Universidade de Aveiro Departamento de Matemática

2010

Paula Cristina Ramos Neves

A Teoria de Valores Extremos na Quantificação de Precipitação Elevada

Paula Cristina Ramos Neves

A Teoria de Valores Extremos na Quantificação de Precipitação Elevada

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requesitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, na especialização em Matemática Empresarial e Tecnológica, realizada sob a orientação científica da Prof.ª Doutora Cláudia Margarida Pedrosa Neves, professora auxiliar do Departamento de Matemática da Universidade de Aveiro.

o júri / the jury

presidente / president

Prof.ª Doutora Isabel Simões Pereira, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro.

vogais / examiners committee

Prof.ª Doutora Cláudia Margarida Pedrosa Neves, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro (orientadora).

Prof.ª Doutora Maria Cristina Souto Miranda, Professora Adjunta do Instituto Superior de Contabilidade e Administração da Universidade de Aveiro.

agradecimentos / acknowledgements

À Profª. Doutora Cláudia Margarida Pedrosa Neves, do departamento de Matemática da Universidade de Aveiro. Agradeço pela forma gentil com que me acolheu e pelo incentivo nos momentos mais difíceis. A sua orientação e sugestões em muito contribuíram para uma melhor exposição de ideias e conceitos.
À Coordenadora do mestrado Profª. Doutora Isabel Pereira pela forma calorosa com que me acolheu.
À Universidade de Cabo-Verde pela oportunidade concedida.
À minha família e amigos que me apoiaram durante este percurso, mesmo aqueles que estando longe sempre me dirigiram palavras de incentivo e encorajamento.

palavras-chave: Distribuição Generalizada de Valores Extremos, Estimação Semi-paramétrica, Índice de valores extremos.

Resumo

O objectivo principal deste trabalho é realçar a importância da Teoria de Valores Extremos na quantificação do risco, condicionalmente a um estado extremo do clima. São apresentados de forma sucinta os principais resultados que alicerçam a Teoria de Valores Extremos. Estatísticas associadas à caracterização do comportamento e reconhecimento do peso da cauda são igualmente abordadas. A modelação da cauda direita da distribuição subjacente aos dados é um assunto de particular interesse. Neste sentido, são apresentadas técnicas de inferência estatística em valores extremos que permitem obter estimativas razoáveis de quantis elevados fora da amplitude da amostra. Seguindo uma abordagem semi-paramétrica, são identificados domínios de atracção e consequentemente famílias de distribuições de valores extremos que melhor se adequam aos dados.

keywords: Generalized Extreme Value distribution, Semi-parametric estimation, Extreme value index.

Abstract

The main objective of this study is to highlight the importance of the Extreme Value Theory in risk quantification, in relation to extreme weather conditions, particularly high rainfall events. A brief summary of the results related to this theory as well as some statistics that enable the characterization of the behavior process and heavy tailed data recognition are presented.

The modeling of the right tail of the underlying distribution of a sample is a subject of special interest. Techniques of statistical inference in extreme values that allow reasonable estimation of extreme precipitation quantiles outside the range of the sample are also presented. Following a semi-parametric approach, we identify domains of attraction and therefore, families of extreme value distributions that best fit the data in study.

Sumário

Capítulo 1

Introdução

Durante as últimas décadas, tem havido mudanças notáveis no clima mundial e Europeu. As temperaturas estão subindo, a precipitação em muitas partes da Europa está a mudar e os extremos climáticos mostram uma frequência crescente em algumas regiões (Painel Intergovernamental da ONU sobre Mudanças Climáticas - IPCC, 2001a) e citado pela EEA-European Environment Agency (2004).
O IPCC afirmou que "é muito provável que a frequência de chuvas intensas venha a aumentar com o aumento da temperatura média global. Isto implica mudanças nos padrões de precipitação, um factor importante, entre outros, para a intensidade e frequência das inundações".

De acordo com o IPCC, "há novas e mais fortes evidências de que a maior parte do aquecimento observado durante os últimos 50 anos pode ser atribuída a actividades humanas, em especial para a emissão de gases de efeito estufa" (IPCC, 2001a).
Alterações climáticas induzidas pelo homem, deverão continuar nas próximas décadas (IPCC, 2001a) com efeitos consideráveis na sociedade humana e no ambiente. A magnitude dos impactos depende fortemente da natureza e da taxa de aumento da temperatura no futuro. Consequências das mudanças climáticas incluem um aumento do risco de inundações e secas, perda de biodiversidade, ameaças à saúde humana e prejuízos para os sectores económicos como a silvicultura, a agricultura, o turismo e a indústria de seguros (IPCC, 2001b).

Muitas realizações em relação às estatísticas de valores extremos e avaliação dos potencias impactos de mudança climática têm sido feitas nos últimos anos e estudos recentes indicam mudanças na intensidade e frequência de eventos extremos pelo globo. (IPCC 2001b, 2007). Mudanças extremas no clima (por exemplo: precipitação extrema, inundações, secas, etc) são particularmente relevantes devido ao seu significante impacto na vida humana e desenvolvimento sócio-económico.

A análise estatística de eventos de precipitação extrema é um pré-requisito na avaliação do risco e pode contribuir para uma melhor previsão do risco de inundações, fundamental na gestão dos recursos hídricos.
Estaremos sempre expostos ao risco de inesperadas e desagradáveis alterações climáticas de causas naturais ou humanas. Um exemplo dos efeitos adversos resultantes é o recente dilúvio ocorrido na ilha da Madeira em Portugal a 20 Fevereiro de 2010. Um caso de inundação jamais visto em Portugal e de proporções catastróficas deixou algumas regiões completamente devastadas pela força das águas, vitimando pessoas e bens.
Não é possível evitar os eventos de precipitação extrema, mas é possível controlar as consequências que elas geram. Assim o presente trabalho aplica metodologias de quantificação do risco condicional a um estado extremo do clima, tais como o VaR (valor de risco).

Um estudo sistemático de eventos extremos é de grande relevância para a climatologia e hidrologia. Assim, estimativas decorrentes são imprescindíveis para o planeamento e desenvolvimento das actividades sujeitas a efeitos adversos, especialmente estruturas de engenharia civil e sistemas agrícolas.

No contexto financeiro, é preciso modelar elevações e quedas abruptas que podem causar grandes perdas. Neste caso, um dos grades desafios para o gestor de risco é implementar modelos de gestão de risco para prever acontecimentos raros e catastróficos e permitir a medição das suas consequências. Sem ferramentas de controle adequadas, decisões impróprias podem ser tomadas. Assim mais atenção deverá ser dada à oscilação de realizações de uma variável aleatória, em particular aos seus maiores valores.

A análise tradicional usa a estatística e probabilidade baseadas no teorema do limite central englobando apenas os eventos associados ao centro da distribuição. Sob condições extremas, esta análise revela-se ineficaz justificando o uso de métodos mais sofisticados de avaliação do risco uma vez que a distribuição Normal não consegue acomodar os riscos associados a eventos extremos.

A Teoria de Valores Extremos (EVT) reveste-se de uma importância fundamental na modelação de eventos extremos. A EVT é frequentemente utilizada na engenharia de recursos hídricos e estudos de gestão (Katz *et al.* 2002; Smith 2001) para obter distribuições de probabilidades com o objectivo de modelar máximos ou mínimos em amostras de dados de variáveis aleatórias, assim como para modelar a distribuição dos excessos acima de certo nível. Outras áreas importantes do conhecimento onde a EVT é aplicada são por exemplo na meteorologia, análise de risco sísmico, ciências do ambiente, telecomunicações, finanças, seguros, longevidade da vida humana. (Reiss and Thomas, 2007).

Os fundamentos desta teoria foram desenvolvidos pela primeira vez por Fisher e Tippett (1928) que introduziram a teoria assintótica das distribuições de valores extremos. Gnedenko (1943) forneceu provas matemáticas de que sob determinadas condições, três famílias de distribuições (Gumbel, Fréchet e Weibull) podem surgir como distribuições limite para

valores extremos em amostras aleatórias. A unificação das famílias Gumbel, Fréchet e Weibull, é conhecida como distribuição Generalizada de Valores Extremos (GEV) (von Mises (1936) e Jenkinson (1955)). Gumbel (1942) aplica a distribuição de frequências de valores extremos na análise de inundações. Além disso Gumbel (1958) desenvolveu a teoria dos máximos anuais, com a concomitante aplicação da distribuição Gumbel a várias situações práticas.

A abordagem GEV é muitas vezes referida na literatura como o método dos Máximos Anuais ou *Block Maxima* quando é usada para modelar distribuições do máximo/mínimo de um dado conjunto de dados (Golstein *et al.* 2003). Uma abordagem alternativa ao método dos Máximos Anuais é o método *Peaks Over Threshold* (POT), baseado nas observações que excedem um determinado nível que se supõe elevado. Esta metodologia é análoga à distribuição Generalizada de Valores Extremos para os Máximos Anuais, mas, lida com a chamada distribuição generalizada de Pareto (GPD). Os fundamentos da análise de valores extremos baseada em valores acima de determinado nível, foram estabelecidos por Balkema e de Haan (1974) e Pickands (1975).

A probabilidade de eventos extremos pode ser estimada usando o índice de valores extremos, que descreve o comportamento da cauda direita da distribuição subjacente a um determinado conjunto de dados. O índice de valores extremos é o parâmetro de forma que rege o comportamento da cauda e pode ser estimado a partir de uma amostra de dados observados com função distribuição desconhecida. Para aplicações práticas das distribuições GEV e GPD, os parâmetros deverão ser estimados. Hill (1975) e Pickands (1975) introduziram os estimadores de Hill e Pickands respectivamente. Dekkers *et al.* (1989), Danielsson e Vries (1997), Ferreira *et al.* (2003) e Fraga Alves *et al.* (2009), propuseram diferentes estimadores para o índice de valores extremos baseados nos momentos.

A EVT proporciona técnicas de inferência estatística orientadas para o estudo de comportamentos extremais de certos factores que ocorrem no Universo. A teoria geral visa avaliar o tipo de distribuições de probabilidade geradas por processos aleatórios. A EVT reveste-se de grande importância na avaliação e modelação do risco de eventos extremos altamente incomuns ou raros. Geralmente, o objectivo da análise de precipitação extrema, é obter estimativas razoáveis de quantis elevados de precipitação que excedem um dado nível. Como foi dito anteriormente, isto pode ser feito ajustando a distribuição generalizada de Pareto (GPD) aos excessos acima de um nível determinístico. Estimativas do índice de valores extremos são determinadas com o objectivo de caracterizar o comportamento da cauda de longas séries históricas de valores diários de precipitação. Os resultados obtidos serão utilizados para extrapolar a função distribuição de valores extremos. Esta extrapolação pode ser empregue para estimar a probabilidade de ocorrência de precipitação acima de um nível aleatório.

O problema prático sobre o qual este trabalho se debruça é com base nas técnicas de inferência estatística fornecidas pela EVT, extrapolar para além da amplitude da amos-

tra. Particularmente, e usando longas séries históricas de valores de precipitação diária em *Berlim* (Alemanha) e *de Bilt* (Holanda), pretende-se no presente trabalho, avaliar o comportamento da cauda da distribuição subjacente às subamostras de valores diários de precipitação acima dos 80 *mm*, para as duas cidades, através de estimativas do índice de valores extremos (parâmetro que dá peso à cauda da distribuição), neste caso, determinadas segundo o estimador dos momentos. O objectivo é, seguindo uma abordagem semi-paramétrica, identificar domínios de atracção e consequentemente famílias de distribuições de valores extremos que melhor se adequam às amostras de *Berlim* e *de Bilt*, bem como determinar a probabilidade de ocorrência de eventos de precipitação extrema. São estimados níveis de precipitação de risco, convista à obtenção de previsões para além dos valores observados na amostra. Pretende-se assim mostrar a utilidade das metodologias desenvolvidas na EVT, na avaliação e quantificação do risco de eventos extremos, como uma ferramenta que pode dar-nos um preciosa indicação do quão expostos podemos estar às condições climáticas adversas.

Uma das recentes abordagens semi-paramétricas utilizada e associada à EVT, é o *Peaks-Over-Random-Threshold*, por forma a obter estimativas dos quantis ou níveis elevados de precipitação da distribuição de excessos para além de um nível aleatório, para duas cidades em estudo: *Berlim*, na Alemanha, e *de Bilt*, na Holanda.
Esta metodologia é também utilizada para determinar estimativas dos níveis de precipitação de risco, considerando para a cidade de *de Bilt*, uma amostra de 100 anos de registro dos níveis diários de precipitação, dividido em sucessivos períodos de 5 anos e utilizando como conjunto de valores extremos em cada período as 25 observações de topo. A estimação é realizada com a incorporação de uma função de tendência, assumindo portanto a não estacionaridade do processo. Pretende-se deste modo determinar medidas de alteração do clima nos subsequentes períodos de 5 anos em função do primeiro período ou instante inicial.

Esta dissertação está dividida em 6 capítulos. Após a introdução, são apresentadas no capítulo 2 algumas das mais usuais medidas de risco que tentam descrever a cauda de uma distribuição. No capítulo 3 são abordados de forma sumária alguns dos principais resultados da Teoria de Valores Extremos, nomeadamente o Teorema fundamental Fisher e Tippett (1928), Gnedenko (1943), que estabelece três domínios de atracção para o máximo convenientemente normalizado e correspondentes distribuições limite para as maiores observações. As abordagens usuais em inferência estatística de valores extremos:
A metodologia clássica de Gumbel dos Máximos por Blocos (ou Máximos Anuais), o método POT dos excessos de nível (*Peaks-Over-Threshold*) e a mais recente metodologia PORT dos excessos de nível aleatório (*Peaks-Over-Random-Threshold*). Alguns dos métodos de estimação paramétrica e semi-paramétrica dos parâmetros associados às distribuições Generalizada de Valores Extremos (GEV) e generalizada de Pareto (GPD) são apresentados no capítulo 4. Os dados em análise nesta dissertação e as metodologias da EVT a eles aplicadas são apresentados no capitulo 5. O capítulo 6 é reservado a conclusões globais.

Capítulo 2

Medidas de Risco Extremo

O presente capítulo é motivado por um grande número de problemas concretos de gestão do risco, nomeadamente riscos financeiros, riscos de seguro, risco de alterações climáticas. Os riscos financeiros englobam: riscos de mercado devido aos movimentos desfavoráveis do mercado, onde a preocupação é a determinação diária do valor em risco para as perdas que incorremos em uma carteira de negociação, riscos de crédito ou gestão de risco operacional associados a perdas irregulares de desvalorizações de crédito e inadimplência ou problemas operacionais imprevistos. Aqui o objectivo pode ser a determinação do capital de risco exigido como caução contra tais riscos.

Nos seguros, um problema comum é a precificação ou a constituição de reservas financeiras para produtos que oferecem protecção contra perdas avultadas como o excesso de perda de tratados de resseguro celebrados com seguradoras primárias. A área dos seguros possui experiência considerável na gestão de riscos extremos, e vários métodos actualmente reconhecidos como pertencentes à Teoria de Valores Extremos possuem um longo historial de uso pelos actuários.
Os riscos de alterações climáticas consistem em mudanças nas condições do clima, decorrentes de factores naturais e/ou humanos, com efeitos consideráveis na sociedade humana e no ambiente. A magnitude dos impactos depende fortemente da natureza e da taxa de aumento da temperatura no futuro. Consequências das mudanças climáticas incluem um aumento do risco de inundações e secas, perda de biodiversidade, ameaças à saúde humana e prejuízos para os sectores económicos como a silvicultura, a agricultura, o turismo e a indústria de seguros (IPCC, 2001b). Um indicador útil para mudanças na magnitude e na frequência de inundações, é a comparação das estimativas de nível de retorno. Para este efeito, a quantificação confiável da incerteza associada às estimativas do nível de retorno é crucial. Alertados para um aumento dos riscos de inundação, tomadores de decisão, com base em resultados quantitativos e explícitos, tentam reajustar a avaliação de riscos e estratégias de gestão. Avaliações de vulnerabilidade regional pode ser uma estratégia para lidar com a ameaça e antecipar cenários extremos, como inundações ou ondas de calor.

É preciso então formalizar os conceitos de risco e de risco extremo. De acordo com (J. MCNeil, A. (2000)), *Riscos* são essencialmente variáveis aleatórias que representam estados futuros e incertos do mundo através de valores que expressam ganhos e perdas. Esses riscos podem ser considerados individualmente ou como parte de um processo estocástico onde riscos presentes dependem de riscos passados. Os riscos aleatórios podem ser, por exemplo:

□ Retornos diários (negativos) de activos financeiros ou portifólio - perdas e ganhos;

□ Perdas operacionais;

□ Indemnizações por sinistros catastróficos;

□ Perdas de crédito.

A abordagem matemática usual para modelar o risco, usa a Teoria da Probabilidade. Aos possíveis valores de um risco, está associada uma distribuição de probabilidade que nunca será observada com exactidão, apesar do conhecimento de valores de perdas passadas devido a riscos similares, uma vez que, quando disponível, esta informação parcial será utilizada no ajustamento a determinada distribuição.
Um *evento extremo* ocorre quando um risco assume valores de cauda da função distribuição subjacente. A Teoria de Valores Extremos é uma ferramenta que tenta fornecer-nos as melhores estimativas possíveis sobre a cauda dessa distribuição.
Segundo (J. MCNeil, A. (2000)), medir um risco significa resumir a sua distribuição num número tido como medida de risco. A média e a variância, por exemplo, medem aspectos centrais do risco mas não fornecem muita informação sobre o risco extremo.

Seja $X_1, X_2, ...$ uma sucessão de variáveis aleatórias i.d.(identicamente distribuídas) com função distribuição (f.d.) subjacente desconhecida dada por $F(x) = P\{X_i \leq x\}$, $\forall x \in \mathbb{R}$.

Algumas das questões mais frequentes sobre a quantificação do risco em muitos campos da ciência moderna, passando pela engenharia, finanças até seguros, envolvem estimativas de quantis elevados. Isto corresponde à determinação de um valor (quantil x_{1-p}) que é excedido por determinada variável com baixa probabilidade (ou com probabilidade p elevada). Seguidamente são apresentadas algumas definições das mais usuais medidas de risco que tentam descrever a cauda de uma distribuição e fornecem o conhecimento geral para aplicações práticas. Um típico exemplo de tais medidas é o Valor de Risco (*Value-at-Risk*). Outras medidas utilizadas com menor frequência são o Défice Esperado (*Expected Shortfall*) e o Nível de Retorno (*Return Level*).

Definição 1 [*Value-at-Risk* (VaR)]
O risco é geralmente expresso como o *valor de risco* (VaR), ou seja, o tamanho da perda que ocorre com uma pequena probabilidade p fixado. A definição de VaR é dada em termos de quantis estatísticos e, pode ser então definido como o $(1-p)$-ésimo quantil da distribuição F. Trata-se portanto de um *quantil* $x_{1-p} := F^{\leftarrow}(1-p)$, $p \in (0,\, 1)$ de uma função distribuição F, onde

$$F^{\leftarrow}(y) := \inf\{x : F(x) \geq y\} \tag{2.1}$$

é a função inversa generalizada de F.
Denote-se por $U(t)$ a função inversa de $1/(1-F)$,

$$U(t) := \left(\frac{1}{1-F}\right)^{\leftarrow}(t) = F\left(1-\frac{1}{t}\right),\, t \geq 1.$$

Então para um valor pequeno de p (eventualmente dependendo da dimensão da amostra n), pretende-se estimar

$$VaR_p = U(1/p)\ , \qquad p = p_n \to 0\ ,\ np_n \leq 1.$$

Por exemplo $VaR_{0.95} = -0.02$ significa que com 95% de probabilidade, a perda máxima esperada de um portifólio é de 2% do seu valor total. VaR_p é um quantil elevado da distribuição das perdas, usualmente a 95% ou 99%, que fornece uma espécie de limite superior para a perda que somente é excedido numa pequena proporção de ocasiões e que pode ser estimado a partir de dados observados.
Porque estamos a lidar com uma pequena probabilidade à direita, somos conduzidos a modelar a cauda da f.d. F subjacente aos dados amostrais. Além disso, e porque geralmente em aplicações práticas encontramos caudas pesadas, podemos supor que a f.d. F subjacente aos dados é tal que:

$$1 - F(x) \sim cx^{-1/\gamma}, \qquad x \to \infty, \tag{2.2}$$

para alguma constante positiva c. Mais genericamente, pode-se supor que $1-F$ é de variação regular no infinito com índice $-1/\gamma$ (notação: $1-F \in VR_{-1/\gamma}$), onde VR_α representa a classe de funções mensuráveis positivas $g(.)$, tal que, para todo $x > 0$, $g(tx)/g(t) \to x^\alpha$, $t \to \infty$. Equivalentemente a (2.2), U varia regularmente de acordo com o índice γ. Assim, podemos escrever $U(t) = t^\gamma L(t)$, onde L é uma função de variação regular lenta, isto é, $L \in VR_0$.

Definição 2 [*Expected Shortfall* (ES)]
Outra medida informativa do risco é o défice esperado ou esperança condicional da cauda, que estima o tamanho potencial da perda que excede o VaR:

$$ES_p = E(X|X > VaR_p)$$

Definição 3 [*Return Level*]
O nível de retorno (*Return Level*) associado a um dado período de retorno t é o nível z_t que se espera exceder em média uma vez a cada t anos. Assim, o nível de retorno z_t é simplesmente um quantil de ordem p, com $p = 1 - 1/t$. Num contexto de análise de séries temporais onde o índice i de X_i corresponde a uma unidade de tempo, por exemplo um ano, pretende-se determinar qual o nível z_t tal que se pode esperar que ocorra um único evento maior que z_t em t anos. Isto é equivalente a $\overline{F}(z_t) = 1/t$ onde $\overline{F} = 1 - F$. Por exemplo se $p = 0.99$, então $t = 100$ e tem-se $z_{0.99}$, tal que $P(X \geq z_{0.01}) = 0.99$. Podemos concluir que em média X será maior ou igual a $z_{0.01}$ a cada 100 anos.

Tal como a função distribuição F, estas medidas são quantidades teóricas que desconhecemos. O objectivo na avaliação e quantificação do risco passa por obter estimativas $\widehat{VaR_p}$, $\widehat{ES_p}$ e $\widehat{z}_t$ destas medidas e eventualmente pela substituição de F pela sua imagem estatística: a função distribuição empírica F_n.

O trabalho de Artzner *et al.* (1997) critica a utilização do VaR como uma medida de risco sob dois aspectos. Primeiramente mostraram que VaR não é necessariamente subadictivo, portanto segundo sua terminologia, VaR não é uma medida de risco coerente. Existem casos onde o portifólio pode ser dividido em sub-portifólios tal que a soma do VaR correspondente é menor do que o VaR do portifólio total. Isto poderá causar problemas se o sistema de gestão de riscos de uma instituição financeira é baseada em limites VaR para carteiras individuais. Além de que o VaR nada nos diz sobre o magnitude potencial da perda que o excede.
Artzner *et al.* (1997) introduz um novo conceito, medida de risco coerente (CRM) e propõe o uso do défice esperado como medida de risco alternativa ao VaR. De acordo com sua definição o défice esperado é uma medida de risco coerente.
Uma medida de risco coerente é uma função real $\rho : \mathbb{R} \to \mathbb{R}$ mensurável, com as seguintes características:

sejam X e Y variáveis aleatórias,

1. Monotonia: $X \geq Y \Rightarrow \rho(X) \geq \rho(Y)$.
2. Invariância para Translações: $\rho(X + \alpha) = \rho(X) + \alpha$, para todo o $\alpha \in \mathbb{R}$.
3. Homogeneidade positiva de ordem 0: Para qualquer $\lambda \geq 0$, $\rho(\lambda X) = \lambda\rho(X)$.
4. Subaditividade: $\rho(X + Y) \leq \rho(X) + \rho(Y)$.
5. Convexidade: Para todo X e Y e $\lambda \geq 0$, $\rho(\lambda X + (1 - \lambda)Y) \leq \rho(X) + (1 - \lambda)\rho(Y)$

Estas características correspondem às exigências óbvias e intuitivas da quantificação do risco de um mercado. Mais concretamente,
(1) Quando a perda de um investimento X é sempre maior ou igual à perda do investimento Y, então o risco do investimento X também é maior ou igual ao risco do investimento Y.

(2) Ao adicionar um investimento livre de risco, ou seja um investimento não aleatório com perdas conhecidas a ($a < 0$, quando o investimento possui pagamentos fixos) à uma carteira, o risco se altera em exactamente a.
(3) Homogeneidade positiva implica que o risco do seu portfolio ou carteira está linearmente relacionado com o tamanho do mesmo.
(4) O risco de uma carteira composta de investimentos em X e Y é no máximo tão grande quanto a soma dos riscos individuais (diversificação do risco).
Segundo o mesmo autor, o VaR não é uma medida de risco coerente. Embora satisfaça os outros três axiomas, não satisfaz a subaditividade em determinadas situações.
(5) Outra medida de risco relacionada é a medida de risco convexa onde os axiomas da subaditividade e da homogeneidade positiva são substituídos pelo axioma da convexidade.

Capítulo 3

Teoria de Valores Extremos

Este capítulo discute os fundamentos básicos da Teoria de Valores Extremos (EVT). A EVT proporciona técnicas de inferência estatística orientadas para o estudo de comportamentos extremais de certos factores que ocorrem no Universo. A teoria geral visa avaliar o tipo de distribuições de probabilidade geradas por processos. A Teoria de Valores Extremos reveste-se de grande importância na avaliação e modelação do risco de eventos extremos, altamente incomuns ou raros. Estes eventos englobam por exemplo, chuvas torrenciais, inundações, ondas de calor, vagas de frio, secas prolongadas, terramotos entre outros. São eventos cuja a probabilidade de ocorrência é baixa, mas que quando ocorrem apresentam graves consequências. Existem hoje dois principais métodos para estimar as distribuições de cauda: a abordagem da teoria de base, em conformidade com o primeiro teorema da Teoria de Valores Extremos (Fisher e Tippett, 1928; Gnedenko, 1943); mais comum actualmente é o ajustamento à cauda de uma distribuição, abordagem baseada no segundo teorema da Teoria de Valores Extremos sobre excessos de nível (Pickans, 1975; Balkema e de Haan, 1974). A diferença entre os dois reside essencialmente na natureza da geração dos dados. As distribuições de valores extremos são as distribuições limite para o mínimo ou o máximo de uma colecção muito grande de variáveis aleatórias independentes da mesma distribuição. O termo valor extremo pode ser interpretado de duas formas: é o máximo ou o mínimo, isto é, o maior (menor) valor de uma série; ou representa os excessos, isto é, os maiores valores de um conjunto de dados acima de um nível suficientemente elevado. Desta forma a Teoria de Valores Extremos (EVT) modela os extremos usando a distribuição limite do máximo (mínimo) convenientemente normalizado ou dos excessos acima de determinado nível, através dos dois principais métodos da EVT utilizados para estimar as distribuições de cauda: *Block Maxima* e *Peaks-Over-Threshold* (POT), respectivamente. Estes conceitos serão formalizados em secções posteriores deste capítulo.

Neste capítulo são apresentados alguns dos resultados fundamentais da Teoria de Valores Extremos, utilizados para modelar as distribuições subjacentes às medidas de risco. A secção 3.1 é reservada a alguns resultados preliminares. Os resultados mais importantes, nomeadamente o teorema fundamental de Fisher e Tippett com as três distribuições de valores extremos associadas são apresentados na secção 3.2. Finalmente, na secção 3.3

falaremos da distribuição limite dos excessos de nível que sustentam a metodologia POT.

3.1 Alguns Resultados Preliminares

Seja X uma v.a. cuja função de distribuição é denotada por F. Para facilitar a exposição, supomos de agora em diante que F é absolutamente contínua. Defina-se x^F o limite superior do suporte de F, *i.e.*,

$$x^F := \sup\{x : F(x) < 1\}. \tag{3.1}$$

Seja $X_1, X_2, ..., X_n, ...$ uma sucessão de v.a.'s i.i.d. e portanto com função distribuição comum F. O máximo das n primeiras v.a.'s da sucessão é denotado por $X_{n:n}$:

$$X_{n:n} = \max(X_1, X_2, ..., X_n).$$

Assim, a título de exemplo, se $X_1, X_2, ..., X_n$ forem valores diários de precipitação (riscos), então $X_{n:n}$ será o valor máximo de precipitação entre n valores de precipitação diários.
Os resultados que se apresentam a seguir referem-se essencialmente ao máximo da amostra, na medida em que, por um lado, importa para o presente estudo as ocorrências de acontecimentos de precipitação anormalmente elevada, por outro lado, a conversão dos resultados na maior parte das vezes é quase imediata, atendendo a que:

$$\min(X_1, X_2, ..., X_n) = -\max(-X_1, -X_2, ..., -X_n)$$

No contexto de v.a.'s i.i.d., a função de distribuição (f.d.) exacta de $X_{n:n}$ é dada por:

$$F_{X_{n:n}}(x) = P(X_{n:n} \leq x) = P(X_1 \leq x, ..., X_n \leq x) = (P\{X \leq x\})^n = F^n(x),$$

para todo $x \in \mathbb{R}$.
Então, é possível concluir a convergência quase certa para o limite superior do suporte x^F definido em (3.1):

$$X_{n:n} \xrightarrow{q.c.} x^F, \quad n \to \infty \tag{3.2}$$

dado que para $x < x^F$, $0 \leq F(x) < 1$ e então

$$F_{X_{n:n}}(x) = F_X^n(x) \to 0, \quad n \to \infty$$

Por conseguinte, $X_{n:n}$ converge em probabilidade para o limite superior do suporte x^F, *i.e.*, $X_{n:n} \xrightarrow{p} x^F$. Como a f.d. F é absolutamente contínua e monótona, então a convergência quase certa em (3.2) é verificada.
No entanto, este resultado não é muito expressivo. É importante conhecer a magnitude que o máximo de uma amostra de dimensão n pode assumir.
Um dos objectivos da Teoria de Valores Extremos, consiste no estudo do comportamento do máximo ou do mínimo de uma amostra, procurando encontrar uma distribuição aproximada quando n é suficientemente grande.
Ao modelar o máximo parcial de uma sucessão de v.a.'s i.i.d. com f.d. F contínua, as distribuições de valores extremos assumem um papel semelhante ao da distribuição Normal no Teorema do Limite Central, para somas de variáveis aleatórias.

3.2 Distribuição do Máximo (GEV)

Para alcançar uma distribuição não degenerada como limite da f.d. do máximo de v.a.'s i.i.d., a expressão (3.2) sugere a transformação de $X_{n:n}$ numa variável $X^*_{n:n}$ mediante linearização:

$$a_n^{-1}(X_{n:n} - b_n) = X^*_{n:n},$$

mediante constantes normalizadoras $a_n > 0$ e $b_n \in \mathbb{R}$.

Teorema 3.2.1 *(Fisher e Tippett (1928), Gnedenko (1943)):*

Seja $X_1, X_2, ..., X_n, ...$ uma sucessão de v.a.'s i.i.d. com a mesma função distribuição F. Se existirem constantes $a_n > 0$ e $b_n \in \mathbb{R}$ tais que

$$\lim_{n\to\infty} P\left\{X^*_{n:n} \leq x\right\} = \lim_{n\to\infty} P\left\{\frac{X_{n:n} - b_n}{a_n} \leq x\right\} = \lim_{n\to\infty} F^n(a_n x + b_n) = H(x),$$

para todo o x pertencente ao conjunto dos pontos de continuidade de H, então as únicas formas possíveis para a f.d. limite são:

Tipo I (Gumbel) $\quad \Lambda(x) = \exp\{-e^{-x}\},\ x \in \mathbb{R};$

Tipo II (Fréchet) $\quad \Phi_\alpha(x) = \exp\{-x^{-\alpha}\},\ x > 0,\ \alpha > 0,$

Tipo III (Weibull) $\quad \Psi_\alpha(x) = \exp\{-(-x)^{\alpha}\},\ x \leq 0,\ \alpha > 0.$

Na figura seguinte estão representadas, a título de exemplo, funções densidade de probabilidade associadas a diferentes formas possíveis de $H(.)$: Gumbel, Fréchet ou Weibull. Observamos que a distribuição de Fréchet possui uma cauda direita com decaimento do tipo polinomial negativo. O decaimento exponencial da cauda direita da distribuição Gumbel, é uma característica comum a distribuições de cauda leve.
Finalmente, a distribuição Weibull possui uma cauda que decresce rapidamente quando x se aproxima do limite superior do suporte, finito neste caso.
Embora os três modelos sejam distintos, é possível unificar as correspondentes funções distribuição mediante a consideração de um parâmetro de forma único (parametrização de von Mises (1936) e Jenkinson (1955)), bastando para isso redefinir as constantes $a_n > 0$ e b_n anteriores:

$$\exists^{a^*_n>0}_{b^*_n\in\mathbb{R}} : F^n(a^*_n x + b^*_n) \xrightarrow[n\to\infty]{} H(x) = \begin{cases} \exp\{-(1+\gamma x)^{-1/\gamma}\}, & 1+\gamma x > 0 \quad \text{se} \quad \gamma \neq 0 \\ \exp\{-e^{-x}\}, & x \in \mathbb{R} \quad \text{se} \quad \gamma = 0 \end{cases}. \tag{3.3}$$

A família $\{H_\gamma,\ \gamma \in \mathbb{R}\}$ é designada por família de distribuições Generalizadas de Valores Extremos, do inglês ("*Generalized Extreme Value distribution*" - GEV), e diz-se que F está

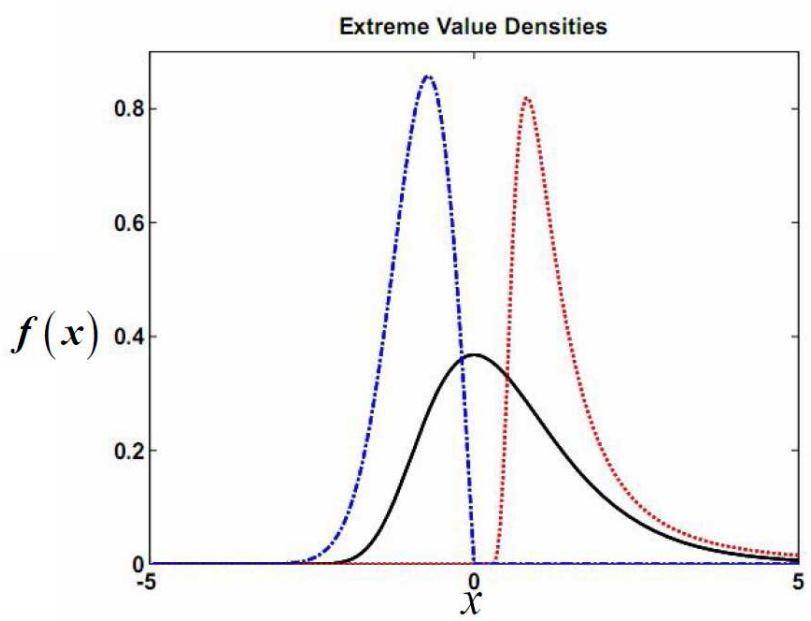

Figura 3.1: Gráfico das funções densidade de probabilidade associadas à distribuição de Gumbel (linha sólida), Fréchet com parâmetro $\alpha = 2$ (linha ponteada) e Weibull com parâmetro $\alpha = -2$ (linha tracejada).

no domínio de atracção da GEV, com a notação $F \in D(H_\gamma)$, $\gamma \in \mathbb{R}$. O parâmetro de forma γ é designado de índice de valores extremos. As distribuições de valores extremos apresentadas no Teorema 3.2.1 podem ser identificadas do seguinte modo:

1. A distribuição *Gumbel* quando $\gamma = 0$, correspondendo ao caso limite em que $\gamma \to 0$;
2. A distribuição *Fréchet* quando $\gamma = \alpha^{-1} > 0$;
3. A distribuição *Weibull* quando $\gamma = -\alpha^{-1} < 0$.

O domínio de atracção Gumbel pode ser encarado como uma classe intermédia de distribuições, entre o domínio de atracção Fréchet e o domínio Weibull, porque para valores de γ muito próximos de zero, as distribuições tipo Fréchet e do tipo Weibull são muito próximas da de Gumbel. Habitualmente não dispomos de informação à priori sobre a distribuição limite do máximo da a.a. $(X_1, \ldots X_n)$ em estudo, por isso a representação generalizada é particularmente útil quando se pretende estimar o índice de valores extremos no caso geral de $\gamma \in \mathbb{R}$, por exemplo através do método da máxima verosimilhança.
De um modo geral existem três metodologias principais no âmbito da inferência estatística em valores extremos: o método clássico de Gumbel ou *Block Maxima* ou Máximos Anuais, o método *Peaks-Over-Threshold* (POT), e o mais recente método *Peaks-Over-Random-Threshold* (PORT), que corresponde a uma variante do método POT, condicionado a um

nível aleatório. Como iremos ver de seguida, estas metodologias determinam classes particulares de distribuições.

3.2.1 Metodologia *Block Maxima* ou dos Máximos Anuais (MA)

Esta metodologia considera o máximo (ou mínimo) que uma variável assume em períodos sucessivos, por exemplo meses ou anos. No estudo do máximo, a amostra aleatória é dividida em blocos. De cada bloco que pode estar identificado com 1 ano de observações, extrai-se o valor máximo para formar o conjunto de valores extremos. As observações máximas constituem acontecimentos extremos, conhecidos como *block maxima* ou máximos anuais (Gumbel). A figura 3.2, representa os máximos em cada um dos quatro períodos ou blocos de dimensão $n = 3$. O problema deste procedimento reside no facto dos blocos terem necessariamente a mesma dimensão elevada, sendo que o valor extremo de um bloco poderá ser menor do que alguns valores inferiores ao extremo de outro. Em consequência a frequência de extremos pode ficar mal medida. Por exemplo, na figura 3.2 a observação X_9, que é inferior ao máximo X_7 considerado no terceiro bloco, é maior que o máximo X_{11} considerado no quarto bloco.

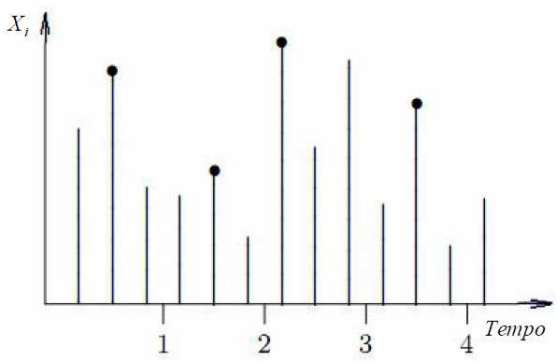

Figura 3.2: Máximos Anuais

3.2.2 Metodologia *Peaks-Over-Threshold* (POT)

A metodologia POT centra-se naquelas observações que excedem um determinado nível u que se supõe elevado. Na figura 3.3 que ilustra a metodologia (POT), as observações X_1, X_2, X_7, X_8, X_9, e X_{11} excedem o nível u fixado e são por isso designadas de excedências de nível.
O método dos Máximos Anuais proposto por Gumbel é o método tradicionalmente usado para analisar dados que apresentam sazonalidade, como por exemplo, em dados hidrológicos. Contudo, a metodologia POT é considerada como mais útil em aplicações práticas devido ao seu uso mais eficiente das observações mais elevadas.

Pode-se observar pela figura 3.2 que o eixo dos xx' representa o tempo de observação dos dados dividido em quatro subperíodos ou blocos de igual dimensão. O eixo dos yy'

indica os valores assumidos pela variável X_i em estudo, em cada bloco. Em cada bloco considera-se apenas o máximo de X_i. De igual modo, na figura 3.3 os eixos xx' e yy' representam respectivamente o tempo de observação dos dados, aqui considerado sem subdivisão em blocos, e a variável X_i em estudo. Nesse período de tempo, considera-se apenas o conjunto de observações X_i que se situam acima do nível $X_i = u$. Uma clara diferença entre as metodologias MA e POT é a de que enquanto na segunda as maiores observações da amostra são tomadas em consideração, na primeira os máximos anuais que constituem o conjunto de valores extremos, não são necessariamente as maiores observações na amostra. Por exemplo, na figura 3.2, os valores de X_i a reter e que constituirão o conjunto de valores extremos são: X_2, X_5, X_7, e X_{11}. A observação X_5, considerada como máximo no segundo bloco segundo a metodologia dos Máximos Anuais, não fará parte do conjunto de valores extremos na metodologia POT, figura (3.3).

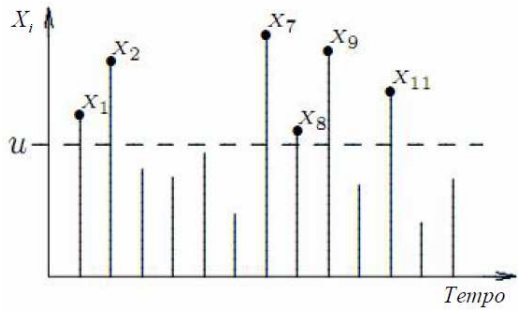

Figura 3.3: Excessos acima do nível u

Define-se como excesso de nível, relativamente ao nível u, toda a v.a. Y tal que $Y = X - u | X > u$. O número de excessos de nível a reter de uma amostra aleatória é obviamente determinado pelo nível determinista u, fixado segundo a metodologia POT. A escolha do nível u não é trivial, porque dele depende o número final k (aleatório) de observações a serem consideradas. Ao reter muitas observações haverá uma grande variabilidade nos dados, o que não é desejável dado que provoca um aumento da imprecisão das estimativas. Por outro lado, se o nível determinar um reduzido número de observações, as estimativas poderão ser pouco fiáveis.

3.2.3 Metodologia *Peaks Over Random Threshold* (PORT)

Sejam:

$$\min_{1 \leq i \leq n} X_i \equiv X_{1:n} \leq X_{2:n} \leq \ ... \ \leq X_{n:n} = \max_{1 \leq i \leq n} X_i \tag{3.4}$$

estatísticas ordinais ($e.o.'s$) ascendentes.
A metodologia PORT, é um método de inferência estatística baseado na amostra dos excessos acima do nível aleatório $X_{n-k:n}$, denotada por

$$\underline{X} := (X_{n:n} - X_{n-k:n},\ X_{n-1:n} - X_{n-k:n},\ ...\ ,\ X_{n-k+1:n} - X_{n-k:n}) \tag{3.5}$$

A classe de estimadores semi-paramétricos do índice de valores extremos γ e consequentemente de quantis elevados, associada a metodologia PORT, é função da amostra de excessos acima do nível aleatório $X_{n-k:n}$ definida em (3.5). O nível aleatório $X_{n-k:n}$, irá depender portanto do número de observações de topo (k) considerado na cauda, para o cálculo das estimativas de $\widehat{\gamma}$ e de quantis elevados.

Na abordagem clássica, muitas vezes consideramos para estimar o índice de valores extremos γ, o estimador de Hill (caso $\gamma > 0$) ou o estimador dos momentos de Dekkers *et al.* (1989), sendo que ambos os estimadores são baseados nas $k+1$ maiores ($e.o.'s$) de topo, definidos em (3.4). O estimador de Hill (1975) é um estimador consistente para o índice de cauda $\gamma > 0$ quando se considera $k = k_n$ uma sucessão intermédia superior, *i.e.*, $k_n \to \infty$ e $k_n/n \to 0$, $n \to \infty$,

$$\widehat{\gamma}^{H}_{k,n} := M^{(1)}_{k,n} = \frac{1}{k}\sum_{i=1}^{k} \log(X_{n-i+1:n}) - \log(X_{n-k:n}). \tag{3.6}$$

Nas mesmas condições, os três estimadores seguintes são consistentes para $\gamma \in \mathbb{R}$:

1. Estimador de Pickands (1975), definido em termos de quantis elevados:

$$\widehat{\gamma}^{P}_{k,n} := \frac{1}{\log 2}\ \log\left(\frac{X_{n-k+1:n} - X_{n-2k+1:n}}{X_{n-2k+1:n} - X_{n-4k+1:n}}\right). \tag{3.7}$$

2. Estimador dos momentos (Dekkers, Einmahl e de Haan, 1989)

   Este é o primeiro estimador dos momentos aqui apresentado. Tem a seguinte expressão funcional:

$$\widehat{\gamma}^{M}_{k,n} := M^{(1)}_{k,n} + 1 - \frac{1}{2}\left(1 - \frac{(M^{(1)}_{k,n})^2}{M^{(2)}_{k,n}}\right)^{-1}, \tag{3.8}$$

   onde

$$M^{(2)}_{k,n} := \frac{1}{k}\sum_{i=1}^{k} \left(\log(X_{n-i+1:n}) - \log(X_{n-k:n})\right)^2.$$

3. O segundo estimador envolvendo momentos é o designado Estimador Mixed-Moment (Fraga Alves *et al.*, 2009)

$$\widehat{\gamma}^{MM}_{n}(k) \equiv \widehat{\gamma}^{MM}_{n}(k;\ X_{n-j+1:n}\,,\ 1 \le j \le k+1) := \frac{\widehat{\varphi}_n(k) - 1}{1 + 2\min(\widehat{\varphi}_n(k) - 1, 0)}, \tag{3.9}$$

   onde

$$\widehat{\varphi}_n(k) := \left(M_n^{(1)}(k) - L_n^{(1)}(k)\right) / \left(L_n^{(1)}(k)\right)^2,$$

$$L_n(k) := \frac{1}{k}\sum_{i=1}^{k}\left(1 - \frac{X_{n-k:n}}{X_{n-i+1:n}}\right).$$

3.3 Distribuição dos Excessos

A função distribuição do excesso acima de um nível $u > 0$, também chamada função distribuição condicional dos excessos é definida por:

$$F_u(y) := P\{X - u \leq y | X > u\}, \qquad y \in \mathbb{R}. \tag{3.10}$$

A função distribuição dos excessos $F_u(y)$ representa assim a probabilidade de um risco X ultrapassar o nível u em pelo menos y, dado que excedeu este nível u fixado.

A figura 3.4, ilustra a representação da f.d. F_u do excesso à custa da f.d. F.

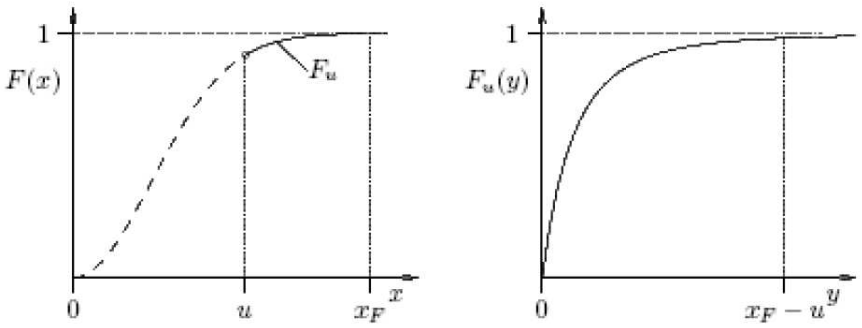

Figura 3.4: Função distribuição F (à esquerda) e Função distribuição condicional F_u (à direita).

Perante o problema da estimação da função distribuição F_u de valores de x acima de um certo nível u, a metodologia *Peaks Over Threshold* (POT) desempenha um papel importante, decorrente do resultado fundamental em EVT que estabelece a distribuição limite para o excesso acima de determinado nível elevado:

Teorema 3.3.1 *(Pickands (1975), Balkema e de Haan, (1974))*

Seja F contínua, $F \in D(H_\gamma)$, $\gamma \in \mathbb{R}$, se e só se existe uma função $\beta > 0$ tal que

$$\lim_{u \uparrow x^F} \sup_{0 \leq y < x^F - u} \left|F_u(y) - G_{\gamma,\beta(u)}(y)\right| = 0.$$

Interpretando $\beta(u)$ como um parâmetro de escala, $G_{\gamma,\beta(u)}$ é a *f.d.* da Generalizada de Pareto (GPD):

$$G_{\gamma,\beta(u)}(y) = \begin{cases} 1-(1+\frac{\gamma y}{\beta(u)})^{-1/\gamma},\ y \geq 0 & \text{se} \quad \gamma > 0 \\ 1-e^{-y/\beta(u)},\ y \geq 0 & \text{se} \quad \gamma = 0 \\ 1-(1+\frac{\gamma y}{\beta(u)})^{-1/\gamma},\ 0 \leq y < -\frac{\beta(u)}{\gamma} & \text{se} \quad \gamma < 0. \end{cases} \tag{3.11}$$

No sentido do teorema 3.3.1, a distribuição generalizada de Pareto é o modelo natural para a desconhecida distribuição dos excessos acima de níveis suficientemente elevados, e este facto é a visão essencial sobre o qual todo o método POT é construído. O nosso modelo para um risco X com distribuição F assume portanto que, para um certo u, a distribuição dos excessos acima deste nível pode ser encarada exactamente como tendo distribuição generalizada de Pareto com parâmetro de forma γ e parâmetro de escala $\beta(u)$ que acomoda a influência do nível u.
Esta distribuição é generalizada no sentido em que absorve algumas distribuições sob uma forma paramétrica comum. Em particular, existe uma estreita relação entre as distribuições GEV em (3.3) e a GPD associada aos excessos:

$$G_{\gamma,1}(y) = 1 + \log(H_\gamma(y)) \tag{3.12}$$

para todo o $y \in \mathbb{R}$ tal que $1 + \log(H_\gamma(y)) > 0$.
Esta característica explica o porquê das funções densidade de probabilidade da GPD e da GEV possuírem caudas extremas assimptoticamente equivalentes. Importante é que, para um determinado nível de u, os parâmetros γ, a_n e b_n da GEV determinam os valores γ e $\beta(u)$ da GPD. Em particular, o índice de cauda γ é o mesmo para a GEV e GPD e independem do nível fixado u. Além disso pode-se inferir em função desta relação entre a GPD e a GEV, que a GPD possui três tipos de distribuições condicionadas ao parâmetro γ.
Se $\gamma = 0$, $G_{0,1}$ corresponde à distribuição Exponencial e pertence ao domínio de atracção Gumbel; se $\gamma > 0$, $G_{\gamma,1}$, é uma distribuição de Pareto e pertence ao domínio de atracção da Fréchet; se $\gamma < 0$, $G_{\gamma,1}$ é do tipo Beta e pertence ao domínio Weibull de atracção para máximos.
O segundo caso, envolvendo distribuições do tipo Pareto, é o mais relevante para os propósitos da avaliação do risco uma vez que trata precisamente de distribuições de cauda pesada.
É útil observar que a função distribuição dos excessos (3.10) pode ser escrita em termos da função F como

$$F_u(y) = \frac{P\{u < X \leq u+y\}}{1-P\{X<u\}} = \frac{F(u+y)-F(u)}{1-F(u)} = 1 - \frac{1-F(u+y)}{1-F(u)},\ y > 0$$

Ao definir $x = u + y$ e aplicando o teorema de Balkema e de Haan, tem-se o modelo GPD inserido na f.d. da cauda de F:

$$F_u(y) \approx G_{\gamma,\beta(u)}(y),$$

$$F_u(y) = \frac{F(u+y) - F(u)}{1 - F(u)} \Leftrightarrow F(x) = F(u) + F_u(y)(1 - F(u)) \Leftrightarrow$$
$$F(x) = F(u) + F_u(x-u)(1 - F(u)).$$

com u suficientemente elevado.

$$1 - F_u(y) = \frac{1 - F(u+y)}{1 - F(u)} \Leftrightarrow 1 - F(x) = (1 - F_u(y))(1 - F_u) \Leftrightarrow$$
$$1 - F(x) = (1 - F_u(x-u))(1 - F(u))$$

e portanto

$$F(x) = (1 - F(u))G_{\gamma,\,\beta(u)}(x-u) + F(u), \quad x > u.$$

A expressão anterior mostra que podemos avançar para uma interpretação do modelo em termos da cauda da distribuição F subjacente, para $x > u$, bastando substituir F pela sua imagem estatística F_n (f.d. empírica).

Capítulo 4

Métodos de Estimação

Supondo que temos realizações $(x_1, x_2, ..., x_n)$, é imperativo estimar γ e $\beta(u)$, escolhendo um valor de u suficientemente elevado.
Seja k o número de observações $(x_1, x_2, ..., x_n)$ que excedem o nível u. A GPD é o modelo adequado para ajustar aos k excessos. Existem diversos métodos disponíveis de estimação dos parâmetros γ e $\beta(u)$ da distribuição generalizada de Pareto que podem ser divididos em duas categorias: métodos paramétricos, pressupondo que os excessos sejam realizações de uma amostra aleatória proveniente da GPD, e estimando γ e $\beta(u)$ pelo método da máxima verosimilhança; métodos semi-paramétricos, sem suposições quanto a natureza da distribuição de probabilidade amostral. Alguns dos estimadores semi-paramétricos mais conhecidos são os estimadores de Hill (1975), de Pickands (1975) e de Dekkers, Einmahl e de Haan (1989), já apresentados no capítulo 3.

4.1 Estimação paramétrica: Estimação de Máxima Verosimilhança

O método da máxima verosimilhança é aplicado no pressuposto de que os excessos acima de determinado nível u seguem distribuição generalizada de Pareto, com f.d. G_{γ,β_u} definida em (3.11). Esta é a forma paramétrica de analisar valores extremos. Neste elenco, o logaritmo da função de verosimilhança da amostra dos excessos $W_i := X_i - u | X_i > u,\ i = 1, ..., k_u$ é dado por

$$\pounds(\gamma,\ \beta_u;\ w_1,\ w_2,\ ...,\ w_{k_u}) = \sum_{i=1}^{k_u} \log g(w_i;\ \gamma,\ \beta_u),$$

onde $g(x;\ \gamma,\ \beta_u) = \frac{\partial}{\partial x} G_{\gamma,\ \beta_u}(x)$ denota a f.d.p. da generalizada de Pareto.
O estimador de máxima verosimilhança de $(\gamma,\ \beta_u)$ verifica as equações

$$\begin{cases} \frac{1}{k_u} \sum_{i=1}^{k_u} \log\left(1 + \frac{\gamma}{\beta_u} w_i\right) = \gamma \\ \frac{1}{k_u} \sum_{i=1}^{k_u} \left(1 + \frac{\gamma}{\beta_u} w_i\right)^{-1} = \frac{1}{1+\gamma}. \end{cases}$$

Portanto o estimador de máxima verosimilhança para $\gamma \in \mathbb{R}$ não tem uma forma fechada explícita. Trata-se no entanto de um estimador consistente para $\gamma > -1$ e assimptoticamente normal quando $\gamma > -1/2$. Neste caso, sob condições não muito restritivas relacionadas com a ordem de convergência de k_u, os estimadores resultantes verificam a seguinte convergência em distribuição:

$$\sqrt{k_u}\left(\hat{\gamma}-\gamma,\ \frac{\hat{\beta}_{(u)}}{\beta_{(u)}}-1\right) \xrightarrow{d} N\left(\mathbf{0},\sum\right), \qquad k_u \to \infty,$$

com matriz de covariâncias dada por

$$\sum = (1+\gamma)\begin{pmatrix} 1+\gamma & 1 \\ 1 & 2 \end{pmatrix}.$$

Na abordagem paramétrica a determinação do nível óptimo de u é essencial. Ao contrário do que acontece na abordagem semi-paramétrica, ainda não existe nenhum método adaptativo para a escolha do nível óptimo de u, aceite de uma forma generalizada. A escolha do nível de u constitui um dos problemas da modelização da GPD.
Na determinação do nível de u, é necessário ter em conta o seguinte dilema:

- Uma escolha do nível u demasiado elevado pode conduzir a uma maior variância nas estimativas, na medida em que o número de observações que excedem u é reduzido;
- ao passo que, uma escolha do nível de u demasiado baixo, pode levar a um maior vies, além de não se poder aplicar o teorema 3.3.1.

A escolha do nível u é basicamente um compromisso entre escolher um nível suficientemente elevado para que o teorema assintótico 3.3.1 possa ser considerado exacto, ou seja que a função dos excessos $F_u(y)$ convirja em probabilidade para a distribuição generalizada de Pareto, e escolher um nível suficientemente baixo para que tenhamos observações suficientes para estimar os parâmetros associados àquela distribuição. Alguns autores propõem uma forma de determinar o nível de u, que consiste na escolha de um valor de X, à direita do qual, a função de excesso médio empírica se assemelha a uma função linear. Existem, no entanto, situações em que o nível óptimo não é evidente, sendo vários os níveis aceitáveis.

4.2 Estimação Semi-paramétrica

Contrariamente ao que acontece na estimação paramétrica, em que se admite uma distribuição predeterminada para os extremos, (uma GPD ou GEV), na abordagem semi-paramétrica, a única suposição de base é a de que a distribuição F subjacente aos dados pertence ao domínio de atracção da GEV ($F \in D(H_\gamma),\ \gamma \in \mathbb{R}$). Neste contexto, são utilizados estimadores semi-paramétricos do índice de valores extremos que determinam o peso da cauda da f.d. F, para qualquer inferência que diga respeito a cauda da distribuição

e que pode ser baseada nas k observações acima de um nível aleatório $X_{n-k:n}$ (PORT). Relembramos que o domínio de atracção Fréchet contém distribuições com cauda polinomial negativa, com limite superior do suporte x^F infinito, enquanto que as do domínio Weibull são de cauda curta e x^F finito. O caso intermédio de f.d.'s no domínio Gumbel abrange simultâneamente x^F finito ou infinito. Exemplos mais relevantes de estimadores semi-paramétricos do índice de valores extremos, são os estimadores de Hill (1975), de Pickands (1975), dos momentos (Dekkers, Einmahl e de Haan (1989)) já apresentados na secção 3.2.3 do capítulo 3.
Seguem-se agora algumas propriedades destes estimadores, nomeadamente sobre consistência e normalidade assimptótica. Sob condições gerais, não muito restritivas envolvendo a ordem de convergência da sucessão intermédia $k = k_n$ tal que $k_n \to \infty$ e $k_n/n \to 0$, $n \to \infty$, tem-se:

Teorema 4.2.1 *(Propriedades do estimador de Hill (1975))*

Seja $\widehat{\gamma}^H = \widehat{\gamma}^H_{k,n}$ o estimador de Hill, definido em (3.6).

1. (*consistência*) Seja X_1, X_2, ... uma sucessão de variáveis aleatórias i.i.d. com uma mesma função distribuição F. Suponhamos que $F \in D(H_\gamma)$ com $\gamma > 0$. Se a sucessão intermédia $k = k_n$ é tal que $k_n \to \infty$, $k_n/n \to 0$, quando $n \to \infty$, então,

$$\widehat{\gamma}^H \xrightarrow{p} \gamma$$

2. (*normalidade assimptótica*) Seja X_1, X_2, ... uma sucessão de variáveis aleatórias i.i.d. com uma mesma função distribuição F, e se $k = k_n \to \infty$ a uma velocidade apropriada, então,

$$\sqrt{k}(\widehat{\gamma}^H - \gamma) \xrightarrow{d} N\left(0,\, \gamma^2\right).$$

Teorema 4.2.2 *(Propriedades do estimador dos Momentos -*

(Dekkers, Einmahl, e de Haan) (1989))

Seja $\widehat{\gamma}^M = \widehat{\gamma}^M_{k,n}$ o estimador dos Momentos, definido em (3.8).

1. (*consistência*) Seja X_1, X_2, ... uma sucessão de variáveis aleatórias i.i.d. com uma mesma função distribuição F. Suponhamos que $F \in D(H_\gamma)$ com $\gamma \in \mathbb{R}$ e $x^F > 0$. Se a sucessão intermédia $k = k_n$ é tal que $k_n \to \infty$, $k_n/n \to 0$, quando $n \to \infty$, então,

$$\widehat{\gamma}^M \xrightarrow{p} \gamma$$

2. (*normalidade assimptótica*) Seja X_1, X_2, ... uma sucessão de variáveis aleatórias i.i.d. com uma mesma função distribuição F com $x^F > 0$, e se $k = k_n \to \infty$ a uma velocidade apropriada, então,

$$\sqrt{k}(\widehat{\gamma}^M - \gamma) \xrightarrow{d} N\left(0,\, var^M_\gamma\right),$$

onde

$$var_\gamma^M := \begin{cases} \gamma^2+1, & \gamma \leq 0, \\ \frac{(1-\gamma)^2(1-2\gamma)(1-\gamma+6\gamma^2)}{(1-3\gamma)(1-4\gamma)}, & \gamma \leq 0. \end{cases}$$

Teorema 4.2.3 *(Propriedades do estimador de Pickands (1975))*

Seja $\widehat{\gamma}^P = \widehat{\gamma}^P_{k,n}$ o estimador de Pickands, definido em (3.7).

1. (*consistência*) Seja X_1, X_2, ... uma sucessão de variáveis aleatórias i.i.d. com uma mesma função distribuição F. Suponhamos que $F \in D(H_\gamma)$ com $\gamma \in \mathbb{R}$. Se a sucessão intermédia $k = k_n$ é tal que $k_n \to \infty$, $k_n/n \to 0$, quando $n \to \infty$, então,

$$\widehat{\gamma}^P \xrightarrow{p} \gamma$$

2. (*normalidade assimptótica*) Seja X_1, X_2, ... uma sucessão de variáveis aleatórias i.i.d. com uma mesma função distribuição F e se $k = k_n \to \infty$ a uma velocidade apropriada, então,

$$\sqrt{k}(\widehat{\gamma}^P - \gamma) \xrightarrow{d} N\left(0,\, var_\gamma^P\right),$$

onde

$$var_\gamma^P := \begin{cases} \frac{\gamma^2(2^{2\gamma+1}+1)}{4(\log 2)^2(2^\gamma-1)^2}, & \gamma \neq 0, \\ \frac{3}{4(\log 2)^4}, & \gamma = 0. \end{cases}$$

Teorema 4.2.4 *(Propriedades do estimador Mixed-Moment, Fraga Alves* et al.*, (2009))*

Seja $\widehat{\gamma}^{MM} = \widehat{\gamma}^{MM}_n(k)$ o estimador Mixed-Moment, definido em (3.9).

1. (*consistência*) Seja X_1, X_2, ... uma sucessão de variáveis aleatórias i.i.d. com uma mesma função distribuição F. Suponhamos que $F \in D(H_\gamma)$ com $\gamma \in \mathbb{R}$. Se a sucessão intermédia $k = k_n$ é tal que $k_n \to \infty$, $k_n/n \to 0$, quando $n \to \infty$, então,

$$\widehat{\gamma}^{MM}_n(k) \xrightarrow{p} \gamma$$

2. (*normalidade assimptótica*) Seja X_1, X_2, ... uma sucessão de variáveis aleatórias i.i.d. com uma mesma função distribuição F e se $k = k_n \to \infty$ a uma velocidade apropriada, então,

$$\sqrt{k}(\widehat{\gamma}^{MM}_n(k) - \gamma) \xrightarrow{d} N\left(0,\, var_\gamma^{MM}\right),$$

onde

$$var_\gamma^{MM} = var^{MM}(\gamma) := \begin{cases} Var_\varphi(\gamma) & \text{se} \quad \gamma \geq 0 \\ (1-2\gamma)^4 Var_\varphi(\gamma) & \text{se} \quad \gamma < 0. \end{cases}$$

e

$$Var_\varphi(\gamma) := \begin{cases} (1+\gamma)^2 & \text{se} \quad \gamma \geq 0 \\ \frac{(1-\gamma)^2(6\gamma^2-\gamma+1)}{(1-2\gamma)^3(1-3\gamma)(1-4\gamma)} & \text{se} \quad \gamma < 0. \end{cases}$$

Capítulo 5

Aplicação a Dados de Precipitação

5.1 Descrição dos Dados

As amostras utilizadas correspondem a níveis de precipitação (em *mm*), coleccionados durante os períodos 1876 a 2007 para a cidade de *Berlim* na Alemanha, e 1906 a 2007 para a cidade de *de Bilt* na Holanda. A fonte dos dados é European Climate Assessment (ECA) and Dataset (http://eca.knmi.nl/).
Os valores de precipitação coleccionados para as duas cidades são diários e começa-se por adoptar um procedimento que os torna independentes. A independência dos valores de precipitação decorre do facto dos dados terem sido desagrupados, para que máximos diários não ocorram em dias consecutivos. A dimensão das amostras resultantes deste processo é de $n_1 = 9210$ e $n_2 = 7139$ para *Berlim* e *de Bilt* respectivamente. Assume-se que estes valores também são identicamente distribuídos, visto que consideramos que estes fenómenos de precipitação são gerados pelo mesmo processo físico.
As estatísticas básicas referentes aos dados estão na tabela 5.1. A diferença entre os resultados da precipitação média diária e mediana das amostras de *Berlim* e *de Bilt*, remete para uma assimetria positiva ou à direita das distribuições subjacentes aos dados das duas amostras, o que sugere para ambas, distribuições de cauda mais pesada do que a normal. Esta constatação indica a rejeição da hipótese de normalidade dos dados das amostras em estudo.
A aplicação prática das metodologias no âmbito da estatística de extremos, será realizada

	n	Mínimo	Máximo	Média	Mediana	Desvio-padrão
Berlim	9210	1	1247	61.96	43	64.19
de Bilt	7139	4	662	79.76	60.0	65.43

Tabela 5.1: Estatísticas Descritivas: *Berlim* e *de Bilt*

em quatro fases distintas.

5.2 Fase I

Numa abordagem preliminar e após ordenar as duas amostras por ordem decrescente de realizações dos X_i, procedeu-se numa perspectiva exploratória, ao cálculo de estimativas grosseiras para um valor de precipitação que é excedido com uma pequena probabilidade. Tais estimativas são quantis, medidas de risco ou valores de risco (VaR_p), adoptando os seguintes valores: $p^* = 0.01$; $p^* = 0.05$; $p^* = 0.001$; $p^* = 0.0001$.
O procedimento consiste primeiramente em calcular a ordem do quantil para aqueles valores de probabilidade utilizando a fórmula $n - [np^*] + 1$, onde $[x]$ designa parte inteira de x. De seguida, passa-se a identificar nas amostras referentes a *Berlim* e *de Bilt*, os quantis ou medidas de risco $X_{n-[np*]+1:n}$, que correspondem aos valores de precipitação que serão excedidos para aqueles valores pequenos de probabilidade.
Assim sendo obteve-se a seguinte tabela de quantis empíricos sobre os valores de precipitação para as duas cidades em estudo,

	$p^* = 0.01$	$p^* = 0.05$	$p^* = 0.001$	$p^* = 0.0001$	$x_{n:n}$
Berlim	314	177	588	1247	1247
de Bilt	322	207	503	662	662

Tabela 5.2: Estimativas para níveis de precipitação elevada associadas à probabilidade p^*

Pelos resultados constantes na tabela anterior é possível constatar por exemplo, que em apenas um dia em cada cem ($p^* = 0.01$), em média, poderá ter ocorrido um valor de precipitação superior a 314 *mm* em *Berlim* e superior a 322 *mm* em *de Bilt*. Os quantis de ordem $p^* = 0.0001$, coincidem com o máximo de cada uma das amostras.

5.3 Fase II

Numa segunda fase e, tendo em conta que poderão ocorrer acontecimentos extremos tais como ocorrência de valores de precipitação anormalmente elevados, pretende-se agora obter um quantil ou uma medida de risco que expresse um nível de precipitação que seja ultrapassado em média, uma vez em cada mil anos ou uma vez em cada dez mil anos, por exemplo.
Parte-se, para o efeito da seguinte definição: "Considera-se que há dias anormalmente chuvosos quando se verificarem ocorrências de mais de 80 *mm* de precipitação". As figuras 5.1 e 5.2 mostram níveis de precipitação anormalmente elevados *i.e.* acima dos 80 *mm* para *Berlim* e *de Bilt*.
Em m anos de registo existirão para cada cidade, K_u acontecimentos de precipitação anormalmente elevada. A partir desta definição, constatou-se que ocorreram ao longo dos períodos de observação considerados para as duas cidades, 2161 acontecimentos de precipitação anormalmente elevada em *Berlim* e 2577 em *de Bilt*. Por ano observou-se em média $\frac{K_u}{m}$ acontecimentos deste tipo para cada cidade. Esses resultados constam da tabela 5.3.

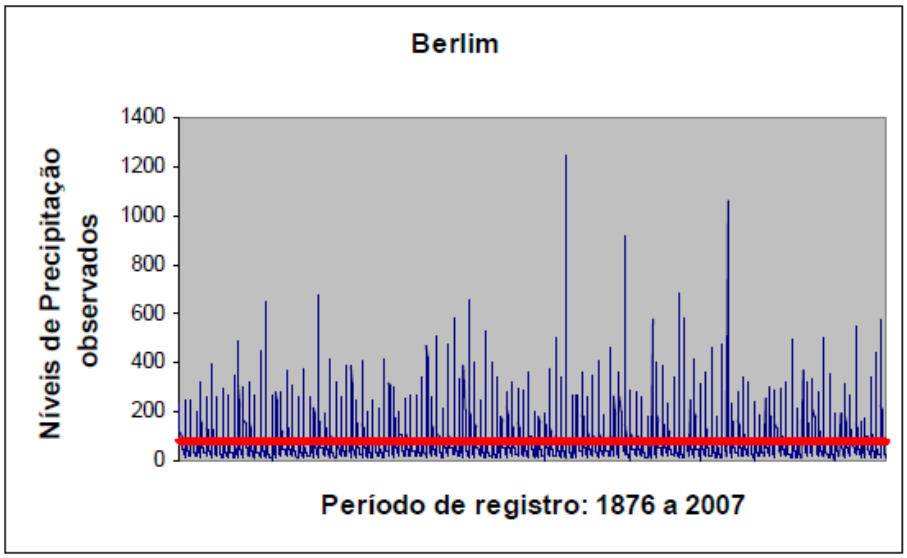

Figura 5.1: Níveis de precipitação acima dos 80 *mm* para *Berlim*.

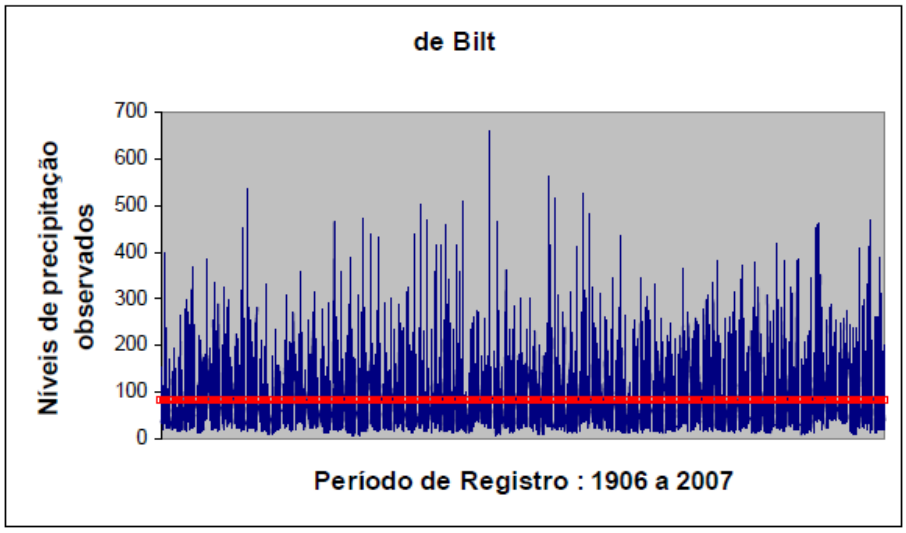

Figura 5.2: Níveis de precipitação acima dos 80 *mm* para *de Bilt*.

	Prec. anormal. elevada (K_u)	Anos de registo (m)	Taxa média anual ($\frac{K_u}{m}$)
Berlim	2161	132	≈ 16
de Bilt	2577	102	≈ 25

Tabela 5.3: Taxa média anual de ocorrência de precipitação anormalmente elevada

O objectivo é, com base nos dados de precipitação (em *mm*), assumidos i.i.d. das duas amostras, calcular um quantil X_{p^*}, ou seja um nível de precipitação que é ultrapassado com uma certa probabilidade num ano ao acaso.
Pretende-se calcular então

$$x_{p^*} : P\{X > x_{p^*}\} = p^*$$

p^* assume os valores 10^{-3} e 10^{-4} que correspondem às probabilidades da variável precipitação ultrapassar esse tal nível x_{p^*} que se pretende calcular, num ano ao acaso.
Porque estamos interessados em ocorrências de dias anormalmente chuvosos, para determinar esse quantil, os valores de p^* serão convertidos em probabilidades em termos de ocorrência de precipitação anormalmente elevada. Desta forma a probabilidade será assim dada:

$$p = \frac{m}{K_u} \times p^*$$

com

$$p^* = 10^{-3} \qquad e \qquad p^* = 10^{-4}$$

o que dará origem às probabilidades:

$$p_1 = \frac{m}{K_u} \times 10^{-3} \qquad e \qquad p_2 = \frac{m}{K_u} \times 10^{-4} \tag{5.1}$$

Os valores de $\frac{m}{K_u}$ são então convertidos numa proporção relativamente a acontecimentos daquele tipo, *i.e.*, a ocorrência por dias anormalmente chuvosos.
A tabela seguinte indica os quantis empíricos sobre os níveis de precipitação que poderão ser excedidos para os valores de probabilidade p_1 e p_2, para as duas cidades em estudo.

	K_u	m	$p_1 = \frac{m}{K_u} \times 10^{-3}$	$p_2 = \frac{m}{K_u} \times 10^{-4}$	$X_{n-[np_1]+1}$	$X_{n-[np_2]+1}$
Berlim	2161	132	6.10828×10^{-5}	6.10828×10^{-6}	1247	1247
de Bilt	2577	102	3.95809×10^{-5}	3.95809×10^{-6}	662	662

Tabela 5.4: Estimativas para ocorrência de precipitação anormalmente elevada num ano ao acaso, com probabilidades p_1 e p_2 definidos em (5.1)

Analisando os valores apresentados na tabela 5.4, em conjunção com a tabela 5.2, pode-se constatar que os valores 1247 e 662 correspondem aos valores máximos de precipitação observados para as cidades de *Berlim* e *de Bilt* respectivamente. Isto implica que se esteja sempre a considerar o máximo de cada amostra e portanto estamos perante estimativas pouco informativas. Em 132 anos de registo dos níveis de precipitação diária para a cidade de *Berlim*, a probabilidade de ocorrer um valor de precipitação superior a 1247 *mm* num ano ao acaso é de 6.10828×10^{-5} e 6.10828×10^{-6} . Enquanto que para *de Bilt*, em 102 anos de registo, a probabilidade de num ano ao acaso, ocorrer um valor de precipitação

superior a 662 *mm* é de 3.95809×10^{-5} e 3.95809×10^{-6}.
É natural que, para valores tão pequenos de probabilidade, os quantis ou valores de risco pretendidos coincidam com o valor máximo de cada amostra, uma vez que nos afastamos da sua parte central. Como estamos limitados pela dimensão da amostra (e porque a distribuição Normal não consegue modelar tais fenómenos), surge a necessidade de métodos de extrapolação para além dos valores de precipitação observados, tal como o(s) que provém da Teoria de Valores Extremos.

5.4 Fase III

Nesta fase expõe-se o método de estimação de características envolvendo a cauda direita da f.d. subjacente F, conhecido como método dos excessos acima de um nível aleatório (PORT - *Peaks Over Random Threshold*), da Teoria Valores Extremos. Esta metodologia utiliza as k maiores observações da amostra, a fim de estimar a forma da cauda da distribuição subjacente aos dados. Qualquer inferência daqui resultante irá reportar-se às duas cidades em estudo.
Para determinarmos qual das três distribuições de valores extremos (Weibull, Gumbel ou Fréchet) melhor se relaciona com as observações de topo das amostras de *Berlim* e *de Bilt*, procedemos a uma análise gráfica preliminar, via QQ-Plot, ou gráfico de análise quantílica, tomando como referência a distribuição exponencial, figura 5.3. A razão que preside ao QQ-plot da exponencial é a relação entre as distribuições GEV para máximos e GPD para as maiores observações, referida em (3.12). Se os dados não apresentarem desvios muito grandes da linha padrão do gráfico QQ-plot exponencial, então é razoável admitir que estes são do tipo exponencial na cauda ou provém de uma distribuição F no domínio de atracção Gumbel, (com $\gamma = 0$). Caso contrário, a distribuição dos dados será do tipo Fréchet, ($\gamma > 0$), ou do tipo Weibull, ($\gamma < 0$) .
Podemos verificar que em ambos os casos ocorrem alguns desvios, porém os mais acentuados ocorrem nos dados de *Berlim*. Este facto sugere que a f.d. F subjacente aos dados de precipitação para *Berlim* é de cauda pesada e está no domínio de atracção Fréchet, $D(H_\gamma)$, $\gamma > 0$.

Para averiguarmos se as caudas das distribuições dos níveis diários de precipitação para as duas cidades possuem a mesma forma, serão construídos os gráficos da função de excesso médio para os dados de precipitação das duas cidades em estudo.
A função do excesso médio é assim definida:

$$e(u) := E(X - u | X > u)$$

O excesso médio empírico dá alguma ideia sobre o comportamento da cauda de F:

$$\widehat{e}_n(x_{n-k:n}) = \frac{1}{k} \sum_{i=1}^{k} X_{n-i+1:n} - X_{n-k:n}$$

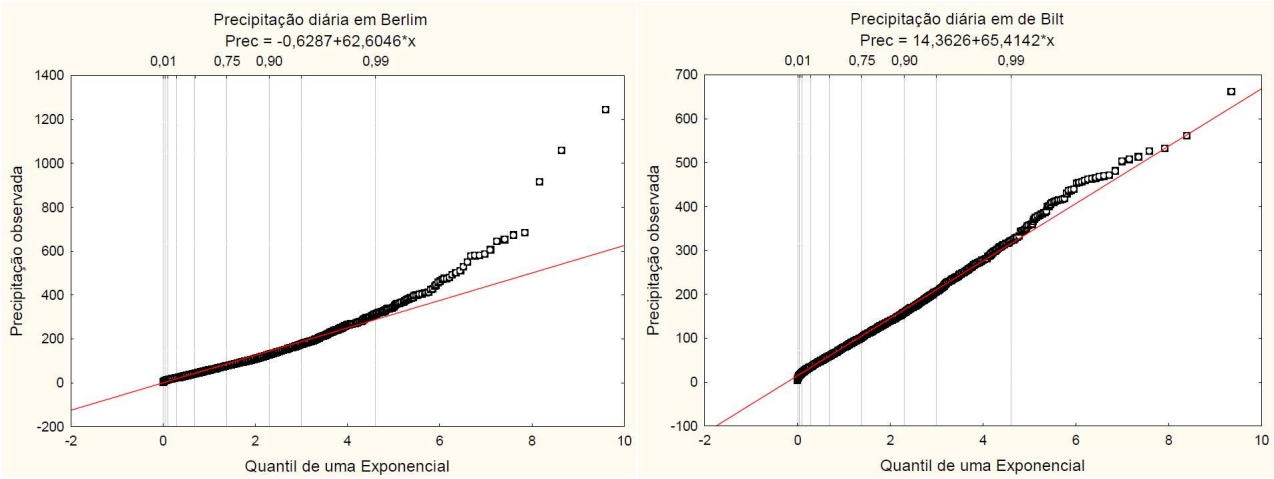

Figura 5.3: Gráfico Quantílico em relação à distribuição Exponencial para *Berlim* (à esquerda) e *de Bilt* (à direita).

O seu gráfico pode ser construído de duas formas alternativas: $e_{k,n}$ versus k ou $e_{k,n}$ versus $x_{n-k:n}$:

$$\{k,\ x_{n-k:n},\ e_{k,n} := \widehat{e}_n(x_{n-k:n}),\quad 5 \leq k \leq n-1\}$$

Porque interessa analisar a cauda da f.d. F, vamos estudar o comportamento dos valores de $\widehat{e}_{k,n}$ para valores decrescentes de k ou para valores crescentes de $x_{n-k:n}$. As figuras 5.4 e 5.5 dão conta disso mesmo.
Quando consideramos a forma da função de excesso médio, a distribuição exponencial assume um papel central. No caso de *Berlim*, o comportamento constante torna-se evidente a partir da figura 5.4. Para os menores valores de k, a função de excesso médio em ultima instância aumenta, e para os maiores valores de $x_{n-k:n}$, ela também aumenta. Este comportamento indica que a distribuição subjacente aos dados de precipitação para *Berlim* é de cauda mais pesada do que a distribuição exponencial. Por outro lado, os dados de precipitação de *de Bilt* mostram um exemplo de distribuições de cauda ligeiramente mais leve do que a da exponencial, figura 5.5.

No âmbito da selecção de domínios de atracção e num contexto de escolha estatística de condições de cauda para valores extremos, seguindo uma abordagem semi-paramétrica, baseada em estatísticas invariantes perante alterações de localização/escala, baseada nos excessos acima de um nível aleatório, são realizados testes de hipóteses sobre a que domínio de atracção pertence a função distribuição subjacente F.
Portanto, qualquer inferência respeitante à cauda da distribuição subjacente F pode ser baseada nas k observações acima de um nível aleatório $x_{n-k:n}$. Este nível aleatório corresponde a uma estatística ordinal intermédia, tomando em consideração a informação crescente acerca da cauda direita, disponibilizada pelo topo da amostra.

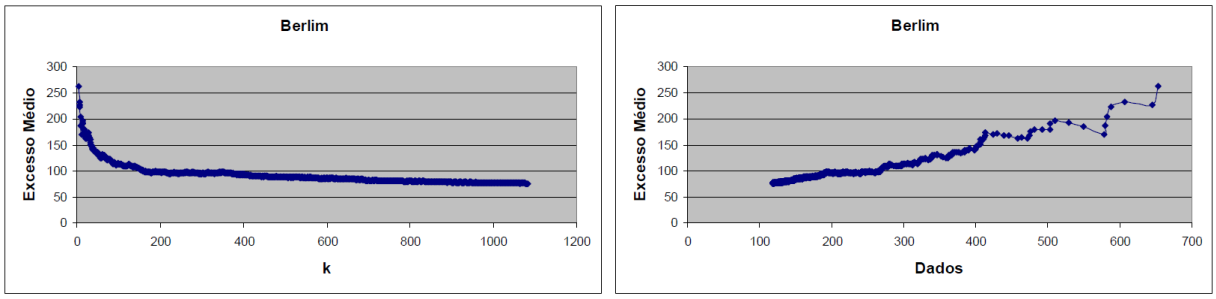

Figura 5.4: Função de excesso médio empírico para os dados de precipitação de *Berlim*: $e_{k,n}$ versus k (à esquerda) e $e_{k,n}$ versus $x_{n-k:n}$ (à direita).

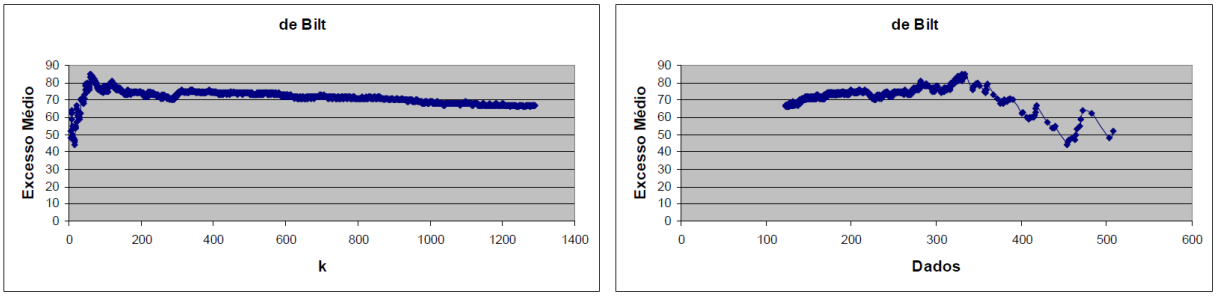

Figura 5.5: Função de excesso médio empírico para os dados de precipitação de *de Bilt*: $e_{k,n}$ versus k (à esquerda) e $e_{k,n}$ versus $x_{n-k:n}$ (à direita).

Na abordagem semi-paramétrica, a única suposição feita é a de que a condição de valores extremos do teorema 3.2.1 (pág. 12) é satisfeita, ou seja, $F \in D(H_\gamma)$, para algum γ real. Neste contexto, o índice de valores extremos é o parâmetro de destaque, pois, em ambas as classes de distribuições GEV e GP, ele determina a forma da cauda da função distribuição subjacente F. Recordemos que o domínio de atracção Fréchet ($\gamma > 0$) contém distribuições com cauda polinomial negativa, com limite superior do suporte x^F infinito, enquanto que as do domínio Weibull ($\gamma < 0$) são de cauda curta e x^F finito. O caso intermédio de f.d.'s no domínio Gumbel, em que $\gamma = 0$ é um valor de referência, abrange simultaneamente x^F finito ou infinito. Em muitas ciências aplicadas onde os extremos são relevantes, o caso mais simples de $\gamma = 0$ é assumido e tendo por base este pressuposto, características extremas tais como probabilidades de excedência ou períodos de retorno são facilmente estimados.

Separar os procedimentos de inferência estatística de acordo com o domínio de atracção para máximos mais conveniente para a distribuição da amostra, tornou-se uma prática usual ao longo da literatura da especialidade, seguindo quer abordagens paramétricas quer semi-paramétricas. Metodologias para testar o domínio Gumbel contra os domínios de atracção para máximos Fréchet ou Weibull têm sido de grande utilidade.

O problema de escolha estatística neste contexto semi-paramétrico pode ser explicitado através de

$$H_0 : F \in D(H_0) \quad versus \quad H_1 : F \in D(H_\gamma)_{\gamma \neq 0} \tag{5.2}$$

ou contra alternativas unilaterais $F \in D(H_\gamma)_{\gamma<0}$ (Domínio Weibull) ou $F \in D(H_\gamma)_{\gamma>0}$ (Domínio Fréchet).

Testes dirigidos a este problema surgiram na literatura a partir dos trabalhos de Galambos (1982) e Castillo et al. (1989). Outros procedimentos para escolha estatística de domínios de atracção para máximos podem ser igualmente sugeridas pelos trabalhos de Hasofer e Wang (1992), Fraga Alves e Gomes (1996), Fraga Alves (1999), Segers e Teugels 2000) entre outros.

Mais recentemente, Neves et al. (2006) e Neves e Fraga Alves (2007), introduziram dois procedimentos de teste que são baseados nas observações da amostra acima de um nível aleatório. Mais especificamente, nas últimas duas referências, os procedimentos para o teste 5.2 baseiam-se nos k excessos acima da $(n-k)$-ésima estatística de ordem intermédia ascendente $x_{n-k:n}$, onde $k = k_n$ é tal que $k \to \infty$ e $k/n = 0$, a medida que $n \to \infty$. Este contexto representa alguma semelhança com a abordagem POT, mas em que $x_{n-k:n}$ assume o papel do nível determinístico u, o que justifica a designação PORT.

Na sequência de uma abordagem semi-paramétrica, apoiada em conceitos da teoria de variação regular, Neves e Alves Fraga (2007), reformularam as propriedades assintóticas das estatísticas de teste de Hasofer e Wang (indicada abaixo por $W_n(k)$), no caso em que $k = k_n$ se comporta como uma sequência intermédia em vez de permanecer fixo, enquanto o tamanho da amostra aumenta (que era o caso abrangido por Hasofer e Wang, 1992). No processo, Grenwood, um novo tipo de teste estatístico $G_n(k)$ (cf. Greenwood, 1946) revela-se útil na avaliação da presença de distribuições de cauda pesada.

Além disso, motivado por diferenças na contribuição relativa do máximo para a soma dos k excessos acima do nível aleatório em diferentes caudas pesadas, um teste estatístico complementar $R_n(k)$ foi introduzido por Neves *et al.* (2006), com o objectivo de discernir entre os três domínios de atracção para máximos.

Sob a hipótese nula de que a f.d. F subjacente aos dados pertence ao domínio Gumbel, o comportamento limite das estatísticas seguintes permite a sua utilização como estatísticas de testes na selecção de domínios de atracção:

$$[Ratio]\ R_n(k) := \frac{X_{n:n} - X_{n-k:n}}{\frac{1}{k}\sum_{i=1}^{k}\left(X_{n-i+1:n} - X_{n-k:n}\right)} - \log k \xrightarrow[n\to\infty]{d} \Lambda, \tag{5.3}$$

onde Λ é uma variável aleatória Gumbel.

A estatística de teste de Hasofer-Wang (HW) é dada por:

$$W_n(k) := \frac{k^{-1}\Big(\sum_{i=1}^{k} X_{n-i+1:n} - X_{n-k:n}\Big)^2}{\sum_{i=1}^{k}\big(X_{n-i+1:n} - X_{n-k:n}\big)^2 - k^{-1}\Big(\sum_{i=1}^{k} X_{n-i+1:n} - X_{n-k:n}\Big)^2}$$

$$\sqrt{k/4}\,(kW_n(k) - 1) \xrightarrow[n\to\infty]{d} N(0,\ 1) \tag{5.4}$$

As regiões críticas para o teste bilateral (5.2), com nível de significância α, são dados por $V_n(k) < v_{\alpha/2}$ ou $V_n(k) > v_{1-\alpha/2}$, onde V tem de ser convenientemente substituída por R ou W e v_ε denota o ε-quantil da distribuição limite correspondente (Gumbel ou Normal).

As figuras 5.6 e 5.7 apresentam as estatísticas de teste (5.3) e (5.4) e suas estimativas, quantil assintótico 95% sob a hipótese nula em (5.2).
Na figura 5.6 todos os testes apontam para a rejeição da condição de valores extremos. Em ambos os casos a condição de valores extremos é rejeitada para quase todos os valores de k, com excepção de alguns valores muito baixos. Os valores observados da estatística de teste (5.3), estão quase sempre acima da linha (a ponteado) determinada pelos pontos críticos correspondentes ao nível de significância $\alpha = 5\%$, *i.e.*, $z_{0.025} = -\log(-\log(0.025)) = -1.30532$ e $z_{0.975} = -\log(-\log(0.975)) = 3.67616$. Para o teste (5.4), os valores observados da estatística de teste estão quase sempre abaixo da linha (a ponteado) determinada pelos pontos críticos correspondentes ao nível de significância $\alpha = 5\%$, nomeadamente $z_{0.025} = -1.96$ e $z_{0.975} = 1.96$. Isto significa que, para quase todos os valores de k, os testes apontam para a rejeição da hipótese nula, de que a função distribuição F subjacente aos dados de *Berlim*, pertence ao domínio Gumbel. Perante os resultados destes testes apenas foram encontradas evidências nos dados de que F poderá pertencer à qualquer domínio de atracção, excepto o domínio Gumbel.

Na figura 5.7, os valores observados das estatísticas de teste (5.3) e (5.4) em função de k, encontram-se entre as linhas determinadas pelos pontos críticos correspondentes ao nível de significância $\alpha = 5\%$, *i.e.* $z_{0.025} = -\log(-\log(0.025)) = -1.30532$ e $z_{0.975} = -\log(-\log(0.975)) = 3.67616$, para o teste (5.3) e $z_{0.025} = -1.96$ e $z_{0.975} = 1.96$ para o teste (5.4). A hipótese de que a distribuição subjacente aos dados de *de Bilt* pertence ao domínio de atracção Gumbel não é rejeitada para quase todo o k.
De agora em diante restringimos a atenção aos valores diários de precipitação acima de 80 *mm*. Esperamos portanto que este nível represente o começo da cauda, *i.e.*, o início dos eventos de precipitação extrema ou seja da ocorrência de precipitação anormalmente elevada. Ao considerarmos em cada amostra, somente os valores de precipitação que ul-

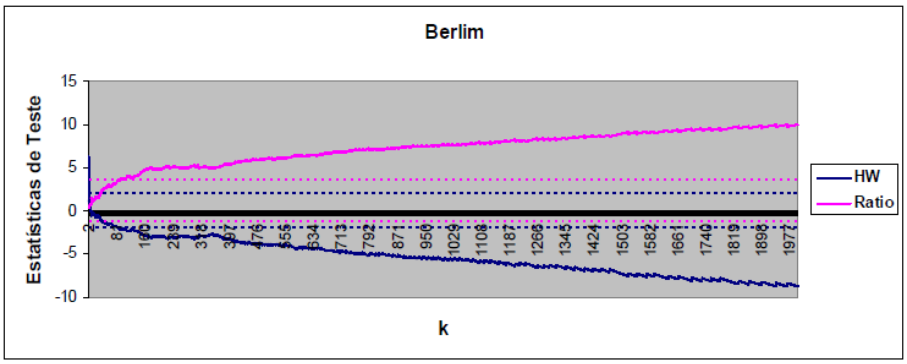

Figura 5.6: Trajectórias das estatísticas de teste (5.3) e (5.4).

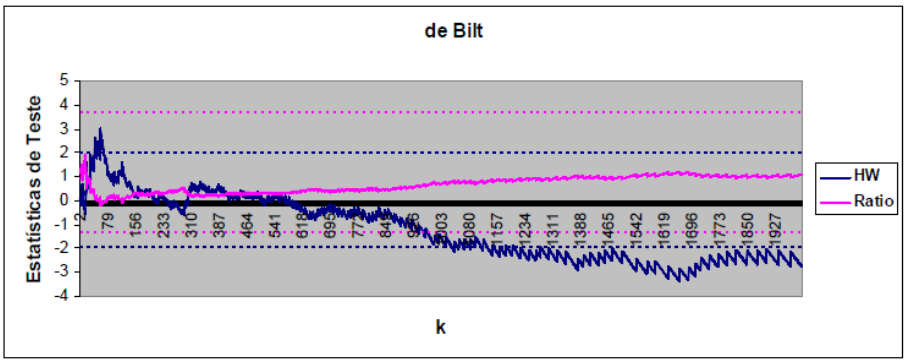

Figura 5.7: Trajectórias das estatísticas de teste (5.3) e (5.4).

trapassam os 80 *mm*, estas reduzem-se a subamostras de topo de dimensão $K_u^1 = 2161$ e $K_u^2 = 2577$ excedências para *Berlim* e *de Bilt*, respectivamente. Consideremos ainda, a sequência de valores diários de precipitação formada pelas $k+1$ estatísticas ordinais de topo nas amostras de excedências de cada cidade. Na sequência, são determinadas as estimativas do índice de valores extremos para ambas as cidades, com o objectivo de verificar se possuem um índice de valores extremos que leve à mesma família de distribuições, bem como estimativas de quantis empíricos (níveis de precipitação de risco).
No contexto da Teoria de Valores Extremos, temos que a função distribuição F subjacente, pertence ao domínio de atracção da Generalizada de Valores Extremos (GEV), *i.e.*, $F \in D(H_\gamma)$, se e somente se existe uma função positiva $a(.)$, de variação regular $a \in RV_\gamma$ e

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \frac{x^\gamma - 1}{\gamma},$$

para todo o $x > 0$, onde U é a função quantil de cauda $U(t) := \left(\frac{1}{1-F}\right)^{\leftarrow}(t)$, $t \geq 1$ (Ver definição 1 - capítulo 2).
Dada esta condição e segundo a Teoria de Valores Extremos, os quantis empíricos sobre os valores de precipitação pretendidos serão estimados de forma alternativa, considerando as

k maiores observações de topo em cada amostra de excessos:

$$x_p^{(n)} = \widehat{U}(p) = \widehat{U}(n/k) + \widehat{a}\left(\frac{n}{k}\right) \frac{\left(p\,\frac{k}{n}\right)^{\widehat{\gamma}} - 1}{\widehat{\gamma}}, \tag{5.5}$$

onde k é o número de observações de topo a serem consideradas na cauda, acima do nível aleatório intermédio $\widehat{U}(n/k) := X_{n-k:n}$, *i.e.*, $k = k_n$ tal que $k_n \to \infty$ e $k_n/n \to 0$, $n \to \infty$. Existem na literatura várias possibilidades para $\widehat{\gamma}$ e $\widehat{a}$, (de Haan e Ferreira, (2006) e Beirlant *et al.* (2004)). Os estimadores da função de escala $a(.)$ e do índice de valores extremos (parâmetro de forma) γ, aqui adoptados e denotados por $\widehat{\gamma}$ e $\widehat{a}$, são os que derivam dos estimadores dos momentos. Além disso $p \in (0,\, 1)$ é a probabilidade do valor de precipitação ultrapassar o quantil que se pretende estimar, dado por $\left(\frac{m}{K_u} \times p^*\right)$, onde m são os anos de registo dos valores diários de precipitação para cada cidade, K_u é a amostra de excedências para cada cidade, e p^* assume os valores 10^{-3} e 10^{-4}. Teremos assim, dois valores de risco, designados por p_1 e p_2 consoante o valor assumido por p^*, tal como anteriormente definido em (5.1).
Após identificar em cada amostra os valores de precipitação superiores a 80 mm, estes são logaritmizados e considerando a soma das k-ésimas observações de topo fixado, calcula-se para cada k assumido, a média das diferenças das k maiores observações:

$$M_{k,n}^{(j)} = \frac{1}{k}\sum_{i=1}^{k}(\log(X_{n-i+1:n}) - \log(X_{n-k:n}))^j, \qquad j = 1,\, 2.$$

De seguida, são calculadas para cada cidade, as estimativas de γ e da escala $a\left(\frac{n}{k}\right)$, denotadas por $\widehat{\gamma}$ e $\widehat{a}\left(\frac{n}{k}\right)$, com base no número k de observações de topo assumido, sendo que $5 \leq k \leq \frac{K_u}{2}$. Embora existam outros estimadores do índice de valores extremos utilizou-se o primeiro estimador dos momentos em (3.8), baseado nas estatísticas de ordem de uma variável aleatória X, já apresentado no capítulo 3 (pág. 16). O estimador para a escala $a\left(\frac{n}{k}\right)$ associado a este mesmo estimador dos momentos, (Dekkers, Einmahl e de Haan, (1989)) é dado por:

$$\widehat{a}\left(\frac{n}{k}\right) = X_{n-k:n}\; M_{k,n}^{(1)}\; \frac{1}{2}\; \left(1 - \frac{(M_{k,n}^{(1)})^2}{M_{k,n}^{(2)}}\right)^{-1}.$$

Note que ao calcular a média das diferenças das k maiores observações logaritmizadas assumimos os excessos que serão utilizados na estimação dos quantis empíricos sobre os níveis de precipitação, conforme a metodologia PORT. O $X_{n-k:n}$ será portanto um valor aleatório dentro de cada amostra de excessos, que irá depender do valor k fixado.
A figura 5.8 apresenta os valores estimados de γ para as cidades em estudo. As estimativas do índice de valores extremos correspondente às cidades de *Berlim* e *de Bilt* são aproximadamente 0.25 e 0.05, respectivamente. Para a cidade de *Berlim* , a série de estimativas é composta maioritariamente por valores positivos. Vale relembrar que um valor positivo

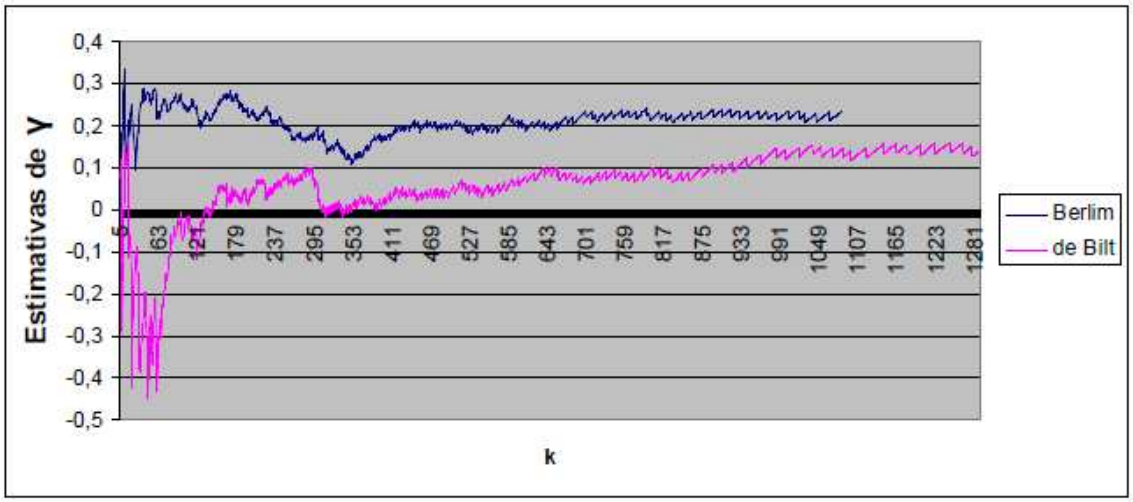

Figura 5.8: Estimativas do Índice de Valores Extremos $\widehat{\gamma}$, segundo o estimador dos momentos para diferentes valores de k: *Berlim* e *de Bilt*

de γ, indica uma distribuição de cauda pesada, pertencente ao domínio de atracção da Fréchet com $\gamma > 0$. Quando $k = 5$, $\gamma < 0$, indicando que a distribuição subjacente aos dados pertence ao domínio Weibull de atracção para máximos. Podemos encontrar patamares de alguma estabilidade para valores de k no intervalo $50 \leq k \leq 120$ e γ tendendo a estabilizar-se em torno de 0.25. A série de estimativas do índice de valores extremos para a cidade de *de Bilt*, apresenta alguma variabilidade, especialmente para os valores mais pequenos de k. No entanto, podemos identificar patamares de alguma estabilidade em $0 < \gamma \leq 0.1$ com $150 \leq k \leq 200$, em que γ tende a estabilizar-se em torno de 0.05 . Para alguns valores de k, a estimativa do índice de valores extremos é aproximadamente zero, pelo que a distribuição subjacente aos dados poderá ser confundida com uma distribuição do tipo Gumbel. Note que o domínio de atracção Gumbel encerra uma grande variedade de distribuições, desde distribuições de cauda leve com limite superior do suporte x^F finito, até distribuições de cauda moderadamente pesada, tal como a log Normal.

Finalmente, em função de cada k fixado e usando a equação (5.5), foram determinados os quantis empíricos sobre os níveis de precipitação, que poderão ser ultrapassados para valores pequenos de probabilidade, convertidos em termos de ocorrência de precipitação anormalmente elevada, através da expressão $\left(\frac{m}{K_u} \times p^*\right)$. Estes valores pequenos de probabilidade são respectivamente, ($p_1 = 6.10828 \times 10^{-5}$ *e* $p_2 = 6.10828 \times 10^{-6}$) para *Berlim* e ($p_1 = 3.95809 \times 10^{-5}$ *e* $p_2 = 3.95809 \times 10^{-6}$) para *de Bilt*. Assim sendo, e porque teremos para cada k fixado um valor $x_p^{(n)}$, o objectivo é construir um gráfico, em função de k, que descreva a trajectória destes quantis $x_p^{(n)}$.
As figuras 5.9 e 5.10 mostram os valores de risco (quantis empíricos) estimados para as cidades de *Berlim* e *de Bilt*, identificados pelas séries de quantis $x_{p_1}^{(n)}$ e $x_{p_2}^{(n)}$.

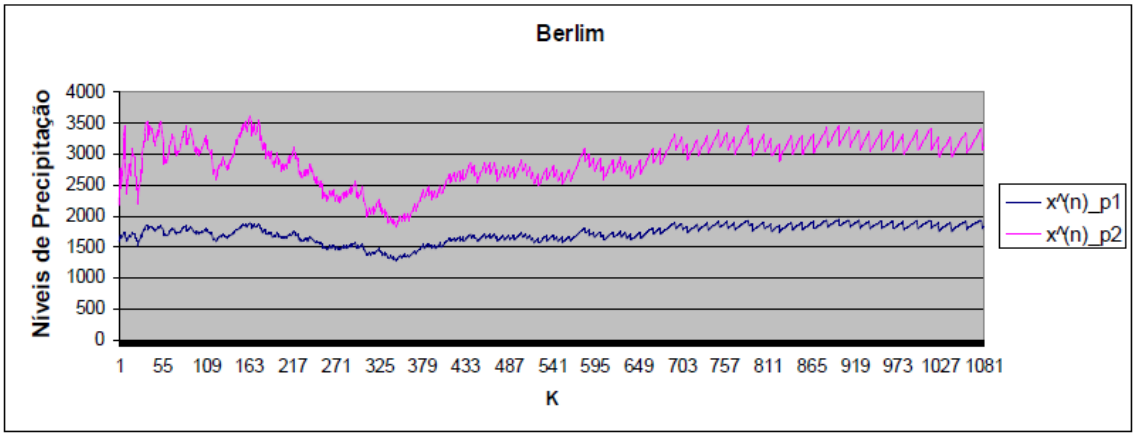

Figura 5.9: Quantis empíricos sobre os níveis de precipitação para *Berlim*

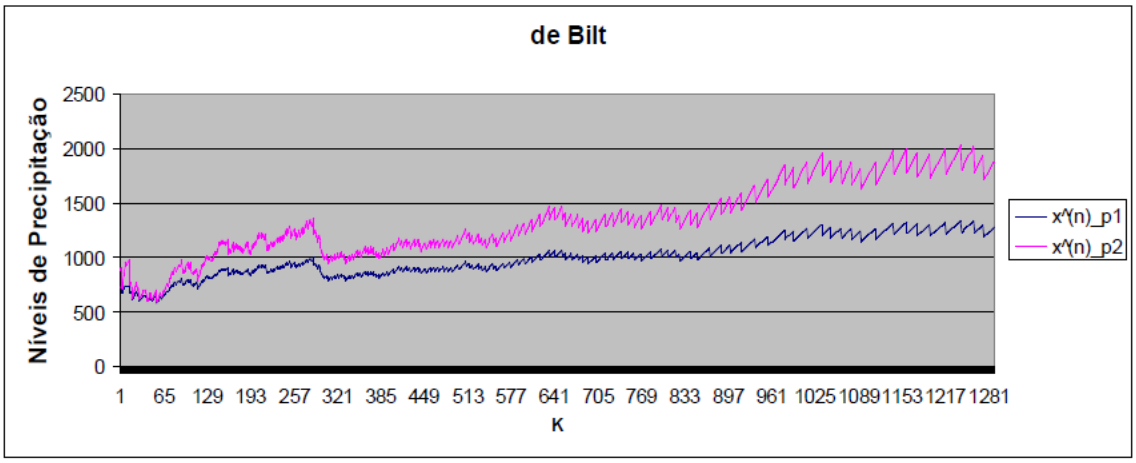

Figura 5.10: Quantis empíricos sobre os níveis de precipitação para *de Bilt*

Pode-se constatar que os níveis de precipitação para *Berlim* apresentam alguma variabilidade em função do número de observações de topo ($5 \leq k \leq 1081$) fixado, particularmente para a série $x_{p_2}^{(n)}$. A série $x_{p_1}^{(n)}$ fornece as menores estimativas, como era de se esperar dado que a probabilidade é menor, muitas vezes oscilando entre os 1500 e 2000 mm de precipitação. Relativamente aos dados de *de Bilt*, também se verifica alguma variabilidade, mais acentuada também na série $x_{p_2}^{(n)}$. Para pequenos valores de k, ($5 \leq k \leq 1289$), verificam-se, para ambas as séries, estimativas dos níveis de precipitação bem próximas umas das outras, entre os 500 e 1000 mm, de precipitação. Isto deve-se ao facto de que, em distribuições de caudas mais leves, para valores cada vez mais pequenos de probabilidade, as diferenças em termos do valor das estimativas dos quantis não seja significativo.
Para os seguintes valores de k, os níveis de precipitação são respectivamente para as séries $x_{p_1}^{(n)}$ e $x_{p_2}^{(n)}$:

k	$x_{p_1}^{(n)}$	$x_{p_2}^{(n)}$
30	1627	2578
50	1753	3132
100	1724	3036

Tabela 5.5: Níveis de precipitação para pequenos valores de k: *Berlim*

k	$x_{p_1}^{(n)}$	$x_{p_2}^{(n)}$
30	685	756
50	643	676
100	761	879

Tabela 5.6: Níveis de precipitação para pequenos valores de k: *de Bilt*

5.5 Fase IV

Até esta fase temos estado a utilizar os dados ao longo dos diferentes anos, sem ter em conta eventuais mudanças climáticas nos valores de precipitação, ou seja, assumindo a hipótese de estacionaridade. Poderá haver uma tendência crescente ou decrescente ao longo dos anos e nós não estamos a ter isso em linha de conta. Os quantis ou valores de risco foram anteriormente calculados assumindo que os dados são identicamente distribuídos. Agora a distribuição pode modificar-se ao longo do tempo e admitiremos que esta modificação, para valores elevados de precipitação se faz de acordo com:

$$\frac{1-F_{s_j}(t)}{1-F_0(t)} \xrightarrow[t\to+\infty]{} e^{cs_j}\,, \quad c \in \mathbb{R} \tag{5.6}$$

onde, $F_{s_j}(t)$ é a função distribuição no instante s_j: $F_{s_j}(t) = P\{X(s_j) \leq t\}$, para cada s_j. Admitiremos no entanto que $F_{s_j} \in D(H_\gamma)$, para todo o s_j. Ou seja para valores grandes ($t \to \infty$), a probabilidade de observar um valor superior a este valor grande aqui fixado, modifica-se em função desta probabilidade inicial no mesmo valor, à custa de uma função do tipo e^{cs_j}. Esta função de risco relativo determinará a tendência.

Para a análise da fase IV consideramos, com base na amostra de observações i.i.d. de *de Bilt*, uma amostra de 100 anos de registo dos valores diários de precipitação, a contar dos anos mais recentes (do ano de 2007 a 1908). Esta amostra foi dividida em $m = 20$ sucessivos períodos de igual dimensão, em que identificamos $j = 0, 1, 2, \ldots, 19$ blocos, ou períodos de 5 anos cada, de modo que, para $j = 0$, $s_j = s_0$ corresponde ao primeiro bloco ou período de 5 anos, para $j = 1$, $s_j = s_1$ corresponde ao segundo bloco ou período de 5 anos, assim sucessivamente. Cada período possui $n = 350$ observações com excepção do último em que $n = 349$. Em cada período de 5 anos ordenamos os dados de precipitação

por ordem decrescente de grandeza e extraímos as 25 observações de topo, para formar o conjunto de valores extremos.
Em m períodos de 5 anos, definimos:
O primeiro período de 5 anos como sendo o instante zero (ou instante inicial): $s_0 = 0$;
os j períodos seguintes: $s_j = j/m - 1\,, \quad j = 0, \ldots, m-1$. Os valores s_j que identificam os períodos variam no intervalo [0, 1].
Os períodos são os apresentados na tabela 5.7.

1º periodo	1908 a 1912	11º periodo	1958 a 1962
2º periodo	1913 a 1917	12º periodo	1963 a 1967
3º Periodo	1918 a 1922	13º periodo	1968 a 1972
4º periodo	1923 a 1927	14º periodo	1973 a 1977
5º periodo	1928 a 1932	15º periodo	1978 a 1982
6º periodo	1933 a 1937	16º periodo	1983 a 1987
7º periodo	1938 a 1942	17º periodo	1988 a 1992
8º periodo	1943 a 1947	18º periodo	1993 a 1997
9º periodo	1948 a 1952	19º periodo	1998 a 2002
10º periodo	1953 a 1957	20º periodo	2003 a 2007

Tabela 5.7: Períodos em estudo.

Para averiguarmos se as caudas das distribuições dos níveis diários de precipitação para cada período possuem um índice de valores extremos que leve a mesma família de distribuições, são determinadas as estimativas do índice de valores extremos para os períodos em estudo. O estimador do índice de valores extremos $\gamma(s_j) = \gamma$ adoptado em cada período e denotado por $\widehat{\gamma}_{n,k}(s_j)$ é dado pelo estimador dos momentos.
A figura 5.11 apresenta as estimativas do índice de valores extremos em cada período s_j. O comportamento da série de estimativas, evidencia uma grande variabilidade, o que sugere diferentes classes de distribuições de valores extremos para a distribuição subjacente aos dados em cada período. Por exemplo, para os períodos s_4 (quinto período) e s_{19} (vigésimo período), o valor é aproximadamente zero, indicando que a distribuição subjacente aos dados poderá pertencer ao domínio de atracção da Gumbel. A menor e a maior das estimativas obtidas para o índice de valores extremos são respectivamente $\widehat{\gamma}_{s_{13}} = -0.4$ e $\widehat{\gamma}_{s_8} = 0.3$.
Sendo γ o parâmetro que determina o peso da cauda da f.d. subjacente aos dados e porque um valor positivo de γ indica uma distribuição de cauda pesada então, nos períodos em que $\gamma > 0$, é de se esperar a ocorrência de níveis de precipitação extrema. Consequentemente, os estimadores utilizados irão ter influência na estimação de quantis elevados e obviamente

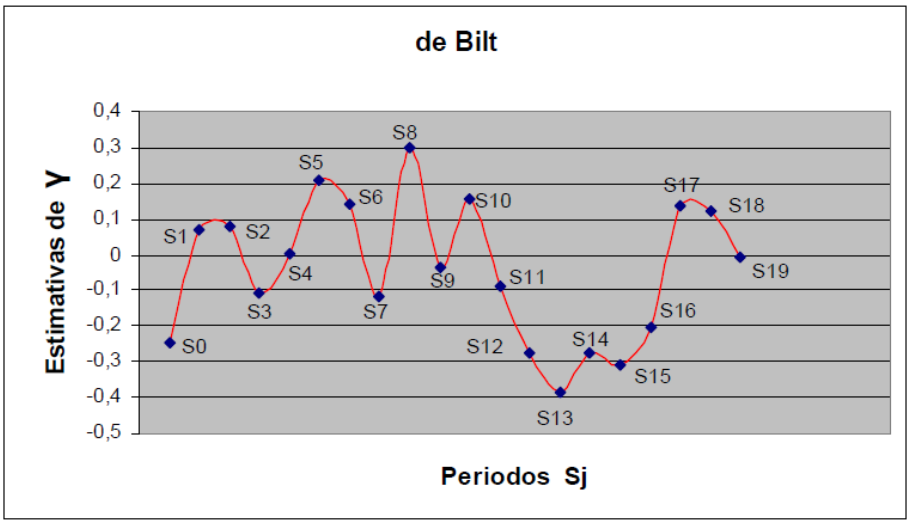

Figura 5.11: Estimativas do índice de valores extremos $\widehat{\gamma}$ para os períodos s_j, segundo o estimador dos momentos com $k = 25$.

também na estimação da probabilidade de excedência de um nível elevado previamente fixado.
Considerando a amostra anteriormente observada para *de Bilt* e as 25 observações de topo são calculados, para cada período s_j, os quantis empíricos $x_{p^*}(s_j)$, com probabilidades 0.01 e 0.0001, como fixado na primeira fase. Estes valores de risco são os níveis de precipitação que poderão ser excedidos somente para aqueles valores de probabilidade. A expressão para determinar o quantil empírico $x_{p^*}(s_j)$ é a seguinte:

$$Q(s_j) \equiv x_{p^*}(s_j) = X_{n-k:n}(0) + \widehat{a}_0 \left(\frac{n}{k}\right) \frac{\left(e^{cs_j} \frac{k}{np^*}\right)^{\widehat{\gamma}(s_j)} - 1}{\widehat{\gamma}(s_j)}.$$

A tabela 5.8 apresenta as estimativas $Q_1(s_j) = x_{p_1}(s_j)$ e $Q_2(s_j) = x_{p_2}(s_j)$, onde $p_1 = 0.01$ e $p_2 = 0.0001$.
As figuras 5.12 e 5.13 apresentam as séries das estimativas $Q_1(s_j)$ e $Q_2(s_j)$, para $s_j = j$, $j = 0, \ldots, 19$.
Pode-se verificar, conforme estimação feita para o índice de valores extremos em cada período, que aos maiores valores estimados do índice, correspondem os níveis de precipitação mais elevados, enquanto que aos menores valores estimados, os níveis mais baixos. Os níveis de precipitação mais elevados ocorrem no nono período (s_8) e os mais baixos no décimo quarto período (s_{13}) para ambas as séries.
Sendo a tendência e^{cs_j} uma função exponencial, para valores próximos de zero, é do tipo linear. Para valores negativos de c, teremos uma tendência decrescente. Analogamente, para $c > 0$, a função tendência e^{cs_j} é uma função crescente. A título de exemplo, fixando $c = 1$ e $s_j = 0.05$, o limite em (5.6) implica que a função distribuição ao final deste mesmo

Períodos	$x_{p_1}(s_j)$	$x_{p_2}(s_j)$	Períodos	$x_{p_1}(s_j)$	$x_{p_2}(s_j)$
1º Periodo	324	478	11º Periodo	445	1343
2º Periodo	378	943	12º Periodo	381	659
3º Periodo	384	974	13º Periodo	345	466
4º Periodo	352	614	14º Periodo	329	403
5º Periodo	376	799	15º Periodo	349	468
6º Periodo	432	1526	16º Periodo	345	446
7º Periodo	419	1228	17º Periodo	368	528
8º Periodo	362	609	18º Periodo	476	1309
9º Periodo	482	2258	19º Periodo	475	1253
10º Periodo	387	742	20º Periodo	431	821

Tabela 5.8: Quantis empíricos sobre os níveis de precipitação para *de Bilt*, em cada período s_j, com probabilidades $p_1 = 0.01$ e $p_2 = 0.0001$.

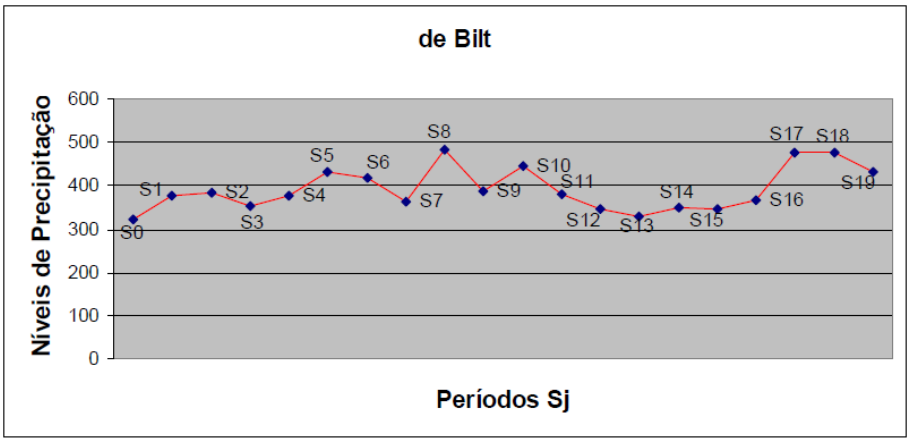

Figura 5.12: Quantis empíricos sobre os níveis de precipitação para *de Bilt*, em cada período s_j, com probabilidade $p_1 = 0.01$.

período, é aproximadamente igual a $e^{0.05}(1 - F_0(t))$, ou seja

$$1 - F_{0.05}(t) \approx e^{0.05}(1 - F_0(t)).$$

A probabilidade de observar um valor superior a t no final do 2º período (ao fim de 10 anos) é igual a 1.05 multiplicado pela probabilidade de observar este mesmo valor no instante

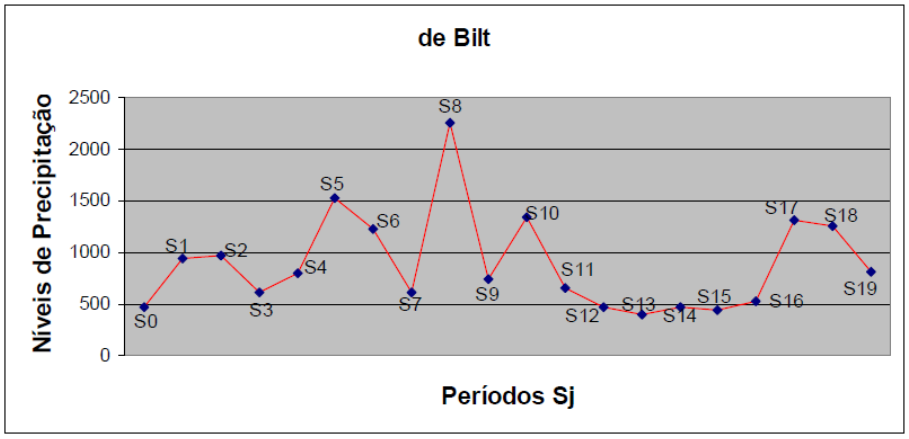

Figura 5.13: Quantis empíricos sobre os níveis de precipitação para *de Bilt*, em cada período s_j, com probabilidade $p_2 = 0.0001$.

inicial. Sendo assim teremos uma tendência crescente.
Para a cidade de *de Bilt*, a estimativa da medida de tendência é $\widehat{c} = 0.8$, decorrente de

$$\widehat{c} := \frac{\sum_{j=0}^{m-1} s_j \ \log\left(1 + \widehat{\gamma}_{n,k}(s_j) \ \frac{X_{n-k:n}(s_j) - X_{n-k:n}(0)}{\widehat{a}_0(n/k)}\right)}{\widehat{\gamma}^{+}_{n,k}(s_j) \ \sum_{j=0}^{m-1} {s_j}^2}.$$

onde $\widehat{\gamma}_{n,k}(s_j)$ é o estimador de $\gamma(s_j) = \gamma$, por exemplo, o estimador dos momentos. Adicionalmente, $\widehat{\gamma}^{+}_{n,k}(s_j)$ é o estimador de $\gamma^{+}(s_j) = \max\,(0,\, \gamma_{(s_j)})$, por exemplo, o estimador de Hill, relativo ao período s_j:

$$\widehat{\gamma}^{+}_{n,k}(s_j) := M^{(1)}_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \log\left(X_{n-i+1:n}(s_j)\right) - \log\left(X_{n-k:n}(s_j)\right),$$

para $j = 0,\, 1,\, 2,\, \ldots,\, 19$ e k fixado em 25 para todos os períodos s_j.
A medida de mudança do clima para os diferentes períodos s_j, avaliada em relação ao instante inicial, possui a seguinte expressão, válida para $\gamma > 0$:

$$\frac{VaR_p(s_j)}{VaR_p(0)} \equiv \frac{Q_j(s_j)}{Q_j(0)} \equiv \frac{x_{p_1}(s_j)}{x_{p_1}(0)} := 1 + \frac{\widehat{a}_0(\frac{n}{k})}{X_{n-k:n}(0)} \frac{e^{\widehat{\gamma}(s_j) c s_j} - 1}{\widehat{\gamma}(s_j)},$$

onde $X_{n-k:n}(0)$ indica a k-ésima observação superior no instante inicial (1º período), $\widehat{a}_0(\frac{n}{k})$ é o estimador para a escala, no instante inicial, associado ao estimador dos momentos de Dekkers, Einmahl e de Haan, (1989) definido na secção 3.2.3 do capítulo 3, e $\widehat{\gamma}(s_j)$ é o estimador do índice de valores extremos, obtido em cada período s_j, pelo primeiro estimador dos momentos.
Para a cidade de *Berlim*, verifica-se que a medida de tendência é nula ($\widehat{c} = 0$), ou seja o processo é estacionário, pelo que não se justifica a aplicação da metodologia aqui descrita.

Períodos	$x_{p_1}(s_j)/x_{p_1}(0)$	Períodos	$x_{p_1}(s_j)/x_{p_1}(0)$
1º periodo	1	11º periodo	1,2173835
2º periodo	1,021054004	12º periodo	1,226633984
3º periodo	1,042184596	13º periodo	1,235590613
4º periodo	1,062637497	14º periodo	1,246503979
5º periodo	1,084108463	15º periodo	1,271862324
6º periodo	1,107433855	16º periodo	1,286738111
7º periodo	1,128428677	17º periodo	1,314556742
8º periodo	1,144622865	18º periodo	1,37522466
9º periodo	1,176891384	19º periodo	1,396274467
10º periodo	1,187950442	20º periodo	1,398076176

Tabela 5.9: Medidas de mudança do clima para os períodos s_j, com $k = 25$.

Na tabela 5.9 registam-se as medidas de alteração climática, para a cidade de *de Bilt*, nos subsequentes períodos, avaliados relativamente ao instante inicial. Pode-se verificar que, ao longo dos períodos há uma tendência para o aumento desses valores, o que sugere um aumento dos níveis de precipitação com o decorrer dos anos. A título de exemplo, se avaliarmos o valor da medida de mudança do clima no 5º período (s_4), podemos dizer que os níveis de precipitação (ao fim de 25 anos), comparativamente aos do 1º período, são em 8% mais elevados. A mesma análise comparativa pode ser feita para os restantes períodos, relativamente ao primeiro.

Capítulo 6

Conclusão

Ao longo deste trabalho tentou-se demonstrar a utilidade da Teoria de Valores Extremos, na quantificação do risco de precipitação elevada, no que respeita à análise da cauda direita das distribuições, onde a abordagem tradicional baseada no Teorema do Limite Central, se revela ineficaz sob condições extremas, justificando o uso de métodos mais sofisticados de avaliação do risco, num contexto de inferência estatística em valores extremos.

Há, no entanto, que ter em conta o número de observações consideradas para estimar a cauda. O número de excessos de nível a reter de uma amostra aleatória é determinado pelo nível determinista "u", fixado segundo a metodologia POT. A escolha do nível "u" não é trivial, porque dele depende o número final k (aleatório) de observações a serem consideradas. Ao reter muitas observações haverá uma grande variabilidade nos dados, o que não é desejável dado que provoca um aumento da imprecisão das estimativas associadas ao parâmetro de cauda, e consequentemente das estimativas de quantis elevados. Por outro lado, se o nível determinar um reduzido número de observações, as estimativas poderão ser pouco fiáveis.
Na metodologia *Peaks-Over-Threshold* (PORT), O nível intermédio $u = X_{n-k:n}$ corresponde à $(k+1)$-ésima estatística ordinal descendente e irá depender do número de observações de topo considerado. Qualquer inferência produzida sobre a cauda da distribuição subjacente será em função deste número k.

As abordagens apresentadas encontram-se ainda em fase de desenvolvimento. No caso da distribuição generalizada de Pareto, em diversas situações o nível óptimo de "u" não é claro. A emergência de um algoritmo adaptativo pode contribuir para a sua determinação. Já no que se refere à abordagem semi-paramétrica os métodos adaptativos de redução de viés têm sido um dos temas de intensa investigação ao longo dos últimos anos.

Na análise aqui considerada, assumiu-se que os níveis diários de precipitação eram independentes e identicamente distribuídos com função distribuição F pertencente a algum domínio de atracção para máximos. Na realidade essa hipótese pode não se verificar. Em

algumas situações os níveis diários de precipitação podem ter uma tendência crescente como se apresenta para a cidade de Bilt em que a estimação é realizada com a incorporação de uma função de tendência, assumindo portanto a não estacionaridade do processo ao longo do tempo. A inovação neste trabalho diz precisamente respeito à incorporação de uma tendência em quantis elevados. Define-se uma medida de mudança do clima. A sua aplicação sugere a presença de uma tendência crescente sobre os valores de precipitação mais elevada em *de Bilt*, na Holanda. De um modo geral, em cada mudança de década, há um aumento de 4% nos valores de precipitação extrema registados em *de Bilt*.

Para a cidade de *Berlim*, conclui-se que os maiores valores de precipitação ocorrem em regime estacionário, ou seja, não foi detectada tendência significativa.

Referências Bibliográficas

[1] Aghakouchak, A.; Nasrollahi, N. (2009). Semi-parametric and parametric inference of extreme value models for rainfall data. *Water Resour Manage*, **24**, 1229-1249.

[2] Artzner, P.; Delbaen, F.; Eber, J.; Heath, D. (1999). Coherent measures of risk. *Mathematical finance* **9**(3), 203-228.

[3] Araújo Santos, P.; Fraga Alves, M.I.; Gomes, M. I. (2006). Peaks Over Random Threshold methodology for tail index and high quantile estimation. *REVSTAT*, **4**, 227-247.

[4] Araújo Santos, P.; Fraga Alves, M.I.; Gomes, M.I. (2006). Comparacão do desempenho de diferentes níveis aleatórios na metodologia PORT. XIV Congresso Anual da *Sociedade Portuguesa de Estatística-SPE*, Universidade da Beira interior.

[5] Balkema, A.; de Haan, L. (1974). Residual life time at great age. *Annals Probab*, **2**, 792-804.

[6] Beirlant, J.; Goegebeur, Y.; Segers, J.; Teugels J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley, London.

[7] Castilho, E.; Galambos, J.; Sarabia, JM. (1989). The selection of the domain of attraction of an extreme value distribution from a set of data. In: Hüsler, J.; Reiss R-D (eds) *Lecture Notes in Statistics*, **51**. Springer, Berlin, 181-190.

[8] Danielsson, J.; Vries, C. (1997). Tail index and quantile estimation with very high frequency data. *J. Empir. Finance*, **4**, 241-257.

[9] de Haan, L.; Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer.

[10] Dekkers, A.; Einmahl, J.; de Haan, L. (1997). A moment estimator for the index of an extreme value distribution. *Annals Stat* **17**, 1833-1855.

[11] Elsi, Z.; Yildirim, I.; Yildirak, K. (2005). Alternative risk measure and extreme value theory in finance: implementacion on ISE 100 index. *International Conference on Business Economics and Management*, Yasar University Izmir Turkey, 2005.

[12] Embrechts, P.; Klüpelberg, C.; Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag Berlin Heidelberg.

[13] European Environment Agency. (N°2/2004) *Impacts of Europe's changing climate*. An indicator-based assessment. Luxembourg: Office for Official Publications of the European communities.

[14] Ferreira, A.; de Haan, L.; Peng, L. (2003). On optimising the estimation of high quantiles of a probability distribution. *Statistics*, **37**(5), 401-434.

[15] Fisher, R.; Tippett, L. (1928). Limiting forms of the frequency distributions of the largest or smallest member of a sample. *Proc Camb Phil. Soc*, **24**, 180-190.

[16] Fraga Alves, M.I. (1999). Asymptotic distribution of Gumbel statistics in a semi-parametric approach. *Port. Math.*, **56**, 282-298.

[17] Fraga Alves, M.I. (2007). Acerca de testes estatísticos para valores extremos. Boletim *SPE-Sociedade Portuguesa de Estatística*. Publicação semestral, primavera de (2007).

[18] Fraga Alves, M.I.; Gomes, M.I. (1996). Statistical choice of extreme value domains of attraction - a comparative analysis. *Commun. Statit. - Theory Math.*, **25**, 789-811.

[19] Fraga Alves, M.I.; Gomes, M.I.; de Hann, L. (2009). Mixed moment estimator and location invariant alternatives. *Extremes* (2009), **12**, 149-185.

[20] Galambos, J. (1982). A Statistical test for extreme value distributions. *In: Nonparametric Statistical Inference*. B.W. Gnedenko *et al.* (eds), North Holland, Amsterdam, 221-230.

[21] Gilli, M.; Kellëzi, E. (2006). An application of extreme value theory for measuring financial risk. *Computational Economics*. **27**(1), 1 - 23.

[22] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals Math*, **44**, 423-453.

[23] Goldstein, J.; Mirza, M.; Etkin, D.; Milton, J. (2003). Hidrologic assessment: aplication of extreme value theory for climate extreme scenarios construction *In: 14th symposium on global change and climate variations*, American meteorological society 83rd annual meeting, Long Beach, 9-13 Feb 2003.

[24] Gomes, M. I.; Fraga Alves, M. I.; Gomes, M. I.; Araújo Santos, P. (2008). Port hill and moment estimators for heavy-tailed models. *Commun. Statist.-Simul. and Compt.* **37**, 1281-1306.

[25] Gomes, M. I.; Pestana, D. (2007). A sturdy reduced-bias extreme quantile (VaR) estimator. *J. American Statistical Association*, **102**, N°477, 280-292.

[26] Greenwood, M. (1946). The statistical study of infections diseases. *J. Roy. Statst. Soc. Ser. A*, **109**, 85-109.

[27] Guillou, A.; Naveau, P.; Diebolt, J.; Ribereau, P. (2008). Return level bounds for discrete and continuous random variables. *Sociedad de Estadística e Investigatión operativa* 2008. **18**, 584-604.

[28] Gumbel, E. (1942). On the frequency distribution of extreme values in meteorological data. *Bull Am Meteorol Soc* **23**, 95-104.

[29] Gumbel, E. (1958). *Statistics of extremes*. Colombia University Press New York.

[30] Hasofer, A.; Wang, J.Z. (1992). A test for extreme value domain of attraction. *J. Amer. Statist. Assoc.*, **87**, 171-177.

[31] Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *Annals Stat* **3**, 1163-1174.

[32] IPCC (2001a). *Climate change 2001: The Scientific Basis*. In: Houghton, JT.; Ding, Y.; Griggs, DJ.; Noguer, M.; Van der Linden, PJ.; Dai, X.; Maskuell, K.; Johnson, CA. (eds) Contributions of working group I to the third assessment report of the intergovernamental panel on climate change. Cambridge University Press, Cambridge.

[33] IPCC (2001b). *Climate change 2001: impacts adaptation, and vulnerability*. In: Mccarty, JJ.; Canziani, OF.; Leary, NA.; Dokken, DJ.; White, KS. (eds) Contributions of working group II to the third assessment report of the intergovernamental panel on climate change. Cambridge University Press, Cambridge.

[34] IPCC (2007). *Climate change 2007: impacts adaptation, and vulnerability*. In: Parry, ML.; Canziani, OF.; Palutikof, JP.; Van der Linden PJ,; Hanson, CE. (eds) Exit EPA disclaimer Contributions of working group II to the third assessment report of the intergovernamental panel on climate change. Cambridge University press, Cambridge.

[35] Jenkinson, A. F. (1955). The frequancy distribution of the annual maximum (or minimum) values of meteorological elements. *Quart. J. Roy. Meteo. Soc.*, **81**, 158-171.

[36] J. MCNeil, A. (2000). Extreme value theory for risk managers. In: Embrechets, P. (ed) *Extremes and Integrated Risk Management*. In association with UBS Warburg.

[37] Katz, R.; Parlange, M.; Naveau, P. (2002). Statistics of extremes in hydrology. *Adv Water Resour* **25**, 1287-1304.

[38] L. Smith, R. (2000). Measuring risk with exreme value theory In: Embrechets, P. (ed) *Extremes and Integrated Risk Management*. In association with UBS Warburg.

[39] Neves, C.; Fraga Alves, M.I. (2007). Semi-parametric approach to Hasofer-Wang and Greenwood statistics in extremes. *Test*, **16**, 297-313.

[40] Neves, C.; Picek, J.; Fraga Alves, M.I. (2006). The contributions of the maximum to the sum of excess for testing max-domains of attraction. *J. Statist. Plann. Inference*, **136** (4), 1281-1301.

[41] Neves, C.; Fraga Alves, M. I. (2008). Testing extreme value conditions-An overview and recent approches. *Revstat-Statistical journal*, **6**, 83-100.

[42] Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals Stat* **3**, 119-130.

[43] Reiss, R. D.; Thomas, M. (2007). *Statistical Analysis of Extreme Values from Insurance, Finance, Hydrology and other fields*. Birkhäuser Verlag. 3rd ed.

[44] Segers, J.; Teugels, J. (2000). Testing the Gumbel hypotesis by Galton's ratio. *Extremes*, **3** (3), 291-303.

[45] Smith, R. (2001). *Extreme value statistics in meteorology and environment*. Environment statistics. Disponível em: http:\www.stat.unc.edu/postscript/rs/envstat/env.html

[46] von Mises, R. (1936). La distribution de la plus grande de n valeurs. Reprinted in selected papers volume II, *American Mathematical Society*, Providence, R. I., 1954, 271-294.