XX SIMPÓSIO BRASILEIRO DE RECURSOS HÍDRICOS
ÁGUA • DESENVOLVIMENTO ECONÔMICO E SOCIOAMBIENTAL
17 – 22 DE NOVEMBRO DE 2013 – BENTO GONÇALVES / RS

ABRH
ASSOCIAÇÃO BRASILEIRA DE RECURSOS HÍDRICOS

UM MODELO PARA SÉRIES DE DURAÇÃO PARCIAL DE VAZÕES DE CHEIA, COM NÚMEROS DE OCORRÊNCIAS SEGUNDO A DISTRIBUIÇÃO ZIP (*ZERO-INFLATED-POISSON*)

Artur Tiago Silva [1] *& Maria Manuela Portela* [2] *& Mauro Naghettini* [3*]

Resumo – As séries de duração parcial, com números de ocorrências de Poisson e excedências segundo a lei Generalizada de Pareto (GP), continua sendo uma ferramenta útil na análise de frequência de extremos hidrológicos. O modelo Poisson-GP exige a validação da hipótese de que o número anual de eventos de cheia, acima de um valor limiar, seja distribuído segundo a lei de Poisson. Observa-se, entretanto, que essa hipótese nem sempre é válida em aplicações práticas. Este artigo propõe uma distribuição alternativa para a modelagem do número anual de cheias, nomeadamante, o modelo distributivo ZIP (*Zero-Inflated Poisson*), de dois parâmetros. O modelo resultante da combinação das distribuições ZIP e GP, é descrito e avaliado. Verificou-se que o modelo ZIP-GP é menos restritivo que o Poisson-GP, uma vez que propicia uma descrição mais precisa do processo de ocorrência de cheias, sob a representação de séries de duração parcial. Descreve-se também uma aplicação do modelo ZIP-GP a amostras de vazões de cheia observadas no norte de Portugal, e faz-se uma avaliação de seu desempenho. Os resultados demonstram a superioridade do modelo ZIP-GP, principalmente para os quantis da cauda inferior, e indicam uma alternativa válida para a análise de frequência de séries hidrológicas de duração parcial.

Palavras-Chave – Análise de frequência, séries de duração parcial, modelo ZIP-GP.

A PEAKS-OVER-THRESHOLD (POT) MODEL FOR FLOODS WITH ZERO-INFLATED POISSON (ZIP) ARRIVALS

Abstract – The peaks-over-threshold (POT) model for hydrological extremes with Poisson arrivals and Generalized Pareto (GP) distributed exceedances remains a useful tool for modeling hydrologic extremes. The Poisson-GP model for flood frequency analysis requires the validation of the hypothesis that the distribution of the annual number of flood events may be described by a Poisson distribution. Such hypothesis is not always valid in practical applications. This paper concerns the use of an alternative distribution for modeling the annual number of floods - the Zero-Inflated Poisson (ZIP) distribution with two parameters. A ZIP-GP model for flood frequency analysis is proposed. This model is less restrictive than the Poisson-GP model since it allows for a more accurate description of the occurrence process in a POT framework. An application of the ZIP-GP model to flood data from Northern Portugal and the evaluation of its performance is presented herein. The results show that the ZIP-GP model outperforms the Poisson-GP, especially for lower quantiles, thus suggesting it is a valid alternative to the Poisson distribution for modeling the annual occurrence counts of peaks in a POT approach for flood frequency analysis.

Key words: frequency analysis, peaks-over-threshold, ZIP-GP model.

[1] Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal. artur.tiago.silva@ist.utl.pt
[2] Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal. mps@civil.ist.utl.pt.
[3*] Universidade Federal de Minas Gerais, Belo Horizonte (MG), Brasil. naghet@netuno.lcc.ufmg.br ou mauronag@superig.com.br.

INTRODUÇÃO

O segundo teorema da teoria de valores extremos, i.e. o teorema de Pickands-Balkema-de Haan (Balkema e de Haan, 1974; Pickands, 1975), define que a cauda superior assintótica de uma variável aleatória X pertence à família de distribuições sintetizada pela distribuição Generalizada de Pareto (GP). Em decorrência, a distribuição GP tem sido largamente empregada como modelo probabilístico das excedências sobre um valor limiar, dentro do quadro de séries de duração parcial, ou POT da terminologia inglesa *peaks-over-threshold*. A função acumulada de probabilidades da distribuição GP é dada por

$$G(x)=\begin{cases}1-\left[1-\kappa\left(\dfrac{x-\mu}{\sigma}\right)\right]^{\frac{1}{\kappa}}, & \kappa\neq 0\\ 1-\exp\left(-\dfrac{x-\mu}{\sigma}\right), & \kappa=0\end{cases} \qquad (1)$$

onde κ representa o parâmetro de forma, σ o de escala e μ denota o parâmetro de posição, dado pelo valor limiar que define as excedências. A distribuição GP reduz-se à exponencial quando κ=0. No contexto de série de duração parcial, se a frequência temporal com que as excedências ocorrem puder ser descrita por um processo homogêneo de Poisson, com parâmetro λ, dado pelo número anual médio de excessos sobre o valor limiar, a distribuição de probabilidades acumuladas dos valores máximos anuais, F(x), é expressa por (Davison e Smith, 1990)

$$F(x)=\exp\{-\lambda[1-G(x)]\} \qquad (2)$$

Combinando as equações (1) e (2), obtém-se

$$F(x)=\begin{cases}\exp\left\{-\lambda\left[1-\kappa\left(\dfrac{x-\mu}{\sigma}\right)\right]^{\frac{1}{\kappa}}\right\}, & \kappa\neq 0\\ \exp\left[-\lambda\exp\left(-\dfrac{x-\mu}{\sigma}\right)\right], & \kappa=0\end{cases} \qquad (3)$$

que representam as formas analíticas da distribuição Generalizada de Valores Extremos (GEV), incluindo o caso particular da distribuição de valores extremos do tipo 1, ou EV1 ou Gumbel, para κ=0. Este resultado está em acordo com o primeiro teorema da teoria de valores extremos, i.e. o teorema de Fisher-Tippett-Gnedenko (Fisher e Tippett, 1928; Gnedenko, 1943), segundo o qual, a distribuição do máximo anual de uma variável aleatória é de um dos três tipos particulares da distribuição GEV, em correspondência aos casos em que o parâmetro κ é nulo, negativo ou positivo.

Neste contexto, o esquema Poisson-GP para a modelagem de séries de duração parcial (SDP) permanece como uma alternativa válida e útil relativamente à análise convencional de séries de máximos anuais (SMA). Contudo, o referido esquema exige a validação da hipótese de que o número de excedências, em relação ao valor limiar, seja distribuído conforme uma lei de Poisson. Cunnane (1979) propôs um teste para verificação desta hipótese, com base no índice de dispersão (quociente entre a variância e a média) do número anual observado de excedências sobre o valor limiar. O referido autor aplicou o teste do índice de dispersão aos dados fluviométricos de 26 estações da Grã-Bretanha, havendo concluído que o número anual de excedências sobre uma vazão

limiar não se conformava a uma variável de Poisson e que os desvios deviam-se ao fato da variância ser significativamente maior do que a média. Como modelo alternativo, Cunnane (1979) propôs a distribuição Binomial Negativa (BN), a qual admite maior variância, embora tenha concluído, em seguida, que o seu emprego também não oferecia resultados plenamente satisfatórios. Neste artigo, propõe-se, como alternativa aos modelos Poisson e BN, a distribuição ZIP (*Zero-Inflated Poisson*) para a modelagem no número anual de excedências sobre um dado valor limiar.

A DISTRIBUIÇÃO ZIP (*ZERO-INFLATED POISSON*)

A distribuição mais simples para modelar a contagem de dados em tempo contínuo, como o número anual de excedências sobre um valor limiar, é a de Poisson com parâmetro λ e função massa de probabilidades (FMP) dada por

$$P(Y=y)=\exp(-y)\frac{\lambda^y}{y!},\quad y=0,1,2,... \tag{4}$$

A distribuição ZIP (*Zero-Inflated Poisson*) é um modelo resultante da mistura de duas componentes, a saber, uma, com massa no ponto zero, e a outra, dada pela distribuição de Poisson, de modo que a FMP resultante seja

$$\begin{cases} P(Y=0)=\phi+(1-\phi)\exp(-\lambda) \\ P(Y=y)=(1-\phi)\exp(-\lambda)\dfrac{\lambda^y}{y!},\quad y=1,2,3... \end{cases} \tag{5}$$

ou

$$P(Y=y)=\left[\phi+(1-\phi)\exp(-\lambda)\right]^{I_{\{0\}}(y)}\left[(1-\phi)\exp(-\lambda)\frac{\lambda^y}{y!}\right]^{1-I_{\{0\}}(y)} \tag{6}$$

Nas equações (5) e (6), o parâmetro ϕ representa a inflação de massa no ponto y=0 e $I_{\{0\}}(y)$ é igual a 1, quando y=0, e igual a 0,quando y≠0. A função de verossimilhança para a distribuição ZIP é

$$L(\phi,\lambda\mid y)=\prod_{i=1}^{n}\left\{\left[\phi+(1-\phi)\exp(-\lambda)\right]^{I_{\{0\}}(y_i)}\left[(1-\phi)\exp(-\lambda)\frac{\lambda^{y_i}}{y_i!}\right]^{1-I_{\{0\}}(y_i)}\right\} \tag{7}$$

e a função log de verossimilhança é

$$\ell(\phi,\lambda\mid y)=\sum_{i=1}^{n}\left\{I_{\{0\}}(y_i)\ln\left[\phi+(1-\phi)\exp(-\lambda)\right]+\left[1-I_{\{0\}}(y_i)\right]\ln\left[(1-\phi)\exp(-\lambda)\frac{\lambda^{y_i}}{y_i!}\right]^{1-I_{\{0\}}(y_i)}\right\} \tag{8}$$

As estimativas de máxima verossimilhança dos parâmetros ϕ e λ são obtidas mediante solução da equação (8).

O MODELO ZIP-GP

Sob o esquema de amostragem de uma série de duração parcial de dados fluviométricos, e fazendo uso da premissa de que as vazões de pico que excederam um dado valor limiar sejam

variáveis IID (independentes e igualmente distribuídas), a probabilidade de não-excedência dos máximos anuais, denotados por X_{AM}, pode ser formalmente expressa por

$$P(X_{AM} \le x) = P(Y=0) + \sum_{y=1}^{\infty} P\left[\bigcap_{k=1}^{y}(X_{OT} \le x) \bigcap (Y=y)\right] \quad (9)$$

onde a variável aleatória Y descreve a contagem do número anual de ocorrências e X_{OT_k} denota a k-ésima vazão de pico que excedeu a vazão limiar, em um dado ano. Se as magnitudes X_{OT_k} forem supostas independentes de seus tempos de ocorrência, a equação (9) torna-se

$$P(X_{AM} \le x) = P(Y=0) + \sum_{y=1}^{\infty} \left\{P(Y=y)[G(x)]^y\right\} \quad (10)$$

onde G(x) é a função de probabilidades acumuladas (FPA) das vazões de pico que excederam a vazão limiar. Sob a hipótese de estacionariedade, os parâmetros ϕ e λ são considerados constantes no tempo. Além disso, a distribuição de probabilidades das excedências sobre um dado valor limiar u, ou seja G(x), também é tomada como invariante no tempo. Substituindo as expressões de P(Y=y) de (5) na equação (10), é possível, mediante manipulação algébrica, obter a seguinte expressão:

$$P(X_{AM} \le x) = \phi + (1-\phi)\exp\{-\lambda[1-G(x)]\} \quad (11)$$

Se $P(X_{AM} \le x)$ for denotado por F(x) e se G(x) for substituído pela expressão da distribuição GP, dada por (1), a equação (11) torna-se

$$F(x) = \begin{cases} \phi + (1-\phi)\exp\left\{-\lambda\left[1-\kappa\left(\dfrac{x-\mu}{\sigma}\right)\right]^{\frac{1}{\kappa}}\right\}, & \kappa \neq 0 \\ \phi + (1-\phi)\exp\left[-\lambda\exp\left(-\dfrac{x-\mu}{\sigma}\right)\right], & \kappa = 0 \end{cases} \quad (12)$$

As funções de quantis correspondentes são

$$x_F(F) = \begin{cases} \mu + \dfrac{\sigma}{\kappa}\left\{1-\left[-\dfrac{1}{\lambda}\ln\left(\dfrac{F-\phi}{1-\phi}\right)\right]^{\kappa}\right\}, & \kappa \neq 0 \\ \mu - \sigma\ln\left[-\dfrac{1}{\lambda}\ln\left(\dfrac{F-\phi}{1-\phi}\right)\right], & \kappa = 0 \end{cases} \quad (13)$$

EXEMPLO DE APLICAÇÃO

A seguir descreve-se um exemplo de análise de frequência de cheias com o emprego do modelo ZIP-GP. A amostra usada no exemplo foi obtida a partir de 64 anos de vazões médias diárias coletadas na estação fluviométrica de Quinta das Laranjeiras, no rio Sabor, no norte de Portugal (41.208ºN, 7.059ºO; bacia hidrográfica com 3464 km^2; http://snirh.pt), as quais encontram-se ilustradas no fluviograma da Figura 1.

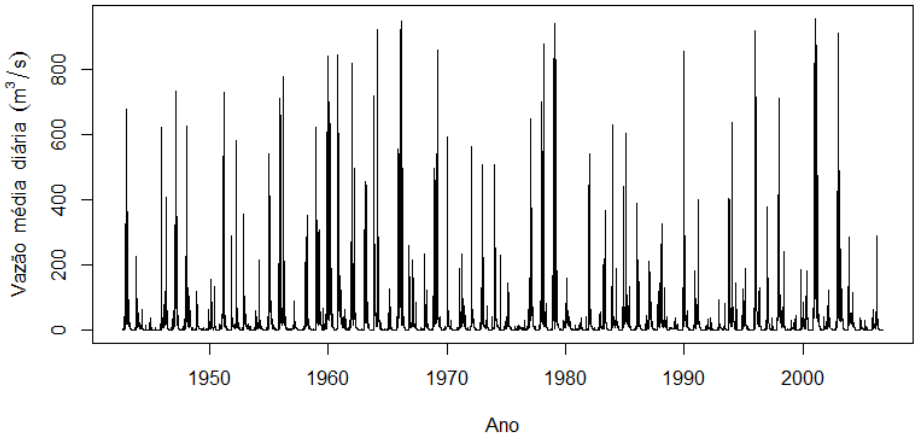

Figura 1 – Vazões médias diárias do rio Sabor em Quinta das Laranjeiras.

Para garantir a independência estatística entre as vazões de pico, selecionaram-se acontecimentos separados no tempo por um período mínimo de 3 dias. Além disso, impôs-se que a vazão remanescente entre dois picos consecutivos deveria decrescer até pelo menos 2/3 da vazão do primeiro pico (Lang et al., 1999).

A seleção da vazão limiar permanece como o aspecto mais subjetivo da modelagem de séries de duração parcial e, de fato, não há regra consensual para fazê-la. Neste artigo, a vazão limiar foi selecionada com o apoio do gráfico da função de "vida remanescente média", ou MRL (*mean residual life*), ilustrado na Figura 2b (Lang et al., 1999). A vazão limiar u=300 m^3/s foi escolhida pelo fato dos limites do intervalo de confiança a 95 % acomodarem um trecho linear da função MRL, para u>300 m^3/s, embora a cauda exponencial, que corresponderia a um trecho constante para a MRL, não pareça provável.

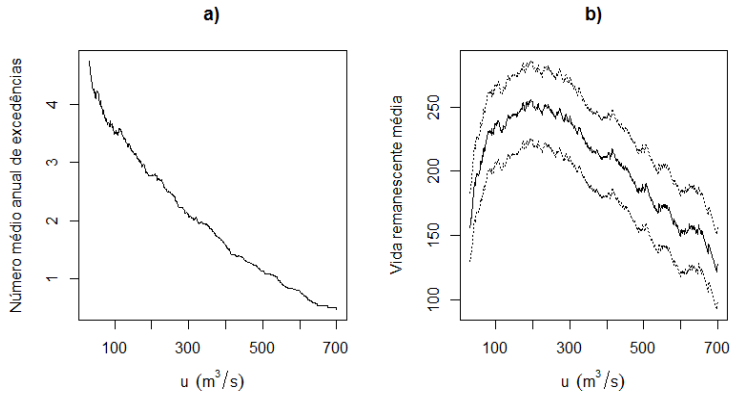

Figura 2 - (a) Número médio anual de excedências em função da vazão limiar u e (b) vida remanescente média (MRL) com os limite do intervalo de confiança a 95 %.

Na Figura 3a encontra-se o gráfico do Índice de Dispersão (DI). Parece evidente que a hipótese de Poisson deva ser rejeitada para todas as vazões limiares maiores do que u=100 m^3/s, incluindo o valor selecionado u=300 m/s. Para comparar as aderências das leis de Poisson e ZIP aos números anuais de ocorrências, Y, para cada vazão limiar, o Critério de Informação de Akaike (AIC), conforme proposto por Akaike (1974), foi aqui empregado, uma vez que as citadas distribuições têm diferentes números de parâmetros. A Figura 3b mostra as diferenças do AIC das distribuições de Poisson e ZIP, para diferentes vazões limiares. É evidente que o modelo ZIP provê a melhor aderência às amostras de **y**, para todas as vazões limiares superiores a u=50 m^3/s, incluindo o valor selecionado de u=300 m^3/s.

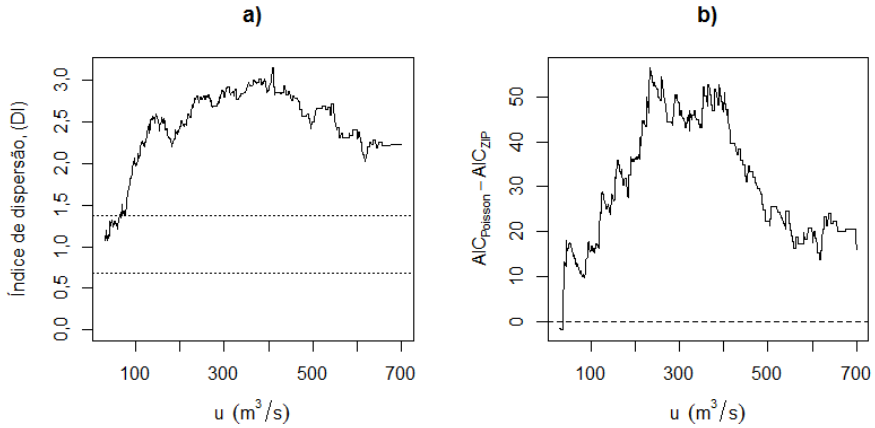

Figura 3 - (a) Gráfico do Índice de Dispersão (DI) e (b) diferenças entre os AICs das leis de Poisson ($AIC_{Poisson}$) e ZIP (AIC_{ZIP}).

A Figura 4 ilustra as distribuições empíricas da FMP e da FPA para u=300 m^3/s, juntamente com as modelos Poisson e ZIP ajustados. Parece evidente a superioridade relativa do modelo ZIP, em termos de aderência à amostra.

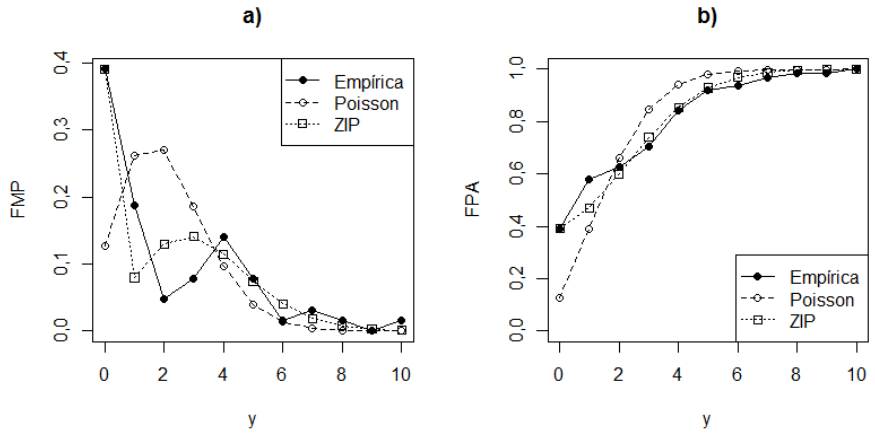

Figura 4 - (a) FMPs empírica, de Poisson e ZIP e (b) FPAs empírica, de Poisson e ZIP.

A Figura 5 mostra as distribuições Exponencial e GP ajustadas aos dados X_{OT}. Como sugerido pela cauda descendente do gráfico da MRL (Fig. 2b), as vazões de pico não parecem ser exponencialmente distribuídas. De fato, o ajuste sugere uma cauda superior limitada superiormente, ou seja, com parâmetro de forma κ positivo.

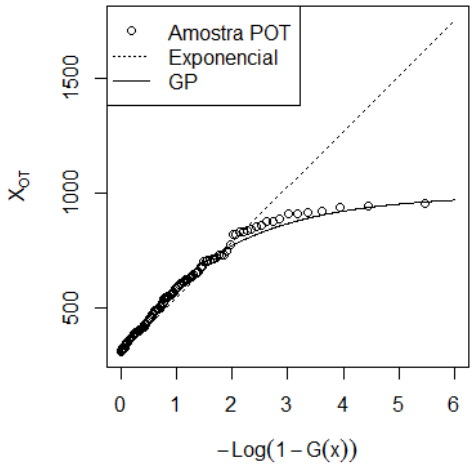

Figura 5 - Ajustamentos dos modelos Exponencial e GP à serie vazões de pico que excederam o valor limiar u=300 m3/s.

Finalmente, na Figura 6, os modelos completos Poisson-GP e ZIP-GP para valores máximos anuais estão grafados juntamente às vazões observadas X_{AM}. É aparente que os dois modelos coincidem para os quantis elevados. Contudo, no cômputo geral, o modelo ZIP-GP claramente provê uma representação mais precisa dos dados observados, particularmente para os quantis de baixo período de retorno.

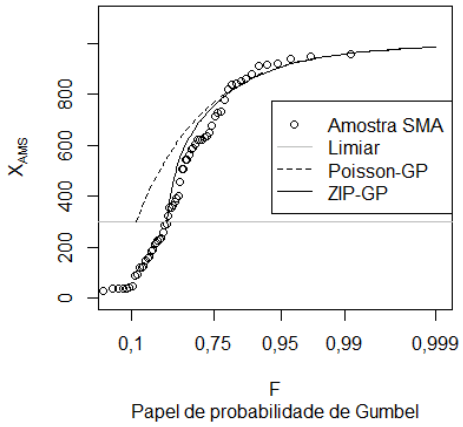

Figura 6 - Ajustamentos dos modelos Poisson-GP e ZIP-GP à serie de vazões máximas anuais.

CONCLUSÕES

Há duas conclusões principais do exemplo de aplicação da metodologia aqui descrita:

1. a distribuição ZIP é uma alternativa válida, relativamente à lei de Poisson, para a modelagem do número anual de vazões de pico que excederam um valor limiar, dentro do contexto da análise de frequência de séries de duração parcial; e
2. embora o maior interesse da análise de frequência de valores máximos concentre-se na cauda superior, onde a lei de Poisson aparenta permanecer como modelo válido, a distribuição ZIP oferece uma modelagem alternativa mais adequada ao conjunto das observações, particularmente para os quantis de baixos períodos de retorno.

Como desenvolvimentos futuros da presente pesquisa, ressaltam-se os seguintes aspectos a serem melhor explorados:

- as conclusões (1) e (2) precisam se apoiar em outros casos de estudo;
- o desempenho do modelo ZIP-GP, do ponto de vista da quantificação das incertezas dos quantis estimados, precisa ser avaliado;
- o modelo ZINB (*Zero-Inflated Negative Binomial*) também é, potencialmente, uma alternativa válida e que necessita de avaliação; e
- a inclusão de eventuais não-estacionariedades na formulação do modelo ZIP-GP é possível, mesmo para G(x) estacionário, a partir do conceito de regressão ZIP (Lambert, 1992) dos números anuais de excedências a partir de covariáveis, tais como índices climáticos.

REFERÊNCIAS

AKAIKE, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6), 716–723.

BALKEMA, A. e DE HAAN, L., 1974. Residual life time at great age. *The Annals of Probability* 2(5), 792–804.

COLES, S., 2001. *An introduction to statistical modeling of extreme values*. Springer.

CUNNANE, C, 1979. A note on the Poisson assumption in partial duration series models. *Water Resources Research* 15(2), 489–494.

DAVISON, A. e SMITH, R, 1990. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)* 52(3), 393– 442.

FISHER, R. e TIPPETT, L., 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In: *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24. Cambridge University Press.

GNEDENKO, B., 1943. Sur la distribution limite du terme maximum d'une série aléatoire. *The Annals of Mathematics* 44(3), 423–453.

LAMBERT, D., 1992. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics* 34(1), 1–14.

LANG, M., OUARDA, T. B. M. J., & Bobée, B., 1999. Towards operational guidelines for over-threshold modeling. *Journal of Hydrology*, 225(3), 103-117.

PICKANDS, J., 1975. Statistical inference using extreme order statistics. *The Annals of Statistics*: 3(1), 119–131.