Universidade Federal do Rio de Janeiro

Modelagem Condicional de Valores Extremos Multivariados

Uma Aplicação no Cálculo de Vazão de Rios em Porto Rico

Valmária Rocha da Silva

Rio de Janeiro

2006

Livros Grátis

http://www.livrosgratis.com.br

Milhares de livros grátis para download.

Valmária Rocha da Silva

Modelagem Condicional de Valores Extremos Multivariados

Uma Aplicação no Cálculo de Vazão de Rios em Porto Rico

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro como parte dos requisitos necessários à obtenção do título de Mestre em Estatística.

Orientadora:
Beatriz Vaz de Melo Mendes

Departamento de Métodos Estatísticos
Instituto de Matemática
Universidade Federal do Rio de Janeiro

março de 2006

Folha de exame

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro como parte dos requisitos necessários à obtenção do título de Mestre em Estatística.

Banca examinadora:

Beatriz Vaz de Melo Mendes

Instituto de Matemática - UFRJ

Francisco Cribari Neto

Departamento de Estatística - UFPE

Denise Britz do Nascimento Silva

ENCE - IBGE

Aos meus pais e a todos os meus professores.

Agradecimentos

Quero começar agradecendo a Michelle Dysman (FGV-RJ), Thaís Fonseca, Alexandra Schmidt e Dani Gamerman pelos incentivos, quer por e-mails ou por palavras proferidas. -Vocês não imaginam o quanto me ajudaram!

Muito obrigada aos professores do Mestrado: Gadelha, Nei, Marco Antônio, Migon e Flávia. -O que seria dos alunos se não fosssem seus mestres?! Vocês são excelentes referenciais!

Agradeço aos meus amigos de turma: Adelmo Inácio, Ana Paula, Carla, Carolina, Carlos Henrique, Geraldo Marcelo, Gilmar, Marcus Vinícius e Marcus Aurélio. - Vocês são amigos mais chegados que irmãos. Ai de mim se não fosse o colo de vocês!

Obrigada também aos amigos da pós-graduação IM-UFRJ: Edna, Romi, Ester, Luis, Gustavo, Aline, Cristiane, Fidel, Luzia, Flávio, Vinícios, Henrique, Francisco, Josiane, Juam, etc. Obrigada também aos amigos da pós de outros estados, Tatiene (PE), Tiago (PE), Carlos (PE), Raquel (RS). -O msn faz milagres!

Aos amigos que infelizmente não continuaram conosco: Lúcia e Mário. -É uma pena não termos ao nosso lado todas as pessoas que amamos!

Ao pessoal da secretaria: Lia, Eduardo, Davi e Rosi. -Valeu pelos "quebra galhos"pra imprimir alguns trabalhos. Muito obrigada mesmo!

Um agradecimento todo especial a minha orientadora Beatriz Mendes pelo incentivo, paciência, oportunidade e confiança. -Não sei se lhe admiro mais como pessoa ou como profissional. É um prazer aprender com você. Você nos orienta para vida e não para um trabalho específico. Não posso agradecer-lhe o bastante!

À CAPES, pelo apoio financeiro.

Muito obrigada também ao pessoal do pensionato que me ajudou no início de minha estadia aqui no Rio. Jamais esquecerei tudo que fizeram por mim. Valeu Kariana, Albert, Evilson e Glaydson.

Às amigas Marilane, Tereza, Joseline e Ediana. -O valor da nossa amizade excede ao de jóias preciosas!

À minha família maravilhosa. -Vocês me ensinaram: a rir com frequência; a conseguir a admiração de pessoas inteligentes; a ganhar o afeto das crianças; a receber críticas de pessoas honestas; a curtir o belo; a descobrir o que há de melhor nos outros; a deixar o mundo melhor quer através de uma ação social bem sucedida ou de uma tarefa bem executada. Sei que vocês ficam felizes pelo simples fato de saber que eu estou bem. Amo vocês com todas as minhas forças!

Aos meus sobrinhos Laila Letícia, Lara Roberta e Pedro Benício. - Se o céu sentir falta de três anjinhos, eu sei onde eles estão!

À família Vasconcelos e Bandeira. -Quando eu tive fome vocês me deram de comer; quando eu tive sede vocês me deram de beber; quando eu não tinha pra onde ir vocês me deram abrigo. Nada melhor que se sentir acolhida em terra estranha. Amo todos vocês!

A Mônica pelo monitor e pelos almoços, e a Júnior e Thiago pela impressora. -Vocês foram refrigério na minha angústia!

À Igreja Batista Nacional de Teresina-PI. Faço das minhas as palavras de Max Lucado "obrigada por enviar chuvas de encorajamento e oceano de orações". -Obrigada também pelo gasto com ligações telefônicas.

À Igreja Batista de Betânia-RJ, por ser um canal de benção em minha vida. -Ah, como é bom entrar no trono da graça com vocês! Só quem conhece Betânia sabe do que estou falando!

Aos amigos e professores da UESPI e UFPI. -Vocês são parte da minha vitória, podem comemorar!

In memory, Albertina Gualter (melhor amiga da graduação), Renan, Artagnan e Rosa Cardoso (avó paterna). -Infelizmente perdemos pessoas que amamos no caminho, mas um dia a gente vai se encontrar, um dia todos nós vamos nos encontrar.

Aplaudo também de pé àquele que não conheço o seu rosto mas que O reconheço em todo lugar; àquele que nunca toquei em suas vestes mas que sinto o seu perfume; àquele que não sei como é a sua voz mas cujo som me guia. À Ti ó Deus, doador da vida e criador de todas as coisas, toda honra e toda glória!

A todos aqueles que de alguma maneira contribuíram para a realização deste trabalho. -Valeu MSN, Crtl+C, Crtl+V, help, google, replace, Kirk Franklin, Michael W. Smith, Hillsong, Oficina G3, Fruto Sagrado, Fernanda Brum, Aline Barros, MaryMary, Tonéx, Diante do Trono, Blinside, Adriano Gospel Funk, André Valadão, Apascentar de Nova Iguaçu, Kleber Lucas, Alda Célia, Renascer Praise, Demon, Yolanda Adams, Filhos do Homem, Moniquinha, Janinha, Adri, galera de Timon-MA, galera do promorar, amigos e vizinhos do parque piauí - PI, aos torcedores anônimos e extravagantes, etc.

Resumo

A Teoria dos Valores Extremos Multivariados e seus métodos tratam da caracterização, estimação e extrapolação das "caudas conjuntas"da distribuição multidimensional. As abordagens existentes são baseadas em argumentos limites em que todas as variáveis são extremas. Iremos aplicar a proposta de um modelo multivariado para situações em que pelo menos uma das variáveis é extrema, desenvolvido por Heffernan and Tawn (2004).

O cálculo das probabilidades dos eventos extremos ocorre em diversas aplicações; a aplicação que consideramos neste trabalho é em relação a riscos hidrológicos. Examinamos a vazão de três rios que provocam inundações em Porto Rico. Os rios são Fajardo, Canóvanas e Espiritu Santo, que monitoram as situações de enchentes nessa região, durante o período de abril de 1967 a setembro de 2003, inclusive.

Abstract

Multivariate extreme value theory and methods concern the caracterization, estimation and extrapolation of the joint tail of the distribuition of a d-dimensional random variable. Existing approaches are basead on limiting arguments in which all components of the variable become large at the same rate. We apply the proposal of a model multivariate for situations in that at least one component is extreme, developed by Heffernan and Tawn (2004).

Accurate assessments of the probabilities of extremes events are sought in a diversity of applications, the application that we considered in this work is in relationship to risks hidrologics. We examined the flooding of three rivers related to floods in Porto Rico. The rivers are Fajardo, Canóvanas and Espírito Santo, that monitor the situations of inundations in that surface, during the period of April of 1967 the September of 2003, inclusively.

Sumário

Introdução **1**

1 Estrutura geral do problema **3**

2 Motivações teóricas **9**

2.1 Suposição de uma representação do limite e suas propriedades . . . 9

2.2 Escolha das funções normalizadoras . . . 11

2.3 Exemplos teóricos . . . 13

3 Estrutura do modelo e suas propriedades **16**

3.1 Modelo de dependência condicional . . . 16

3.2 Diagnósticos . . . 18

4 Inferência **20**

4.1 Estimação marginal . . . 21

4.2 Modelos condicionais . . . 21

4.3 Incerteza . . . 22

5 Análise da vazão de três rios de Porto Rico **24**

5.1 Rio Canóvanas . . . 25

5.2 Rio Fajardo . . . 25

5.3 Rio Espiritu Santo . . . 26

6 Análise sem considerar sazonalidade **27**

6.1 Máximos mensais . . . 29

6.1.1 Dados completos . . . 33

6.1.2 Ajuste do modelo marginal . . . 33

6.1.3 Ajuste do modelo de dependência . . . 37

6.2 Dados sem outlier . . . 41

7 Análise considerando sazonalidade **46**

7.1 Análise descritiva 46

7.2 Ajuste do modelo marginal 49

7.3 Ajuste do modelo de dependência 50

Conclusão **55**

Referências Bibliográficas **56**

Apêndice **58**

Lista de Figuras

1.1 Ilustração do conjunto extremo C bivariado. 4

6.1 acf (esquerda) e pacf (direita) das médias diárias dos fluxos dos rios. 28

6.2 acf (esquerda) e pacf (direita) dos máximos mensais. 29

6.3 Gráfico dos máximos mensais de cada série em função do tempo. 31

6.4 Boxplot dos máximos mensais, separados em meses. 32

6.5 Gráfico dos máximos mensais. 33

6.6 Gráfico da função Média dos Excessos empírica. 34

6.7 Histograma do ajuste dos máximos mensais. 36

6.8 Máximos mensais transformados para a distribuição Gumbel. 36

6.9 Comparação dos parâmetros de dependência estimados usando como limiar o quantil do modelo de dependência 70%. A linha contínua representa o *convex hull* de $(\hat{a}_{j|i}, \hat{b}_{j|i})$ e a linha pontilhada $(\hat{a}_{i|j}, \hat{b}_{i|j})$. 39

6.10 Gráficos dos Resíduos. 40

6.11 Dados simulados do modelo condicional estimado e para os dados na escala original. Nas linhas $i = 1, 2, 3$, apresentamos os pares das distribuições condicionais simuladas, dado que a variável i excede o quantil marginal empírico $x_i(0,99)$. As linhas verticais correspondem ao limiar $x_i(0,99)$. Os pontos abaixo e acima dos limiares são os dados originais e os dados simulados, respectivamente. Os círculos menores representam os pontos que pertencem ao conjunto $C^3(v_p)$. Os círculos grandes representam os seis pontos com maiores valores de $\sum_{i=1}^{3} y_i$. As curvas representam a igualdade dos quantis marginais. 43

6.12 Níveis de retorno estimados para o conjunto $C^3(v_p)$ usando os modelos condicionais estimados. Em (a) dado que o rio Canóvanas é extremo; em (b) dado que o rio Fajardo é extremo; e em (c) dado que o rio Espiritu Santo é extremo. O nível de retorno é $\sum_{i=1}^{3} y_i$, representado pela linha contínua. As linhas pontilhadas correspondem ao intervalo de confiança de 95%. Os pontos "o"representam os níveis de retorno empíricos. 44

6.13 Densidade GPD com os parâmetros estimados da série original (linha contínua) superposta por uma GPD com os novos parâmetros estimados (linha pontilhada). 45

7.1 Função de autocorrelação e autocorrelação parcial. 48
7.2 Gráfico dos fluxos dos máximos mensais separados por períodos. 52
7.3 Histograma do ajuste dos máximos mensais nos dois períodos. 53
7.4 Máximos mensais de cada período transformados para a distribuição Gumbel. . . . 54

Lista de Tabelas

2.1 Exemplos de estruturas de dependência multivariada classificados pelo comportamento de dependência extremal. 14

6.1 Teste KPSS e teste R/S para as médias diárias dos fluxos dos rios. 27
6.2 Teste KPSS sem tendência, KPSS com tendência e teste R/S para os máximos mensais dos fluxos dos rios. 30
6.3 Sumário dos máximos mensais do fluxo dos rios. 30
6.4 Teste de Mann-Kendall de tendência. 31
6.5 Parâmetros estimados da GPD para as distribuições marginais dos máximos mensais. 35
6.6 Teste GOF para o ajuste GPD do fluxo dos máximos mensais. 35
6.7 Parâmetros estimados para o modelo de dependência. 37
6.8 Teste τ de Kendall para os pares $\mathbf{Z}_{|i}$. 38
6.9 Esperanças empírica e baseada no modelo. 41
6.10 Parâmetros estimados para o modelo de dependência sem outlier. 42

7.1 Testes de Mann-Kendall de tendência e sazonalidade. 47
7.2 Teste KPSS sem tendência (KPSS), KPSS com tendência (KPSS-T) e teste R/S de memória longa para os máximos mensais dos fluxos dos rios em cada período. . 47
7.3 Sumário dos máximos mensais nos dois períodos. 49
7.4 Parâmetros do ajuste marginal dos fluxos dos rios nos dois períodos. 50
7.5 Parâmetros estimados do modelo de dependência para os dois períodos. 51
7.6 Valores críticos do teste de Kolmogorov 60

Introdução

A Teoria dos Valores Extremos Multivariados e seus métodos tratam da caracterização, estimação e extrapolação das "caudas conjuntas"da distribuição multidimensional. O cálculo das probabilidades dos eventos extremos ocorre em diversas aplicações, tais como em impacto ambiental Coles and Tawn (1994); Joe (1994); Haan and Ronde (1998); Schlather and Tawn (2000), em finanças Embrechts, Kluppelberg, and Mikosch (1997); Longin (2000); Staricã (2000); Poon, Rockinger, and Tawn (1981) e em modelagem do trânsito na Internet Maulik, Resnick, and Rootzen (2002); Resnick and Rootzen (2000).

Todos os modelos existentes de eventos extremos são baseados em situações em que todas as variáveis são extremas. Nesta tese estudamos o artigo de Janet E. Heffernan e Jonathan A. Tawn, *J. R. Statist Soc B, [2004],66, Part 3.* A novidade dessa abordagem é que ela pode ser aplicada em situações em que pelo menos uma das variáveis é extrema; além disso esse modelo também é aplicável a situações de qualquer dimensão.

A aplicação considerada neste trabalho é em relação a riscos hidrológicos. Examinamos três rios que provocam inundações em Porto Rico; são eles os rios Fajardo, Canóvanas e Espiritu Santo, que monitoram as situações de enchentes nessa região, durante o período de abril de 1967 a setembro de 2003, inclusive.

Todos os anos na época de chuva algumas regiões apresentam problemas de enchentes, resultando em desabrigados, danos materiais dos mais variados e, o que é mais grave, mortes. As principais causas das enchentes são: alto índice pluviométrico da região; alto grau de impermeabilização do solo pela malha asfáltica e de concreto; a falta de educação ambiental dos administradores e da população em geral; e falta de saneamento básico adequado.

Apresentamos no Capítulo 1 a estrutura geral do problema. No Capítulo 2 apresenta-

mos a motivação teórica do modelo, uma nova abordagem assintótica na qual esse modelo condicional é baseado e alguns exemplos teóricos. No Capítulo 3 mostramos a estrutura do modelo e suas propriedades, detalhando o modelo de dependência condicional e os diagnósticos. No Capítulo 4 a inferência para o modelo é discutida. Toda a parte teórica está introduzida nesses quatro primeiros capítulos.

Os Capítulos 5, 6 e 7 consistem na aplicação do nosso trabalho. No Capítulo 5 falamos um pouco a respeito dos rios. No Capítulo 6, analisamos e fizemos uma aplicação sem considerar sazonalidade e no Capítulo 7 analisamos os dados considerando sazonalidade. Finalmente apresentamos as considerações finais e comentamos a importância dos resultados obtidos.

Capítulo 1

Estrutura geral do problema

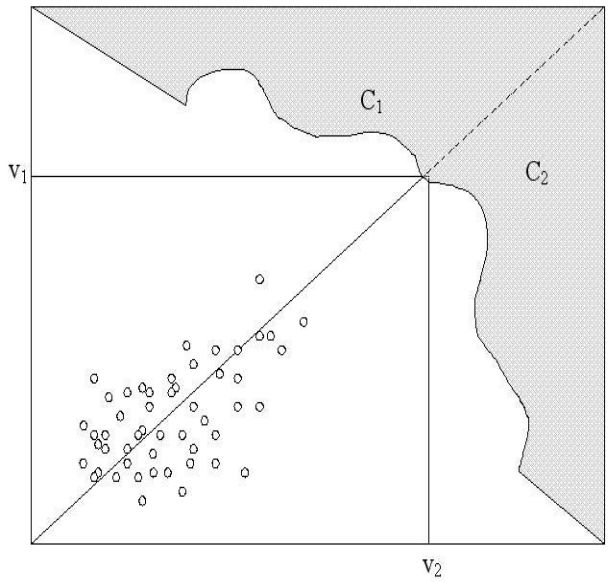
Considere um vetor de variáveis contínuas $\mathbf{X} = (X_1, \ldots, X_d)$ com função de distribuição desconhecida $F(x)$. Para o caso de n observações independentes e identicamente distribuídas, desejamos estimar uma função distribuição de $\mathbf{X}$ quando $\mathbf{X}$ é extremo em pelo menos uma de suas componentes. Por exemplo, esta função poderia ser: $E(X_j \mid X_i > v)$; $Pr(X_1 + X_2 + X_3 > v)$ ou $Pr(\mathbf{X} \in C)$ com C sendo um conjunto extremo detalhado a seguir e v sendo $v = inf_{x \in C_i}(x_i)$. Os métodos que são desenvolvidos aqui permitem que qualquer função seja considerada. Portanto, para simplificar nosso trabalho vamos concentrar nossa discussão na estimação de $Pr(\mathbf{X} \in C)$ onde C é um conjunto extremo se para todo $\mathbf{x} \in C$ pelo menos uma das componentes de $\mathbf{x}$ é extrema. Tipicamente nenhuma observação ocorre em C. A estrutura de C motiva o surgimento natural da partição de C em subconjuntos $C = \cup_{i=1}^{d} C_i$. Aqui, C_i é uma parte de C o qual X_i é a maior componente de $\mathbf{X}$, medido pelos quantis das distribuições marginais. Especificamente, para cada $i = 1, \ldots, d$, denota-se F_{X_i} como a distribuição marginal de X_i; então,

$$C_i = C \cap \{x \in R^d : F_{X_i}(x_i) > F_{X_j}(x_j)\}; j = 1, ..., d; j \neq i.$$

No caso bivariado,

$$\begin{aligned} C_1 &= C \cap \{x \in R^2 : F_{X_1}(x_1) > F_{X_2}(x_2)\} \\ C_2 &= C \cap \{x \in R^2 : F_{X_2}(x_2) > F_{X_1}(x_1).\} \end{aligned}$$

Figura 1.1: Ilustração do conjunto extremo C bivariado.



Ilustramos o conjunto C, no caso bivariado, com a Figura 1.1.

Assumimos que os subconjuntos de C da forma $C \cap \{x \in R^d : F_{X_i}(x_i) = F_{X_j}(x_j)\}$ para algum $j \neq i$ podem ser ignorados; pois são conjuntos nulos.

Com a partição de C definida, C é um conjunto extremo se para todo x no C_i não-vazio incidir no extremo da cauda de F_X, i.e., se $v_{X_i} = inf_{x_i \in C_i}(x_i)$, quando $F_{X_i}(v_{X_i})$ é fechado para $i = 1, \ldots, d$. Então

$$Pr(\mathbf{X} \in C) = \sum_{i=1}^{d} Pr(\mathbf{X} \in C_i) = \sum_{i=1}^{d} Pr(\mathbf{X} \in C_i \mid X_i > v_{X_i}) Pr(X_i > v_{X_i}). \qquad (1.1)$$

Considere a estimação de $Pr(\mathbf{X} \in C)$ usada na decomposição (1.1). Precisamos estimar $Pr(X_i > v_{X_i})$ e $Pr(\mathbf{X} \in C_i | X_i > v_{X_i})$; a primeira parte requer um modelo marginal de valores extremos e a segunda necessita de um modelo de dependência para valores extremos. Nosso trabalho foi concentrado em torno dessas duas partes.

No caso univariado a distribuição condicional de excesso $Y = X - u$, denotada aqui por F_u, é dada por

$$F_u(y) = Pr(Y \leq y \mid X > u) = Pr(X - u < y \mid X > u).$$

A função de sobrevivência dos excessos Y é dada por

$$\begin{aligned}\overline{F_u}(y) &= Pr(X-u>y|X>u),\\ \overline{F}_u(y) &= \frac{\overline{F}_X(u+y)}{\overline{F}_x(u)},\end{aligned}$$

. Segue então que

$$\overline{F}_{X(u+y)} = \overline{F}_{u(y)}\overline{F}_{x(u)}.$$

Métodos para marginais extremas são relativamente conhecidos; veja Davison and Smith (1990), Smith (1989) e Dekkers, Einmahl, and Haan (1989). A Teoria dos Valores Extremos univariados fornece a justificação assintótica pela distribuição de Pareto Generalizada por ser um modelo apropriado para as distribuições de excessos acima de um limiar grande.

De acordo com Haan (1984), F_X pertence ao domínio máximo de atração da GEV (Distribuição de Valores Extremos Generalizada) se e somente se existe uma função $a(.)$ positiva e mensurável tal que para $1+\xi x>0$ tem-se

$$\lim_{u\to x^{F_X}} \frac{\overline{F_X}(u+a(u))}{\overline{F_X}(u)} = \begin{cases} (1+\xi_i x)^{\frac{-1}{\xi}} & \text{se } \xi_i \neq 0 \\ e^{-x} & \text{se } \xi = 0. \end{cases}$$

Considere a distribuição de Pareto generalizada padrão, cuja função distribuição, denotada por P_ξ é dada por

$$P_\xi(x) = \begin{cases} 1-(1+\xi_i x)^{\frac{-1}{\xi}} & \text{se } \xi_i \neq 0, \\ 1-e^{-x} & \text{se } \xi = 0, \end{cases}$$

onde $x \geq 0$ se $\xi \geq 0$ e $0 \leq y \leq -1/\xi$ se $\xi < 0$.

Temos então que a adequação da GPD para modelar excessos além de um limiar grande se justifica a partir do fato de que a distribuição F_X pertence ao domínio de atração da GEV.

Desta maneira modelamos a cauda da marginal de X_i para $i=1,\ldots,d$ por

$$Pr(X_i > x + u_{X\,i}|X_i > u_{X_i}) = (1+\frac{\xi_i x}{\beta_i})_+^{\frac{-1}{\xi_i}}, \text{ onde } x>0. \tag{1.2}$$

Aqui, u_{X_i} é um limiar para a variável X_i, com β_i e ξ_i sendo parâmetros de escala e forma respectivamente com $\beta_i > 0$ e $s_+ = \max(s, 0)$ para qualquer $s \in \mathbb{R}$. Requeremos um modelo para completar a distribuição marginal F_{X_i} de X_i para cada $i = 1, \ldots, d$, desde que para estimar $Pr(\mathbf{X} \in C_i | X_i > v_{X_i})$ precisamos descrever todos os valores X_j que podem ocorrer com qualquer X_i grande. Adaptamos o modelo semiparamétrico $\widehat{F}_{X_i}$ por F_{X_i} de Coles, Heffernan, and Tawn (1991), i.e.,

$$\widehat{F}_{Xi}(x) = \begin{cases} 1 - (1 - \widetilde{F}_{X_i}(u_{X_i}))(1 + \xi_i(\frac{x - u_{X_i}}{\beta_i}))^{\frac{-1}{\xi_i}} & \text{para } x > u_{Xi}, \\ \widetilde{F}_{X_i}(x) & \text{para } x \leq u_{X_i}, \end{cases} \tag{1.3}$$

onde $\widetilde{F}_{X_i}$ é a distribuição empírica dos valores de X_i. O limite do suporte da distribuição por $x^{F_{X_i}}$ é ∞ se $\xi_i \geq 0$ e $\mu_{X_i} - \frac{\beta_i}{\xi_i}$ se $\xi_i < 0$. O modelo (1.3) fornece as bases para estimação do termo $Pr(X_i > v_{X_i})$ da decomposição (1.1).

Para determinar $Pr(\mathbf{X} \in C_i | X_i > v_{X_i})$ necessitamos das marginais e da estrutura de dependência de $\mathbf{X}$. Separamos nosso trabalho em duas partes e focamos primeiro no modelo de dependência para trabalhar com marginais que assumimos ser conhecidas. Transformamos todas as distribuições marginais univariadas em distribuições marginais Gumbel padrão, que é dada por

$$\begin{aligned} Y_i &= -\log(-\log \widehat{F}_{Xi}(x_i)) \text{ para } i = 1, \ldots, d \\ Y_i &= t_i(X_i; \psi_i, \widetilde{F}_{X_i}) \\ Y_i &= t_i(X_i) \end{aligned} \tag{1.4}$$

onde $\psi_i = (\beta_i, \xi_i)$ são os parâmetros marginais. Essa transformação resulta em $Pr(Y_i \leq y) = \exp(-\exp(-y))$ para cada i, então $Pr(Y_i > y) \sim \exp(-y)$ quando $y \to \infty$ e Y_i tem cauda exponencial no extremo. Para diferenciar qual das variáveis marginais estamos usando, $\mathbf{X}$ denota a distribuição com marginais dos dados originais e $\mathbf{Y}$ a distribuição com marginais Gumbel.

Agora focamos no modelo de dependência extremo de uma variável com marginais Gumbel. Modelar a dependência de valores extremos multivariado é mais complexo do que modelar valores extremos univariados. Apesar de já existirem várias propostas, as metodologias evoluem lentamente. Quando o interesse é o extremo superior de cada

componente de $\mathbf{Y}$, toda a estrutura de dependência fica dentro de duas categorias: dependência assintótica e independência assintótica. A variável $\mathbf{Y}_{-i}$ é assintoticamente dependente e assintoticamente independente da variável Y_i quando o limite

$$\lim_{y\to\infty}\{Pr(\mathbf{Y}_{-i} > y|Y_i > y\}$$

é diferente de zero ou igual a zero, respectivamente. Aqui $\mathbf{Y}_{-i}$ denota o vetor $\mathbf{Y}$ excluído a componente Y_i e $\mathbf{y}$ é um vetor de valores y. Todos os métodos existentes para valores extremos multivariados são apropriados para estimar $Pr(\mathbf{X} \in C)$ sob dependência assintótica ou independência assintótica associada a $\mathbf{Y}$, dado que todo $\mathbf{x} \in C$ é extremo em todas as componentes.

Heffernan and Tawn (2004) apresentam uma abordagem para valores extremos multivariados que consiste em uma mudança na direção dos métodos de valores extremos existentes. A estratégia do modelo é baseada na suposição da forma assitótica da distribuição condicional da variável, dado que esta tem uma componente extrema, i.e., a distribuição de $\mathbf{Y}_{-i} \mid Y_i = y_i$ quando y_i torna-se grande. Essa abordagem condicional fornece uma extensão natural do modelo da distribuição condicional generalizada de Pareto (1.2), para um caso multivariado como $Pr(\mathbf{X} \in C_i|X_i > v_{X_i})$ pode ser expresso como

$$Pr(\mathbf{X} \in C_i|X_i > v_{X_i}) = \int_{v_{X_i}}^{x^{F_{X_i}}} Pr(X_i \in C_i|X_i = x)d\widehat{F}_{X_i}(x)/(1-\widehat{F}_{X_i}(v_{X_i})), \qquad (1.5)$$

onde a integral é estimada usando a distribução $\mathbf{Y}_{-i}|Y_i = y_i$ depois da transformação marginal. Quando $v_{X_i} > u_{X_i}$, a derivada de $\widehat{F}_{X_i}(x)/(1-\widehat{F}_{X_i}(v_{X_i}))$ é uma função densidade da distribuição generalizada de Pareto com os parâmetros escala e forma iguais a $\beta_i + \xi_i(v_{X_i} - u_{X_i})$ e ξ_i, respectivamente.

Essa abordagem condicional se aplica às variáveis que são assintoticamente dependentes ou assintoticamente independentes; ela pode ser usada para estimar $Pr(\mathbf{X} \in C)$ para qualquer conjunto extremo C e também pode ser aplicada a uma variável de qualquer dimensão.

O modelo proposto pelos autores para a distribuição condicional foi motivado pela suposição da distribuição assintótica e é auxiliada por uma classe de exemplos teóricos. O

modelo é semiparamétrico; uma regressão paramétrica é usada para estimar os parâmetros de forma e escala das distribuições marginais da cauda; e métodos não paramétricos são usados para estimar a estrutura residual multivariada.

Embora nessa abordagem falte uma completa caracterização assintótica da estrutura de probabilidade, como as que justificam os métodos dos valores extremos existentes, ela apresenta uma grande força matemática e vantagens práticas em comparação com os métodos de valores extremos existentes.

Capítulo 2

Motivações teóricas

Aqui, apresentamos grandes resultados teóricos que motivaram a escolha do modelo estatístico. Primeiramente, na Seção 2.1, fizemos uma suposição sobre a forma assintótica da distribuição condicional e examinamos as conseqüências desta suposição. Na Seção 2.2, identificamos as funções normalizadoras que devem satisfazer as condições subjacentes. Na Seção 2.3 discutimos alguns exemplos teóricos os quais sugerem que a suposição assintótica é apropriada para uma grande classe de distribuições. Apresentamos também uma classe de funções normalizadoras, que é pequena, ao passo que a classe das distribuiçòes limite é grande.

2.1 Suposição de uma representação do limite e suas propriedades

Considere a estrutura assintótica das distribuições condicionais estabelecidas a partir de uma variável aleatória d-dimensional $\mathbf{Y} = (Y_1, \ldots, Y_d)$ com distribuições marginais Gumbel. Para cada $i = 1, \ldots, d$, examinamos a distribuição condicional $Pr(\mathbf{Y}_{-i} \leq y_{-i} | Y_i = y_i)$, onde aqui, a álgebra dos vetores será sempre componente-a-componente. Para examinar o comportamento no limite destas distribuições quando $y_i \to \infty$ requeremos que a distribuição limite seja não degenerada em todas as marginais e assim precisamos controlar o crescimento de y_{-i} de acordo com a dependência de $\mathbf{Y}_{-i}$ em Y_i.

Especificamente, assumimos que: dado i existem funções normalizadoras vetoriais $\mathbf{a}_{|i}(y_i)$ e $\mathbf{b}_{|i}(y_i)$, isto é, R^{d-1}, os quais podem ser escolhidos tais que, para todo $z_{|i}$ fixo e para qualquer seqüência de valores y_i tal que, quando $y \to \infty$,

$$\lim_{y\to\infty}(Pr\{\mathbf{Y}_{-i} \leq \mathbf{a}_{|i}(y_i) + \mathbf{b}_{|i}(y_i)\mathbf{z}_{|i}|Y_i = y_i\}) = G_{|i}(\mathbf{z}_{|i}), \tag{2.1}$$

onde todas as distribuições marginais limite $G_{|i}$ são não degeneradas.

Uma expressão alternativa dessa suposição que tem uma interpretação estatística mais fácil, é que as variáveis padronizadas

$$\mathbf{Z}_{|i} = \frac{\mathbf{Y}_i - \mathbf{a}_{|i}(y_i)}{\mathbf{b}_{|i}(y_i)} \tag{2.2}$$

têm a propriedade

$$\lim_{y_i\to\infty}\{Pr(\mathbf{Z}_{|i} \leq \mathbf{z}_{|i} \mid Y_i = y_i)\} = G_{|i}(\mathbf{z}_{|i}), \tag{2.3}$$

onde todas as distribuições limite $G_{|i}$ são não degeneradas.

Sob a suposição (2.1) ou, equivalentemente sob a suposição (2.3), temos que, condicionalmente a $Y_i > u_i$, quando $u_i \to \infty$ as variáveis $Y_i - u_i$ e $\mathbf{Z}_{|i}$ são independentes no limite com $Y_i - u_i$ sendo exponencial e $\mathbf{Z}_{|i}$ sendo $G_{|i}(\mathbf{z}_{|i})$. Para ver que este resultado se verifica, seja $y_i = u_i + y$ com $y > 0$ fixo, então

$$\begin{aligned}
Pr(\mathbf{Z}_{|i} \leq \mathbf{z}_{|i}, Y_i - u_i = y \mid Y_1 > u_1) &= Pr(\mathbf{Z}_{|i} \leq \mathbf{z}_{|i} \mid Y_i = u_i + y).\frac{Pr(Y_i = u_i + y \mid Y_1 > u_1)}{Pr(Y_1 > u_1)} \\
&= Pr(\mathbf{Z}_{|i} \leq \mathbf{z}_{|i} \mid Y_i = u_i + y).\frac{f_{Y_i}(u_i + y)}{Pr(Y_1 > u_1)} \\
&= Pr(\mathbf{Y}_{-i} \leq \mathbf{a}_{|i}(y + u_i) + \mathbf{b}_{|i}(y + u_i)\mathbf{z}_{|i} \mid Y_i = u_i + y).\frac{f_{Y_i}(u_i + y)}{Pr(Y_i > u_i)} \\
&\to G_{|i}(z_{|i})\exp(-y) \text{ quando } u_i \to \infty,
\end{aligned}$$

ou seja,

$$Pr(\mathbf{Z}_{|i} \leq \mathbf{z}_{|i}, Y_i - u_i = y \mid Y_1 > u_1) \to G_{|i}(z_{|i})\exp(-y) \text{ quando } u_i \to \infty, \tag{2.4}$$

onde f_{Y_i} é a função de densidade marginal de Y_i. O resultado final de convergência segue do fato de que a cauda de uma Gumbel é exponencial (isto é, para um limiar suficientemente grande o excesso de uma variável aleatória Gumbel segue uma distribuição esponencial) e, de (2.1), e não depende de como $y_i \to \infty$.

Considerando agora as distribuições marginais e a estrutura de dependência das distribuições $G_{|i}(\mathbf{z}_{|i})$, para cada $j \neq i$, definimos $G_{j|i}(z_{j|i})$ como sendo a distribuição limite da distribuição condicional de

$$Z_{j|i} = \frac{Y_j - a_{j|i}(y_i)}{b_{j|i}(y_i)} \text{ dado } Y_i = y_i \text{ quando } y_i \to \infty,$$

onde cada $a_{j|i}(y_i)$ são componentes de $\mathbf{a}_{|i}(y_i)$ e cada $b_{j|i}(y_i)$ são componentes de $\mathbf{b}_{|i}(y_i)$ associados com a variável Y_j. Assim, $G_{j|i}$ é a distribuição marginal de $G_{|i}$ associada à variável Y_j. Se

$$G_{|i}(\mathbf{z}_{|i}) = \prod_{j \neq i} G_{j|i}(z_{j|i}),$$

então dizemos que os elementos de $\mathbf{Y}_{-i}$ são assintoticamente independentes dado a condicional Y_i.

2.2 Escolha das funções normalizadoras

Agora identificamos as funções normalizadoras $\mathbf{a}_{|i}(y_i)$ e $\mathbf{b}_{|i}(y_i)$ de acordo com as características da distribuição condicional de $\mathbf{Y}_{-i}|Y_i$, de tal forma que essas funções possam ser identificadas nos exemplos teóricos. As funções normalizadoras e a distribuição limite não são únicas no seguinte sentido: se as funções normalizadoras $\mathbf{a}_{|i}(y_i)$ e $\mathbf{b}_{|i}(y_i)$ resultam em uma distribuição limite não-degenerada $G_{|i}(\mathbf{z}_{|i})$, então usando as funções normalizadoras

$$\begin{aligned} \mathbf{a}^*_{|i}(y_i) &= \mathbf{a}_{|i}(y_i) + \mathbf{A}\mathbf{b}_{|i}(y_i)e \\ \mathbf{b}^*_{|i}(y_i) &= \mathbf{B}\mathbf{b}_{|i}(y_i), \end{aligned} \tag{2.5}$$

com $\mathbf{A}$ e $\mathbf{B}$ sendo vetores constantes arbitrários, com $\mathbf{B} > 0$, temos como resultado uma distribuição limite $G_{|i}(\mathbf{B}\mathbf{z}_{|i} + \mathbf{A})$ também não-degenerada. Isto é, as constantes normalizadoras e a distribuição limite não degenerada são únicas a menos de uma transformação afim nas constantes normalizadoras.

Verificando o resultado acima:

$$\begin{aligned} Pr\{\mathbf{Y}_{-i} \leq \mathbf{a}^*_{|i}(y_i) + \mathbf{b}^*_{|i}(y_i)\mathbf{z}_{|i}|Y_i = y_i\} &= Pr\{\mathbf{Y}_{-i} \leq \mathbf{a}_{|i}(y_i) + A\mathbf{b}_{|i}(y_i) + B\mathbf{b}_{|i}(y_i)\mathbf{z}_{|i}|Y_i = y_i\} \\ &= Pr\{\mathbf{Y}_{-i} \leq \mathbf{a}_{|i}(y_i) + \mathbf{b}_{|i}(y_i)(A + B\mathbf{z}_{|i})|Y_i = y_i\} \end{aligned}$$

usando (2.1) $\rightarrow G_{|i}(A + B\mathbf{z}_{|i})$ quando $y_i \rightarrow \infty$.

Contudo, de acordo com os argumentos usuais, tais como aqueles usado por Leadbetter, Lindgren, and Rootzén (1983), página 7, esta é o única maneira que dois limites diferentes sem massa podem aparecer. Assim temos que a classe de distribuições limite é única, a menos de transformações afim, e que as funções normalizadoras podem ser identificadas a menos das constantes $\mathbf{A}$ e $\mathbf{B}$ na expressão (2.5).

Para i fixo, a escolha das funções vetoriais $\mathbf{a}_{|i}(y_i)$ e $\mathbf{b}_{|i}(y_i)$ pode ser feita para cada componente, resultando em $d-1$ condições baseadas no comportamento limite de $Y_j|Y_i = y_i$, para cada $j \neq i$, já que a suposição vista em (2.1) especifica que cada distribuição marginal de $G_{|i}$ deve ser não-degenerada. Assim, estamos interessados na distribuição condicional de $Y_j|Y_i = y_i$, a qual denotamos por $F_{j|i}(y_j|y_i)$. As funções *harzand* condicionais $h_{j|i}$ associadas às distribuições univariadas condicionais $F_{j|i}(y_j|y_i)$ são dada por

$$h_{j|i}(y_j|y_i) = \frac{f_{j|i}(y_j|y_i)}{1 - F_{j|i}(y_j|y_i)}, \text{ para } -\infty < y_j < \infty,$$

onde $f_{j|i}(y_j|y_i)$ é a densidade condicional de $Y_j|Y_i = y_i$.

Teorema 1. Suponha que o vetor aleatório $\mathbf{Y}$ tem densidade conjunta absolutamente contínua. Se, para um determinado i, as funções vetoriais $\mathbf{a}_{|i}(y_i)$ e $\mathbf{b}_{|i}(y_i)$ satisfazem a propriedade limite (2.1) ou, equivalentemente (2.3), então as componentes destas funções vetoriais correspondendo às variáveis Y_j, para cada $j \neq i$, satisfazem, a menos de transformações afim, as propriedades (2.6) e (2.7) dadas por:

$$\lim_{y_i \rightarrow \infty} (F_{j|i}\{a_{j|i}(y_i)|y_i\}) = p_{j|i}, \tag{2.6}$$

onde $p_{j|i}$ é uma constante no intervalo $(0, 1)$ e

$$b_{j|i}(y_i) = h_{j|i}\{a_{j|i}(y_i)|y\}^{-1}. \tag{2.7}$$

Devido à flexibilidade na forma da função normalizadora dada pela expressão (2.5), uma simplificação da estrutura das funções normalizadoras pode ser obtida, como ilustrado pelo *Corolário* 1.

Corolário 1. Se as funcões $\mathbf{a}_{|i}(y_i)$ e $\mathbf{b}_{|i}(y_i) > 0$ satisfazem as condições do Teorema 1 e se existe uma constante $s_{j|i} < \infty$ tal que

$$\lim_{y_i \rightarrow \infty} \left[\frac{a_{j|i}(y_i)}{b_{j|i}(y_i)}\right] = s_{j|i},$$

então a relação limite de (2.1) se verifica com $a_{j|i}(y_i) = 0$. Mais ainda; se $b_{j|i}(y_i) = t_{j|i}k_{j|i}(y_i)$ para qualquer constante $t_{j|i} > 0$ independente de y_i, e $k_{j|i}(y_i)$ qualquer função de y_i, então a relação de limite (2.1) se verifica com $b_{j|i}(y_i)$ substituído por $k_{j|i}(y_i)$.

2.3 Exemplos teóricos

Para uma seleção de distribuições multivariadas com distribuições marginais Gumbel, iremos apresentar as funções normalizadoras $\mathbf{a}_{|i}(y)$ e $\mathbf{b}_{|i}(y)$, dadas pelo *Teorema* 1 e *Corolário* 1, e algumas propriedades das respectivas distribuições limite condicionais não-degeneradas $G_{|i}$. Estes exemplos teóricos foram selecionados de forma a representar as quatro classes de dependência extremal que foram indicadas no Capítulo 2. A primeira classe se refere a distribuições assintoticamente dependentes e as outras três são distribuições assintoticamente independentes, apresentando:

- dependência extremal negativa ($0 < \eta_{\mathbf{Y}} < 1/d$);
- quase independência extremal ($\eta_{\mathbf{Y}} = 1/d$);
- dependência extremal positiva ($1/d < \eta_{\mathbf{Y}} < 1$).

Como a dependência entre os pares determina cada uma das componentes das funções normalizadoras, nós apresentamos os resultados de acordo com o valor do coeficiente de dependência de cauda η_{ij} de cada par (Y_i, Y_j). O valor de $\eta_{ij} = 1/2$ corresponde ao caso de quase independência do par (Y_i, Y_j).

A Tabela 2.1 apresenta dois exemplos de cada uma das quatro classes. Os casos especiais de dependência extremal perfeita, negativa ($\eta = 0$) ou positiva ($\eta = 1$), estão incluídos aqui para identificar limites inferiores e superiores para o comportamento das funções normalizadoras, embora rigorosamente os métodos apresentados no capítulo anteior não se apliquem a essas duas distribuições, pois para cada uma delas a distribuição condicional é degenerada. Neste estágio o interesse reside somente na estrutura das funções normalizadoras e nas distribuições limite. Além disso, como as distribuições limite bivariadas $G_{|i}$ são freqüentemente complicadas, neste estágio identificamos somente

Tabela 2.1: Exemplos de estruturas de dependência multivariada classificados pelo comportamento de dependência extremal.

Estrutura de Dependência	η_{ij}	$a_{j\|i}(y)$	$b_{j\|i}(y)$	$G_{j\|i}$
i. Perfeita	1	y	1	Degenerada
ii. VEM	1	y	1	complexa
iii. Gaussiana ($\rho_{ij} > 0$)	$(1+\rho_{ij})/2$	$\rho_{ij}^2 y$	$y^{1/2}$	Gaussiana
iv. VEM cauda leve	$2^{-\alpha}$	0	$y^{1/2}$	Gumbel
v. Morgenstern	0.5	0	1	Gumbel
vi. Independente	0.5	0	1	Gumbel
vii. Gaussiana ($\rho_{ij} < 0$)	$(1+\rho_{ij})/2$	$-log(\rho_{ij}^2 y)$	$y^{1/2}$	Gaussiana
viii. Dependência perfeita negativa	0	$-log(y)$	1	Degenerada

as distribuições marginais univariadas $G_{j|i}$ e dizemos se (ou não) as marginais da $G_{|i}$ são independentes.

Os exemplos listados na Tabela 2.1 satisfazem (2.1), têm estrutura simples para as funções normalizadoras e resultam em várias distribuições limite $G_{|i}$ que não estão contidas em nenhuma família de distribuição simples. As funções normalizadoras são todas casos epeciais da família paramétrica

$$\mathbf{a}_{|i}(y) = \mathbf{a}_{|i} y - I_{\{\mathbf{a}_{|i}=0, \mathbf{b}_{|i}<0\}}\{\mathbf{c}_{|i} - \mathbf{d}_i \log(y)\},$$
$$\mathbf{b}_{|i}(y_i) = y^{\mathbf{b}_{|i}}, \tag{2.8}$$

onde $\mathbf{a}_{|i}$, $\mathbf{b}_{|i}$, $\mathbf{c}_{|i}$ e $\mathbf{d}_i$ são constantes em R^{d-1} e I é uma função indicadora. As constantes vetoriais têm componentes tais que $a_{j|i}$, $b_{j|i}$, $c_{j|i}$ e $d_{j|i}$ são constantes com $0 \leq a_{j|i} \leq 1$, $-\infty < b_{j|i} < 1$, $-\infty < c_{j|i} < \infty$ e $0 \leq d_{j|i} \leq 1$ para todo $i \neq j$. Para $\mathbf{a}_{|i}(y)$ de acordo com o par sendo positivamente ou negativamente associado, a família paramétrica tem formulação estrutural diferente e isto é devido à assimetria da distribuição marginal Gumbel, que tem cauda superior mais pesada que a inferior. Se fosse simétrica, o efeito da locação para dependência positiva ou dependência negativa seria o mesmo.

A construção das operações limite que resultam nas funções normalizadoras e nas distribuições limite não garante a continuidade dessas funções normalizadoras nem da função distribuição, conforme o parâmetro da distribuição original se modifica. Dois exemplos ilustram esse ponto, em caso particular quando os parâmetros da distribuição subjacente se aproximam dos valores correspondentes a independência. Um caso especial da distribuição ii é a distribuição bivariada de valores extremos com estrutura de dependência logística [1], é assintoticamente dependente quando o parâmetro α está entre 0 e 1 (mais detalhes em Heffernan and Tawn (2004)). Quando $\alpha = 1$, as variáveis são independentes. Conseqüentemente, a normalização que é requerida é descontínua em $\alpha = 1$. Contudo, quando $\alpha \to 1$ a distribuição limite de $G_{j|i}$ coloca quase toda a massa de probabilidade cada vez mais perto de $-\infty$, indicando que a locação relacionada com a locação está se tornando muito forte. Similarmente, a distribuição normal multivariada iii resulta $G_{j|i}$ como uma normal com variância $2\rho_{ij}^2(1-\rho_{ij}^2)$; assim, quando o $\rho_{ij} \to 0$ o limite é degenerado e a normalização relacionada com a escala torna-se mais forte.

[1] Gumbel (1960) introduziu uma mudança na distribuição logística multivariada

$$V^d(\mathbf{y}) = (\sum_{j=1}^{d} y_j^{-1/\alpha})^{\alpha},$$

para qualquer $d \geq 2$ e $0 < \alpha \leq 1$. A independência ocorre quando $\alpha = 1$ e a independência positiva perfeita ocorre no limite quando $\alpha \to 0$.

Capítulo 3

Estrutura do modelo e suas propriedades

Na Seção 3.1 apresentamos um modelo semiparamétrico para descrever a dependência dos valores extremos em problemas multivariados. Esse modelo é apresentado para variáveis com distribuições marginais univariadas Gumbel. Combinado com o nosso modelo marginal, descrito no Capítulo 1, esse modelo resulta em um modelo completo para os valores extremos de um vetor aleatório $\mathbf{X}$. Os problemas relacionados com a *self-consistency* dos vários modelos condicionais são discutidos na Seção 3.2.

3.1 Modelo de dependência condicional

A estrutura do modelo é motivada pelos resultados do Capítulo 2. Para obter o comportamento de $\mathbf{Y}_{-i}$ com Y_i grande, usamos a formulação da distribuição condicional limite, obtida em (2.1), e sua propriedade de independência no limite, obtida em (2.4). Iremos assumir que para cada i existe um limiar u_{Y_i} grande para o qual modelamos

$$Pr\{(\mathbf{Y}_{-i} \leq \mathbf{a}_{|i}(y_i) + \mathbf{b}_{|i}(y_i)\mathbf{z}_{|i})|Y_i = y_i\} = G_{|i}(\mathbf{z}_{|i}), \quad \text{para todo } y_i > u_{Y_i},$$

onde o $\mathbf{Z}_{|i}$ é o resíduo padronizado definido pela expressão (2.2) com função distribuição $G_{|i}$, e $\mathbf{Z}_{|i}$ é independente de Y_i dado $y_i > u_{Y_i}$. O comportamento da dependência extremal é então caracterizado pelas funções: i) locação $\mathbf{a}_{|i}(y_i)$; ii) escala $\mathbf{b}_{|i}(y_i)$; e iii) função

distribuição $G_{|i}$.

Consideramos primeiro a especificação dos modelos condicionais individuais, i.e., $\mathbf{a}_{|i}(y_i)$, $\mathbf{b}_{|i}(y_i)$ e $G_{|i}(\mathbf{z}_{|i})$ para um dado i. Foi adotado o modelo paramétrico visto em (3.8), que é uma família paramétrica simples de funções normalizadoras, as quais são apropriadas para várias classes de exemplos teóricos que são apresentados na Tabela 2.1. Adotamos os parâmetros de $\mathbf{a}_{|i}(y_i)$ e $\mathbf{b}_{|i}(y_i)$ por $\theta_{|i} = (\mathbf{a}_{|i}, \mathbf{b}_{|i}, \mathbf{d}_{|i}, \mathbf{d}_{|i})$ e adotamos a convenção de que $c_{j|i} = d_{j|i} = 0$ exceto se $a_{j|i} = 0$ e $b_{j|i} < 0$. Discutimos a estimação de $\theta_{|i}$, denotando o estimador de $\theta_{|i}$ por $\hat{\theta}_{|i}$, e os estimadores correspondentes das funções normalizadoras por $\hat{\mathbf{a}}_{|i}$ e $\hat{\mathbf{b}}_{|i}$. Como (3.1) não impõe nenhuma estrutura em $G_{|i}$, adotamos um modelo não-paramétrico para $G_{|i}$. Estimamos essa distribuição usando a distribuição empírica da variável aleatória $\hat{\mathbf{Z}}_{|i}$, definida por

$$\hat{\mathbf{Z}}_{|i} = \frac{\hat{\mathbf{Y}}_{-i} - \hat{\mathbf{a}}_{|i}(y_i)}{\hat{\mathbf{b}}_{|i}(y_i)}, \quad \text{para } Y_i = y_i > u_{Y_i}.to$$

Os exemplos teóricos sugerem que os $\mathbf{Z}_{|i}$ são em geral condicionalmente assintoticamente independentes. Então sugerimos modelar as componentes de $\hat{Z}_{|i}$ como sendo independentes i.e., $\hat{G}_{|i}(\mathbf{z}_{|i}) = \prod_{j=i} \hat{G_{j|i}}(\mathbf{z}_{j|i})$, onde $G_{j|i}$ é a função distribuição empírica dos $_{|i}$.

Em resumo, para cada i nosso modelo de dependência é um modelo de regressão semiparamétrico multivariado da forma

$$\mathbf{Y}_{-i} = \mathbf{a}_{|i}(y_i) + \mathbf{b}_{|i}(y_i)\mathbf{Z}_{|i}, \quad \text{para } Y_i = y_i > u_{Y_i}, \tag{3.3}$$

onde $\mathbf{a}_{|i}(y_i)$ e $\mathbf{b}_{|i}(y_i)$ são obtidos através do modelo paramétrico (2.8) e a distribuição de resíduos é modelada de forma não-paramétrica. O conjunto total de parâmetros do modelo será representado por $\Theta = (\theta_{|1}, .., \theta_{|d})$. Cada modelo de regressão se aplica apenas além do limiar μ_{Y_i}, para o qual a estrutura de dependência está bem descrita por (3.3).

Categorizamos a estrutura de dependência que é dada pelo modelo (3.3) através do uso das quatro classes que identificam o comportamento dos quantis da distribuição de $Y_j \mid Y_i = y_i$ quando $y_i \to \infty$. Se os quantis da distribuição condicional crescem à mesma

taxa que y_i, i.e., $a_{j|i} = 1$ e $b_{j|i} = 0$, então as variáveis (Y_i, Y_j) são assintoticamente dependentes. Para as distribuições assintoticamente independentes, os quantis condicionais tendem para: ∞, um limite finito ou para $-\infty$ quando $y_i \to y$, se (Y_i, Y_j) exibir dependência extremal positiva, quase independência extremal ou dependência extremal negativa, respectivamente. Deste modo, as variáveis exibem: dependência extremal positiva quando $0 < a_{j|i} < 1$ ou $b_{j|i} > 0$; quase independência extremal ocorre quando $c_{j|i} = d_{j|i} = 0$ e $b_{j|i} \leq 0$; e dependência extremal negativa ocorre quando $a_{j|i} = 0$, $d_{j|i} > 0$ e $b_{j|i} < 0$.

Tratando os d modelos condicionais separadamente obtém-se a versão mais geral do nosso modelo com parâmetro Θ, onde este é um vetor de tamanho $4d(d-1)$, embora para cada par, $c_{j|i}$ e $d_{j|i}$ só serão não-nulos se não existir associação positiva entre o par (Y_j, Y_i). Existem muitas distribuições multivariadas cuja estrutura de dependência é permutável de alguma maneira. A forma mais comum de permutabilidade é de par-a-par, i.e., Y_i depende de Y_j da mesma maneira que Y_j depende de Y_i. Dizemos que as variáveis Y_i e Y_j exibem fraca permutabilidade extremal se $\theta_{j|i} = \theta_{i|j}$, e forte permutabilidade extremal se adicionalmemte $G_{j|i} = G_{i|j}$.

3.2 Diagnósticos

Os exemplos da Seção 2.3 indicam que a taxa de convergência da distribuição condicional de $\mathbf{Y}_{-i} \mid Y_i = y$, quando $y \to \infty$, para a distribuição limite pode ser lenta. Porém a forma de convergência para a distribuição condicional limite é usada somente para motivar nossa estrutura do modelo e não estamos interessados nos verdadeiros valores limite de $\theta_{|i}$ e $G_{|i}$. O que é importante na prática é se a distribuição condicional da variável normalizada $\mathbf{Z}_{|i}$ é estável em relação às escolhas dos limiares.

Essa exigência sugere que os diagnósticos para que nossa estrutura do modelo sejam baseados na análise de sensibilidade de vários outros limiares. Para estimação marginal usamos diagnósticos que são baseados no gráfico da função média dos excessos; veja Smith (1989). Para a estimação do modelo de dependência, uma suposição fundamental para a modelagem é que $\mathbf{Z}_{|i}$ é independente de Y_i dado $Y_i > u_{Y_i}$, para um limiar u_{Y_i}

grande para cada i. Ajustando o modelo condicional sobre vários outros limiares grandes, a estabilidade dos valores estimados para $\theta_{|i}$ e as extrapolações resultantes podem ser verificados. Além disso, vários outros testes padrão que verificam independência podem ser aplicados para $\mathbf{Z}_{|i}$ para determinar se as variáveis podem ser tratadas como sendo condicionalmente assintoticamente independentes.

Capítulo 4

Inferência

O modelo geral inclui o modelo da distribuição marginal visto em (1.3) e o modelo de dependência (2.1). Ambos os modelos são semiparamétricos, ou seja, uma parte é modelada através de parâmetros e a outra não; nessa segunda parte nenhum modelo paramétrico apropriado foi encontrado.

As inferências sobre a estrutura marginal e sobre a estrutura de dependência foram feitas através de uma série de passos: primeiro o parâmetro ψ da distribuição marginal foi estimado e depois o parâmetro θ da distribuição de dependência foi estimado assumindo que os parâmetros da marginal eram conhecidos. O passo da estimação é muito mais simples que a estimação conjunta de todos os parâmetros. Encontramos em Shi, Smith, and Coles (1992) a sugestão de que a perda de eficiência em relação a estimação conjunta é provavelmente pequena a menos que os valores de ξ_i, $i = 1, 2, \ldots, d$ difiram grandemente.

Breves detalhes do passo da estimação marginal são dados na Seção 4.1. Seguindo a estimação marginal, os dados foram transformados em uma distribuição com marginais Gumbel através da transformação (1.4) com ψ substituído pelo valor estimado $\hat{\psi}$. Descrevemos o porquê de usarmos a estimação gaussiana para os parâmetros das funções normalizadoras $\theta_{|i}$ para cada distribuição condicional separada sob a suposição de que não há nenhuma restrição entre $\theta_{|i}$ e $\theta_{|j}$ para quaisquer i e j. O ajuste de submodelos requer a estimação conjunta de todos os parâmetros θ do modelo condicional. Discutimos uma abordagem para estimação conjunta que tem semelhança com a pseudoverossimilhança de Besag (1975) e apresentamos técnicas para avaliar a incerteza na estimação do

modelo geral (global). Assumimos que os dados são realizações de variáveis aleatórias $X_1, \ldots, X_n$ independentes e identicamente distribuídas.

4.1 Estimação marginal

Estima-se as d distribuições marginais univariadas ignorando a dependência entre as componentes. Especificamente, assumimos independência entre as componentes da variável construindo a função de log-verossimilhança

$$\log\{L(\Psi)\} = \sum_{i=1}^{d} \sum_{k=1}^{\eta_{u_{X_i}}} \log\{\hat{f}_{X_i}(x_{i|i,k})\}, \tag{4.1}$$

onde $\hat{f}_{X_i}$ é a densidade que é associada com a distribuição (1.3), $\eta_{u_{X\,i}}$ é o número de observações com o i-ésima componente excedendo ao limiar da marginal $u_{X\,i}$ e a j-ésimo componente de cada k-ésima observação é denotada por $x_{j|i,k}$ com $j = 1, \ldots, d$; $k = 1, \ldots, \eta_{u_{X\,i}}$. Se não existe nenhuma função de ligação entre os parâmetros das várias componentes (4.1) então a maximização da log-verossimilhança (4.1) é equivalente a ajustar a GPD para os excessos acima de um limiar para cada marginal. Quando existe alguma ligação entre os parâmetros das marginais, a maximização conjunta da função log-verossimilhança (4.1) é mais eficiente.

4.2 Modelos condicionais

Para cada i, desejamos estimar $\theta_{|i}$ sob mínimas suposições de $G_{|i}$. Se assumirmos que $\mathbf{Z}_{|i}$ tem dois momentos finitos, então $\theta_{|i}$ determina as médias marginais e variância da variável condicional $\mathbf{Y}_{-i} \mid Y_i = y_i$ quando $y_i > u_{Y_i}$. Especificamente se $\mathbf{Z}_{|i}$ tem médias marginais e a desvios-padrão denotados pelos vetores $\mu_{|i}$ e $\sigma_{|i}$, respectivamente, então as variáveis aleatórias $\mathbf{Y}_{-i} \mid Y_i = y_i$, para $y > u_{Y_i}$, têm vetor média e desvio padrão respectivamente dados por

$$\mu_{|i}(y) = \mathbf{a}_{|i}(y) + \mu_{|i}\mathbf{b}_{|i}(y),$$
$$\sigma_{|i}(y) = \sigma_{|i}\mathbf{b}_{|i}(y),$$

ambos sendo funções de y, $\theta_{|i}$ e das constantes $\lambda_{|i} = (\mu_{|i}, \sigma_{|i})$. Esses $(\theta_{|i}, \lambda_{|i})$ são os parâmetros dos modelos de uma regressão linear multivariados com variância não constante e distribuição dos erros não especificada. Especificamente, maximizamos a função objetivo associada sob o espaço paramétrico para produzir um estimador pontual válido para $\theta_{|i}$. O modelo paramétrico para $G_{|i}$ é escolhido por conveniência e por simplicidade computacional. Assumimos que as componentes de $\mathbf{Z}_{|i}$ são independentes e normais e conseqüentemente, nossa inferência sobre $\theta_{|i}$ é baseada na estimação da normal. A simplificação da independência torna-se razoável ao passo que os $\theta_{|i}$ determinam somente as características marginais da distribuição condicional. Heffernan and Tawn (2004) consideraram uma série de distribuições paramétricas para as marginais de $\mathbf{Z}_{|i}$ e selecionaram a distribuição normal por ser mais simples e ter uma boa performance e nas simulações feitas.

Portanto, a função objetivo que foi usada para estimação pontual de $\theta_{|i}$ e $\lambda_{|i}$ é

$$Q_{|i}(\theta_{|i}, \lambda_{|i}) = \sum_{j \neq i} \sum_{k=1}^{n_{u_{Y_i}}} \left(\log \left\{ \sigma_{j|i} \left(y_{i|i,k} \right) \right\} + \frac{1}{2} \left\{ \frac{y_{j|i,k} - \mu_{j|i}(y_{i|i,k})}{\sigma_{j|i}(y_{i|i,k})} \right\}^2 \right), \tag{4.2}$$

onde a notação segue as convenções que são adaptadas no Capítulo 2 e para log-verossimilhança (4.1). Maximizamos $Q_{|_i}$ conjuntamente com respeito a $\theta_{|i}$ e $\lambda_{|i}$ para obter a estimativa pontual de $\hat{\theta}_{|i}$. Ajustamos o modelo de dependência em dois estágios: primeiro fixamos $c_{j|i} = d_{j|i} = 0$. Só estimamos $c_{j|i}$ e $d_{j|i}$ se $\hat{a}_{j|i} = 0$ e $\hat{b}_{j|i} < 0$.

Agora consideramos a estimação conjunta dos parâmetros θ do modelo condicional. Por razões que são similares àquelas discutidas até aqui, assumimos independência entre as distribuições condicionais para obter a função objetivo:

$$Q(\theta, \lambda) = \sum_{i=1}^{d} Q_{|i}(\theta_{|i}, \lambda_{|i}), \tag{4.3}$$

onde $Q_{|i}(\theta_{|i}, \lambda_{|i})$ está expresso em (4.2) e $\lambda = (\lambda_{|1}, \lambda_{|2}, \ldots, \lambda_{|i})$.

4.3 Incerteza

A incerteza aparece através da estimação do modelo marginal semiparamétrico, das funções normalizadoras paramétricas da estrutura de dependência condicional e nos mo-

delos não-paramétricos da distribuição dos resíduos padronizados. O método bootstrap semiparamétrico é usado para calcular a incerteza e para estimar os erros-padrão das estimativas dos parâmetros e de outras quantidades tais como $Pr(\mathbf{X} \in C)$. Assumimos que os limiares do modelo marginal e do modelo de dependência são fixos, portanto as incertezas relacionadas a estas escolhas não são verificadas pelo bootstrap.

O bootstrap tem três passos: generação de dados de acordo com o modelo ajustado, estimação do modelo paramétrico, e cálculo das quantidades de interesse. Estes estágios são repetidos independentemente. A novidade do algoritmo está na geração dos dados. Para garantir que as amostras obtidas pelo bootstrap conservam as características marginais e de dependência dos dados, usamos um algoritmo em dois passos para a geração dos dados. Primeiro um bootstrap não-paramétrico é empregado, garantindo a preservação da estrutura de dependência; depois um passo paramétrico é usado nas marginais.

O procedimento é como segue. Os dados originais são primeiramente tranformados em uma variável aleatória com marginais Gumbel, usando o modelo visto em (1.3), que é estimado através os dados originais. Uma amostra bootstrap não-paramétrica é então obtida pela amostragem com reposição dos dados transformados. Trocamos então os valores da marginal por essa amostra bootstrap, garantindo que as distribuições marginais são todas Gumbel e preservando as associações entre os pontos ranqueados em cada componente. Especificamente para cada $i = 1, \ldots, d$ substituímos a amostra ordenada da componente Y_i por uma amostra do mesmo tipo da distribuição Gumbel padrão. O resultado da amostra é então transformado, voltando a ter marginais originais pelo uso do modelo marginal que foi estimado com os dados originais. Os dados que são gerados pelo uso dessa abordagem têm distribuições marginais univariadas com caudas superiores simuladas pelo ajuste do modelo GPD e estrutura de dependência totalmente consistente entre os *ranks* das componentes das variáveis.

Capítulo 5

Análise da vazão de três rios de Porto Rico

Os dados representam as médias diárias dos fluxos de três rios: Canóvanas[1], Espiritu Santo[2] e Fajardo[3], situados na região nordeste de Porto Rico. Os dados são medidos em pés cúbicos por segundo (equivale a 23,317 litros por segundos). As médias diárias estudadas aqui compreendem o período de abril de 1967 a novembro de 2003, inclusive, resultando em uma amostra de tamanho 13332. As médias diárias são disponibilizadas pela *National Water Information Sistem*, *http://waterdata.usgs.gov*.

O U.S.Geological é a maior agência da ciência natural da nação e tem a maior responsabilidade dentro do governo federal para: representar a nação; avaliar seus recursos e perigos geológicos; compreender seus recursos biológicos naturais; e fornecer informações hidrológicas para estimação dos riscos hidrológicos à nação. O centro da ciência da água situado no Cararibe é envolvido com as investigações dos recursos das águas em Porto Rico.

O USGS investiga a ocorrência, a quantidade, a qualidade, a distribuição e o movi-

[1]http://nwis.waterdata.usgs.gov/nwis/discharge?site_no=50061800&agency_cd=USGS&format=rdb&begin_date=04/01/1967&end_date=09/30/2003&period=

[2]http://nwis.waterdata.usgs.gov/nwis/discharge?site_no=50063800&agency_cd=USGS&format=rdb&begin_date=04/01/1967&end_date=09/30/2003&period=

[3]http://nwis.waterdata.usgs.gov/nwis/discharge?site_no=50071000&agency_cd=USGS&format=rdb&begin_date=04/01/1967&end_date=09/30/2003&period=

mento das águas da superfície e subterrâneas. Propaga os dados ao público, ao estado e aos governos locais, às utilidades públicas e confidenciais e a outras agências federais envolvidas com controle dos recursos de águas.

A região de Porto Rico sofre constantes ações de ciclones, com incidências de tornados. A temperatura média é de 82° F (28° C), com ventos regulares vindos do oeste. Devido o clima tropical pode-se esperar frequentes enchurradas dispersas durante todo o ano. É conhecido, através de evidências empíricas, que as variáveis exibem variação sazonal regular, a saber, período chuvoso e de seca. Banhado, ao norte pelo Oceano Atlântico e ao sul pelo Mar das Caraíbas, Porto Rico tem fronteira marítima, a leste com as Ilhas Virgens Americanas e a oeste com a República Dominicana.

Neste Capítulo, nas seções seguintes, falamos um pouco dos três rios: Canóvanas, Espiritu Santo e Fajardo.

5.1 Rio Canóvanas

O rio Canóvanas fica situado a oeste da Floresta Experimental Luquillo, em Río Grande na bacia de Loíza, nordeste de Porto Rico. A área de drenagem é de 25,5 Km^2. A terra da proximidade do rio é usada tipicamente para pasto, com agrupamentos de pequenas de casas. Precipitações de chuvas fortes acontecem devido à alta intensidade dos ventos vindos do leste e às precipitações tropicais.

5.2 Rio Fajardo

O rio Fajardo nasce ao oeste do Rio Abajo do município de Ceiba-PR a uma elevação de aproximadamente 2,624 pés (800 m) acima do nível do mar. Sua extensão é de aproximadamente 14,9 milhas (24 km) de onde ele nasce até onde ele desemboca (na Sonda de Vieques ao oeste de Porto Rico). Ele passa pelos municípios de Ceiba e Fajardo.

O río Fajardo é situado na ponta do nordeste de Porto Rico com uma bacia de aproximadamente 27 milhas quadradas de área (43,5 km^2). A cidade de Fajardo é situada aproximadamente 37 milhas (43,57 km) ao leste de San Juan. Grande parte da vazão

da planície do rio é usada para a produção de açúcar. O rio flui ao longo da borda do sul da cidade. Inundações sérias ocorreram em 1960, com a passagem do furacão Dora, e também em 1969, em 1970, em 1974, em 1975, em 1978, em 1989 e, mais recentemente, em janeiro 1992. A cidade de Fajardo é o principal centro urbano do oriente de Porto Rico. Cerca de 600 famílias estão sujeitas a inundações. Outras propriedades também estão sujeitas a inundações, incluindo 29 estabelecimentos comerciais, uma indústria farmacéutica, uma escola, a planta de tratamento da perda de água e aproximadamente 400 hectares de terra utilizados em agricultura.

5.3 Rio Espiritu Santo

O rio Espiritu Santo nasce na serra de Luquilo, no município de Rio Grande e tem uma elevação de aproximadamente 2,427 pés (740 m) acima do nível do mar. Tem comprimento de aproximadamente 12 milhas (19,2 km) de onde nasce até onde desemboca, no oceano Atlântico ao norte de Porto Rico dentro dos limites do povoado de Rio Grande. Cruza o município de Rio Grande. O rio Espiritu Santo é rico e belo devido à sua vegetação viva (mangues).

Capítulo 6

Análise sem considerar sazonalidade

Analisamos aqui os dados originais sem considerar sazonalidade. Na análise estudamos estacionaridade, dependência temporal e aplicamos alguns testes. Ajustamos o modelo marginal e o modelo de dependência para os dados completos. Repetimos toda a análise para os dados sem um outlier. Todos os testes aplicados neste capítulo estão descritos no Apêndice.

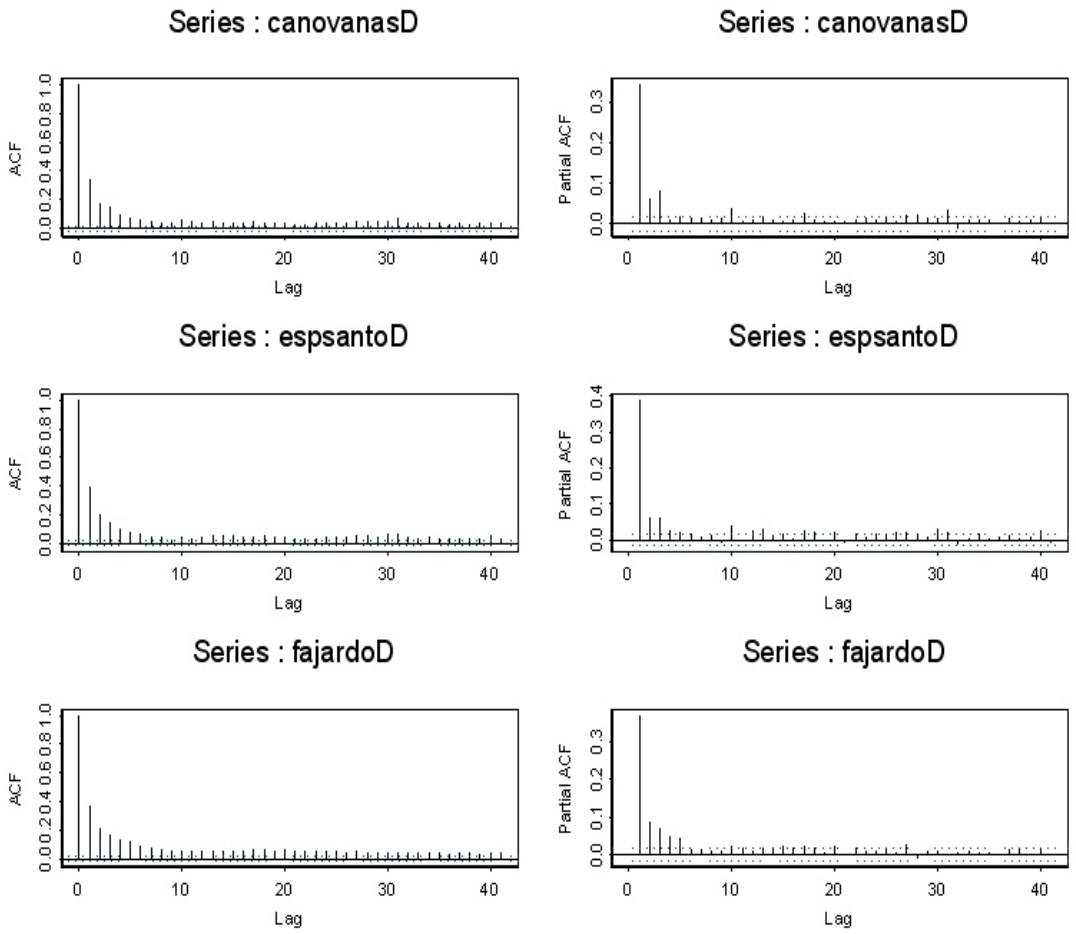
Primeiro fizemos uma análise em relação à estacionaridade dos dados. Analisamos se cada série se desenvolve no tempo aleatoriamente ao redor de uma média constante, refletindo alguma forma de equilíbrio estável. Aplicamos o teste KPSS de Kwiatkowski, Phillips, Schmidt and Shin (1992) para estacionaridade da série com e sem tendência. A hipótese nula é de estacionaridade da série.

Nas três séries aceitamos a hipótese de que são estacionárias sem tendência ao nível de significância de 5%. Os valores da estatística teste KPSS encontram-se listados na Tabela 6.1 com os respectivos p-valores entre parênteses.

Tabela 6.1: Teste KPSS e teste R/S para as médias diárias dos fluxos dos rios.

Rios	KPSS	R/S
Canóvanas	0,1845(0,1)	1,4989
Fajardo	0,2940(0,1)	1,9460
Espiritu Santo	0,2227(0,1)	1,9225

Figura 6.1: acf (esquerda) e pacf (direita) das médias diárias dos fluxos dos rios.



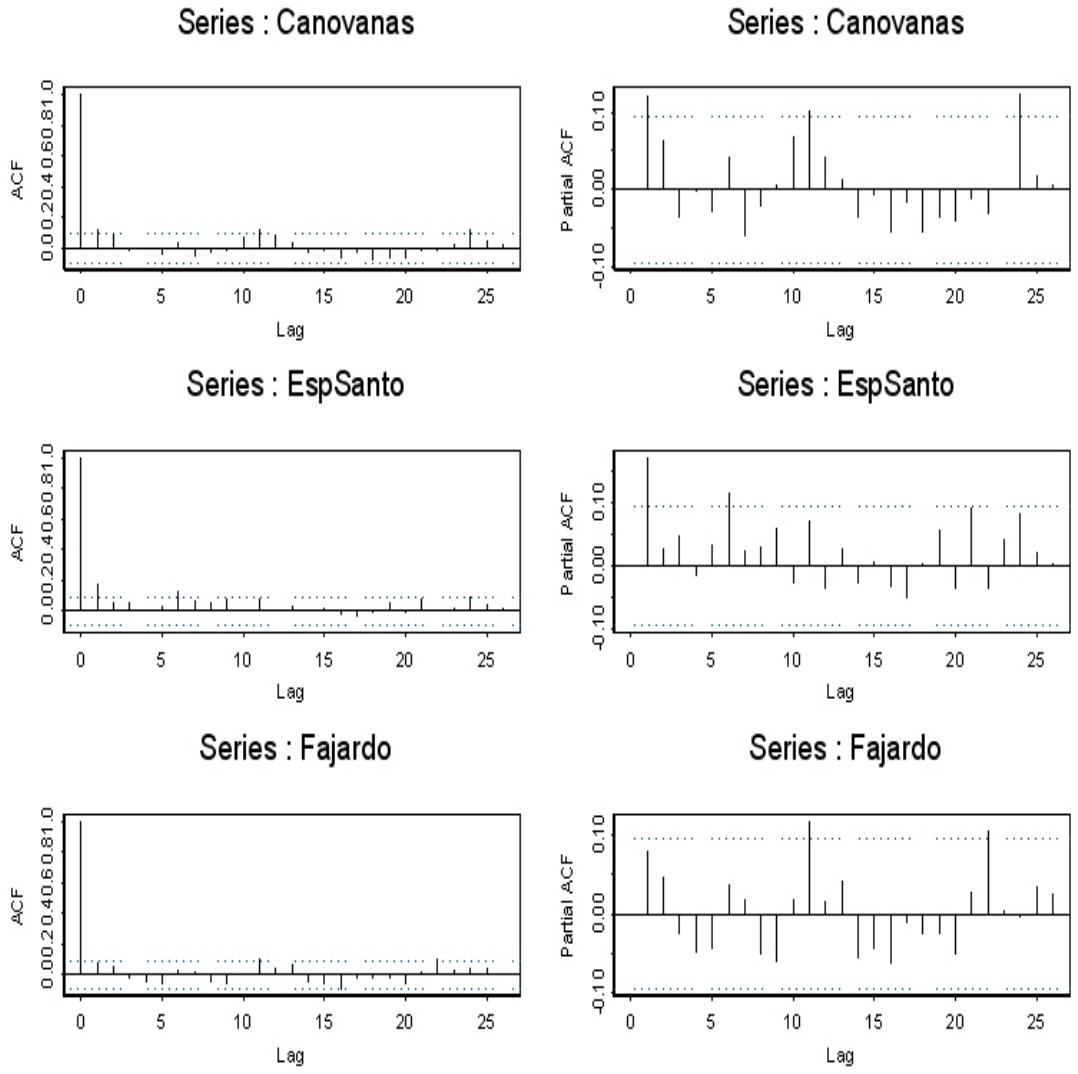
Foi aplicado também o teste R/S de Lo para a hipótese nula de inexistência de memória longa. Os valores da estatística teste R/S encontram-se também listados na Tabela 6.1, sendo os seus valores críticos iguais a $1,747$ para o nível de 5% de significância, e $1,62$ para 10% de significância. A hipótese nula foi rejeitada somente para o rio Canóvanas ao nível de significância de 5%.

Na Figura 6.1 notamos que as funções de autocorrelação de cada uma das séries apresentam valores significativamente diferentes de zero. Isso mostra que há dependência nas séries.

6.1 Máximos mensais

Para reduzir a dependência temporal tomamos os máximos mensais das médias diárias. Prosseguimos nossa análise com base nos máximos mensais dos fluxos dos rios, tomados a partir das médias diárias.

Figura 6.2: acf (esquerda) e pacf (direita) dos máximos mensais.



A Figura 6.2 mostra as funções de autocorrelação (acf) e autocorrelação parcial (pacf) dos máximos mensais de cada série. Comparamos a Figura 6.2 com a Figura 6.1 e notamos

que houve redução na dependência, como esperado.

Aplicamos os mesmos testes que foram aplicados nas séries das médias diárias, de estacionaridade (KPSS) e de memória longa (R/S). Através desses testes, observamos que as séries dos máximos mensais dos fluxos dos rios podem ser consideradas estacionárias e sem memória longa, ao nível de 5%. Os resultados estão na Tabela 6.2.

Tabela 6.2: Teste KPSS sem tendência, KPSS com tendência e teste R/S para os máximos mensais dos fluxos dos rios.

Rios	KPSS	R/S
Canóvanas	0,0720(0,1)	0,8613
Fajardo	0,0957(0,1)	1,1172
Espiritu Santo	0,1644(0,1)	1,2227

A Tabela 6.3 apresenta algumas estatísticas descritivas dos máximos mensais (a média, a moda, a mediana, o máximo, o mínimo e o desvio-padrão) para cada série. Notamos que a série do rio Fajardo apresenta o maior desvio-padrão.

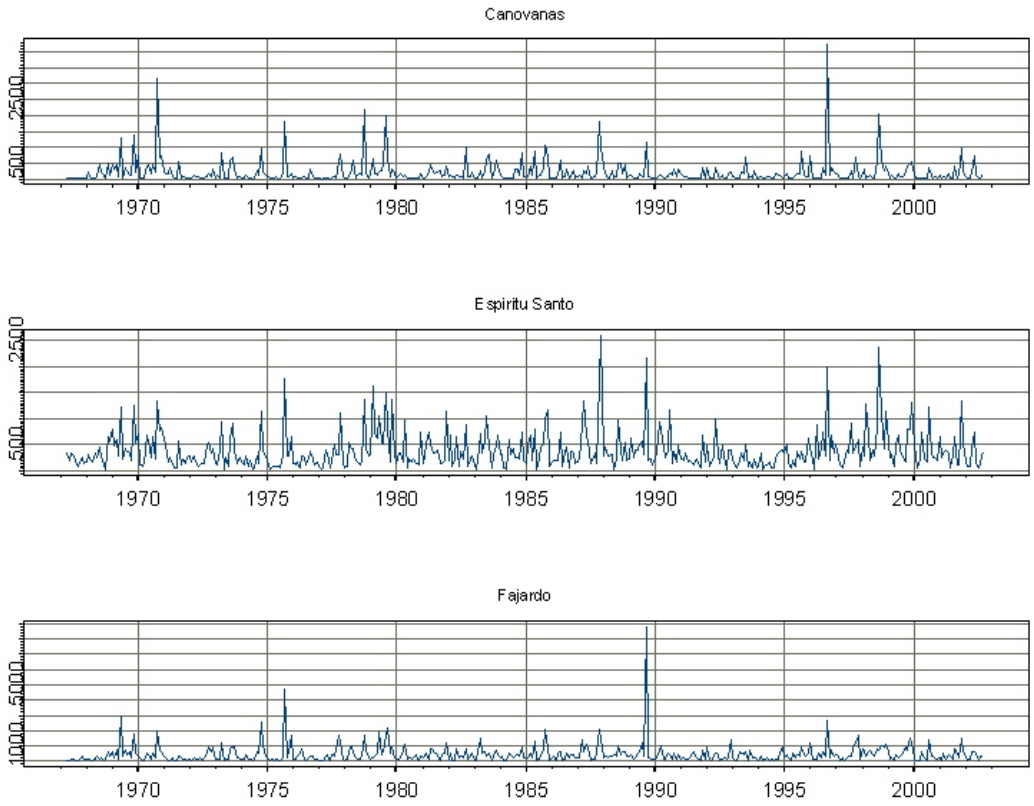
Tabela 6.3: Sumário dos máximos mensais do fluxo dos rios.

Rios	Mínimo	1°Q	Mediana	Média	3°Q	Máximo	dp
Canovanas	3,90	30,25	85,50	207,57	211,75	4230,00	379,47
Fajardo	17,0	138,3	291	455,3	552,3	8800,00	632,60
Esp Santo	17,0	153,5	275,0	385,3	489,5	2600,00	369,02

No gráfico da Figura 6.3 temos os máximos mensais em função do tempo. Aparentemente não há indicação de tendência para nenhuma das três séries. Para verificar a existência de tendência nas séries, aplicamos um teste.

O teste aplicado foi o de Mann-Kendall (Davison and Hinkley (1997)) para observar se existe tendência nas séries. Este teste tem como hipótese nula a inexistência de tendência.

Figura 6.3: Gráfico dos máximos mensais de cada série em função do tempo.



A Tabela 6.4 apresenta os valores da estatística do teste, representados por τ, e seus respectivos p-valores. Rejeitamos a hipótese nula apenas para a série do rio Espiritu Santo ao nível de 5% de significância. Este teste também está descrito no Apêndice.

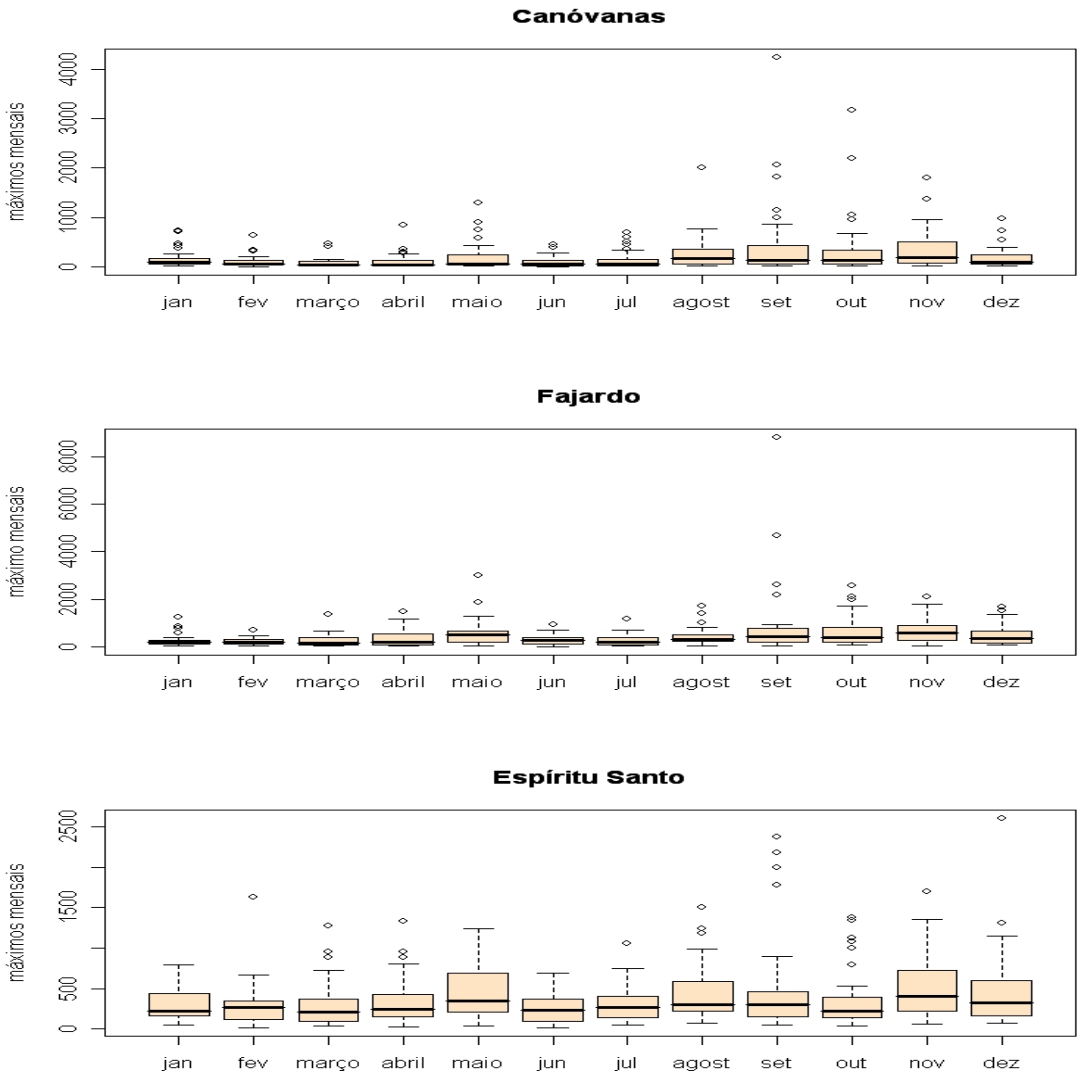
Tabela 6.4: Teste de Mann-Kendall de tendência.

Rios	τ	p-valor
Canovanas	0,0293	0,44846
Fajardo	0,0059	0,85431
Esp Santo	0,0773	0,015689

Olhando para Figura 6.4, notamos que a maioria dos valores discrepantes, para as três séries, estão no mês de setembro. Existe uma observação (8800) na série do rio Fajardo

que é discrepante. Esta observação é referente ao máximo do mês de setembro de 1989. Através de registros na internet (www.elexpresso.com/site/ver_notíciaasp?n_id=2859) vimos que Porto Rico e outras antilhas da América Central foram surpreendidos por furacões e ciclones no período de 10 a 22 de setembro de 1989. Esses fatos dão sentido ao valor discrepante; mesmo assim não podemos descartar a possibilidade de um erro de digitação.

Figura 6.4: Boxplot dos máximos mensais, separados em meses.

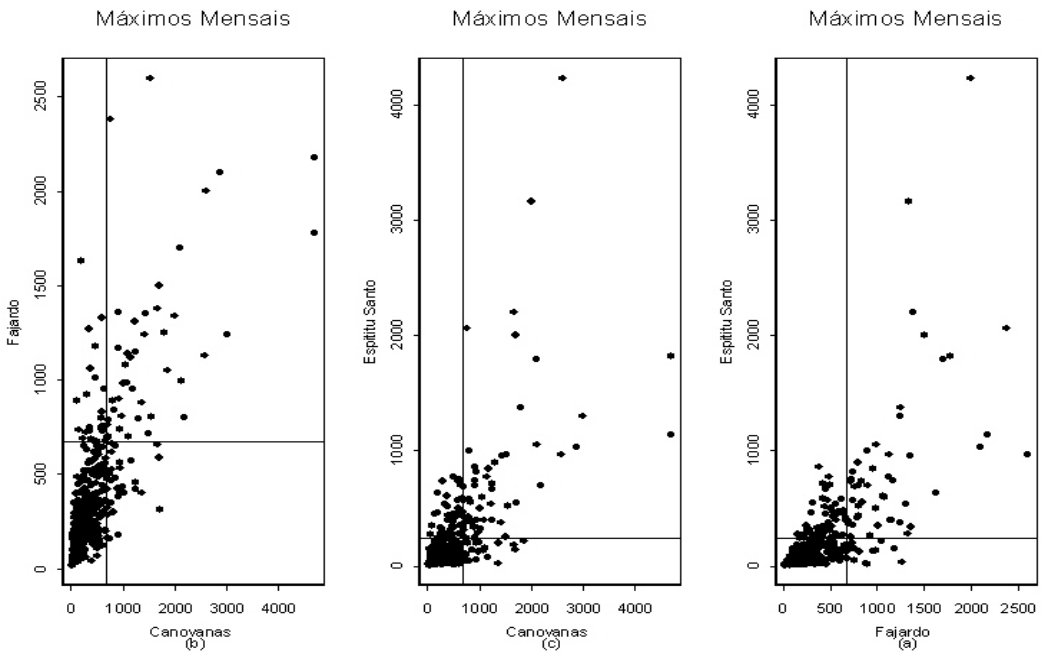


A análise agora prossegue de duas maneiras: análise dos máximos mensais completos na Seção 6.1.1 e sem o outlier do rio Fajardo na Seção 6.2.

6.1.1 Dados completos

Nesta seção ajustamos o modelo marginal e o modelo de dependência para as variáveis com os dados completos. A Figura 6.5 mostra o gráfico dos máximos mensais dos fluxos de cada rio. Observamos, através da figura, que as variáveis parecem ser assintoticamente independentes. As linhas verticais e horizontais da Figura 6.5 representam os limiares escolhidos. Essa escolha é detalhada na Seção 6.1.2.

Figura 6.5: Gráfico dos máximos mensais.



6.1.2 Ajuste do modelo marginal

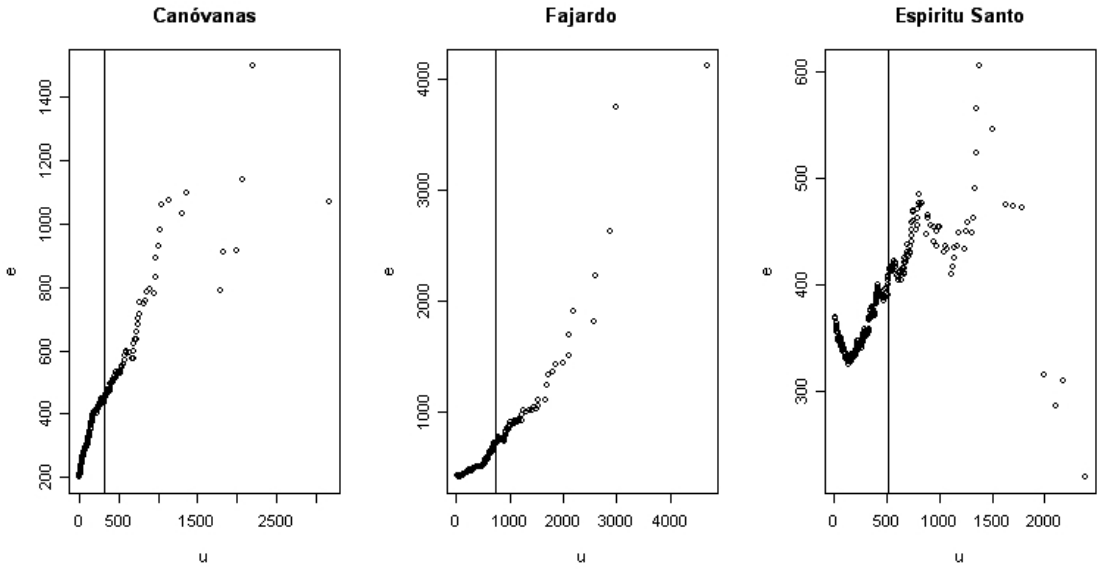
Focamos agora no modelo marginal (1.3). Esse modelo consiste primeiramente na transformação da distribuição dos dados em uma distribuição com marginais Gumbel.

Estimamos os parâmetros da GPD para cada componente.

Para fazer o ajuste da GPD foi necessário primeiramente escolher um limiar para cada série. É claro que um valor de u muito "alto"' implica um número pequeno de observações na cauda, podendo resultar em maior variabilidade dos estimadores. Por outro lado, um limiar que não seja suficientemente alto não satisfaz às suposições teóricas. Temos aqui um *trade-off* entre vício e variância. Usamos a função média dos excessos empírica. A técnica gráfica exploratória baseada nessa função nos auxilou na tarefa da escolha de um limiar u alto suficiente para que a aproximação da distribuição dos excessos por uma GPD fosse justificada.

Na Figura 6.6 temos três gráficos da média dos excessos versus o limiar. Em cada gráfico, procuramos por aqueles valores de u tais que a função média dos excessos empíricas se torne mais ou menos linear.

Figura 6.6: Gráfico da função Média dos Excessos empírica.



Olhando para Figura 6.6 e para o histograma dos excessos (não mostrado aqui), escolhemos como limiar o quantil empírico 82% para a série do rio Canóvanas, o quantil empírico 77% para o rio Espiritu Santo e o quantil empírico 85% para o rio Fajardo.

Ajustamos o modelo marginal (1.3) para as três componentes. A Tabela 6.5 apresenta: os valores do limiar escolhido u_{X_i}; o quantil estimado do limiar dado por $\hat{F}_{X_i}(u_{X_i})$; os parâmetros escala ($\hat{\beta}_i$) e forma ($\hat{\xi}_i$) estimados pela GPD com os respectivos erros-padrão entre parênteses; e o quantil $0,99$ da marginal estimada $\hat{x}(0,99) = \hat{F}_{X_i}^{-1}(0,99)$. A estimação dos parâmetros forma e escala foi feita por máxima verossimilhança.

Tabela 6.5: Parâmetros estimados da GPD para as distribuições marginais dos máximos mensais.

Parâmetros	$Canovanas$	$Fajardo$	$EspSanto$
u_{X_i}	319,68	725,8	518,98
$\widehat{F}_{Xi}(u_{X_i})$	0,82	0,85	0,77
$\widehat{\beta}_i$	288,62(53,9)	495,72(94,91)	344,3(56,21)
$\widehat{\xi}_i$	0,36(0,15)	0,31(0,15)	0,12(0,13)
$\widehat{x}_i(0,99)$	1778,72	2822,76	1829,93

Para ver o ajuste do modelo na cauda, aplicamos o teste GOF de Kolmogorov (estatística Dn de Kolmogorov). A hipótese nula do teste é de que o ajuste é bom. Rejeitamos H_0 se o valor da estatística teste D_n for maior que um valor crítico.

A Tabela 6.6 mostra as estatísticas D_n de Kolmogorov para as três séries e os valores críticos. De acordo com os valores obtidos, aceitamos a hipótese nula para as três séries ao nível de 5% de significância.

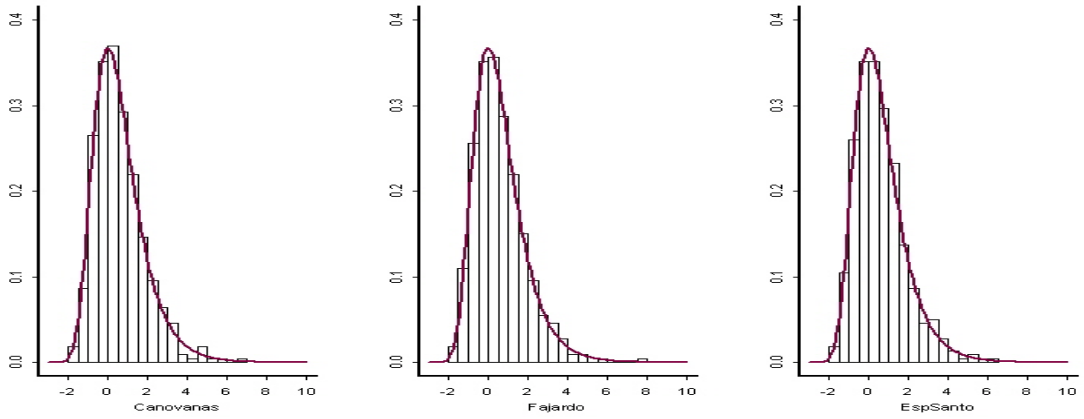
Tabela 6.6: Teste GOF para o ajuste GPD do fluxo dos máximos mensais.

Rios	D_n	Valor Crítico
Canovanas	0,03560520	0,1505448
Fajardo	0,05988047	0,1644550
Esp Santo	0,0554383	0,1333881

Fizemos também histogramas dos dados transformados em uma distribuição Gumbel com a densidade Gumbel superposta, como mostra a Figura 6.7. Notamos que o ajuste

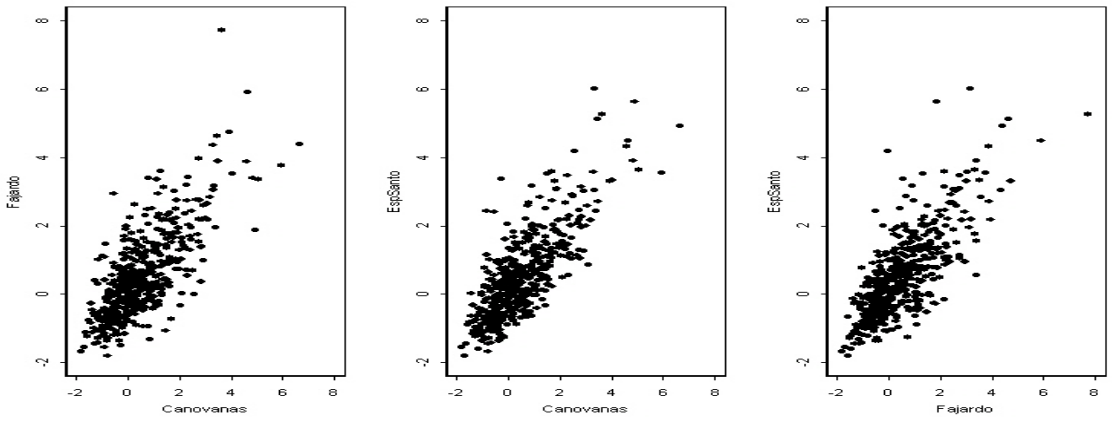
é adequado.

Figura 6.7: Histograma do ajuste dos máximos mensais.



A Figura 6.8 mostra os dados transformados, através da relação vista em (1.4). Essa figura parece indicar que as variáveis são assintoticamente independentes em todas as combinações.

Figura 6.8: Máximos mensais transformados para a distribuição Gumbel.



6.1.3 Ajuste do modelo de dependência

Agora consideramos o modelo de dependência (4.1) aplicado aos dados depois da transformação (1.4). Estimamos os parâmetros do modelo de dependência condicional separadamente, para cada i. Lembrando, $\hat{\mathbf{\Theta}}$ é o vetor de todos os parâmetros a serem estimados do modelo de dependência geral. O $\hat{\mathbf{\Theta}}$ do nosso modelo geral é um vetor de tamanho 12, embora $c_{j|i}$ e $d_{j|i}$ só sejam estimados se $\hat{a}_{j|i} = 0$ e $\hat{b}_{j|i} < 0$.

Tabela 6.7: Parâmetros estimados para o modelo de dependência.

Parâmetros	FAJ| CAN	ES| CAN	CAN| FAJ	ES| FAJ	CAN| ES	FAJ| ES
$\hat{a}_{j\|i}$	0,783	0,962	0,952	0,920	0,557	0,552
$\hat{b}_{j\|i}$	0,264	0,384	0,477	0,187	0,425	0,497

Novamente necessitamos escolher um limiar para o nosso modelo. Por simplicidade, limitamos nossa busca por um limiar da estrutura de dependência $u_{Y_i} = u$ para todo i. O limiar da estrutura de dependência escolhido é tal que $Pr(Y_i \leq u) = 0,7$.

A Tabela 6.7 mostra os valores estimados de $(\hat{a}_{j|i}, \hat{b}_{j|i})$, $\forall$i$\neq j$. Chamamos rio Fajardo de FAJ, rio Canóvanas de CAN e rio Espiritu Santo de ES. Para nenhum dos modelos de dependência condicional foi necessário estimar os valores de $c_{j|i}$ e $d_{j|i}$, pois não obtivemos nenhum par em que $\hat{a}_{j|i} = 0$ e $\hat{b}_{j|i} < 0$. Pela Tabela 6.7, todos os modelos são assintoticamente independentes com dependência extremal positiva.

Os valores $(\hat{\mathbf{a}}_{|\mathbf{i}}, \hat{\mathbf{b}}_{|\mathbf{i}})$ e as distribuições amostrais dos pares $(\hat{a}_{j|i}, \hat{b}_{j|i})$ para todo $i \neq j$, são mostrados na Figura 6.9. Em particular, os pares das distribuições amostrais são apresentados por *convex hulls* obtidos através de 100 realizações bootstrap da distribuição amostral de $\hat{\Theta}$. Gráficos desse tipo poderiam ser usados para investigar a estabilidade de $\hat{\Theta}$ na escolha de um limiar. Mudanças significantes nas regiões dos *convex hulls* indicam sensibilidade dos parâmetros estimados para a escolha de um limiar. Um limiar apropriado seria aquele que tem a propriedade de que seu aumento não resulta em mudança significativa no *convex hull*, uma vez que a variabilidade aumenta quando o cálculo das estimativas é feito pelo uso desses limiares maiores. O limiar selecionado é o de menor

valor que satisfaz essa propriedade.

Tendo decidido o limiar do modelo de dependência, consideramos simplificações possíveis para a estrutura de dependência estimada. Através da Figura 6.9 e dos gráficos dos dados com marginais Gumbel (Figura 6.8), parece existir diferenças nos níveis de dependência extremal entre os pares das variáveis. Nos *convex hulls*, se a área dos pares $(\hat{a}_{j|i}, \hat{b}_{j|i})$ e $(\hat{a}_{i|j}, \hat{b}_{i|j})$ não se cruzam, então dizemos que há diferenças significantes entre os pares das variáveis. Aqui, não tivemos nenhum caso em que os *convex hulls* não se cruzam. A Figura 6.9 mostra que (Can, Faj) e (Can, ES) exibem permutabilidade no par. Isso indica que uma mudança no par estimado pelo modelo de dependência é apropriada para esses dados. Embora para os pares dos fluxos dos rios não exista nenhuma evidência para rejeitar a fraca permutabilidade, na ausência de conhecimentos mais detalhados sobre o processo não tentamos identificar os subconjuntos de pares para os quais podemos assumir um modelo de dependência simplificado.

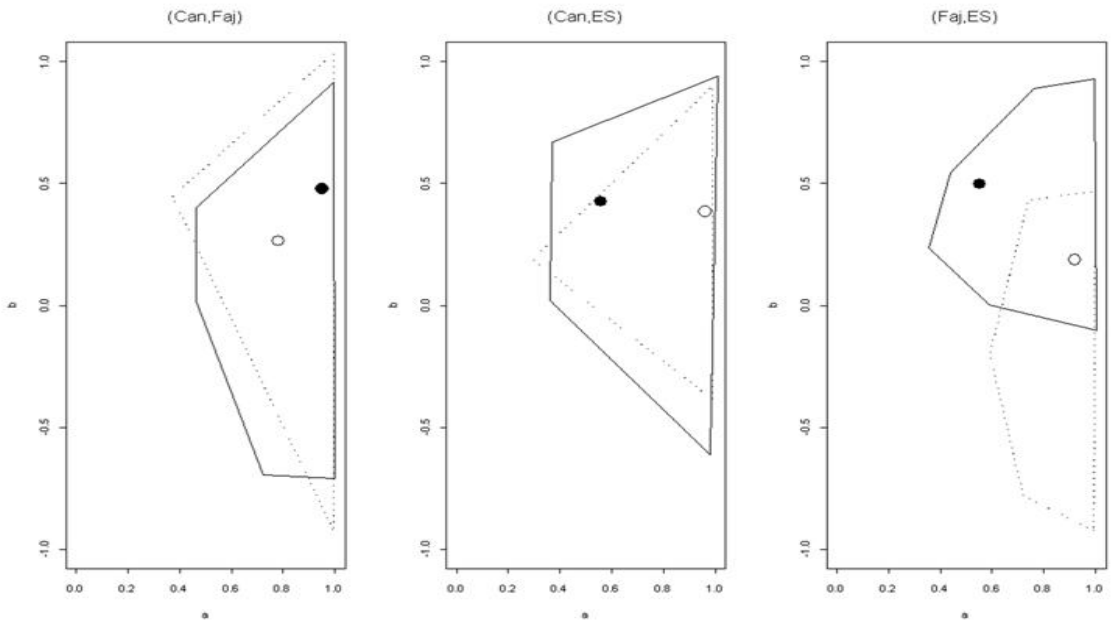
Finalmente estudamos se os $\mathbf{Z}_{|i}$ são independentes para qualquer i, isto é, se podemos assumir independência condicional assintótica entre as distribuições marginais $G_{|i}$ dos resíduos para qualquer i. Calculamos os resíduos e fizemos o gráfico dos pares das componentes de $\mathbf{Z}_{|i}$. A Figura 6.10 mostra que é razoável admitirmos que as distribuições marginais dos resíduos são condicionalmente assintoticamente dependentes.

Aplicamos o teste τ de Kendall (Best e Roberts (1975)) que tem como hipótese nula que as variáveis não são correlacionadas (ρ igual a zero). Aplicamos o teste e para cada par $\mathbf{Z}_{|i}$ e rejeitamos a hipótese nula, ou seja, o teste revelou que as variávies podem ser consideradas dependentes. Os resultados estão dispostos na Tabela 6.8, que contém a estatística teste, ρ estimado e o p-valor do teste.

Tabela 6.8: Teste τ de Kendall para os pares $\mathbf{Z}_{|i}$.

Pares	estatística	$\hat{\rho}$	p-valor
$\mathbf{Z}_{\|Can}$	5,11	0,41	$1,130e^{-06}$
$\mathbf{Z}_{\|Faj}$	8,19	0,58	$2,123e^{-13}$
$\mathbf{Z}_{\|ES}$	4,72	0,38	$5,973e^{-06}$

Figura 6.9: Comparação dos parâmetros de dependência estimados usando como limiar o quantil do modelo de dependência 70%. A linha contínua representa o *convex hull* de $(\hat{a}_{j|i}, \hat{b}_{j|i})$ e a linha pontilhada $(\hat{a}_{i|j}, \hat{b}_{i|j})$.

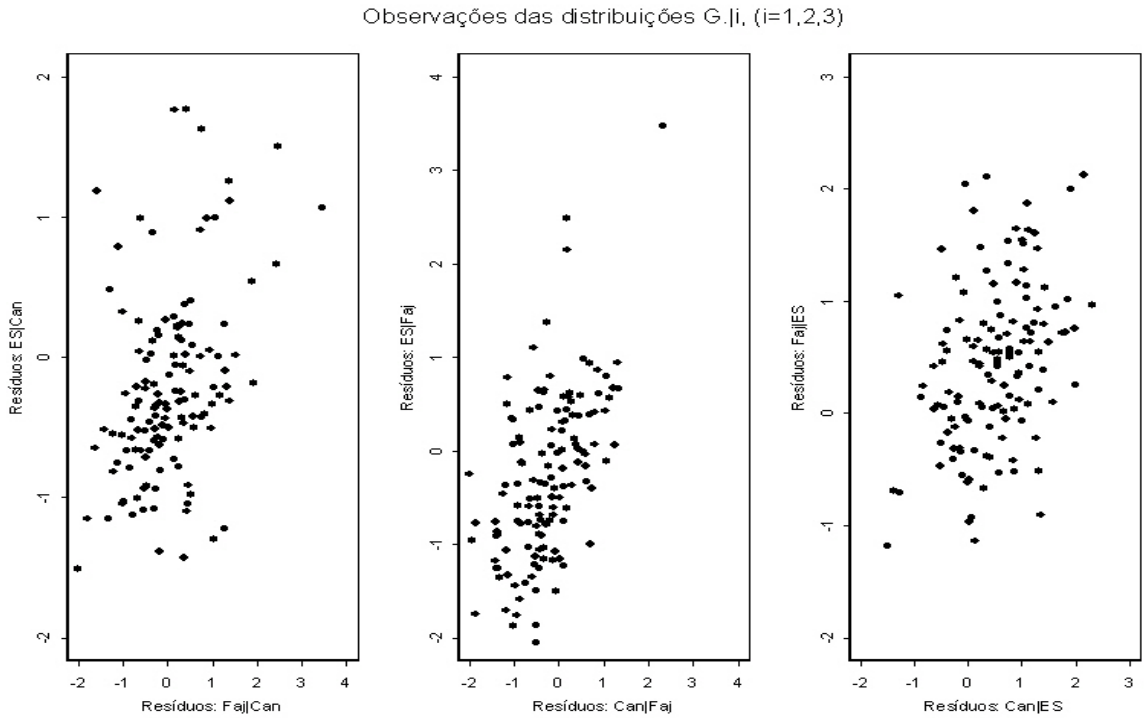


A Figura 6.9 também mostra diferenças entre os parâmetros estimados que são obtidos para os conjuntos dos dados. Todos os pares, aparentemente, exibem forte dependência, com razoáveis evidências de que essas variáveis são assintoticamente dependentes.

Para ilustrar as implicações dos diferentes níveis de dependência entre os pares e as diferentes distribuições marginais, na Figura 6.11 apresentamos as pseudo-amostras das distribuições condicionais das variáveis dado que X_i excede a um determinado limiar. No gráfico de cada par, a curva corresponde à igualdade dos quantis marginais. Observamos que uma quase dependência assintótica dos pares (Can, ES) e (Faj, ES) é vista através da disposição dos pontos simulados em torno dessa curva. Os pontos dos outros pares são mais espalhados apresentando, aparentemente, independência assintótica com dependência extremal positiva.

Calculamos a esperança condicional para cada componente dado X_i excedendo a um

Figura 6.10: Gráficos dos Resíduos.



limiar. Estas estimações refletem as características do modelo marginal e do modelo de dependência dos fluxos dos rios. A Figura 6.11 mostra as pseudo amostras para a distribuição condicional de cada váriável dado que X_i excede o quantil 0,99. A Tabela 6.9 mostra as esperanças condicionais estimadas para cada variável condicionada a um X_i excedendo a um limiar. Quando condicionamos X_i excedendo a um quantil de 0,95, as esperanças empíricas dessas funções são confiáveis se comparadas com as estimações das esperanças baseada no modelo.

Agora concentramos na estimação dos nívies de retorno das combinações lineares das variáveis na escala marginal Gumbel. A escolha dessa função é feita para enfatizar o efeito de dependência nas combinações extremas. Focamos no conjunto multidimensional $C^3(v) = \{\mathbf{y} \in R^m : \sum_{i \in M} y_i > v\}$ e reportamos o nível de retorno estimado v_p definido por $Pr\{\mathbf{Y} \in C(v_p)\} = p$. A escolha desse conjunto permite uma extrapolação da dependência extremal em partes do espaço no qual nem todas as variáveis são simultaneamente extremas. Na Figura 6.11 destacamos os pontos simulados pertencentes a

Tabela 6.9: Esperanças empírica e baseada no modelo.

X_j	E[X_j]l	E[$X_j \mid X_i > x_i(0.95)$]		E[$X_j \mid X_i > x_i(0.99)$]
	Empírica	Empírica	Baseada no modelo	Baseada no modelo
A variável X_i é Canóvanas				
Canovanas	207,57(18,6)	1454,32(181,3)	1022,85(121,4)	1651,97(175,5)
Fajardo	455,27(31,8)	2111,23(384,5)	1709,73(209,4)	2178,59(289,9)
Espiritu Santo	385,34(18,7)	1376,78(128,7)	1171,58(129,7)	1497,26(158,1)
A variável X_i é Fajardo				
Canovanas	207,57(18,6)	1222,96(218,2)	1141,52(205,4)	1348,59(244,4)
Fajardo	455,27(31,8)	2385,91(343,7)	1723,75(261,0)	2112,76(297,0)
Espiritu Santo	385,34(18,7)	1305,18(123,2)	1170,24(144,8)	1465,67(287,1)
A variável X_i é Espiritu Santo				
Canovanas	207,57(18,6)	1250,36(220,5)	977,462(204,5)	1454,54(293,5)
Fajardo	455,27(31,8)	1921,91(406,1)	1636,18(371,6)	2363,88(390,7)
Espiritu Santo	385,34(18,7)	1565,46(91,4)	1151,88(95,4)	1781,35(114,8)

$C^3(15)$ e os maiores valores $\sum_{i=1}^{3} y_i$ e círculos grandes.

A Figura 6.12 mostra os valores dos níveis de retorno $C^3(v_p)$ obtidos empiricamente e baseados no modelo. Os níveis de retorno baseado no modelo não diferem muito dos empíricos; além disso há uma boa quantidade de níveis de retorno empírico dentro do intervalo de confiança de 95%. Os gráficos (a), (b) e (c) da Figura 6.12 mostram que as estimativas empíricas e baseadas no modelo para os níveis de retorno são próximas indicando que o modelo de dependência é razoável.

6.2 Dados sem outlier

Analisamos os dados sem o outlier (8800) da série do rio Fajardo. Essa observação é o máximo do mês de setembro de 1989. Substituímos o outlier pelo segundo maior valor observado do mês de setembro de 1989 (1520).

Os testes de estacionaridade KPSS e memória longa R/S foram aplicados e os resultados foram similares aos da série com outlier, ou seja, aceitamos que a série é estacionária e sem memória longa ao nível de 5% de significância.

O sumário das estatísticas descritivas para a série do rio Fajardo sem o outlier indicou que o desvio-padrão diminuiu de 632,60 para 493,11 e que a série se tornou um pouco mais parecida com as demais, com caudas menos pesadas e menor variação.

Fizemos o ajuste do modelo (1.3) para a série do rio Fajardo sem considerar o outlier. Primeiro fizemos uma análise de sensibilidade, e escolhemos como limiar o quantil empírico 85%. Depois, estimamos através da GPD os parâmetros forma e escala do modelo marginal. O valor estimado do parâmetro forma (0,10) diminuiu e o erro permaneceu praticamente o mesmo. Em relação ao parâmetro escala, o valor estimado (551,08) e o erro (102,25) aumentaram.

A Figura 6.13 mostra que a cauda da distribuição ajustada pela GPD para a série do rio Fajardo sem o outlier não sofre grandes mudanças em relação à cauda da distribuição dos dados originais. Observamos que as caudas das duas distribuições com e sem o outlier, a partir de um certo ponto, ficam praticamente iguais.

Na estimação dos parâmetros do modelo de dependência, os valores estimados não se alteraram muito com a remoção do outlier, como mostra a Tabela 6.10. A estrutura do modelo permanece a mesma, ou seja, continuam sendo assintoticamente independentes com dependência extremal positiva em todos os pares. Todos os parâmetros estimados de $\hat{a}_{j|i}$ e $\hat{b}_{j|i}$ foram positivos e não há a necessidade de estimar os valores de $c_{j|i}$ e $d_{j|i}$.

Os resultados em relação à estrutura geral do modelo, tanto no modelo de marginal quanto no modelo de dependência, não se alteram muito com a remoção do outlier.

Tabela 6.10: Parâmetros estimados para o modelo de dependência sem outlier.

Parâmetros	FAJ| CAN	ES| CAN	CAN| FAJ	ES| FAJ	CAN| ES	FAJ| ES
$\hat{a}_{j\|i}$	0,80	0,96	0,94	0,92	0,56	0,50
$\hat{b}_{j\|i}$	0,19	0,38	0,43	0,27	0,42	0,52

Figura 6.11: Dados simulados do modelo condicional estimado e para os dados na escala original. Nas linhas $i = 1, 2, 3$, apresentamos os pares das distribuições condicionais simuladas, dado que a variável i excede o quantil marginal empírico $x_i(0,99)$. As linhas verticais correspondem ao limiar $x_i(0,99)$. Os pontos abaixo e acima dos limiares são os dados originais e os dados simulados, respectivamente. Os círculos menores representam os pontos que pertencem ao conjunto $C^3(v_p)$. Os círculos grandes representam os seis pontos com maiores valores de $\sum_{i=1}^{3} y_i$. As curvas representam a igualdade dos quantis marginais.

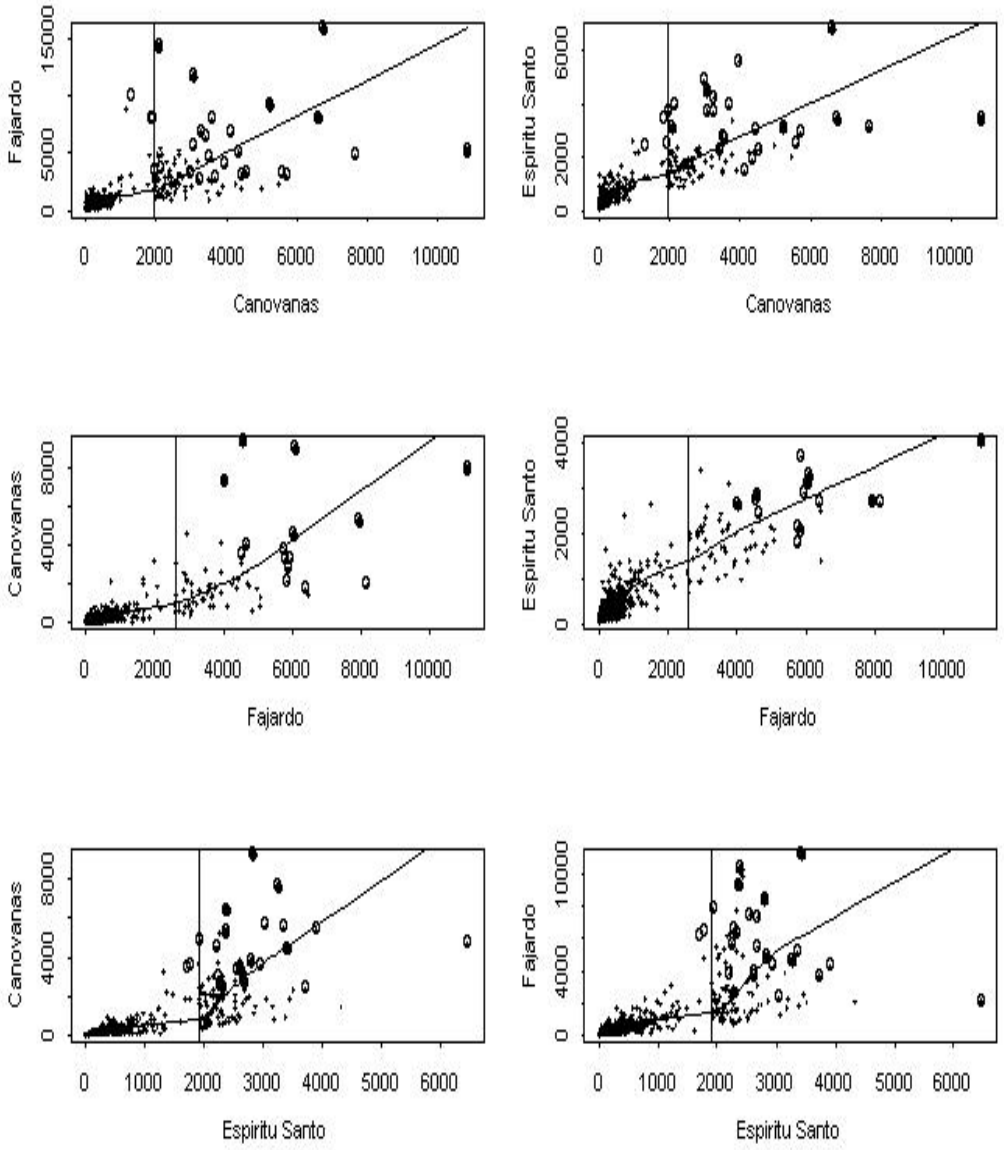


Figura 6.12: Níveis de retorno estimados para o conjunto $C^3(v_p)$ usando os modelos condicionais estimados. Em (a) dado que o rio Canóvanas é extremo; em (b) dado que o rio Fajardo é extremo; e em (c) dado que o rio Espiritu Santo é extremo. O nível de retorno é $\sum_{i=1}^{3} y_i$, representado pela linha contínua. As linhas pontilhadas correspondem ao intervalo de confiança de 95%. Os pontos "o"representam os níveis de retorno empíricos.

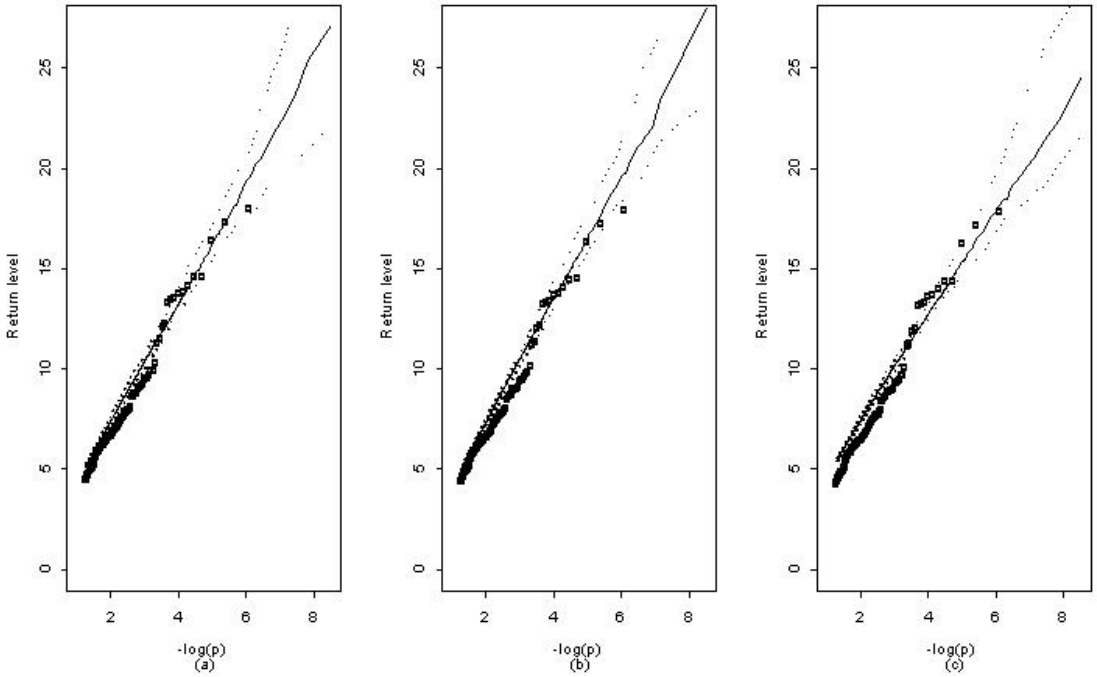
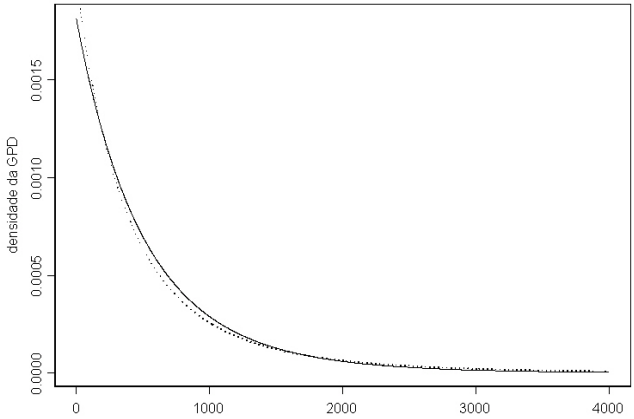


Figura 6.13: Densidade GPD com os parâmetros estimados da série original (linha contínua) superposta por uma GPD com os novos parâmetros estimados (linha pontilhada).



Capítulo 7

Análise considerando sazonalidade

Separamos os dados em dois períodos: chuvoso e de seca. Segundo Héctor, Glenn, and Frederic (2002), o período de chuvas em Porto Rico corresponde aos meses de maio, agosto, setembro, outubro, novembro e dezembro. O período de seca corresponde aos meses de janeiro a abril e de junho e julho. Essa separação dos períodos é coerente com a Figura 7.4 (boxplots dos máximos mensais separados por meses).

O principal objetivo dessa análise é estudar o que há por trás da estrutura de dependência das variáveis. Através dessa estrutura podemos identificar se as relações entre os valores extremos dessas variáveis estão de acordo com o conhecimento científico e se a interação entre os fluxos dos rios e os períodos de chuva e seca éestão bem definida.

Neste Capítulo foi feita uma rápida análise descritiva dos dados em cada período. Aplicamos alguns testes de estacionaridade, tendência e sazonalidade. Depois ajustamos o modelo marginal e o modelo de dependência.

7.1 Análise descritiva

Na Tabela 7.1 temos os valores das estatísticas teste e seus p-valores entre parênteses. Segundo os testes, aceitamos as hipóteses nulas dos dois testes para as três séries no período de chuva ao nível de significância de 5%. No período de seca rejeitamos as hipóteses nulas dos dois testes para as séries do rio Canóvanas e do rio Fajardo. Aqui, τ é a estatística teste de Kendall.

Tabela 7.1: Testes de Mann-Kendall de tendência e sazonalidade.

	Chuva		Seca	
Rios	τ-Tendência	τ-Sazonal	τ-Tendência	τ-Sazonal
Canóvanas	0,0005(0,99)	0,0011(0,98)	0,112(0,01)	0,116(0,019)
Fajardo	0,0481(0,29)	0,0318(0,52)	0,112(0,01)	0,116(0,019)
Espiritu Santo	0,0061(0,89)	0,0047(0,92)	0,04(0,38)	0,0482(0,33)

Aplicamos os mesmos testes que foram aplicados para os dados completos para estacionaridade com tendência e sem tendência e também o teste para inexistência de memória longa.

Através dos valores da Tabela 7.2 aceitamos a hipótese de que as séries, no período de chuva, são estacionárias, ao nível de 5% de significância. No período de seca, rejeitamos a hipótese de estacionaridade sem tendência para as séries do rio Canóvanas e Fajardo. No teste de inexistência de memória longa, aceitamos a hípótese de que não há memória longa para os três rios nos dois períodos. Nos demais casos aceitamos a hipótese de inexistência de memória longa. O valor crítico desse teste é 1,747 para o nível de 5% de significância.

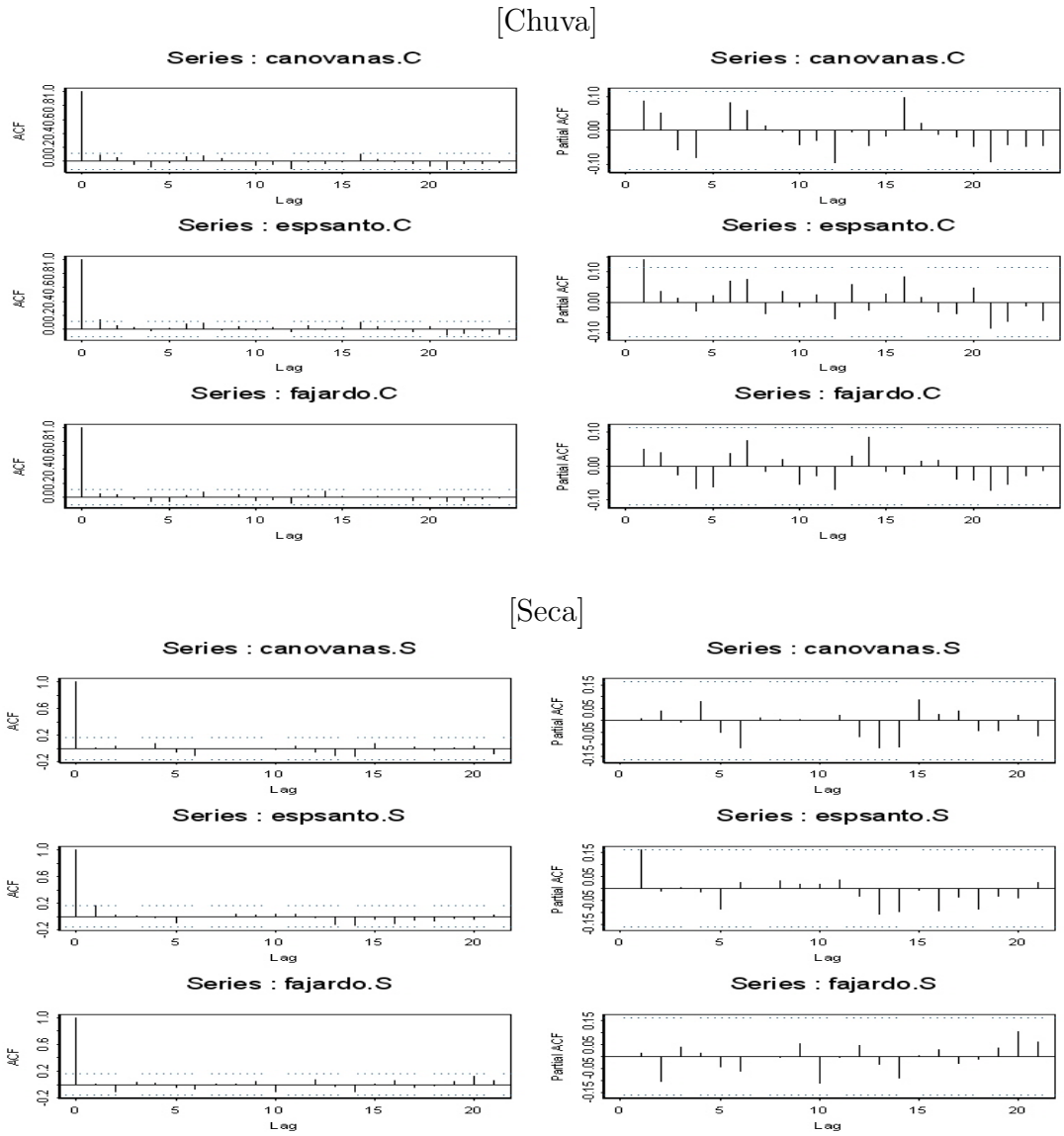
Tabela 7.2: Teste KPSS sem tendência (KPSS), KPSS com tendência (KPSS-T) e teste R/S de memória longa para os máximos mensais dos fluxos dos rios em cada período.

	Chuva			Seca		
Rios	KPSS	KPSS-T	R/S	KPSS	KPSS-T	R/S
Canóvanas	0,0575(0,1)	0,0517(0,1)	0,9186	0,5374(0,03)	0,0757(0,1)	0,8303
Fajardo	0,0527(0,1)	0,0352(0,1)	0,9285	0,5374(0,03)	0,0757(0,1)	1,1344
Espiritu Santo	0,0754(0,1)	0,0524(0,1)	1,1727	0,1724(0,1)	0,0556(0,1)	1,2180

Na Figura 7.1 mostramos a função de autocorrelação e autocorrelação parcial para

cada série em cada período. Notamos que não há, aparentemente, dependência entre as variáveis. Obtivemos ruído branco para todas as séries dos dois períodos.

Figura 7.1: Função de autocorrelação e autocorrelação parcial.



Na Tabela 7.3 temos as estatísticas descritivas dos fluxos dos rios nos dois períodos. O rio Fajardo apresenta o maior desvio-padrão nos períodos de seca e de chuva. Os desvios-padrão dos rios no período de chuva são maiores que no período de seca.

Tabela 7.3: Sumário dos máximos mensais nos dois períodos.

	Rios	Mínimo	1°Q	Mediana	Média	3°Q	Máximo	dp
Chuva	Canóvanas	7,4	48,0	127,0	300,8	335,5	4230,0	496,58
	Fajardo	32,0	206,0	419,0	614,6	726,0	8800,0	808,64
	Esp Santo	37,0	181,5	322,0	461,5	597,0	2600,0	430,23
Seca	Canóvanas	3,90	22,5	56,0	114,4	142,5	1030,0	156,88
	Fajardo	17,0	102,0	197,0	295,9	391,0	2870,0	311,94
	Esp Santo	17,0	125,0	237,0	309,2	406,0	2100,0	276,12

7.2 Ajuste do modelo marginal

O conjunto dos dados contém alguns valores grandes em torno do mês de setembro de cada ano. Em especial, como já citado, existe um valor bem discrepante em relação aos demais valores observados (8800) da série do rio Fajardo no período chuvoso. Nossa análise foi feita com e sem essa observação que contém esse outlier. Fizemos o ajuste dos modelos e os estágios de inferência das análises excluindo e incluindo esse ponto discrepante, para avaliar a sensibilidade dos resultados em relação à sua presença. A estrutura de dependência estimada não foi afetada com a remoção do outlier. Reportamos somente a análise com os dados sem retirar o outlier.

Para ajustar o modelo marginal, primeiro escolhemos o limiar através de uma análise de sensibilidade, como a feita para os dados sem considerar sazonalidade. Ajustamos o modelo marginal (1.3) para cada componente e em cada período. A Tabela 7.4 apresenta: os valores do limiar escolhido u_{X_i}; o quantil estimado do limiar dado por $\hat{F}_{X_i}(u_{X_i})$; os parâmetros estimados de escala ($\hat{\beta}_i$) e forma ($\hat{\xi}_i$) da GPD com os respectivos erros-padrão entre parênteses; e o quantil de $0,99$ da marginal estimada $\hat{x}(0,99) = \hat{F}^{-1}_{X_i}(0,99)$ para cada componente em cada período.

Os valores de $\hat{x}(0,99)$ deferem grandemente da mesma componente sobre os períodos. O parâmetro forma do rio Canóvanas no período de chuva resultou em um valor bem diferente do período de seca. O parâmetro escala estimado tem maior variação na série

do rio Fajardo.

Tabela 7.4: Parâmetros do ajuste marginal dos fluxos dos rios nos dois períodos.

Período	Parâmetros	*Canóvanas*	*Fajardo*	*EspSanto*
Chuva	u_{X_i}	479,4	968,6	650,74
	$\widehat{F}_{Xi}(u_{X_i})$	0,82	0,85	0,77
	$\widehat{\beta}_i$	373,53(102,83)	677,18(175,47)	430,86(95,78)
	$\widehat{\xi}_i$	0,38(0,23)	0,28(0,2)	0,02(0,17)
	$\widehat{x}_i(0,99)$	2435	3700,63	2054,86
Seca	u_{X_i}	165	547,2	418,58
	$\widehat{F}_{Xi}(u_{X_i})$	0,82	0,85	0,77
	$\widehat{\beta}_i$	228,62(58,36)	191,08(54,22)	359,08(48,47)
	$\widehat{\xi}_i$	-0,07(0,2)	0,37(0,2)	0,13(0,15)
	$\widehat{x}_i(0,99)$	766,02	1435,9	1805,02

A Figura 7.3 nos mostra o histograma dos dados transformados em uma distribuição Gumbel com a densidade da Gumbel superposta. Vemos que o ajuste é razoável para as três componentes em cada período. A Figura 7.4 mostra o gráfico dos dados transformados para distribuição Gumbel como vista em (1.4), parecendo indicar que as variáveis são assintoticamente independentes em todas as componentes nos dois períodos.

7.3 Ajuste do modelo de dependência

Agora consideramos o modelo de dependência (4.1) aplicado para os dados depois da transformação (1.4). Modelamos os máximos mensais dos fluxos dos rios de forma univariada em cada período. $\hat{\Theta}$ é um vetor de tamanho 48, embora para cada par ordenado $\hat{c}_{j|i}$ e $\hat{d}_{j|i}$ só serão estimados se $\hat{a}_{j|i} = 0$ e $\hat{b}_{j|i} < 0$.

Novamente necessitamos escolher um limiar para o nosso modelo. Como para os dados completos, por simplicidade, limitamos nossa busca por um limiar da estrutura

de dependência $u_{Y_i} = u \ \forall i$. O limiar da estrutura de dependência escolhido é tal que $Pr(Y_i \leq u) = 0,7$ para cada um dos dois períodos.

Tabela 7.5: Parâmetros estimados do modelo de dependência para os dois períodos.

	Parâmetros	ES| Faj	Can| Faj	Faj| Esp	Can| ES	Faj| Can	ES| Can
chuva	$\hat{a}_{j\|i}$	1,00	0.87	0.62	0.73	0,96	0.73
	$\hat{b}_{j\|i}$	0,32	0,39	0,38	0,02	0,56	0,66
Seca	$\hat{a}_{j\|i}$	0,43	0,62	0,58	0,63	0,50	0,30
	$\hat{b}_{j\|i}$	0,41	0,15	0,40	-0,13	0,30	0,36

A Tabela 7.5 mostra os valores estimados $(\hat{a}_{j|i}, \hat{b}_{j|i})$, $\forall$i$\neq j$, nos dois períodos. Novamente, para nenhum dos modelos de dependência foi necessário estimar os valores de $c_{j|i}$ e $d_{j|i}$, pois não obtivemos nenhum par em que $\hat{a}_{j|i} = 0$ e $\hat{b}_{j|i} < 0$. Pela Tabela 7.5, todos os modelos são assintoticamente independentes com estrutura de dependência extremal positiva.

Figura 7.2: Gráfico dos fluxos dos máximos mensais separados por períodos.

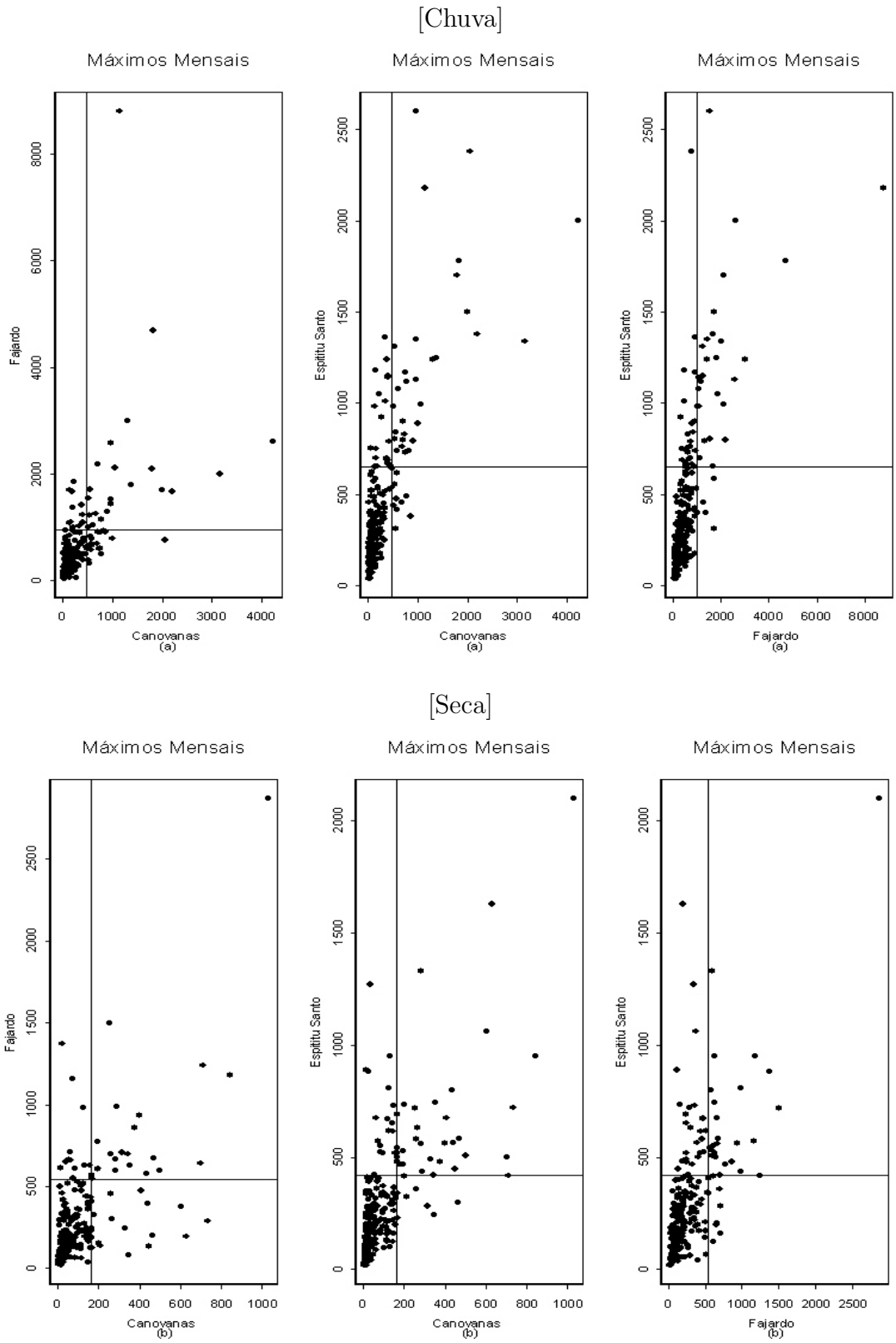
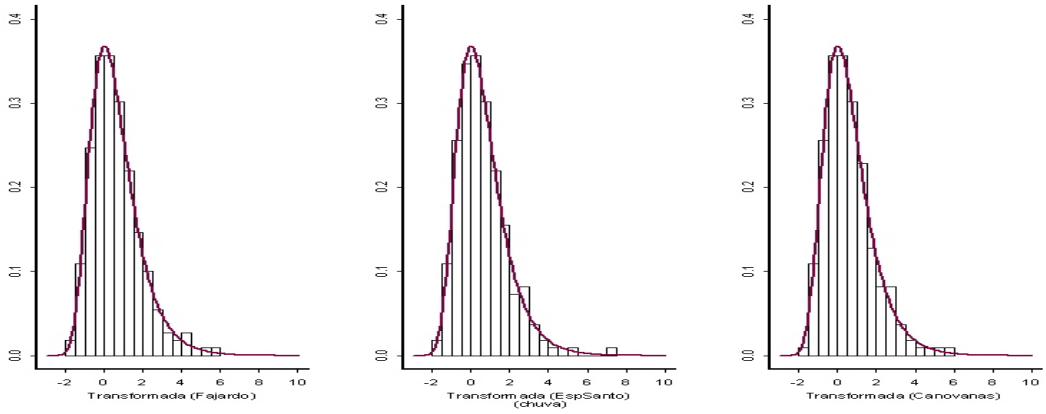


Figura 7.3: Histograma do ajuste dos máximos mensais nos dois períodos.

[Chuva]



[Seca]

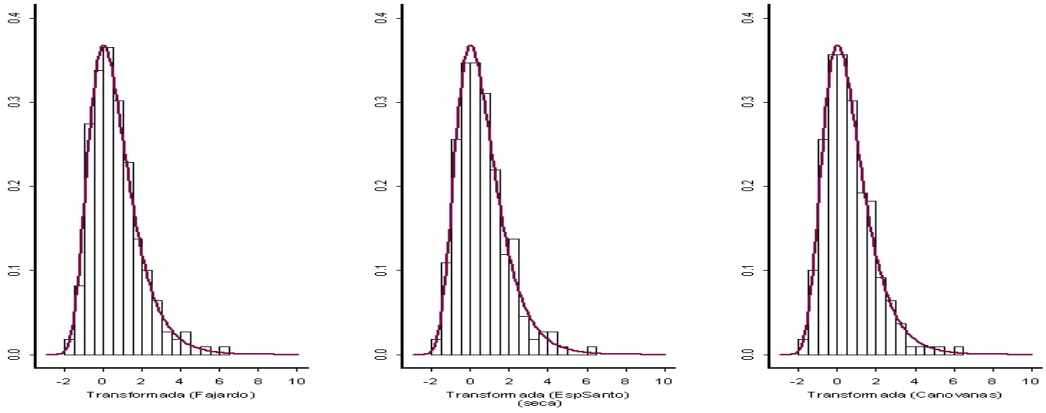
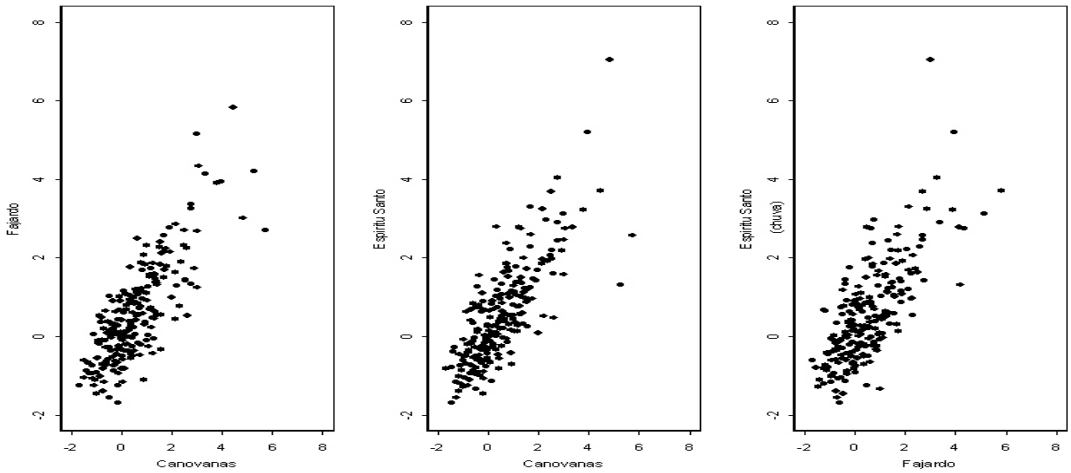
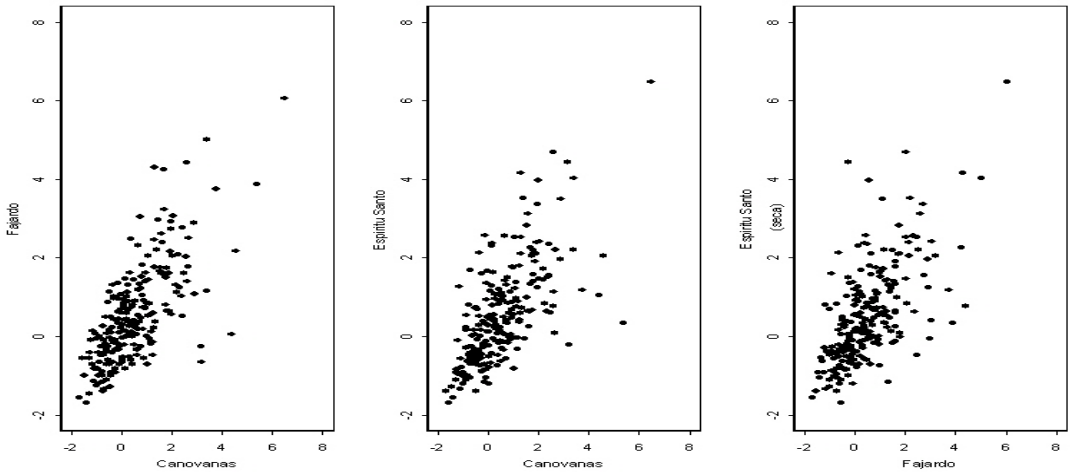


Figura 7.4: Máximos mensais de cada período transformados para a distribuição Gumbel.

[Chuva]



[Seca]



Conclusão

A aplicação do modelo condicional para valores extremos multivariados resultou em estimativas precisas para a vazão dos rios Canóvanas, Fajardo e Espiritu Santo, considerando ou não sazonalidade.

Além da vantagem de ser um modelo que pode ser aplicado a situações de qualquer dimensão, esse modelo também mostrou que podemos "quantificar"a dependência entre as variáveis, bem como estimar os efeitos de dependência. Os resultados dessa abordagem podem nos auxiliar em tomadas de decisão.

Dentre os resultados obtidos nesta análise, destacamos a importância da influência de diversos fatores hidrológicos em relação à vazão dos rios, além de evidenciar a importância dos estudos de previsão de riscos extremos.

Referências Bibliográficas

Besag, J. (1975). Statistical analysis of non-lattice data. *Statisticin*(24), 179-195.

Coles, S. G., Heffernan, J. E., & Tawn, J. A. (1991). Modelling extreme multivariate events. *J. R. Statist Soc. B*(53), 377-392.

Coles, S. G., & Tawn, J. A. (1994). Statistical methods for multivariate extremes: an application to structural desig with discussion. *Appl. Statist.*(43), 1-48.

Davison, A. C., & Smith, R. L. (1990). Models for exceedances over high thresholds (with discussion). *J. R. Statist. Soc. B*(52), 393-442.

Dekkers, A. l. M., Einmahl, J. H. J., & Haan, L. de. (1989). A moment estimator for the index of an extreme value distribution. *Ann Statist.*(17), 1833-1855.

Embrechts, P., Kluppelberg, C., & Mikosch, T. (1997). *Modelling extremal events.* Londom: Springer.

Haan, L. de. (1984). Variation and the characterization of domains of attaction. *Statistics Extrems and Applications*(40), 31-48.

Haan, L. de, & Ronde, J. de. (1998). Sea and wind: multivariate extremes at work. *Extremes*(1), 7-45.

Héctor, D., Glenn, S., & Frederic, N. (2002). Prediction of master recession curves and baseflow recessions in the luquilo mountains of puerto rico. *Journal of the American Water Resources Association*, *38*(3), 693-704.

Heffernan, J. E., & Tawn, J. A. (2004). A conditional approach for multivariate extreme values. *J. R. Statist. Soc. B*(66), 497-546.

Joe, H. (1994). Multivariate extreme-value distributions with applications to environmental data. *Can. J. Statist.*(22), 47-64.

Leadbetter, M. R., Lindgren, G., & Rootzén, H. (1983). *Extremes and related properties of random sequences and series.* New York: Springer.

Longin, F. M. (2000). From value at risk to stress testing: the extreme value approach. *J. Bnkng. Finan.*(24), 1097-1130.

Maulik, K., Resnick, S. I., & Rootzen, H. (2002). Asymptotic independence and a network trafic model. *J. Appl. Probal.*(39), 671-699.

Poon, S. H., Rockinger, M., & Tawn, J. A. (1981). Extreme-value dependence in financial markets: diagnostics, model and financial implications. *Rev. Finan. Stud.*(17), 581-610.

Resnick, S. I., & Rootzen, H. (2000). Self-similar communication models and very heavy tails. *Ann Appl, Probab.*(10), 753-778.

Schlather, M., & Tawn, J. A. (2000). A dependence for multivariate and spatial extreme values: properties and inference. *Biometrika.*, *90*(90), 139-156.

Shi, D., Smith, R. L., & Coles, S. G. (1992). *Joint versus marginal estimation for bivariate extremes.* Technical Report 2074. Department of Statistics, University of North Carolina at Chapel Hill, Chapell Hill.

Smith, R. L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground level ozone (with discussion). *Statist. Sci.*(4), 367-393.

Staricã, C. (2000). Multivariate extremes for models with constant conditional correlations. *J. Emp. Finan.*(6), 513-553.

Apêndice

Teste KPSS (Kwiatkowski et al,1992)

Para testar a estacionariedade das séries do fluxo dos rios foi utilizado o teste KPSS. O modelo de regressão com tendência linear tem a forma

$$X_t = c + \mu\, t + k \sum_{i=1}^{t} \xi_i + \eta_t, \tag{7.1}$$

com η_t estacionário e ξ_t i.i.d. com média zero e variância. Quando $k \neq 0$, o processo é integrado e quando $k = 0$ o processo é estacionário em torno de uma tendência linear. A hipótese nula é $H_0 : k = 0$ e a hipótese alternativa é $H_1 : k \neq 0$.

Sob H_0, a regressão (7.1) é ajustada pelo método dos mínimos quadrados, obtendo-se os resíduos $\hat{\eta}_t$. Usando esses resíduos a soma parcial

$$S_t = \sum_{i=1}^{t} \eta_i$$

é construída sob H_0.

A estatística teste $KPSS$ é definida como

$$\frac{\sum_{t=1}^{n} S_t^2}{n^2 \hat{w}^2},$$

onde

$$\hat{w}_T^2 = \hat{\sigma}_\eta^2 + 2 \sum_{\tau=1}^{n} \left(1 - \frac{\tau}{\tau - 1}\right) \hat{\gamma}_T$$

é o estimador da densidade espectral na frequência zero quando $\hat{\sigma}_\eta^2$ é estimador da variância de η_t e $\hat{\gamma}_T = \frac{1}{n}\sum_{t=\tau+1}^{n} \hat{\eta}_t\, \hat{\eta}_{t-\tau}$ é o estimador da covariância.

O problema deste teste consiste em determinar o ponto de referência T: para T muito pequeno, o teste é viciado quando há autocorrelação e, para T muito grande, o teste perde o poder.

Teste modificado R/S de Lo (1991)

Para verificar a existência de memória longa nas séries foi utilizado o teste modificado R/S de Lo. Existem inúmeros testes para detectar a existência de memória longa em séries temporais.

A hipótese nula é H_0: não há memória longa e a hipótese alternativa é H_1: há memória longa. A estatística teste RS_T é a amplitude das somas parciais dos desvios da média reescalonados pelo desvio padrão:

$$RS = \frac{1}{\sqrt{S_{N,q}^2}} \left[\max_{1\leq k\leq n} \sum_{j=1}^{k} (X_j - \bar{X}_N) - \min_{1\leq k\leq n} \sum_{j=1}^{k} (X_j - \bar{X}_N) \right],$$

onde N representa o tamanho da amostra, $\bar{X_N}$ a média amostral da variável X e $S_{N,q}^2$ é estimador HAC *(heterocedastic and autocorrelation consistent)* para a variância, $\sigma_X^2 = cov(X_j, X_0)$, dado por:

$$S_{N,q}^2 = \frac{1}{N}\sum_{j=1}^{N}(X_j - \bar{X}_N)^2 + 2\sum_{j=1}^{q} w_j(q)\gamma_j,$$

onde γ_j representa as autocovariâncias amostrais e $w_j(q) = \frac{1-j}{q+1}$ representa os pesos aplicados às autocovariâncias amostrais no *lag* j para responder a uma possível depêndencia de amplitude curta (*short range dependence*) até a q-ésima ordem. Lo (1991) mostrou que a estatística teste é consistente contra uma classe de processos estacionários de memória longa alternativos.

Teste de Kolmogorov

Usamos esse teste para avaliar a bondade do ajuste da GPD.

Definição: As hipóteses nula e alternativa do teste de Kolmogorov são:

- Hipótese nula: $H : F = F_0$
- Hipótese alternativa: $K : F \neq F_0$

A estatística teste de Kolmogorov está definida por:

$$D_n = sup_{-\infty<x<\infty}|\hat{F}_n(x) - F_0(x)|.$$

Rejeita-se H em favor de K quando os valores de D_n são grandes.

A Tabela 7.6 apresenta os valores críticos K_α do teste de Kolmogorov:

$$Pr(Dn \geq k_\alpha) = \alpha.$$

$n > 80$	α
1,682/$(\sqrt{n} + 0,12 + 0,11/\sqrt{n})$	0,10
1,358/$(\sqrt{n} + 0,12 + 0,11/\sqrt{n})$	0,05

Tabela 7.6: Valores críticos do teste de Kolmogorov

Teste de Mann Kendall para tendência e sazonalidade

O teste Mann-Kendall é uma técnica não-paramétrica que testa se há tendência crescente ou decrescente. Uma vantagem deste teste é que os resultados são fáceis de interpretar e se aplicam a todo período em estudo. Porém, uma desvantagem é que a variação sazonal nos dados não é levada em conta.

O teste Mann-Kendall para sazonalidade determina se existe uma tendência durante cada mês; i.e., se os valores apresentaram em um mês particular (por exemplo, janeiro) aumento ou diminuição em relação ao período estudado. Este teste é útil porque leva

em conta o fato de que pode haver variações sazonais nos dados. O teste de Mann-Kendall para sazonalidade também foi executado separadamente para os fluxos dos rios nos periodos de seca e chuva.

Usamos os testes Kendall e Mann-Kendall sazonal através do pacote computacional Kendall(R). Há dois resultados principais nestes testes. O primeiro é o p-valor, ou nível de significância. Este número descreve se a tendência é significativamente diferente de zero. O segundo valor é τ que é uma medida da força da tendência. τ é determinado como um número entre -1 e $+1$, onde o sinal ($-$ ou $+$) indica a direção de tendência

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Administração
Baixar livros de Agronomia
Baixar livros de Arquitetura
Baixar livros de Artes
Baixar livros de Astronomia
Baixar livros de Biologia Geral
Baixar livros de Ciência da Computação
Baixar livros de Ciência da Informação
Baixar livros de Ciência Política
Baixar livros de Ciências da Saúde
Baixar livros de Comunicação
Baixar livros do Conselho Nacional de Educação - CNE
Baixar livros de Defesa civil
Baixar livros de Direito
Baixar livros de Direitos humanos
Baixar livros de Economia
Baixar livros de Economia Doméstica
Baixar livros de Educação
Baixar livros de Educação - Trânsito
Baixar livros de Educação Física
Baixar livros de Engenharia Aeroespacial
Baixar livros de Farmácia
Baixar livros de Filosofia
Baixar livros de Física
Baixar livros de Geociências
Baixar livros de Geografia
Baixar livros de História
Baixar livros de Línguas

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo