A DISTRIBUIÇÃO GENERALIZADA DE PARETO-POISSON NO ESTUDO DA PRECIPITAÇÃO PLUVIAL TOTAL DIÁRIA MÁXIMA EM PIRACICABA, SP

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

▪ RESUMO: Utilizou-se o modelo da distribuição generalizada de Pareto-Poisson para analisar a série de dados de precipitações diárias totais excedentes em relação a um valor limiar, registrados durante o período de 1917 a 2004 em Piracicaba, SP. Escolheu-se o valor limiar tal que os excessos seguiram distribuição generalizada de Pareto e que o tamanho médio dos grupos de excedentes não se diferiu estatisticamente de 1. Uma vez escolhido o valor limiar, os parâmetros do modelo foram estimados por meio do método da máxima verossimilhança e baseando-se nessas estimativas foram obtidos os níveis de retorno para os períodos de 10, 25, 50 e 100 anos e respectivos intervalos de 95% de confiança por meio do método do perfil da máxima verossimilhança. Além disso, foram obtidos os períodos de retorno para precipitações diárias totais de 40, 60, 80 e 100 mm. A qualidade do ajuste aos dados foi avaliada por meio dos gráficos probabilidade-probabilidade e quantil-quantil. Analisando os resultados observou-se que em um período de 50 anos, espera-se em média, que precipitações diárias totais acima de 85 mm ocorram apenas nos meses de novembro a março e, então, concluiu-se que para esses meses esperam-se as maiores precipitações diárias totais.

▪ PALAVRAS-CHAVE: Teoria dos valores extremos; período de retorno; nível de retorno.

1 Introdução

O excesso de precipitação pode ocasionar inúmeros prejuízos que podem ser evitados ou amenizados construindo-se estruturas hidráulicas de controle de águas naturais, para o dimensionamento e construção dessas estruturas é levada em

[1]Departamento Ciências Exatas, Universidade de São Paulo, campus de Piracicaba – ESALQ/USP, Caixa Postal 9, CEP 13418-900 , Piracicaba, SP, Brasil. E-mail: *rrsilva@esalq.usp.br / sszocchi@esalq.usp.br*

consideração a probabilidade de ocorrência de precipitações pluviais máximas. Estas são geralmente determinadas utilizando-se a teoria dos valores extremos e, neste contexto, há várias metodologias propostas.

Um método que pode ser utilizado é o ajuste da distribuição generalizada de Pareto-Poisson a uma série histórica de dados, constituída somente pelos eventos que excedem um determinado valor limiar, denominada série de duração parcial por Cunnane (1973). No entanto, a distribuição generalizada de Pareto-Poisson é apropriada apenas para observações independentes e identicamente distribuídas, e segundo Coles (2001) precipitações extremas normalmente ocorrem em grupos. Este pesquisador propõe a estratégia de estabelecer uma regra para definir o tamanho dos grupos de excedentes, identificar os máximos de cada grupo e aplicar o modelo distribuição generalizada Pareto-Poisson.

Para definir o tamanho dos grupos, Davison e Smith (1990) propuseram a utilização de um modelo duplamente estocástico de Poisson. Leadbetter et al. (1989) propuseram uma regra empírica que consiste em considerar dois excedentes sucessivos pertencentes ao mesmo grupo, caso o número de observações abaixo do valor limiar entre eles for menor que um valor pré-fixado arbitrariamente, denominado r e, caso contrário, considera-os como pertencentes a grupos distintos.

Como uma alternativa aos métodos mencionados, propõe-se selecionar um valor limiar tal que o tamanho médio dos grupos não difere de 1, por meio do teste de hipótese para o parâmetro "extremal index", de forma que os excessos sigam a distribuição generalizada de Pareto.

2 Material

Os dados de precipitação pluvial diária total, em mm, relativos ao período de 1917 a 2004, foram obtidos 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, disponíveis em *http* : *//www.lce.esalq.usp.br/postocon.html*. Para cada ano, a série de dados foi subdividida em 12 meses, obtendo-se, assim, uma série de dados para cada mês.

A precipitação total de um dia refere-se ao total de precipitação ocorrida das 7 horas da manhã do dia anterior até as 7 horas da manhã do dia seguinte.

3 Métodos

Seja X a variável aleatória precipitação pluvial diária total e $Y = X - u$, a variável aleatória excesso em relação ao valor limiar u, condicionada a $X > u$. Segundo Smith(2003) um modelo para a distribuição dos excessos de Y pode ser a distribuição generalizada de Pareto, DGP, com função de distribuição acumulada

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

função densidade de probabilidade

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

com esperança, para $\xi < 1$, dada por

$$E(Y) = \frac{\bar{\sigma}}{1-\xi},$$

e variância, para $\xi < \frac{1}{2}$, por

$$Var(Y) = \frac{\bar{\sigma}^2}{(1-\xi)^2(1-2\xi)} = \frac{1}{1-2\xi}\ [E(Y)]^2,$$

sendo ξ o parâmetro de forma e $\bar{\sigma}$ o parâmetro de escala, tais que $\xi > 0$, $\bar{\sigma} > 0$ e $1 + \xi\frac{y}{\bar{\sigma}} > 0$, ou seja, para $\xi < 0$, $0 < y < -\frac{\bar{\sigma}}{\xi}$, em que $-\frac{\bar{\sigma}}{\xi}$ é o limite superior para os excessos.

Considere agora, K a variável aleatória do número de ocorrência de excedentes acima do valor limiar u, num período de T anos, com distribuição de Poisson com média λT cuja função de probabilidades é dada por

$$Pr(K = k) = \frac{(\lambda T)^k e^{-\lambda T}}{k!}$$

para $k = 0, 1, 2, \ldots$, sendo λ o número esperado de excedentes por ano. Então a variável aleatória número de ocorrências de excedentes em relação a um nível $x > u$, denotada por K_x, segue também uma distribuição de Poisson, porém com média dada por

$$\lambda_x = \lambda Pr(Y > x - u) = \lambda\left[1 - Pr(Y \leq x - u)\right] = \lambda\left[1 - G(x - u)\right]$$

e assim,

$$Pr(K_y = 0) = e^{-\lambda_y} = \exp\left\{ -\lambda\Big[1 + \xi\left(\frac{x-u}{\bar{\sigma}}\right)\Big]^{-\frac{1}{\xi}}\right\}, \tag{2}$$

que é uma das formas de se apresentar o modelo da distribuição generalizada de Pareto-Poisson, útil para demonstrar sua relação com a distribuição generalizada dos valores extremos apresentada por Jekinson (1955) que pode ser obtida subtituindo-se

$$\bar{\sigma} = \sigma + \xi(u - \mu),$$

e

$$\lambda = \left[1 + \xi\Big(\frac{u-\mu}{\sigma}\Big)\right]^{-\frac{1}{\xi}},$$

em (2).

Uma vez definido o modelo, o próximo passo é a escolha do valor limiar, que depende do valor do parâmetro "extremal index". Para definição deste parâmetro considere $X_1^*, \ldots, X_n^*$ uma seqüência de variáveis aleatórias obtidas a partir de um processo estacionário com função distribuição acumulada marginal F e M_n^* o máximo dessa série, ou seja,$M_n^* = \max\{X_1^*, \ldots, X_n^*\}$.

O parâmetro "extremal index" θ, é definido como a solução da equação

$$Pr(M_n^* \leq x^*) = F(x^*)^{n\theta}$$

sendo $\theta \in [0,1]$ que, segundo Leadbetter, Lindgren e Rootzén (1983), pode ser interpretado como o tamanho médio dos grupos de excedentes.

Para estimar θ, adotou-se neste trabalho o estimador intervalar proposto por Ferro e Segers (2003), dado por:

$$\widehat{\theta} = \begin{cases} \text{mín}\left\{1, \frac{2\left(\sum_{i=1}^{k-1} T_i\right)^2}{(k-1)\sum_{i=1}^{k-1} T_i^2}\right\}, & \text{se máx}\{T_i; 1 \leq i \leq k-1\} \leq 2 \\ \text{mín}\left\{1, \frac{2\left[\sum_{i=1}^{k-1}(T_i-1)\right]^2}{(k-1)\sum_{i=1}^{k-1}(T_i-1)(T_i-2)}\right\}, & \text{se máx}\{T_i; 1 \leq i \leq k-1\} \geq 2 \end{cases} \quad (3)$$

sendo T_i o tempo entre as ocorrências dos excedentes e k o número de excedentes observados.

Como método para a escolha do limiar u, propõe-se neste trabalho, que se escolha o menor valor do limiar tal que θ não difere estatisticamente de 1, e tal que o modelo DGP-Poisson seja adequado.

Para isso, dados q valores limiares candidatos $u_1, \ldots, u_q$, testa-se, para cada um deles a hipótese $H_0 : \theta = 1$ *versus* $H_a : \theta < 1$, cujo procedimento proposto para um valor limiar u pré-estabelecido tem os seguintes passos:

1. Obter a estimativa $\widehat{\theta}$ do θ para a série original de dados por meio de (3);
2. Retirar B permutações dos n elementos da série original de dados;
3. Para cada permutação, estimar θ por meio de (3) obtendo o conjunto de B estimativas de θ, $\left\{\widehat{\theta}^{*1}, \ldots, \widehat{\theta}^{*B}\right\}$;
4. Calcular o valor-p do teste, por meio da proporção de estimativas do "extremal index" θ, inferiores a $\widehat{\theta}$, ou seja,

$$\text{valor-p} = \frac{1}{B}\sum_{b=1}^{B} I_{\left\{\widehat{\theta}^{*b} < \widehat{\theta}\right\}}(\widehat{\theta}^{*b}),$$

sendo $I(.)$ uma função indicadora que assume o valor 1 se $\widehat{\theta}^{*b} < \widehat{\theta}$ e 0, caso contrário.

Uma vez escolhidos os valores limiares cujo "extremal index" não diferem de 1, propõe-se escolher o menor deles que leve a um bom ajuste do modelo DGP-Poisson. A qualidade do ajuste pode ser avaliada por meio de análise do gráfico das médias

dos excessos em relação a um valor $w > u$ em função de w, que tem interpretação baseada no fato de

$$E(Y - w|y > w) = \frac{\bar{\sigma} + \xi(w-u)}{1-\xi} = \frac{\bar{\sigma} - \xi u}{1-\xi} + \frac{\xi}{1-\xi}w, \tag{4}$$

válida para $\xi < 1$ (Smith, 2003). Assim, se os dados tiverem uma distribuição generalizada de Pareto com parâmetros $\bar{\sigma}$ e ξ, o gráfico deverá seguir uma linha reta com intercepto $\frac{\bar{\sigma}-\xi u}{1-\xi}$ e coeficiente angular $\frac{\xi}{1-\xi}$, o que sugere que se podem estimar $\bar{\sigma}$ e ξ, além de testar a qualidade do ajuste dos dados, baseados na linearidade do gráfico.

De modo a tornar sua interpretação menos subjetiva, Smith (2003) propõe a construção de uma faixa de confiança ou envelope simulado utilizando o procedimento descrito a seguir. Considere que para um certo limiar u, a verdadeira distribuição dos excessos sobre u é GPD com parâmetros $\bar{\sigma}$ e ξ e que $\widehat{\bar{\sigma}}$ e $\widehat{\xi}$, são as estimativas de máxima verossimilhança.

Seja $\widehat{\mu}(w)$ a média amostral dos excessos acima de w, sendo $w > u$. Então, sob a hipótese de que os excessos têm distribuição generalizada de Pareto, a diferença estatística entre a média empírica e a média teórica dos excessos para cada w, dada por

$$T(w) = \widehat{\mu}(w) - \frac{\widehat{\bar{\sigma}} + \widehat{\xi}(w-u)}{1-\widehat{\xi}}$$

tem esperança nula. Assim, dado um valor limiar u e uma série de valores de w tais que $w > u$, o envelope pode ser construído usando-se o procedimento:

1. obter as estimativas de máxima verossimilhança $\widehat{\bar{\sigma}}$ e $\widehat{\xi}$ dos parâmetros $\bar{\sigma}$ e ξ da distribuição generalizada de Pareto para os dados da amostra original e para cada w, calcular $\widehat{\mu}(w)$;

2. gerar 99 amostras aleatórias da distribuição generalizada de Pareto com parâmetros $\bar{\sigma} = \widehat{\bar{\sigma}}$ e $\xi = \widehat{\xi}$, de tamanho igual ao da amostra original;

3. para cada amostra, obter as estimativas de máxima verossimilhança e a média amostral dos excessos de w para cada w, ou seja, $\widehat{\bar{\sigma}}^{(j)}$, $\widehat{\xi}^{(j)}$ e $\widehat{\mu}^{(j)}(w)$, para $j = 1, \ldots, 99$;

4. para cada w, calcular os percentis de ordem $\frac{\alpha}{2}$ e $1 - \frac{\alpha}{2}$ de

$$\widehat{\mu}(w) - \frac{\widehat{\bar{\sigma}}^{(j)} + \widehat{\xi}^{(j)}(w-u)}{1-\widehat{\xi}^{(j)}} + \frac{\widehat{\bar{\sigma}} + \widehat{\xi}(w-u)}{1-\widehat{\xi}}$$

que formam, para cada w, os limites inferiores e superiores do envelope, respectivamente.

Agora será apresentado como estimar os parâmetros do modelo DGP-Poisson. Segundo Smith (2001), a função de verossimilhança pode ser escrita em duas partes.

A primeira parte correspondendo ao processo Poisson da ocorrência dos excedentes e a outra referente à distribuição dos excessos com DGP.

Seja $\{x_1, \ldots, x_k\}$ um conjunto de k dados observados em um período de T anos, tais que $x_i > u$ para todo $i = 1, \ldots, k$ e $\{y_1, \ldots, y_k\}$ os excessos em relação a u. Então a função de verossimilhança para o modelo da DGP-Poisson é dada por

$$L(\lambda, \bar{\sigma}, \xi) = L(\lambda) \times L(\bar{\sigma}, \xi) = \frac{(\lambda T)^k \exp\{-\lambda T\}}{k!} \times \prod_{i=1}^{k} \frac{1}{\bar{\sigma}} \left(1 + \xi \frac{y_i}{\bar{\sigma}}\right)^{-\left(\frac{1+\xi}{\xi}\right)} \tag{5}$$

cujo logaritmo é

$$\begin{aligned} l(\lambda, \bar{\sigma}, \xi) = l(\lambda) + l(\bar{\sigma}, \xi) \quad &= \quad k \log \lambda + k \log T - \lambda T - \log k! - \\ &\quad k \log \bar{\sigma} - \frac{1+\xi}{\xi} \sum_{i=1}^{k} \log \left(1 + \xi \frac{y_i}{\bar{\sigma}}\right), \end{aligned} \tag{6}$$

válido para $\left(1 + \xi \frac{y_i}{\bar{\sigma}}\right) > 0$.

Derivando-se (6) em relação a λ tem-se que

$$\frac{\partial l(\lambda, \bar{\sigma}, \xi)}{\partial \lambda} = \frac{\partial l(\lambda)}{\lambda} = \frac{k}{\lambda} - T \tag{7}$$

e igualando-se a equação (7) a zero, obtém-se a equação $\frac{k}{\lambda} - T = 0$, cuja solução para λ é

$$\widehat{\lambda} = \frac{k}{T}, \tag{8}$$

estimador de máxima verossimilhança para λ.

Considerando-se que

$$l(\bar{\sigma}, \xi) = -k \log \bar{\sigma} - \frac{1+\xi}{\xi} \sum_{i=1}^{k} \log \left(1 + \xi \frac{y_i}{\bar{\sigma}}\right) \tag{9}$$

e assumindo que (9) é diferenciável, então o estimador de máxima verossimilhança é dado pela solução do sistema de equações obtido igualando-se a zero as derivadas de primeira ordem de (9) em relação a $\bar{\sigma}$ e ξ dadas por,

$$\frac{\partial l(y_i, \bar{\sigma}, \xi)}{\partial \bar{\sigma}} = -\frac{k}{\bar{\sigma}} + \left(\frac{1+\xi}{\xi}\right) \sum_{i=1}^{k} \frac{1}{\left(1 + \xi \frac{y_i}{\bar{\sigma}}\right)} \frac{\xi y_i}{\bar{\sigma}^2}$$

e

$$\frac{\partial l(y_i, \bar{\sigma}, \xi)}{\partial \xi} = \frac{1}{\xi^2} \sum_{i=1}^{k} \log \left(1 + \xi \frac{y_i}{\bar{\sigma}}\right) - \left(\frac{1+\xi}{\xi}\right) \sum_{i=1}^{k} \frac{1}{\left(1 + \xi \frac{y_i}{\bar{\sigma}}\right)} \frac{y_i}{\bar{\sigma}},$$

ou seja, os estimadores de máxima verossimilhança $\widehat{\bar{\sigma}}$ e $\widehat{\xi}$ de $\bar{\sigma}$ e ξ são as soluções do sistema de equações

$$\begin{cases} \frac{\partial l(y_i, \bar{\sigma}, \xi)}{\partial \bar{\sigma}} & = 0 \\ \frac{\partial l(y_i, \bar{\sigma}, \xi)}{\partial \xi} & = 0 \end{cases} \tag{10}$$

A matriz Hessiana calculada para $\bar{\sigma} = \widehat{\bar{\sigma}}$ e $\xi = \widehat{\xi}$, por sua vez, chamada de matriz de informação observada I, é dada por

$$\mathbf{I} = \begin{bmatrix} -\frac{\partial l^2(\bar{\sigma},\xi)}{\partial \bar{\sigma}} & -\frac{\partial l(\bar{\sigma},\xi)}{\partial \bar{\sigma}\partial \xi} \\ -\frac{\partial l(\bar{\sigma},\xi)}{\partial \bar{\sigma}\partial \xi} & -\frac{\partial l^2(\bar{\sigma},\xi)}{\partial \xi} \end{bmatrix}$$

sendo

$$-\frac{\partial l^2(\bar{\sigma},\xi)}{\partial \bar{\sigma}}\bigg|_{\bar{\sigma}=\widehat{\bar{\sigma}},\xi=\widehat{\xi}} = -\frac{k}{\widehat{\bar{\sigma}}^2} - \frac{1+\widehat{\xi}}{\widehat{\xi}}\sum_{i=1}^{k}\frac{1}{\left(1+\widehat{\xi}\frac{y_i}{\widehat{\bar{\sigma}}}\right)^2}\frac{\widehat{\xi}^2 y_1^2}{\widehat{\bar{\sigma}}^4} - \frac{2\widehat{\xi}y_i}{\widehat{\bar{\sigma}}^3}\frac{1}{\left(1+\widehat{\xi}\frac{y_i}{\widehat{\bar{\sigma}}}\right)},$$

$$-\frac{\partial l(\bar{\sigma},\xi)}{\partial \bar{\sigma}\partial \xi}\bigg|_{\bar{\sigma}=\widehat{\bar{\sigma}},\xi=\widehat{\xi}} = \frac{2}{\widehat{\xi}^3}\sum_{i=1}^{k}\log\left(1+\widehat{\xi}\frac{y_i}{\widehat{\bar{\sigma}}}\right) - \frac{2}{\widehat{\xi}^2}\sum_{i=1}^{k}\frac{1}{\left(1+\widehat{\xi}\frac{y_i}{\widehat{\bar{\sigma}}}\right)} - \frac{1+\widehat{\xi}}{\widehat{\xi}}\sum_{i=1}^{k}\frac{1}{\left(1+\widehat{\xi}\frac{y_i}{\widehat{\bar{\sigma}}}\right)}\frac{y_1^2}{\widehat{\bar{\sigma}}}$$

e

$$-\frac{\partial l^2(\bar{\sigma},\xi)}{\partial \xi}\bigg|_{\bar{\sigma}=\widehat{\bar{\sigma}},\xi=\widehat{\xi}} = \frac{1}{\widehat{\xi}^2}\sum_{i=1}^{k}\frac{1}{\left(1+\widehat{\xi}\frac{y_i}{\widehat{\bar{\sigma}}}\right)}\frac{\widehat{\xi}y_i}{\widehat{\bar{\sigma}}^2} - \left(\frac{1+\widehat{\xi}}{\widehat{\xi}}\right)\sum_{i=1}^{k}\left[\frac{-1}{\left(1+\widehat{\xi}\frac{y_i}{\widehat{\bar{\sigma}}}\right)^2}\frac{\widehat{\xi}y_i^2}{\widehat{\bar{\sigma}}^3} + \frac{y_i}{\widehat{\bar{\sigma}}^2}\frac{1}{\left(1+\widehat{\xi}\frac{y_i}{\widehat{\bar{\sigma}}}\right)}\right].$$

A importância de I está no fato de que para uma amostra de tamanho n suficientemente grande, $\widehat{\theta} = (\widehat{\bar{\sigma}}, \widehat{\xi})$ segue distribuição normal multivariada com média $\widehat{\theta}$ e matriz de covariância $\mathbf{\Psi}$ dada pelo inverso da matriz de informação observada $\mathbf{I}^{-1}$, ou seja,

$$\mathbf{\Psi} = \mathbf{I}^{-1} = \begin{bmatrix} Var(\widehat{\bar{\sigma}}) & Cov(\widehat{\bar{\sigma}}, \widehat{\xi}) \\ Cov(\widehat{\bar{\sigma}}, \widehat{\xi}) & Var(\widehat{\xi}) \end{bmatrix},$$

utilizada neste trabalho para obtenção dos erros padrões das estimativas dos parâmetros do modelo.

Importante frisar que uma das aplicações mais importantes dos modelos de valores extremos não são as estimativas dos parâmetros do modelo, mas sim as estimativas dos níveis de retorno para um certo período de tempo pré-especificado.

Para definir o que venha ser nível de retorno, suponha que o excesso sobre u de uma variável aleatória X tenha distribuição generalizada de Pareto com parâmetros $\bar{\sigma}$ e ξ. Então para todo $x > u$ têm-se que

$$Pr(X > x | X > u) = \left[1 + \xi\left(\frac{x-u}{\bar{\sigma}}\right)\right]^{-\frac{1}{\xi}} \tag{11}$$

Segundo Rasmussem e Rosbjerg (1991), nível de retorno x_N é o valor x que pode ser excedido, em média, uma vez a cada N anos, ou seja,

$$Pr(X > x_N | X > u) = \frac{1}{\lambda N}. \tag{12}$$

Assim, fazendo-se $x = x_N$ em (11) e igualando a (12), obtém-se a equação

$$\left[1 + \xi\Big(\frac{x_N - u}{\bar{\sigma}}\Big)\right]^{-\frac{1}{\xi}} = \frac{1}{\lambda N}$$

com solução dada por

$$x_N = u + \frac{\bar{\sigma}}{\xi}\left[(N\lambda)^{\xi} - 1\right], \tag{13}$$

cuja estimativa $\widehat{x}_N$ é obtida substituindo-se $\bar{\sigma}$, ξ e λ por suas estimativas de máxima verossimilhança $\widehat{\bar{\sigma}}, \widehat{\xi}$ e $\widehat{\lambda}$, respectivamente.

A construção do intervalo de confiança para o nível de retorno x_N, por outro lado, pode ser obtida por meio do método do perfil de verossimilhança considerando-se a reparametrização:

$$\bar{\sigma} = \begin{cases} \dfrac{(x_N - u)\xi}{\lambda^{\xi} - 1}, & \text{se } \xi \neq 0; \\ \dfrac{x_N - u}{\log(\lambda)}, & \text{se } \xi = 0. \end{cases}, \tag{14}$$

sugerida por Coles (2001).

Para obter-se o perfil de verossimilhança para x_N, atribuem-se vários valores para x_N e para cada um desses valores maximiza-se (6) com respeito a ξ, obtendo-se assim o valor do máximo do logaritmo da verossimilhança. Então o intervalo de $100(1-\alpha)\%$ de confiança para ξ é dado por todos os pontos para os quais

$$l(\widehat{x}_N) - l(\xi | \widehat{x}_N) \leq \frac{1}{2}\chi^2_{\{1;1-\alpha\}}. \tag{15}$$

Outra aplicação prática importante é o período de retorno que segundo Chow (1964), citado por Wang (1991), é definido como intervalo médio de tempo dentro do qual um evento de nível de magnitude x será igualado ou excedido ao menos uma vez e é dado por

$$T_p = \frac{1}{\lambda\left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-\frac{1}{\xi}}} \tag{16}$$

cuja estimativa é obtida substituindo-se $\bar{\sigma}$, ξ e λ por suas estimativas de máxima verossimilhança $\widehat{\bar{\sigma}}, \widehat{\xi}$ e $\widehat{\lambda}$.

Por fim, para verificar a qualidade do ajuste, Davison e Smith (1990) sugerem que sejam construídos adicionalmente ao gráfico das médias dos excessos, os gráficos de probabilidade-probabilidade, ("PP-plot") e de quantil-quantil ("QQ-plot").

Sejam os excedentes do valor limiar colocados em ordem crescente $\{x_{(1)} < \ldots < x_{(k)}\}$. Segundo Coles (2001), o gráfico de probabilidade-probabilidade é definido pelos pontos de coordenadas

$$\left\{\frac{1}{i+1}, \widehat{G}(x_{(i)})\right\}$$

para $i = 1, \ldots, k$, sendo

$$\widehat{G}(x_{(i)}) = 1 - \left[1 + \widehat{\xi}\left(\frac{x-u}{\widehat{\sigma}}\right)\right]^{-\frac{1}{\widehat{\xi}}}.$$

Por outro lado, o gráfico quantil-quantil é definido pelos pontos de coordenadas

$$\left\{\widehat{G}\left(\frac{1}{k+1}\right), x_{(i)}\right\}$$

para $i = 1, \ldots, k$.

A interpretação destes gráficos é a seguinte. Se os pontos desses gráficos seguirem uma tendência linear significa que os excessos seguem a distribuição generalizada de Pareto, portanto, pode-se dizer que o modelo ajusta-se bem ao conjunto de dados analisado.

4 Resultados e discussão

Nesta seção, são apresentados e discutidos os resultados obtidos para o mês de fevereiro com a finalidade de ilustrar mais detalhadamente a metodologia e, posteriormente, serão discutidos de forma geral, os resultados para os outros meses do ano.

A Tabela 1 apresenta, para o mês de fevereiro, diversos valores limiares candidatos, o número de eventos excedentes para cada um desses valores e os valores-p dos testes da hipótese $H_0 : \theta = 1$ *versus* $H_a : \theta < 1$ para cada uma das séries parciais formadas a partir dos valores limiares. Os resultados apresentados nessa tabela revelam que não há evidências estatísticas para rejeitar a hipótese de que θ seja igual a 1 para séries parciais formadas a partir do valor limiar $u = 25$ mm, considerando-se um nível de significância 5%, ou seja, os excedentes são aproximadamentes independentes para $u \geq 25$ mm. Nota-se também, que para $u \geq 39$ mm, o número de excedentes é menor que o número de anos observados, ou seja, a partir do valor limiar 39 mm as séries parciais têm um menor número de observações do que uma série de máximos anuais considerando o período de 88 anos observados.

Uma vantagem de se escolher o valor limiar baseado no teste sobre θ em relação ao método proposto por Leadbetter *et al.* (1989) é que não é preciso supor subjetivamente o tamanho dos grupos dos excedentes, além de que a série parcial contém todos os excedentes acima do valor limiar estabelecido e não apenas o máximo de cada grupo de excedentes.

Uma vez determinados os valores limiares candidatos, a próxima etapa consiste na escolha, dentre esses valores, do menor, cujo ajuste do modelo da DGP-Poisson seja adequado.

Tabela 1 - Valores limiares candidatos (u), em mm, números de observações excedentes (k) a cada limiar e valores-p para o teste de hipótese $H_0 : \theta = 1$ *versus* $H_a : \theta < 1$ considerando-se 10.000 permutações dos dados para o mês de fevereiro

u	k	Valor-p	u	k	Valor-p	u	k	Valor-p
13	437	0,000	33	114	1,000	53	34	0,226
14	403	0,000	34	109	1,000	54	32	0,279
15	376	0,000	35	103	1,000	55	30	0,332
16	355	0,000	36	100	1,000	56	27	1,000
17	330	0,000	37	95	1,000	57	26	1,000
18	308	0,001	38	87	0,377	58	24	1,000
19	292	0,001	39	80	1,000	59	24	1,000
20	272	0,004	40	77	1,000	60	23	1,000
21	252	0,010	41	75	0,311	61	23	1,000
22	232	0,010	42	70	0,191	62	22	1,000
23	218	0,020	43	67	0,251	63	21	1,000
24	201	0,025	44	62	1,000	64	20	1,000
25	186	0,053	45	59	0,293	65	17	1,000
26	183	0,065	46	55	1,000	66	17	1,000
27	170	0,111	47	51	1,000	67	15	1,000
28	161	0,145	48	48	1,000	68	14	1,000
29	151	0,256	49	46	1,000	69	14	1,000
30	141	0,299	50	41	0,317	70	14	1,000
31	134	0,198	51	39	0,217	71	13	1,000
32	125	0,245	52	36	0,294	72	13	1,000

Como método para avaliar o ajuste do modelo citado, utilizou-se uma inspeção visual do gráfico das médias dos excessos. Para o mês de fevereiro a Figura 1 apresenta o gráfico das médias dos excessos para alguns valores limiares sendo, para a elaboração dos envelopes simulados, utilizadas 99 simulações de Monte Carlo. Analisando-se essa figura, não se observa falta de ajuste do modelo em nenhum dos quatros gráficos apresentados e, além disso, a reta teórica apresentada tem coeficiente angular praticamente nulo sugerindo que ξ seja aproximadamente nulo. Uma vez que para o limiar $u = 25$ mm não foi observada falta de ajuste, escolheu-se esse valor para o modelo final.

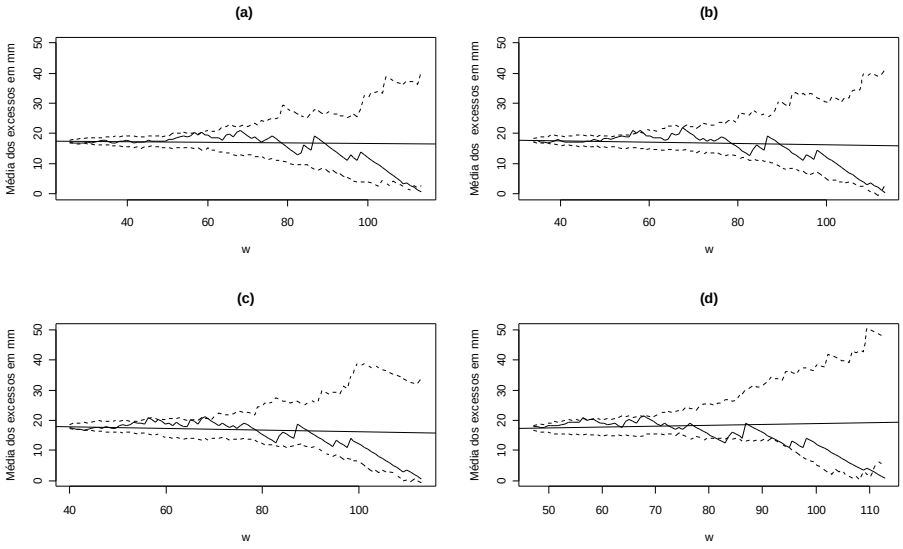

Figura 1 - Gráfico das médias dos excessos para quatro valores limiares: (a) 25 mm, (b) 33 mm, (c) 39 mm e (d) 46 mm para os dados de precipitação total diária, durante o mês de fevereiro, em Piracicaba – SP, com envelopes simulados utilizando-se 99 simulações de Monte Carlo.

Para cada um dos outros meses do ano, embora tenham sidos construídos aproximadamente 4 a 5 gráficos de média dos excessos, são apresentados, na Figura 2, somente os gráficos finais, após a escolha do valor limiar u. Analisando a Figura 2 observa-se que, com exceção dos meses de novembro e fevereiro, para todos os outros meses a estimativa da reta teórica tem coeficiente angular negativo, sugerindo que $\xi < 0$, ou seja, que se trata de uma distribuição Weibull.

Na Tabela 2, são apresentadas as estimativas de máxima verossimilhança dos parâmetros do modelo da DGP-Poisson para cada mês do ano, cujos valores limiares foram escolhidos utilizando-se o mesmo procedimento utilizado para a análise dos dados de fevereiro.

Analisando a Tabela 2, vê-se que em média, para o mês de fevereiro, há aproximadamente 2,11 observações excedentes de 25 mm por ano e que, para todos os meses do ano, $\widehat{\xi}$ está entre -0,5 e 0,5 o que, segundo Smith (1985) significa que o estimador de máxima verossimilhança é regular e a teoria clássica de estimação de máxima verossimilhança pode ser aplicada sem problemas. Além disso, embora os erros padrão das estimativas de ξ sejam elevados, o que poderia sugerir a exclusão desse parâmetro do modelo, optou-se por deixá-lo no modelo final, seguindo a orientação de Smith (2001).

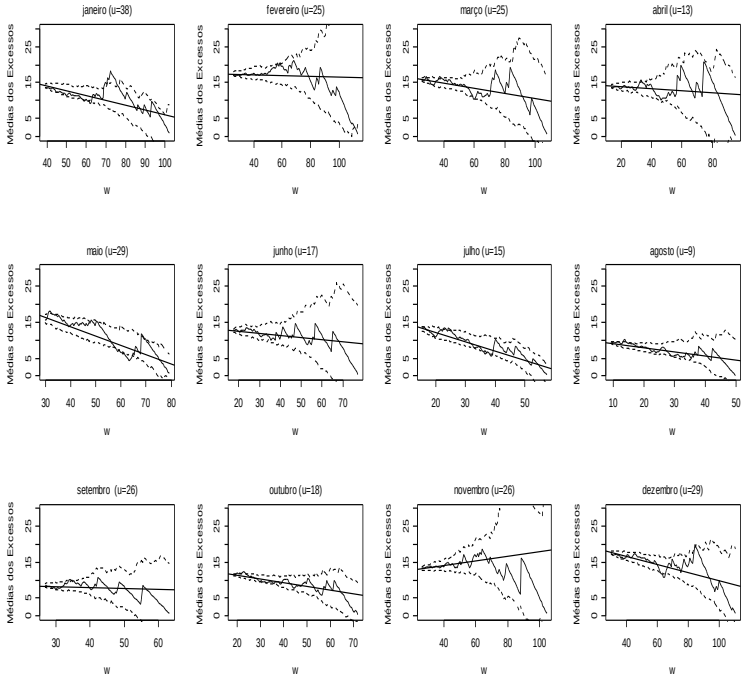

Figura 2 - Gráfico das médias dos excessos para todos os meses do ano.

Tabela 2 - Limiares, número de excedentes e estimativas de máxima verossimilhança dos parâmetros λ, σ e ξ da distribuição generalizada Pareto-Poisson e respectivos erros padrões entre parênteses

Meses	u	k	$\widehat{\lambda}$		$\widehat{\widetilde{\sigma}}$		$\widehat{\xi}$	
Janeiro	38	114	1,295	(0,121)	16,43	(1,96)	-0.1549	(0,0748)
Fevereiro	25	186	2,114	(0,155)	17,65	(1,85)	-0.0116	(0,0747)
Março	25	159	1,807	(0,143)	17,14	(1,89)	-0.0770	(0,0767)
Abril	13	147	1,670	(0,138)	14,37	(1,66)	-0.0294	(0,0811)
Maio	29	30	0,341	(0,062)	22,73	(5,86)	-0.3597	(0,1933)
Junho	17	72	0,818	(0,096)	13,53	(2,29)	-0.0627	(0,1214)
Julho	15	40	0,455	(0,072)	18,32	(3,80)	-0.3531	(0,1452)
Agosto	9	99	1,125	(0,113)	10,59	(1,51)	-0.1388	(0,1019)
Setembro	26	52	0,591	(0,082)	8,62	(1,76)	-0.0294	(0,1494)
Outubro	18	199	2,261	(0,160)	12,82	(1,28)	-0.1133	(0,0702)
Novembro	26	139	1,580	(0,134)	12,34	(1,58)	0.0600	(0,0961)
Dezembro	29	177	2,011	(0,151)	20,03	(2,03)	-0.1249	(0,0686)

Além disso, observa-se que em agosto e outubro a abril, para cada série parcial, o número de observações é maior que 88, resultando em uma média anual de excedentes maior que 1, não observada para os meses de maio a julho e setembro. Este fato mostra que a metodologia proposta neste trabalho para seleção do valor limiar pode gerar séries de duração parcial com número de observações menor do que o número de anos observados. Na Tabela 3, são apresentados os níveis de retornos e respectivos intervalos de 95% de confiança baseados no método do perfil de verossimilhança máxima. Assim, por exemplo, para fevereiro, espera-se que o total diário de precipitação pluvial de 116,6 mm seja excedido, em média, uma vez a cada 100 anos.

Tabela 3 - Níveis de retornos e respectivos intervalos de 95% de confiança construídos por método do perfil de verossimillhança, em mm, para os períodos de retorno de 5, 10, 25, 50 e 100 anos para cada um dos meses do ano

Meses	5 anos	10 anos	25 anos	50 anos	100 anos
Janeiro	64,7	72,7	82,2	88,5	94,2
	(60,8;69,3)	(68,0;79,1)	(76,3;92,5)	(81,6;103,1)	(86,3;113,7)
Fevereiro	66	77,9	93,4	105,1	116,6
	(60,8;73,0)	(70,8;88,9)	(83,0;113,2)	(91,4;134,2)	(99,0;157,5)
Março	59,7	69,5	81,6	90,2	98,4
	(55,2;65,3)	(63,7;78,1)	(73,8;96,8)	(80,4;112,3)	(86,3;129,1)
Abril	42,6	51,8	63,8	72,6	81,3
	(38,4;47,9)	(46,3;60,3)	(56,0;79,3)	(62,6;95,8)	(68,7;114,1)
Maio	40,0	51,5	63,0	69,4	74,4
	(30,0;60,0)	(40,0;70,0)	(55,3;74,4)	(61,6;89,6)	(66,6;92,2)
Junho	35,2	43,6	54,2	61,8	69,1
	(31,5;39,9)	(38,6;50,8)	(47,3;68,7)	(53,2;85,3)	(58,4;105,3)
Julho	28,1	36,5	44,9	49,7	53,4
	(24,4;32,7)	(31,7;42,0)	(39,7;53,1)	(44,1;62,2)	(47,8;66,9)
Agosto	25,3	30,8	37,3	41,7	45,7
	(22,6;28,4)	(27,5;35,4)	(33,1;45,8)	(36,7;54,8)	(39,7;64,4)
Setembro	35,2	40,9	48,3	53,8	59,1
	(33,0;38,7)	(37,4;45,8)	(43,4;58,8)	(47,5;71,9)	(51,2;88,7)
Outubro	45,2	51,7	59,5	64,9	69,9
	(42,2;49,0)	(47,9;57,5)	(54,4;69,3)	(58,6;78,8)	(62,2;88,6)
Novembro	53,1	63,0	77,7	87,6	99,0
	(48,9;58,8)	(57,0;72,9)	(67,4;96,6)	(74,6,;119,1)	(81,7;146,9)
Dezembro	69,2	79,1	91,1	99,2	106,7
	(64,5;75,0)	(73,3;87,6)	(83,6;104,9)	(90,1;118,7)	(95,6;132,9)

A Tabela 4 mostra os períodos de retornos para os níveis de 40, 60, 80 e 100 mm respectivamente e analisando-a vê-se que para fevereiro espera-se, em média, que ao menos 1,1 vez por ano, ocorra um total de 40 mm de precipitação pluvial diária e espera-se, ao menos uma vez a cada aproximadamente 37 anos, um total de 100 mm.

Tabela 4 - Períodos de retornos, em anos, de precipitações totais diárias de 40, 60, 80 e 100 mm para cada um dos meses do ano

Meses	40 mm	60 mm	80 mm	100 mm
Janeiro	0,9	3,5	20,0	> 200
Fevereiro	1,1	3,5	11,3	37,0
Março	1,4	5,1	22,1	115,0
Abril	4,1	18,6	90,1	> 200
Maio	3,2	19,1	> 200	–
Junho	7,4	42,3	> 200	> 200
Julho	14,2	> 200	–	–
Agosto	38,0	> 200	> 200	–
Setembro	8,9	112,0	> 200	> 200
Outubro	3,0	26,5	> 200	> 200
Novembro	1,9	8,1	30,9	106,3
Dezembro	0,9	2,8	10,6	53,6

Como forma complementar de se verificar a qualidade do ajuste da distribuição generalizada de Pareto para as séries de excessos em relação ao limiar adotado, as Figuras 3 e 4 apresentam, para cada mês do ano, os gráficos quantil-quantil e gráficos probabilidade-probabilidade, respectivamente. Esses gráficos sugerem um bom ajuste da distribuição generalizada de Pareto aos excessos em relação ao limiar utilizado, para todos os excessos do ano.

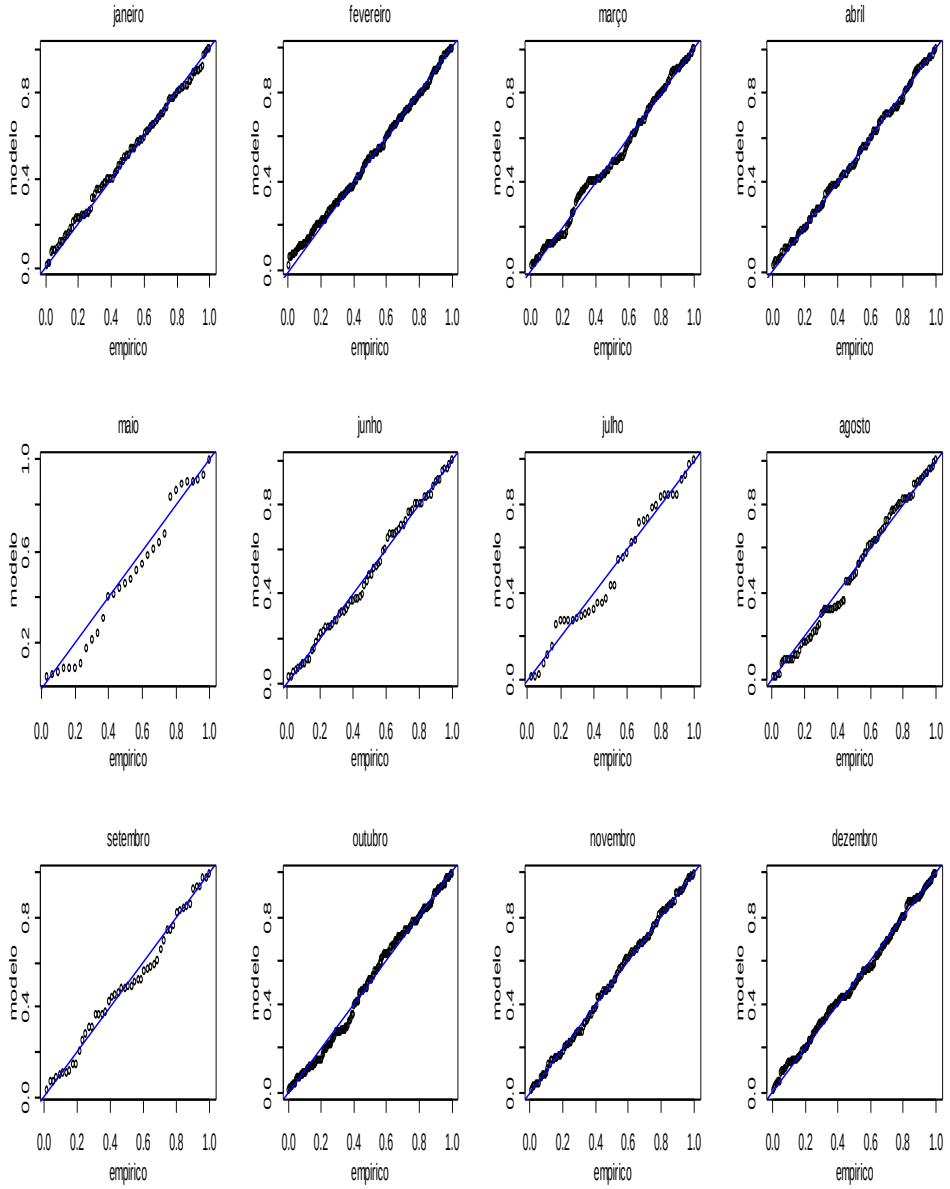

Figura 3 - Gráfico probabilidade-probabilidade para o ajuste da distribuição generalizada de Pareto para todos os meses do ano.

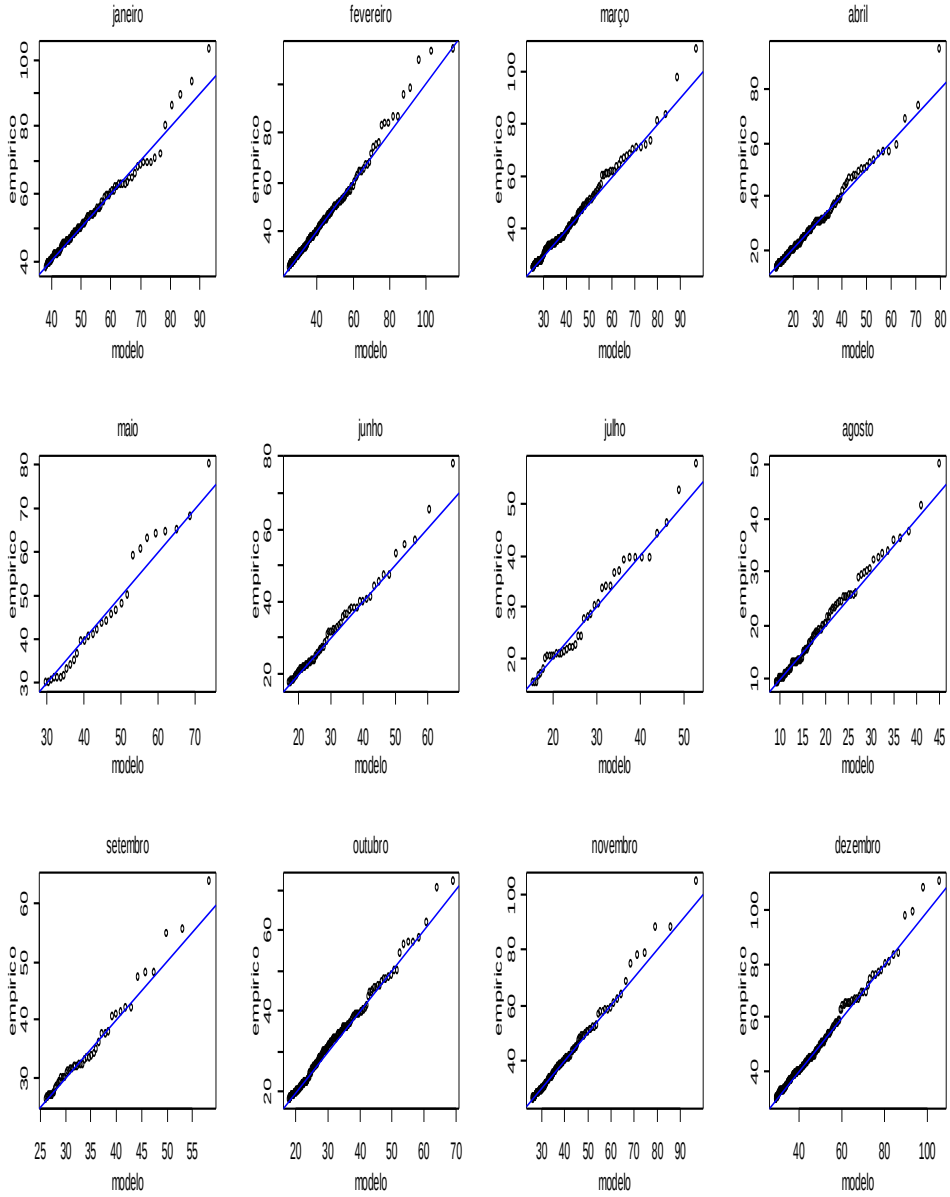

Figura 4 - Gráfico quantil-quantil para para todos os meses do ano.

Conclusões

A metodologia apresentada para seleção do valor limiar é uma alternativa aos métodos já existentes sendo os meses que apresentam maiores totais de precipitação pluvial diárias são os meses de novembro a março. E, por fim, em um período de 50 anos, espera-se, em média, precipitações diárias totais acima de 85 mm apenas para os meses de novembro a março.

SILVA, R. R.; ZOCCHI, S. S. The generalized Pareto-Poisson distribution in study of maximum daily total rainfall in Piracicaba, SP. *Rev. Mat. Estat.,* São Paulo, v.24, n.3, p.77-94, 2005.

- *ABSTRACT: In this study we used the Generalized Pareto-Poisson distribution to analyze a data set of daily total rainfall exceeding a threshold value as recorded from 1917 to 2004 in Piracicaba, SP. The threshold value was chosen so that the excess would follow a generalized Pareto distribution and the mean cluster size would not differ from one. Once the threshold value was chosen, the parameters of the model were estimated using a maximum likelihood method. Based on its estimates, we estimated the return levels for 10, 25, 500 and 100 years and their 95% confidence intervals using the likelihood profile method. Moreover, return periods for daily total rainfall of 40, 60, 80 and 100 mm were calculated. The goodness of fit of the models were visually evaluated by means of probability-probability plots and quantile-quantile plots. As a result of the analysis, a total daily rainfall over 85 mm is expected once every 50 years from November to March on average, and thus we conclude that, in this period, the year's highest total daily rainfall is expected.*

- *KEYWORDS: Theory value extreme; return period; level return.*

Referências

CHOW, V. T. *Handbook of applied hydrology*: A compendium of water resources tecnology. New York: McGraw-Hill, 1964. 1486p.

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

CUNNANE, C. A particular comparison of annual maxima and partial series methods of flood frequency prediction. *J. Hidrol.*, Amsterdam, v.18, p.257-271, 1973.

DAVISON, A. C.; SMITH, R. L. Models for excedances over high thresholds. *J. R. Stat. Soc., Ser. B.*, London, v.52, p.393-442, 1990.

FERRO, C. A. T.; SEGERS, J. Inference for clusters of extremes values. *J. R. Stat. Soc., Ser. B.*, London, v.65, p.545-556, 2003.

JEKINSON, A. F. The frequency distribution of the annual maximun (or minimum) values of metereological elements. *Q. J. R. Meteorol. Soc.*, Brackneel, v.81, p.158-171, 1955.

LEADBETTER, M. R.; LINDGREN, G.; ROOTZÉN, H. *Extremes and related properties for random sequences and series.* New York: Springer, 1983. 336p.

LEADBETTER, M. R. et al. *On clustering of high levels in statistically stationary series.* In: INT. MEETING ON STATIST. CLIM., 4, 1989, Wellington. *Prociding ...* Wellington: Ed. J. Sansom. 1998.

RASMUSSEN, P. F.; ROSBJERG, D. Risk estimation in partial duration series. *Water Res. Res.*, Washington, v.25, p.2319-2330, 1989.

RASMUSSEN, P. F.; ROSBJERG, D. Prediction uncertainty in seasonal partial duration series. *Water Res. Res*, Washington, v.27, p.2875-2883, 1991.

ROSBJERG, D.; MADSEN, H.; RASMUSSEN, P. F. Prediction in partial duration series with generalized Pareto-distributed exceedances. *Water Res. Res.*, Washington, v.28, p.3001-3010, 1992.

SMITH, R. L. Maximum likehood estimation in a class of nonregular cases. *Biometrika*, London, v.72, p.67-90, 1985.

SMITH, R. L. *Enviroment Statistics*: extreme value theory section. Disponível em: http: // www.stat.unc.edu/postscript/rs/envstat/env.html. Acesso em: 15 out. 2003.

WANG, Q. J. The POT model described by the generalized Pareto distribution with Poisson arrival rate. *J. Hydrol.*, Amsterdam, v.129, p.263-280, 1991.

Recebido em 02.05.2006.

Aprovado após revisão em 25.09.2006.