MEMÓRIAS
DA
ACADEMIA DAS CIÊNCIAS
DE
LISBOA

CLASSE DE CIÊNCIAS

TOMO XLVIII

Desafios em Estatística de Extremos

M. Ivette Gomes

ACADEMIA DAS CIÊNCIAS
DE LISBOA

LISBOA • 2022

Desafios em Estatística de Extremos

M. Ivette Gomes[1]

ABSTRACT

Cheias, fogos, furacões, secas e outros acontecimentos extremos têm fornecido uma razão para os desenvolvimentos recentes da *teoria de valores extremos* (EVT, do inglês, *extreme value theory*). A *estatística de extremos* é hoje em dia confrontada com muitos desafios, especialmente em tópicos relacionados com a modelação de risco e a eficiência e robustez das metodologias que nos permitem compreender a complexidade dos acontecimentos extremos nas mais diversas áreas. O compromisso entre robustez e extremos necessita pois de novos desenvolvimentos e de novas abordagens. Para além da estimação do *índice de valores extremos*, o parâmetro fundamental em EVT, consideraremos a estimação de *quantis extremais* e de *períodos de retornos* de níveis elevados.

Key-Words: Caudas; estatística de extremos; médias generalizadas; risco.

1. INTRODUÇÃO

A *teoria de valores extremos*, muito frequentemente denotada EVT, do inglês *extreme value theory* ajuda-nos a controlar acontecimentos potencialmente desastrosos, de grande relevo para a sociedade e de elevado impacto social. Em EVT, a ordenação da amostra é primordial, e isso permitiu-nos chegar a uma vasta metodologia estatística e associada teoria distribucional relativas a amostras ordenadas, como pode ser visto nos variados livros sobre *estatísticas ordinais* (EO's) e sobre EO's extremais, de que referimos Arnold *et al.* (1992; 2008).

Existe, por um lado, um *interesse natural pela ordenação*: os extremos são importantes como expressão do pior ou do melhor que pode ser encontrado numa amostra (temperaturas mínimas, níveis máximos de barragens, tempos de vida mínimos em teoria da fiabilidade ou análise de sobrevivência). Alternativamente, um conjunto de observações pode ser *deliberadamente ordenado*, de forma a facilitar a análise estatística pretendida. Mencionamos, por exemplo a existência de vários métodos rápidos para estimação de parâmetros ou para testes de significância baseados em estatísticas sistemáticas (combinações lineares de EO's), como a amplitude e a semi-amplitude.

Os domínios de aplicação da EVT são muito variados. Mencionamos, entre outras, as áreas de *bioestatística, engenharia estrutural, finanças, seguros*, e também *ambiente (hidrologia, meteorologia, sismologia*…). Terramotos, fogos, cheias e outros acontecimentos extremos têm levado a

[1] CEAUL and DEIO, FCUL, Universidade de Lisboa (ivette.gomes@fc.ul.pt), Academia de Ciências de Lisboa, Portugal.

re-desenvolvimentos recentes da *análise de valores extremos* (EVA, do inglês *extreme value analysis*), *estatística de extremos univariados* (SUE, do inglês, *statistics of univariate extremes*), e também extremos multivariados e espaciais. Embora seja possível encontrar alguns artigos de interesse histórico relacionados com acontecimentos extremos, o campo remonta a Gumbel, em artigos publicados a partir de 1935, e sumariados em Gumbel (1958; 2004). Gostaríamos ainda de mais uma vez realçar o nome de um português pioneiro da área de extremos, J. Tiago de Oliveira, membro efectivo da Academia das Ciências de Lisboa desde 1985 até à sua morte prematura em 1992, com 63 anos de idade (veja-se Gomes, 1993a, 1994; Tiago de Oliveira, J.C., ed., 1993, entre outros).

Figura 1
Tiago de Oliveira com o seu cachimbo

Até ao início dos anos oitenta a estatística de extremos era essencialmente de índole paramétrica, baseada em resultados assintóticos da teoria de valores extremos. Deu-se então uma mudança para abordagens semi-paramétricas e mesmo não-paramétricas. Mas a modelação paramétrica tornou-se recentemente muito popular, particularmente em aplicações espaciais da EVT.

Os tópicos a abordar são os seguintes: começaremos por uma breve *motivação* para a necessidade da EVT, tentando responder à questão, *Porquê a teoria de valores extremos*? Procederemos em seguida a uma também breve referência à *estatística de extremos*, com a apresentação de dois *estudos de casos*, que amplamente justificam os modelos extremais. Para além da estimação do *índice de valores extremos* (EVI, do inglês *extreme value index*), um dos principais parâmetros em EVT, referiremos brevemente a estimação de outros parâmetros de acontecimentos extremos, como o *value-at-risk* (VaR), dando ênfase à utilização de *médias generalizadas* (GMs, do inglês *generalized means*) e à metodologia PORT para a estimação do EVI e do VaR, onde PORT é o acrónimo de *peaks over random thresholds*.

2. MOTIVAÇÃO PARA A NECESSIDADE DA EVT

É perfeitamente natural perguntar qual o porquê da EVT. Para motivar o interesse por este tema, damos alguns exemplos de grande relevância para a sociedade, e que envolvem esta teoria. Para alguns destes exemplos, e outros, veja-se Beirlant *et al.* (2004) e Gomes *et al.* (2013), também entre outros livros na área de extremos aplicados, de que referimos Coles (2001), Reiss & Thomas (2001, 2007) Castillo *et al.* (2005) e Markovich (2007).

2.1. Extremos e Ambiente

2.1.1. As cheias no Mar do Norte

Na madrugada de 1 de Fevereiro de 1953, o nível das águas excedeu os 5.6 metros acima do nível do mar, destruiu as defesas marítimas, tendo inundado áreas na Holanda, Inglaterra, Bélgica, Dinamarca, França, e cerca de 2500 pessoas morreram.

Figura 2
A cheia no Mar do Norte, 1 de Fevereiro de 1953

2.1.2. O furacão Katrina

Nova Orleães encontra-se situada abaixo do nível do mar, no meio de dois lagos, a norte e a este, e do rio Mississipi a sul. A inundação provocada pelo furacão Katrina, no dia 29 de Agosto de 2005, deveu-se, sobretudo, a uma brecha de 60 metros num dique junto ao lago Pontchartrain.

Figura 3
Nova Orleães após o Katrina, 29 de Agosto de 2005

2.1.3. Terramoto no centro de Itália

Este terramoto ocorreu a 30 de Outubro de 2016 numa região que, apenas quatro dias antes, já havia sido castigada por uma série de tremores de terra. Contudo o novo sismo causou danos, mas NÃO deixou mortos, inclusive em cidades que, em Agosto de 2016, tinham sido destruídas por um tremor de terra que matou mais de três centenas de pessoas.

Figura 4
Destruição em L'Aquila, 30 de Outubro de 2016

2.1.4. Ciclone Idai

Segundo a ONU, tratou-se da *pior tempestade de sempre no Hemisfério Sul*. Este ciclone teve origem numa depressão tropical que se formou na costa leste de Moçambique no início de Março de 2019 e foi ganhando força à medida que seguiu rumo ao continente, com ventos até 177 km/h, centenas de mortos e milhares de desalojados.

Figura 5
Ciclone Idai, Março 2019

2.1.5. Alguns comentários

Traduzimos de forma livre parte de uma notícia do *New York Times*, *Sept'05*, intitulada *New Orleans After Hurricane Katrina: An Unnatural Disaster*? Dizia o redator que teriam de construir um sistema de diques adequado, para o que necessitariam de engenheiros holandeses, capazes de desenhar essas estruturas. A primeira estrutura deveria ser uma barragem com pelo menos 40-50 pés de altura, construída ao longo do lago e de cada canal com ligação ao lago.

Tratar-se-ia de um plano que custaria biliões, mas seria sensato que se aprendesse a lição, de modo a NÃO se ter uma repetição dentro dos próximos 20 anos.

Na realidade, como resultado das cheias do Mar do Norte, o governo holandês constituiu uma comissão (*Delta Committee*). E decretou que os diques deveriam ser construídos com uma altura tal que "a probabilidade de uma inundação num determinado ano fosse de 1 em 10.000". Mas o período de observação dos dados é muitíssimo mais curto!... É então necessário proceder a uma extrapolação para além dos dados observados! E a EVT consegue dar respostas fidedignas sobre a altura da referida barragem, entrando em linha de conta com aquilo a que chamamos *período de retorno* de um acontecimento extremo, que não é mais do que o intervalo de tempo médio entre ocorrências de um determinado valor extremo, como o furacão Katrina ou a cheia no Mar do Norte ou o terramoto no centro de Itália ou o recente ciclone Idai, já seguido pelo ciclone Keneth. A estimação desse parâmetro de acontecimentos raros depende de forma crucial de uma estimação fiável do chamado EVI, já atrás referido.

O mesmo tipo de extrapolação é necessária e desejável relativamente a sismos que ocorrem em locais específicos, tal como o que ocorreu em 2016, no centro de Itália, entre muitos outros registados pelos sismologistas por todo o mundo. E relativamente a este último tópico várias questões podem ser colocadas: (1) Com base nos dados, o que é que se pode aprender sobre a distribuição dos terramotos no espaço, no tempo e em magnitude? (2) Conseguimos encontrar modelos estatísticos que descrevam de forma fidedigna a distribuição dos abalos sísmicos? (3) E podem esses modelos ser usados para prever futuros sismos? A EVT consegue responder parcialmente às questões postas anteriormente (veja-se Pisarenko and Sornette, 2003, Beirlant *et al.*, 2004, 2016b, 2019, e Gomes *et al.*, 2013 e Gomes & Pestana, 2019, entre outros autores). O grande desafio para sismologistas e estatísticos consiste exatamente em estimar de forma precisa quão frequentes e de que magnitude esses grandes abalos sísmicos podem ser.

2.2. Extremos no mercado financeiro

O Comité de Basileia sobre controlo bancário formula normas e directrizes de supervisão e recomenda boas práticas para as instituições financeiras. Entre outras medidas de risco, essa regulamentação envolve a estimação do chamado VaR, que não é mais do que um quantil extremo da distribuição de perdas e ganhos. E existem contribuições altamente positivas da EVT, que entram em linha de conta com as caudas pesadas (EVI > 0) dos log-retornos diários (em percentagem), i.e. de $R_t = 100 \log (P_t/P_{t-1})$, com P_t o preço de fecho no dia t (veja-se Embrechts *et al.*, 1997, entre outros).

2.3. Extremos em Biologia e Medicina

Em análise de sobrevivência estamos interessados em X = tempo de sobrevivência do doente após a detecção de determinada doença maligna, e possivelmente na estimação de um limite superior do suporte de X (EVI ≤ 0). Mas, com Z = tempo até morte ou tempo de duração do estudo, temos usualmente dados sujeitos a *censura aleatória*, i.e. temos de admitir a existência de uma variável não-observada, Y, sendo observados valores de $Z = \min(X, Y)$ e de $\delta = \mathrm{I}_{[X \le Y]}$, onde a variável indicatriz, δ, determina se X foi sujeita a censura. A estimação de parâmetros de

acontecimentos extremos relativos a X depende pois da estimação de parâmetros equivalentes de Z e de metodologia inovadora, com inúmeros desenvolvimentos recentes (veja-se, entre outros, Beirlant *et al.*, 2007, 2010, 2016a, Einmahl *et al.*, 2008, Gomes and Neves, 2010, 2011, Ndao *et al.*, 2014, 2016, Worms & Worms, 2014, 2018, Stupfler, 2016, 2019).

2.4. Porquê EVT?

À questão, 'Porquê EVT?' podemos responder: a EVT é necessária porque nem tudo é normal! Muitas questões da vida real requerem estimação relativa a acontecimentos acerca dos quais os dados são inexistentes ou se existem são escassos, mesmo quando se trabalha com *big data* – são os designados *acontecimentos extremos ou raros.*

A resposta à questão, *Existirá um padrão escondido subjacente a este tipo de acontecimentos extremos?*, é positiva, e será parcialmente referida mais adiante. Na realidade os acontecimentos extremos, embora improváveis por hipótese, são mais frequentes do que seria de esperar segundo o modelo gaussiano ou normal, com caudas leves, de tipo exponencial quadrática. Independentemente da forma do centro da distribuição, a *cauda assume formas sempre muito especiais*, desde que estejamos suficientemente longe nessa cauda. A EVT é um ramo probabilístico de suporte à *Estatística* que lida exatamente com tais situações, ajudando a descrever e a quantificar os ditos acontecimentos raros, *extrapolando para além da amostra.*

Na análise de dados clássica os extremos podem vir a ser rotulados de *outliers*, chegando por vezes mesmo a ser ignorados no estudo, uma vez que se afastam do modelo 'ajustado'. Se o objectivo for inferir acerca de acontecimentos do dia-a-dia, poderá ser irrelevante suprimir tais dados das pontas, mas se a questão fulcral residir em eventos que não ocorrem com muita frequência, dever-se-á aplicar o contexto EVT, dando relevância exactamente a esses valores extremos.

Apresentamos, na Figura seguinte, seis dos estatísticos pioneiros na área de extremos.

Figura 6
Sir Ronald **Fisher** (1890-1962, *cima à esquerda*), Leonard **Tippett** (1902-1985, *cima ao centro*), Richard **von Mises** (1883-1953, *cima à direita*), Ernst Waloddi **Weibull** (1887-1979, *baixo à esquerda*), Emil **Gumbel** (1891-1966, *baixo ao centro*), Maurice **Fréchet** (1878-1973, *baixo à direita*)

Referimos em seguida algumas das frases célebres de Emil Gumbel, um dos nomes sonantes e pioneiros na área de *estatística de extremos*: "Il est impossible que l'improbable n'arrive jamais; Il y aura toujours une valeur qui dépassera toutes les autres; It seems that the rivers know the theory. It only remains to convince the engineers of the validity of this analysis." E a esta última frase, atrevemo-nos a acrescentar: "Não só os rios, mas também os movimentos da crosta terrestre, os mercados financeiros, a biologia, a medicina, e assim por diante, conhecem a teoria de valores extremos". Em Coles (2001), Beirlant *et al.* (2004), Castillo *et al.* (2004) e Gomes *et al.* (2013), entre outros, são tratados diversos estudos de casos, num leque variado de áreas de aplicação da *modelação de acontecimentos raros* (MAR).

2.5. Uma breve referência aos principais modelos assintóticos em EVT

Alguns dos resultados chave que têm levado à 'explosão' da componente estatística da EVT nas últimas décadas são o resultado obtido por Fréchet (1927), sobre a equação funcional de estabilidade para máximos, resolvido mais tarde, com algumas restrições, por Fisher & Tippett (1928), que derivaram as possíveis leis limites da sucessão de máximos normalizados, $(X_{n:n} - b_n) / a_n$, associados a uma amostra aleatória simples, $(X_1,\ldots, X_n)$, proveniente de uma *função de distribuição cumulativa* (FDC), F. Essas leis limites foram formalizadas inicialmente por Gnedenko (1943), e mais tarde por de Haan (1970), e usadas por Gumbel (1958; 2004) para aplicações da EVT em climatologia, engenharia e hidrologia.

As principais questões a ter em consideração são essencialmente as seguintes: existem usualmente poucas observações na cauda da distribuição; são requeridas estimativas muito para além do máximo observado; necessitamos de recorrer a modelos para a cauda, usualmente baseados em resultados assintóticos. Será sensato usar esses modelos em todas as situações reais envolvendo acontecimentos raros? É preciso não esquecer, parafraseando George Box, …*all models are wrong but some models are useful* (Box & Draper, 1987, p. 424).

A *estatística de extremos* baseia-se essencialmente no teorema de Gnedenko (Gnedenko, 1943), o chamado *teorema de tipos extremais*, um dos resultados limite fundamentais em EVT, que em linhas muito gerais permite identificar a distribuição de máximos, linearmente normalizados, com a chamada *lei geral de valores extremos* (GEV, do inglês *general extreme value*). Trata-se da também chamada lei *max-estável* (MS, do inglês *max-stable*), definida como uma lei para a qual é válida a equação funcional $\mathrm{MS}^n(\alpha_n x + \beta_n) = \mathrm{MS}(x)$, $n \geq 1$, $\alpha_n > 0$, $\beta_n \in \mathbb{R}$. Essa lei é do tipo de:

$$\mathrm{MS}_\xi(x) \equiv \mathrm{GEV}_\xi(x) = \begin{cases} \exp(-(1+\xi x)^{-1/\xi}), & \text{se } \xi \neq 0, \\ \exp(-\exp(-(x))), & \text{se } \xi = 0. \end{cases} \tag{2.1}$$

Contrariamente ao modelo normal ou gaussiano, muito frequente em estatística clássica, devido ao *teorema limite central* (TLC) para somas, o modelo MS adapta-se de forma bastante fidedigna à cauda direita do modelo F subjacente aos dados. Na realidade, sempre que é possível encontrar $a_n > 0$ e $b_n \in \mathrm{R}$ tais $(X_{n:n} - b_n) / a_n$ converge para uma variável aleatória não degenerada, essa variável aleatória é necessariamente do tipo $\mathrm{MS}_\xi \equiv \mathrm{GEV}_\xi$, em (2.1). Dizemos então que a FDC F está no *domínio de atração para máximos* da GEV_ξ, e escrevemos $F \in D_M(\mathrm{GEV}_\xi)$ ou $F \in D_M(\mathrm{MS}_\xi)$.

O parâmetro ξ, o chamado EVI, já atrás referido, é o parâmetro fundamental em *Estatística de Extremos*.

- Se $\xi < 0$ (max-Weibull), a cauda é curta, $x_F := \{\inf x : F(x) \geq 0\} < \infty$.
- Se $\xi = 0$ (Gumbel), a cauda é exponencial, $x_F < \infty$ ou $x_F = \infty$.
- Se $\xi > 0$ (Fréchet), temos uma cauda pesada, de tipo-Pareto, e $x_F = \infty$.

Inicialmente surgiram 3 distribuições possíveis:

$$\begin{aligned} &\text{Tipo I:} \quad \Lambda(x) = e^{-e^{-x}},\ x \in \mathrm{R} \quad [\text{Gumbel}],\\ &\text{Tipo II:} \quad \Phi_\alpha(x) = e^{-x^{-\alpha}},\ x > 0,\ \alpha > 0 \quad [\text{Fréchet}],\\ &\text{Tipo III:} \quad \Psi_\alpha(x) = e^{-(-x)^{\alpha}},\ x < 0,\ \alpha > 0 \quad [\text{Max-Weibull}], \end{aligned}$$

também chamadas distribuições de *valores extremos* (EV), associadas respectivamente com $\xi = 0$, $\xi = 1/\alpha > 0$ e $\xi = -1/\alpha < 0$, que podem obviamente ser unificadas na $\mathrm{GEV}_\xi \equiv \mathrm{MS}_\xi$, definida em (2.1).

Na figura 7 ilustramos o comportamento dos três tipos de densidades de valores extremos, $g_\xi(x) = \mathrm{dEV}_\xi(x)/\mathrm{d}x$, comparativamente com a densidade normal, $\varphi(x) = \exp(-x^2/2)/\sqrt{2\pi}$, $x \in \mathbb{R}$.

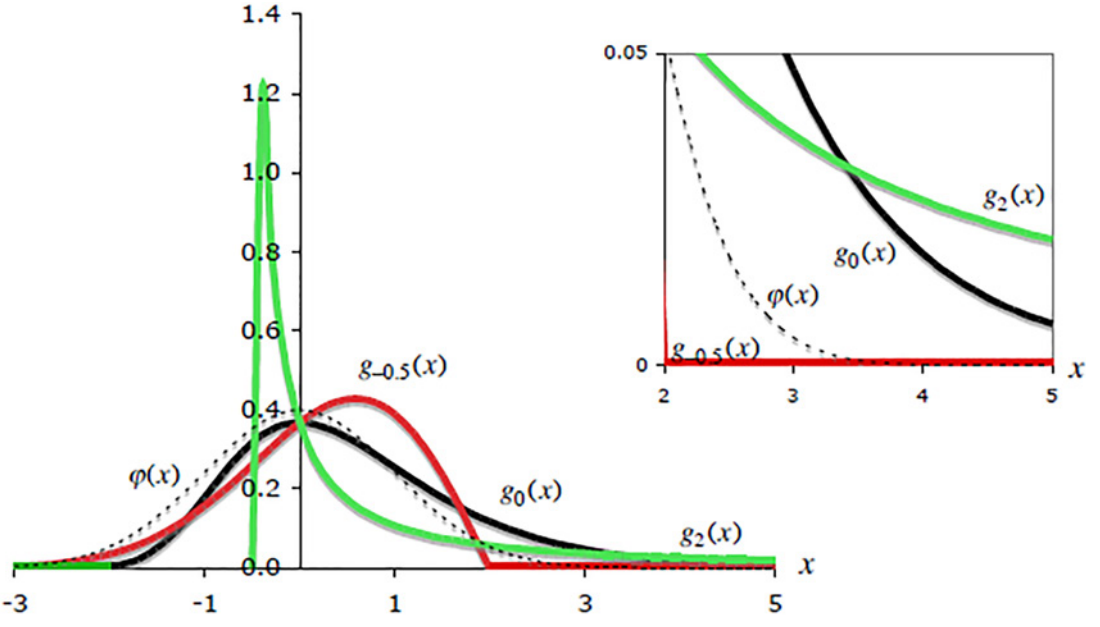

Figura 7
Densidades de valores extremos, g_ξ, $\xi = -0.5, 0, 2$, e normal, φ

Observação 2.1. Note-se que os resultados que aqui referimos para máximos podem ser de imediato convertidos em resultados para mínimos, pois $\min_{1\leq i\leq n} X_i = -\max_{1\leq i\leq n}(-X_i)$.

Mais geral do que a classe de modelos MS, podemos considerar a classe dos modelos *max-semi-estáveis* (MSS, do inglês *max-semi-stability*), introduzida em Grienvich (1992a, 1992b), Pancheva

(1992), e amplamente estudados em Canto e Castro *et al.* (2001) e em Temido & Canto e Castro (2003), com a forma funcional,

$$\mathrm{MSS}_{\xi,\nu}(x) = \begin{cases} \mathrm{e}^{-\nu\{\ln(1+\xi x)/\xi\}(1+\xi x)^{-1/\xi}}, \ 1+\xi x > 0 \ \text{ se } \ \xi \neq 0, \\ \mathrm{e}^{-\nu(x)\exp(-x)}, \ x \in \mathbb{R}, \qquad\qquad\qquad\quad \text{se } \ \xi = 0, \end{cases}$$

onde $v(\cdot)$ é uma função positiva, limitada e periódica, sendo $\mathrm{MS}_{\xi} = \mathrm{MSS}_{\xi,1}$. E vários autores têm detectado que os modelos MSS parecem ser mais interessantes do que os modelos MS para modelar algumas das variáveis relativas a tremores de terra (veja-se Sornette, 1998, entre outros artigos). Na realidade, e essencialmente através dos modelos MSS, mas também MS (veja-se Beirlant, *et al.*, 2016b, 2019, entre outros), a EVT permite-nos prever (pelo menos parcialmente, sendo este um desafio em que pensamos ainda ser necessário investir) qual a frequência e dimensão de sismos de relevo em determinada zona. Contudo, face a postularmos o modelo MSS, temos então dificuldades adicionais com a estimação dos parâmetros desconhecidos (veja-se Canto e Castro & Dias, 2011; Canto e Castro *et al.*, 2000, 2011, artigos dedicados ao desenvolvimento de métodos de estimação para os parâmetros de um modelo MSS). No nosso entender, o controlo de tremores de terra é de extrema dificuldade, e requer um esforço multidisciplinar, que pensamos não ter sido totalmente conseguido até à data, particularmente quando tentamos abordar o carácter espacial e temporal do processo subjacente a tremores de terra. Gostaria no entanto de referir uma tese de mestrado (Rosário, 2013), relacionada com dados também analisados em Beirlant *et al.* (2004, 2016b), de que usaremos alguns dos gráficos a apresentar mais adiante, e os trabalhos sobre o assunto já atrás referidos.

A *estatística de extremos* tem usualmente como base *resultados assintóticos* não só relacionados com o comportamento limite não-degenerado da sucessão de valores máximos, mas também de outras EO's de topo, quer individualmente quer em conjunto, e dos excessos acima de níveis elevados. Consequentemente, em *estatística de extremos* são fundamentais os modelos EV e GEV, que já referimos, os *processos extremais*, relacionados com a distribuição limite conjunta das k maiores (ou menores) EO's, e os modelos *generalizados de Pareto* (GP = 1+ln GEV). Na realidade, com $\mathrm{GP}_{\xi}(\cdot) = 1 + \ln \mathrm{GEV}_{\xi}(\cdot)$, e F_u a distribuição dos excessos acima de um nível elevado u, condicional a termos $X > u$, tem-se (Balkema & de Haan, 1974; Pickands, 1975),

$$F_u(y) := \mathbb{P}(X - u \leq y \mid X > u) \approx \mathrm{GP}_{\xi}(y/\sigma_u).$$

Para $\xi = 0$ obtemos o modelo Gumbel para máximos e um modelo exponencial para os excessos. O modelo exponencial, com função de cauda direira (ou de sobrevivência), $\overline{F}(x) := 1 - F(x) = \exp(-x/\sigma_u)$, $x > 0$, goza pois de um papel extraordinariamente importante em EVT, na chamada metodologia POT (do inglês, *peaks over threshold*), onde a escolha de u é fundamental.

As aplicações estatísticas da EVT têm dado ênfase à relaxação da condição de independência, à consideração de contextos multivariados e espaciais e a uma utilização cada vez mais aprofundada de abordagens relacionadas com variação regular e processos pontuais, de que não iremos falar. Para recensões críticas de muitos dos tópicos desta área, podemos ver alguns volumes das

revistas *Extremes* (13:2,3, 2010) e *Revstat* (10:1, 2012), entre outros. Por entre artigos de recensão crítica na área de SUE, mencionamos Gomes *et al.* (2008a), Neves & Fraga Alves (2008), Hüsler & Peng (2008), Beirlant *et al.* (2012), Scarrot & MacDonald (2012) e Gomes & Guillou (2015).

3. BREVE REFERÊNCIA À SUE

3.1. Parâmetros de acontecimentos extremos

O EVI é o parâmetro crucial em *estatística de extremos*. Em amostras dependentes, temos ainda o *índice extremal* (EI, do inglês, *extremal index*), θ, relacionado com a dimensão dos grupos de excedências de um nível elevado (Leadbbetter *et al.*, 1983). A influência do EI na estimação de outros parâmetros de acontecimentos extremos pode ser encontrada em Gomes (1993b), entre outros artigos na área. Para além destes parâmetros fundamentais, e com

$$U(t) := F^{\leftarrow}(1-1/t) = \inf\{x: F(x) \geq 1-1/t\}, t \in [1,\infty], \tag{3.1}$$

a função quantil de cauda associada ao modelo subjacente F, com inversa generalizada $F^{\leftarrow}$, referimos unicamente os *quantis extremais* ou VaR, i.e.

$$\mathrm{VaR}_q \equiv \chi_{1-q} := F^{\leftarrow}(1-q) = U(1/q), \tag{3.2}$$

para q pequeno (usualmente inferior a $1/n$, sendo n a dimensão da amostra), e o *período de retorno de um nível elevado u*, o número médio de excedências de u, dado por

$$T(u) = 1/(1-F(u)), \tag{3.3}$$

numa situação de observações independentes e identicamente distribuídas provenientes de uma FDC, F.

3.2. Abordagem clássica de Gumbel à Estatística de Extremos

A inferência estatística sobre acontecimentos extremos está directamente ligada a observações que são extremas em determinado sentido. As diferentes maneiras de definir esse tipo de observações levam a diferentes abordagens à SUE. Por vezes, só temos acesso a máximos anuais ou *máximos de blocos* (BM, do inglês *block maxima*). É então sensato usar o resultado limite principal em EVT, que fornece as leis MS (ou EV) como as únicas possíveis FDC's para máximos linearmente normalizados. Mais geralmente, se estamos interessados em cheias ($x_{n:n}$) ou em secas ($x_{1:n}$) de um rio num certo lugar, parece sensato trabalhar com máximos anuais ou mínimos, mesmo quando temos acesso a descargas diárias. Estamos então a usar a chamada *abordagem de Gumbel* ou o *método BM*.

Quando a dimensão da amostra $n \to \infty$, e devido ao resultado limite para a sucessão normalizada de valores máximos, i.e. ao teorema de Gnedenko, podemos escrever

$$\mathbb{P}[X_{n:n} \leq x] = F^n(x) \approx \mathrm{GEV}_{\xi}((x-\lambda_n)/\delta_n),$$

com a $GEV_{\xi}(x)$ a função GEV, em (2.1), e $(\lambda_n, \delta_n) \in (R, R^+)$ um vector de parâmetros desconhecidos de localização e escala, que substituem os chamados coeficientes de atracção, (b_n, a_n), na sucessão normalizada de valores máximos, $(X_{n:n} - b_n)/ a_n$. A aproximação anterior foi usada por Gumbel em vários artigos que culminaram no seu livro de 1958, para validar a utilização de distribuições do tipo EV (Fréchet, Gumbel e max-Weibull). Gumbel sugeriu pois o primeiro modelo em *estatística de extremos univariados*, frequentemente designado por *modelo dos máximos anuais*, ou *modelo* GEV *univariado*, ou ainda *modelo de* Gumbel.

De acordo com este modelo, dividimos os N dados, $(X_1,\ldots, X_N)$, em m sub-amostras (usualmente correspondentes a m anos) de dimensão n ($N = nm$) e ajustamos um dos modelos extremais Tipo I ou II ou III ou GEV à amostra formada pelos m máximos de cada sub-amostra, associados pois a $Y = \max(X_1,\ldots, X_n)$. Toda a inferência estatística se resume à inferência associada aos modelos em questão. Como a utilidade prática de uma FDC depende da existência de bons métodos para estimação dos seus parâmetros, e tal estimação não é (era?) sempre fácil para o modelo GEV, é usual, quando estamos face a uma amostra de máximos, tentar o ajustamento de uma das três distribuições, Gumbel (a mais simples), Fréchet ou max-Weibull, fazendo primeiro um teste de escolha estatística de um dos três modelos, que pode ser tão simples como um *quantil vs quantil* (QQ)-plot, de que falaremos mais adiante. Hoje em dia é frequente usar o método de máxima verosimilhança, implementado em vários 'R-packages', tais como evd, evdbayes, evir, evt0, extRemes, extremevalues, fExtremes, ismev, POT and SpatialExtremes, entre outros, que nos podem ajudar em muitos dos processos inferenciais relacionados com modelação de risco e extremos.

3.3. Outras abordagens

Embora a abordagem de Gumbel se tenha revelado muito útil nas mais diversas situações, várias críticas têm sido colocadas:

– Uma delas tem a ver com o facto de estarmos a perder informação quando usamos só o máximo observado e não outras EOs, caso tenhamos acesso a essas observações.
– Por outro lado, em muitas áreas de aplicação não existe sazonalidade dos dados, e o método de sub-amostras pode parecer algo artificial.

Para inferir sobre a cauda direita, parece na realidade mais sensato considerar um pequeno número k de EO's de topo associadas aos dados originais. Uma das abordagens paramétricas não-clássicas está então relacionada com a distribuição limite conjunta das k maiores observações, que nos conduz ao chamado *modelo extremal multivariada* (MEV, do inglês '*multivariate extreme value*') (Pickands, 1975; Weissman, 1978; Gomes, 1978, 1981). Uma segunda abordagem paramétrica não clássica é baseada na distribuição limite dos excessos de níveis elevados, determinísticos (e temos a chamada metodologia POT, introduzida em Smith, 1987, e em certo modo análoga ao modelo MEV) ou aleatórios, a que está associada a metodologia PORT, do inglês *peaks over random thresholds*, terminologia introduzida em Araújo Santos *et al.* (2006). Recentemente, estes métodos têm vindo a ser abordados sob um ponto de vista não-paramétrico. O tipo de ajustamento utilizado para as maiores observações não se identifica então com uma forma paramétrica dependente de

localização e escala, (λ, δ), e consideram-se abordagens semi paramétricas, baseadas nas k EO's de topo (veja-se de Haan & Ferreira, 2006, entre outros), admitindo unicamente que $F \in D_M(GEV_\xi)$.

3.4. Estudos de casos

Antes da análise de dois conjuntos de dados, procederemos a uma breve referência ao método gráfico mais usual de validação de um modelo F.

3.4.1. Como proceder à validação de um modelo de forma simples?

É óbvio que a *linearidade num gráfico* pode ser facilmente constatada por observação directa de uma núvem de pontos, e quantificada em termos do *coeficiente de correlação*. A ideia subjacente aos QQ-plots surgiu da necessidade de responder de forma simples à pergunta: *Será que um determinado modelo probabilístico fornece um ajustamento sensato à distribuição subjacente aos dados?*

Um QQ-plot tem como objetivo a *obtenção de uma confirmação visual rápida do ajustamento de um modelo probabilístico sugerido por exemplo pelo histograma, a dados* $(y_1, \ldots, y_n)$, *permitindo ainda a estimação grosseira de parâmetros*. Trata-se de um método de linearização da FDC, F, que postulamos como subjacente aos dados: face à amostra ordenada observada, $(y_{1:n} \leq \ldots \leq y_{n:n})$, e $F(y) = F((y-\lambda)/\delta)$, os pontos,

$$(F^{\leftarrow}(p_i), y_{i:n}),\ p_i := \frac{i}{n+1},\ 1 \leq i \leq n,$$

devem ser aproximadamente lineares, face à validade de F. Se o gráfico resultante mostrar que existe uma relação linear entre $y_{i:n}$ e $F^{\leftarrow}(p_i)$, para $1 \leq i \leq n$, temos uma validação informal de $F(\cdot)$. A intersecção com o eixo das abcissas e a inclinação da recta fornecem-nos então estimativas grosseiras do parâmetro de localização λ e do parâmetro de escala δ. A estimação dos parâmetros pode pois ser feita através do módulo de regressão de qualquer *package* estatístico.

Um exemplo simples, que veremos em seguida, é o de um QQ-plot Gumbel, um modelo muito usual em *estatística de extremos*.

Exemplo 3.1. Admitamos que pretendemos validar $F \equiv \Lambda$, com

$$\Lambda(y; \lambda, \delta) = e^{-e^{-(y-\lambda)/\delta}} =: p \Rightarrow y = \lambda + \delta(-\ln(-\ln(p))).$$

A geração de observações ou *números pseudo-aleatórios* (NPA's) Gumbel é simples $[-\ln(-\ln U)]$, com U denotando um NPA uniforme (0,1). Procedemos pois à geração de 250 NPA's Gumbel (0,1). O gráfico da Figura 8 (*esquerda*) é um QQ-plot normal, que fornece indicação imediata da não-normalidade dos dados.

O traçado da núvem de pontos:

$$(-\ln(-\ln(i/(n+1))), y_{i:n}),\ 1 \leq i \leq n,$$

forneceu o gráfico da Figura 8 (*direita*). A recta dos mínimos quadrados fornece-nos então as estimativas, $\lambda^{**} = -0.10$, $\delta^{**} = 0.99$.

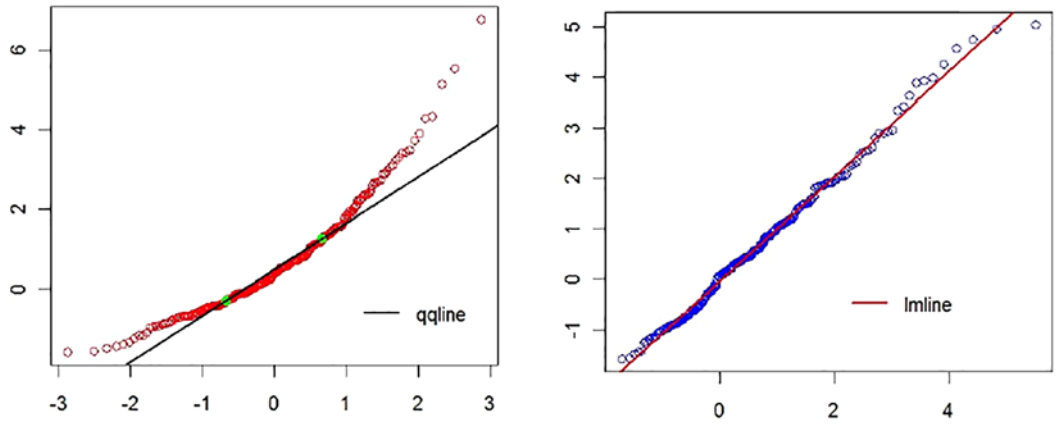

Figura 8
Dados Gumbel em QQ-plot normal (*esquerda*) e Gumbel (*direita*), com quantis teóricos em abcissa

3.4.2. Dados 'maasmax.txt'

Trata-se de um conjunto de dados, $Y_1,\ldots, Y_m$, de descargas anuais máximas do rio Meuse (Borgharen, NL), em m^3/s, no período 1911-1995, num total de $m = 85$ anos (Beirlant *et al.*, 2004; http://lstat.kuleuven.be/Wiley/). Tratam-se de réplicas da variável aleatória $Y \equiv M_n$, com M_n:= máximo anual = max $(X_1,\ldots, X_n)$, $n = 365$ (*dias*), e foram analisados com todo o detalhe em Gomes *et al.* (2013). Esses dados são apresentados na Figura 9, onde também apresentamos a caixa-com-bigodes associada.

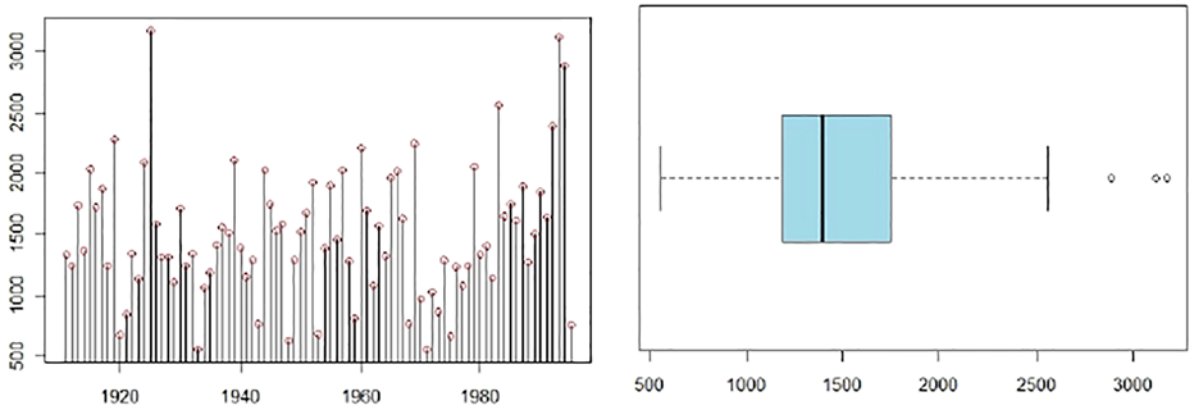

Figura 9
Máximos anuais do rio Meuse em 1911-1995 (*esquerda*) e caixa-com-bigodes associada (*direita*)

A caixa-com-bigodes exibe uma leve assimetria direita, mas vejamos o que se passaria se se seguisse uma 'análise estatística tradicional', fazendo um ajustamento do *máximo anual* ao modelo normal, $N(\mu, \sigma)$, para parâmetros convenientes. Na Figura 10 apresentamos o QQ-plot normal associado aos dados em análise: $\{(\Phi^{\leftarrow}(i/(m+1)), y_{i:m}): i = 1,\ldots, m\}$.

Apesar do mau ajustamento nas caudas, o modelo normal não é rejeitado pelos testes clássicos de ajustamento, aos níveis de significância usuais. Ajustando uma recta de mínimos quadrados,

obtém-se $\hat{\mu} = 1495.962$ e $\hat{\sigma} = 551.006$, a que corresponde um coeficiente de correlação amostral de $r_Q = 0.979$, bastante elevado. Sobrepondo ao histograma dos dados (máximos anuais) a normal estimada $Y \frown N\ (\hat{\mu} = 1495.962, \hat{\sigma} = 551.006)$, obtém-se o resultado da Figura 11.

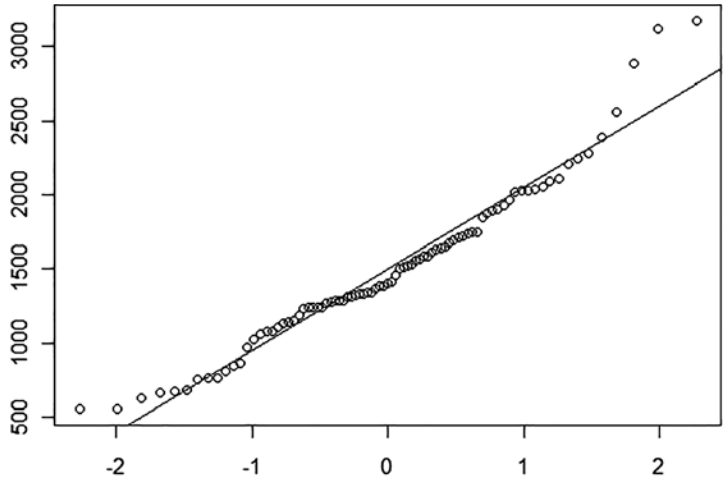

Figura 10
QQ-plot normal para os dados 'maasmax.txt'

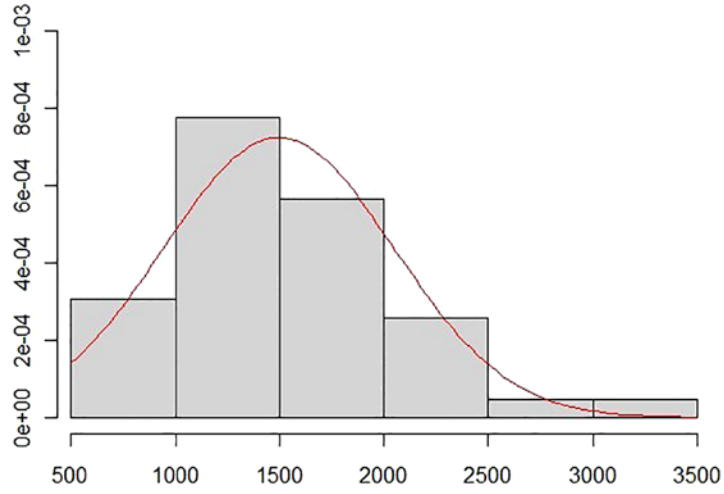

Figura 11
Histograma e normal ajustada (dados 'maasmax.txt')

Dada a natureza dos dados, a importância do modelo Gumbel em EVT, e ainda ao facto de se verificar uma assimetria à direita, vejamos o que se passaria se se considerasse um ajustamento do *máximo anual* ao modelo Gumbel, para parâmetros convenientemente estimados. Representamos na Figura 12 o QQ-plot Gumbel associado aos dados. Ajustando uma recta de mínimos quadrados, obtém-se para parâmetros de localização e escala, respectivamente, $\hat{\lambda} = 1247.363$ e $\hat{\delta} = 445.688$, a que corresponde um coeficiente de correlação amostral $r_Q = 0.992$ (superior ao encontrado no caso do ajustamento à normal). Note-se que, com γ a constant de Euler, o valor

estimado de (μ, σ) é $(\tilde{\mu}, \tilde{\sigma}) = (\hat{\lambda} + \gamma\,\hat{\delta}, \pi\hat{\delta}/\sqrt{6}) = (1504.621, 571.617)$, relativamente próximo do valor (1495.962, 551.006), obtido através de um ajustamento normal.

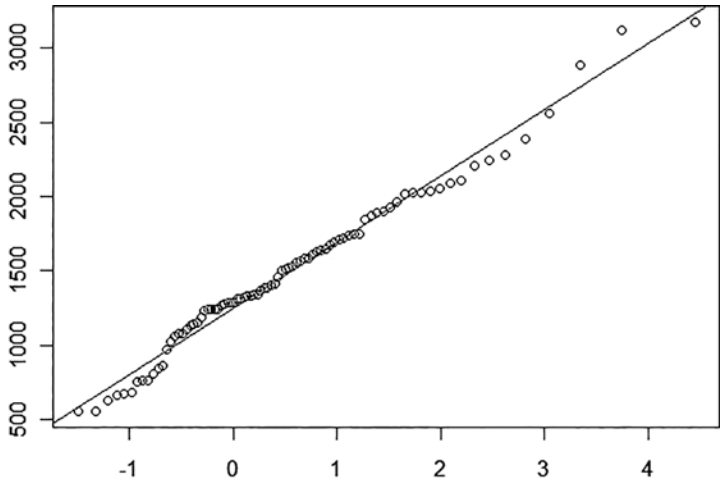

Figura 12
QQ-plot Gumbel: $\{(\Lambda^{\leftarrow}(i/(m+1)), y_{i:k}): i = 1,\ldots, m\}$

Pensemos agora na estimação do período de retorno do nível $y_T = 3175 = y_{85:85}$, o máximo das descargas ao longo dos 85 anos, i.e. o parâmetro definido em (3.3):

$$T = \frac{1}{\mathbb{P}[Y > y_T]} = \frac{1}{1 - F_Y(y_T)}, \quad \text{e} \quad \hat{T} = \frac{1}{1 - \Phi\left(\frac{3175-\hat{\mu}}{\hat{\sigma}}\right)} \simeq 866 \text{ anos.}$$

Consideremos em seguida a FDC Gumbel estimada, e pensemos na estimação do mesmo período de retorno do nível $y_T = 3175$. Obtemos então:

$$\hat{T} = \frac{1}{1 - \Lambda\left(\frac{3175-\hat{\lambda}}{\hat{\delta}}\right)} \simeq 76 \text{ anos,}$$

valor consideravelmente inferior a 866 anos, o valor encontrado no caso do tratamento com a normal.

Pensemos em seguida na estimação do nível de retorno a $T = 100$-anos, face ao modelo normal. Como:

$$U(T) = F_Y^{\leftarrow}\left(1 - 1/T\right),$$

o nível médio de descargas ultrapassado pelo máximo anual todos os $T = 100$ anos seria estimado por:

$$\widehat{U}(100) = \hat{\mu} + \hat{\sigma}\,\Phi^{\leftarrow}\left(1 - 1/100\right) = 2777.793.$$

Consideremos então a Gumbel estimada, e pensemos na estimação do mesmo nível de retorno a T = 100-anos. Este nível de retorno é estimado por:

$$\widehat{U}(100) = \hat{\lambda} + \hat{\delta}\,\Lambda^{\leftarrow}(1 - 1/100) = 3297.596,$$

acima do valor estimado face a um ajustamento ao modelo normal.

3.4.3. Momentos sísmicos

A base de dados utilizada para ilustração é constituída por medições de momentos sísmicos acima de exp(+23) dyne-cm. Temos então acesso ao registo de 8123 terramotos superficiais de duas zonas de forte intensidade sísmica, 6458 em zonas de subducção e 1665 em zonas de dorsais oceânicas. Estas regiões podem ser facilmente identificadas na Figura 13. Estes dados foram tratados por Pisarenko and Sornette (2003), e retirados do catálogo sísmico de Harvard, no período de 1 de Janeiro de 1977 a 31 de Maio de 2000. Foram ainda utilizados em Beirlant *et al.* (2004, 2016b) e em Rosário (2013).

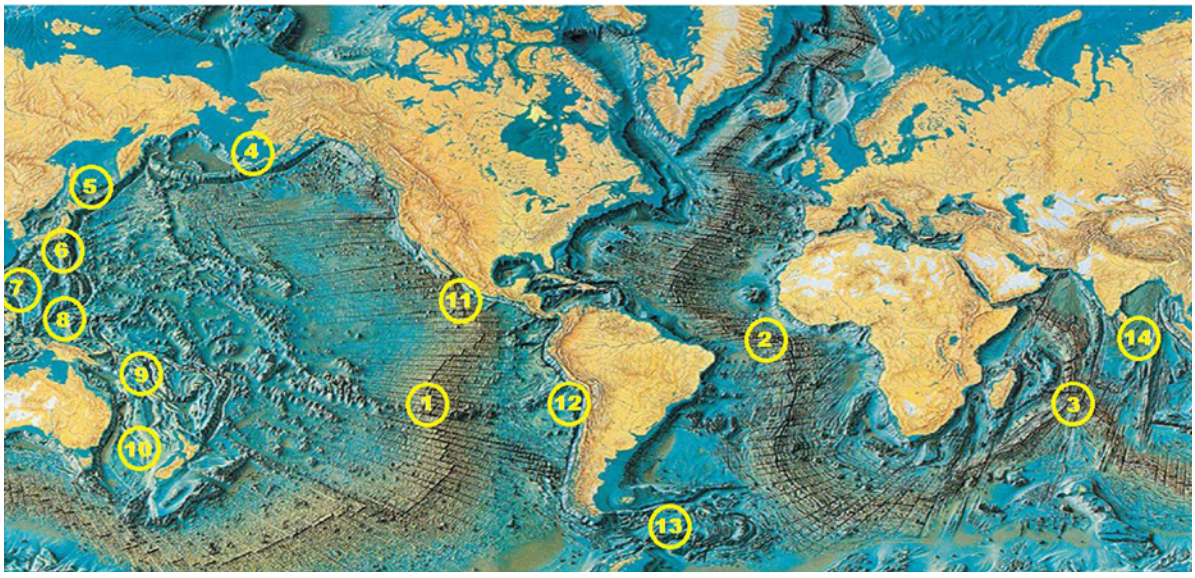

Figura 13
Principais zonas sísmicas: Zona dorsal oceânica (1–3): 1-Sudeste Pacífico, 2-Meso-Atlântica, 3-Sudeste Indiano; Zona de subducção (4–14): 4-Alasca, 5-Ilhas Curila, 6-Japão, 7-Tailândia, 8-Ilhas Mariana, 9-Ilhas Salomão, 10-Ilhas Tonga, 11-México, 12-América do Sul, 13-Ilhas Sandwich, 14-Sunda Arc

Os momentos sísmicos (dyne-cm) foram convertidos em magnitudes na escala de Richter. A assimetria direita dos dados é notória, e pode ser visualmente confirmada quer pela caixa-com-bigodes quer pelo histograma representados na Figura 14. A cauda direita aparenta no entanto ser leve ou do tipo exponencial ($\xi \leq 0$), pois não é demasiado extensa e as barras não decrescem de forma demasiado lenta.

Dada a natureza dos dados, a importância do modelo Gumbel em EVT, e ainda ao facto de se verificar mais uma vez uma assimetria à direita, vejamos o que se passaria se se considerasse um ajustamento ao modelo Gumbel, para parâmetros convenientemente estimados. Representamos

na Figura 15 o QQ-plot Gumbel associado aos dados. Ajustando uma recta de mínimos quadrados, obtém-se para parâmetros de localização e escala, respectivamente, $\tilde{\lambda} = 5.431$ e $\tilde{\delta} = 0.288$, a que corresponde um coeficiente de correlação amostral $r_Q = 0.998$. No caso das mesmas regiões de dorsais oceânicas, e se tentarmos o ajustamento de um modelo EV_ξ, obtém-se a estimativa $\tilde{\xi} = -0.013$ e um coeficiente de correlação amostral $r_Q = 0.999$, ligeiramente superior ao anteriormente obtido, e que não justifica enveredarmos pela estimação adicional de ξ.

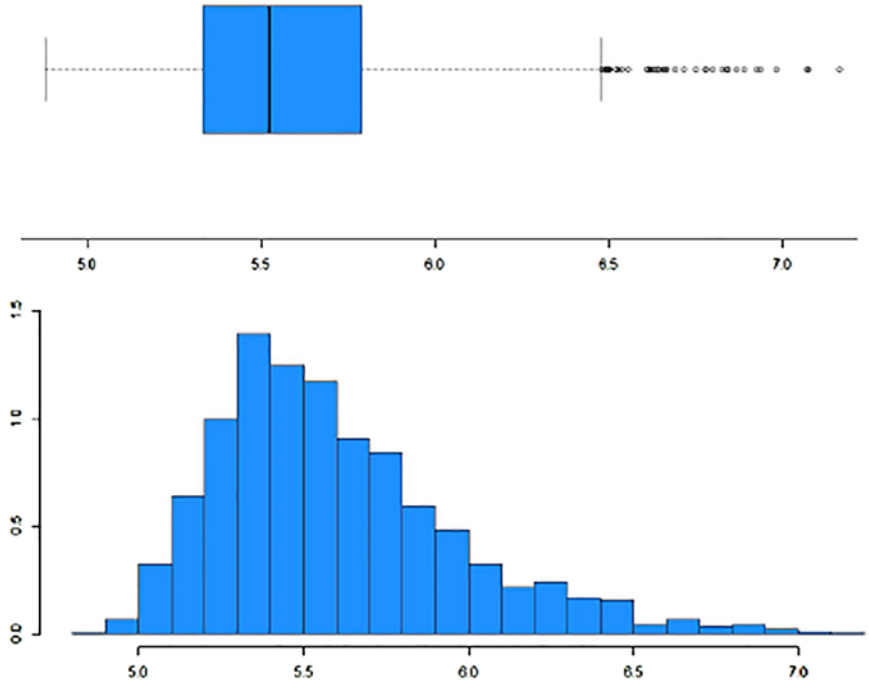

Figura 14
Caixa-com-bigodes (*cima*) e histograma (*baixo*) das magnitudes da amostra referente a sismos das zonas de dorsais oceânicas

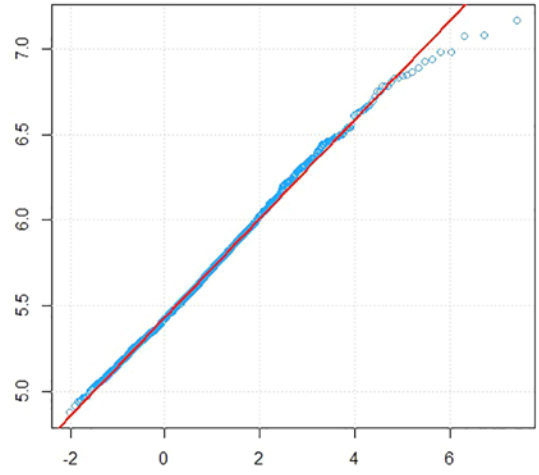

Figura 15
QQ-Plot Gumbel das magnitudes da amostra referente a sismos das zonas de dorsais oceânicas (quantis teóricos em abcissa)

Se tivessemos tentado o ajustamento de um modelo normal, e pensássemos na forma mais simples de estimar o período de retorno do nível $y_T = 7.163$, o máximo da amostra, teríamos:

$$\hat{\mu} = 5.597,\ \hat{\sigma} = 0.368 \quad \text{e } \hat{T} = \frac{1}{1 - \Phi\left(\frac{7.163 - \hat{\mu}}{\hat{\sigma}}\right)} \simeq 98167.8 \text{ anos.}$$

Se considerarmos a Gumbel estimada, e pensarmos na forma equivalente de estimação do período de retorno do nível $y_T = 7.163$, obtemos:

$$\hat{T} = 1/\left(1 - \Lambda\left(\frac{7.163 - \hat{\lambda}}{\hat{\delta}}\right)\right) \simeq 409.5 \text{ anos,}$$

valor este consideravelmente inferior ao encontrado no caso do tratamento com a normal, e que amplamente justifica a utilização de um modelo de extremos.

4. BREVE REFERÊNCIA À SUE SEMI-PARAMÉTRICA

Recentemente, os métodos BM, MEV e POT têm sido considerados em ambiente semi-paramétrico. Não há então qualquer ajustamento de um modelo paramétrico adequado. Admitimos apenas que $F \in D_M(\text{GEV}_\xi)$, sendo ξ o único parâmetro de acontecimentos extremos a ser inicialmente estimado, com base em algumas observações de topo e de acordo com metodologia adequada. É usual considerar as k observações de topo, $X_{n:n} \geq \ldots \geq X_{n-k+1:n}$, acima de um *nível aleatório* $X_{n-k:n}$, que necessita de ser uma EO *intermedia superior*, i.e.

$$k = k_n \to \infty,\ k \in [1, n),\ k = o(n) \text{ quando } n \to \infty. \tag{4.1}$$

Como os riscos são mais elevados quando estamos com uma cauda direita pesada, consideraremos unicamente modelos de *cauda pesada*, i.e. FDC's de tipo-*Pareto*, com um EVI positivo, trabalhando pois em $D_M + := D_M(\text{GEV}_{\xi>0})$.

4.1. Estimadores GM do EVI

Por entre a grande variedade de estimadores do EVI, mencionamos apenas o estimador de Hill (H), introduzido em Hill (1975), a média dos excessos das log-observações,

$$V_{ik} := \ln X_{n-i+1:n} - \ln X_{n-k:n}, \qquad 1 \leq i \leq k < n,$$

i.e.

$$\text{H}(k) = \text{H}(k; \underline{\mathbf{X}}_n) := \frac{1}{k}\sum_{i=1}^{k} V_{ik}, \qquad 1 \leq k < n, \tag{4.2}$$

e uma generalização competitiva de H(k), introduzida recentemente na literatura.

4.1.1. Estimador média-de-ordem p (H_p) do EVI

Note-se que podemos escrever:

$$\mathrm{H}(k) = \sum_{i=1}^{k} \ln\left(\frac{X_{n-i+1:n}}{X_{n-k:n}}\right)^{1/k} = \ln\left(\prod_{i=1}^{k} \frac{X_{n-i+1:n}}{X_{n-k:n}}\right)^{1/k}.$$

O estimador H, em (4.2), é pois o logaritmo da *média geométrica* (ou *média-de- ordem*-0) de

$$U_{ik} := \frac{X_{n-i+1:n}}{X_{n-k:n}}, \quad 1 \le i \le k < n.$$

Brilhante *et al.* (2013), e quasi simultaneamente Paulauskas & Vaičiulis (2013), e Beran *et al.* (2014), consideraram como estatísticas básicas a *media-de-ordem-p* de U_{ik}, $1 \le i \le k$, para $p \ge 0$. Mais geralmente, Gomes & Caeiro (2014) e também Caeiro *et al.* (2016) consideraram essas mesmas estatísticas para $p \in \mathbb{R}$, i.e.

$$\mathrm{M}_p(k) = \begin{cases} \left(\frac{1}{k}\sum_{i=1}^{k} U_{ik}^{p}\right)^{1/p}, & \text{se } p \neq 0, \\ \left(\prod_{i=1}^{k} U_{ik}\right)^{1/k}, & \text{se } p = 0, \end{cases}$$

e a classe associada de estimadores H_p do EVI:

$$\mathrm{H}_p(k) = \mathrm{H}_p(k; \underline{\mathbf{X}}_n) := \begin{cases} \left(1 - \mathrm{M}_p^{-p}(k)\right)/p, & \text{se } p < 1/\xi,\ p \neq 0, \\ \ln \mathrm{M}_0(k) = \mathrm{H}(k), & \text{se } p = 0. \end{cases} \tag{4.3}$$

E outras generalizações recentes, que não vamos referir, têm-se revelado altamente competitivas.

4.2. Estimadores PORT do EVI

Os estimadores do EVI anteriormente mencionados dependem do parâmetro de *controlo p* ∈ R, são altamente flexíveis, mas, tal como frequentemente desejado, não são invariantes a mudanças na localização, dependendo fortemente de possíveis alterações da localização do modelo subjacente aos dados, contrariamente ao que acontece com o EVI, que é independente dessa localização. É pois sensato sugerir a utilização da classe de estimadores PORT-H_p do EVI, introduzida em Gomes *et al.* (2016). Esses estimadores são semelhantes aos estimadores PORT-H do EVI, estudados em Araújo Santos *et al.* (2006), e também considerados em Gomes *et al.* (2008b).

As classes de estimadores PORT são baseadas numa *amostra de excessos* acima de um nível aleatório $X_{n_s:n}$, $n_s := \lfloor ns \rfloor + 1$, $0 \le s < 1$,

$$\underline{\mathbf{X}}_n^{(s)} := \left(X_{n:n} - X_{\lfloor ns \rfloor+1:n}, \ldots, X_{\lfloor ns \rfloor+2:n} - X_{\lfloor ns \rfloor+1:n}\right). \tag{4.4}$$

Para $0 \le s < 1$ e $k < n - n_s$, os estimadores PORT-H_p do EVI têm a mesma forma funcional de H_p, em (4.3), mas com $\underline{\mathbf{X}}_n = (X_1, \ldots, X_n)$ substituída pela amostra dos excessos $\underline{\mathbf{X}}_n^{(s)}$, em (4.4). São pois dados por

$$\mathrm{H}_p(k, s) := \mathrm{H}_p\big(k; \underline{\mathbf{X}}_n^{(s)}\big). \tag{4.5}$$

4.3. Estimação semi-paramétrica de outros parâmetros

Apesar da estimação de *períodos de retorno de níveis elevados, probabilidades de execedência e coeficiente de dependência na cauda*, entre outros parâmetros de acontecimentos extremos, serem tão importantes na modelação de risco como a estimação do EVI, iremos unicamente referir brevemente a estimação semiparamétrica de um quantil elevado, VaR_q, já definido em (3.2), que pode ser facilmente obtida através da aproximação $U(tx) \approx U(t)x^{\xi}$, com $U(\cdot)$ definida em (3.1). Tal como se fez para a estimação do EVI, vamos basear essa estimação nas k EO's superiores, assumindo que k é uma sucessão *intermédia*, i.e. que se tem a validade de (4.1).

Sendo $\hat{\xi}$ qualquer estimador consistente do EVI, o estimador semi-paramétrico mais simples de VaR_q é dado por

$$Q_{\hat{\xi}}^{(q)}(k) := X_{n-k:n}\,(k/(nq))^{\hat{\xi}} \qquad \text{(Weissman, 1978)}. \tag{4.6}$$

Para modelos de cauda pesada, os estimadores 'clássicos' do EVI, i.e. os estimadores H, definidos em (4.2), são os mais frequentemente usados em (4.6), de modo a obtermos os estimadores 'clássicos' do VaR, para os quais se usa a notação óbvia, $Q_{\mathrm{H}}^{(q)}(k)$. Sugerimos agora a substituição de $\hat{\xi}$ por $\mathrm{H}_p(k)$ em (4.3), para o qual usamos a notação $Q_{\mathrm{H}p}^{(q)}(k)$ (veja-se Gomes *et al.*, 2015, para detalhes sobre estes estimadores).

4.3.1. Estimação PORT-H_p do VaR

Muitos dos estimadores semi-paramétricos do VaR, tais como os estimadores H_p do VaR em Gomes *et al.* (2015) (veja-se também os livros de Beirlant *et al.*, 2004, e de Haan and Ferreira, 2006), não gozam do comportamento adequado face a transformações lineares dos dados, um comportamento relacionado com o facto de se ter para qualquer quantil elevado, VaR_q,

$$\mathrm{VaR}_q(\lambda + \delta X) = \lambda + \delta \mathrm{VaR}_q(X),$$

para qualquer modelo X, λ real, e δ real positivo. Para $\xi > 0$, Araújo Santos *et al.* (2006) avançaram com estimadores do VaR que têm a propriedade linear atrás referida, e que são baseados na *amostra de excessos*, $\underline{\mathbf{X}}_n^{(s)}$, $0 \le s < 1$, em (4.4). Tais estimadores foram denominados PORT-VaR, e têm como base os estimadores PORT-H do EVI, $\mathrm{H}(k; \underline{\mathbf{X}}_n^{(s)})$, $k < n - n_s$. Agora, sugerimos para uma estimação adequada do VaR, a consideração dos estimadores PORT-H_p do VaR, definidos por

$$\widehat{\mathrm{VaR}}_q(k; p, s) := (X_{n-k:n} - X_{n_s:n}) \left(\frac{k}{nq}\right)^{\mathrm{H}_p(k,s)} + X_{n_s:n},$$

com $H_p(k; s)$ definidos em (4.5). As simulações de Monte-Carlo em Figueiredo *et al.* (2017) mostram a elevada potencialidade dos estimadores PORT-H_p de VaR_q.

Note-se que a metodologia PORT não provoca qualquer mudança na variância assintótica, um ponto a favor destes estimadores, que podem ter viés assintótico nulo.

5. COMENTÁRIOS FINAIS

- Muito mais haveria a dizer sobre o papel da EVT na modelação de acontecimentos raros e risco.
- Referimos essencialmente uma das abordagens paramétricas à estatística de extremos univariados. Mas devemos encarar as metodologias paramétricas e semi-paramétricas não como concorrentes, mas sim como complementares, ambas com desafios variados.
- O caso da não independência da amostra, que tem recentemente merecido grande destaque, não foi sequer aflorado, sendo um tema com muitos desafios.
- Nesta introdução ao estudo de valores extremos, focámos a nossa atenção no caso univariado, mas como facilmente se antevê a EVT nos campos multivariado e/ou espacial tem igualmente relevância para a *modelação de acontecimentos raros.*
- Mas a própria SUE (*estatística de extremos univariados*) é ainda um tópico de grande importância na modelação de risco.
- E em ambiente semi-paramétrico, tópicos como a seleção do nível elevado (veja-se Caeiro & Gomes, 2015) e a metodologia PORT (veja-se Gomes *et al.*, 2016) suscitam também grandes desafios.
- Por outro lado, a falta de eficiência dos estimadores H_p para $p < 0$, em conjugação com os resultados em Stehlík *et al.* (2010), relacionados com a robustez dos estimadores H_{-1} do EVI, alerta-nos para a necessidade de discussão adicional do tópico *robustez vs efficiência*.
- As análises de risco associadas a acontecimentos extremos requerem a experiência combinada e multidisciplinar de estatísticos e de especialistas em climatologia, hidrologia, finanças, seguros, medicina, desporto, e outras áreas.
- Esperamos ter aguçado o vosso apetite por um tema relativamente recente em termos históricos, e com tantas áreas de aplicação quantas as que possamos conceber.

(Comunicação apresentada à Classe de Ciências
na sessão de 16 de maio de 2019)

REFERÊNCIAS

[1] Araújo Santos, P., Fraga Alves, M.I. & Gomes, M.I. (2006). Peaks over random threshold methododlogy for tail index and high quantile estimation. *Revstat* **4**:3, 227–247.

[2] Arnold, B., Balakhrishna, N. & Nagaraja, H. N. (1992; 2008). *A First Course in Order Statistics.* 1st Ed., Wiley; 2nd Ed., SIAM.

[3] Balkema, A.A. & de Haan, L. (1974). Residual life time at great age. *Ann. Probab.* **2**, 792–804.

[4] Beirlant, J., Goegebeur, Y., Segers, J. & Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley, England.

[5] Beirlant, J., Guillou, A., Dierckx, G. & Fils-Villetard, A. (2007). Estimation of the extreme value index and extreme quantiles under random censoring. *Extremes* **10**, 151–174.

[6] Beirlant, J., Guillou, A. & Toulemonde, G. (2010). Peaks-Over-Threshold modeling under random censoring. *Communications in Statistics–Theory and Methods* **39**, 1158–1179.

[7] Beirlant, J., Caeiro, F. & Gomes, M.I. (2012). An overview and open researh topics in statistics of univariate extremes. *Revstat* **10**:1, 1–31.

[8] Beirlant, J., Bardoutsos, A., de Wet, T. & Gijbels, I. (2016a). Bias reduced tail estimation for censored Pareto type distributions, *Statistics and Probability Letters* **109**, 78–88.

[9] Beirlant, J., Fraga Alves, M.I. & Gomes, M.I. (2016b). Tail fitting for truncated and non-truncated Pareto-type distributions, *Extremes* **19**:3, 429–462.

[10] Beirlant, J., Kijko, A.J., Reynkens, T. & Einmahl, J.H.J. (2019). Estimating the maximum possible earthquake magnitude using extreme value methodology: the Groningen case. *Natural Hazards* **98**, 1091-1113.

[11] Beran, J., Schell, D. & Stehlik, M. (2014). The harmonic moment tail index estimator: asymptotic distribution and robustness. *Ann. Inst. Statist. Math.* **66**, 193–220.

[12] Box, G.E.P. & Draper, N.R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley.

[13] Brilhante, M.F., Gomes, M.I. & Pestana, D. (2013). A simple generalisation of the Hill estimator. *Comput. Statistics and Data Analysis* **57**:1, 518–535.

[14] Caeiro, F. and Gomes, M.I. (2015). Threshold selection in extreme value analysis. In Dipak Dey & Jun Yan, *Extreme Value Modeling and Risk Analysis: Methods and Applications*, Chapman-Hall/CRC, Chapter 4, 69–87.

[15] Caeiro, F., Gomes, M.I., Beirlant, J. & de Wet, T. (2016). Mean-of-order-*p* reduced-bias extreme value index estimation under a third-order framework. *Extremes* **19**:4, 561–589.

[16] Canto e Castro, L. & Dias, S. (2011). Generalized Pickands' estimators for the tail index parameter and max-semistability. *Extremes* **14**:4, 429–449.

[17] Canto e Castro, L., Temido, G. & Gomes, M.I. (2000). Inferência estatística em modelos max-semiestáveis. Em P. Oliveira & E. Athayde eds., *Um Olhar sobre a Estatística*, 291–305, Edição S.P.E.

[18] Canto e Castro, L., Haan, L. de & Temido, M.G. (2001). Rarely observed maxima, *Th. Prob. Appl.* **45**, 658–662.

[19] Canto e Castro, L., Dias, S. & Temido, M.G. (2011). Looking for maxsemistability: a new test for the extreme value condition. *Journal of Statistical Planning and Inference* **141**, 3005–3020.

[20] Castillo, E., Hadi, A., Balakrishnan, N. & Sarabia, J.M. (2005). *Extreme Value and Related Models with Applications in Engineering and Science*. Wiley, Hoboke, New Jersey.

[21] Coles S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag.

[22] Einmahl, J.H.J., Fils-Villetard, A. & Guillou, A. (2008). Statistics of extremes under random censoring. *Bernoulli* **14**(1), 207–227.

[23] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin, Heidelberg.

[24] Figueiredo, F., Gomes, M.I., & Henriques-Rodrigues, L. (2017). Value-at-risk estimation and the PORT mean-of-order-p methodology. *Revstat* **15**:2, 187–204.

[25] Fisher, R.A. & Tippett, L.H.C. (1928). Limiting forms of the frequency distributions of the largest or smallest member of a sample. *Proceedings Cambridge Philosophical Society* **24**, 180-190.

[26] Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Polon. Math.* **6**, 93–116.

[27] Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.* **44**, 423-453.

[28] Gomes, M.I. (1978). *Some Probabilistic and Statistical Problems in Extreme Value Theory*. Ph. D. Thesis, Univ. Sheffield.

[29] Gomes, M.I. (1981). An i-dimensional limiting distribution function of largest values and its relevance to the statistical theory of extremes. In C. Taillie *et al.* (eds.), *Statistical Distributions in Scientific Work*, Vol. **6**, 389–410, D. Reidel.

[30] Gomes, M.I. (1993a). A obra científica de J. Tiago de Oliveira. In D. Pestana (ed.), *Estatística Robusta, Extremos e Mais Alguns Temas*, 241–248, Edições Salamandra.

[31] Gomes, M.I. (1993b). On the estimation of parameters of rare events in environmental time series. In V. Barnett and K.F. Turkman (eds.). *Statistics for the Environment*, Wiley, New York, 225–241.

[32] Gomes, M.I. (1994). J. Tiago de Oliveira: Obituary. *J. Royal Statist. Soc. A* **157**, 499–500.

[33] Gomes, M.I. & Caeiro, F. (2014). Efficiency of partially reduced-bias mean-of- order-*p* versus minimum-variance reduced--bias extreme value index estimation. In M. Gilli *et al.* (eds.). *Proceedings of* COMPSTAT 2014, ISI/IASC, 289–298.

[34] Gomes, M.I. & Guillou, A. (2015). Extreme value theory and statistics of univariate extremes: A review. *International Statistical Review* **83**:2, 263–292.

[35] Gomes, M.I. & Neves, M.M. (2010). A note on statistics of extremes for censoring schemes on a heavy right tail. In Luzar--Stifller, V. *et al.* (eds.), *Proceedings of* ITI 2010, SRCE Univ. Computing Centre Editions, 539–544.

[36] Gomes, M.I. & Neves, M.M. (2011). Estimation of the extreme value index for randomly censored data. *Biometrical Letters* **48**, 1–22.

[37] Gomes, M.I. & Pestana D. (2019). Estatística de Extremos: Um instrumento para predição de tremores de terra? In Memórias da Academia das Ciências de Lisboa, Classe de Ciências, Tomo XLVI, pp. 305-312. "Sessão: À Conversa sobre o Terramoto de 1755", Academia das Ciências de Lisboa, November 2, 2016, Lisboa, Portugal.

[38] Gomes, M.I., Canto e Castro, L., Fraga Alves, M.I. & Pestana, D. (2008a). Statistics of extremes for iid data and breakthroughs in the estimation of the extreme value index: Laurens de Haan leading contributions. *Extremes* **11**:1, 3–34.

[39] Gomes, M.I., Fraga Alves, M.I. & Araújo Santos, P. (2008b). PORT Hill and moment estimators for heavy-tailed models. *Commun. Statist. – Simul. and Comput.* **37**, 1281-1306.

[40] Gomes, M.I., Fraga Alves, M.I. & Neves, C. (2013). *Análise de Valores Extremos: uma Introdução*. Edições SPE & INE.

[41] Gomes, M.I., Brilhante, F., & Pestana, D. (2015). A mean-of-order-*p* class of value-at-risk estimators. In C. Kitsos *et al.* (eds.), *Theory and Practice of Risk Assessment*, Springer Proceedings in Mathematics and Statistics 136, Springer International Publishing, Switzerland, pp. 305–320.

[42] Gomes, M.I., Henriques-Rodrigues, L. & Manjunath, B.L. (2016). Mean-of-order-*p* location-invariant extreme value index estimation. *Revstat* **14**:3, 273–296.

[43] Grienvich, I.V. (1992a). Max-semistable laws corresponding to linear and power normalizations. *Th. Probab. Appl.* **37**, 720–721.

[44] Grienvich, I.V. (1992b). Domains of attraction of max-semistable laws under linear and power normalizations. *Th. Probab. Appl.* **38**, 640–650.

[45] Gumbel, E.J. (1958; 2004). *Statistics of Extremes*. Columbia University Press, New York.

[46] de Haan, L. (1970). *On Regular Variation and its Application to the Weak Convergence of Sample Extremes*. Mathematical Centre Tract 32, Amsterdam.

[47] de Haan, L. & Ferreira, A. (2006). *Extreme Value Theory: an Introduction*. Springer-Verlag.

[48] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3**, 1163-1174.

[49] Hüsler, J. & Peng, L. (2008). Review of testing issues in extremes: in honor of Professor Laurens de Haan. *Extremes* **11**, 99–111.

[50] Leadbetter, R., Lindgren, G. & Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer Series of Statistics.

[51] Markovich, N. (2007). *Nonparametric Analysis of Univariate Heavy-tailed Data*, John Wiley & Sons, England.

[52] Ndao, P., Diop, A. & Dupuy, J.-F. (2014). Nonparametric estimation of the conditional tail index and extreme quantiles under random censoring. *Computational Statistics and Data Analysis* **79**, 63–79.

[53] Ndao, P., Diop, A. & Dupuy, J.-F. (2016). Nonparametric estimation of the conditional extreme-value index with random covariates and censoring. *J. Statistical Planning and Inference* **168**, 20–37.

[54] Neves, C. & Fraga Alves, M.I. (2008). Testing extreme value conditions – an overview and recent approaches. *Revstat* **6**, 1, 83–100.

[55] Pancheva, E. (1992). Multivariate max-semistable distributions, *Th. Probab. and Appl.* **37**, 731–732.

[56] Paulauskas, V. & Vaičiulis, M. (2013). On the improvement of Hill and some others estimators. *Lithuanian Mathematical J.* **53**, 336–355.

[57] Pickands III, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist.* **3**, 119-131.

[58] Pisarenko, V.F. & Sornette, D. (2003). Characterization of the frequency of extreme events by the generalized Pareto distribution. *Pure and Applied Geophysics* **160**, 2343–2364.

[59] Reiss, R.-D. & Thomas, M. (2001; 2007). *Statistical Analysis of Extreme Values, with Application to Insurance, Finance, Hydrology and Other Fields*, 2nd edition; 3rd edition, Birkhäuser Verlag.

[60] Rosário, P. (2013). *Valores Extremos em Sismologia–Caso Estudo*. Mestrado em Estatística e Investigação Operacional, DEIO, FCUL.

[61] Scarrot, C. & MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *Revstat* **10**:1, 33–60.

[62] Smith, R.L. (1987). Estimating tails of probability distributions. *Ann. Statist.* **15**, 1174–1207.

[63] Sornette, D. (1998). Discrete scale invariance and complex dimensions. *Physics Reports* 297, 239–270.

[64] Stehlík M., Potocký R., Waldl H. & Fabián Z. (2010). On the favourable estimation of fitting heavy tailed data. *Computational Statistics* **25**, 485–503.

[65] Stupfler, G. (2016). Estimating the conditional extreme-value index under random right-censoring. *J. Multivariate Analysis* **144**, 1–24.

[66] Stupfler, G. (2019). On the study of extremes with dependent random right- censoring. *Extremes* **22**, 97–129.

[67] Temido, M.G. & Canto e Castro, L. (2003). Max-semistable laws in extremes of stationary random sequences. *Theory Probab. Appl.* **47**:2, 365–374.

[68] Tiago de Oliveira, J.C. (1993), ed., *J. Tiago de Oliveira: O Homem e a Obra*. Colecção Grandes Mestres, Edições Colibri.

[69] Weissman, I. (1978). Estimation of parameters and large quantiles based on the *k* largest observations. *J. Amer. Statist. Assoc.* **73**, 812-815.

[70] Worms, J. & Worms, R. (2014). New estimators of the extreme value index under random right censoring, for heavy-tailed distributions. *Extremes* **17**, 337–358.

[71] Worms, J. & Worms, R. (2018). Extreme value statistics for censored data with heavy tails under competing risks. *Metrika* **81**:7, 849–889.