AJUSTE DA DISTRIBUIÇÃO GENERALIZADA DE PARETO ALIADA A TÉCNICA "DECLUSTERING" PARA ANÁLISE DE DADOS DE VAZÃO MÉDIA DIÁRIA DO POSTO DE ARTEMIS, PIRACICABA, SÃO PAULO, BRASIL

Renato Rodrigues SILVA[1]
Silvio Sandoval ZOCCHI[1]

▪ 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 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; mistura de distribuições; nível de retorno

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.

[1]Departamento Ciências Exatas, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Caixa Postal 9, CEP 13418-900 , Piracicaba, São Paulo, Brasil. E-mail: *rrsilva@esalq.usp.br* / *sszocchi@esalq.usp.br*

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 predizer, 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 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. Então, segundo Pickands (1975), para u suficientemente grande, a distribuição de Y dado u pode ser bem aproximada por meio de

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

que é a função de distribuição acumulada generalizada de Pareto cuja 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 (1) em que $\xi > 0$ e $\xi < 0$, respectivamente. Como limite de $G(y)$ com ξ tendendo a zero tem-se

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

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

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

Para se fazer o ajuste da distribuição generalizada de Pareto, inicialmente 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

$$E(Y - \delta|Y > \delta) = E(X - (u + \delta)|X > (u + \delta)) = \frac{\sigma + \xi\delta}{1 - \xi} \tag{2}$$

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

Logo, fazendo $w = u + \delta$ tem-se que

$$E(Y - \delta|Y > \delta) = E(X - w|X > w) = \frac{\sigma + \xi(w - u)}{1 - \xi} = \frac{\sigma - \xi u}{1 - \xi} + \frac{\xi}{1 - \xi}w,$$

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 1.

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

$$\mu(w) = \frac{1}{n_w}\sum_{h=1}^{n_w}(x_h - w) \tag{3}$$

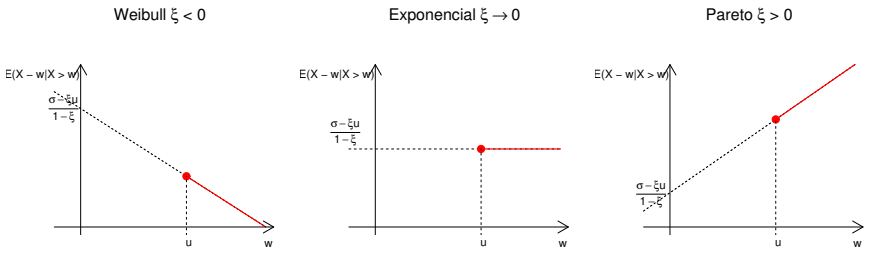

Figura 1 - Gráficos da esperança de $X - w|X > w$ em função de $w > u$, para os casos em que $Y = X - u|X > u$ segue a distribuição Weibull, $(\xi < 0)$, exponencial $(\xi \to 0)$ e Pareto, $(\xi > 0)$, respectivamente.

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 (3);

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 $s = 1, \ldots, 99$;

4. Para cada $w \in [u; x_{max}]$, calcular os percentis de ordem $\frac{\alpha}{2}$ e $1 - \frac{\alpha}{2}$ de

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

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 valores limiares candidatos, 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 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 é 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 $\left\{x_1^{(*)}, \ldots, x_k^{(*)}\right\}$, ou seja, ajusta-se o modelo

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

cuja estimativa 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 $\left\{x_1^{(*)},\ldots,x_k^{(*)}\right\}$ o conjunto de k vazões médias diárias excedentes de u observadas durante um período de N anos e $\{y_1,\ldots,y_k\}$ os excessos em relação a u. Considere, ainda, que $\{d_i\} = \{l : x_l \geq u\}$, válido para todo $i = 2,\ldots,k$ e $l = 1,\ldots,n$, 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

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

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

Para $i = 1$, $y_1 \in H_1$ e por sua vez para $i = 2,\ldots,k$ 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 $i = 2,\ldots,k$.

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

$\{z_1, \ldots, z_J\}$ dos J grupos e que esses valores sejam independentes e identicamente distríbuidos. 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

$$l(\sigma, \xi) = -J \log \sigma - \frac{1+\xi}{\xi} \sum_{j=1}^{J} \log \left(1 + \xi \frac{z_j}{\sigma}\right). \tag{4}$$

Derivando - se (4) em relação a σ e ξ e igualando essas derivadas a zero obtém-se o seguinte sistema de equações

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

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

$$l(\sigma) = -J \log(\sigma) - \frac{1}{\sigma} \sum_{j=1}^{J} z_i \tag{5}$$

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 (5) 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

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

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

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

em que σ_0 é um valor fixo de σ, ξ_0 é um valor fixo de ξ e $\chi^2_{1,(1-\alpha)}$ 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

$$D(\sigma) = 2\left[l(\hat{\sigma}) - l(\sigma)\right] = 2\left[J\log\left(\frac{\sigma}{\hat{\sigma}}\right) + \sum_{j=1}^{J}\left(\frac{1}{\sigma} - \frac{1}{\hat{\sigma}}\right)\right] \tag{6}$$

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

$$I.C_{(100\times(1-\alpha)\%)}(\sigma) = \left\{\sigma : D(\sigma) \leq \chi^2_{1,(1-\alpha)}\right\}. \tag{7}$$

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

$$Pr(X > \tau | X > u) = \left[1 + \xi\left(\frac{\tau - u}{\sigma}\right)\right]^{-\frac{1}{\xi}} = \frac{1}{t} = \frac{1}{T365,25} \tag{8}$$

o que implica que

$$Pr(X > \tau) = p\left[1 + \xi\left(\frac{\tau - u}{\sigma}\right)\right]^{-\frac{1}{\xi}} = \frac{1}{T365,25}. \quad (9)$$

Dessa forma, define-se o nível de retorno isolando τ em (9), ou seja,

$$\tau = u + \frac{\sigma}{\xi}\left[(T365,25p)^{\xi} - 1\right] \quad (10)$$

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

$$\hat{\tau} = u + \frac{\hat{\sigma}}{\hat{\xi}}\left[(T365,25\hat{p})^{\hat{\xi}} - 1\right] \quad (11)$$

que é válida apenas quando considera-se $X_1, \ldots, X_n$ 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 (10) da seguinte forma

$$\tau = u + \frac{\sigma}{\xi}\left[(T365,25\Theta p)^{\xi} - 1\right] \quad (12)$$

em que $\Theta \in (0,1)$ é o parâmetro "extremal index" que pode ser interpretado como o tamanho médio dos grupos para $n \to \infty$, 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,

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

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 (12), 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

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

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

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

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, $\left\{z_{(1)}, \ldots, z_{(J)}\right\}$, o gráfico probabilidade - probabilidade é definido pelos pontos de coordenadas $\left\{\frac{j}{J+1}, \hat{G}(z_{(j)})\right\}$, sendo

$$\hat{G}(z_{(j)}) = 1 - \left[1 + \hat{\xi}\left(\frac{y}{\hat{\sigma}}\right)\right]^{-\frac{1}{\hat{\xi}}}. \tag{13}$$

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 $\left\{z_1^{*(1)}, \ldots, z_J^{*(1)}\right\}, \ldots, \left\{z_1^{*(B)}, \ldots, z_J^{*(B)}\right\}$,

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 $\left(1 - \frac{\alpha}{2}\right)$ de $\left(z_j^{(1)}, \ldots, z_j^{(B)}\right)$ 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

$$\hat{G}^{-1}\left(\frac{j}{J+1}\right) = u + \frac{\hat{\sigma}}{\hat{\xi}}\left[z^{-\hat{\xi}} - 1\right] \tag{14}$$

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.

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. Analisando essa figura, observa-se que para $w > 730\ m^3.s^{-1}$, o gráfico da média dos excessos é aproximadamente linear e descrescente com um aspecto "serreado" evidente. No entanto, segundo Smith (2004), valores de w cuja a média do excessos tem uma 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 $u = 400\ m^3.s^{-1}$. 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 $u = 400\ m^3.s^{-1}$ e $u = 730\ m^3.s^{-1}$, apresentados na Figura 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 $u = 400\ m^3.s^{-1}$ confirmando a escolha anterior. Além disso, observa-se que o gráfico da estimativa da esperança da média do excessos é uma reta descrecente, 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 de $400\ m^3.s^{-1}$. Para esta finalidade, primeiramente foi feita uma inspeção visual do gráfico do logaritmo da vazão média diária excedente de $400\ m^3.s^{-1}$ conforme mostra a Figura 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

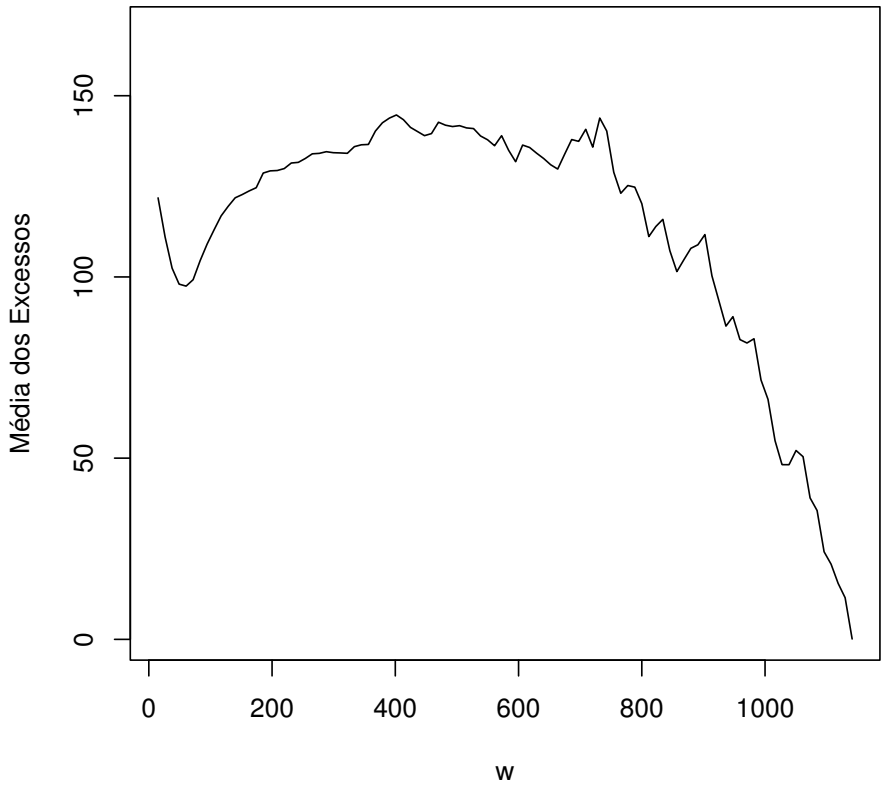

Figura 2 - Médias dos excessos de vazão média diária ($m^3.s^{-1}$) em função de w.

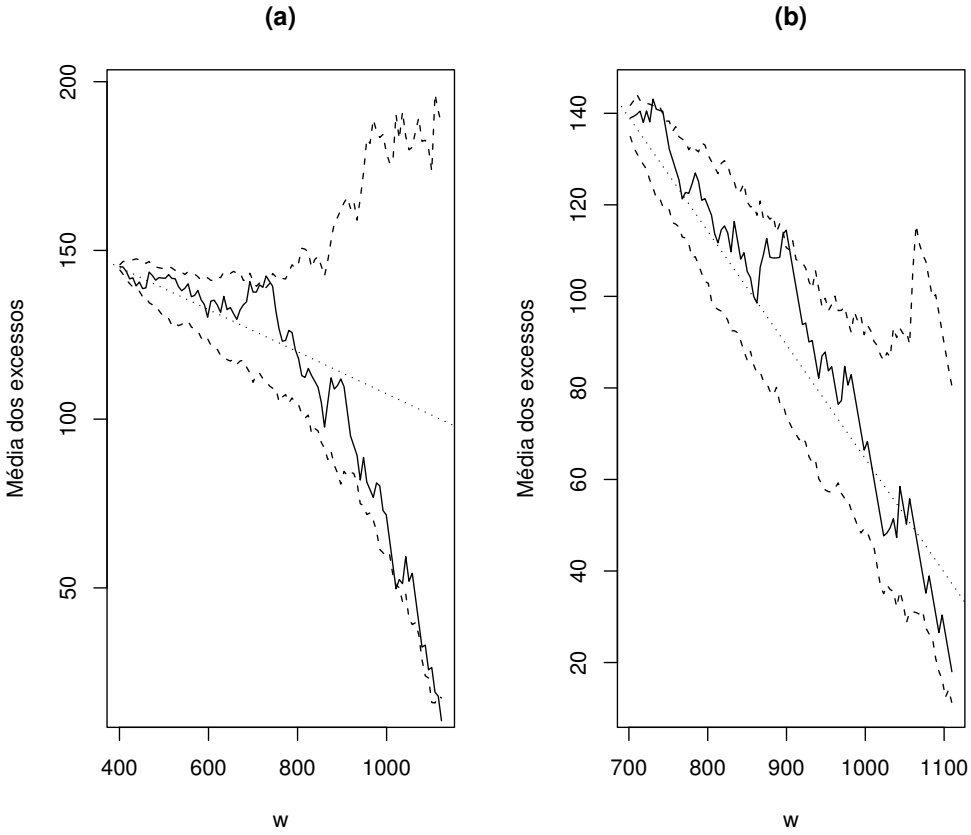

Figura 3 - Médias dos excessos para dois valores limiares candidatos: (a) u = 400 $m^3.s^{-1}$ e (b) u = 730 $m^3.s^{-1}$ para dados de vazões médias diárias com envelopes simulados de 95 $\%$ de confiança utilizando-se 99 simulações de Monte Carlo.

aproximadamente igual a 1, o que implica que não há evidência estatística para rejeitar a hipótese nula de estacionariedade ao nível de significância de 5 $\%$. Portanto, pode-se concluir que, embora Moraes et. al (1997) e Groppo et. al (2005) tenham relatado uma tendência linear descrescente 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 $400\ m^3.s^{-1}$.

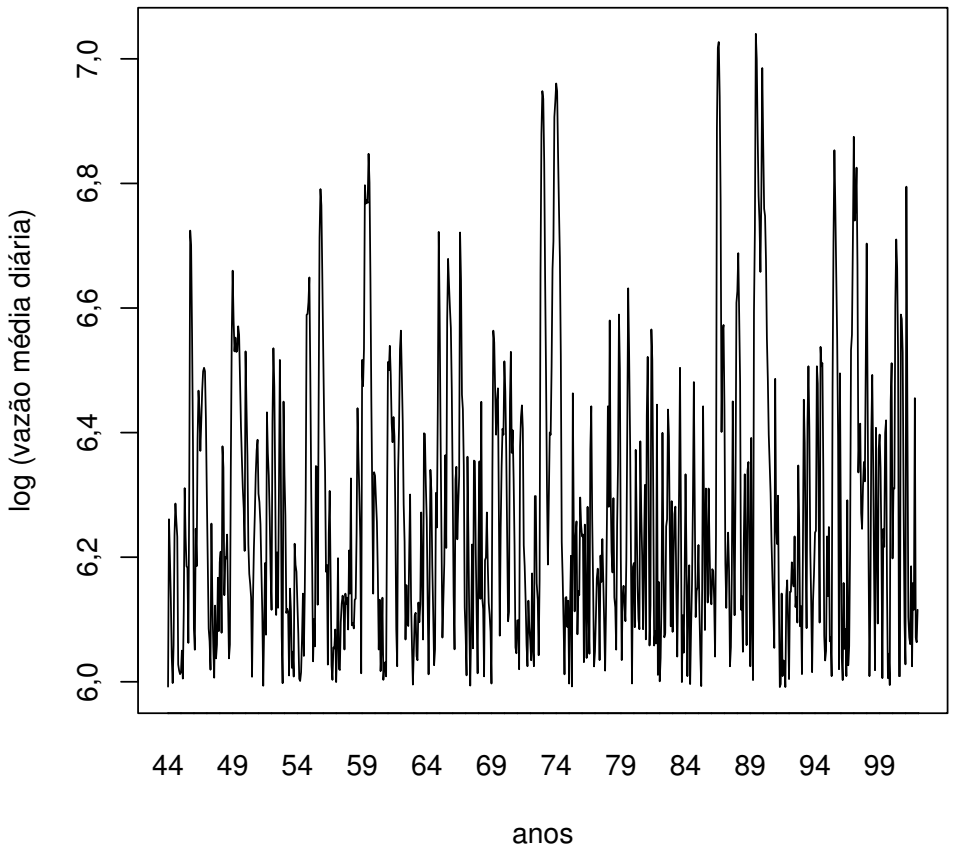

Figura 4 - Série temporal do logaritmo da vazão média diária excedentes de $400\ m^3.s^{-1}$.

A Figura 5, por sua vez, apresenta o gráfico de autocorrelação dos excedentes acima de $400\ m^3.s^{-1}$, e os gráficos de autocorrelação dos máximos dos grupos de excedentes

acima 400 $m^3.s^{-1}$ para $r = 1$, $r = 2$ e $r = 4$, respectivamente.

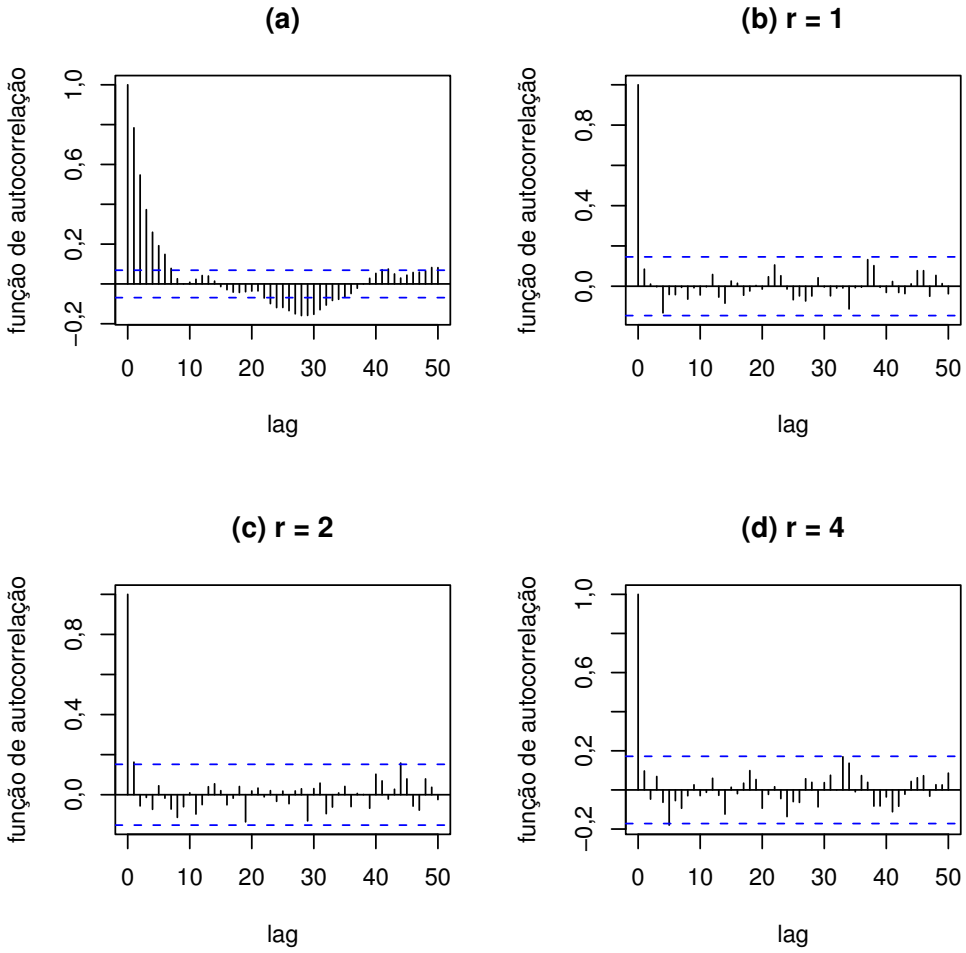

Figura 5 - Autocorrelação dos excedentes acima do limiar $u = 400\ m^3.s^{-1}$ (a) e gráficos de autocorrelação dos máximos dos grupos de excedentes acima de 400 $m^3.s^{-1}$ dos dados de vazão média diária para (b) r = 1, (c) r = 2 e (d) r = 4.

Analisando o gráfico de autocorrelção 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 aproximadamentes 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 $m^3.s^{-1}$ 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 (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 6.

A Tabela 1 apresenta as 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. 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 $m^3.s^{-1}$ uma vez a cada 60 anos, valor próximo da máxima vazão média diária registrada em 59 anos, 1141,5 $m^3.s^{-1}$, o que significa que essa estimativa é consistente com os dados observados. Por sua vez, analisando a Tabela 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 conseqüência 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.

Por fim, a Figura 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.

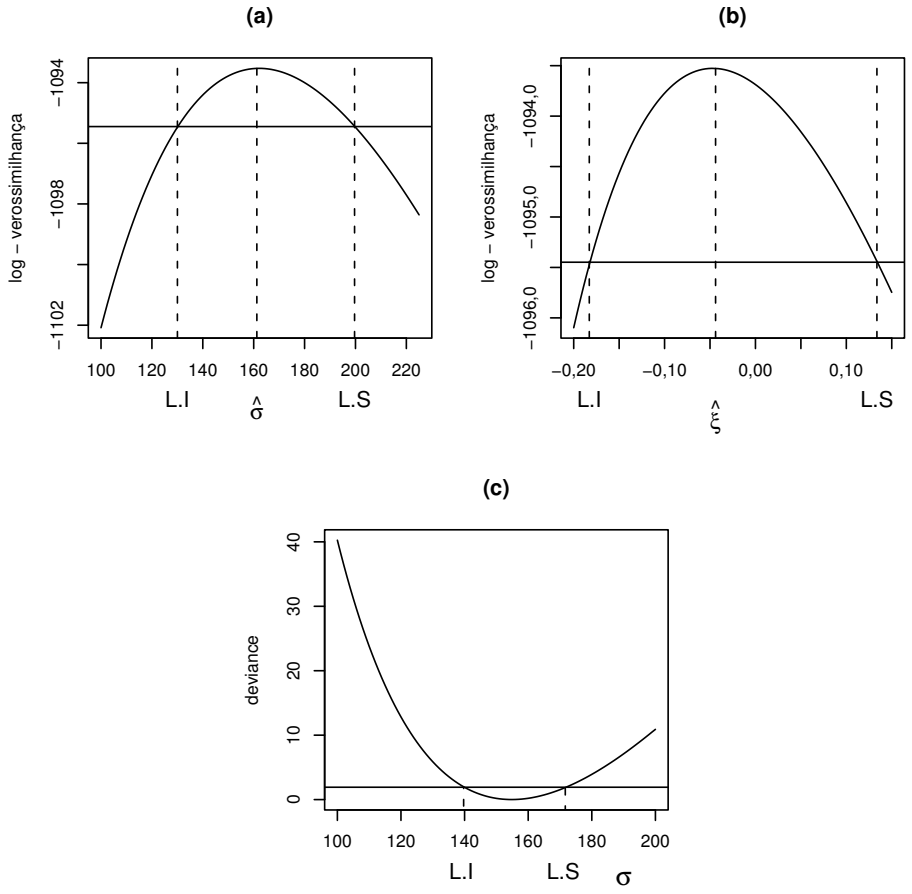

Figura 6 - Perfis de verossimilhança e função deviance para a obtenção dos intervalos de 95% de confiança (LI; LS) dos parâmetros σ, (a), e ξ, (b), da distribuição GP e do parâmetro σ, (c) da distribuição exponencial.

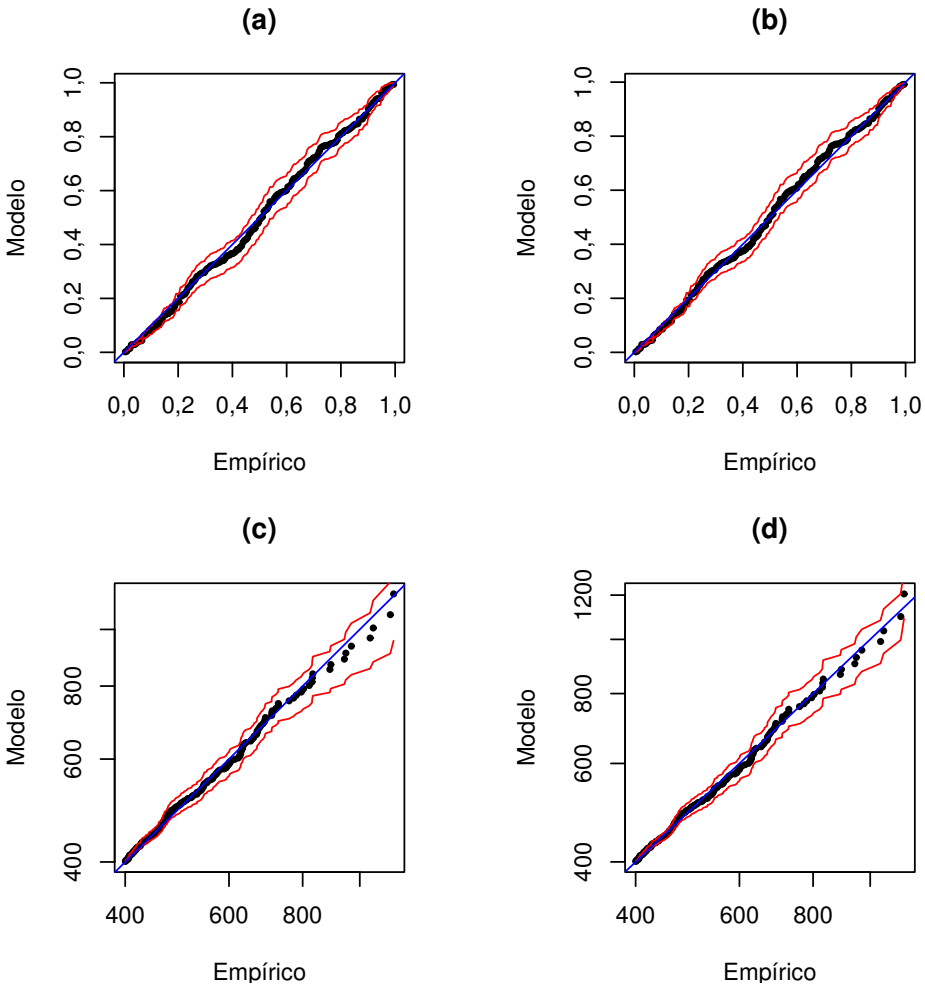

Figura 7 - Gráfico probabilidade - probabilidade para o ajuste da distribuição generalizada de Pareto (a) e distribuição exponencial (b) com envelope simulado de 95% de confiança, gráfico quantil-quantil para o ajuste da distribuição generalizada de Pareto (c) e distribuição exponnecial (d) com envelope simulado de 95% de confiança em escala logarítmica.

Tabela 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 - Estimativas dos níveis de retorno e seus respectivos intervalos de 95% de confiança utilizando a distribuição Exponecial 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

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 $m^3.s^{-1}$ uma vez a cada 60 anos para o município de Piracicaba, SP.

Agradecimentos

Este trabalho é parte da dissertação de Mestrado em Estatistica e Experimentacao Agronomica do primeiro autor no Departamento de Ciencias Exatas da ESALQ/USP, Piracicaba.

SILVA, R. R.; ZOCCHI, S. S. The fitting of the generalized distribution Pareto join to declustering methods to analyze to dataset of mean daily flow of Artemis station, Piracicaba, São Paulo, Brazil. *Rev. Mat. Estat.,* São Paulo, v.26, n.4, p.45-65, 2008.

- *ABSTRACT: One of the most used methodologies in the context of the extreme 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, to fit the generalized Pareto distribution to a dataset, it must consider the data is a sequence independent and having a common distribution that it is usually an unrealistic assumption. An way of to solve this problem is to use methods of declustering propose by Leadbetter et al. (1989), that in general identify groups of ocorrence of extreme flow for after fitting of the generalized distribution Pareto to maximum of the groups. In this work, was mades the fit generalized distribution Pareto and expoencial distribution, particular case of GP, join to declustering methods to dataset mean daily flow of Artemis station, Piracicaba, SP, Brazil and after was compared the estimates the return levels of 5, 10, 50 and 100 years. Concludes the intervals estimates of return levels of 50 and 100 years obtained through the fitting the exponencial distribution are more precision than obtained through the fitting the generalized Pareto distribution.*

- *KEYWORDS: Generalized distribution Pareto; generalized extremes values distribution; return levels.*

Referências

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. *Rev. Mat. Estat.*, São Paulo, v.22, n.1, p.95-111, 2004.

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

DAVISON, A. C.; SMITH, R. L. Models for exceeedances over high thresholds. *J. R. Stat. Soci., Stat. Methodol., Ser.B*, London, v.52, 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. *J.Wind Eng. Ind. Aerodynam.*, Amsterdan, v.83, p.1-10, 1999.

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

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. *Rev. Bras. Recur. Hídr.*, Porto Alegre, 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. *Rev. Bras. Agrometeorol.*, Santa Maria, v.9, n.1, p.125-135, 2001.

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

PICKANDS, J. Statistical inference using extreme order statistics. *Ann. Stat.*, Hayward, v.3, 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, Rio de Janeiro, 2006, 114f. 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, Vienna, Austria, 2006. Disponível em ¡http://www.R-project.org¿.

SILVA, R. R; ZOCCHI, S. S. Densidades não paramétricas no estudo da velocidade máxima do vento em Piracicaba, SP. In: REUNIÃO DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA, 50, e SIMPÓSIO DE ESTATÍSTICA APLICADA À AGRICULTURA, 11., 2005. Londrina: RBRAS e UEL, 2005, 1 CD-ROM.

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

Recebido em 30.05.2008.

Aprovado após revisão em 03.12.2008.