\alpha}{2})\%$ da distribuição empírica das estimativas “bootstrap” do nível de retorno para um período de N anos.

Uma observação importante é que para o caso particular de mistura de duas distribuições de Gumbel em que $\sigma_1 = \sigma_2$, obtém-se os intervalos de $100 \times (1 - \alpha)\%$ de confiança dos níveis de retorno de N anos de maneira similar ao que foi apresentado para mistura de duas distribuições de Gumbel com distintos parâmetros de escala, no entanto, deve-se estimar os níveis de retorno por meio de (3.20).

3.3 Resultados e Discussão

O primeiro passo da análise dos dados de velocidade máxima do vento dos meses de abril, maio, agosto e setembro, foi verificar a hipótese de independência por meio do teste de chorrilho (“run test”) cujos resultados são apresentados na tabela 3.1. Analisando essa tabela vê-se que, ao nível de significância de 5%, não há evidências estatísticas para rejeitar a hipótese nula de independência.

Tabela 3.1 – Números totais de valores maiores n_1 , e menores, n_2 , que a mediana, número total de chorrilhos r e valores críticos do teste de chorrilho para os dados de velocidade máxima do vento dos meses de abril, maio, agosto e setembro, considerando-se o nível de significância 5%

Mês	n_1	n_2	r	valores críticos
abr.	25	25	27	18; 34
maio	25	25	22	18; 34
ago.	25	25	28	18; 34
set.	24	24	26	17; 33

Como próxima etapa, conforme mostra a Tabela 3.2, foram estimados os parâmetros μ e σ da distribuição de Gumbel por meio do método de máxima verossimilhança e seus respectivos intervalos de 95% de confiança obtidos por meio do método do perfil da verossimilhança para os dados de velocidade máxima do vento dos meses de abril, maio, agosto e setembro. Convém ressaltar que utilizou-se como valores iniciais para o algoritmo “BFGS” as estimativas obtidas por meio do método dos momentos (50, 079; 10, 068), (50, 260; 12, 615), (50, 899; 10, 677) e (62, 095; 13, 513) para os meses de abril, maio, agosto e setembro, respectivamente.

Tabela 3.2 – Estimativas de máxima verossimilhança dos parâmetros da distribuição de Gumbel (Modelo 1), da mistura de duas distribuições de Gumbel com mesmo parâmetro de escala (Modelo 2), da mistura de duas distribuições de Gumbel com distintos parâmetros de escala (Modelo 3) com seus respectivos intervalos de 95 % de confiança, e estimativa do critério de informação de Akaike e do critério de informação Bayesiano dos diversos modelos para os dados de velocidade máxima do vento dos meses de abril, maio, agosto e setembro

Mês	Modelo	Parâmetros					AIC	BIC
		μ_1	σ_1	μ_2	σ_2	p		
Abr.	1	50,079	10,068				396,8	400,7
		(47,7; 54,1)	(8,1; 12,8)					
	2	47,231	6,609	79,596		0,837	392,8	400,4
		(45,5; 49,4)	(5,6; 7,8)	(72,4; 84,9)		(0,71; 0,93)		
	3	47,474	7,076	78,966	3,690	0,842	392,5	402,0
		(45,1; 50,2)	(5,1; 8,9)	(74,7; 83,2)	(0,7; 6,7)	(0,70; 0,94)		
Maio	1	50,260	12,615				418,4	422,3
		(46,5; 54,1)	(10,1; 16,1)					
	2	45,412	7,407	77,663		0,746	413,1	420,8
		(43,1; 48,2)	(5,9; 9,4)	(71,5; 84,7)		(0,60; 0,88)		
	3	45,633	7,736	77,897	6,399	0,755	414,9	424,4
		(42,6; 49,1)	(4,8; 10,2)	(69,6; 84,0)	(2,1; 10,8)	(0,56; 0,88)		
Ago.	1	50,899	10,677				396,6	400,4
		(47,7; 54,1)	(8,7; 13,4)					
	2	45,809	8,845	57,234		0,483	399,7	407,3
		(37,2; 52,9)	(5,7; 11,2)	(49,7; 67,1)		(0,07; 0,75)		
	3	48,697	8,930	75,036	2,711	0,856	395,8	405,4
		(45,5; 51,9)	(5,9 10,9)	(71,2; 78,7)	(0,4; 6,0)	(0,68; 0,95)		
Set.	1	62,095	13,513				414,0	417,8
		(58,1; 66,3)	(10,9; 17,1)					
	2	58,569	10,297	89,969		0,822	416,1	423,6
		(53,8; 62,5)	(6,9; 12,8)	(72,3; 110,3)		(0,51; 0,96)		
	3	59,527	11,165	95,957	3,843	0,874	414,1	423,6
		(55,7; 63,6)	(7,5; 13,8)	(88,5; 102,3)	(0,3; 8,5)	(0,72; 0,96)		

A tabela 3.2 também apresenta as estimativas de máxima verossimilhança dos parâmetros das misturas de duas distribuições de Gumbel obtidas por meio do algoritmo ECM e seus respectivos intervalos de 95% de confiança obtidos por meio da metodologia “bootstrap” paramétrico. De uma forma geral, vê-se que para o ajuste de mistura de duas distribuições de Gumbel com distintos parâmetros de escala, as estimativas do parâmetro σ_2 são menores do que as estimativas do parâmetro σ_1 e, além disso, as amplitudes dos intervalos de 95% de confiança do parâmetro σ_2 , 6, 8, 7, 5, 6, 8, 2, respectivamente, são maiores do que as amplitudes dos intervalos de 95% de confiança do parâmetro σ_1 , 3, 8, 5, 4, 5, 6, 3, respectivamente. Observa-se também que as estimativas do parâmetro p são maiores que 0, 74, exceto para o ajuste da mistura de duas distribuições de Gumbel com mesmo parâmetro de escala aos dados do mês de agosto, sugerindo que o grupo de velocidades mais extremas são menos freqüentes e menos dispersos.

Uma vez estimados os parâmetros de cada uma das distribuições candidatas, foi feita a seleção do modelo probabilístico mais adequado por meio dos testes de razão de verossimilhanças “bootstrap” e assintótico, conforme mostra a tabela 3.3, e por meio das estimativas do critério de informação de Akaike e critério de informação Bayesiano apresentados na tabela 3.2.

Tabela 3.3 – Teste de razão de verossimilhanças ‘bootstrap’ 1 e 2 para testar que sob H_0 , os dados seguem distribuição de Gumbel versus H_a que os dados seguem uma mistura de duas distribuições de Gumbel com mesmo e distintos parâmetros de escala, respectivamente, e teste de razão de verossimilhanças 3 para testar $H_0 : \sigma_1 = \sigma_2$ versus $H_a : \sigma_1 \neq \sigma_2$ para os modelos de misturas de duas distribuições de Gumbel

Mês	Teste	Deviance	Valor - p
Abr.	1	8,070	0,025
	2	10,368	0,055
	3	2,298	0,130
Maio	1	9,293	0,014
	2	9,568	0,113
	3	0,275	0,600
Ago.	1	0,874	0,660
	2	6,715	0,197
	3	5,842	0,016
Set.	1	1,954	0,494
	2	5,917	0,264
	3	3,964	0,046

Analisando a tabela 3.2 vê-se que para os dados de velocidade máxima de vento

do mês de abril, a estimativa do critério AIC é menor para o ajuste do modelo de mistura de duas distribuições de Gumbel com distintos parâmetro de escala, mas por outro lado, vê-se que é o ajuste da mistura de duas distribuições de Gumbel com mesmo parâmetro de escala que minimiza o critério BIC. Por sua vez, por meio da tabela 3.3 observa-se que para os dados do mês de abril rejeita-se a hipótese H_0 para o teste 1 ao nível de significância de 5% mas não rejeita-se H_0 para os teste 2, logo, escolhe-se a distribuição candidata que melhor ajusta a essa série de dados é a mistura de duas distribuição de Gumbel com mesmo parâmetro de escala.

Para os dados do mês de maio, rejeita-se H_0 para o teste 1, mas não se rejeita H_0 para o teste 2, então nesse caso, considerando apenas esses testes, escolhe-se a mistura de duas distribuições de Gumbel com mesmo parâmetro de escala, e vê-se também que essa escolha se confirma usando tanto o critério AIC como o BIC.

Ainda analisando a tabela 3.2 observa-se que para os dados de velocidade máxima de vento referente ao mês de agosto, é o ajuste do modelo de mistura de duas distribuições de Gumbel com distintos parâmetros de escala que minimiza o critério AIC, no entanto, é o ajuste da distribuição de Gumbel que minimiza o critério BIC. Além disso, analisando a tabela 3.3 vê-se que não se rejeita H_0 nem para o teste 1 nem para o teste 2, no entanto, rejeita-se H_0 para o teste 3, sendo assim, utilizando o princípio da parcimônia, a distribuição candidata escolhida é a distribuição de Gumbel.

E por fim, vê-se que a distribuição candidata que melhor ajusta-se aos dados referentes ao mês de setembro é a distribuição de Gumbel. Uma vez que, ao nível de significância de 5% não se rejeita H_0 para os testes 1 e 2 e o que se confirma ao analisar os critérios de seleção AIC e BIC. Convém ressaltar que embora rejeita-se H_0 para o teste 3 para considerando os dados referentes aos meses de agosto e setembro, esses resultado foi desconsiderados utilizando o princípio da parcimônia, uma vez que não rejeitou-se a hipótese de que os dados seguem a distribuição de Gumbel para os testes 1 e 2.

Com a finalidade de avaliar a qualidade do ajuste das distribuições escolhidas para os dados de velocidade máxima de vento dos meses de abril, maio, agosto e setembro, foram construídos os gráficos probabilidade - probabilidade e quantil - quantil com envelope simulado conforme mostram as figuras 3.1 e 3.2 respectivamente. Analisando os gráficos probabilidade - probabilidade e quantil-quantil vê-se um boa qualidade do ajuste dos modelos para cada um dos meses, uma vez que os pontos desses gráficos encontram-se dentro dos limites inferiores e superiores do envelope simulado de 95% de confiança. Além disso, na figura 3.3 são apresentadas as densidades não paramétricas e o ajuste das distribuições que melhor ajusta-se aos dados de velocidade máxima do vento dos meses de abril, maio, agosto e setembro. Analisando essa figura, observa-se que a distribuição dos dados dos meses de abril e maio é aparentemente bimodal.

A tabela 3.4 apresentam as estimativas das probabilidades de ocorrência de veloci-

Figura 3.1 – Gráfico probabilidade - probabilidade com envelope simulado de 95% de confiança para o ajuste da mistura de duas distribuições de Gumbel com mesmo parâmetros de escala para os meses de abril e maio, e para o ajuste da distribuição de Gumbel para os meses de agosto e setembro

Figura 3.2 – Gráfico quantil - quantil com envelope simulado de 95% de confiança para o ajuste da mistura de duas distribuições de Gumbel com mesmo parâmetros de escala para os meses de abril e maio, e para o ajuste da distribuição de Gumbel para os meses de agosto e setembro

Figura 3.3 – Densidades não paramétricas para os dados de velocidade máxima do vento dos meses de abril e maio, agosto e setembro (---) e distribuições ajustadas correspondentes (—)

dades máxima mensais de vento acima 40, 50, 60, 70, 80, 90 e 100 $km.h^{-1}$ para os meses de abril, maio, agosto e setembro. Analisando essa tabela, observa-se que, para os meses de abril e maio respectivamente, as estimativas de probabilidades de ocorrência de rajadas de ventos com velocidades acima de 60 $km.h^{-1}$ são por volta de 0,28 e 0,35, e comparando com as estimativas obtidas por Bautista et al (2004), que são aproximadamente de 0,35 e 0,39, conclui-se que as estimativas das probabilidades de ocorrências com ventos acima 60 $km.h^{-1}$ são superestimadas quando se utiliza a distribuição de Gumbel.

Por outro lado, quando se analisa para esses mesmos meses as estimativas de probabilidades de ocorrência de rajadas de ventos com velocidade acima de 80 $km.h^{-1}$, vê-se que se obtém valores de aproximadamente de 0,10 e 0,14, que são maiores que as obtidas por Bautista et. al (2004), dadas por 0,06 e 0,1. Logo, ao usar a distribuição de Gumbel para estimar probabilidades de ocorrências de ventos acima 80 $km.h^{-1}$ pode-se estar subestimando as mesmas.

Tabela 3.4 – Probabilidades de ocorrência de rajada máxima mensais de vento com velocidade acima de 40, 50, 60, 70, 80, 90 e 100 $km.h^{-1}$, a 10 m acima do nível do solo, para os meses de abril, maio, agosto e setembro, em Piracicaba, SP

Mês	> 40	> 50	> 60	> 70	> 80	> 90	> 100
Abr.	0,95780	0,56625	0,27551	0,18660	0,10498	0,03171	0,00753
Maio	0,90646	0,56452	0,35116	0,26527	0,13849	0,04557	0,01262
Ago.	0,93768	0,66306	0,34714	0,15390	0,06340	0,02535	0,01001
Set.	0,99408	0,91349	0,68892	0,42714	0,23340	0,11910	0,05871

A tabela 3.5 apresenta as estimativas de níveis de retornos e seus respectivos intervalos de 95% de confiança obtidos por meio da metodologia “bootstrap” paramétrico para os dados de velocidade máxima de vento dos meses de abril e maio, e obtidas por meio do perfil da verossimilhança para os meses de agosto e setembro.

Então, analisando essa tabela, vê-se que as amplitudes dos intervalos de 95% de confiança do nível de retorno para o período de 100 anos são, para os meses de abril e maio respectivamente, 15,8 e 17,1, ou seja, menores que as amplitudes dos intervalos de 95% de confiança do nível de retorno obtidos por Bautista et.al (2004) para esses mesmos meses e período pré fixado, 26,22 e 31,93 respectivamente. Sendo assim, pode-se concluir que os comprimentos dos intervalos de confiança dos níveis de retorno são menores quando utiliza-se a mistura de duas distribuições de Gumbel com mesmo parâmetro de escala.

Tabela 3.5 – Níveis de Retorno $km.h^{-1}$ estimados e limites inferior (L.I) e superior (L.S) de seus respectivos intervalos de 95 % de confiança para os períodos pré fixados de 5, 10, 50 e 100 anos

Mês	5 anos	10 anos	50 anos	100 anos
Abr.	67,6 (59,0; 78,9)	80,5 (68,4; 85,9)	93,3 (84,5; 100,4)	98,1 (89,3; 105,1)
Maio	75,5 (63,5; 81,7)	83,2 (75,2; 89,0)	96,5 (87,8; 104,0)	101,8 (92,8; 109,9)
Ago.	66,9 (62,1; 73,0)	74,9 (68,9; 82,8)	92,6 (83,4; 104,7)	100,0 (89,6; 114,9)
Set.	82,4 (76,2; 90,2)	92,5 (84,6; 102,7)	114,8 (102,9; 130,8)	124,3 (110,7; 142,4)

3.4 Conclusões

Diante dos resultados obtidos pode concluir que:

- A mistura de duas distribuições de Gumbel é a distribuição que melhor se ajusta aos dados de velocidades máxima de ventos para os meses de abril e maio.
- A distribuição de Gumbel é a distribuição que melhor se ajusta aos dados de velocidades máxima de ventos para os meses de agosto e setembro.
- A utilização da distribuição de Gumbel para estimar as probabilidades de ocorrência de rajadas de ventos com velocidade acima de $60 km.h^{-1}$ podem acarretar em estimativas superestimadas enquanto que para velocidades acima de $80 km.h^{-1}$ acarretam estimativas subestimadas.
- As amplitudes dos intervalos de confiança dos níveis de retorno foram menores quando utiliza-se a mistura de duas distribuições de Gumbel com mesmo parâmetro de escala.

Referências

ATKINSON, A. C. **Plots, Transformations and Regressions**. Oxford: Statistical Science Series; 1985. 296 p.

AKAIKE A new look at the statistical model identification, **IEEE Transactions on Automatic Control**, New York, v. 19, n. 6, p. 716-723, 1974.

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, n. 1, p. 95-111, 2004.

BOTTOLO, L.; CONSONNI, G.; DELLAPORTAS, PETROS.; LIJOI, ANTONIO. Bayesian analysis of extremes values by mixture modeling. **Extremes**, New York, v. 6, n. 1, p. 25 - 47, 2003.

BROWNLIE, K.A. **Statistical theory and methodology in science and engineering**. 2 ed. New York: John Wiley, 1965. 590 p.

DAY, N. E. Estimating the components of a mixture of normal distributions. **Biometrika**, Cambridge, v. 56, n. 3, p. 463-474. 1969.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM Algorithm. **Journal of the Royal Statistical Society B**, London, v. 39, n. 1, p. 1-38, 1977.

EFRON, B.; TIBSHIRANI, R. **An Introduction to the Bootstrap**. London: Chapman & Hall, 1993. 436 p.

EISENHART, C.; SWED, F. Tables for testing randomness of grouping in a sequence of alternatives. **Annals of Mathematical Statistics**, Ann Arbor, v. 14, n. 1, p. 66-87. 1941

FAHMI, K.J.; ALABBASI, J.N. Mixture distributions - an alternative approach for estimating maximum magnitude earthquake occurrence. **Geophysical Journal International**, New York, v. 89, n. 2, p. 741-747, 1987.

FAHMI, K.J.; ALABBASI, J.N. Application of a mixture distribution of extreme values to earthquake magnitudes in Iraq and conterminous regions. **Geophysical Journal International**, New York, v. 107, n. 2, p. 209-217, 1991.

FENG, Z. D.; McCULLOCH, C. E. Likelihood ratio tests of the number of components in a normal mixture with unequal variances. **Biometrics**, Washington, v. 50, n. 4, p. 1158-1162. 1994.

FENG, Z. D.; McCULLOCH, C. E. Using bootstrap likelihood ratios in finite mixture models. **Journal of the Royal Statistical Society B**, London, v. 58, n.3, p. 609-617, 1996.

LIN, T. I.; LEE, J. C.; YEN, S. Y. Finite mixture modelling using the skew normal distributions. **Statistica Sinica**, Taipei, v. 17, p. 909-927, 2007.

LIN, T. I.; LEE, J. C.; YEN, S. Y. Robust mixture modeling using the skew t distribution. **Journal Statistical Comptuing**, Philadelphia, v. 17, p. 81-92, 2007.

JENKINSON, A.F The frequency distribution of the annual maximun (or minimum) values of metereological elements. **Quartely Journal of the Royal Meteorological Society**, London, v.81, n. 348, p. 158-171, 1955.

MENG, X. L.; RUBIN, D. B. Maximum likelihood estimation via the ECM algorithm: A general framework. **Biometrika**, Cambridge, v.80, n. 2, p. 267-278. 1993.

McLACHLAN, G. J. On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. **Applied Statistics**, London, v. 36, n. 3, p. 318-324. 1987.

McLACHLAN, G. J.; PEEL, D. **Finite mixture models**. New York: Wiley, 2000. 419 p.

NOCEDAL, J.; WRIGHT, S.J. **Numerical optimization**. New York: Springer Verlag, 1999. 636 p.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, ISBN 3-900051-07-0, Vienna, Austria, 2006. Disponível em: <<http://www.R-project.org>>. Acesso em: 01 out. 2007.

RAO, C.R. The utilization of multiple measurements in problems of biological classification. **Journal of the Royal Statistical Society B**, London, v. 10, n. 2, p. 159-203, 1948.

RENCER, A. C.; **Methods of multivariate analysis**. New York: John Wiley, 2002. 708 p.

SCHWARZ, G.E. Estimating the dimension of a model. **Annals of Statistics**, Hayward, v. 6, n. 2, p. 461-464, 1978.

SILVA, R.R.; ZOCCHI, S.S. Densidades não paramétricas no estudo da velocidade máxima do vento em Piracicaba, SP. In: RBRAS, 50.; SEAGRO, 11., 2005, Londrina. **Resumos ...** Londrina: UEL, 2005, 1 CD-ROM.

TARTAGLIA, V.; CAPORALI, E.; CAVIGLI, E.; MORO, A. L - moments based assessment of a mixture model for frequency analysis of rainfall extremes. **Advanced in Geosciences**, Katlenburg - Lindau, v.2, p. 331-334, Jan. 2006. Disponível em: <<http://www.adv-geosci.net/11/index.html>>. Acesso em: 01 maio 2007.

TITTERINGTON, D.M.; SMITH, A.F.M.; MAKOV, U. E. **Statistical analysis of finite mixture distributions**. New York: Wiley, 1985. 243 p.

TITTERINGTON, D. M. Some recent research in the analysis of mixture distributions. **Statistics**, Berlin, v. 21, n. 4, p. 619-641, 1990.

THOM, H. C. S. Toward a universal climatological extreme wind distribution. In: PROCEEDINGS OF THE INTERNATIONAL RESEARCH SEMINAR: WIND EFFECTS ON BUILDINGS AND STRUCTURES. Ottawa, **Proceedings...**, Ottawa: Univ. of Toronto Press, 1967, p. 669-684.

WALSHAW, D. Modelling extreme wind speeds in regions prone to hurricanes. **Applied Statistics**, London, v. 49, n. 1, p. 51-62, 2000.

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

ZOCCHI, S. S. **Misturas de modelos logit, probit e complemento log-log**. 1993. 129 p. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura Piracicaba, "Luiz de Queiroz", Universidade de São Paulo, 1993.

Apêndices

CÓDIGO FEITO EM LINGUAGEM R PARA ELABORAÇÃO DO CAPÍTULO 1

```
#####
# BIBLIOTECAS
#####
library(sm)
library(survival)
library(extRemes)
library(evd)
library(evdbayes)
library(tseries)
library(MASS)
options(OutDec = ",")

#####
# FUNÇÕES
#####

#####
# SELEÇÃO DO VALOR LIMIAR
#####

Mrl.statis = function (data, w, u, sigma, xi)
{
  mle = gpd.fit(data,threshold = u, method="BFGS")$mle
  muw = numeric(length(w))
  for(i in 1: length(w)){
    exc = data[data > w[i]]
    muw[i] = mean(exc - w[i], na.rm =T)
  }
  esp = (sigma + xi*(w - u)) / (1 - xi)
  espSim = (mle[1] + mle[2]*(w - u) ) / ( 1 - mle[2])
  T = muw - espSim + esp
  T
}

Mrl.muw = function(data, w)
{
  y = numeric(length(w))
  for(i in 1:length(w) ) {
    y[i] = mean(data[data > w[i]] - w[i] )
  }
  y
}

Env.mrlplot = function(data, u, titulo, nint)
{
  mle= gpd.fit (data,threshold=u,method="BFGS")$mle
  n = nint * length(data[data > u])
  m = length(data[data > u])
  amostra = matrix(numeric(n), nint, m)
  wmin<-u+0.0025*u

```

```

wmax<-max(data)-0.045*u
x = seq(wmin, wmax, l = 100)
b = length(x)
y = matrix(numeric(nint*b), nint, b)
muw = Mrl.muw(data,x)
quan = matrix(numeric(2*b), 2, b)
for(i in 1: nint) {
  amostra[i, ] = rgpd(n = m, loc = u, scale = mle[1],
    shape = mle[2])
  y[i,] = Mrl.statis(amostra[i,],w= x, u = u, mle[1],mle[2])
}
for(j in 1:b) {
  quan[,j] = quantile(y[,j], probs=c(0.025,0.975), na.rm =T)
}
if(missing(titulo)){
  titulo = ""
}
plot(x, muw, type="l", ylab = "Média dos excessos",
  xlab = "w", ylim = c(min(quan[1,]), max(quan[2,])),cex = 0.1 )
points(x, quan[1,], lty=2,type="l",cex=3)
points(x, quan[2,], lty=2,type="l",cex=3)

a = mle[2] /(1 - mle[2])
b = (mle[1] - mle[2]*u)/(1- mle[2])
abline(b, a, lty = 3,cex=3)
}

Mrl.plot = function (data,umin,umax,titulo)
{
  nint = 100
  if(missing(umin)) {
    umin = min(data)
  }
  if(missing(umax)){
    umax = max(data) - 0.1
  }
  if(missing(titulo)){
    titulo = ""
  }
  x = xl = xu = numeric(nint)
  u = seq(umin, umax, length = nint)
  for (i in 1:nint) {
    data = data[data > u[i]]
    x[i] = mean(data - u[i])
    n <- length(data)
    xu[i] <- x[i] + (qnorm((1 + 0.95)/2) * sd(data))/sqrt(n)
    xl[i] <- x[i] - (qnorm((1 + 0.95)/2) * sd(data))/sqrt(n)
  }
  plot(u, x, type = "l", xlab = "w",
    ylab = "Média dos Excessos", ylim = c(min(xl[!is.na(xl)]),
      max(xu[!is.na(xu)])))
  title(titulo)
}

```

```
#####
# MÉTODO DO DESAGRUPAMENTO
#####

IdeGro = function(dados, u, r)
{
  d = which(dados > u)
  groups = numeric(length(d))
  j = 1
  groups[1] = j
  for(i in 2: length(d) ) {
    condit = (d[i] - d[i - 1] - 1)
    if( condit < r) {
      groups[i] = groups[i - 1]
    } else {
      j = j + 1
      groups[i] = j
    }
  }
  return(cbind(dados[dados > u], groups))
}

ResGro = function(mgrupo, r)
{
  ideGro = unique(mgrupo[,2])
  k = length(mgrupo[,1])
  J = length(ideGro)
  maxGro = numeric(J)
  xdat = numeric
  for(j in 1:J ) {
    boolea = which(mgrupo[,2] == ideGro[j])
    maxGro[j] = max(mgrupo[boolea,1])
  }
  resul = list(
    ncluster = length(ideGro),
    dcluster = maxGro ,
    excedentes = k,
    extremalindex = J / k)
  return(resul)
}

Declust= function(dados, u, r)
{
  gro = IdeGro(dados, u, r)
  maxExc = ResGro(gro, r)
  return(maxExc)
}

#####
# ESTIMAÇÃO POR MEIO DA MÁXIMA VEROSSIMILHANÇA - GPD
#####
```

```

Gpd.likelihood= function(mle, data, u)
{
  if(abs(mle[2]) < 10^(-4)) {
    l = length(data) * log(mle[1]) + sum(data - u)/mle[1]
  } else {
    y = (data - u)/mle[1]
    y = 1 + mle[2] * y
    if(any(y <= 0) || mle[1] <= 0) {
      l = 10^6
    } else {
      l = length(data) * sum( log(mle[1])) + sum(log(y))
        * (1/mle[2] + 1)
    }
  }
  l
}

Gpd.fit = function (data, u, method, distribution, r)
{
  maxit = 10000
  if (missing(method) ) {
    stop("especifique o método!!!")
  }
  if (missing(distribution) ) {
    distribution = "GPD"
  }
  z = list()
  z$threshold = u
  z$n = length(data)
  z$exc = data[data > u]
  z$nexc = length(z$exc)
  z$rate = z$nexc/z$n
  in2 = sd((z$exc-u))
  in1 = 0.1
  ini = c(in2,in1)
  if(method == "independence") {
    if(distribution == "GPD") {
      opt = optim(ini, Gpd.likelihood, data = z$exc, u = u,
        method = "BFGS", hessian=T)
      z$cov = solve(opt$hessian)
      z$se = sqrt(diag(z$cov))
      z$nllh= opt$value
      z$mle = opt$par
    } else
    if(distribution == "EXP") {
      expmle = mean(z$exc - u)
      z$mle = expmle
      nllh = -z$nexc* log(expmle) - (1/expmle)*sum(z$exc-u)
      z$nllh = nllh
    } else {
      stop("Erro")
    }
  } else
  if(method == "declustering") {

```

```

dclust = Declust(data, u, r)
z$dclust = dclust$dcluster
z$J = dclust$ncluster
if(distribution=="GPD") {
  z$extremalindex = dclust$extremalindex
  opt = optim(ini, Gpd.likelihood, data = z$dclust, u = u,
    method = "BFGS", hessian=T)
  z$cov = solve(opt$hessian)
  z$se = sqrt(diag(z$cov))
  z$llh = opt$value
  z$mle = opt$par
} else
if(distribution=="EXP") {
  z$extremalindex = dclust$extremalindex
  expmle = mean(z$dclust - u)
  z$mle = expmle
  llh = -z$J * log(z$mle) - (1/z$mle)*sum(z$dclust-u)
  z$llh = -llh
} else {
  stop("Erro")
}
} else {
  stop("Erro")
}
z
}

#####
# NIVEL DE RETORNO
#####

Level.return = function(z, T, type ) {
  if(missing(type)) {
    stop("deve-se especificar o tipo !!!")
  }
  mle = z$mle
  u = z$threshold
  m = T * 365.25
  p = z$rate
  theta = z$extremalindex
  if(type == "GPD") {
    return(u + (mle[1]/mle[2]) * ((m * p * theta)^mle[2] - 1))
  } else {
    return(u + mle * log(m * theta * p))
  }
}

#####
# INTERVALOS DE CONFIANÇA PARA OS PARÂMETROS -
# PERFIL DA VEROSSIMILHANÇA
#####

Gpd.likecondit= function(xi, sigma, data, u)
{

```

```

if(abs(xi) < 10^(-4)) {
  l = -length(data) * log(sigma) - sum(data - u)/sigma
} else {
  y = (data - u)/sigma
  y = 1 + xi * y
  if(any(y <= 0) || sigma <= 0) {
    l = -10^6
  } else {
    l = -length(data) * log(sigma) - sum(log(y)) *
      (1/xi + 1)
  }
}
l
}

Gpd.newlikecondit= function(xi, xp, m, p , u, theta, data)
{
  sig = (xi * (xp - u))/((m * p * theta)^xi - 1)
  l = Gpd.likecondit(xi,sig, data,u=u)
  l
}

Gpd.prof = function (z, xlow, xup, method, T, xlab, ylab,
dclust)
{
  if(missing(dclust)) {
    dclust = 0
  }
  conf = 0.95
  nint = 100
  if(missing(method)) {
    stop("deve-se especificar o método!!!")
  }
  if( ( method == "return level" && missing(T) )) {
    stop("deve-se especificar T!!!")
  }
  if(dclust == 0 ) {
    theta = 1
    xdat = z$exc
  } else {
    theta = z$extremalindex
    xdat = z$dclust
  }
  mle = z$mle
  u = z$threshold
  v = numeric(nint)
  x = seq(xlow, xup, length = nint)
  if(method == "sigma") {
    ini = z$mle[2]
  } else
  if(method == "xi") {
    ini = z$mle[1]
  } else {
    ini = z$mle[2]
  }
}

```

```

    pi = z$rate
    nm = T * 365.25
  }
  for (i in 1:nint) {
    if(method == "sigma") {
      opt = optim(ini, Gpd.likecondit, sigma = x[i],
        data = xdat, u = u, method = "BFGS",
        control=list(fnscale=-1))
    } else
    if(method == "xi") {
      opt = optim(ini, Gpd.likecondit, xi = x[i],
        data = xdat, u = u, method = "BFGS",
        control=list(fnscale=-1))
    } else {
      opt = optim(par = ini, fn = Gpd.newlikecondit,
        xp = x[i], data = xdat, u=u, p = pi, m = nm,
        theta = theta, method = "BFGS",
        control=list(fnscale=-1))
    }
    v[i] = opt$value
  }
  plot(x, v, type = "l", xlab = xlab, ylab = ylab)
  ma <- -z$nlh
  abline(h = ma - 0.5 * qchisq(conf, 1), lty = 1)
  invisible()
}

#####
# INTERVALOS DE CONFIANÇA PARA OS PARÂMETROS - DEVIANCE
#####

Get.Data = function(z, method)
{
  if(method == "independence") {
    data = z$exc
    J = z$nexc
    theta = 1
  } else {
    data = z$dclust
    J = z$J
    theta = z$extremalindex
  }
  res = list(data=data, J=J, theta = theta)
  res
}

Exp.likelihood = function(xlow, xup, z, method, ylab, xlab)
{
  typedata = Get.Data(z, method)
  mle = z$mle
  data = typedata$data
  J = typedata$J
  u = z$threshold

```

```

l = function(x) {2 * ((-z$nlh) + J*log(x)+(1/x)*
  sum(data-u) )}
curve(l, x = c(xlow, xup), ylab = ylab, xlab = xlab)
abline(h=0.5 * qchisq(0.95, 1))
}

Exp.newlikelihood = function(xlow, xup, z, method, ylab,
  xlab, T)
{
  typedata = Get.Data(z, method)
  data = typedata$data
  J = typedata$J
  theta = typedata$theta
  u = z$threshold
  p = z$rate
  l =function(x){
    mle = ( u - x) / log( 1 / ( p * theta * T * 365.25 ))
    l = 2 * ((-z$nlh) + J*log(mle)+(1/mle)* sum(data-u) )
    l
  }
  curve(l, x =c(xlow,xup), ylab = ylab, xlab = xlab)
  abline(h=0.5 * qchisq(0.95, 1))
}

Dev.intconf = function(xlow, xup, z, method,
  type, T, ylab, xlab)
{
  if(type == "return level") {
    Exp.newlikelihood(xlow, xup, z, method, ylab, xlab, T)
  } else {
    Exp.likelihood(xlow, xup, z, method, ylab, xlab)
  }
}

#####
# QQ-PLOT / PP-PLOT
#####

Fgraphics = function(data, u, type, mle){
  if(type == "PPplot") {
    emp = sort(data)
    x = (1:length(data))/(length(data)+1)
    y = Pgpd(emp, mle, u)
  } else {
    emp = (1:length(data)/(length(data) + 1))
    x = sort(data)
    y = Qgpd(emp, mle, u)
  }
  eixo = list(x = x, y = y, emp= emp)
}

Rgpd = function(n, mle, u) {
  loc = u

```



```

    scale = mle[1]
    if(length(mle) == 1) {
      shape = 0
    } else {
      shape = mle[2]
    }
    rgpd(n, loc = u, scale = scale, shape = shape)
  }

Pgpd = function(x, mle, u) {
  loc = u
  scale = mle[1]
  if(length(mle) == 1) {
    shape = 0
  } else {
    shape = mle[2]
  }
  pgpd(x, loc = u, scale = scale, shape = shape)
}

Qgpd = function(x, mle, u) {
  loc = u
  scale = mle[1]
  if(length(mle) == 1) {
    shape = 0
  } else {
    shape = mle[2]
  }
  qgpd(x, loc = u, scale = scale, shape = shape)
}

Gen.env = function(z, method, type, distr, nint = 1000,
  xlab, ylab)
{
  typedata = Get.Data(z, method)
  mle = z$mle
  u = z$threshold
  data = sort(typedata$data)
  n = length(data)
  npar = length(mle)
  eixo = Fgraphics(data, u, type, mle)
  emp =
  x = eixo$x
  y = eixo$y
  emp = eixo$emp
  b = length(x)
  asimul = matrix(numeric(nint*n), nint, n)
  parsim = matrix(numeric(nint*length(mle)), nint,
    length(mle))
  quan = matrix(numeric(b*2), 2, b)
  ysim = matrix(numeric(nint*b), nint, b)
  for(i in 1:nint) {
    asimul[i,] = Rgpd(n, mle, u)
    parsim[i,] = Gpd.fit(data = asimul[i,], u = u,

```

```

        method="independence", distribution=distr)$mle
    if(type == "PPplot" ){
        ysim[i, ] = Pgpdp(emp, parsim[i,], u)
    } else
        if(type == "QQplot" ) {
            ysim[i, ] = Qgpdp(emp, parsim[i,], u)
        }
    }
    for(j in 1:b) {
        quan[,j] = quantile(ysim[,j], probs=c(0.025,0.975),
            na.rm =T)
    }
    plot(x,y, xlab=xlab, ylab=ylab, cex = 0.5, pch = 16 )
    points(x, quan[1,], col=2,type="l",cex=3)
    points(x, quan[2,], col=2,type="l",cex=3)
    abline(0, 1, col = 4,cex=3)
}

#####
#  ANALISE DE DADOS
#####

#####
#  CARREGAR OS DADOS
#####
pathpre = "C:/Documentos/Renato/Pós graduação Esalq
/2ano_mestrado/"
file = "Conjunto de dados/Vazões/vazaolim.txt"
path = paste(pathpre,file, sep="")
data = read.table(path,h=T)
attach(data)

#####
#  SELEÇÃO DO VALOR LIMIAR
#####

Mrl.plot(vazao)
par(mfrow=c(1,2))
Env.mrlplot(data = vazao, u = 400, nint = 99)
title("(a)")
Env.mrlplot(data = vazao, u = 700, nint = 99)
title("(b)")

#####
#  ANALISE EXPLORATORIA
#####

temp = c( as.character(44:99),paste(c(0,0,0),1:3,sep=""))
par(mar=c(3.9,3.9,3.9,3.9))
plot.ts(log(vazao[vazao > 400]),type="l",
    xaxt = "n",xlab="anos",ylab =
expression(paste("log", " ", , "(", "vazão", " ",
    "média", " ", "diária", ")") ))
)

```

```

eixox = axis(1,seq(1,813,14),
labels =temp,   tcl =-0.005, cex = 0.5 )

#####
# MÉTODO DO DESAGRUPAMENTO
#####

dadosdcr1 = Declust(vazao, 400, 1)
dadosdcr2 = Declust(vazao, 400, 2)
dadosdcr7 = Declust(vazao, 400, 7)

#####
# GRÁFICO DE AUTOCORRELAÇÃO
#####

par(mfrow=c(2,2))

acf(vazao[vazao > 400],
lag.max =50, ylab="função de autocorrelação",
main = "", xlab = "lag")
title("(a)")

acf(dadosdcr1\,$dcluster ,
lag.max =50, ylab="função de autocorrelação",
main = "", xlab = "lag")
title("(b) r = 1")

acf(dadosdcr2\,$dcluster ,
lag.max =50, ylab="função de autocorrelação",
main = "", xlab= "lag" )
title("(c) r = 2")

acf(dadosdcr7\,$dcluster,
lag.max =50, ylab="função de autocorrelação",
main = "", xlab = "lag")
title("(d) r = 4")

#####
# ESTIMAÇÃO POR MEIO DO MÉTODO DA MÁXIMA VEROSSIMILHANÇA DA GPD
#####

mod1 = Gpd.fit(data=vazao,u=400,method="declustering",r=1)
mod1$mle

#####
# INTERVALO DE CONFIANÇA
#####

nf<-layout(matrix(c(1,1,2,2,3,4,4,5), 2, 4, byrow=TRUE))
layout.show(nf)

Gpd.prof (z = mod1, xlow = 100, xup = 225,
method = "sigma", xlab = "",

```

```

  ylab = "log - verossimilhança",dclust=1)
title(" (a) ")
abline(v = 130.0173,lty = 2)
abline(v = 199.6440,lty = 2)
abline(v = mod1$mle[1],lty = 2)
mtext(side = 1, expression(hat(sigma)),
at = c( mod1$mle[1], -1095.429 ), padj = 2.7, cex = 0.8 )
mtext(side = 1,paste("L.I"),at = c(130.0173 ),
padj = 3.3, cex = 0.8 )

mtext(side = 1,paste("L.S"),at = c(199.6440 ),
padj = 3.3, cex = 0.8 )

mtext(side = 1,expression(sigma),at = c(220),
padj = 4.0, cex = 1 )

Gpd.prof (z = mod1, xlow = -0.2,xup = 0.15,
  method = "xi", xlab = "",
  ylab = "log - verossimilhança",dclust=1)
title(" (b) ")
abline(v = -0.1828399,lty = 2)
abline(v = 0.1336557,lty = 2)
abline(v = mod1$mle[2],lty = 2)
mtext(side = 1, expression(hat(xi)),
at = c( mod1$mle[2], -1095.444 ), padj = 2.0, cex = 0.8 )
mtext(side = 1,paste("L.I"),at = c(-0.1828399 ),
padj = 3.3, cex = 0.8 )
mtext(side = 1,paste("L.S"),at = c( 0.1336557 ),
padj = 3.3, cex = 0.8 )
mtext(side = 1,expression(xi),at = c(.17),
padj = 2.6, cex = 1 )

#####
# ESTIMAÇÃO POR MEIO DO MÉTODO DA MÁXIMA VEROSSIMILHANÇA
# DISTRIBUIÇÃO EXPONENCIAL
#####

mod2= Gpd.fit(data=vazao,u=400,method="declustering",r=1,
distribution="EXP")

#####
# INTERVALO DE CONFIANÇA
#####

plot(0,0,axes =F,type="n",frame.plot=F,xlab="",ylab="")

dev = Dev.intconf(100, 200, mod2, method= "declustering",
type="sigma", ylab ="deviance", xlab = "")
title(" (c) ")
segments(139.7019,-1,139.7019,1.946873, lty=2)
segments(171.6616,-2,171.6616,1.946873, lty=2)
mtext(side = 1,paste("L.I"),at = c(139,7019 ),
padj = 3.3, cex = 0.8 )

```

```

mtext(side = 1,paste("L.S"),at = c(171,6616),
padj = 3.3, cex = 0.8 )
mtext(side = 1,expression(sigma),at = c(190),
padj = 4.0, cex = 1 )
plot(0,0,axes =F,type="n",frame.plot=F,xlab="",ylab="")

```

```

#####
# NÍVEIS DE RETORNO - GPD
#####

```

```

T = c(5, 10, 50, 60, 100)
nretorno1 = Level.return(mod1, T, "GPD" )
nretorno1 = nr(400,model$mle,N,365.25,model$rate)
nretorno1 = round(nretorno1,0)
attr(nretorno1,"class")= "matrix"
attr(nretorno1,"dim")= c(1,5)
colnames(nretorno1) = c("5 anos","10 anos", "50 anos",
"60 anos", "100 anos")
nretorno1

```

```

#####
# INTERVALO DE CONFIANÇA NÍVEIS DE RETORNO - GPD
#####

```

```

Gpd.prof (z = mod1, xlow = 700, xup = 1000,
method = "return level", T = 5, xlab = "",
ylab = "log - verossimilhança",dclust=1)

```

```

Gpd.prof (z = mod1, xlow = 800, xup = 1100,
method = "return level", T = 10, xlab = "",
ylab = "log - verossimilhança",dclust=1)

```

```

Gpd.prof (z = mod1, xlow = 900, xup = 1500,
method = "return level", T = 50, xlab = "",
ylab = "log - verossimilhança",dclust=1)

```

```

Gpd.prof (z = mod1, xlow = 900, xup = 1500,
method = "return level", T = 60, xlab = "",
ylab = "log - verossimilhança",dclust=1)

```

```

Gpd.prof (z = mod1, xlow = 950, xup = 1700,
method = "return level", T = 100, xlab = "",
ylab = "log - verossimilhança",dclust=1)

```

```

#####
# NÍVEIS DE RETORNO - EXPONENCIAL
#####

```

```

nretorno2 = Level.return(mod2, T, "EXP" )
nretorno2 = round(nretorno2,0)
attr(nretorno2,"class")= "matrix"
attr(nretorno2,"dim")= c(1,5)
colnames(nretorno2) = c("5 anos","10 anos","50 anos",

```

```

    "60 anos", "100 anos")
nretorno2

#####
# INTERVALO DE CONFIANÇA NÍVEIS DE RETORNO - EXPONENCIAL
#####

Dev.intconf(700, 900, mod2, method= "declustering",
  type="return level", T =5, ylab="", xlab="")
Dev.intconf(850, 1150, mod2, method= "declustering",
  type="return level", T =10, ylab="", xlab="")
Dev.intconf(1050, 1350, mod2, method= "declustering",
  type="return level", T =50, ylab="", xlab="")
Dev.intconf(1050, 1350, mod2, method= "declustering",
  type="return level", T =60, ylab="", xlab="")
Dev.intconf(1050, 1650, mod2, method= "declustering",
  type="return level", T =100, ylab="", xlab="")

#####
# PP-PLOT E QQ-PLOT
#####

par(mfrow=c(2,2))

par(mar=c(3.9,3.9,3.9,3.9))
nf<-layout(matrix(c(1,2,3,4), 2, 2, byrow=TRUE))
layout.show(nf)
Gen.env(mod1, method = "declustering" , type = "PPplot",
  distr= "GPD", nint = 1000, "Empírico", "Modelo")
title("(a)")
Gen.env(mod2, method = "declustering" , type = "PPplot",
  distr= "EXP", nint = 1000, "Empírico", "Modelo")
title("(b)")
Gen.env(mod1, method = "declustering" , type = "QQplot",
  distr= "GPD", nint = 1000, "Empírico", "Modelo")
title("(c)")
Gen.env(mod2, method = "declustering" , type = "QQplot",
  distr= "EXP", nint = 1000, "Empírico", "Modelo")
title("(d)")

#CÓDIGO FEITO EM LINGUAGEM R PARA ELABORAÇÃO DO CAPÍTULO 2

#####
# CARREGAR OS DADOS
#####

options(OutDec = ",")
pathpre = "C:/Documentos/Renato/Pós graduação Esalq/
  2ano_mestrado/"
file = "Conjunto de dados/Ventos/ventos.txt"
path = paste(pathpre,file, sep="")
data = read.table(path,h=T)

```

```

attach(data)
janeiro = janeiro[janeiro!="NaN"]
setembro = setembro[setembro!="NaN"]

#####
# BIBLIOTECAS
#####

library(evd)
library(extRemes)
library(sm)
library(MASS)
library(lawstat)

#####
# FUNÇÕES
#####

#####
# RUN TEST
#####

runs.test =
function (y)
{
  alternative <- match.arg(alternative)
  DNAME = deparse(substitute(y))
  y <- na.omit(y)
  med <- median(y, na.rm = TRUE)
  y = y[y != med]
  q <- rep(0.05, length(y))
  p <- rep(-0.05, length(y))
  d <- y
  q[I(d < med) | I(d == med)] <- NA
  p[I(d >= med)] <- NA
  m <- length(na.omit(q))
  n <- length(na.omit(p))
  R <- 1
  s <- sign(y - med)
  for (k in 1:(length(y) - 1)) {
    if (s[k] != s[k + 1]) {
      R <- R + 1
    }
  }
  list(statistic = STATISTIC, p.value = p.value,
       n1 =m , n2 =n,r = R)
}

#####
# VALORES INICIAIS
#####

Chumod1 = function(data) {

```

```

aa = kmeans(data,centers=2)$cluster
x1 = data[aa==which(table(aa)==max(table(aa)))]
x2 = data[aa==which(table(aa)==min(table(aa)))]
p = length(x1)/length(data)
sigma = sqrt(6 * var(data))/pi
mu1 = mean(x1) - 0.57722 * sigma
mu2 = mean(x2) - 0.57722 * sigma
mle = c(mu1,sigma,mu2, p)
mle
}

Chumod2 = function(data) {
  aa = kmeans(data,centers=2)$cluster
  x1 = data[aa==which(table(aa)==max(table(aa)))]
  x2 = data[aa==which(table(aa)==min(table(aa)))]
  p = length(x1)/length(data)
  sigma1 = sqrt(6 * var(x1))/pi
  sigma2 = sqrt(6 * var(x2))/pi
  mu1 = mean(x1) - 0.57722 * sigma1
  mu2 = mean(x2) - 0.57722 * sigma2
  mle = c(mu1,sigma1,mu2, sigma2,p)
  mle
}

#####
# MISTURA DE DUAS DISTRIBUIÇÕES DE GUMBEL COM MESMO PARÂMETROS
# DE ESCALA
#####

modgum1 = function(mle,x)
{
  if((mle[2] <= 0) || (mle[4] < 0) ||
    (mle[4] > 1) || (mle[3] < mle[1])){
    return(0)
  } else {
    mle[4]*dgumbel(x, loc = mle[1], scale = mle[2])+
    (1-mle[4])* dgumbel(x, loc = mle[3], scale = mle[2])
  }
}

#####
# MISTURA DE DUAS DISTRIBUIÇÕES DE GUMBEL COM DISTINTOS
# PARÂMETROS DE ESCALA
#####

modgum2 = function(mle,x)
{
  if((mle[2] <= 0) || (mle[4] <= 0) || (mle[5] < 0) ||
    (mle[5] > 1) || (mle[3] < mle[1])){
    return(0)
  } else {
    mle[5]*dgumbel(x, loc = mle[1], scale = mle[2])+
    (1-mle[5])* dgumbel(x, loc = mle[3], scale = mle[4])
  }
}

```



```

}

#####
# LOG - VEROSSIMIHANÇA MISTURA DE DUAS DISTRIBUIÇÕES DE GUMBEL
# COM MESMO PARÂMETRO DE ESCALA
#####

logmod1 = function(mle,x)
{
  if((mle[2] <= 0) || (mle[4] < 0) || (mle[4] > 1) ||
    (mle[3] < mle[1])){
    return(-10^6)
  } else {
    sum(log(mle[4]*dgumbel(x, loc = mle[1],scale = mle[2],log=F)+
      (1-mle[4])* dgumbel(x, loc = mle[3], scale = mle[2],log=F)))
  }
}

#####
# LOG - VEROSSIMIHANÇA MISTURA DE DUAS DISTRIBUIÇÕES DE GUMBEL
# COM DISTINTOS PARÂMETRO DE ESCALA
#####

logmod2 = function(mle,x)
{
  if((mle[2] <= 0) || (mle[4] <= 0) || (mle[5] < 0) ||
    (mle[5] > 1) || (mle[3] < mle[1])){
    return(-10^6)
  } else {
    sum(log(mle[5]*dgumbel(x, loc = mle[1],scale = mle[2],log=F)+
      (1-mle[5])* dgumbel(x, loc = mle[3], scale = mle[4],log=F)))
  }
}

#####
# ALGORITMO ECM - MISTURA DE DUAS DISTRIBUIÇÕES DE GUMBEL
#####

#####
# PASSO M - (mul, sigmal)
#####

MulSigmal = function(mle, x,z, aux ) {
  if ((mle[2] <= 0) || (aux[2] <= 0) || (mle[1] >= aux[1])) {
    return(-10^6)
  } else {
    y <- (x - mle[1])/mle[2]
    yaux = (x - aux[1])/aux[2]
    l = sum(z*log(aux[3])) -sum(z*log(mle[2])) - sum(z*y) -
      sum(z*exp(-y))+sum((1 - z)*log((1 - aux[3]))))
      -sum((1 - z) *log(aux[2]))- sum((1 - z)*yaux)
      - sum((1 - z)*exp(-yaux))
    l = max(l, -10^6, na.rm = T)
  }
}

```

```

1

}

}

#####
# PASSO M - (mu2, sigma2)
#####

Mu2Sigma2 = function(mle, x,z, aux ) {
  if ((mle[2] <= 0) || (aux[2] <= 0) || (mle[1] <= aux[1])) {
    return(-10^6)
  } else {
    y <- (x - mle[1])/mle[2]
    yaux = (x - aux[1])/aux[2]
    l = sum((1-z)*log(1-aux[3])) -sum((1 - z)*log(mle[2]))
      - sum((1 - z)*y) - sum((1 - z)*exp(-y))
      + sum(z*log(aux[3])) -sum(z*log(aux[2])) - sum(z*yaux)
      - sum(z*exp(-yaux))
    l = max(l, -10^6, na.rm = T)
    l
  }
}

#####
# PASSO M - (mu1, mu2, sigma)
#####

logM = function(mle,x,z, aux)
{
  if((mle[2] <= 0) || (mle[3] <= mle[1]) ){
    return(-10^6)
  } else {
    l = sum(z*log(aux*dgumbel(x, loc = mle[1],scale = mle[2],log=F)))+
      sum((1-z)*log((1-aux)*
        dgumbel(x, loc = mle[3], scale = mle[2],log=F)))
    l = max(l, -10^6, na.rm = T)
    l
  }
}

#####
# AJUSTE DE MISTURAS
#####

ecm = function(obj) {
  x = obj$data
  h = obj$choose
  resul = vector("list", 7)
  mleant = obj$ini[h,]
  resul$ini = mleant
  k = length(mleant)
  mleatual = numeric(k)

```

```

i = 0
lant = 1
latual = 0
while( ( abs(latual - lant) > 10^(-5) ) && ( i < 1000 ) ) {
  z = mleant[k]* dgumbel(x ,mleant[1], mleant[2]) / obj$mod(mleant, x)
  mleatual[k] = mean(z)
  lant = obj$log(mleant, x)
  if(k == 5) {
    mleatual[1:2] = optim(par = mleant[1:2], x = x, z = z, aux
      = c(mleant[3:4],mean(z)),
      fn = obj$M1step, method="BFGS", control=list(fnscale=-1),
      hessian = T)$par
    mleatual[3:4] = optim(par = mleant[3:4], x = x, z = z,
      aux = mleatual[c(1,2,5)], fn = obj$M2step, method="BFGS",
      control=list(fnscale=-1),hessian = T)$par
  } else {
    mleatual[1:3] = optim(par = mleant[1:3], x = x, z = z,
      aux = mleatual[4], fn = obj$Mstep, method="BFGS",
      control=list(fnscale=-1),hessian = T)$par
  }
  mleant = mleatual
  latual = obj$log (mleatual, x)
  i = i + 1
}
resul$k = length(mleatual)
resul$iter = i
resul$mle = mleatual
resul$z = z
resul$lh = latual
resul$n = length(x)
resul
}

Mixture.fit = function(obj) {
  n = nrow(obj$ini)
  k = ncol(obj$ini)
  mod = vector("list", n)
  l = numeric(n)
  ini = matrix(numeric(n*k),n,k)
  mle = matrix(numeric(n*(k)),n,(k))
  for(i in 1:n){
    obj$choose = i
    mod[[i]] = ecm(obj)
    ini[i,] = mod[[i]]$ini
    mle[i,] = mod[[i]]$mle
    l[i] = mod[[i]]$lh
  }
  othermle = mle
  otherl = l
  ini = ini[which( l == max(l)),]
  mle = mle[which(l == max(l)),]
  l = max(l)
  resul = list(ini = ini, mle = mle,
    lh = l, othermle = othermle, otherl = otherl, data = obj$data)
}

```

```

    resul
  }

#####
# INTERVALO DE CONFIANÇA DOS PARÂMETROS BOOTSTRAP PARAMÉTRICO - GUMBEL
#####

Gumbel.like= function(mu, sigma, data)
{
  data = (data - mu)/sigma
  if (is.infinite(mu) || sigma <= 0) {
    l <- -10^6
  } else {
    l <- -(length(data) * log(sigma))
      - sum(exp(-data)) - sum(data)
  }
  l
}

Gumbel.prof.mu= function (z, xlow, xup, nint)
{
  mle = z$mle
  xdat = z$data
  v = numeric(nint)
  x = seq(xlow, xup, length = nint)
  ini = sqrt(6 * var(xdat))/pi
  for (i in 1:nint) {
    opt = optim(par = ini, fn = Gumbel.like, data = xdat, mu=x[i],
      method = "BFGS", control=list(fnscale=-1))
    v[i] = opt$value
  }
  plot(x, v, type = "l", xlab = expression(mu),
    ylab = "log - verossimilhança")
  ma <- z$lh
  abline(h = ma - 0.5 * qchisq(0.95, 1), lty = 1)
  invisible()
}

Gumbel.prof.sigma = function (z, xlow, xup, nint)
{
  mle = z$mle
  xdat = z$data
  sig = sqrt(6 * var(xdat))/pi
  in1 = mean(xdat) - 0.57722 * sig
  v = numeric(nint)
  x = seq(xlow, xup, length = nint)
  ini = sqrt(6 * var(xdat))/pi
  for (i in 1:nint) {
    opt = optim(par = ini, fn = Gumbel.like, data = xdat, sigma = x[i],
      method = "BFGS", control=list(fnscale=-1))
    v[i] = opt$value
  }
}

```

```

    plot(x, v, type = "l", xlab = expression(sigma),
         ylab = "log - verossimilhança")
    ma <- z$lh
    abline(h = ma - 0.5 * qchisq(0.95, 1), lty = 1)
    invisible()
}

#####
# INTERVALO DE CONFIANÇA DOS PARÂMETROS BOOTSTRAP PARAMÉTRICO - MISTURAS
#####

ICpBoot = function(B,obj, label) {
  mle = obj$mle
  n = length(obj$data)
  k = length(obj$mle)
  intconf = matrix(numeric(k*2), 2, k)
  boot =matrix(numeric(k*B), B, k)
  for(i in 1:B) {
    sample = obj$simul(n=n,mle)
    obj$data = sample
    obj$ini = matrix(obj$mle,1,k,byrow=T)
    boot[i, ] = Mixture.fit(obj)$mle
  }
  intconf= apply(boot, 2, quantile , probs = c(0.025, 0.975))
  intconf = t(intconf)
  colnames(intconf) = c("L.I", "L.S")
  rownames(intconf) = label
  intconf
}

#####
# AIC E BIC
#####

AIC = function(mod) {
  l = mod$lh
  k = length(mod$mle)
  aic = -2*l + 2*k
  aic
}

BIC = function(mod) {
  l = mod$lh
  k = length(mod$mle)
  n = length(mod$data)
  bic = -2*l + k*log(n)
  bic
}

#####
# RAZÃO DE VEROSSIMILHANÇAS BOOTSTRAP PARAMÉTRICO
#####

```

```

Boot = function(B,obj, ...) {
  mleH0 = obj$mleH0
  lH0 = obj$lhH0
  mleHa = obj$mle
  lHa = obj$lh
  k = length(mleH0)
  n = length(obj$data)
  mod = obj$mod
  boot = vector("list", l=B)
  boot$lambda = 2*(lHa - lH0)
  for(i in 1:B) {
    sample = obj$sample(n=n,...)
    obj$ini =matrix( mleHa, 1, length(mleHa), byrow=T)
    boot[[i]]$H0 = gum.fit(sample, "BFGS")
    obj$data = sample
    boot[[i]]$Ha = Mixture.fit(obj)
    boot[[i]]$lamboot = 2*(boot[[i]]$Ha$lh + boot[[i]]$H0$nlh )
    boot$pvalue = (1 /B)*sum(boot$lambda < boot$lamboot)
  }
  boot
}

#####
# QQ-PLOT  E PP-PLOT
#####

fdp1 = function(y, mu1, mu2, sigma, pi){
  pi*dgumbel(y, loc = mu1,scale = sigma)+(1-pi)*
  dgumbel(y, loc = mu2, scale = sigma)
}

fd1 = function(y, mu1, mu2, sigma, pi,p){
  integrate(fdp1, mu1 = mu1, mu2 = mu2, sigma= sigma,
  pi =pi, lower = -Inf, upper =y )$value - p
}

Dmixgumbell1 = function(x, mle) {
  fdp1(x, mle[1], mle[3], mle[2], mle[4])
}

Pgumbel = function(x, mle){
  pgumbel(x, loc=mle[1],scale = mle[2])
}

Pmixgumbell1 = function(x, mle) {
  integrate(fdp1, mu1 = mle[1], mu2 = mle[3],
  sigma = mle[2], pi = mle[4],lower = -Inf, upper =x )$value
}

Qgumbel = function(p, mle){
  qgumbel(p, loc=mle[1],scale = mle[2])
}

Qmixgumbell1 = function(p, mle, min, max) {

```

```

    if(missing(min)) { min = 0}
    if(missing(max)) {max = 2.5*mle[3] }
    uniroot(fdl, mu1 = mle[1], mu2 = mle[3],
      sigma= mle[2], pi = mle[4], p = p, interval =c(min,max),
      tol = 0.001)$root
  }

Rmixgumbell1 = function(n,mle) {
  z = rbinom (n ,1, mle[4])
  x = numeric(n)
  for(i in 1:n) {
    if (z[i] == 1) {
      x[i] = rgumbel (1, mle[1], mle[2])
    } else {
      x[i] = rgumbel (1, mle[3], mle[2])
    }
  }
  return (x)
}

Rmixgumbel2 = function(n,mle) {
  z = rbinom (n ,1, mle[5])
  x = numeric(n)
  for(i in 1:n) {
    if (z[i] == 1) {
      x[i] = rgumbel (1, mle[1], mle[2])
    } else {
      x[i] = rgumbel (1, mle[3], mle[4])
    }
  }
  return (x)
}

QQplot = function (obj, s, min, max)
{
  x = sort(obj$data)
  y = numeric(length(x))
  n = length(x)
  B = 1000
  asimul = matrix(numeric(B*n),B, n)
  parsim = matrix(numeric(B*length(obj$mle)),B,
    length(obj$mle))
  quan = matrix(numeric(n*2), 2,n)
  ysim = matrix(numeric(1000*n),B, n)
  for(j in 1:B) {
    if(length(obj$mle) == 2) {
      asimul[j,] = obj$simul(n, obj$mle[1], obj$mle[2])
      parsim[j,] = gum.fit(asimul[j,])$mle
      for(i in 1:length(x)) {
        ysim[j,i] = obj$quantil(c(i /(1 + n)),parsim[j,])
      }
    } else {
      asimul[j,] = obj$simul(n, obj$mle)
      obj$ini = matrix(obj$mle,1, length(obj$mle), byrow=T)
    }
  }
}

```

```

    obj$data = asimul[j,]
    parsim[j,] = Mixture.fit(obj)$mle
    for(i in 1:length(x)) {
        ysim[j,i] = obj$quantil(c(i / (1 + n)), parsim[j,],
            min, max)
    }
}
}
for(i in 1:length(x)) {
    quan[,i] = quantile(ysim[,i], probs=c(0.025,0.975),
        na.rm =T)
    y[i] = obj$quantil(c(i / (1 + n)), obj$mle)
}
plot(x, y, xlab = "Empírico", ylab = "Modelo",
    main = s, cex=0.7, pch=16)
points(x, quan[1,], col=2, type="l", cex=3)
points(x, quan[2,], col=2, type="l", cex=3)
abline(0, 1, col = 4)
}

PPplot = function (obj, s, ...)
{
    y = numeric(length(obj$data))
    n = length(obj$data)
    B = 1000
    x = 1:n / (n + 1)
    asimul = matrix(numeric(B*n), B, n)
    parsim = matrix(numeric(B*length(obj$mle)), B,
        length(obj$mle))
    quan = matrix(numeric(n*2), 2, n)
    ysim = matrix(numeric(1000*n), B, n)
    for(j in 1:B) {
        asimul[j,] = obj$simul(n, ...)
        if(length(obj$mle) == 2) {
            parsim[j,] = gum.fit(asimul[j,])$mle
            ord = sort(asimul[j,])
            for(i in 1:length(x)) {
                ysim[j,i] = obj$prob(ord[i], parsim[j,])
            }
        } else {
            obj$ini = matrix(obj$mle, 1, length(obj$mle), byrow=T)
            parsim[j,] = Mixture.fit(obj)$mle
            ord = sort(asimul[j,])
            for(i in 1:length(x)) {
                ysim[j,i] = obj$prob(ord[i], parsim[j,])
            }
        }
    }
}
orde = sort(obj$data)
for(i in 1:length(x)) {
    quan[,i] = quantile(ysim[,i], probs=c(0.025,0.975),
        na.rm =T)
    y[i] = for(i in 1:n) { y[i] = obj$prob(orde[i], obj$mle) }
}

```



```

    }
    plot(x, y , xlab = "Empírico", ylab = "Modelo",
         main = s, cex=0.7, pch=16)
    points(x, quan[1,], col=2,type="l",cex=3)
    points(x, quan[2,], col=2,type="l",cex=3)
    abline(0, 1, col = 4)
}

#####
# NÍVEL DE RETORNO
#####

Level.return = function(z, T ) {
  mle = z$mle
  return(mle[1]- mle[2]*log(-log(1 - 1/T)) )
}

#####
# INTERVALOS DE CONFIANÇA PARA OS NÍVEIS DE RETORNO
#####

Gumbel.likecondit= function(xp, sigma, T, data)
{
  mu = xp + sigma * log(-log(1 - 1/T))
  data = (data - mu)/sigma
  if (is.infinite(mu) || sigma <= 0) {
    l <- -10^6
  } else {
    l <- -(length(data) * log(sigma)) - sum(exp(-data))
    -sum(data)
  }
  l
}

Gumbel.prof = function (z, xlow, xup, T, xlab, ylab, nint)
{
  mle = z$mle
  xdat = z$data
  v = numeric(nint)
  x = seq(xlow, xup, length = nint)
  ini = sqrt(6 * var(xdat))/pi
  for (i in 1:nint) {
    opt = optim(par = ini, fn = Gumbel.likecondit,
               xp = x[i], data = xdat, T = T, method = "BFGS",
               control=list(fnscale=-1))
    v[i] = opt$value
  }
  plot(x, v, type = "l", xlab = xlab, ylab = ylab)
  ma <- z$lh
  abline(h = ma - 0.5 * qchisq(0.95, 1), lty = 1)
  invisible()
}

#####

```

```

# INTERVALO DE CONFIANÇA USANDO BOOTSTRAP PARAMÉTRICO
#####

NRBoot = function(B,T,obj, min,max) {
  mle = obj$mle
  n = length(obj$data)
  resul = numeric(B)
  for(i in 1:B) {
    sample = obj$simul(n=n,mle)
    obj$data = sample
    obj$ini = matrix(mle,1, length(mle), byrow=T)
    mleboot = Mixture.fit(obj)$mle
    resul[i] = obj$quantil( c(1 - 1 /T), mleboot,min, max)
  }
  inf = quantile(resul, 0.025)
  sup = quantile(resul, 0.975)
  res = list(intconf=c(inf,sup), boot= resul)
  res
}

#####
# ANALISE DOS DADOS - EXEMPLO MÊS DE ABRIL
#####
rabr = runs.test(abril)

#####
# ESTIMATIVAS DISTRIBUIÇÃO DE GUMBEL
#####

abrmod1 = gum.fit(abril, "BFGS")
abrmod1$lh = -abrmod1$nlh
abrmod1$quantil = Qgumbel
abrmod1$prob = Pgumbel
abrmod1$simul = rgumbel

#####
# VALORES INICIAIS PARA MISTURA
#####

Inimod1abr = matrix(numeric(80),20,4)
for(i in 1:20) { Inimod1abr[i,] = Chumod1(sample(abril,replace=T)) }

Inimod2abr = matrix(numeric(100),20,5)
for(i in 1:20) { Inimod2abr[i,] = Chumod2(sample(abril,replace=T)) }

#####
# OBJETOS
#####

abrmod2 = list(data = abril, mod = modgum1, ini = Inimod1abr, log = logmod1,
  Mstep = logM, choose = 1)
abrmod2$mleH0 = abrmod1$mle
abrmod2$lhH0 = abrmod1$lh

```

```

abrmod2$simul= Rmixgumbel1
abrmod2$sample = rgumbel

abrmod3 = list(data = abril, mod = modgum2, ini = Inimod2abr, log = logmod2,
  M1step = Mu1Sigma1, M2step = Mu2Sigma2, choose = 1)
abrmod3$mleH0 = abrmod1$mle
abrmod3$lhH0 = abrmod1$lh
abrmod3$simul = Rmixgumbel2
abrmod3$sample = rgumbel

#####
# AJUSTE DAS MISTURAS ECM
#####

resabrmod2 = Mixture.fit(abrmod2)
abrmod2$mle = resabrmod2$mle
abrmod2$lh = resabrmod2$lh

resabrmod3 = Mixture.fit(abrmod3)
abrmod3$mle = resabrmod3$mle
abrmod3$lh = resabrmod3$lh

#####
# INTERVALO DE CONFIANÇA DOS PARÂMETROS - GUMBEL
#####

par(mfrow=c(2,2))

Gumbel.prof.mu(abrmod1, 45, 55, 100)
abline(v = 47.087,lty = 2)
abline(v = 53.193,lty = 2)
abline(v = abrmod1$mle[1],lty = 2)
mtext(side = 1, expression(hat(mu)), at = c(abrmod1$mle[1])
, padj = -1.0, cex = 1.2 )
mtext(side = 1,paste("L.I"),at = c( 47.087),
padj = -2.5, cex = 0.7 )
mtext(side = 1,paste("L.S"),at = c(53.193),
padj = -2.5, cex = 0.7 )
title("abril")

Gumbel.prof.sigma(abrmod1, 7, 13.5, 100)
abline(v = 8.091,lty = 2)
abline(v = 12.818,lty = 2)
abline(v = abrmod1$mle[2],lty = 2)
mtext(side = 1, expression(hat(sigma)), at = c(abrmod1$mle[2])
, padj = -1.0, cex = 1.2 )
mtext(side = 1,paste("L.I"),at = c( 8.091),
padj = -2.5, cex = 0.7 )
mtext(side = 1,paste("L.S"),at = c(12.818),
padj = -2.5, cex = 0.7 )
title("abril")

#####

```

```

# INTERVALO DE CONFIANÇA DOS PARÂMETROS - MISTURAS
#####

abrmod2$int = ICpBoot(1000, abrmod2, label=label1)
abrmod3$int = ICpBoot(1000, abrmod3, label=label2)

#####
# AIC E BIC
#####

abrmod1$aic = AIC(abrmod1)
abrmod1$bic = BIC(abrmod1)
abrmod2$aic = AIC(abrmod2)
abrmod2$bic = BIC(abrmod2)
abrmod3$aic = AIC(abrmod3)
abrmod3$bic = BIC(abrmod3)

#####
# RAZAO DE VEROSSIMILHANÇA BOOTSTRAP
#####

abr1resul1000 = Boot(1000, abrmod2, abrmod1$mle[1], abrmod1$mle[2])
abr2resul1000 = Boot(1000, abrmod3, abrmod1$mle[1], abrmod1$mle[2])

#####
# QQ-PLOT E PP-PLOT
#####

abrmod2$quantil = Qmixgumbell1
abrmod2$prob = Pmixgumbell1
abrmod2$simul = Rmixgumbell1
abrmod2$dens = Dmixgumbell1

QQplot(abrmod2, s = "abril", min =
  c(min(abril)-3*sd(abril)), max = c(max(abril)+ 3*sd(abril) ) )

PPplot(abrmod2, s = "abril", abrmod2$mle )

#####
# PROBABILIDADES DE OCORRÊNCIA
#####

prob = seq(40, 100, by = 10)
probabr = numeric(length(prob))

for(i in 1:7) {
  probabr[i] = 1 - Pmixgumbell1(prob[i], abrmod2$mle)
}

#####
# INTERVALO DE CONFIANÇA - NIVEL DE RETORNO MISTURA
#####
nivel = c(5, 10, 50, 100)

```

```

zpabr = numeric(4)

B = 1000
nivelabr = list(nr5 = NULL, nr10 = NULL, nr50 = NULL, nr100 = NULL)

for(i in 1:length(nivel) ) {
  zpabr[i] = abrm2$quantil(c(1 - 1/nivel[i]), abrm2$mle)
  nivelabr[[i]] = NRBoot(B,nivel[i],abrm2, min(abril),
    c(max(abril)+ 3*sd(abril) ) )
}

#####
# AJUSTE DA DISTRIBUIÇÃO
#####

plot(density(abril, kernel = "rectangular"), lty=2,
main= "abril", ylab= "densidade", xlab = "",
ylim = c(0,0.05))
points(abril, rep(0.0001, length(abril)), cex=0.7,
pch = 16)
curve(abrm2$mle[4]*dgumbel(x,abrm2$mle[1], abrm2$mle[2]) +
(1 - abrm2$mle[4])*dgumbel(x,abrm2$mle[3], abrm2$mle[2]),
  20, 100, ylab="", xlab="", add=T )
mtext(text=expression(paste("velocidade", " ", "do", " ", "vento",
" ", km*.h^-1)),
side=1,padj=1.95,cex=0.9)

```