UNIVERSIDADE FEDERAL DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM SANEAMENTO, MEIO AMBIENTE E RECURSOS HÍDRICOS

ANÁLISE BAYESIANA DE FREQÜÊNCIA DE VAZÕES MÁXIMAS ANUAIS COM INFORMAÇÕES HISTÓRICAS:

APLICAÇÃO À BACIA DO RIO SÃO FRANCISCO EM SÃO FRANCISCO

Fernando Alves Lima

Belo Horizonte
2005

ANÁLISE BAYESIANA DE FREQÜÊNCIA DE VAZÕES MÁXIMAS ANUAIS COM INFORMAÇÕES HISTÓRICAS:

APLICAÇÃO À BACIA DO RIO SÃO FRANCISCO EM SÃO FRANCISCO

Fernando Alves Lima

Fernando Alves Lima

ANÁLISE BAYESIANA DE FREQÜÊNCIA DE VAZÕES MÁXIMAS ANUAIS COM INFORMAÇÕES HISTÓRICAS:

APLICAÇÃO À BACIA DO RIO SÃO FRANCISCO EM SÃO FRANCISCO

Dissertação apresentada ao Programa de Pós-graduação em Saneamento, Meio Ambiente e Recursos Hídricos da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Saneamento, Meio Ambiente e Recursos Hídricos.

Área de concentração: Hidráulica e Recursos Hídricos

Linha de pesquisa: Modelos Estocásticos em Hidrologia

Orientador: Mauro da Cunha Naghettini

Belo Horizonte
Escola de Engenharia da UFMG
2005

UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Escola de Engenharia
Programa de Pós-Graduação em Saneamento, Meio Ambiente e Recursos Hídricos
Av. Contorno 842 – 7º andar 30110-060 Belo Horizonte – BRASIL
Tel: 55 (31) 3238-1882 Fax: 55 (31) 3238-1882 posgrad@desa.ufmg.br
www.smarh.eng.ufmg.br

FOLHA DE APROVAÇÃO

Análise Bayesiana de Frequência de Vazões Máximas
Anuais com Informações Históricas:
Aplicação à Bacia do Rio São Francisco em São Francisco

FERNANDO ALVES LIMA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Prof. MAURO DA CUNHA NAGHETTINI

Prof. LUIZ RAFAEL PALMIER

Prof. EDUARDO SÁVIO PASSOS RODRIGUES MARTINS

PESQ. EBER JOSÉ DE ANDRADE PINTO

Aprovada pelo Colegiado do PG SMARH

Profa. Mônica Maria Diniz Leão
Coordenadora

Versão Final aprovada por

Prof. Mauro da Cunha Naghettini
Orientador

Belo Horizonte, 17 de outubro de 2005.

AGRADECIMENTOS

Gostaria de agradecer:

em especial, aos meus pais, exemplos de dedicação e perseverança, responsáveis por todos os passos de uma jornada iniciada 26 anos atrás;

aos meus irmãos, pelo forte laço que nos une e pelo aprendizado que a convivência diária me proporciona;

à Wanessa, não só pelo amor e carinho, que tanto me ajudam, mas também pela paciência e compreensão na etapa final desse trabalho;

ao professor e amigo Mauro, pelo exemplo de seriedade, compromisso e satisfação com o seu trabalho, pela grande ajuda nessa pesquisa, pelas contribuições à minha formação profissional e pela motivação daquelas palavras "*fica tranqüilo!*";

à vovó Benvinda, que, embora não tenha recebido da vida uma saudação digna de seu nome, é um grande símbolo de força e coragem;

a todos os meus familiares, certo de que, lá no fundo, cada um sabe bem a sua importância;

aos amigos da sala 909, pois o sentimento de que estávamos todos no mesmo barco me serviu de incentivo;

aos professores e funcionários do Departamento de Engenharia Hidráulica e Recursos Hídricos, pelas contribuições diversas;

aos amigos da CPRM, pela presteza com que sempre me atenderam.

RESUMO

A determinação da probabilidade de que uma certa vazão seja igualada ou excedida em um ao qualquer, objeto da análise de freqüência de cheias, constitui um problema corrente na engenharia. No entanto, essa vazão de interesse está freqüentemente associada a um período de retorno substancialmente mais longo que o dos registros fluviométricos regulares. A análise de freqüência tradicional de vazões de enchentes, baseada exclusivamente nas amostras usualmente curtas de dados sistemáticos, pode conduzir a estimativas pouco realistas da probabilidade de excedência de eventos extremos. Para reduzir as incertezas envolvidas e aumentar o nível de confiança nas estimativas dos parâmetros e quantis, deve-se utilizar toda a informação disponível. Além dos dados sistemáticos, informações sobre cheias históricas e paleohidrológicas (paleovazões) podem ser encontradas e incorporadas à análise de freqüência. Essas informações permitem aumentar substancialmente o tamanho da amostra, reduzindo o grau de extrapolação e tornando menos incertas as inferências feitas no domínio das cheias extremas. A teoria bayesiana também constitui uma ferramenta estatística importante para a análise de freqüência de cheias. Na abordagem bayesiana, os parâmetros do modelo probabilístico são tratados como variáveis aleatórias, modeladas por uma distribuição de probabilidades *a priori*, formulada, por exemplo, à luz de algum conhecimento subjetivo acumulado por especialistas ou de informações provenientes de análise regional. Utilizando o teorema de Bayes, essa distribuição *a priori*, que pode ser informativa ou não, é então "atualizada" pelos dados fluviométricos locais, resultando na distribuição *a posteriori*. Assim, usando a distribuição *a posteriori* dos parâmetros e a distribuição de probabilidades das cheias, condicionada aos parâmetros, é possível determinar uma distribuição de probabilidades marginal para as cheias, independente das estimativas dos parâmetros. Essa dissertação busca avaliar o ganho, em termos de redução de incertezas, do uso da abordagem bayesiana e das cheias históricas na análise de freqüência de vazões máximas anuais. Para isso, um estudo de caso é apresentado, o qual consiste na realização da análise de freqüência de vazões máximas anuais no rio São Francisco em São Francisco, utilizando: (1) os registros sistemáticos de vazão do referido posto fluviométrico, (2) as informações sobre cheias históricas coletadas na bacia hidrográfica do São Francisco, e (3) a teoria bayesiana, com uma distribuição *a priori* informativa (tanto do ponto de vista da prescrição do modelo distributivo como das estimativas de seus parâmetros) ou não informativa.

ABSTRACT

The estimation of the exceedance probability of a rare flood is a current problem in engineering. However, rare floods are associated to high return periods which are much longer than the time span covered by systematic streamflow records. The conventional flood frequency analysis, which is based on the usually short systematic streamflow data samples, may yield unrealistic estimates of the risks associated with extreme events. In order to reduce the uncertainties and to produce more reliable parameter and quantile estimates, every piece of available information should be used. Beyond systematic streamflow data, information on historical floods and paleofloods may be found and incorporated into the flood frequency analysis. Such pieces of information may augment the sample size, thus reducing the range of extrapolation and yielding more reliable inferences in the domain of extreme floods. Bayesian theory is also an important statistical tool for flood frequency analysis. According to the Bayesian approach, the distribution parameters are treated as random variables, being modeled by a prior probability distribution, which may be formulated on the basis, for instance, of a subjective prior knowledge, as provided by a specialized professional, or additional information from regional analysis. By using Bayes' theorem, this prior distribution, which may be informative or not, is then updated by the local streamflow data, thus producing the posterior distribution of the parameters. Hence, by using this posterior distribution along with the flood probability distribution, conditioned to the parameters, it is possible to find a marginal probability function to floods, independently of parameter estimates. The present MSc thesis aims to evaluate the gain, in terms of uncertainty reduction, from using the Bayesian approach, along with information on historical floods, into the frequency analysis of annual flood maxima. In order to perform it, a case study for the São Francisco river basin, at the location of São Francisco, is presented according to the following scenarios: (1) only the systematic flood records are used; (2) information on historical floods are incorporated; and (3) Bayesian theory is employed, with and without an informative prior distribution on parameters and on the probabilistic model.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

$E[X]$	Valor esperado de *X*
$Var[X]$	Variância de *X*
μ	Média
σ	Desvio-padrão
φ	Variância
C_V	Coeficiente de variação
γ	Coeficiente de assimetria
κ	Curtose
μ_r	Momento populacional central de ordem *r*
$\mu_r^{'}$	Momento populacional de ordem *r* em relação à origem
β_r	Momento populacional ponderado por probabilidade de ordem *r*
λ_r	Momento-L populacional de ordem *r*
τ	Coeficiente de variação-L populacional
τ_3	Coeficiente de assimetria-L populacional
τ_4	Coeficiente de curtose-L populacional
m_r	Momento amostral de ordem *r*
b_r	Momento amostral ponderado por probabilidade de ordem *r*
l_r	Momento-L amostral de ordem *r*
t_r	Quociente de momentos-L amostral de ordem *r*
EXP	Distribuição Exponencial
LNO	Distribuição Log-Normal
GUM	Distribuição Gumbel
GEV	Distribuição Generalizada de Valores Extremos
GLO	Distribuição Logística Generalizada
GPA	Distribuição Generalizada de Pareto
PE3	Distribuição Pearson III
WEI	Distribuição Weibull
T	Período de retorno
x_T	Quantil de período de retorno *T*
$P[A]$	Probabilidade de ocorrência do evento *A*
$P[A^c]$	Probabilidade de ocorrência do evento *A complemento*

$P[A \cap B]$	Probabilidade de ocorrência da interseção do evento *A* com o evento *B*
$P[A \mid B]$	Probabilidade de ocorrência do evento *A*, dado que o evento *B* ocorreu
N_S	Número de anos do período sistemático
$N_S^{\bullet}$	Número de anos do período sistemático com cheias de magnitude conhecida
$N_S^{<}$	Número de anos do período sistemático com cheias de magnitude inferior a um limiar de referência superior x_U
$N_S^{>}$	Número de anos do período sistemático com cheias de magnitude superior a um limiar de referência inferior x_L
$N_S^{<>}$	Número de anos do período sistemático com cheias de magnitude compreendida no intervalo $[x_L; x_U]$
N_H	Número de anos do período não sistemático
$N_H^{\bullet}$	Número de anos do período não sistemático com cheias de magnitude conhecida
$N_H^{<}$	Número de anos do período não sistemático com cheias de magnitude inferior a um limiar de referência superior y_U
$N_H^{>}$	Número de anos do período não sistemático com cheias de magnitude superior a um limiar de referência inferior y_L
$N_H^{<>}$	Número de anos do período não sistemático com cheias de magnitude compreendida no intervalo $[y_L; y_U]$
y_H	Limiar de referência (*threshold*) da amostra censurada
W	Fator de ponderação do método dos momentos ponderados historicamente
ERL	Tamanho efetivo da amostra
MSE	Erro quadrático médio
MG	Ganho marginal
$F_X(x)$	Função acumulada de probabilidade
$f_X(x)$	Função densidade de probabilidade
$L(\Theta \mid X)$	Função de verossimilhança dos parâmetros, dada a amostra
Θ	Vetor de parâmetros do modelo distributivo
$\hat{\Theta}$	Estimadores tradicionais dos parâmetros do modelo distributivo
$\tilde{f}(x)$	Distribuição bayesiana de probabilidades
$f_0(\Theta)$	Distribuição de probabilidades *a priori* dos parâmetros (ou apenas *prior*)
$f_1(\Theta \mid X)$	Distribuição de probabilidades *a posteriori* dos parâmetros (ou apenas *posterior*)

μ_Θ	Vetor das médias dos parâmetros
σ_Θ	Vetor dos desvios-padrão dos parâmetros
R_Θ	Matriz dos coeficientes de correlação dos parâmetros
Σ_Θ	Matriz de covariância dos parâmetros
$I(\Theta)$	Distribuição de probabilidades usada na técnica *importance sampling*
Θ^*	Estimadores bayesianos dos parâmetros do modelo distributivo
ParE	Distribuição dos parâmetros esperados
ProE	Distribuição da probabilidade esperada
Di	Medida de discordância
H	Medida de heterogeneidade regional
Z^{DIST}	Medida de aderência da distribuição *"DIST"*
Γ	Função gama

1 INTRODUÇÃO

A presente dissertação de mestrado foi desenvolvida no âmbito do Programa de Pós-graduação em Saneamento, Meio Ambiente e Recursos Hídricos da Universidade Federal de Minas Gerais (UFMG), o qual é conduzido conjuntamente pelo Departamento de Engenharia Sanitária (DESA) e pelo Departamento de Engenharia Hidráulica e Recursos Hídricos (EHR).

1.1 Contextualização

Desde os primórdios da história, é possível constatar a preferência natural da espécie humana pela ocupação das áreas situadas às margens de cursos d'água. De fato, essas áreas favorecem o avanço econômico e social das comunidades, oferecendo não só as condições propícias para o desenvolvimento de atividades agrícolas, industriais e turísticas, mas também contribuindo com o bem estar social, por meio do fornecimento de água e alimentos. No entanto, as comunidades ribeirinhas podem, algumas vezes, pagar um alto preço por esses benefícios, uma vez que, ao ocuparem indiscriminadamente as planícies fluviais, estão sujeitas aos riscos de inundação. As enchentes mais severas podem causar verdadeiras catástrofes naturais, com graves prejuízos econômicos (danos materiais, destruição da infra-estrutura local e perturbação das atividades econômicas) e sociais (problemas de saúde pública e perda de vidas humanas). Um exemplo é a grande enchente ocorrida em 1979 na bacia do rio São Francisco. Esse rio, um dos mais importantes do Brasil, experimentou naquele ano uma cheia de grandes proporções, levando destruição a diversas cidades (veja, por exemplo, as Figuras 1.1, 1.2, 1.3 e 1.4).

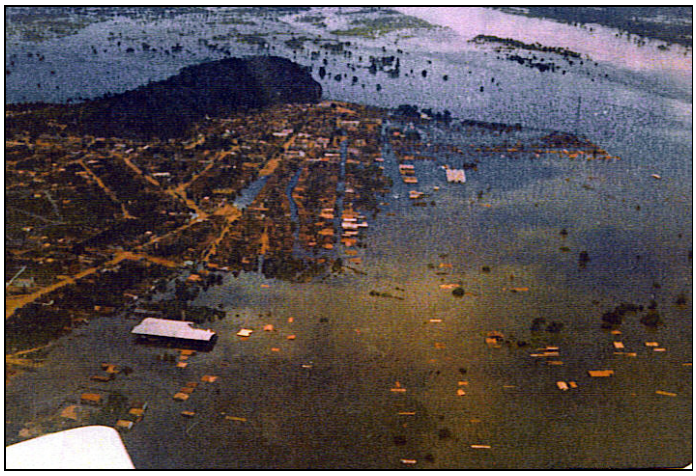


FIGURA 1.1: Cidade de Bom Jesus da Lapa (BA) inundada pela enchente de 1979.
Fonte: CPRM, 1979.

FIGURA 1.2: Cidade de Malhada (BA) inundada pela enchente de 1979.
Fonte: CPRM, 1979.

FIGURA 1.3: Cidade de Januária (MG) inundada pela enchente de 1979.
Fonte: CPRM, 1979.

FIGURA 1.4: Cidade de São Francisco (MG) inundada pela enchente de 1979.
Fonte: CPRM, 1979.

Vale salientar que o conceito de risco de inundação está associado a dois fatores distintos: a aleatoriedade do fenômeno natural, que está ligada às características hidrológicas e hidrometeorológicas da bacia, e a vulnerabilidade a que se expõe uma certa comunidade, que depende das estratégias de uso e ocupação do solo. Sendo assim, o risco de inundação se refere, simultaneamente, à ocorrência de uma vazão de enchente, que constitui um processo estocástico, e à presença de danos econômicos e sociais causados pelo evento. A Figura 1.5 ilustra o conceito de risco de inundação.

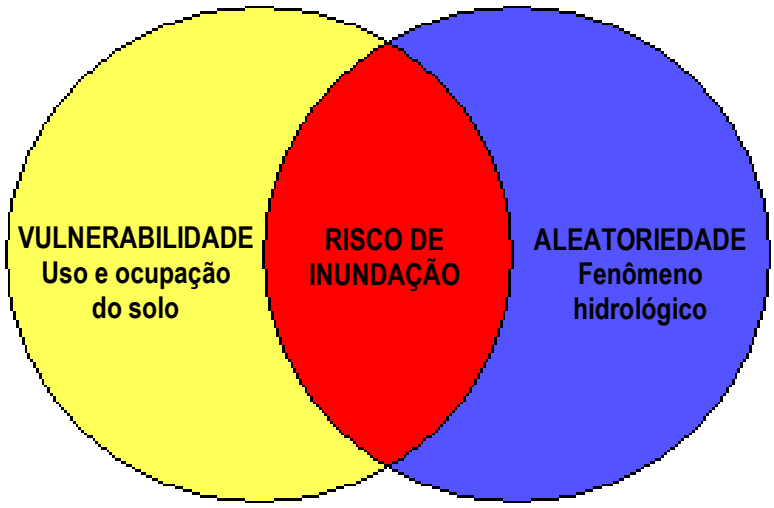


FIGURA 1.5: Ilustração do conceito de risco de inundação.
Fonte: adaptado de NAULET, 2002.

Nesse contexto, a determinação da magnitude e da freqüência de ocorrência de vazões de enchentes constitui uma das mais importantes e complexas aplicações da teoria de probabilidades e da estatística matemática na hidrologia: a chamada *análise de freqüência*. Essa aplicação fornece subsídios não só para a adoção de medidas que visam a redução dos riscos de inundação, dentre as quais cita-se a construção de reservatórios, o erguimento de diques de proteção e o ordenamento do uso e ocupação das planícies fluviais, mas também para o projeto e a operação de estruturas hidráulicas destinadas ao aproveitamento de recursos hídricos.

Em suma, a análise de freqüência busca extrair inferências quanto à probabilidade com que uma variável aleatória irá igualar ou exceder um determinado quantil de interesse, a partir de um conjunto de observações daquela variável. No entanto, os métodos convencionais de análise de freqüência de vazões de enchentes, baseados no ajuste de uma distribuição de

probabilidades à série de dados observados e na extrapolação de sua cauda superior para estimar a probabilidade de excedência de eventos extremos, estão sujeitos a um elevado grau de incertezas. Segundo Benjamim e Cornell (1970), as incertezas envolvidas na estimação da probabilidade de excedência de vazões extremas, as quais interferem no processo de decisão, podem ser divididas em três categorias:

1) *incerteza natural*, em virtude da aleatoriedade presente no fenômeno hidrológico em estudo, caracterizado por um processo estocástico;
2) *incerteza estatística*, associada à estimação dos parâmetros da distribuição de probabilidades adotada para modelar o processo estocástico, utilizando amostras de pequeno tamanho e sujeitas a erros de medição ou de extrapolação da curva-chave;
3) *incerteza do modelo*, devido à inexistência de leis dedutivas para seleção da distribuição de probabilidades a ser empregada, especialmente levando-se em conta que alguns modelos distributivos, cuja qualidade do ajuste aos eventos ordinários não apresenta variação significativa, podem exibir comportamentos bastante distintos em suas caudas superiores.

Para obter estimativas mais confiáveis, deve-se tentar minimizar os efeitos das incertezas mencionadas, utilizando, racionalmente, todo tipo de informação que estiver disponível.

Especialmente no caso da análise de freqüência de vazões de enchentes, as séries de dados sistemáticos são usualmente muito curtas, dificultando a obtenção de estimativas confiáveis da probabilidade de excedência de vazões extremas. Assim, a utilização de séries de vazões máximas anuais, constituídas apenas pelas observações fluviométricas regulares, submetidas à análise local de freqüência, pode deixar de considerar alguma informação adicional valiosa. Felizmente, estas séries não são a única fonte de informação disponível. Um aumento do tamanho da amostra pode ser obtido por meio da utilização de séries de duração parcial (*peaks over threshold*), da análise conjunta dos dados fluviométricos de diferentes estações de medição, empregando técnicas de regionalização hidrológica, ou ainda da incorporação de informações históricas e/ou paleohidrológicas sobre cheias (dados não sistemáticos). Essa última alternativa será analisada na presente dissertação, especialmente no que se refere ao uso das cheias históricas. Um ponto delicado quanto à utilização dos dados não sistemáticos de cheias é o problema da estacionariedade dos fenômenos naturais ao longo do tempo. A esse respeito, Duband (1994) apresenta uma opinião controversa ao afirmar que, embora uma

eventual mudança climática seja tão mais evidente quanto mais longa for a série disponível, seu impacto é maior em parâmetros meteorológicos, tais como temperatura e pressão atmosférica. Ainda segundo o referido autor, ao nível das principais variáveis hidrológicas, chuva e vazão, resultantes da interação de inúmeros processos, essas variações podem ser imperceptíveis. Essa opinião está longe de representar um consenso no meio técnico e científico; certas correntes defendem a posição de que as variáveis chuva e vazão também sejam claramente influenciadas por possíveis mudanças climáticas.

Outra ferramenta importante é a teoria bayesiana, que, ao tratar explicitamente os parâmetros do modelo distributivo como variáveis aleatórias, modeladas por uma certa distribuição, permite obter, para as vazões de enchentes, uma distribuição marginal de probabilidades. Essa distribuição leva em conta todos os valores possíveis dos parâmetros, minimizando as incertezas estatísticas envolvidas. O uso da abordagem bayesiana na análise de freqüência de vazões máximas anuais também será discutido nessa dissertação.

Finalmente, um estudo de caso é realizado, com o intuito de aplicar a uma bacia hidrográfica brasileira os conhecimentos referentes à incorporação dos dados não sistemáticos e ao uso da abordagem bayesiana na análise de freqüência de vazões máximas anuais.

1.2 Organização da dissertação

No Capítulo 1, é feita uma contextualização do problema associado à ocorrência de grandes cheias, apresentando seus impactos, bem como algumas alternativas para o tratamento da questão de forma mais precisa. No Capítulo 2, são apresentados os objetivos da pesquisa. O Capítulo 3 constitui uma revisão sobre os procedimentos de análise de freqüência de vazões máximas anuais, enfocando, especialmente, o uso de informações não sistemáticas sobre cheias. O Capítulo 4 trata do emprego da abordagem bayesiana na análise de freqüência de vazões máximas anuais. No Capítulo 5, um estudo de caso é desenvolvido, aplicando-se à bacia do rio São Francisco os conhecimentos apresentados nos capítulos anteriores. O Capítulo 6 refere-se às conclusões e recomendações do presente trabalho.

2 OBJETIVOS

2.1 Objetivo geral

O objetivo geral dessa pesquisa é avaliar o ganho, em termos de redução de incertezas, do emprego da abordagem bayesiana e da incorporação de cheias históricas na análise de freqüência local de vazões máximas anuais.

O emprego da abordagem bayesiana visa prover uma distribuição marginal de probabilidades, supostamente livre das incertezas envolvidas na estimação dos parâmetros. Nesse contexto, pode-se utilizar ou não alguma informação qualificada *a priori*, relativa não só à prescrição do modelo distributivo como também às estimativas de seus parâmetros. Já o uso de cheias históricas, em conjunto com as observações fluviométricas regulares, busca estender o tamanho da amostra, possibilitando a obtenção de estimativas paramétricas mais confiáveis e reduzindo o grau de extrapolação da cauda superior da distribuição de probabilidades.

2.2 Objetivos específicos

Os objetivos específicos dessa pesquisa são:

- buscar, interpretar e processar informações sobre cheias históricas, de forma a poder utilizá-las na análise de freqüência de vazões de enchentes;
- formular uma distribuição de probabilidades capaz de expressar o que se conhece *a priori* sobre os parâmetros do modelo, utilizando algum conhecimento subjetivo acumulado por especialistas ou informações provenientes de análise regional;
- realizar um estudo de caso, utilizando os dados sistemáticos, as cheias históricas e a teoria bayesiana na análise de freqüência local de vazões máximas anuais;
- avaliar o impacto do uso da abordagem bayesiana e das cheias históricas na determinação do risco associado a eventos extremos.

3 ANÁLISE DE FREQÜÊNCIA DE VAZÕES MÁXIMAS ANUAIS

3.1 Introdução

O objetivo da análise de freqüência de vazões de enchentes é inferir a probabilidade de que eventos de determinada magnitude sejam igualados ou excedidos em um intervalo de tempo futuro, a partir da análise de uma série temporal da variável hidrológica em questão. As séries temporais reúnem as observações ou medições da referida variável hidrológica, organizadas seqüencialmente de acordo com sua ocorrência no tempo, registrando, dessa forma, a variabilidade do fenômeno natural de interesse. No caso específico de eventos hidrológicos extremos, tais como vazões máximas, as séries temporais empregadas na análise de freqüência podem ser anuais, quando são constituídas pelos máximos registrados em cada ano hidrológico em uma determinada estação fluviométrica, ou de duração parcial (*peaks over threshold*), quando são constituídas por todos os registros de magnitude superior a um determinado limiar de referência, independentemente do número de eventos selecionados por ano hidrológico. Nessa dissertação, serão abordados apenas os conceitos relativos à análise de freqüência de vazões máximas utilizando séries de duração anual. Maiores detalhes referentes ao uso das séries de duração parcial podem ser encontrados em Todorovic e Zelenhasic (1970) e Bradley e Potter (1992).

No contexto da análise de freqüência, as séries temporais constituem um sub-conjunto com um número limitado de observações (*amostra*), extraído do conjunto de todas as realizações possíveis de uma variável hidrológica (*população*). Pressupõe-se que uma dada amostra tenha sido sorteada aleatoriamente, dentre um número infinito de outras amostras que também poderiam, com igual chance, ser extraídas da população. Além disso, os dados hidrológicos que compõem as séries temporais devem satisfazer as seguintes condições: (1) *independência*, que significa a inexistência de correlação entre os elementos da série; (2) *estacionariedade*, que se refere ao fato dos dados serem identicamente distribuídos, ou seja, terem sido extraídos de uma mesma população; e (3) *representatividade*, isto é, capacidade de representar adequadamente a variabilidade inerente ao fenômeno natural de interesse.

Uma vez que se disponha de uma amostra com as características descritas anteriormente, ela poderá ser utilizada na análise de freqüência para extrair conclusões sobre o comportamento populacional da variável hidrológica em questão. Na abordagem estatística clássica, uma

função paramétrica de probabilidades é previamente selecionada para representar a distribuição da população; os parâmetros do modelo são, então, estimados a partir da informação disponível, sumariada pela amostra. Esse procedimento caracteriza os chamados métodos paramétricos de análise de freqüência. Estimados os parâmetros, tem-se, portanto, uma distribuição de probabilidades particularizada para uma situação, que pode, agora, ser usada para inferir sobre probabilidades de cenários não observados.

Em adição aos métodos paramétricos da abordagem estatística clássica, outras técnicas podem ser empregadas na análise de freqüência, tais como os métodos não paramétricos e os métodos bayesianos. Os métodos não paramétricos consistem no ajuste gráfico de uma função contínua aos dados da série hidrológica, os quais, associados às respectivas probabilidades de excedência empíricas, são plotados em papéis de probabilidades. Os métodos bayesianos, que serão apresentados detalhadamente no Capítulo 4, permitem combinar à amostra de dados hidrológicos, informações objetivas e/ou subjetivas, avaliadas *a priori*, para expressar com menor grau de incerteza o comportamento paramétrico *a posteriori*. Os procedimentos de inferência bayesiana fornecem ainda ferramentas para a formulação de uma distribuição marginal de probabilidades para a variável hidrológica em análise, levando-se em conta todos os valores possíveis dos parâmetros.

Quanto à sua abordagem no espaço, a análise de freqüência pode ser classificada como local ou regional. Quando são utilizados apenas dados pontuais, registrados em uma única estação de medição, tem-se a análise de freqüência local. Em contrapartida, na análise de freqüência regional, são usados dados provenientes de vários postos hidrométricos/hidrometeorológicos, associados a certas similaridades fisiográficas e/ou climáticas de uma área geográfica, permitindo assim estimar a distribuição de freqüência da variável hidrológica em locais com amostras muito pequenas ou desprovidos de observações. No Anexo D, é apresentada uma síntese da metodologia de análise de freqüência regional utilizando momentos-L, proposta por Hosking e Wallis (1997).

3.2 Tipos de informações disponíveis para a análise de freqüência de vazões de enchentes

Nesse item, serão apresentados os diferentes tipos de informações que podem ser explorados ao se realizar uma análise de freqüência local de vazões máximas anuais. Conforme citado anteriormente, as séries de duração parcial estão fora do escopo desse trabalho.

Ouarda *et al.* (1998) classificou a informação não contemporânea, relativa às cheias, em três categorias: (1) evidências geológicas de cheias pré-históricas (paleohidrologia), (2) evidências botânicas de cheias pré-históricas, especialmente aquelas encontradas nos troncos das árvores (dendrohidrologia), e (3) observações registradas em jornais, arquivos e outros documentos. Baker (1987) propôs uma classificação diferente, em função da ordem cronológica dos eventos, de forma que o período completo de informações sobre cheias seja dividido em três períodos distintos: pré-histórico, histórico e contemporâneo. Essa classificação será adotada ao longo da presente dissertação, e as principais características de cada um dos três períodos serão descritas a seguir.

3.2.1 O período pré-histórico (paleovazões)

Esse é o período de domínio da paleohidrologia, técnica que fornece informações sobre eventos antigos de cheias, os quais não foram diretamente observados por seres humanos. Os dados paleohidrológicos são obtidos de maneira indireta, por meio de evidências físicas da ocorrência das chamadas "paleovazões".

As técnicas empregadas para a determinação das paleovazões combinam sedimentologia, geomorfologia e geobotânica, e se dividem em dois grupos principais: estudos de depósitos de sedimentos e estudos botânicos. Os depósitos de sedimentos são locais em que materiais característicos dos leitos dos cursos d'água (areia, silte e/ou pedregulho), que se encontravam em suspensão durante a ocorrência de grandes cheias, sofreram acumulação, particularmente onde as condições morfológicas favoreceram a redução da velocidade do escoamento. Cavidades rochosas são especialmente valiosas devido à preservação das camadas de sedimentos depositados ao longo dos anos. Terraços formados nas planícies de inundação também constituem elementos importantes a serem investigados, permitindo a identificação dos níveis atingidos pelas cheias pré-históricas. As informações botânicas, coletadas ao longo de um curso d'água, resultam da identificação e interpretação de anomalias nos anéis de crescimento das árvores e de "cicatrizes" em seus troncos, provocadas pela ocorrência de grandes cheias.

Os estudos de depósitos de sedimentos e os estudos botânicos fornecem indicadores físicos, baseados em evidências geológicas e botânicas, da ocorrência de cheias antigas. Esses indicadores, associados a técnicas de datação paleológica e de modelagem hidráulica, permitem avaliar a magnitude e a freqüência de cheias ocorridas durante um longo período de

tempo (desde séculos até milênios), possibilitando a construção de uma série de paleovazões, as quais excederam, ou não, diferentes limiares de referência (também chamados limites paleohidrológicos), em intervalos de tempo específicos.

Segundo Benito *et al.* (2004), as principais incertezas associadas às paleovazões são: (1) o nível d'água atingido durante a cheia, uma vez que o nível dos depósitos de sedimentos representa apenas uma estimativa inferior do nível máximo alcançado pelo pico da cheia, (2) as incertezas inerentes ao processo de datação das evidências físicas, e (3) a continuidade e completude dos registros de paleovazões, de forma que o período pré-histórico possa ser adequadamente representado. Outra dificuldade se refere à estacionariedade da população de cheias pré-históricas, hipótese dificilmente satisfeita, tendo em vista as variações climáticas que ocorreram em grande escala no planeta, como, por exemplo, as diferentes eras glaciais.

Maiores detalhes a respeito dos conhecimentos da paleohidrologia podem ser encontrados em Baker (1987) e House *et al.* (2001).

3.2.2 O período histórico

Em locais habitados por um longo período de tempo, as informações sobre cheias históricas podem estar disponíveis, constituindo o conjunto de dados do período histórico. Essas informações se referem a eventos de cheias diretamente observados por seres humanos, que ocorreram e foram documentados antes do período relativamente curto de observações fluviométricas regulares.

A coleta de informações sobre cheias históricas requer um trabalho meticuloso, que envolve a consulta a diversos volumes manuscritos, jornais e outros periódicos antigos, diários de viagens, órgãos administrativos locais, arquivos históricos e, até mesmo, relatos pessoais. É possível, inclusive, que técnicas auxiliares sejam necessárias, por exemplo, para a conversão de unidades de medida e para a correta interpretação de estilos de escrita.

Conforme descrito por Glaser (1996), a recuperação de informações sobre cheias históricas pode se deparar com algumas dificuldades. Primeiramente, há o grande volume e variedade de documentos em que as informações podem, esporadicamente ou sistematicamente, terem sido registradas. Além disso, essas informações nem sempre descrevem diretamente um fenômeno particular, mas, ao invés disso, expõem os seus efeitos ou impactos. Finalmente, nem todas as

informações refletem a realidade, uma vez que cópias sucessivas, reinterpretações, traduções e resumos podem distorcer o conteúdo do texto original. Assim, para assegurar a qualidade da informação, a coleta de dados deve, sempre que possível, se basear no documento primário.

Em geral, as informações históricas podem ir desde descrições dos danos causados pela cheia, até a data de ocorrência do evento, ou a última vez que uma cheia daquela magnitude ocorreu, ou, ainda, referências ao nível atingido durante a passagem do pico da cheia, o que pode ser confirmado por meio de investigações locais, buscando a presença de marcas em construções antigas (pontes, casas, igrejas, portões, muros, dentre outras). Essas informações, associadas à modelagem hidráulica, permitem avaliar a magnitude e a freqüência de cheias ocorridas durante o período histórico (em geral, alguns séculos), possibilitando a construção de uma série de cheias históricas, as quais, na maioria dos casos, são documentadas especialmente por terem excedido certos limiares de referência (também chamados limites de percepção) em intervalos de tempo específicos.

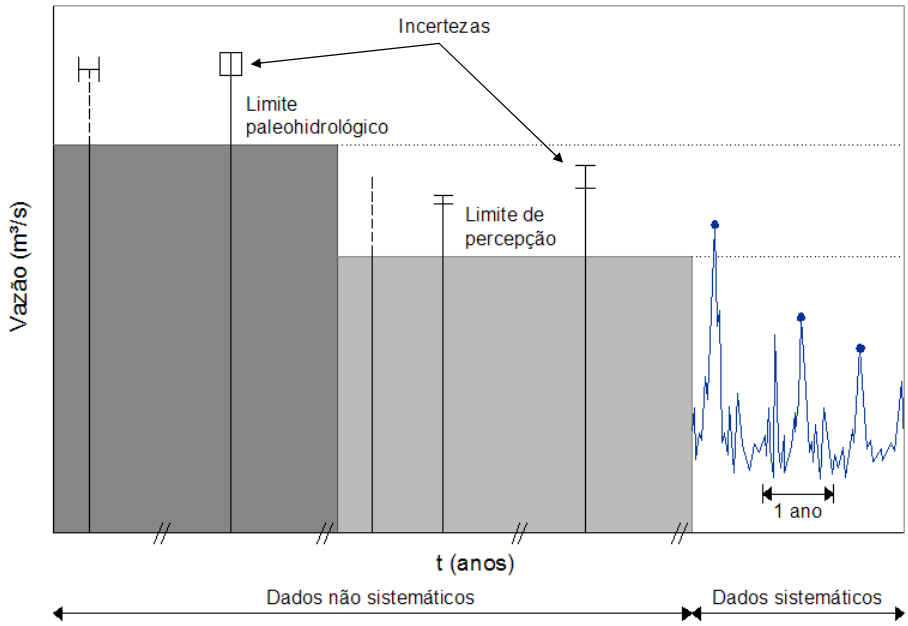
Assim como as paleovazões, as cheias históricas estão sujeitas às incertezas referentes à continuidade e à completude dos registros, bem como à estacionariedade da população de cheias históricas, afetada pelas possíveis variações climáticas e por ações antrópicas, como urbanização, desmatamentos e intervenções na calha fluvial.

Maiores detalhes a respeito da utilização de cheias históricas na análise de freqüência podem ser encontrados em Sutcliffe (1987) e Bayliss e Reed (2001).

3.2.3 O período contemporâneo

As informações do período contemporâneo correspondem às observações fluviométricas regulares, provenientes das estações de medição. Essas observações são geralmente expressas em termos de elevações do nível d'água, as quais são convertidas em vazões por meio das respectivas curvas-chave. Os dados do período contemporâneo estão sujeitos a erros de medição, que se devem principalmente à leitura, à transcrição ou ao processamento incorretos. Por isso, antes de serem utilizados na análise de freqüência, eles devem ser verificados com o intuito de identificar e eliminar erros, tendências, heterogeneidades e dependência serial, que podem estar presentes na amostra. Uma descrição dos principais testes e procedimentos a serem realizados durante a etapa de verificação dos dados contemporâneos é apresentada por Cândido (2003) e Naghettini e Pinto (no prelo).

A Figura 3.1 mostra os diferentes tipos de informações relativas às cheias, de acordo com a classificação cronológica dos eventos descrita anteriormente.



PERÍODO	PRÉ-HISTÓRICO	HISTÓRICO	CONTEMPORÂNEO
Técnica de coleta de dados	Paleohidrologia (paleovazões)	Observações ocasionais documentadas ou relatadas (cheias históricas)	Estações de medição (cotas e vazões)
Tipos de dados	Amostras censuradas, com eventos de intensidade conhecida ou não		Amostras não censuradas
Incertezas	Relativas à data e à intensidade	Relativas à intensidade	

FIGURA 3.1: Classificação cronológica das informações relativas às cheias.
Fonte: adaptado de NAULET, 2002.

3.2.4 Classificação dos dados

3.2.4.1 Dados sistemáticos

Os *dados sistemáticos* representam as informações do período contemporâneo, ou seja, uma seqüência de observações fluviométricas contínuas, realizadas em estações de medição, durante um período de N_S anos. Esses dados serão denotados pela variável aleatória *X*, que representa a vazão máxima observada em cada ano hidrológico, cujo conjunto constitui uma série de vazões máximas anuais. As séries de duração parcial também são constituídas pelos dados sistemáticos, porém, como citado anteriormente, estão fora do escopo dessa dissertação.

3.2.4.2 Dados não sistemáticos

Os *dados não sistemáticos* representam as informações dos períodos pré-histórico e histórico, compreendendo, respectivamente, as paleovazões, identificadas por meio de evidências físicas no ambiente natural, e as cheias históricas, documentadas de forma não sistemática por seres humanos. A duração, em anos, do período não sistemático será denotada por N_H, e a variável aleatória *Y* será usada para representar os eventos desse período.

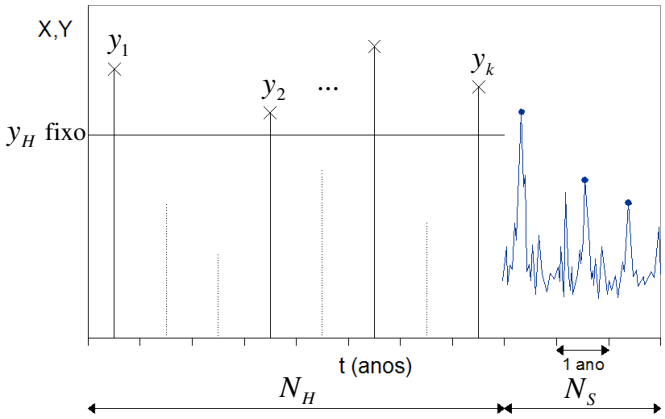
Uma vez que a paleohidrologia requer investigações geológicas e botânicas avançadas, bem como a utilização de modernas técnicas de datação, a identificação de paleovazões poderia se tornar um entrave financeiro e tecnológico para essa pesquisa. Portanto, para a realização do estudo de caso apresentado no Capítulo 5, optou-se por explorar apenas as cheias históricas. No entanto, ainda que nos itens 3.5 e 3.6 sejam usadas as expressões "período histórico" e "cheias históricas", os procedimentos descritos podem ser estendidos a um período não sistemático composto por informações pré-históricas e históricas, ou seja, a uma situação em que paleovazões também estejam disponíveis.

3.2.4.3 Tratamento estatístico dos dados não sistemáticos

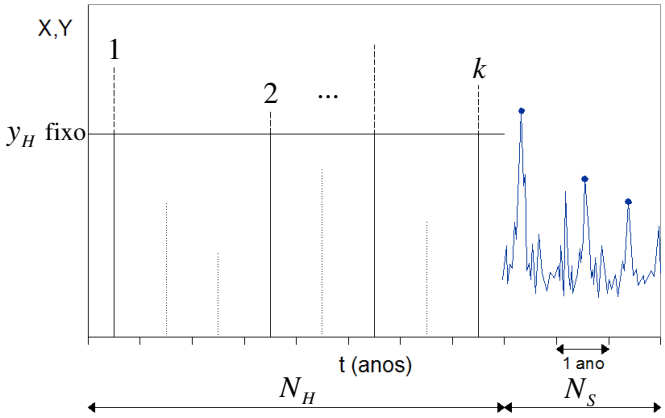
Do ponto de vista estatístico, as informações paleohidrológicas e históricas são tratadas da mesma maneira. Elas permitem construir amostras truncadas ou censuradas que, supondo a existência de um único limiar de referência y_H, apresentam as seguintes características:

1) *k* cheias de intensidade conhecida y_j, superiores ao limiar y_H fixo, constituindo uma amostra *censurada* ou *truncada tipo I* (Figura 3.2a);
2) *k* excedências do limiar y_H fixo, cujas intensidade não são conhecidas, constituindo uma amostra *binomial censurada* ou *truncada tipo I* (Figura 3.2b);
3) um número fixo *k* de cheias superiores ao limiar y_H, agora variável, $y_1 \geq y_2 \geq ... \geq y_k$, constituindo uma amostra *truncada tipo II* (o valor do limiar passa a depender do número de eventos acima dele: $y_H = y_k$, conforme Figura 3.2c).

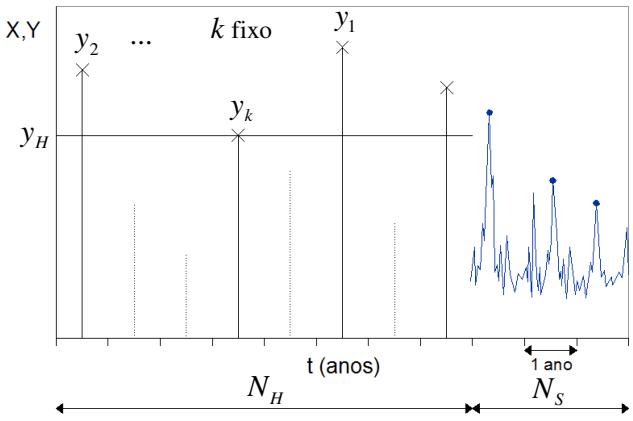
Maiores detalhes referentes ao tratamento estatístico dos dados não sistemáticos, de maneira a incorporá-los na análise de freqüência de vazões máximas anuais, são apresentados no item 3.5. Os aspectos referentes à utilização de uma amostra truncada tipo II não serão abordados nessa dissertação.



(a) Truncada tipo I



(b) Truncada tipo I



(c) Truncada tipo II

FIGURA 3.2: Diferentes tipos de amostras truncadas.
Fonte: adaptado de NAULET, 2002.

3.3 Prescrição do modelo de distribuição de probabilidades

Uma das etapas mais importantes da análise de freqüência de vazões máximas anuais é a escolha da distribuição de probabilidades que melhor representa o verdadeiro comportamento populacional da variável hidrológica. Embora Cândido (2003) afirme que não existem leis dedutivas para a seleção de uma distribuição de probabilidades particular ou de uma família de distribuições, alguns fatores subjetivos e testes estatísticos podem ser considerados para a prescrição do modelo distributivo mais adequado. A esse respeito, Hosking e Wallis (1997) e Davis e Naghettini (2001) apresentam as seguintes ponderações:

1) *Limite superior*:

   Mesmo que, sob certas condições, se possa considerar fisicamente impossível a ocorrência de valores extremamente elevados para uma variável aleatória (digamos, por exemplo, uma vazão de 100.000 m³/s em uma bacia hidrográfica com área de drenagem equivalente a poucas centenas de quilômetros quadrados), impor um limite superior ao modelo probabilístico pode comprometer a obtenção de boas estimativas de quantis para os tempos de retorno que realmente interessam. Ao se empregar uma distribuição ilimitada superiormente, as premissas implícitas são: (1) que o limite superior não é conhecido e nem pode ser estimado com a precisão necessária, e (2) que no intervalo de tempos de retorno de interesse do estudo, a distribuição de probabilidades da população pode ser melhor aproximada por uma função ilimitada do que por uma que possua limite superior. Certamente, quando existem evidências empíricas de que a distribuição populacional possui um limite superior, ela deve ser aproximada por uma distribuição limitada superiormente. Seria o caso, por exemplo, do ajuste da distribuição generalizada de valores extremos a uma certa amostra, cuja tendência de possuir um limite superior estaria refletida na estimativa de um valor positivo para o parâmetro de forma k.

2) *Cauda superior*:

   O "peso" da cauda superior de uma distribuição de probabilidades determina a intensidade com que os quantis aumentam, à medida que os tempos de retorno tendem para valores muito elevados. Em outras palavras, o peso da cauda superior reflete a intensidade com que a função densidade de probabilidade decresce na região de valores extremos da variável aleatória. Os pesos das caudas superiores de algumas das principais distribuições de probabilidades são mostrados na Tabela 3.1.

Embora a correta prescrição da cauda superior do modelo distributivo tenha importância fundamental na análise de freqüência de cheias, os tamanhos das amostras de dados hidrológicos são invariavelmente insuficientes para se determinar tal característica com exatidão. Assim, é aconselhável utilizar um grande conjunto de distribuições candidatas, cujos pesos das respectivas caudas superiores se estendam por uma ampla faixa.

TABELA 3.1: Peso da cauda superior de algumas distribuições de probabilidades.

Cauda superior	Forma de $f_X(\,)$ para valores elevados de x	Distribuição
Pesada	x^{-a}	GEV, GPA e GLO com parâmetro de forma $k<0$.
↑	$x^{-a\ln(x)}$	LNO com assimetria positiva.
	$\exp(-x^a) \quad 0<a<1$	WEI com parâmetro de forma $\lambda<1$.
	$x^a \exp(-bx)$	PE3 com assimetria positiva.
	$\exp(-x)$	EXP e GUM.
↓	$\exp(-x^a) \quad a>1$	WEI com parâmetro de forma $\lambda>1$.
Leve	Limite superior	GEV, GPA e GLO com parâmetro de forma $k>0$; LNO e PE3 com assimetria negativa.

Fonte: adaptado de HOSKING e WALLIS, 1997.
NOTAS:
1) *a* e *b* representam constantes positivas.
2) GEV: Generalizada de Valores Extremos, GPA: Generalizada de Pareto, GLO: Logística Generalizada, LNO: Log-Normal, WEI: Weibull, PE3: Pearson III, EXP: Exponencial, GUM: Gumbel.

3) *Cauda inferior e limite inferior*:

Considerações semelhantes àquelas feitas anteriormente se aplicam à cauda e ao limite inferior da distribuição. Porém, se o interesse está centrado em se prescrever a melhor aproximação da cauda superior, a forma da cauda inferior é irrelevante. A presença de *outliers* baixos em uma dada amostra pode, inclusive, vir a comprometer a correta estimação das características da cauda superior (NRC, 1987). Com relação ao limite inferior, seu valor pode, em geral, ser conhecido ou igualado a zero, e o parâmetro de posição do modelo distributivo pode ser ajustado de modo a permitir essa imposição. Entretanto, Hosking e Wallis (1997) ressaltam que, em diversos casos, a prescrição de um limite inferior nulo é inútil, e que melhores resultados podem ser obtidos sem nenhuma prescrição *a priori*.

Em geral, a seleção da "melhor" distribuição de probabilidades se baseia na qualidade e consistência de seu ajuste aos dados disponíveis. Nesse sentido, diversos testes de aderência podem ser empregados, sendo o do Qui-Quadrado, o de Kolmogorov-Smirnov, o de Filliben e o de comparação entre quocientes de momentos-L os mais conhecidos. Embora os dois

primeiros sejam mais comuns, alguns autores recomendam o uso dos dois últimos (STEDINGER *et al.*, 1993; HOSKING e WALLIS, 1997; CÂNDIDO, 2003; NAGHETTINI e PINTO, no prelo). Vale ressaltar que, especialmente no caso das amostras tipicamente curtas encontradas em hidrologia, os testes de aderência não são suficientemente potentes e conclusivos na seleção da distribuição de probabilidades a ser utilizada. As características dos principais testes de aderência, bem como suas aplicações e limitações, podem ser encontradas em Naghettini e Pinto (no prelo).

Cândido (2003) desenvolveu um sistema especialista que, utilizando a técnica de inteligência artificial, é capaz de selecionar, dentre algumas distribuições candidatas, aquela(s) mais apropriada(s) para modelar uma população de eventos hidrológicos máximos anuais, da qual foi extraída uma amostra particular.

3.4 Estimação dos parâmetros utilizando apenas dados sistemáticos

Nesse item, apresenta-se uma descrição sucinta dos principais métodos existentes para a estimação dos parâmetros da distribuição de probabilidades escolhida para representar o comportamento populacional das vazões máximas anuais, utilizando-se apenas os dados sistemáticos, ou seja, as observações fluviométricas regulares.

O Anexo A apresenta as principais propriedades das distribuições de probabilidades mais utilizadas para modelar eventos máximos de variáveis hidrológicas, bem como seus estimadores paramétricos, calculados pelos três métodos descritos a seguir.

3.4.1 Método dos momentos (MMO)

Os momentos populacionais de ordem *r*, em relação à origem, de uma variável aleatória *X*, descrita pela distribuição de probabilidades $f_X(x \mid \Theta)$, cujo vetor de parâmetros é Θ, são definidos por:

$$\mu'_r = E[X^r] = \int_{-\infty}^{\infty} x^r f_X(x \mid \Theta)\, dx \tag{3.1}$$

O momento de ordem 1, em relação à origem, é igual ao valor esperado de *X*, e representa a média populacional, ou seja:

$$\mu'_1 = \mu = E[X] \tag{3.2}$$

Analogamente, os momentos populacionais de ordem *r*, em relação à média, ou simplesmente momentos centrais, são definidos por:

$$\mu_r = E[(X-\mu)^r] = \int_{-\infty}^{\infty} (x-\mu)^r f_X(x \mid \Theta)\, dx \qquad r \geq 2 \tag{3.3}$$

Os momentos centrais de ordem 2, 3 e 4 são particularmente importantes em hidrologia estatística. O momento central de ordem 2 representa a variância populacional de *X*, dada pela equação (3.4), e constitui uma medida de dispersão da variável aleatória em relação à média. Outras grandezas capazes de representar a dispersão da variável *X* em torno da média são o desvio-padrão e o coeficiente de variação, dados respectivamente pelas equações (3.5) e (3.6).

$$\mu_2 = \sigma^2 = E[(X-\mu)^2] = E[X^2] - \mu^2 \tag{3.4}$$

$$\sigma = \sqrt{\sigma^2} \tag{3.5}$$

$$C_V = \frac{\sigma}{\mu} \tag{3.6}$$

A partir dos momentos centrais de ordem 3 e 4, são definidos os coeficientes de assimetria e de curtose, expressos respectivamente pelas equações (3.7) e (3.8).

$$\gamma = \frac{\mu_3}{\sigma^3} \tag{3.7}$$

$$\kappa = \frac{\mu_4}{\sigma^4} \tag{3.8}$$

O método dos momentos para estimação dos parâmetros de uma distribuição de probabilidades consiste em igualar os momentos amostrais, calculados a partir de uma amostra de tamanho *n*, aos correspondentes momentos populacionais. Sejam m_r e μ_r, respectivamente, os momentos amostrais e populacionais de ordem *r*, e $\Theta = (\theta_1, \theta_2, ..., \theta_k)$ o vetor dos *k* parâmetros desconhecidos. O sistema de equações fundamental do método dos momentos é:

$$m_r = \mu_r(\Theta) \qquad r = 1,2,...,k \tag{3.9}$$

A solução desse sistema de *k* equações e *k* incógnitas resulta no vetor $\hat{\Theta} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k)$, cujos elementos são os estimadores paramétricos pelo método dos momentos. Rao e Hamed (2000) apresentam uma exposição detalhada das estimativas pelo método dos momentos para as principais distribuições de probabilidades empregadas em hidrologia.

3.4.2 Método do máximo de verossimilhança (MVS)

O método do máximo de verossimilhança consiste em maximizar uma função dos parâmetros da distribuição de probabilidades. O equacionamento para a condição de máximo resulta em um sistema de igual número de equações e incógnitas, cujas soluções produzem os estimadores paramétricos de máxima verossimilhança.

Seja $(x_1, x_2, ..., x_n)$ uma amostra retirada da população da variável hidrológica de interesse, modelada pela seguinte distribuição de probabilidades, a qual depende do vetor de parâmetros $\Theta = (\theta_1, \theta_2, ..., \theta_k)$:

$$f_X(x \mid \Theta) \tag{3.10}$$

Como os elementos constituintes da amostra são, por pressuposto, independentes e identicamente distribuídos, eles possuem uma distribuição de probabilidades conjunta dada por:

$$f_X(x_1, x_2, ..., x_n \mid \Theta) = f_X(x_1 \mid \Theta) \cdot f_X(x_2 \mid \Theta) \cdot ... \cdot f_X(x_n \mid \Theta) \tag{3.11}$$

Essa densidade conjunta, conhecida como função de verossimilhança, é proporcional à probabilidade de que a amostra tenha sido extraída da população distribuída conforme a equação (3.10). Portanto, a função de verossimilhança é formalmente expressa por:

$$L(\Theta \mid x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f_X(x_i \mid \Theta) \tag{3.12}$$

Essa equação é uma função do vetor de parâmetros Θ, exclusivamente. Os valores dos parâmetros que maximizam a função de verossimilhança são aqueles que também maximizam

a probabilidade de que a amostra em questão tenha sido sorteada da população distribuída conforme a equação (3.10). A busca da condição de máximo para a função de verossimilhança resulta no seguinte sistema de *k* equações e *k* incógnitas:

$$\frac{\partial L(\Theta \mid x_1, x_2, ..., x_n)}{\partial \theta_j} = 0 \qquad j = 1,2,...k \tag{3.13}$$

A solução desse sistema produz os estimadores de máxima verossimilhança dos parâmetros da distribuição, denotados por $\hat{\theta}_j$. Uma vez que, em muitos casos, é mais simples maximizar o logaritmo de uma função, é freqüente a substituição da função de verossimilhança pela função logaritmo de verossimilhança, dada por:

$$\ln L(\Theta \mid x_1, x_2, ..., x_n) = \sum_{i=1}^{n} \ln f_X(x_i \mid \Theta) \tag{3.14}$$

Não há nenhuma restrição quanto à adoção desse procedimento, pois, como a função logaritmo é contínua, monótona e crescente, maximizar o logaritmo da função de verossimilhança é o mesmo que maximizar a própria função. Portanto, o sistema de equações (3.13) pode ser reescrito da seguinte forma:

$$\frac{\partial \ln L(\Theta \mid x_1, x_2, ..., x_n)}{\partial \theta_j} = \frac{1}{L} \cdot \frac{\partial L}{\partial \theta_j} = 0 \qquad j = 1,2,...k \tag{3.15}$$

Os estimadores de máxima verossimilhança são, geralmente, superiores àqueles obtidos pelo método dos momentos, especialmente para distribuições com mais de dois parâmetros, já que os momentos de ordem superior estão sujeitos a um elevado viés quando calculados utilizando amostras de pequeno tamanho. Além disso, em termos assintóticos, o método do máximo de verossimilhança é considerado o mais eficiente, pois provê estimadores paramétricos com menor variância. Em alguns casos, a solução dos sistemas de equações (3.13) e (3.15) pode ser bastante complexa, sendo necessária a implementação de procedimentos computacionais para sua resolução numérica. No entanto, com a utilização crescente de computadores de alto desempenho, essa desvantagem não constitui mais um problema significativo. Maiores detalhes a respeito do método do máximo de verossimilhança para estimação dos parâmetros das principais distribuições de probabilidades usadas em hidrologia podem ser encontrados em Clarke (1994) e Rao e Hamed (2000).

3.4.3 Método dos momentos-L (MML)

Greenwood *et al.* (1979) introduziram os momentos ponderados por probabilidades (MPP's), dos quais são deduzidos os momentos-L. Os MPP's de uma variável aleatória *X*, descrita pela função de probabilidades acumulada $F_X(x \mid \Theta) = P[X < x]$, cujo vetor de parâmetros é Θ, são definidos por:

$$M_{p,r,s} = E[X^p F^r (1-F)^s] = \int_0^1 [x(F)]^p F^r (1-F)^s dF \tag{3.16}$$

onde $x(F)$ é a função de quantis, e *p*, *r* e *s* representam números reais. Quando *r* e *s* são nulos e *p* é um número não negativo, os MPP's $M_{p,0,0}$ são iguais aos momentos convencionais μ'_p, de ordem *p*, em relação à origem. Os MPP's $M_{1,r,0}$ são os de utilização mais freqüente na caracterização das distribuições de probabilidades. Eles são definidos por:

$$M_{1,r,0} = \beta_r = E[XF^r] = \int_0^1 x(F) F^r dF \tag{3.17}$$

Hosking (1986) demonstrou que os MPP's β_r, que constituem combinações lineares de *X*, possuem a generalidade suficiente para a estimação dos parâmetros das distribuições de probabilidades, além de estarem menos sujeitos a flutuações amostrais e, portanto, serem mais robustos que os correspondentes momentos convencionais. Para uma amostra de tamanho *n*, ordenada de modo crescente, $x_1 \le x_2 \le ... \le x_n$, as estimativas não-enviesadas de β_r podem ser calculadas pela seguinte expressão:

$$b_r = \hat{\beta}_r = \frac{1}{n} \sum_{i=1}^{n} \frac{\binom{i-1}{r}}{\binom{n-1}{r}} x_i = \frac{1}{n} \sum_{i=1}^{n} \frac{(i-1)!}{(i-r-1)!} \cdot \frac{(n-r-1)!}{(n-1)!} \cdot x_i \tag{3.18}$$

Embora possam ser usados na estimação dos parâmetros, os MPP's β_r não são de fácil interpretação como descritores de escala e forma das distribuições de probabilidades. Alternativamente, Hosking (1990) introduziu o conceito dos momentos-L, que constituem grandezas diretamente interpretáveis, obtidas por meio de combinações lineares de β_r.

Os momentos-L de ordem *r*, denotados por λ_r, são formalmente definidos por:

$$\lambda_r = \sum_{k=0}^{r-1} p^*_{r-1,k} \beta_k \tag{3.19}$$

onde:

$$p^*_{r-1,k} = (-1)^{r-k-1} \binom{r-1}{k} \binom{r+k-1}{k} = \frac{(-1)^{r-k-1}(r+k-1)!}{(k!)^2 (r-k-1)!} \tag{3.20}$$

A aplicação das equações (3.19) e (3.20) para os quatro primeiros momentos-L resulta nas seguintes expressões:

$$\lambda_1 = \beta_0 \tag{3.21}$$

$$\lambda_2 = 2\beta_1 - \beta_0 \tag{3.22}$$

$$\lambda_3 = 6\beta_2 - 6\beta_1 + \beta_0 \tag{3.23}$$

$$\lambda_4 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 \tag{3.24}$$

Os momentos-L amostrais, denotados por l_r, são calculados pela substituição de β_r, nas equações (3.21) a (3.24), por suas estimativas b_r.

O momento-L λ_1 é equivalente à média μ e, portanto, constitui uma medida populacional de posição. Para ordens superiores a 1, os quocientes de momentos-L são particularmente úteis na descrição de escala e forma das distribuições de probabilidades. Como medida equivalente ao coeficiente de variação convencional, define-se o coeficiente τ, dado por:

$$\tau = \frac{\lambda_2}{\lambda_1} \tag{3.25}$$

Essa grandeza pode ser interpretada como uma medida populacional de dispersão ou de escala. Analogamente aos coeficientes de assimetria e curtose convencionais, podem ser definidos os coeficientes τ_3 e τ_4, dados por:

$$\tau_3 = \frac{\lambda_3}{\lambda_2} \tag{3.26}$$

$$\tau_4 = \frac{\lambda_4}{\lambda_2} \tag{3.27}$$

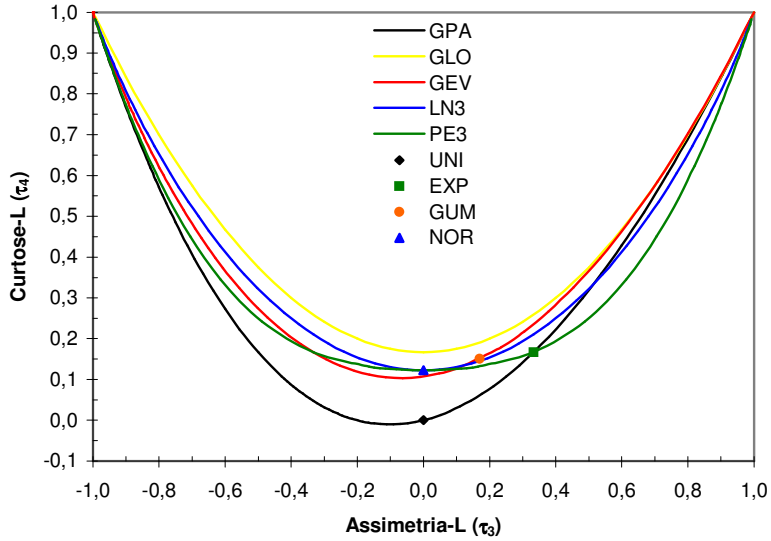
Os quocientes de momentos-L amostrais, cujas notações são t, t_3 e t_4, são calculados pela substituição de λ_r, nas equações (3.25) a (3.27), por suas estimativas l_r.

De acordo com Naghettini e Pinto (no prelo), os momentos-L apresentam diversas vantagens em relação aos momentos convencionais, entre as quais destacam-se os limites de variação de t, t_3 e t_4. De fato, se *X* é uma variável aleatória não negativa, pode-se demonstrar que $0 < \tau < 1$. Quanto a τ_3 e τ_4, é um fato matemático que esses coeficientes estão compreendidos no intervalo $[-1,+1]$, em oposição aos correspondentes coeficientes convencionais, que podem assumir valores arbitrariamente mais elevados. Outras vantagens dos momentos-L em relação aos momentos convencionais são discutidas por Vogel e Fennessey (1993).

Uma maneira conveniente de representar os momentos-L das diversas distribuições de probabilidades é o diagrama de quocientes de momentos-L, mostrado na Figura 3.3. Nesse diagrama, uma distribuição de dois parâmetros (posição e escala) é plotada como um único ponto, em decorrência do fato de que duas distribuições que diferem entre si apenas pelos valores dos parâmetros de posição e escala são as distribuições das variáveis aleatórias *X* e $Y = aX + b$, com $a > 0$, cujos quocientes de momentos-L, denotados respectivamente por τ_r^X e τ_r^Y, se relacionam conforme a equação (3.28), e, portanto, são iguais.

$$\tau_r^Y = (\langle \text{sinal de a} \rangle 1)^r \tau_r^X \qquad r \geq 3 \tag{3.28}$$

Já as distribuições de três parâmetros (posição, escala e forma) são plotadas como curvas, cujos pontos correspondem aos diferentes valores do parâmetro de forma. Plotando-se no diagrama de quocientes de momentos-L os coeficientes t_3 (assimetria-L) e t_4 (curtose-L), calculados a partir de uma amostra particular, é possível identificar qual distribuição melhor representa o comportamento populacional da variável hidrológica de interesse.



GPA: Generalizada de Pareto, GLO: Logística Generalizada, GEV: Generalizada de Valores Extremos, LN3: Log-Normal 3 Parâmetros, PE3: Pearson III, UNI: Uniforme, EXP: Exponencial, GUM: Gumbel, NOR: Normal

FIGURA 3.3: Diagrama de quocientes de momentos-L.
Fonte: adaptado de HOSKING e WALLIS, 1997.

O método dos momentos-L para estimação dos parâmetros das distribuições de probabilidades é semelhante ao método dos momentos convencionais, e consiste em igualar os momentos-L amostrais, calculados a partir de uma amostra de tamanho *n*, aos correspondentes momentos-L populacionais. De fato, os momentos-L e seus quocientes, a saber λ_1, λ_2, τ, τ_3 e τ_4, podem ser expressos como funções dos parâmetros das distribuições de probabilidades. Se $(\lambda_1, \lambda_2, \tau_j)$ e (l_1, l_2, t_j) representam, respectivamente, os momentos-L (e seus quocientes) populacionais e amostrais, e $\Theta = (\theta_1, \theta_2, ..., \theta_k)$ representa o vetor dos k parâmetros desconhecidos, o sistema de equações fundamental do método dos momentos-L é:

$$\begin{aligned} l_i &= \lambda_i(\Theta) \qquad i = 1,2 \\ t_j &= \tau_j(\Theta) \qquad j = 3,4,...,k \end{aligned} \tag{3.29}$$

A solução desse sistema de k equações e k incógnitas resulta no vetor $\hat{\Theta} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k)$, cujos elementos são os estimadores paramétricos pelo método dos momentos-L. Hosking e Wallis (1997) mostram que, para amostras de tamanho pequeno a moderado, o método dos momentos-L é geralmente mais eficiente que o do máximo de verossimilhança.

3.5 Estimação dos parâmetros utilizando dados sistemáticos e não sistemáticos

Conforme descrito por Naulet (2002), para um modelo de vazões máximas anuais, a informação hidrológica disponível pode ser representada pelo esquema geral mostrado na Figura 3.4, observando-se três casos distintos:

1) *Informação não censurada*:

   A intensidade das vazões máximas anuais é conhecida, e os eventos são denotados por x_i, com $i = 1$ a $N_S^{\bullet}$, para cheias do período sistemático, e y_j, com $j = 1$ a $N_H^{\bullet}$, para cheias do período histórico (o ponto sobrescrito indica que a intensidade da cheia é conhecida).

2) *Informação censurada*:

   A intensidade das vazões máximas anuais não é conhecida precisamente, mas dispõe-se de elementos suficientes para considerá-la:

   a) inferior a um limiar de referência:

      $N_S^{<}$ eventos de magnitude inferior a x_{Ui} e $N_H^{<}$ eventos de magnitude inferior a y_{Uj};

   b) superior a um limiar de referência:

      $N_S^{>}$ eventos de magnitude superior a x_{Li} e $N_H^{>}$ eventos de magnitude superior a y_{Lj};

   c) compreendida em um intervalo:

      $N_S^{<>}$ eventos compreendidos no intervalo $[x_{Li};x_{Ui}]$ e $N_H^{<>}$ eventos compreendidos no intervalo $[y_{Lj};y_{Uj}]$.

3) *Ausência de informação*:

   Os anos em que não se dispõe de informações suficientes para inferir sobre a magnitude das vazões máximas anuais (representados pelo símbolo "?" na Figura 3.4) não devem ser considerados na análise de freqüência.

A amostra total de vazões máximas anuais será, então, constituída de N anos, distribuídos conforme mostrado na equação (3.30):

$$\begin{aligned} N &= N_S + N_H \\ N &= N_S^{\bullet} + N_S^{<} + N_S^{>} + N_S^{<>} + N_H^{\bullet} + N_H^{<} + N_H^{>} + N_H^{<>} \end{aligned} \tag{3.30}$$

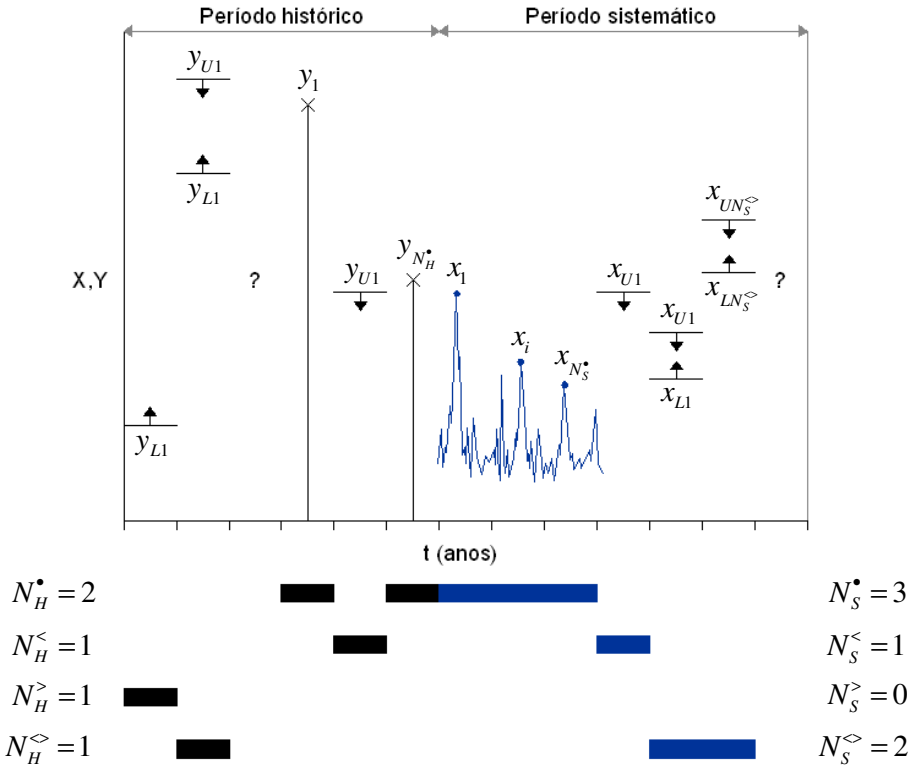


FIGURA 3.4: Esquema dos diversos tipos de informações relativas às cheias, para um modelo de vazões máximas anuais.
Fonte: adaptado de NAULET, 2002.

Os dados dos períodos sistemático e histórico constituem, dessa forma, uma única amostra truncada. Deve-se ressaltar que os métodos de incorporação de informação histórica na análise de freqüência pressupõem que as cheias sistemáticas (representadas pela variável aleatória *X*) e não sistemáticas (representadas pela variável aleatória *Y*) sejam identicamente distribuídas, isto é, que elas sejam descritas pela mesma função densidade de probabilidade $f_X(\,)$, cuja função acumulada de probabilidade é denotada por $F_X(\,)$.

Na seqüência, serão apresentados os principais métodos existentes para a incorporação dos dados não sistemáticos na estimação dos parâmetros da distribuição de probabilidades escolhida para representar o comportamento populacional das vazões máximas anuais. Para isso, serão adotados os critérios e notações utilizados por Naulet (2002), considerados adequadamente didáticos.

3.5.1 Método dos momentos ponderados historicamente (MPH)

O método dos momentos ponderados historicamente, também denominado método dos momentos ajustados, foi apresentado pelo USWRC (1982), na edição revisada do Boletim 17 (Boletim 17B). Assim como o método clássico dos momentos (item 3.4.1), o método MPH consiste em igualar os momentos populacionais aos correspondentes momentos amostrais, calculados por meio da atribuição de "pesos" distintos aos elementos constituintes da amostra de vazões máximas anuais. Esse procedimento resulta em um sistema de equações, cuja solução fornece os estimadores $\hat{\Theta}$ dos parâmetros da distribuição.

Nesse método, não são permitidos eventos censurados no período das cheias sistemáticas ($N_S^< = N_S^> = N_S^{<>} = 0$), nem cheias históricas de intensidade conhecida y_j, inferiores a um limiar de referência y_U. Além disso, a amostra é dividida em eventos de magnitude superior e inferior a um único limiar y_U, definido como a menor cheia histórica de intensidade conhecida. Dessa forma, eventos maiores que y_U, observados no período sistemático, são tratados como cheias históricas, e o período de observações sistemáticas fica reduzido aos eventos menores que y_U. A Figura 3.5 mostra o esquema de interpretação das cheias históricas e sistemáticas, para aplicação do método dos momentos ponderados historicamente.

Kirby (1981) afirma que todo ano cuja vazão máxima excede o limiar y_U pode pertencer ao período histórico ou sistemático, e que o ajuste previsto no cálculo dos momentos amostrais visa, de fato, preencher a porção do período histórico cujas magnitudes dos eventos são desconhecidas (e, por pressuposto, inferiores a y_U) com um número apropriado de cópias da porção inferior a y_U pertencente ao período sistemático. Esse preenchimento é alcançado aplicando-se um fator de ponderação (W) às cheias sistemáticas inferiores ao limiar de referência y_U, definido por:

$$W = \frac{N - N^{\bullet >}}{N_S^{\bullet <}} \geq 1 \tag{3.31}$$

Considerando o exemplo da Figura 3.5, tem-se: $N = 8$, $N^{\bullet >} = 3$ e $N_S^{\bullet <} = 2$. Assim, o fator de ponderação W será:

$$W = \frac{8 - 3}{2} = 2{,}5$$

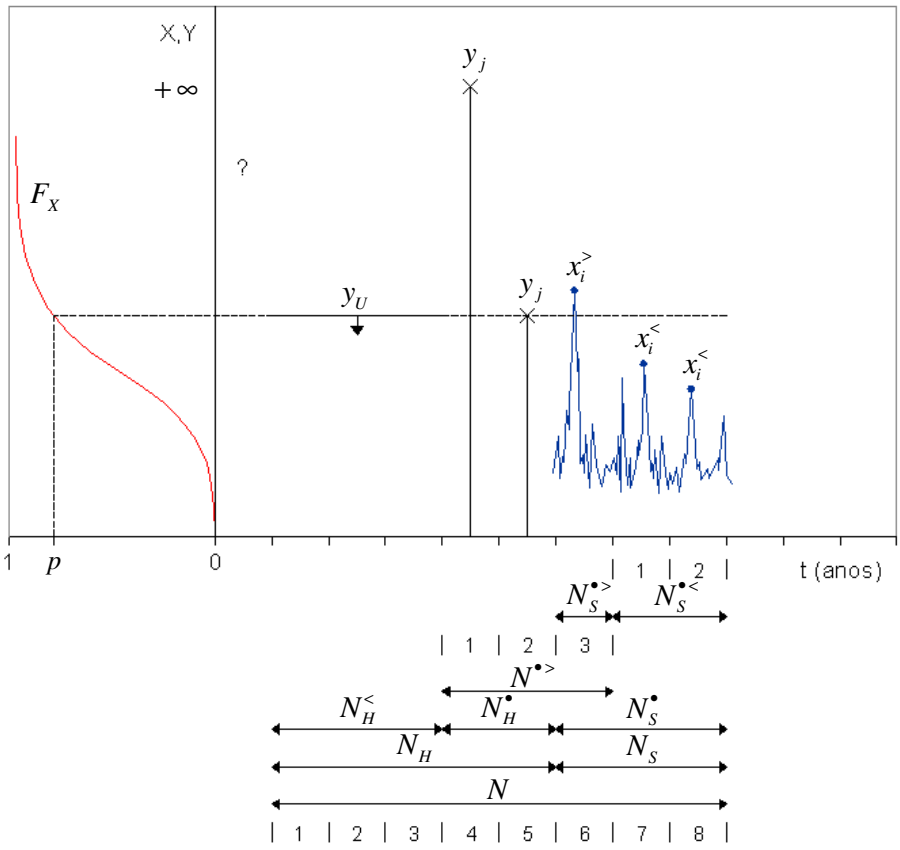


FIGURA 3.5: Esquema de aplicação do método dos momentos ponderados historicamente.
Fonte: adaptado de NAULET, 2002.

Portanto, para o cálculo dos momentos amostrais pelo método MPH, utiliza-se:

- um fator de ponderação W para os $N_S^{\bullet <}$ eventos máximos anuais $x_i^<$ do período sistemático para representar artificialmente os $N_H^<$ anos truncados do período histórico;
- um fator de ponderação igual a 1 para os $N^{\bullet >} = N_H^\bullet + N_S^{\bullet >}$ eventos maiores que o limiar de referência y_U.

Se p denota a probabilidade de não-excedência de y_U, tal que $P[X < y_U] = P[Y < y_U] = p$, então, dado que $N^{\bullet >}/N$ representa a probabilidade empírica de excedência de y_U, p pode ser estimado por $\hat{p} = 1 - (N^{\bullet >}/N)$. Calculam-se, então, as médias das cheias de intensidade conhecida:

- superiores ao limiar y_U (período histórico e sistemático):

$$m^{\bullet >} = \frac{1}{N^{\bullet >}} \left(\sum_{i=1}^{N_S^{\bullet >}} x_i^{>} + \sum_{j=1}^{N_H^{\bullet}} y_j \right) \tag{3.32}$$

- inferiores ao limiar y_U (período sistemático):

$$m^{\bullet <} = \frac{1}{N_S^{\bullet <}} \left(\sum_{i=1}^{N_S^{\bullet <}} x_i^{<} \right) \tag{3.33}$$

A média empírica da amostra global, denominada média ponderada historicamente e denotada por $\tilde{m}$, é obtida ponderando-se as médias calculadas anteriormente pelas respectivas probabilidades de ocorrência:

$$\tilde{m} = \hat{p} \cdot m^{\bullet <} + (1 - \hat{p}) \cdot m^{\bullet >}$$

$$\tilde{m} = \frac{1}{N} \left(W \sum_{i=1}^{N_S^{\bullet <}} x_i^{<} + \sum_{i=1}^{N_S^{\bullet >}} x_i^{>} + \sum_{j=1}^{N_H^{\bullet}} y_j \right) \tag{3.34}$$

Os mesmos argumentos podem ser utilizados para a obtenção de momentos de ordem superior, tais como a variância e o coeficiente de assimetria ponderados historicamente, expressos respectivamente pelas equações (3.35) e (3.36):

$$\tilde{s}^2 = \frac{1}{(N-1)} \left[W \sum_{i=1}^{N_S^{\bullet <}} \left(x_i^{<} - \tilde{m} \right)^2 + \sum_{i=1}^{N_S^{\bullet >}} \left(x_i^{>} - \tilde{m} \right)^2 + \sum_{j=1}^{N_H^{\bullet}} \left(y_j - \tilde{m} \right)^2 \right] \tag{3.35}$$

$$\tilde{g} = \frac{N}{\tilde{s}^3 (N-1)(N-2)} \left[W \sum_{i=1}^{N_S^{\bullet <}} \left(x_i^{<} - \tilde{m} \right)^3 + \sum_{i=1}^{N_S^{\bullet >}} \left(x_i^{>} - \tilde{m} \right)^3 + \sum_{j=1}^{N_H^{\bullet}} \left(y_j - \tilde{m} \right)^3 \right] \tag{3.36}$$

Embora o Boletim 17B (USWRC, 1982) recomende a utilização do método dos momentos ponderados historicamente para a distribuição Log-Pearson III, estudos demonstram sua aplicação a outras distribuições, tais como a Log-Normal 2 parâmetros (STEDINGER e COHN, 1986), a Log-Normal 3 parâmetros (COHN e STEDINGER, 1987) e a Gumbel (GUO e CUNNANE, 1991).

3.5.2 Método do máximo de verossimilhança (MVS)

O método do máximo de verossimilhança para estimação dos parâmetros utilizando dados sistemáticos e não sistemáticos consiste em maximizar uma função dos parâmetros da distribuição, chamada função de verossimilhança. A diferença em relação ao procedimento apresentado no item 3.4.2 (método MVS para estimação dos parâmetros utilizando apenas dados sistemáticos) reside, portanto, no conjunto de informações utilizadas para a formulação da referida função. Diversos autores (HOSKING e WALLIS, 1986a, 1986b; STEDINGER e COHN, 1986; COHN e STEDINGER, 1987; SUTCLIFFE, 1987; GUO e CUNNANE, 1991; FRANCÉS *et al.*, 1994; KUCZERA, 1996; COHN *et al.*, 1997; MARTINS e STEDINGER, 2000, 2001a, 2001b; O'CONNELL *et al.*, 2002; NAULET, 2002) incorporaram os dados não sistemáticos na formulação da função de verossimilhança e demonstraram que esse método, além de apresentar a flexibilidade necessária para lidar com os diferentes tipos de informações relativas às cheias, produz estimadores paramétricos eficientes e relativamente robustos.

Considere, inicialmente, que a informação histórica disponível constitua um período de N_H anos, em que foram observadas:

- k cheias de intensidade conhecida $(y_1, y_2, ..., y_k)$, superiores a um limiar de referência fixo y_H (item 3.5.2.1); ou
- k excedências de um limiar de referência fixo y_H (item 3.5.2.2); ou
- k cheias de intensidade compreendida em um certo intervalo $[y_{Lm}; y_{Um}]$, com $m = 1,2,...,k$, superiores a um limiar de referência fixo y_H (item 3.5.2.3).

As formulações das funções de verossimilhança para essas três situações serão descritas na seqüência. Naulet (2002) apresentou um modelo geral que permite incorporar: (1) os três tipos de informações citados anteriormente (não só para o período histórico, mas também para o período sistemático) e (2) diferentes limiares de referência sucessivos ao longo dos anos. Essa abordagem geral será descrita no item 3.5.2.4. Os demais procedimentos do método MVS são os mesmos apresentados para o caso em que se utilizam apenas dados sistemáticos (item 3.4.2).

3.5.2.1 Cheias de intensidade conhecida, superior a um limiar fixo

Num período histórico de N_H anos, a amostra é constituída por *k* cheias de intensidade conhecida $(y_1, y_2, ..., y_k)$, as quais foram registradas exatamente por terem excedido um certo limiar de referência y_H, e $(N_H - k)$ cheias de intensidade desconhecida, porém inferiores a y_H (amostra *censurada*, ou truncada tipo I, conforme Figura 3.2a).

Sejam os eventos $A:[y_H \le Y < +\infty]$, $A^c:[0 < Y < y_H]$ e $B:[y \le Y < y + dy]$. Dada a condição que as cheias históricas e sistemáticas sejam descritas pela mesma função densidade de probabilidade $f_X(\,)$, as probabilidades de ocorrência dos eventos A e A^c são expressas respectivamente por:

$$P[y_H \le Y < +\infty] = \int_{y_H}^{+\infty} f_X(y \mid \Theta)\, dy = 1 - F_X(y_H \mid \Theta) \tag{3.37}$$

$$P[0 < Y < y_H] = \int_{0}^{y_H} f_X(y \mid \Theta)\, dy = F_X(y_H \mid \Theta) \tag{3.38}$$

A probabilidade de ocorrência do evento *B*, dado que *A* ocorreu, é expressa por:

$$P[(y \le Y < y + dy) \mid (y_H \le Y < +\infty)] = \frac{P[(y \le Y < y + dy) \cap (y_H \le Y < +\infty)]}{P[y_H \le Y < +\infty]} \tag{3.39}$$

onde $P[(y \le Y < y + dy) \cap (y_H \le Y < +\infty)] = P[y \le Y < y + dy]$, uma vez que *B* está contido em *A*. Assim:

$$P[(y \le Y < y + dy) \mid (y_H \le Y < +\infty)] = \frac{f_X(y \mid \Theta)}{1 - F_X(y_H \mid \Theta)}\, dy \tag{3.40}$$

Considerando que os elementos da amostra são independentes, a probabilidade P_C de se observar, nos N_H anos, exatamente *k* cheias de intensidade conhecida y_m, superiores a y_H, e $(N_H - k)$ cheias de intensidade desconhecida, inferiores a y_H, é expressa por:

$$P_C = \binom{N_H}{k} \cdot [1 - F_X(y_H \mid \Theta)]^k \cdot [F_X(y_H \mid \Theta)]^{N_H - k} \cdot \prod_{m=1}^{k} \frac{f_X(y_m \mid \Theta)}{1 - F_X(y_H \mid \Theta)} dy_m$$

$$P_C = \binom{N_H}{k} \cdot [F_X(y_H \mid \Theta)]^{N_H - k} \cdot \prod_{m=1}^{k} f_X(y_m \mid \Theta)\, dy_m \qquad (3.41)$$

Portanto, a função de verossimilhança L_C, que é proporcional à probabilidade de que a amostra descrita no parágrafo anterior tenha sido observada, é dada pela seguinte expressão:

$$L_C = \binom{N_H}{k} \cdot [F_X(y_H \mid \Theta)]^{N_H - k} \cdot \prod_{m=1}^{k} f_X(y_m \mid \Theta) \qquad (3.42)$$

3.5.2.2 Cheias de intensidade desconhecida, superior a um limiar fixo

Num período histórico de N_H anos, a amostra é constituída por *k* cheias superiores e $(N_H - k)$ cheias inferiores a um certo limiar de referência y_H. Todos os eventos são de intensidade desconhecida (amostra *binomial censurada*, ou truncada tipo I, conforme Figura 3.2b).

Considerando as propriedades dos eventos A e A^c, descritos no item anterior, e a independência entre os elementos da amostra, a probabilidade P_{BC} de se observar, nos N_H anos, exatamente *k* cheias superiores e $(N_H - k)$ cheias inferiores a y_H, é dada pela expressão (3.43), que corresponde a uma distribuição binomial:

$$P_{BC} = \binom{N_H}{k} \cdot [1 - F_X(y_H \mid \Theta)]^k \cdot [F_X(y_H \mid \Theta)]^{N_H - k} \qquad (3.43)$$

Portanto, a função de verossimilhança L_{BC} é dada por:

$$L_{BC} = \binom{N_H}{k} \cdot [1 - F_X(y_H \mid \Theta)]^k \cdot [F_X(y_H \mid \Theta)]^{N_H - k} \qquad (3.44)$$

3.5.2.3 Cheias de intensidade compreendida em um intervalo, superior a um limiar fixo

Num período histórico de N_H anos, a amostra é constituída por *k* cheias de intensidade compreendida em um intervalo $[y_{Lm}; y_{Um}]$, as quais foram registradas exatamente por terem

excedido um certo limiar de referência y_H, e $(N_H - k)$ cheias de intensidade desconhecida, porém inferiores a y_H (amostra *censurada em um intervalo*, ou truncada tipo I, conforme Figura 3.2a).

Sejam os eventos A e A^c, cujas propriedades foram descritas anteriormente, e o evento $C : [y_L \le Y < y_U]$. A probabilidade de ocorrência do evento *C*, dado que *A* ocorreu, é expressa por:

$$P[(y_L \le Y < y_U) \mid (y_H \le Y < +\infty)] = \frac{P[(y_L \le Y < y_U) \cap (y_H \le Y < +\infty)]}{P[y_H \le Y < +\infty]} \tag{3.45}$$

onde $y_L \ge y_H$ e $P[(y_L \le Y < y_U) \cap (y_H \le Y < +\infty)] = P[y_L \le Y < y_U]$, uma vez que C está contido em *A*. Assim:

$$P[(y_L \le Y < y_U) \mid (y_H \le Y < +\infty)] = \frac{F_X(y_U \mid \Theta) - F_X(y_L \mid \Theta)}{1 - F_X(y_H \mid \Theta)} \tag{3.46}$$

Considerando que os elementos da amostra são independentes, a probabilidade P_{CI} de se observar, nos N_H anos, exatamente *k* cheias de intensidade compreendida em um intervalo $[y_{Lm}; y_{Um}]$, cujo limite inferior excede y_H, e $(N_H - k)$ cheias de intensidade desconhecida, inferiores a y_H, é expressa por:

$$P_{CI} = \binom{N_H}{k} \cdot [1 - F_X(y_H \mid \Theta)]^k \cdot [F_X(y_H \mid \Theta)]^{N_H - k} \cdot \prod_{m=1}^{k} \frac{F_X(y_{Um} \mid \Theta) - F_X(y_{Lm} \mid \Theta)}{1 - F_X(y_H \mid \Theta)}$$

$$P_{CI} = \binom{N_H}{k} \cdot [F_X(y_H \mid \Theta)]^{N_H - k} \cdot \prod_{m=1}^{k} [F_X(y_{Um} \mid \Theta) - F_X(y_{Lm} \mid \Theta)] \tag{3.47}$$

Portanto, a função de verossimilhança L_{CI} é dada por:

$$L_{CI} = \binom{N_H}{k} \cdot [F_X(y_H \mid \Theta)]^{N_H - k} \cdot \prod_{m=1}^{k} [F_X(y_{Um} \mid \Theta) - F_X(y_{Lm} \mid \Theta)] \tag{3.48}$$

3.5.2.4 Generalização

Os casos particulares apresentados nos três itens anteriores podem ser generalizados, de forma a levar em conta a existência de diferentes limiares y_{Hj} e diferentes intervalos $[y_{Lm,j}; y_{Um,j}]$, associados a t períodos históricos de N_{Hj} anos $(j = 1,2,...,t)$, tal que k_j eventos tenham igualado ou excedido y_{Hj} e $\sum_{j=1}^{t} N_{Hj} = N_H$. Nessas condições, as funções de verossimilhança são dadas por:

1) *Cheias de intensidade conhecida, superior a um limiar fixo*:

$$L_C = \prod_{j=1}^{t} \left\{ \binom{N_{Hj}}{k_j} \cdot [F_X(y_{Hj} \mid \Theta)]^{N_{Hj}-k_j} \cdot \prod_{m=1}^{k_j} f_X(y_{m,j} \mid \Theta) \right\} \tag{3.49}$$

Se considerarmos vários períodos sucessivos de um ano $(N_{Hj} = 1)$, com $k_j = 1$ ou 0 caso o evento do ano *j* seja, respectivamente, superior ou inferior ao limiar y_{Hj}, o período histórico de N_H anos será composto por:

- $N_H^{\bullet}$ anos de cheias de intensidade conhecida y_j, superior ao limiar;
- $N_H^{<}$ anos de cheias de intensidade desconhecida, inferior ao limiar, o qual, nesse caso, é denotado por y_{Uj}.

A expressão da função de verossimilhança pode, então, ser reescrita como:

$$L_C = \prod_{j=1}^{N_H^{<}} F_X(y_{Uj} \mid \Theta) \cdot \prod_{j=1}^{N_H^{\bullet}} f_X(y_j \mid \Theta) \tag{3.50}$$

2) *Cheias de intensidade desconhecida, superior a um limiar fixo:*

$$L_{BC} = \prod_{j=1}^{t} \left\{ \binom{N_{Hj}}{k_j} \cdot [1 - F_X(y_{Hj} \mid \Theta)]^{k_j} \cdot [F_X(y_{Hj} \mid \Theta)]^{N_{Hj}-k_j} \right\} \tag{3.51}$$

Se considerarmos vários períodos sucessivos de um ano $(N_{Hj} = 1)$, com $k_j = 1$ ou 0 caso o evento do ano *j* seja, respectivamente, superior ou inferior ao limiar y_{Hj}, o período histórico de N_H anos será composto por:

- $N_H^{>}$ anos de cheias de intensidade desconhecida, superior ao limiar, o qual, nesse caso, é denotado por y_{Lj};

- $N_H^{<}$ anos de cheias de intensidade desconhecida, inferior ao limiar, o qual, nesse caso, é denotado por y_{Uj}.

A expressão da função de verossimilhança pode, então, ser reescrita como:

$$L_{BC} = \prod_{j=1}^{N_H^{>}} [1 - F_X(y_{Lj} \mid \Theta)] \cdot \prod_{j=1}^{N_H^{<}} F_X(y_{Uj} \mid \Theta) \tag{3.52}$$

3) *Cheias de intensidade compreendida em um intervalo, superior a um limiar fixo:*

$$L_{CI} = \prod_{j=1}^{t} \left\{ \binom{N_{Hj}}{k_j} \cdot [F_X(y_{Hj} \mid \Theta)]^{N_{Hj} - k_j} \cdot \prod_{m=1}^{k_j} [F_X(y_{Um,j} \mid \Theta) - F_X(y_{Lm,j} \mid \Theta)] \right\} \tag{3.53}$$

Se considerarmos vários períodos sucessivos de um ano $(N_{Hj} = 1)$, com $k_j = 1$ ou 0 caso o evento do ano *j* seja, respectivamente, superior ou inferior ao limiar y_{Hj}, o período histórico de N_H anos será composto por:

- $N_H^{<>}$ anos de cheias de intensidade compreendida no intervalo $[y_{Lj}; y_{Uj}]$, superior ao limiar;

- $N_H^{<}$ anos de cheias de intensidade desconhecida, inferior ao limiar, o qual, nesse caso, é denotado por y_{Uj}.

A expressão da função de verossimilhança pode, então, ser reescrita como:

$$L_{CI} = \prod_{j=1}^{N_H^{<}} F_X(y_{Uj} \mid \Theta) \cdot \prod_{j=1}^{N_H^{<>}} [F_X(y_{Uj} \mid \Theta) - F_X(y_{Lj} \mid \Theta)] \tag{3.54}$$

Para os dados do período sistemático, as funções de verossimilhança podem ser obtidas utilizando um raciocínio análogo ao desenvolvido para as cheias históricas. As expressões (3.50), (3.52) e (3.54) podem ser decompostas em expressões elementares, que correspondem

às funções de verossimilhança dos diferentes tipos de informações apresentadas no esquema geral da Figura 3.4:

1) *Informação não censurada*:

$$L_H^{\bullet} = \prod_{j=1}^{N_H^{\bullet}} f_X(y_j \mid \Theta) \qquad \text{Período histórico} \qquad (3.55)$$

$$L_S^{\bullet} = \prod_{i=1}^{N_S^{\bullet}} f_X(x_i \mid \Theta) \qquad \text{Período sistemático} \qquad (3.56)$$

2) *Informação censurada*:

a) inferior a um limiar:

$$L_H^{<} = \prod_{j=1}^{N_H^{<}} F_X(y_{Uj} \mid \Theta) \qquad \text{Período histórico} \qquad (3.57)$$

$$L_S^{<} = \prod_{i=1}^{N_S^{<}} F_X(x_{Ui} \mid \Theta) \qquad \text{Período sistemático} \qquad (3.58)$$

b) superior a um limiar:

$$L_H^{>} = \prod_{j=1}^{N_H^{>}} [1 - F_X(y_{Lj} \mid \Theta)] \qquad \text{Período histórico} \qquad (3.59)$$

$$L_S^{>} = \prod_{i=1}^{N_S^{>}} [1 - F_X(x_{Li} \mid \Theta)] \qquad \text{Período sistemático} \qquad (3.60)$$

c) compreendida em um intervalo:

$$L_H^{<>} = \prod_{j=1}^{N_H^{<>}} [F_X(y_{Uj} \mid \Theta) - F_X(y_{Lj} \mid \Theta)] \qquad \text{Período histórico} \qquad (3.61)$$

$$L_S^{<>} = \prod_{i=1}^{N_S^{<>}} [F_X(x_{Ui} \mid \Theta) - F_X(x_{Li} \mid \Theta)] \qquad \text{Período sistemático} \qquad (3.62)$$

Portanto, levando-se em conta os diversos tipos de informações que podem estar disponíveis, a função de verossimilhança da amostra completa, denotada por L_{TOTAL}, é dada por (NAULET, 2002):

$$L_{TOTAL} = \begin{pmatrix} L_H^{\bullet} \cdot L_H^{<} \cdot L_H^{>} \cdot L_H^{<>} \cdot \\ \cdot L_S^{\bullet} \cdot L_S^{<} \cdot L_S^{>} \cdot L_S^{<>} \end{pmatrix} \tag{3.63}$$

Vale lembrar que, como discutido no item 3.4.2, não há restrições quanto à utilização da função logaritmo de verossimilhança, ao invés da função de verossimilhança propriamente dita.

3.5.3 Método do algoritmo dos momentos esperados (EMA)

O método do algoritmo dos momentos esperados, apresentado por Lane e Cohn (1996) e Cohn *et al.* (1997), é um procedimento de estimação de parâmetros que, tal como o método dos momentos (item 3.4.1), tem como princípio a igualdade entre os momentos populacionais e os correspondentes momentos amostrais. Ao contrário do método MPH (item 3.5.1), no método do algoritmo dos momentos esperados as cheias históricas e sistemáticas são tratadas da mesma maneira, e todos os tipos de informações censuradas, mostrados no esquema geral da Figura 3.4, podem ser utilizados. Para os eventos de magnitude desconhecida (vazões censuradas), os momentos são estimados por meio dos valores esperados da variável aleatória, descrita por uma função densidade de probabilidade truncada. Por isso, eles são denominados momentos esperados.

Lembrando que as cheias sistemáticas e históricas são descritas pela mesma função densidade de probabilidade $f_X(\,)$, os momentos esperados para os três casos de informações censuradas da Figura 3.4 são dados por:

1) *Período Histórico*:

   a) cheia inferior a um limiar y_{Uj}:

$$\mu'^{<}_{H,r,j} = E[Y^r \mid Y < y_{Uj}] = \frac{1}{F_X(y_{Uj} \mid \Theta)} \int_0^{y_{Uj}} y^r f_X(y \mid \Theta)\, dy \tag{3.64}$$

   b) cheia superior a um limiar y_{Lj}:

$$\mu'^{>}_{H,r,j} = E[Y^r \mid Y > y_{Lj}] = \frac{1}{1 - F_X(y_{Lj} \mid \Theta)} \int_{y_{Lj}}^{+\infty} y^r f_X(y \mid \Theta)\, dy \tag{3.65}$$

c) cheia compreendida em um intervalo [y_{Lj};y_{Uj}]:

$$\mu'^{<>}_{H,r,j} = E[Y^r \mid y_{Lj} < Y < y_{Uj}] = \frac{1}{F_X(y_{Uj} \mid \Theta) - F_X(y_{Lj} \mid \Theta)} \int_{y_{Lj}}^{y_{Uj}} y^r f_X(y \mid \Theta)\, dy \tag{3.66}$$

2) *Período Sistemático*:

a) cheia inferior a um limiar x_{Ui}:

$$\mu'^{<}_{S,r,i} = E[X^r \mid X < x_{Ui}] = \frac{1}{F_X(x_{Ui} \mid \Theta)} \int_{0}^{x_{Ui}} x^r f_X(x \mid \Theta)\, dx \tag{3.67}$$

b) cheia superior a um limiar x_{Li}:

$$\mu'^{>}_{S,r,i} = E[X^r \mid X > x_{Li}] = \frac{1}{1 - F_X(x_{Li} \mid \Theta)} \int_{x_{Li}}^{+\infty} x^r f_X(x \mid \Theta)\, dx \tag{3.68}$$

c) cheia compreendida em um intervalo [x_{Li};x_{Ui}]:

$$\mu'^{<>}_{S,r,i} = E[X^r \mid x_{Li} < X < x_{Ui}] = \frac{1}{F_X(x_{Ui} \mid \Theta) - F_X(x_{Li} \mid \Theta)} \int_{x_{Li}}^{x_{Ui}} x^r f_X(x \mid \Theta)\, dx \tag{3.69}$$

Portanto, para a amostra completa, os momentos de ordem *r*, em relação à origem, denotados por $\breve{m}'_r$, são calculados somando-se os dados de intensidade conhecida (x_i e y_j), conforme é feito para o cálculo dos momentos amostrais com base em uma amostra não censurada, e os momentos esperados, dados pelas equações (3.64) a (3.69). Assim:

$$\breve{m}'_r = \frac{1}{N}\left(\begin{array}{l} \sum_{i=1}^{N_S^{\bullet}} x_i^r + \sum_{i=1}^{N_S^{<}} \mu'^{<}_{S,r,i} + \sum_{i=1}^{N_S^{>}} \mu'^{>}_{S,r,i} + \sum_{i=1}^{N_S^{<>}} \mu'^{<>}_{S,r,i} + \\ + \sum_{j=1}^{N_H^{\bullet}} y_j^r + \sum_{j=1}^{N_H^{<}} \mu'^{<}_{H,r,j} + \sum_{j=1}^{N_H^{>}} \mu'^{>}_{H,r,j} + \sum_{j=1}^{N_H^{<>}} \mu'^{<>}_{H,r,j} \end{array} \right) \tag{3.70}$$

Como a expressão para o cálculo de $\breve{m}'_r$ é função dos parâmetros desconhecidos Θ, o método do algoritmo dos momentos esperados constitui um processo iterativo, como mostrado na Figura 3.6. No instante inicial, são utilizados os parâmetros estimados a partir da amostra de dados sistemáticos não censurada.

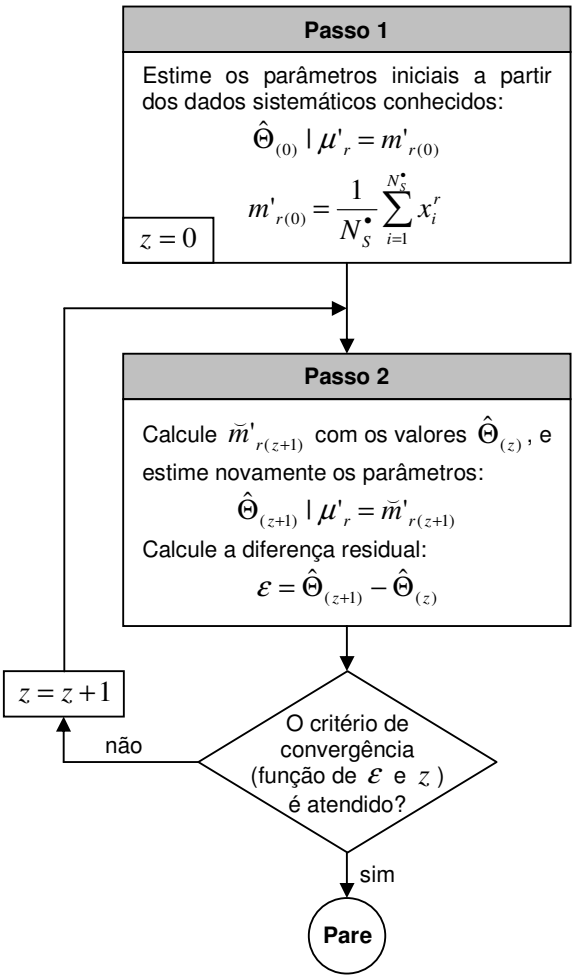


FIGURA 3.6: Processo iterativo do método do algoritmo dos momentos esperados.

Cohn *et al.* (1997) afirmam que, quanto à incorporação de cheias históricas na estimação dos parâmetros, o método EMA é mais eficiente que o método MPH (item 3.5.1), e se aproxima da eficiência alcançada pelo método MVS (item 3.5.2), requerendo, no entanto, menor esforço computacional. Lane e Cohn (1996) aplicaram o método EMA para a distribuição Log-Normal 2 parâmetros, enquanto Cohn *et al.* (1997), England (1999) e Cohn *et al.* (2001) o fizeram para a distribuição Log-Pearson III. Naulet (2002) adaptou o método para as distribuições Gumbel e GEV.

3.6 Probabilidades de excedência empíricas (posições de plotagem)

3.6.1 Probabilidades empíricas sem informação censurada

As fórmulas de posição de plotagem especificam, com base em uma amostra de tamanho *N*, ordenada de maneira decrescente, a freqüência com que determinado evento (por exemplo, a *i*-ésima maior vazão da série) é igualado ou excedido. Essa freqüência, conhecida como probabilidade empírica de excedência, é denotada por $\hat{p}_i = P[X \geq x_i] = 1 - \hat{F}_i$.

Cunnane (1978) apresentou uma revisão das principais fórmulas para cálculo da probabilidade empírica de excedência sem informação censurada. As expressões propostas na literatura têm a seguinte forma geral:

$$\hat{p}_i = \frac{i - \alpha}{N_S^{\bullet} + 1 - 2\alpha} \qquad i = 1,2,...,N_S^{\bullet} \tag{3.71}$$

onde $N_S^{\bullet}$ representa o tamanho de uma amostra de dados sistemáticos (todos de intensidade conhecida), ordenada de maneira decrescente ($x_1 \geq x_2 \geq ... \geq x_{N_S^{\bullet}}$); e α é uma constante de generalização, cujos diferentes valores correspondem aos casos particulares mostrados na Tabela 3.2.

TABELA 3.2: Valores de α para as fórmulas de probabilidade empírica de excedência.

α	Fórmula de	Distribuição mais adequada
0	Weibull	-
3/8	Blom	Normal
0,4	Cunnane	-
0,44	Gringorten	Exponencial, Gumbel
0,5	Hazen	-

3.6.2 Probabilidades empíricas com informação censurada

Hirsch e Stedinger (1987) apresentaram uma revisão das diferentes fórmulas existentes para cálculo da probabilidade empírica de excedência com cheias históricas. Esses autores propuseram uma formulação, baseada no conceito de "excedências", coerente com o método do máximo de verossimilhança utilizando informação censurada (item 3.5.2). Essa abordagem será descrita na presente dissertação, com algumas modificações introduzidas por Naulet (2002).

Considere que os dados mostrados no esquema geral da Figura 3.4 sejam representados pela variável aleatória *Z*, que corresponde às vazões máximas anuais da amostra constituída pelos dois períodos (histórico e sistemático). O cálculo das probabilidades empíricas de excedência segue, então, as etapas descritas a seguir:

1) Classifique, em ordem decrescente, a amostra das $N^{\bullet}$ cheias $(z_1 \geq z_2 \geq ... \geq z_{N^{\bullet}})$, constituída:

   - pelas $N_S^{\bullet}$ e $N_H^{\bullet}$ cheias sistemáticas x_i e históricas y_j de intensidade conhecida;
   - pelos valores médios das cheias de intensidade compreendida nos intervalos $[x_{Li}; x_{Ui}]$ e $[y_{Lj}; y_{Uj}]$ (hipótese simplificadora para representar o valor de máxima densidade).

2) Classifique, em ordem crescente, o conjunto dos *t* limiares de referência distintos $(y_{H_1} = 0 < y_{H_2} < ... < y_{H_t} < y_{H_{t+1}} = +\infty)$, de forma que o limiar y_{H_j} se aplique aos $N_j^{<}$ anos.

3) Calcule a probabilidade de excedência do limiar *j*, $p_{e_j} = P[Z \geq y_{H_j}]$, que, de acordo com a abordagem introduzida por Hirsch e Stedinger (1987) e modificada por Naulet (2002), é dada por:

$$\hat{p}_{e_j} = \hat{p}_{e_{j+1}} + \hat{p}_{c_j} \cdot (1 - \hat{p}_{e_{j+1}}) \qquad j = t, t-1, ..., 1 \tag{3.72}$$

   Conforme mostrado na Figura 3.7, $p_{e_{t+1}} = 0$ (pois $y_{H_{t+1}} = +\infty$) e $p_{e_1} = 1$ (pois $y_{H_1} = 0$). A probabilidade condicional $p_{c_j} = P[(y_{H_j} \leq Z < y_{H_{j+1}}) \mid (Z < y_{H_{j+1}})]$ é estimada por:

$$\hat{p}_{c_j} = \frac{A_j}{A_j + B_j + C_j} \qquad j = t, t-1, ..., 1 \tag{3.73}$$

   onde:

   A_j é o número de cheias conhecidas no intervalo $[y_{H_j}; y_{H_{j+1}}]$;

   B_j é o número de cheias conhecidas inferiores a y_{H_j};

   $C_j = \sum_{k=1}^{j} N_k^{<}$ é o número de cheias desconhecidas inferiores a y_{H_j}.

4) Finalmente, calcule as probabilidades empíricas de excedência ($\hat{p}_i$) das A_j cheias compreendidas entre os limiares $[y_{H_j}; y_{H_{j+1}}]$, utilizando a seguinte expressão:

$$\hat{p}_i = p_{e_{j+1}} + (p_{e_j} - p_{e_{j+1}}) \cdot \left(\frac{i - \alpha}{A_j + 1 - 2\alpha} \right) \qquad \begin{matrix} j = t, t-1, ..., 1 \\ i = 1, 2, ..., A_j \end{matrix} \tag{3.74}$$

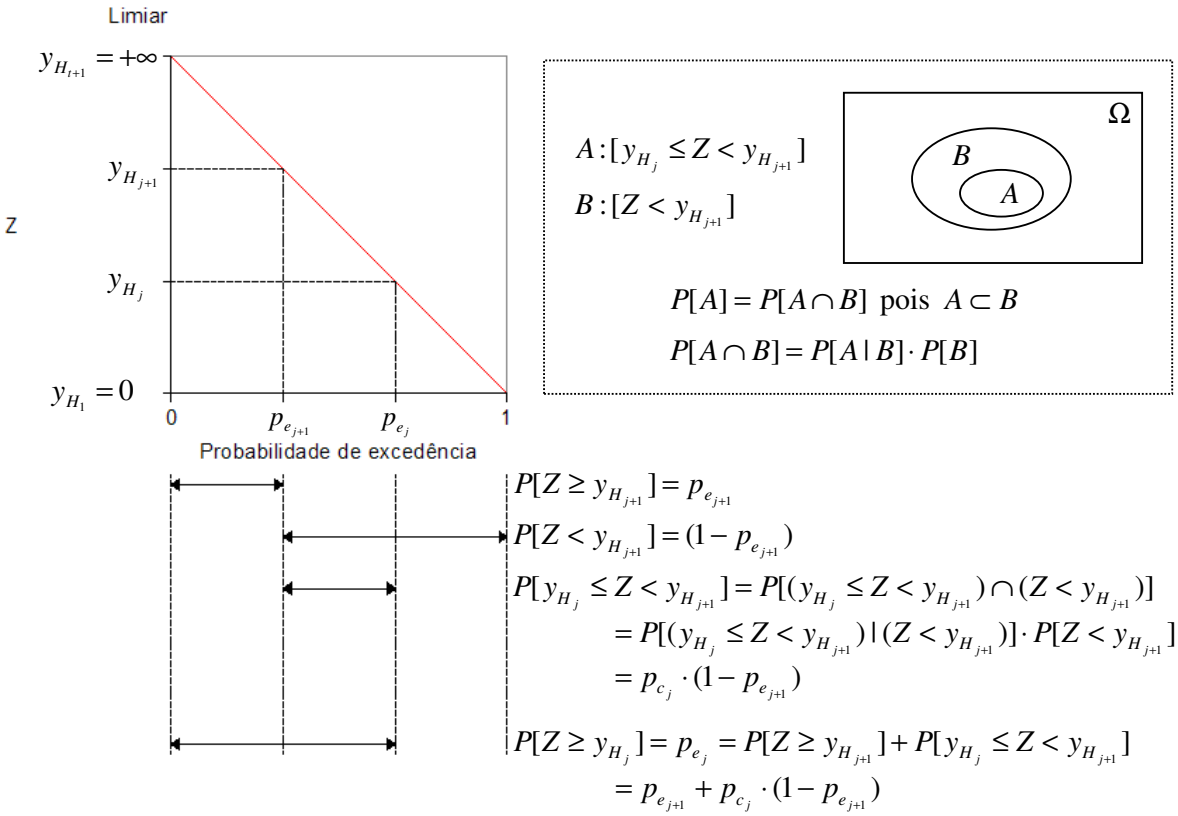


FIGURA 3.7: Esquema para o cálculo das probabilidades de excedência dos limiares.
Fonte: adaptado de NAULET, 2002.

3.7 Redução da incerteza na análise de freqüência devido ao uso de dados não sistemáticos

Os dados não sistemáticos (paleovazões e cheias históricas) podem ser bastante valiosos na análise de freqüência de vazões de enchentes. Vários autores (HOSKING e WALLIS, 1986a, 1986b; STEDINGER e COHN, 1986; COHN e STEDINGER, 1987; SUTCLIFFE, 1987; GUO e CUNNANE, 1991; FRANCÉS *et al.*, 1994; COHN *et al.*, 1997; MARTINS e STEDINGER, 2000, 2001a, 2001b; BAYLISS e REED, 2001; O'CONNELL *et al.*, 2002; NAULET, 2002; BENITO *et al.*, 2004) demonstraram que a incorporação de dados não

sistemáticos na análise de freqüência pode estender significativamente o período de observação, reduzindo o grau de extrapolação e melhorando o nível de confiabilidade na estimação do risco associado a eventos raros. Segundo Salas *et al.* (1994), o uso de informações históricas e paleohidrológicas provê a mais abrangente reconstrução da magnitude e freqüência das cheias ocorridas antes do período de registros sistemáticos. Parent e Bernier (2003) afirmam que, ainda que esses dados sejam esparsos e muitas vezes imprecisos, eles provêem valiosas informações sobre o comportamento das distribuições de freqüência no domínio das cheias extremas. De acordo com Thorndycraft *et al.* (2003), a acurácia das estimativas de vazões deduzidas de evidências históricas e paleohidrológicas de cheias tem aumentado recentemente, com a melhoria da capacidade computacional e com o desenvolvimento de novas tecnologias para obtenção de dados topográficos, o que permite elevar a eficiência da modelagem hidráulica usando modelos uni e bidimensionais.

É claro que a redução das incertezas envolvidas na estimação dos parâmetros e quantis depende de alguns fatores, dentre os quais destacam-se: (1) a distribuição de probabilidades selecionada (modelos de dois ou três parâmetros), (2) o método de estimação dos parâmetros (MPH, MVS ou EMA), (3) o tipo de informação não sistemática disponível (censurada, binomial censurada ou censurada em um intervalo), e (4) a abordagem usada na determinação da amostra (série de vazões máximas anuais ou série de duração parcial).

De maneira geral, os trabalhos realizados demonstram que o valor da informação não sistemática é maior quando três parâmetros precisam ser estimados, e que, em termos de redução da variância dos estimadores paramétricos, o método do máximo de verossimilhança é o mais eficiente.

Cohn e Stedinger (1987) definiram o tamanho efetivo da amostra (*effective record length – ERL*) como o número de anos de dados sistemáticos que produziriam, para um determinado quantil X_T, o mesmo erro quadrático médio (*mean square errror – MSE*) obtido utilizando-se uma combinação de dados sistemáticos e não sistemáticos. O tamanho efetivo da amostra é dado por:

$$ERL(N_S, N_H) = N_S \left(\frac{MSE(\hat{X}_T \mid N_S, 0)}{MSE(\hat{X}_T \mid N_S, N_H)} \right) \tag{3.75}$$

onde $MSE(\hat{X}_T) = E[(\hat{X}_T - X_T)^2]$, N_S é o número de anos do período sistemático e N_H é o número de anos do período não sistemático. Para um período sistemático de tamanho definido, o tamanho efetivo da amostra (*ERL*) cresce de forma aproximadamente linear à medida que o número de anos do período não sistemático aumenta.

Cohn e Stedinger (1987) definiram ainda o ganho marginal (*marginal gain – MG*), que representa uma medida da eficiência com que a informação não sistemática é explorada para reduzir a incerteza associada à utilização de uma amostra relativamente curta de dados sistemáticos. O ganho marginal é dado por:

$$MG = \frac{ERL(N_S, N_H) - N_S}{N_H} \tag{3.76}$$

Caso não se disponha de dados não sistemáticos precisos, melhorias significativas são alcançadas com o uso de informação binomial censurada. Especialmente quando o período de retorno do limiar de referência é elevado, a diferença entre o ganho obtido com a utilização de informação censurada ou binomial censurada é desprezível.

Martins e Stedinger (2001b) concluíram que o uso de informação não sistemática pode ser de grande importância, e que a escolha da abordagem baseada em séries de vazões máximas anuais ou em séries de duração parcial não resulta em diferenças consideráveis.

4 ABORDAGEM BAYESIANA NA ANÁLISE DE FREQÜÊNCIA DE VAZÕES MÁXIMAS ANUAIS

4.1 Introdução

Conforme descrito no Capítulo 1, algumas incertezas (*natural*, *estatística* e *do modelo*) são inerentes aos procedimentos de análise de freqüência de cheias, os quais permitem inferir sobre a ocorrência de vazões futuras, usando um modelo probabilístico conveniente, cujos parâmetros desconhecidos são estimados a partir de uma amostra particular. No entanto, os métodos clássicos de análise de freqüência não permitem considerar adequadamente as incertezas envolvidas, podendo conduzir a decisões incorretas.

Uma maneira coerente de lidar com esse problema é utilizar os procedimentos de inferência bayesiana, que permitem levar em conta a incerteza estatística e reunir toda a informação disponível, seja ela histórica, paleohidrológica, regional ou subjetiva. Na análise bayesiana de freqüência, a distribuição de probabilidades da variável aleatória de interesse (vazão máxima anual, no caso do presente estudo) é combinada com alguma informação avaliada *a priori*, que possa parecer pertinente, para produzir estimativas mais confiáveis da probabilidade de ocorrência de cenários futuros. Os principais aspectos referentes à aplicação da estatística bayesiana na análise de freqüência de vazões máximas anuais serão apresentados nesse capítulo.

4.2 Teorema de Bayes

No contexto da análise bayesiana de freqüência, o teorema de Bayes, devido ao matemático inglês Thomas Bayes (1702-1761), constitui uma ferramenta fundamental. Ele resulta de uma interessante combinação entre a regra da multiplicação e o teorema da probabilidade total. As argumentações de Naghettini e Pinto (no prelo), relativas a esses dois teoremas, são descritas a seguir.

Suponha que o espaço amostral Ω de um certo experimento seja o resultado da união de k eventos mutuamente excludentes $(B_1, B_2, ..., B_k)$, todos com probabilidade de ocorrência diferente de zero. Considere, também, um evento A, cuja probabilidade de ocorrência é $P[A] = P[B_1 \cap A] + P[B_2 \cap A] + ... + P[B_k \cap A]$. Essa situação é ilustrada na Figura 4.1.

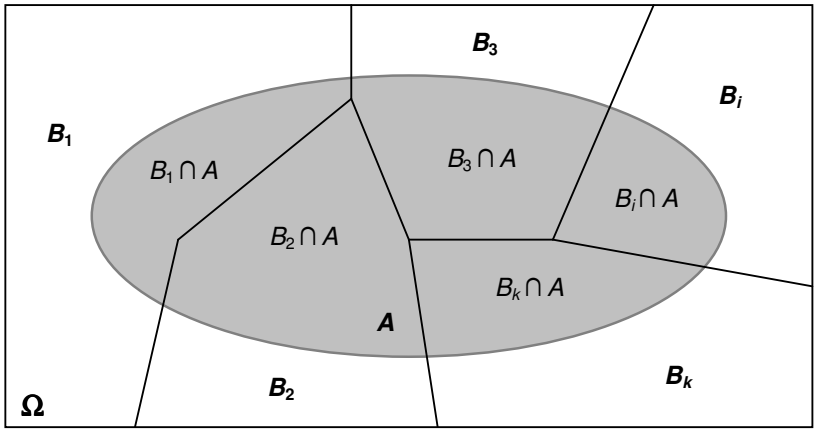


FIGURA 4.1: Diagrama de Venn para o Teorema da Probabilidade Total.

Usando a definição de probabilidade condicional, a probabilidade de ocorrência do evento A é dada pela equação (4.1), que é a expressão formal do chamado *teorema da probabilidade total*:

$$P[A] = P[B_1] \cdot P[A \mid B_1] + P[B_2] \cdot P[A \mid B_2] + ... + P[B_k] \cdot P[A \mid B_k]$$

$$P[A] = \sum_{i=1}^{k} P[B_i] \cdot P[A \mid B_i] \tag{4.1}$$

Considerando novamente a situação mostrada na Figura 4.1, a probabilidade de ocorrência de qualquer um dos eventos mutuamente excludentes (B_r, por exemplo), condicionada à ocorrência do evento A, é expressa por:

$$P[B_r \mid A] = \frac{P[B_r \cap A]}{P[A]} \tag{4.2}$$

Pela regra da multiplicação, o numerador do segundo membro da equação (4.2) pode ser expresso por $P[A \mid B_r] \cdot P[B_r]$, enquanto o denominador pode ser posto na forma do teorema da probabilidade total. A equação resultante é a expressão do *teorema de Bayes*:

$$P[B_r \mid A] = \frac{P[A \mid B_r] \cdot P[B_r]}{\sum_{i=1}^{k} P[A \mid B_i] \cdot P[B_i]} \tag{4.3}$$

De acordo com Lee (1989), a estatística bayesiana envolve sempre a definição de duas probabilidades para uma certa hipótese: uma incondicional (a probabilidade *a priori*) e outra condicionada a alguma evidência (a probabilidade *a posteriori*). Nesse sentido, Naghettini e Pinto (no prelo) argumentam que o teorema de Bayes constitui um quadro lógico importante para a "revisão" ou a "atualização" de probabilidades previamente estabelecidas, à luz de novas informações.

Para exemplificar tal possibilidade, considere a necessidade hipotética de se calcular a probabilidade de que a temperatura mínima em um dia qualquer de Janeiro, num dado local, esteja acima de 15ºC. Nesse caso, seja B o evento correspondente às temperaturas mínimas diárias superiores a 15ºC e B^c o evento complementar, de tal maneira que eles sejam mutuamente excludentes e que, portanto, $B \cup B^c = \Omega$. Se a única informação disponível é que o evento B ocorreu em 25 dos 31 dias do mês de Janeiro, é natural que a probabilidade $P[B]$ seja estimada pela freqüência relativa dos dias de Janeiro em que a temperatura mínima excedeu 15ºC, nesse caso $(25/31) = 80{,}65\%$. Dentro do contexto do teorema de Bayes, essa estimativa é denominada probabilidade *a priori* ou subjetiva, indicando o grau de confiança inicial do meteorologista quanto à ocorrência de B. Entretanto, a temperatura mínima diária pode ser afetada pela ocorrência de precipitações naquele dia e, supondo que se preveja um dia chuvoso, tal cenário certamente irá modificar a probabilidade *a priori* $P[B]$. Para incorporar essa nova informação, é preciso conhecer as estimativas de $P[A]$ e $P[A \mid B]$, que denotam, respectivamente, as probabilidades de ocorrer chuva em um dia qualquer de Janeiro e apenas nos dias com temperatura mínima maior que 15ºC. Suponha que a análise de freqüência dos registros históricos resulte na estimativa de 18 dias chuvosos em Janeiro, dos quais 15 apresentam temperatura mínima diária superior a 15ºC. Assim, $P[A] = (18/31)$ e $P[A \mid B] = (15/25)$. Com essas estimativas na equação (4.3), e lembrando que o denominador da referida equação é, de fato, a $P[A]$, tem-se:

$$P[B \mid A] = \frac{P[A \mid B] \cdot P[B]}{P[A]} = \frac{(15/25) \cdot (25/31)}{(18/31)} = 83{,}33\%$$

Essa é a probabilidade *a posteriori*, "revisada" pela incorporação da informação de que o evento A ocorreu.

4.3 Análise bayesiana de freqüência de vazões máximas anuais

A incorporação da estatística bayesiana na análise de freqüência de cheias tem sido objeto de estudo de vários autores. Adotando essa abordagem, é possível desenvolver uma distribuição marginal de probabilidades para eventos futuros de cheias, que permite levar em conta as incertezas do fenômeno natural, bem como aquelas associadas à estimação dos parâmetros da distribuição (BENJAMIN e CORNELL, 1970; WOOD *et al.*, 1974; VICENS *et al.*, 1975; WOOD e RODRIGUEZ-ITURBE, 1975a; WOOD, 1978). Alguns trabalhos incorporam ainda as incertezas referentes à escolha do modelo probabilístico (WOOD *et al.*, 1974; WOOD e RODRIGUEZ-ITURBE, 1975b). Embora estejam fora do escopo da presente dissertação, existem estudos que, sob a ótica bayesiana, estabelecem também procedimentos para a determinação de uma vazão de projeto "ótima" q_T, estimada de forma que um determinado critério econômico seja obedecido (BENJAMIN e CORNELL, 1970; DAVIS *et al.*, 1972, 1979; WOOD e RODRIGUEZ-ITURBE, 1975a).

Em muitos casos, a análise bayesiana de freqüência de cheias requer o uso de procedimentos numéricos avançados, como os implementados nos *softwares* FLIKE (KUCZERA, 1999) e FLDFRQ3 (O'CONNELL *et al.*, 2002). O *software* FLIKE, descrito brevemente no item 4.5, será utilizado para a realização do estudo de caso relatado no Capítulo 5. Os procedimentos de inferência bayesiana apresentados por Wood *et al.* (1974) e Lee (1989) serão adotados nessa dissertação e descritos nos próximos itens.

4.3.1 Inferência bayesiana

Considere a situação em que um hidrólogo é solicitado a elaborar estimativas sobre a freqüência de ocorrência de vazões extremas. Antes de utilizar os dados fluviométricos do local de interesse, ele poderá reunir toda a informação disponível sobre o conjunto de parâmetros do modelo, representado pelo vetor $\Theta = (\theta_1, \theta_2, ..., \theta_k)$. Essa informação será descrita *a priori* por uma distribuição de probabilidades, representando aquilo que o hidrólogo pôde inferir sobre os parâmetros do modelo distributivo, à luz, por exemplo, de modelos regionais e/ou de seu conhecimento acumulado ao longo dos anos. Assim, dado o mesmo conjunto de informações iniciais, é bastante plausível que duas pessoas formulem distribuições de probabilidades *a priori* distintas. Quando, inicialmente, nada se pode inferir sobre os parâmetros ou as informações disponíveis não são suficientemente fortes, pode-se

adotar uma distribuição de probabilidades *a priori* não informativa, expressa, por exemplo, por uma distribuição uniforme.

O hidrólogo, agora, poderá lançar mão da amostra de vazões máximas anuais do local de interesse, denotada por $Q = (q_1, q_2, ..., q_N)$, a qual presume-se ter sido retirada da população descrita pela distribuição $f(q \mid \Theta)$, condicionada ao vetor de parâmetros Θ. É importante salientar que a amostra Q deve ser constituída pela maior quantidade possível de informações sobre cheias, podendo, conforme exposto no Capítulo 3, incorporar dados sistemáticos e não sistemáticos em um contexto censurado $(N = N_S + N_H)$.

O teorema de Bayes, apresentado no item 4.2, pode também ser utilizado para se realizar inferências sobre uma determinada variável aleatória, ao invés de inferências sobre a ocorrência de eventos específicos. Dessa forma, a distribuição de probabilidades *a priori* dos parâmetros pode ser "atualizada" pelos dados fluviométricos locais, resultando na distribuição de probabilidades *a posteriori* dos parâmetros, como mostra a equação (4.4), que representa a expressão do *teorema de Bayes para variáveis aleatórias contínuas*:

$$f_1(\Theta \mid Q) = \frac{f(q \mid \Theta) \cdot f_0(\Theta)}{\int_{\Theta} f(q \mid \Theta) \cdot f_0(\Theta)\, d\Theta} \tag{4.4}$$

Nessa equação, $f_1(\Theta \mid Q)$ é a distribuição de probabilidades *a posteriori* dos parâmetros, condicionada aos dados fluviométricos; $f(q \mid \Theta)$ é a distribuição de probabilidades das vazões máximas anuais, condicionada ao vetor de parâmetros; e $f_0(\Theta)$ é a distribuição de probabilidades *a priori* dos parâmetros.

No contexto do teorema de Bayes, as variáveis aleatórias para as quais busca-se formular uma distribuição de probabilidades são os parâmetros do modelo. Assim, uma vez que os dados fluviométricos já foram observados, a amostra é considerada fixa e a distribuição de probabilidades das vazões máximas anuais passa a ser denotada por $L(\Theta \mid Q)$, interpretada como a função de verossimilhança dos parâmetros desconhecidos, dada a amostra. Considerando-se que o denominador da equação (4.4) resulta em uma função dependente apenas de *q*, e que, como enfatizado anteriormente, na aplicação do teorema de Bayes, os

dados fluviométricos não são considerados variáveis, esse termo pode ser tratado como uma constante de normalização, e a equação (4.4) pode ser reescrita da seguinte forma:

$$f_1(\Theta \mid Q) \propto L(\Theta \mid Q) \cdot f_0(\Theta) \tag{4.5}$$

Pode-se perceber que a distribuição de probabilidades *a posteriori* dos parâmetros é obtida por meio de uma combinação entre a informação avaliada *a priori* e a informação fornecida pelos dados fluviométricos. Dessa forma, fica claro que, para uma dada amostra de vazões máximas anuais, a formulação de diferentes distribuições *a priori* resultará em distribuições *a posteriori* distintas (veja Figura 4.2). Entretanto, à medida que aumenta a quantidade de informação contida na função de verossimilhança, a utilização de diferentes distribuições *a priori* se torna cada vez menos impactante, conduzindo a distribuições *a posteriori* cada vez mais parecidas. Em outras palavras, assintoticamente, a função de verossimilhança "domina" a distribuição *a priori*.

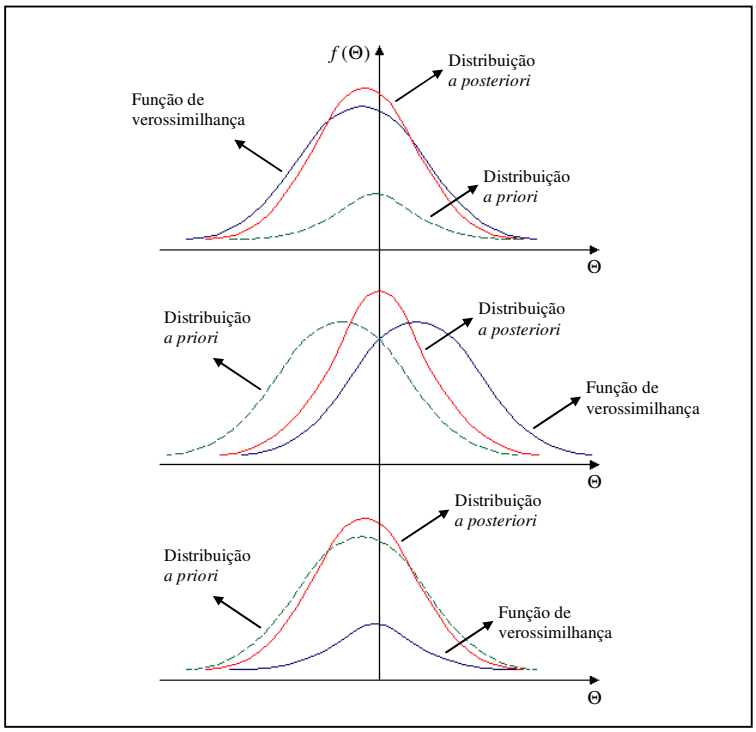


FIGURA 4.2: Diferentes distribuições *a posteriori*, resultantes da combinação de diferentes distribuições *a priori* e funções de verossimilhança.
Fonte: adaptado de WOOD *et al.*, 1974.

A Figura 4.3 mostra como o processamento, por meio do teorema de Bayes, da informação *a priori* sobre os parâmetros resulta em informações com menor grau de incerteza *a posteriori*.

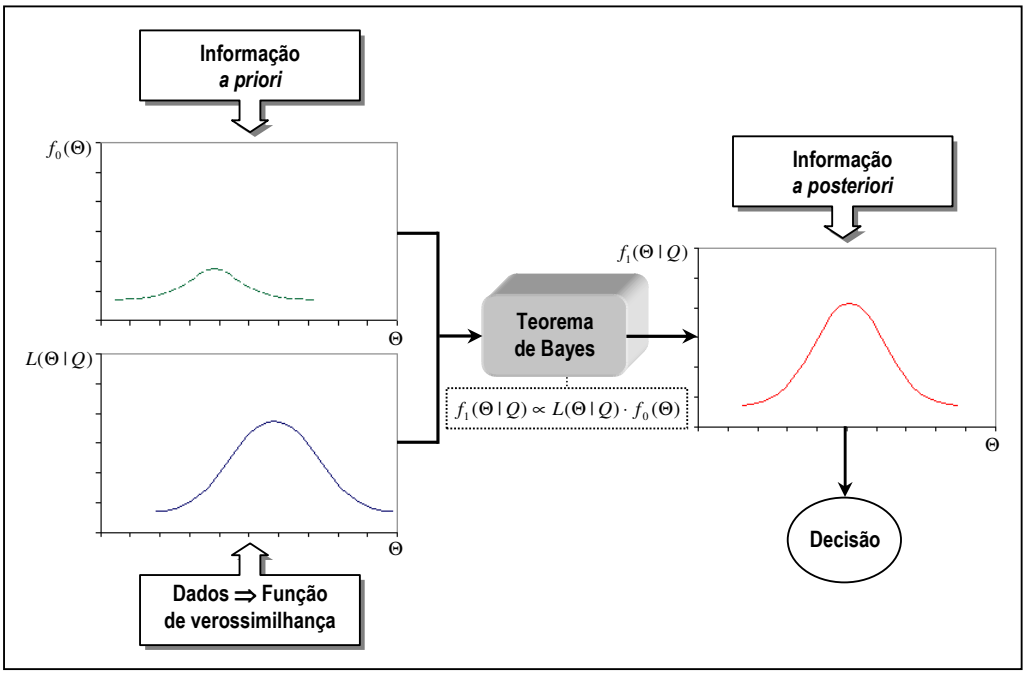


FIGURA 4.3: Aplicação do Teorema de Bayes na análise de freqüência de cheias.
Fonte: adaptado de PARENT e BERNIER, 2003.

Nesse momento, o hidrólogo dispõe de duas distribuições de probabilidades: $f(q \mid \Theta)$, que modela as vazões máximas anuais e depende do vetor de parâmetros, cujos verdadeiros valores populacionais são desconhecidos, e $f_1(\Theta \mid Q)$, que representa o comportamento dos parâmetros do modelo distributivo, após a análise de toda a informação disponível. Para inferir sobre a ocorrência de eventos futuros de cheias, o hidrólogo deve, então, prosseguir com a análise bayesiana de freqüência utilizando uma das duas abordagens apresentadas na literatura. A primeira consiste em encontrar o estimador pontual ótimo (ou *estimador bayesiano*) do vetor de parâmetros, denotado por Θ^*, e utilizar a distribuição $f(q \mid \Theta^*)$ para modelar as vazões máximas anuais. A segunda busca formular uma distribuição de probabilidades independente das estimativas dos parâmetros, chamada *distribuição bayesiana* ou *preditiva*. Essas abordagens são descritas a seguir.

4.3.1.1 Estimação pontual bayesiana

Seja $\ell(\hat{\Theta}, \Theta)$ uma função que representa a perda verificada quando o estimador $\hat{\Theta}$ é usado no lugar do vetor Θ, que contém os verdadeiros valores populacionais dos parâmetros. Uma vez que, na análise bayesiana, os parâmetros são tratados como variáveis aleatórias, a função de perda também o será, e o seu valor esperado pode ser expresso por:

$$E[\ell(\hat{\Theta},\Theta)] = \int_{\Theta} \ell(\hat{\Theta},\Theta) \cdot f_1(\Theta \mid Q)\, d\Theta \tag{4.6}$$

O estimador bayesiano ótimo Θ^* é aquele que minimiza a perda esperada em todo o domínio de Θ, ou seja:

$$\Theta^* \leftarrow \min E[\ell(\hat{\Theta},\Theta)] \tag{4.7}$$

Dependendo da forma da função de perda, diferentes estimadores bayesianos irão minimizar a equação (4.6). Conforme descrito em Wood *et al.* (1974), para uma função de perda quadrática, do tipo $\ell(\hat{\Theta},\Theta) = c(\hat{\Theta}-\Theta)^2$, onde *c* é uma constante, o estimador bayesiano ótimo é o valor esperado de Θ. Portanto:

$$\Theta^* = E[\Theta] = \mu_{\Theta} = \int_{\Theta} \Theta \cdot f_1(\Theta \mid Q)\, d\Theta \tag{4.8}$$

Nessas condições, como o estimador bayesiano minimiza uma função de perda quadrática, ele é considerado o estimador ótimo em termos de mínimos erros quadrados.

4.3.1.2 Distribuição bayesiana

A distribuição bayesiana ou preditiva pode ser encontrada integrando-se, em todo o domínio de Θ, a distribuição conjunta das vazões máximas anuais e dos parâmetros, de forma a obter uma distribuição marginal de *q*, ou seja:

$$\tilde{f}(q) = \int_{\Theta} f(q \mid \Theta) \cdot f_1(\Theta \mid Q)\, d\Theta \tag{4.9}$$

onde $\tilde{f}(q)$ é a distribuição bayesiana de probabilidades das vazões máximas anuais (distribuição marginal de *q*) e $f(q \mid \Theta) \cdot f_1(\Theta \mid Q) = f(q,\Theta)$ representa a distribuição conjunta de probabilidades das vazões máximas anuais e dos parâmetros do modelo.

A distribuição bayesiana de probabilidades pode ser vista como o valor esperado da distribuição $f(q \mid \Theta)$, levando-se em conta todos os valores possíveis dos parâmetros. Como demonstrado por Wood *et al.* (1974), apenas a abordagem baseada na formulação da distribuição bayesiana permite considerar adequadamente toda a incerteza relativa aos

parâmetros do modelo. De fato, o procedimento de estimação pontual bayesiana, ao empregar apenas o primeiro momento μ_Θ, não utiliza todo o conhecimento disponível sobre o vetor de parâmetros, representado pela distribuição $f_1(\Theta \mid Q)$. Isso irá resultar na subestimação da variância das vazões máximas anuais, e as inferências sobre *q* não refletirão completamente as incertezas existentes.

Para a grande maioria das distribuições de probabilidades utilizadas na análise de freqüência de vazões máximas anuais, alguns métodos numéricos precisam ser implementados para a resolução da equação (4.9). Entretanto, se a distribuição de probabilidades *a priori* dos parâmetros for expressa em um formato adequado, definido como a forma natural conjugada de $L(\Theta \mid Q)$, $f_1(\Theta \mid Q)$ apresentará o mesmo formato que $f_0(\Theta)$. Nessas condições, como demonstrado por Wood *et al.* (1974), Stedinger (1983) e Lee (1989), as distribuições $f_1(\Theta \mid Q)$ e $\tilde{f}(q)$ podem ser avaliadas analiticamente para os casos em que são empregados certos modelos distributivos comuns, tais como o Normal e o Log-Normal 2 parâmetros.

4.3.2 Inferência bayesiana para a distribuição Normal

4.3.2.1 Formulação da distribuição *a priori* e *a posteriori* para os parâmetros

Seja Q uma amostra não censurada, constituída por *n* vazões máximas anuais observadas no período sistemático. Admitindo que essa amostra tenha sido extraída de uma população modelada pela distribuição Normal, com média μ e variância $\varphi = \sigma^2$ desconhecidas, a função densidade de probabilidade das vazões máximas anuais será:

$$f(q \mid \mu, \varphi) \propto \varphi^{-1/2} \cdot \exp\left[-\frac{1}{2\varphi}(q-\mu)^2 \right] \tag{4.10}$$

Considerando o princípio da independência serial, a função de verossimilhança é dada pela equação (4.11):

$$L(\mu, \varphi \mid Q) = \prod_{i=1}^{n} f(q_i \mid \mu, \varphi)$$

$$L(\mu, \varphi \mid Q) \propto \varphi^{-n/2} \cdot \exp\left[-\frac{1}{2\varphi} \sum_{i=1}^{n} (q_i - \mu)^2 \right]$$

$$L(\mu,\varphi \mid Q) \propto \varphi^{-n/2} \cdot \exp\left[-\frac{1}{2\varphi}\left(\sum_{i=1}^{n}(q_i-\overline{q})^2 + n(\mu-\overline{q})^2\right)\right]$$

$$L(\mu,\varphi \mid Q) \propto \varphi^{-n/2} \cdot \exp\left[-\frac{1}{2\varphi}\left(s^2\nu + n(\mu-\overline{q})^2\right)\right] \qquad (4.11)$$

onde:

$$\overline{q} = \frac{1}{n}\sum_{i=1}^{n} q_i \qquad \text{(média amostral)}$$

$$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(q_i-\overline{q})^2 \qquad \text{(variância amostral)}$$

$$\nu = n-1$$

Suponha que o conhecimento *a priori* sobre a variância φ seja expresso por um múltiplo da distribuição Qui-Quadrado inversa, com $\nu_0 = n_0 - 1$ graus de liberdade, tal que:

$$\varphi \sim s_0^2 \nu_0 \chi_{\nu_0}^{-2}$$

$$f_0(\varphi) \propto \varphi^{-(\nu_0/2)-1} \cdot \exp\left(-\frac{1}{2\varphi}s_0^2\nu_0\right)$$

$$E[\varphi] = \frac{s_0^2\nu_0}{\nu_0-2} \qquad \nu_0 > 2$$

$$VAR[\varphi] = \frac{2(E[\varphi])^2}{\nu_0-4} = \frac{2(s_0^2\nu_0)^2}{(\nu_0-2)^2(\nu_0-4)} \qquad \nu_0 > 4 \qquad (4.12)$$

Suponha, também, que a distribuição *a priori* da média μ, condicionada a φ, seja Normal, com média m_0 e variância φ/n_0, tal que:

$$\mu \mid \varphi \sim N[m_0, \varphi/n_0]$$

$$f_0(\mu \mid \varphi) \propto \left(\frac{\varphi}{n_0}\right)^{-1/2} \cdot \exp\left[-\frac{n_0}{2\varphi}(\mu-m_0)^2\right]$$

$$E[\mu \mid \varphi] = m_0$$

$$VAR[\mu \mid \varphi] = \frac{\varphi}{n_0} = \frac{s_0^2\nu_0}{n_0(\nu_0-2)} \qquad \nu_0 > 2 \qquad (4.13)$$

Portanto, para os parâmetros μ e φ, a distribuição conjunta *a priori* será uma Normal/Qui-Quadrado inversa, dada pela equação (4.14), que constitui uma forma natural conjugada da função de verossimilhança $L(\mu,\varphi \mid Q)$.

$$f_0(\mu,\varphi) = f_0(\varphi)\cdot f_0(\mu \mid \varphi)$$

$$f_0(\mu,\varphi) \propto \varphi^{-[(\nu_0+1)/2]-1}\cdot \exp\left\{-\frac{1}{2\varphi}\left[s_0^2\nu_0 + n_0(\mu-m_0)^2\right]\right\} \tag{4.14}$$

Aplicando o teorema de Bayes para variáveis aleatórias contínuas, pode-se, então, encontrar a distribuição conjunta *a posteriori* dos parâmetros, que também será uma Normal/Qui-Quadrado inversa:

$$f_1(\mu,\varphi \mid Q) \propto L(\mu,\varphi \mid Q)\cdot f_0(\mu,\varphi)$$

$$f_1(\mu,\varphi \mid Q) \propto \varphi^{-[(\nu_0+n+1)/2]-1}\cdot \exp\left\{-\frac{1}{2\varphi}\left[s^2\nu + s_0^2\nu_0 + n(\mu-\overline{q})^2 + n_0(\mu-m_0)^2\right]\right\}$$

$$f_1(\mu,\varphi \mid Q) \propto \varphi^{-[(\nu_1+1)/2]-1}\cdot \exp\left\{-\frac{1}{2\varphi}\left[s_1^2\nu_1 + n_1(\mu-m_1)^2\right]\right\} \tag{4.15}$$

onde:

$$n_1 = n_0 + n$$

$$m_1 = \frac{1}{n_1}(n_0\cdot m_0 + n\cdot \overline{q})$$

$$\nu_1 = \nu_0 + \nu + 1 = \nu_0 + n$$

$$s_1^2 = \frac{1}{\nu_1}(s_0^2\cdot\nu_0 + s^2\cdot\nu + n_0\cdot m_0^2 + n\cdot\overline{q}^2 - n_1\cdot m_1^2)$$

4.3.2.2 Determinação da distribuição bayesiana das vazões máximas anuais

A distribuição bayesiana das vazões máximas anuais é encontrada aplicando-se a equação (4.9). Se a distribuição *a posteriori* dos parâmetros μ e φ é uma Normal/Qui-Quadrado inversa, dada pela equação (4.15), e se as vazões máximas anuais são modeladas por uma

distribuição Normal, conforme a equação (4.10), tal que $q \sim N[\mu,\varphi]$, é possível demonstrar que a distribuição bayesiana de *q* será *t* de Student (veja Anexo B):

$$\tilde{f}(q) = \int_{\mu}\int_{\varphi} f(q \mid \mu,\varphi) \cdot f_1(\mu,\varphi \mid Q)\, d\varphi\, d\mu$$

$$\tilde{f}(q) = B(\nu_1/2, 1/2)^{-1} \cdot \left[1 + \frac{(q-m_1)^2}{(s_1^2\nu_1)/r_1}\right]^{-(\nu_1+1)/2} \cdot \left(\frac{s_1^2\nu_1}{r_1}\right)^{-1/2} \tag{4.16}$$

onde:

$$B(\nu_1/2, 1/2) = \frac{\pi^{1/2}\Gamma(\nu_1/2)}{\Gamma(\nu_1/2 + 1/2)}$$

$$r_1 = \frac{n_1}{n_1+1}$$

Em conseqüência, as vazões máximas anuais *q* serão distribuídas conforme a equação (4.17):

$$\frac{q-m_1}{s_1/\sqrt{n_1/(n_1+1)}} \sim t_{\nu_1} \tag{4.17}$$

Caso seja adotada uma distribuição *a priori* não informativa, m_1, s_1, n_1 e ν_1 assumirão seus valores amostrais $\bar{q}$, s, n e ν. Dessa forma, inferências sobre um processo Normal, com média e variância desconhecidas, serão feitas usando uma distribuição *t* de Student, que permite considerar completamente as incertezas relativas aos parâmetros. Para os mesmos dados amostrais, a forma da distribuição *t* de Student é parecida com a da distribuição Normal. A primeira, no entanto, possui maior variância (WOOD *et al.*, 1974).

4.3.3 Inferência bayesiana para a distribuição Log-Normal 2 parâmetros

Considere, agora, que as vazões máximas anuais sejam modeladas pela distribuição Log-Normal 2 parâmetros. Nesse caso, os procedimentos da análise bayesiana de freqüência serão análogos àqueles discutidos no item 4.3.2, que se refere à situação em que os dados fluviométricos são modelados pela distribuição Normal. A única diferença é que a amostra a ser utilizada será constituída pelos logaritmos das *n* observações de vazões máximas anuais do local de interesse.

4.3.3.1 Exemplo de aplicação: rio São Francisco em Manga

Com o intuito de demonstrar a aplicação dos procedimentos descritos anteriormente, foi realizada a análise bayesiana de freqüência de cheias no rio São Francisco, em Manga (posto fluviométrico 44500000, com 202.400 km² de área de drenagem). As informações disponíveis são: a série de vazões máximas anuais do referido posto fluviométrico, mostrada no Anexo E, e um conjunto hipotético de informações *a priori* sobre os parâmetros do modelo distributivo, provenientes, por exemplo, de uma análise regional.

Como, por pressuposto, a variável aleatória q é modelada pela distribuição Log-Normal 2 parâmetros, foi utilizada uma amostra constituída pelos logaritmos das vazões máximas anuais observadas em Manga, tal que $x = \ln(q)$.

As estimativas amostrais dos parâmetros, obtidas pelo método dos momentos, são:

$$n = 64$$
$$\bar{x} = 8{,}7718$$
$$\nu = 63$$
$$s^2 = 0{,}0799$$

As características *a priori* da variância e da média dos logaritmos das vazões, supostamente obtidas por análise regional, são:

$$E[\varphi] = 0{,}090$$
$$VAR[\varphi] = 0{,}015$$
$$E[\mu \mid \varphi] = 8{,}600$$
$$VAR[\mu \mid \varphi] = 0{,}015$$

Portanto, usando as equações (4.12) e (4.13), tem-se os parâmetros da distribuição *a priori*, dada pela equação (4.14):

$$\nu_0 = [(2 \cdot 0{,}090^2)/0{,}015] + 4 = 5{,}08 \approx 5$$
$$n_0 = 5 + 1 = 6$$
$$m_0 = 8{,}600$$
$$s_0^2 = [0{,}015 \cdot 6 \cdot (5-2)]/5 = 0{,}054$$

Combinando as estimativas amostrais e as estimativas baseadas na informação *a priori*, como mostrado na equação (4.15), tem-se os parâmetros da distribuição *a posteriori*:

$$n_1 = 70$$

$$m_1 = 8{,}7571$$

$$\nu_1 = 69$$

$$s_1^2 = 0{,}0792$$

Para esse exemplo, a função de verossimilhança e as distribuições *a priori* e *a posteriori* dos parâmetros μ e φ são mostradas nas Figuras 4.4, 4.5 e 4.6, respectivamente. Pode-se perceber que a utilização de informações avaliadas *a priori*, embora relativamente menos precisas que a informação contida na função de verossimilhança, resulta em uma distribuição *a posteriori* com menor grau de incertezas.

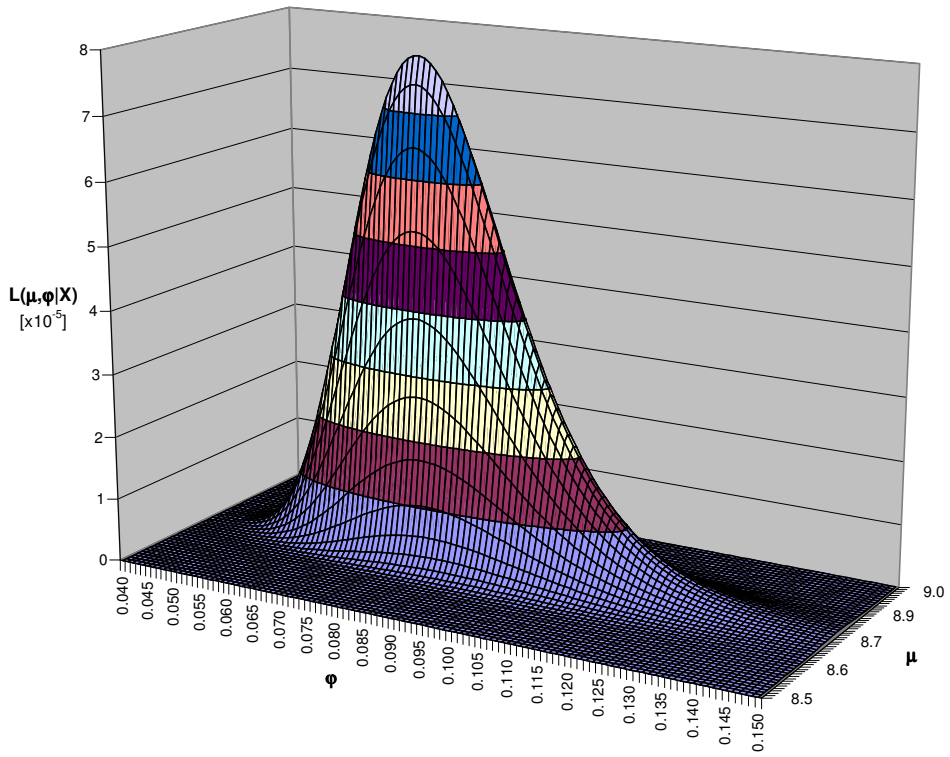


FIGURA 4.4: Função de verossimilhança formulada para a análise bayesiana de freqüência de vazões máximas anuais no rio São Francisco em Manga.

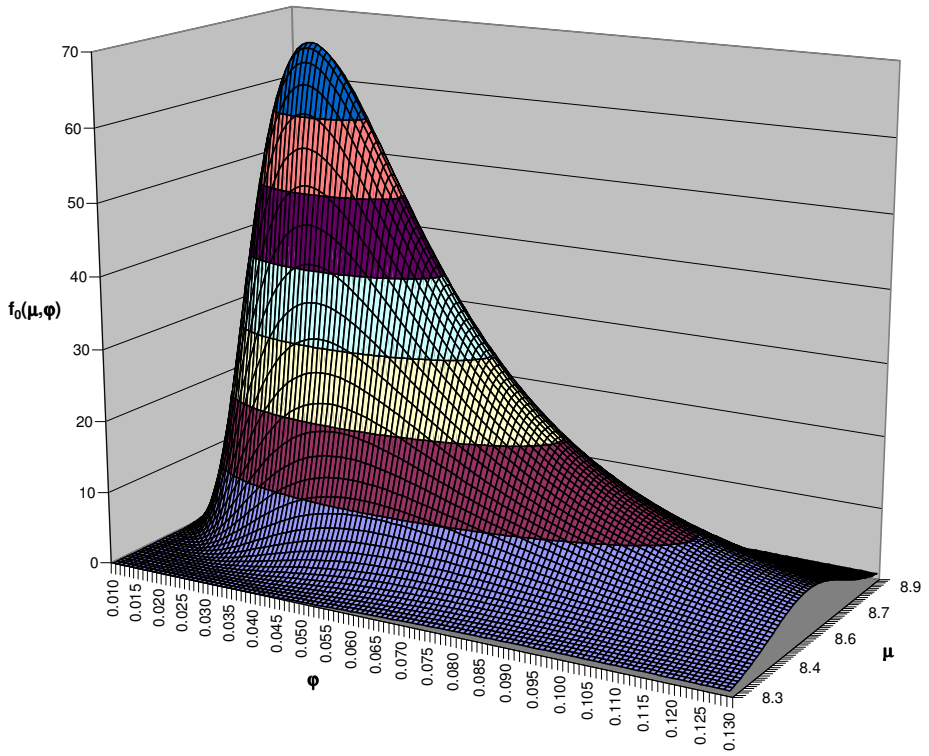


FIGURA 4.5: Distribuição *a priori* formulada para a análise bayesiana de freqüência de vazões máximas anuais no rio São Francisco em Manga.

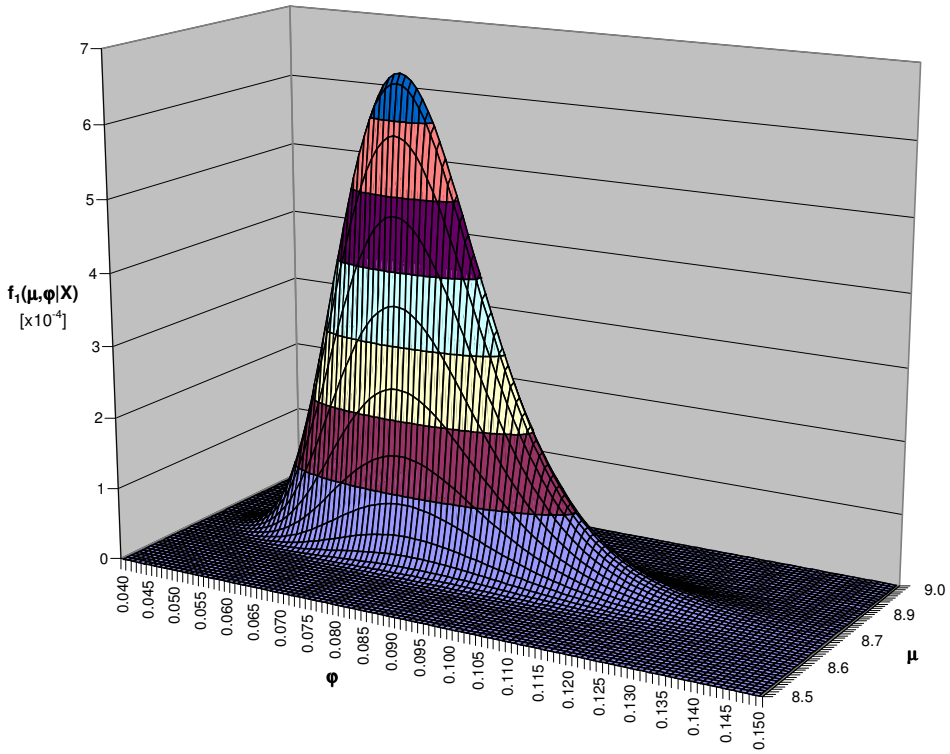


FIGURA 4.6: Distribuição *a posteriori* formulada para a análise bayesiana de freqüência de vazões máximas anuais no rio São Francisco em Manga.

A distribuição bayesiana dos logaritmos das vazões máximas anuais pode ser encontrada usando a equação (4.17), e, para a distribuição *a posteriori* formulada anteriormente, será expressa por:

$$\frac{\ln(q)-8{,}7571}{\sqrt{0{,}0792/(70/71)}} \sim t_{69}$$

$$\frac{\ln(q)-8{,}7571}{0{,}2834} \sim t_{69}$$

A Tabela 4.1 apresenta os períodos de retorno, associados aos quantis de interesse, para a distribuição bayesiana (considerando a distribuição *a priori* informativa ou não informativa) e para a distribuição clássica (com os parâmetros estimados pelo método dos momentos). As respectivas curvas de quantis são mostradas na Figura 4.7.

TABELA 4.1: Resultado da análise bayesiana de freqüência de vazões máximas anuais no rio São Francisco em Manga, usando a distribuição Log-Normal 2 parâmetros.

Vazão (m^3/s)	Tempo de retorno (anos)		
	AF clássica (MOM)	AF bayesiana (Distr. *a priori* não informativa)	AF bayesiana (Distr. *a priori* informativa)
2000	1.0000	1.0001	1.0001
3000	1.0034	1.0046	1.0050
4000	1.0477	1.0518	1.0565
5000	1.2253	1.2307	1.2503
6000	1.6641	1.6674	1.7237
7000	2.5902	2.5813	2.7235
8000	4.4833	4.4205	4.7652
9000	8.3834	8.1097	8.9329
10000	16.5515	15.5353	17.4824
11000	33.9167	30.4922	35.0457
12000	71.2362	60.4988	70.9958
13000	151.9601	120.1959	143.9816
14000	327.0390	237.5685	290.4314
15000	706.6150	465.0772	580.1398
16000	1527.2592	899.1318	1144.2129
17000	3293.3126	1713.4326	2224.0914
18000	7071.1545	3214.8713	4255.7737
19000	15095.8345	5935.4027	8011.6327
20000	32009.1888	10780.1797	14834.6611

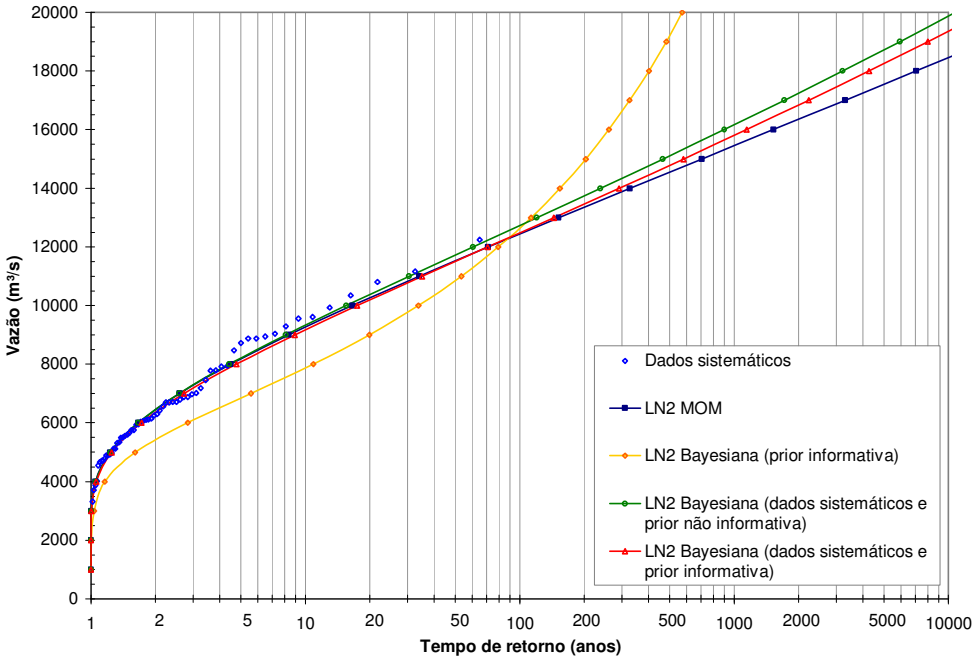


FIGURA 4.7: Análise bayesiana de freqüência de vazões máximas anuais no rio São Francisco em Manga, usando a distribuição Log-Normal 2 parâmetros.

Pode-se notar que, na região em que a distribuição resultante do uso exclusivo das informações avaliadas *a priori* apresenta um comportamento coerente com os dados amostrais (digamos, até cerca de 100 anos de tempo de retorno), a distribuição bayesiana obtida combinando-se os dados sistemáticos e a distribuição *a priori* informativa é o resultado de uma ponderação entre as distribuições formuladas usando apenas as informações *a priori* ou apenas os dados sistemáticos (isto é, considerando uma distribuição *a priori* não informativa). É interessante perceber ainda que, ao se levar em conta a incerteza relativa aos parâmetros, vazões maiores são obtidas para o mesmo tempo de retorno. Em geral, quando a mesma informação é utilizada, a abordagem bayesiana conduz a uma distribuição mais conservadora.

4.4 Redução da incerteza na análise de freqüência devido ao uso da teoria bayesiana

A redução da incerteza devido ao uso da teoria bayesiana está ligada ao fato de que essa abordagem permite considerar adequadamente a incerteza estatística envolvida em uma análise de freqüência. Assim, pode-se dizer que a abordagem bayesiana conduz, na verdade, a uma melhor descrição da incerteza. Ao se considerar os parâmetros do modelo probabilístico

como variáveis aleatórias, eles podem ser modelados, inicialmente, por uma distribuição *a priori* (representando algum conhecimento avaliado num primeiro momento sobre Θ) e, após a análise dos dados fluviométricos, por uma distribuição *a posteriori*. Os dados fluviométricos são representados por uma amostra de dados sistemáticos e, caso tenham sido encontrados, dados não sistemáticos de cheias. Quando não estiverem disponíveis informações confiáveis sobre Θ, pode-se adotar uma distribuição *a priori* não informativa, e a distribuição *a posteriori* será proporcional à própria função de verossimilhança dos parâmetros, dada a amostra observada.

A distribuição *a posteriori* pode, então, ser usada para a obtenção de uma distribuição marginal de *q*, resultante de uma integração em todo o domínio de Θ. Essa distribuição, livre das incertezas associadas à estimação dos parâmetros, é chamada distribuição bayesiana. A distribuição bayesiana permite fazer inferências sobre *q*, levando-se em conta a incerteza sobre o verdadeiro valor populacional de Θ e a incerteza residual sobre *q*, quando Θ é conhecido (LEE, 1989).

4.5 O software FLIKE: uma ferramenta computacional para análise bayesiana de freqüência de variáveis hidrológicas

Kuczera (1999) desenvolveu o *software* FLIKE, que permite utilizar, na análise bayesiana de freqüência de vazões máximas anuais, dados sistemáticos, dados não sistemáticos e informações avaliadas *a priori* sobre os parâmetros. Cinco distribuições de probabilidades freqüentemente empregadas em hidrologia foram implementadas no FLIKE: Log-Normal 2 parâmetros, Gumbel, Log-Pearson III, Generalizada de Valores Extremos e Generalizada de Pareto. Segue uma breve descrição dos procedimentos incorporados nesse *software*, o qual encontra-se disponibilizado a partir da URL http://www.eng.newcastle.edu.au/~cegak/.

4.5.1 Formulação das distribuições *a priori* e *a posteriori* dos parâmetros

Uma vez que o FLIKE dispõe de várias distribuições de probabilidades para a análise de freqüência de cheias, a distribuição *a posteriori* dos parâmetros, ao contrário do que foi exposto nos itens 4.3.2 e 4.3.3, é obtida por meio da resolução numérica da equação (4.5), com $L(\Theta|Q)$ e $f_0(\Theta)$ formuladas como apresentado a seguir.

Considere uma amostra de N_S vazões máximas anuais (dados sistemáticos) e *t* conjuntos de dados não sistemáticos censurados, cada um deles caracterizado por um período histórico de N_{Hj} anos, nos quais um limiar de referência y_{Hj} foi excedido k_j vezes, por eventos cuja magnitude não se pode estimar com precisão (amostra binomial censurada). Nessa situação, conforme discutido no item 3.5.2, a função de verossimilhança dos parâmetros é dada por:

$$L(\Theta \mid Q) = \prod_{i=1}^{N_S} f(q_i \mid \Theta) \cdot \prod_{j=1}^{t} \left\{ \binom{N_{Hj}}{k_j} \cdot [1 - F(y_{Hj} \mid \Theta)]^{k_j} \cdot [F(y_{Hj} \mid \Theta)]^{N_{Hj} - k_j} \right\} \tag{4.18}$$

onde $f(q \mid \Theta)$ é a função densidade de probabilidade e $F(y_H \mid \Theta) = P(q < y_H \mid \Theta)$ é a função acumulada de probabilidade das vazões máximas anuais.

A distribuição de probabilidades *a priori* dos parâmetros é representada por uma distribuição Normal multivariada, tal que $\Theta \sim N(\mu_\Theta, \Sigma_\Theta)$, onde μ_Θ é o vetor das médias e Σ_Θ é a matriz de covariância dos parâmetros, a qual pode ser expressa por:

$$\Sigma_\Theta = R_\Theta \cdot \sigma_\Theta \cdot \sigma_\Theta^T \tag{4.19}$$

Na equação (4.19), R_Θ é a matriz dos coeficientes de correlação e σ_Θ é o vetor dos desvios-padrão dos parâmetros. Para um modelo distributivo de três parâmetros, tem-se:

$$\mu_\Theta = \begin{bmatrix} m_{\theta_1} \\ m_{\theta_2} \\ m_{\theta_3} \end{bmatrix} \qquad \sigma_\Theta = \begin{bmatrix} s_{\theta_1} \\ s_{\theta_2} \\ s_{\theta_3} \end{bmatrix} \qquad R_\Theta = \begin{bmatrix} 1 & r_{\theta_1\theta_2} & r_{\theta_1\theta_3} \\ r_{\theta_2\theta_1} & 1 & r_{\theta_2\theta_3} \\ r_{\theta_3\theta_1} & r_{\theta_3\theta_2} & 1 \end{bmatrix} \tag{4.20}$$

O FLIKE permite realizar a análise bayesiana de freqüência utilizando uma distribuição de probabilidades *a priori* informativa ou não informativa. Quando, inicialmente, não se dispõe de informações relevantes sobre os parâmetros, μ_Θ, σ_Θ e R_Θ são expressos da seguinte forma:

$$\mu_\Theta = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \qquad \sigma_\Theta = \begin{bmatrix} \infty \\ \infty \\ \infty \end{bmatrix} \qquad R_\Theta = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.21}$$

4.5.2 Determinação da distribuição bayesiana da variável hidrológica

No *software* FLIKE, a distribuição bayesiana de probabilidades $\tilde{f}(q)$ é representada pela chamada distribuição da probabilidade esperada (ProE), que consiste na determinação do valor esperado da probabilidade de excedência do quantil q_T, associado ao tempo de retorno *T*. Seja $P(q \geq q_T)$ definida por:

$$P(q \geq q_T) = \int_{q_T}^{+\infty} \tilde{f}(q)\, dq = \frac{1}{T} \tag{4.22}$$

Como $\tilde{f}(q)$ é dado pela equação (4.9), pode-se reescrever a equação (4.22) da seguinte forma:

$$P(q \geq q_T) = \int_{\Theta} \left(\int_{q_T}^{+\infty} f(q \mid \Theta)\, dq \right) \cdot f_1(\Theta \mid Q)\, d\Theta$$

$$P(q \geq q_T) = \int_{\Theta} P(q \geq q_T \mid \Theta) \cdot f_1(\Theta \mid Q)\, d\Theta \tag{4.23}$$

onde $P(q \geq q_T \mid \Theta)$ representa a probabilidade de excedência do quantil q_T, dado o vetor de parâmetros Θ, e $P(q \geq q_T)$ denota o valor esperado da probabilidade de excedência do quantil q_T, levando-se em conta todos os valores possíveis de Θ.

Para integrar a equação (4.23), utiliza-se uma técnica de Monte Carlo que permite a geração de amostras aleatórias de uma distribuição de probabilidades. Essa técnica, conhecida como *importance sampling*, produz uma amostra de *n* vetores de parâmetros Θ_i e seus respectivos pesos normalizados, expressos por $p_i = w_i \Big/ \sum_{j=1}^{n} w_j$. Para sua aplicação, é necessário definir uma função densidade de probabilidade $I(\Theta)$ (*importance probability density function*), capaz de prover uma aproximação razoável de $f_1(\Theta \mid Q)$, tal que $w_i = f_1(\Theta_i \mid Q)/I(\Theta_i)$. Kuczera (1999) demonstrou que, expressando-se $I(\Theta)$ como uma distribuição Normal multivariada conveniente, é possível alcançar aproximações satisfatórias de $f_1(\Theta \mid Q)$, à medida que o período de dados sistemáticos se aproxima de 50 anos. A função $I(\Theta)$, dada pela equação (4.24), pode ser avaliada mediante uma aproximação de segunda ordem da

distribuição $f_1(\Theta|Q)$, o que resulta em uma distribuição Normal multivariada, cuja matriz de covariância deve ser multiplicada por uma fator de escala γ (com valores típicos na faixa de 1,5 a 2,5), para assegurar que $I(\Theta)$ produza amostras representativas de toda a distribuição $f_1(\Theta|Q)$.

$$I(\Theta) \sim N[\hat{\Theta}, \gamma^2 \Sigma_\Theta] \tag{4.24}$$

No *software* FLIKE, a aproximação dada por $I(\Theta)$ pode ser verificada visualmente, como exemplificado na Figura 4.8. Essa figura mostra a superfície da distribuição *a posteriori* dos parâmetros de posição e forma da distribuição GEV, com o parâmetro de escala fixo em seu valor mais provável. A elipse tracejada é o limite da região de 90% de probabilidade, baseada na aproximação Normal multivariada da equação (4.24). O contorno marcado pela letra "R" engloba a região de 90% de probabilidade da distribuição *a posteriori* $f_1(\Theta|Q)$. Se a aproximação dada por $I(\Theta)$ for adequada, a elipse tracejada e o contorno "R" devem ser virtualmente coincidentes.

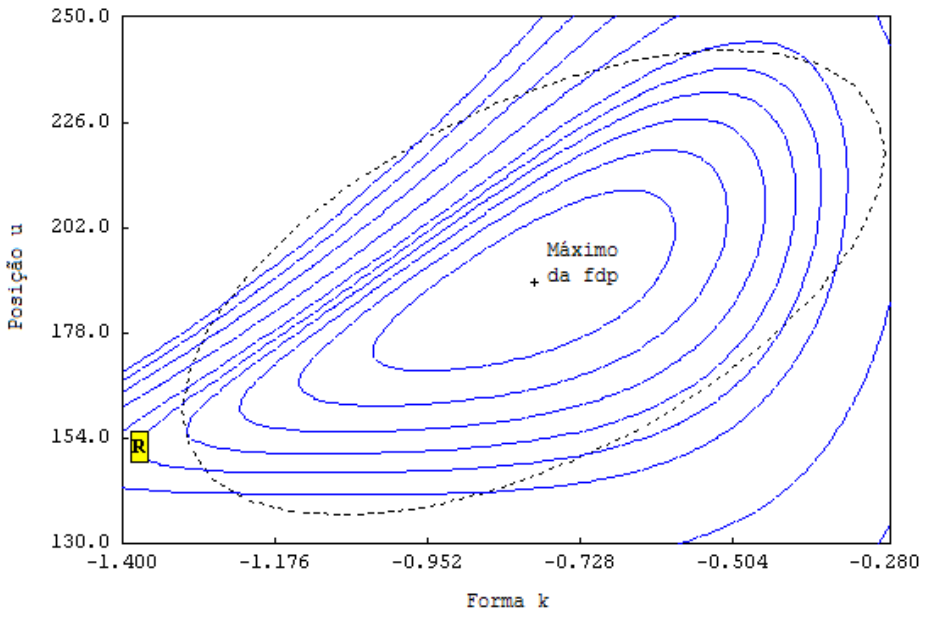


FIGURA 4.8: Superfície da distribuição condicional *a posteriori* dos parâmetros de posição e forma da GEV e região de 90% de probabilidade da aproximação Normal multivariada.

Utilizando a técnica *importance sampling*, os quantis da distribuição da probabilidade esperada, correspondente à equação (4.23), podem ser estimados por:

$$\hat{P}(q \geq q_T) = \sum_{i=1}^{n} P(q \geq q_T \mid \Theta_i) \cdot p_i \tag{4.25}$$

onde:

$$p_i = w_i \Big/ \sum_{j=1}^{n} w_j$$

$$w_i = f_1(\Theta_i \mid Q) \Big/ I(\Theta_i)$$

A Figura 4.9 mostra a distribuição da probabilidade esperada, calculada usando o *software* FLIKE, para o exemplo do item 4.3.3.1. As curvas apresentadas se referem à situação em que nenhuma informação *a priori* sobre os parâmetros foi considerada. Como era esperado, a distribuição bayesiana calculada numericamente pelo FLIKE coincide com a distribuição obtida analiticamente.

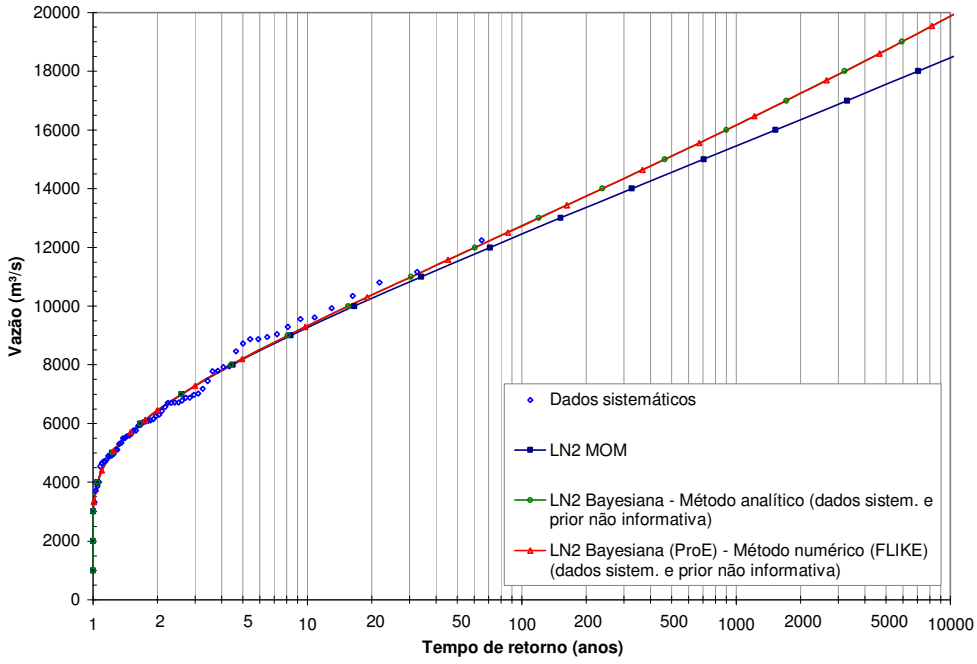


FIGURA 4.9: Análise bayesiana de freqüência de vazões máximas anuais no rio São Francisco em Manga, usando a distribuição Log-Normal 2 parâmetros e o *software* FLIKE.

4.5.3 Estimação pontual bayesiana e cálculo dos intervalos de credibilidade dos quantis

No *software* FLIKE, foi implementado também um procedimento de estimação pontual bayesiana, que permite obter a chamada distribuição dos parâmetros esperados (ParE). A técnica *importance sampling* é novamente utilizada, possibilitando a resolução da equação (4.8) por meio da seguinte expressão:

$$\Theta^* = E[\Theta] = \sum_{i=1}^{n} \Theta_i \cdot p_i \tag{4.26}$$

onde:

$$p_i = w_i \Big/ \sum_{j=1}^{n} w_j$$

$$w_i = f_1(\Theta_i \mid Q) / I(\Theta_i)$$

A distribuição $f(q \mid \Theta^*)$ será, então, a distribuição dos parâmetros esperados, cujos quantis de tempo de retorno *T* são denotados por $q_T(\Theta^*)$. Os intervalos de $100(1-\alpha)\%$ de credibilidade (ou intervalos de confiança bayesianos) são calculados ordenando-se os *n* quantis $q_T(\Theta_i)$ e encontrando-se aqueles cujas probabilidades de excedência são aproximadamente $\alpha/2$ e $1-(\alpha/2)$. Esses quantis serão, respectivamente, o limite superior e o limite inferior do intervalo.

5 ESTUDO DE CASO: BACIA DO RIO SÃO FRANCISCO EM SÃO FRANCISCO

5.1 Introdução

Nesse capítulo, um estudo de caso é desenvolvido com o intuito de aplicar os conhecimentos referentes à utilização dos dados não sistemáticos e da abordagem bayesiana na análise de freqüência de vazões máximas anuais. Para isso, foi escolhida a bacia do rio São Francisco, uma das mais importantes do País, e que, por ter sido explorada desde a época do Brasil Império, apresenta grande potencial relativo à disponibilidade de informações sobre cheias históricas.

A metodologia empregada nessa pesquisa consiste na realização, por meio do *software* FLIKE, da análise bayesiana de freqüência de vazões máximas anuais no rio São Francisco em São Francisco, utilizando-se: (1) dos registros sistemáticos de vazão do referido posto fluviométrico, (2) de informações sobre cheias históricas coletadas na região, e (3) de uma distribuição de probabilidades *a priori* informativa, tanto do ponto de vista da prescrição do modelo probabilístico, como das estimativas de seus parâmetros.

5.2 Caracterização da bacia hidrográfica do São Francisco

A bacia hidrográfica do São Francisco abrange 521 municípios e 7 unidades da Federação, a saber: Bahia (48,2% da área de drenagem), Minas Gerais (36,8%), Pernambuco (10,9%), Alagoas (2,3%), Sergipe (1,1%), Goiás (0,5%) e Distrito Federal (0,2%). Com 2.800 km de extensão, e drenando uma área de aproximadamente 641.000 km², o rio São Francisco nasce na Serra da Canastra, em Minas Gerais, e escoa no sentido sul-norte pela Bahia e Pernambuco, quando altera seu curso para sudeste, desembocando no Oceano Atlântico entre Alagoas e Sergipe. Devido à sua extensão e aos diferentes ambientes percorridos, a bacia hidrográfica está dividida em 4 trechos: Alto São Francisco, Médio São Francisco, Sub-Médio São Francisco e Baixo São Francisco (veja Figura 5.1).

O rio São Francisco atravessa regiões com condições naturais bastante diversas. Os extremos superior e inferior da bacia apresentam índices pluviométricos relativamente altos, enquanto os trechos Médio e Sub-Médio estão inseridos em ambientes mais secos. Assim, cerca de 75% do deflúvio do São Francisco é gerado em Minas Gerais, numa área correspondente a 37% da

área total da bacia. A área compreendida entre a fronteira Minas-Bahia e a cidade de Juazeiro (BA) representa 45% do vale e contribui com apenas 20% do deflúvio anual. Os ecossistemas observados na maior parte da bacia do São Francisco são o cerrado e a caatinga, verificando-se também fragmentos de mata atlântica (que ocorre no Alto São Francisco, principalmente nas cabeceiras) e de ambientes costeiros. No que se refere às características geológicas, os aluviões recentes, arenitos e calcários, que dominam boa parte da bacia, funcionam como verdadeiras esponjas, retendo água na estação chuvosa, e liberando-a nos meses de estiagem.

Essa bacia hidrográfica é de fundamental importância para o País, devido ao volume de água transportado num ambiente semi-árido, o que tem contribuído para o desenvolvimento econômico da região. Um aspecto significativo no cenário social e econômico da bacia se refere à agricultura, intimamente ligada à irrigação de grandes áreas. O rio São Francisco é utilizado também para outras finalidades, destacando-se a geração de energia, o abastecimento de água, a navegação, o turismo e a pesca.

Em grande parte do vale do São Francisco, as áreas mais propícias à utilização econômica situam-se às margens do rio e, por esse motivo, encontram-se densamente povoadas. Embora, durante vários anos, o São Francisco fique restrito à sua calha em boa parte da bacia, periodicamente ocorrem grandes cheias, nas quais as águas extravasam a calha do rio, inundando as áreas mais baixas e causando graves prejuízos sociais e econômicos à região.

O posto fluviométrico de São Francisco, localizado no trecho médio do rio São Francisco, em Minas Gerais, conforme ilustrado na Figura 5.1, possui uma área de drenagem de 182.537 km² e está em operação desde 1934 (sua série de vazões máximas anuais inicia com o evento do ano hidrológico de 1934/1935, ou seja, de 01/10/1934 a 30/09/1935). Situado cerca de 370 km a jusante do reservatório de Três Marias, com um incremento na área de drenagem da ordem de 130.000 km², estudos, tais como o realizado pela Comissão Interministerial de Estudos para Controle das Enchentes do Rio São Francisco (1980), demonstram a influência quase nula da operação desse reservatório nos hidrogramas das grandes cheias ocorridas em São Francisco. Ainda de acordo com o estudo da referida comissão, foi atribuída à barragem de Três Marias a função de proteger a cidade de Pirapora, situada às margens do rio São Francisco, contra enchentes de até 50 anos de recorrência. A existência de afluentes caudalosos e sem controle, a jusante de Pirapora, torna aquela barragem ineficiente para a proteção de outras cidades, tais como São Francisco.

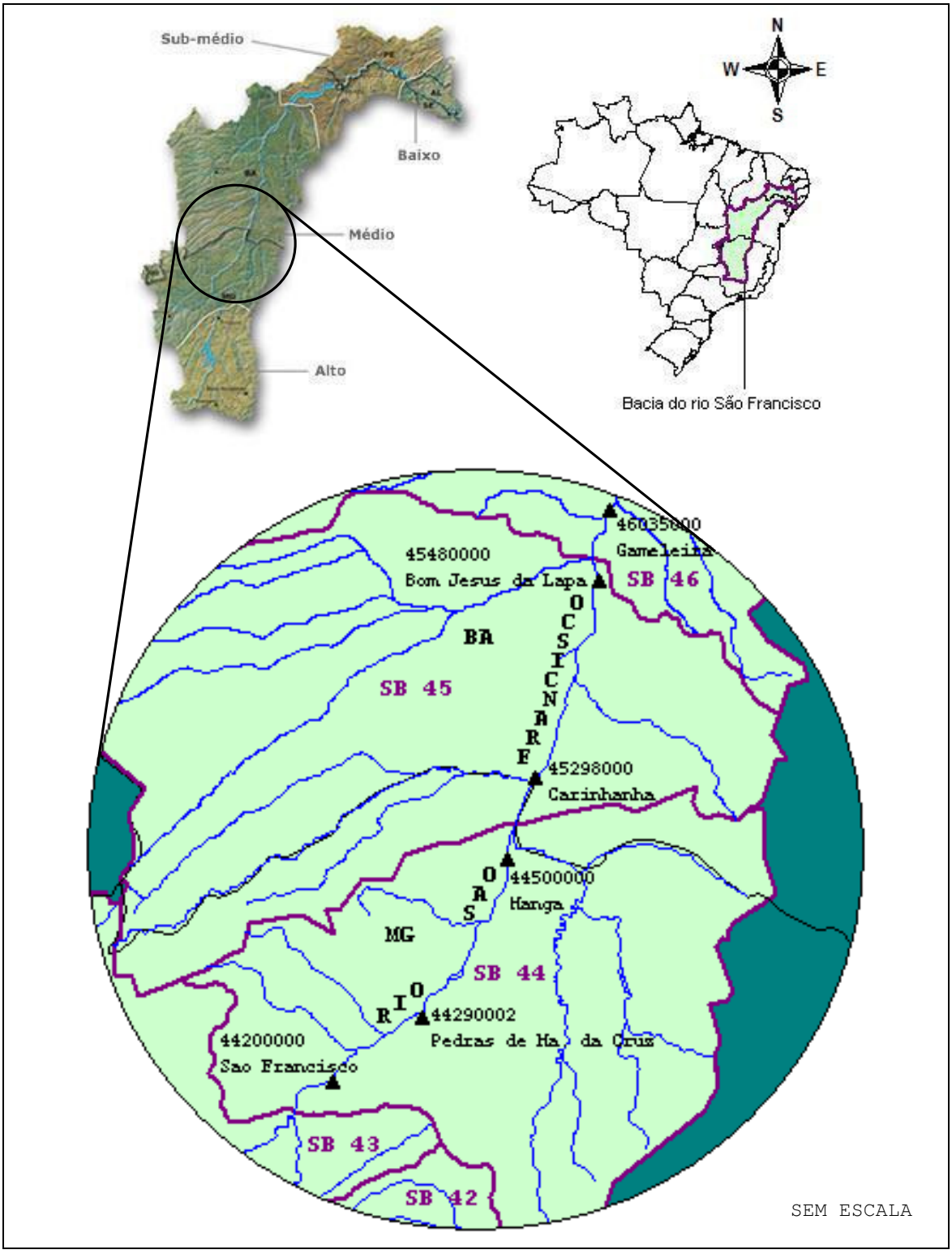


FIGURA 5.1: Bacia do rio São Francisco.
Fonte: adaptado de www.ana.gov.br.

5.3 Coleta de informações históricas sobre cheias na bacia do rio São Francisco

5.3.1 Aspectos da metodologia

A pesquisa por dados não sistemáticos de vazões de enchentes foi realizada por meio da busca de registros históricos sobre cheias em arquivos e jornais dos séculos XVIII, XIX e início do século XX, encontrados no Arquivo Público Mineiro e na Hemeroteca Histórica, destacando-se os relatórios da Comissão Geográfica e Geológica do Estado, produzidos no final do século XIX. Foram estudados também os registros feitos por aventureiros durante suas expedições científicas a diversas regiões do Brasil, especialmente à então Província das Minas Gerais, realizadas na primeira metade de século XIX. As principais referências consultadas são as obras dos naturalistas Auguste de Saint-Hilaire e José de Souza Azevedo Pizarro; e do geólogo Wilhelm Ludwig Von Eschwege. Nessas obras, os autores relatam detalhadamente suas viagens, apresentando informações de notável interesse para a História Natural, a Geologia, a Geografia, a Hidrologia, a Etnologia e, em menor escala, para a Economia. Maiores detalhes sobre essas fontes de informações históricas são apresentados no Anexo B. O relatório produzido pela Comissão Interministerial de Estudos para Controle das Enchentes do Rio São Francisco (1980), que apresenta detalhes sobre a enchente ocorrida naquele rio em 1979, também fornece informações relevantes sobre as maiores cheias do São Francisco.

Durante a busca por informações históricas sobre cheias, priorizou-se, a princípio, a bacia do rio das Mortes, havendo, nesse momento, a expectativa de que tais informações fossem mais facilmente encontradas devido à existência de cidades historicamente importantes (tais como Tiradentes e São João del Rei) na referida bacia. Entretanto, muito embora nos relatórios da Comissão Geográfica e Geológica do Estado alguns dados climatológicos (como temperatura média mensal, total mensal de precipitação e número de dias chuvosos no mês) tenham sido encontrados para locais e anos específicos, como, por exemplo, Diamantina (1894) e Uberaba (1895), constatou-se que, devido à forma como se apresentavam e às possíveis interferências antrópicas na bacia, inúmeras dificuldades seriam encontradas para transformar esses dados em informações passíveis de serem incorporadas à análise de freqüência de cheias, da maneira como foi descrito no Capítulo 3. Além disso, com o desenvolver da pesquisa, percebeu-se que a grande maioria das informações disponíveis se referiam a aspectos econômicos, sociais e culturais da região.

Optou-se, então, por estender a busca à bacia do rio São Francisco, especialmente às suas cabeceiras, onde se localiza o posto fluviométrico de Iguatama, um dos mais antigos da rede hidrométrica do Estado. A referida bacia foi explorada por diversos naturalistas, dentre os quais destacam-se Saint-Hilaire e Pizarro, que descreveram suas principais observações em ricas obras literárias.

Durante parte do período que permaneceu no Brasil (1816 a 1822), Saint-Hilaire percorreu, em sua viagem pelas províncias do Rio de Janeiro e Minas Gerais (relatada com detalhes na obra de mesmo nome), o trajeto mostrado na Figura 5.2. Nesse trajeto, que corresponde em grande parte ao *Caminho Novo* da *Estrada Real* (veja Figura 5.3), o referido naturalista segue do Rio de Janeiro ao *Distrito Diamantino* (onde hoje se encontra o município de Diamantina), explorando, também, os vales dos rios Doce e Jequitinhonha. Posteriormente, Saint-Hilaire dirigiu-se à bacia do rio São Francisco, alcançando sua margem direita na localidade denominada *Capão do Cleto*, de onde seguiu para jusante e montante, só abandonando a margem do rio para retornar à região do *Distrito Diamantino*. Algumas peculiaridades do itinerário percorrido na parte final da viagem são apresentadas nos trechos transcritos a seguir, extraídos de Saint-Hilaire (1975):

"(...)

Itinerário aproximado desde a entrada do sertão pelo lado do Pé do Morro até o Rio S. Francisco, passando por S. Eloi, Formigas e Contendas:

		léguas
Do Jequitinhonha a	*Taioba (choupana), cerca de .....*	*6½*
″	*Ribeirão (fazenda) .....*	*2*
″	*S. Eloi (fazenda) .....*	*4*
″	*Veados (ao relento) .....*	*2½*
″	*Pindaíba (lugar abandonado) .....*	*4*
″	*Formigas (povoação) .....*	*3*
″	*Veados (choupana), cerca de .....*	*3*
″	*Caiçara (habitação) .....*	*2½*
″	*Riachão (habitação) .....*	*4*
″	*Riacho de S. Lourenço (choupana) ...*	*3½*
De R. de S. Lour. a	*Contendas (povoação) .....*	*4*
De R. de S. Lour. a	*Tamanduá (pequena habitação) .....*	*3*
De R. de S. Lour. a	*Tapera (choupana) .....*	*3½*
De R. de S. Lour. a	*Capão do cleto, fazenda à margem do Rio S. Francisco .....*	*5*
	Total	*50½*

(...)"

"(...)

A cerca de duas ou três léguas da fazenda chamada Capão do Cleto*, desci por uma encosta íngreme a uma vasta planície. O calor tornou-se mais forte ainda do que nos dias precedentes; o céu não apresentava mais esse matiz puro e brilhante que admirei tantas vezes, mas estava carregado de vapores avermelhados. Perto de um retiro, situado a meia légua de Capão, a vegetação mudou bruscamente; a terra, gretada, estava ainda extremamente seca, mas era evidente que pouco antes estivera coberta de água; duas plantas espinhosas, que não encontrara ainda em parte alguma, formavam aqui e ali largas moitas: uma me pareceu idêntica a essa mimosa de flores amarelas e odoríferas, que se cultiva no Rio de Janeiro sob o nome de esponjeira, a outra era uma espécie de bauinia de folhas muito pequenas e flores esverdeadas (*Bauhinia inundata*, N.). Esses vegetais não tinham perdido o verdor como as árvores das caatingas, e a mimosa apresentava uma multidão de flores douradas em meio à folhagem de um verde escuro. Algumas espécies de pequenas aves que até então não se me tinham oferecido à vista, esvoaçavam sobre os arbustos. Tudo anunciava a influência de uma causa que ainda não percebia: cheguei, enfim, à habitação de Capão do Cleto, à margem do Rio S. Francisco, e reconheci que as mudanças que tanto me impressionavam eram devidas à vizinhança do rio.*

(...)"

"(...)

Foi, como já disse, em Capão do Cleto, que admirei pela primeira vez o Rio S. Francisco. Fui recebido hospitaleiramente pelo proprietário de Capão, o Sr. Capitão Cleto, e passei alguns dias em sua casa.

O Sr. Cleto descendia de um dos paulistas que primeiro se fixaram às margens do rio acima e abaixo de Capão. Estes não faziam parte dos bandos que fugiram no combate do Rio das Mortes. Eram dois primos, Matias Cardoso e Manoel Francisco de Toledo, homens poderosos, que, depois de matar um ouvidor, fugiram de sua pátria com a família e os escravos. Encontraram nos arredores de Capão uma tribo indígena, a dos Chicriabás ou Xicriabás; fizeram-lhes a princípio guerra; em seguida, porém, trataram com eles e firmaram pazes. O Rei concedeu aos dois primos a posse de uma e outra margem do Rio S. Francisco em toda a extensão que pudessem percorrer num dia, embarcando-se no rio, e além disso, deu a um dos primos o título de mestre de campo dos índios por duas gerações. Matias Cardoso e Manoel Francisco de Toledo tinham, ao que parece, reduzido grande número de índios à escravidão, como então se praticava; serviram-se desses infelizes para abrir fazendas e construir várias igrejas, entre outras a de Morrinhos, de que já falei. Foi a supressão da escravidão dos índios que deu o primeiro golpe na prosperidade de ambas as famílias. Pouco a pouco foram vendendo suas imensas possessões, e seu descendente, o Capitão Cleto, só me pareceu possuir medíocres bens. Ignorava em que ano Cardoso e Toledo tinham chegado às margens do S. Francisco; mas encontrara em papéis de família uma carta datada de 1712, que um dos primos escrevia ao outro das próprias margens do rio.

(...)"

"(...)

Deixei Capão do Cleto para dirigir-me a Salgado, que está situado mais abaixo, do lado oposto do S. Francisco. O caminho não se estende pela própria margem do rio; todavia é-lhe paralelo. Até a povoação de Pedras de Baixo (atual município de Pedras de Maria da Cruz)*, atravessei algumas vezes terrenos cobertos das duas plantas espinhosas a que já me referi; e, de tempos em tempos, avistava no meio desses terrenos lagos habitados por grande número de aves aquáticas, no meio das quais se distinguiam sempre as garças brancas, os jaburus e as colhereiras.*

(...)"

"(...)

Salgado *ou* Brejo do Salgado *é a sede de uma paróquia que tem quarenta léguas de comprimento por vinte de largura, cuja população ascende a oito mil almas, e que se estende pela margem do S. Francisco até o Rio* Carunhanha*, limite entre as províncias de Minas Gerais e de Pernambuco.*

(...)"

"(...)

Salgado, onde encontrei a hospitalidade mais acolhedora, foi o termo de minha viagem pela parte setentrional do sertão. Deixei essa povoação aos 29 de agosto de 1817, a fim de me dirigir ao Distrito Diamantino, e, até Capão, segui o caminho por onde passara alguns dias antes.

(...)"

"(...)

Itinerário aproximativo de Salgado ao Distrito Diamantino, passando pelas povoações de Pedras dos Angicos (atual município de São Francisco) *e Coração de Jesus:*

		léguas
De Salgado a	*Pedras de Baixo, povoação*	*3½*
″	*Capão do Cleto, habitação*	*5*
″	*Riachão de Cana Brava, hab.*	*6*
″	*Pedras dos Angicos, povoação*	*2*
″	*Logrador, habitação*	*5*
″	*Canoas,* id.	*3*
″	*Macaúba,* id.	*3*
″	*Rancharia,* id.	*3*
″	*Santa Clara,* id.	*3*
″	*Coração de Jesus, povoação*	*4*
″	*S. Bento, habitação*	*3*
″	*Buraco,* id.	*3½*
″	*Boa Vista, cabana*	*3*
″	*Fazenda do Negro, habitação*	*3*
″	*Sucuriu, cabana*	*5*
″	*Catônio, habitação*	*6*
″	*Curmataí, povoação*	*5*
″	*Serviço dos Diamantes do Rio Pardo, primeira lavra diamantina*	*6*
	Total	*72*

(...)"

"(...)

Tivera, a princípio, a intenção de seguir o S. Francisco até a confluência do Rio das Velhas, por um espaço de trinta e sete léguas, a partir de Riachão de Cana Brava: como, porém, encontrava apenas duas ou três plantas por dia, meus empregados e eu nos fatigávamos inutilmente, e as próprias montarias sofriam muito com a seca, decidi-me a não ir além de Pedras dos Angicos, e tomar um caminho que me conduzisse diretamente até o Tijuco. Devo lamentar, todavia, não ter visto a povoação de Barra, que está situada na confluência do Rio das Velhas, (...). Talvez deva lastimar não ter podido visitar o julgado de S. Romão, que se acha à margem esquerda do S. Francisco, a doze ou quinze léguas de Angicos, (...).

(...)"

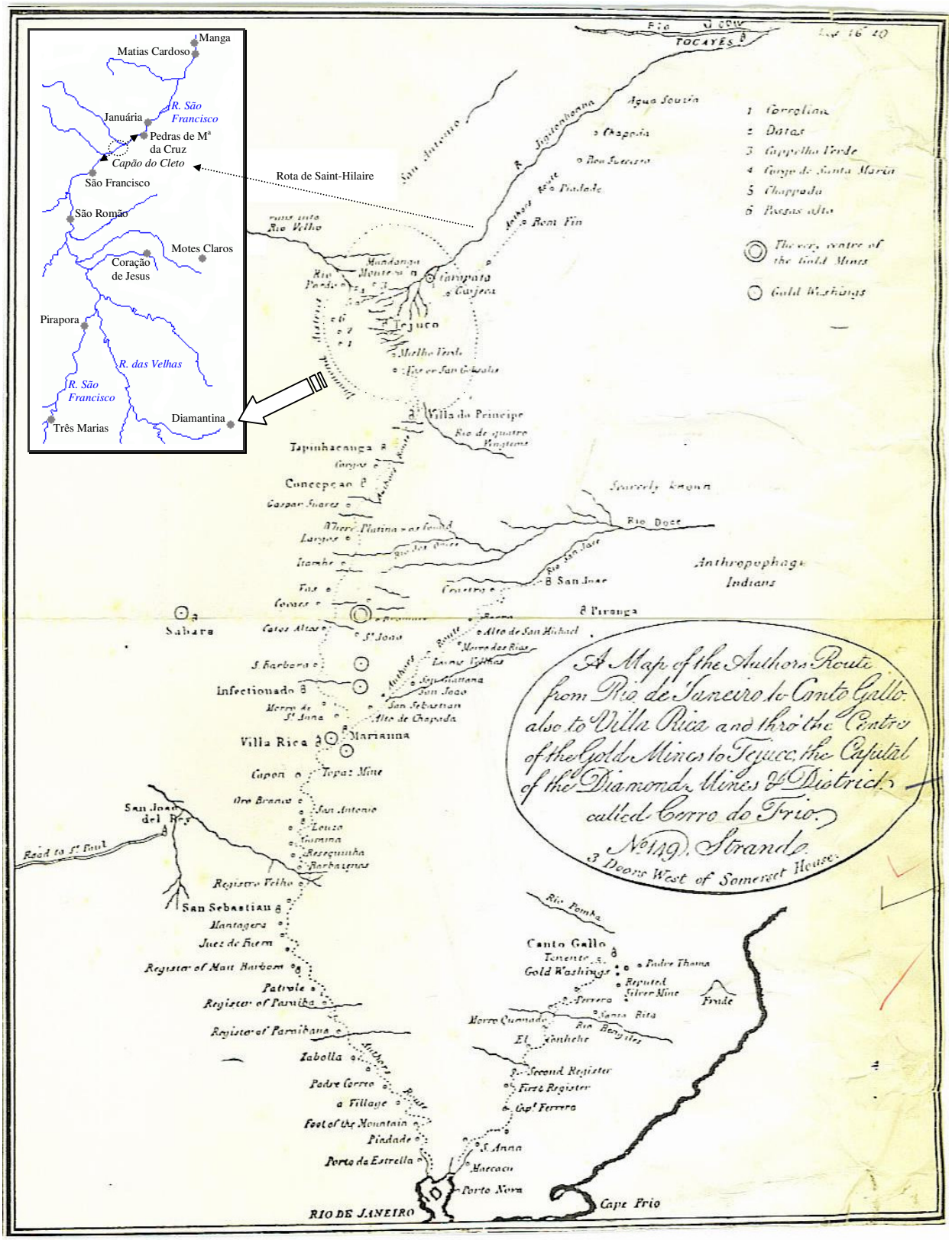


FIGURA 5.2: Trajeto da viagem realizada por Saint-Hilaire pelas províncias do Rio de Janeiro e Minas Gerais.

Fonte: adaptado de SAINT-HILAIRE, 1975.

FIGURA 5.3: Mapa da Estrada Real.
Fonte: www.descubraminas.com.br.

5.3.2 Apresentação das informações históricas de cheias

A Tabela 5.1 apresenta as principais informações relativas às enchentes históricas no rio São Francisco, encontradas durante a pesquisa. Utilizando essas informações e considerando algumas hipóteses plausíveis, é possível construir uma amostra binomial censurada de vazões máximas anuais, que será usada na análise bayesiana de freqüência de cheias.

TABELA 5.1: Informações sobre cheias históricas no rio São Francisco.

Ano	Comentários
-	Saint-Hilaire (1975) afirma, em um dos relatos de suas viagens ocorridas entre 1816 e 1822, o seguinte: *"Na estação das chuvas, que começa pelos fins de setembro e dura até janeiro, o rio* (São Francisco) *engrossa pouco a pouco; acaba por transbordar e, nos lugares em que a Serra não fica a muito pequena distância, as águas se estendem por uma légua e mais até. Em Capão do Cleto, cobrem meia légua de terreno, quer dizer, vão até esse retiro, onde reconheci que a vegetação mudava de natureza. A margem esquerda, mais elevada que a direita, é geralmente menos exposta às inundações.*"
-	Ainda no mesmo relato, Saint-Hilaire (1975) afirma também que: *"Não poderia dizer a que distância da nascente do S. Francisco começa a inundação desse rio, nem se ela se estende até a foz; entretanto, uma passagem do* Brasil, Novo Mundo*, de Eschwege, prova que a inundação já se faz sentir no Porto de S. Miguel, que deve estar situado, penso, entre os 19 e 20º de lat. S.; e parece, por um artigo do* Journal *do mesmo autor, que a região das salinas também fica inundada."*
-	Eschwege (1996), nos relatos da viagem ocorrida em 1816, diz: *"O rio São Miguel e os seus tributários originam-se na serra calcária, e com a água calcária irrigam a região, boa parte da qual é inundada durante a estação chuvosa. (...). Nesse ponto, o rio* (São Francisco) *abriu um corte de 30 pés em nível inferior ao da margem. Sua largura é apenas de 50 a 60 passos, mas é profundo. Por ocasião das chuvas continuadas, transborda e inunda grandes extensões, tornando a travessia muito perigosa."*
1773	Pizarro (1948), nas memórias de uma das suas viagens, relata o seguinte: *"Supera êste rio de S. Francisco a todos os da capitania* (de Minas Gerais) *na soberba com que eleva as águas fora do seu leito, quando as inundações o volumam; pois que chega a estender-se espraiando por mais de seis léguas, e às vezes muito além delas, como aconteceu no ano 1773, em que passou a mais de vinte, cobrindo as fazendas distantes das suas margens dez léguas, e levando consigo a maior parte do gado que povoava os campos. Por êle navegam as barcas condutoras do sal fabricado nos sertões de Pernambuco, de que se utilizam os povos mineiros. Abunda de tôda qualidade de peixe, principalmente de Surubís, e Dourados os mais monstruosos."*
1712 1790	Em outra parte, mas ainda nos mesmos relatos, Pizarro (1948) diz: *"Fronteira ao Arraial está uma ilha, que se diz de S. Rumão, com meia légua de comprido e quase 400 passos geométricos de largo, onde consta por tradição constante, e não controvertida, que houve uma aldeia de índios, os quais a desampararam depois de destroçados (...) em dia de S. Rumão. Não havendo certeza do ano dêsse fato, sabe-se contudo, que fôra antes de 1712, porque esta época é bem conhecida dos mais antigos habitantes do país, entre quem ficou memorável a grande enchente do rio* (São Francisco)*, que no ano referido houve, bem como a do ano 1790, que fêz outra época, excedendo a primeira. Daquele acontecimento, em dia assinalado, teve origem a denominação do distrito, e lugar, intitulado de S. Rumão."*

Ano	Comentários
1919 1926	No relatório da Comissão Interministerial de Estudos para Controle das Enchentes do Rio São Francisco (1980), encontra-se a seguinte afirmação: *"A enchente de 1979, ocorrida no rio São Francisco, conforme informações obtidas no médio, sub-médio e baixo São Francisco, deve ser considerada, em magnitude, a terceira mais importante do século, sendo a maior aquela ocorrida em 1919, seguida pela de 1926. Cabe ressalvar que, no período de observações hidrométricas regulares, que teve início em 1925, ela estava classificada em segundo lugar."*
-	Ainda de acordo com o relatório da referida comissão: *"É no médio São Francisco que se observam os maiores transbordamentos – 2 a 18 vezes o seu leito, representando largura média de 9 km, atingindo 16 km na região de Januária* (MG) *e chegando mesmo à enorme largura de cerca de 14 léguas (84 km) próximo de Xique-Xique* (BA) *– e também as vazantes mais lentas, com permanência de mais de 120 dias."*

Dos três primeiros relatos mostrados na Tabela 5.1, pode-se perceber que as enchentes do São Francisco, muitas vezes associadas ao extravasamento da calha principal, eram observadas em grande parte da bacia, desde a "*região das salinas*", correspondente ao sertão pernambucano e norte baiano, passando por "*Capão do Cleto*", já no norte mineiro, e estendendo-se até o "*Porto de S. Miguel*", no Alto São Francisco. Como pode ser visto no primeiro relato, em Capão do Cleto, inundações de meia légua (3 km) na margem direita eram comuns durante a estação chuvosa, e, portanto, não constituem um limiar de referência adequado para a construção da amostra a ser utilizada na análise de freqüência de cheias. Com o intuito de estimar a vazão correspondente à referida inundação, foi realizada uma simulação hidráulica no trecho entre São Francisco e Pedras de Maria da Cruz, usando o modelo HEC-RAS (HEC, 2002). Nesse processo, algumas simplificações foram adotadas, a saber:

1. Para a calha principal do rio, as seções transversais de montante e jusante são aquelas utilizadas como seções medidoras pela Agência Nacional de Águas (ANA), nos postos fluviométricos de São Francisco e Pedras de Maria da Cruz. Além disso, empregando o balizamento feito em 2000 pela Associação da Hidrovia do São Francisco (AHSFRA), foi construída uma seção intermediária aproximada em Capão do Cleto. Já as planícies de inundação, foram estimadas com base em cartas topográficas da região (veja Figura 5.5).

2. A declividade foi considerada constante ao longo do trecho, e o valor adotado (6 cm/km) foi estimado pela declividade média da linha d'água, em condições tais que se pudesse assumir escoamento permanente uniforme.

3. No trecho em questão, assumiu-se pequena variação do coeficiente de rugosidade de uma seção para outra. Para as seções de São Francisco e Pedras de Maria da Cruz, o coeficiente de rugosidade da calha principal foi calibrado, utilizando os dados das curvas-chave dos referidos postos fluviométricos. Os valores encontrados foram, respectivamente, 0,023 e 0,025. Para as planícies de inundação, o valor adotado em todas as seções foi 0,040, tendo em vista a vegetação predominantemente rasteira da região.

4. A simulação foi realizada em regime de escoamento permanente, considerando-se que, em um certo intervalo de tempo, durante a passagem do pico da cheia, as condições de vazão ao longo do trecho não se alteraram significativamente.

As Figuras 5.4 e 5.5 apresentam, respectivamente, o trecho do rio São Francisco, modelado para a realização da simulação hidráulica, e as seção transversais utilizadas.

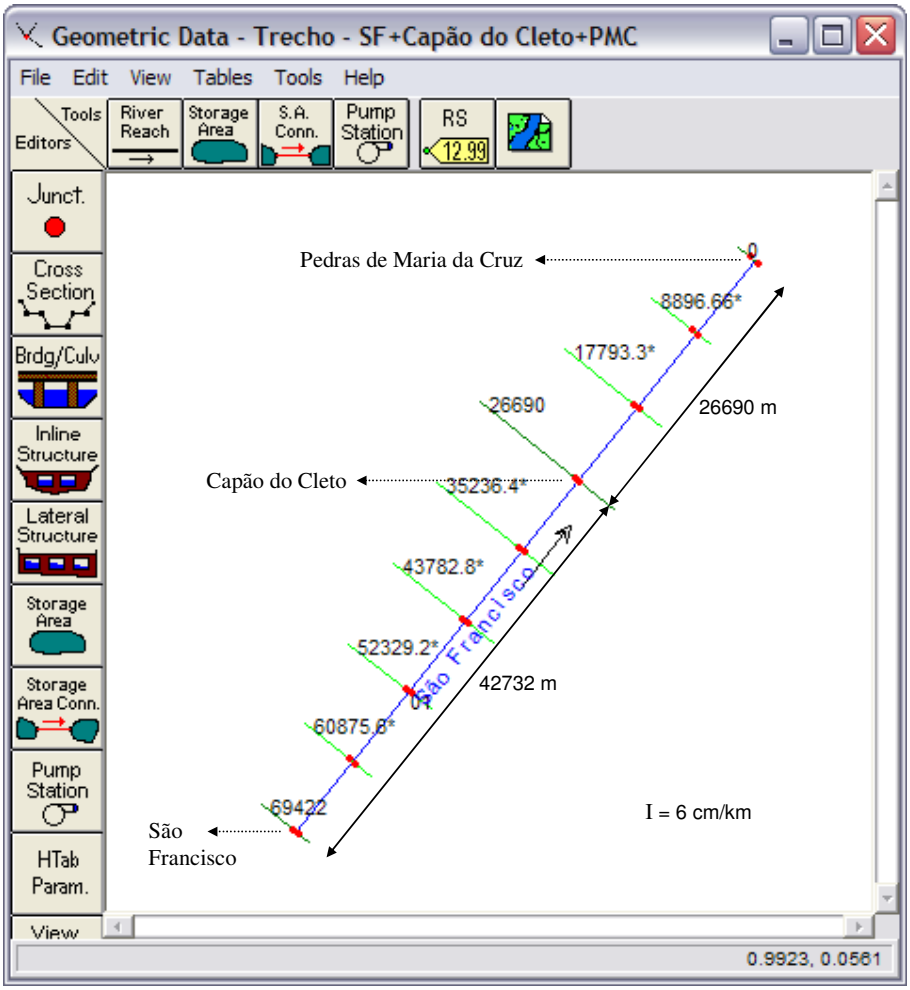
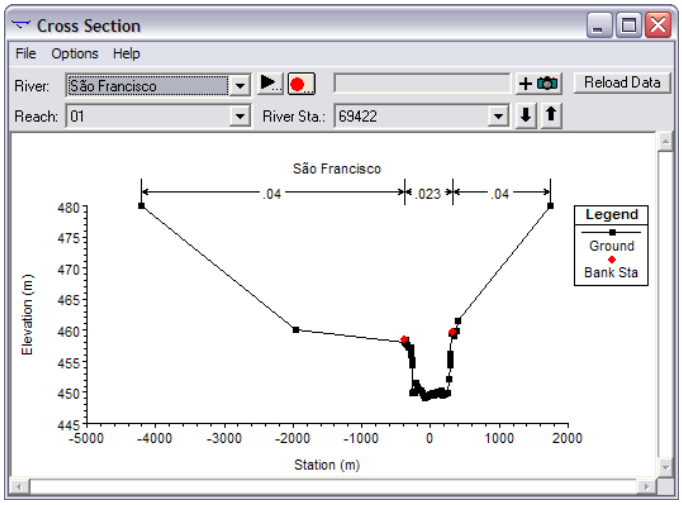
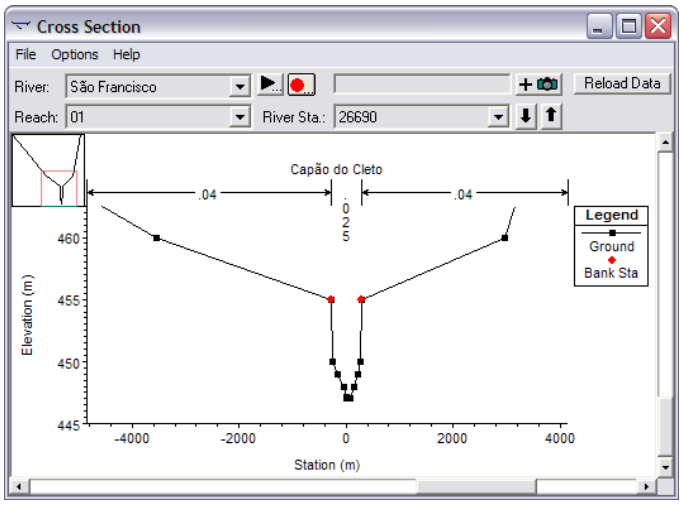


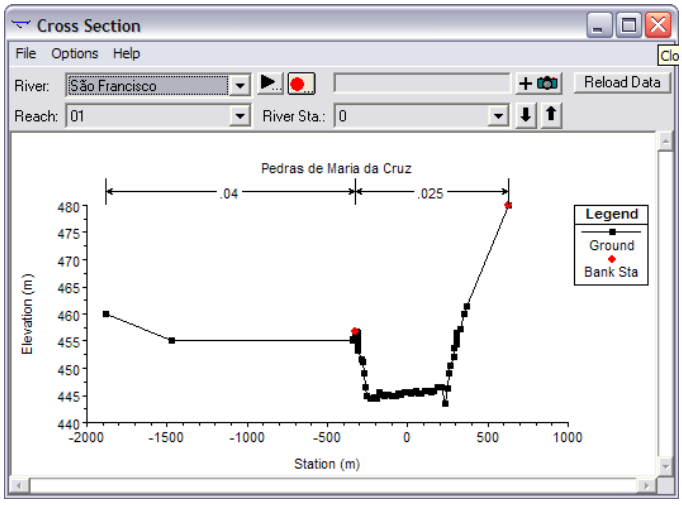
FIGURA 5.4: Modelagem do trecho em que foi realizada a simulação hidráulica.



(a) São Francisco



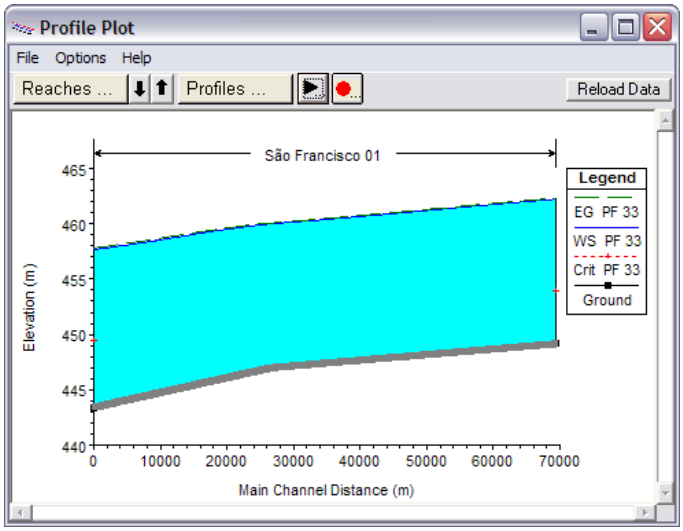
(b) Capão do Cleto



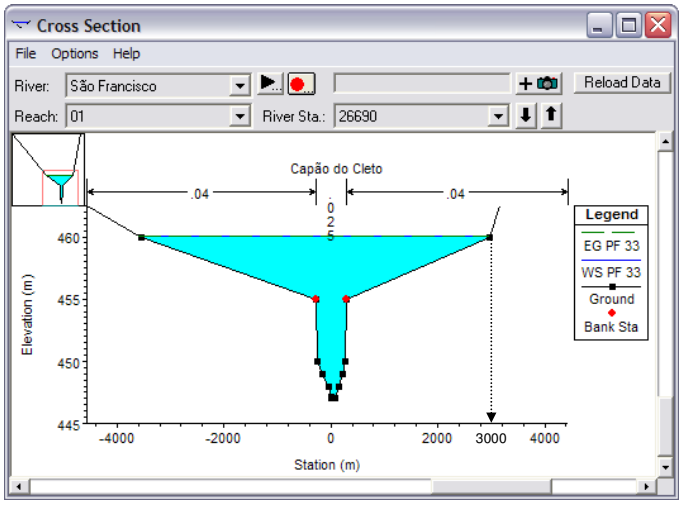
(c) Pedras de Maria da Cruz

FIGURA 5.5: Seções transversais construídas para a simulação hidráulica.

A Figura 5.6 mostra o resultado da simulação para um cenário de vazão igual a 13.250 m³/s, que conduziu à cota correspondente a meia légua (3 km) de inundação na margem direita, em Capão do Cleto. Entretanto, a despeito das inúmeras incertezas envolvidas e tendo em vista as séries de vazões máximas anuais dos postos de São Francisco e Pedras de Maria da Cruz, pode-se concluir que uma vazão da ordem de 13.000 m³/s não constitui um limiar de referência com período de retorno suficientemente elevado para ser empregado na construção de uma amostra binomial censurada. De fato, o uso de tal amostra com um limiar muito baixo pode conduzir à subestimação dos quantis de interesse.



(a) Perfil longitudinal



(b) Seção transversal em Capão do Cleto

FIGURA 5.6: Resultado da simulação hidráulica para a vazão de 13.250 m³/s.

Ainda analisando as informações apresentadas na Tabela 5.1, pode-se identificar dois eventos (1919 e 1926), cujas intensidades não se pode estimar com precisão, reconhecidamente maiores que a enchente de 1979. O evento de 1773, que, conforme relatado, provocou uma inundação correspondente a mais de vinte léguas de largura (120 km), também foi maior que o de 1979, como demonstrado a seguir. Suponhamos, por exemplo, que essas vinte léguas de inundação tenham ocorrido no Médio São Francisco, próximo à cidade de Xique-Xique (BA), onde, de acordo com o último relato da Tabela 5.1, acontecem os maiores transbordamentos. Mesmo após a enchente de 1979, a maior inundação relatada nessa região corresponde a uma largura de cerca de 14 léguas (84 km), menor que aquela provocada pelo evento de 1773. Com relação ao evento de 1712, descrito como uma memorável "*(...) grande enchente do rio, (...)*", e ao evento de 1790, que teria sido ainda maior que o de 1712, apesar de não se ter elementos para compará-los à enchente de 1979, serão considerados superiores a ela, pois, pelos adjetivos que lhes são atribuídos, há indícios de que foram consideravelmente grandes.

Consideremos, ainda, que os eventos de 1712, 1773, 1790, 1919 e 1926 correspondam a cheias generalizadas na bacia do rio São Francisco. Primeiramente, observe na Tabela 5.1 a maneira como foram relatados os dois últimos eventos. Descreve-se de forma explícita que as informações foram "*(...) obtidas no médio, sub-médio e baixo São Francisco, (...)*", o que permite concluir que em todos esses trechos as cheias de 1919 e 1926 foram mais severas que a de 1979. Observa-se também que os três primeiros eventos, relatados pelo naturalista José de Souza Azevedo Pizarro, o qual é referenciado por Auguste de Saint-Hilaire (ambos viajantes que percorreram grande parte da bacia do São Francisco), não foram descritos como cheias ocorridas em um local específico, o que constitui um indício de que foram observadas de maneira generalizada na bacia. Os eventos de 1712 e 1790, inclusive, tiveram claramente proporções nacionais, o que se pode inferir do trecho que diz: "*(...), porque esta época é bem conhecida dos mais antigos habitantes do país, entre quem ficou memorável a grande enchente do rio, que no ano referido houve, (...)*".

Para a construção da amostra binomial censurada, é necessário agora definir o intervalo de tempo em que as referidas cheias excederam o limiar de referência (enchente de 1979). Assim, é possível determinar o número de anos nos quais os eventos máximos anuais, embora desconhecidos, não foram alarmantes o suficiente para serem registrados historicamente, e, por isso, serão considerados inferiores ao limiar de referência. Como Pizarro e Eschwege são citados por Saint-Hilaire, fica claro que suas viagens aconteceram antes do período em que

Saint-Hilaire esteve viajando pelo Brasil. Como esse francês aqui chegou em 1816 e permaneceu até 1822, pode-se considerar que os eventos de 1712, 1773 e 1790 correspondem às excedências ocorridas entre 1712, no mínimo, e 1822. É claro que alguns anos anteriores a 1712 podem também ter apresentado cheias inferiores ao limiar, até que se verificasse novamente outra excedência. Porém, essa é uma hipótese que necessita de investigações mais profundas e, por isso, não será utilizada nessa pesquisa. Ao considerarmos 1712 como o ano inicial, estamos adotando uma estimativa mínima do tamanho da amostra. Como o relatório da Comissão Interministerial de Estudos para Controle das Enchentes do Rio São Francisco (1980) afirma que a enchente de 1979 foi a terceira maior do século, os eventos de 1919 e 1926 correspondem às excedências ocorridas entre 1901 e o início do período de medições sistemáticas. Desse forma, nada se conhece sobre os eventos ocorridos entre 1823 e 1900, os quais, por isso, não farão parte da amostra.

5.3.2.1 Determinação da amostra binomial censurada de vazões máximas anuais

As informações relatadas no item anterior viabilizaram a construção de uma amostra binomial censurada de vazões máximas anuais para o posto fluviométrico de São Francisco, no rio São Francisco, conforme ilustrado na Figura 5.7. Vale ressaltar que, uma vez que as informações históricas se referem a enchentes generalizadas no São Francisco, ou seja, ocorridas em grande parte da bacia, e que o período histórico constitui uma amostra binomial censurada, com o limiar de referência igual à vazão máxima anual de 1979 em um dado local de interesse, esses dados não sistemáticos podem ser associados às observações sistemáticas de qualquer posto fluviométrico, localizado na região explorada pelos autores dos registros apresentados na Tabela 5.1.

Portanto, a amostra binomial censurada de vazões máximas anuais do posto fluviométrico de São Francisco, mostrada na Figura 5.7, é composta por:

- um período histórico que vai de 1712 a 1934 (223 anos), sub-dividido em um período de 78 anos (de 1823 a 1900), sobre o qual nada se pode inferir, e um período de 145 anos (de 1712 a 1822 e, depois, de 1901 a 1934), no qual 5 eventos (1712, 1773, 1790, 1919 e 1926) igualaram ou excederam 17.380 m³/s, que corresponde à vazão máxima anual de 1979 em São Francisco (limiar de referência):

$$N_H = N_H^{\bullet} + N_H^{<} + N_H^{>} + N_H^{<>} = 0 + 140 + 5 + 0 = 145 \qquad \text{e} \qquad k = 5$$

- e um período de observações sistemáticas regulares, iniciado em 1935, representado por uma série de 68 anos de vazões máximas anuais (veja Anexo E):

$$N_S = N_S^{\bullet} = 68$$

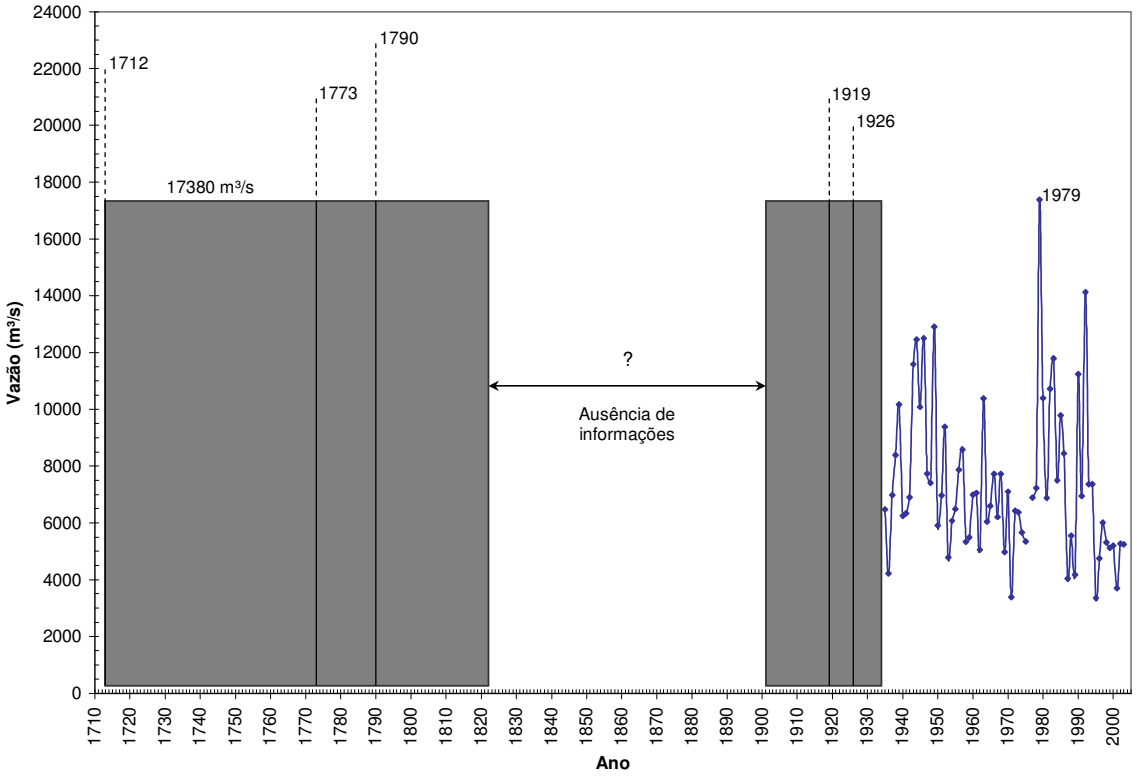


FIGURA 5.7: Amostra binomial censurada de vazões máximas anuais no rio São Francisco em São Francisco.

5.4 Formulação da distribuição a priori informativa

Duas possibilidades foram analisadas para a formulação da distribuição *a priori* informativa. A primeira consiste em utilizar os resultados de uma análise regional de freqüência para formular uma distribuição capaz de expressar o que se conhece *a priori* sobre os parâmetros (distribuição *a priori* regional). Na segunda, é adotada uma adaptação da distribuição proposta por Martins e Stedinger (2000), por meio da qual busca-se restringir os valores dos parâmetros a intervalos estatisticamente e/ou fisicamente plausíveis para o fenômeno geofísico em questão (distribuição *a priori* geofísica). Ambos os procedimentos serão apresentados na seqüência.

5.4.1 Distribuição *a priori* regional

Hosking e Wallis (1997) propuseram uma metodologia de análise regional de freqüência, utilizando momentos-L, que não só fornece subsídios para a escolha do modelo probabilístico, mas também possibilita a formulação da distribuição *a priori* dos parâmetros. Maiores detalhes a respeito dessa metodologia são apresentados no Anexo D. O cálculo da medida de aderência *Z* dos dados observados às distribuições candidatas, implementado na referida metodologia, permite a seleção da distribuição de freqüência que melhor se ajusta aos dados, ou seja, aquela cujo valor de *Z* mais se aproxima de zero. Além disso, os resultados da análise regional de freqüência podem ser utilizados para o ajuste de uma distribuição Normal multivariada ao vetor de parâmetros Θ, constituindo a distribuição de probabilidades *a priori* informativa, no formato requerido pelo *software* FLIKE.

5.4.1.1 Caracterização da região

Para a análise regional de freqüência, foram utilizados os postos fluviométricos apresentados na Tabela 5.2. As respectivas séries de vazões máximas anuais (dados sistemáticos) são mostradas no Anexo E. Todos os postos estão localizados no rio São Francisco, em seu trecho médio, conforme ilustrado na Figura 5.1.

TABELA 5.2: Postos fluviométricos utilizados na análise regional de freqüência.

Código	Nome da estação	Município	Estado	Área de drenagem (km²)
44200000	São Francisco	São Francisco	MG	182.537
44290002	Pedras de Maria da Cruz	Januária	MG	191.063
44500000	Manga	Manga	MG	202.400
45298000	Carinhanha	Carinhanha	BA	255.700
45480000	Bom Jesus da Lapa	Bom Jesus da Lapa	BA	273.750
46035000	Gameleira	Sítio do Mato	BA	309.540

5.4.1.2 Prescrição do modelo probabilístico

Os resultados da análise regional demonstraram que a região pode ser considerada como "aceitavelmente homogênea", não apresentando amostras com características estatísticas muito discrepantes das grupais. Entretanto, vale ressaltar que foram observados valores negativos da medida de heterogeneidade *H*, indicando a possível presença de correlação espacial entre os dados dos diferentes postos. Dentre as distribuições candidatas, três foram selecionadas pelo critério da medida de aderência *Z*, como mostra a Tabela 5.3. Uma vez que a distribuição Normal Generalizada não foi implementada no *software* FLIKE, a distribuição

escolhida é a Generalizada de Valores Extremos (GEV). A adequação dessa distribuição às características regionais pode ser vista no diagrama de quocientes de momentos-L, mostrado na Figura 5.8.

TABELA 5.3: Qualidade do ajuste das distribuições candidatas aos dados.

Distribuição	Medida de aderência (Z)	Situação
Logística Generalizada	1,93	Descartada
Generalizada de Valores Extremos	0,33	Selecionada
Normal Generalizada	-0,06	Selecionada
Pearson III	-0,86	Selecionada
Generalizada de Pareto	-3,34	Descartada

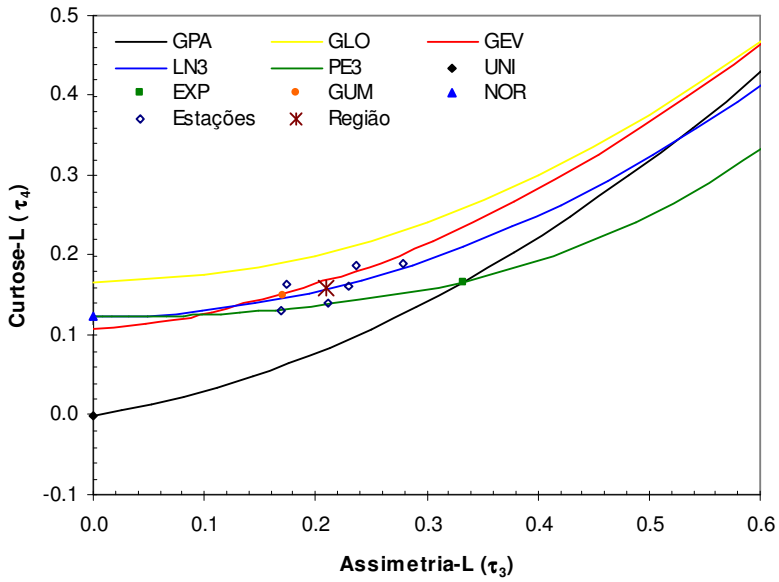


FIGURA 5.8: Diagrama de quocientes de momentos-L.

5.4.1.3 Formulação da distribuição *a priori* regional

A Tabela 5.4 apresenta um sumário dos resultados obtidos na análise regional de freqüência, utilizados para o ajuste da distribuição Normal multivariada aos parâmetros da GEV. Tendo em vista que os parâmetros de posição (β) e escala (α) são características essencialmente locais, os resultados da Tabela 5.4 serão usados para formular uma distribuição *a priori* informativa apenas a respeito do parâmetro de forma (k). Essa distribuição será uma Normal, com média igual ao valor regional de k e desvio-padrão calculado com base na "amostra" de parâmetros de forma, cujo tamanho é igual ao número de postos da região.

TABELA 5.4: Resultados da análise regional de freqüência e ajuste da distribuição Normal multivariada aos parâmetros da GEV.

Posto	Código	n	$Q_{média}$ (1) (m³/s)	CV-L (τ)	Assimetria-L (τ_3)	Curtose-L (τ_4)		Parâmetros		
								Posição (β)	Escala (α)	Forma (k)
1	44200000	68	7.385,2	0,1974	0,2372	0,1877		0,82318	0,25665	-0,10254
2	44290002	30	6.812,2	0,2151	0,2788	0,1886		0,79997	0,26041	-0,16335
3	44500000	64	6.710,3	0,1614	0,1694	0,1295		0,86568	0,23303	0,00082
4	45298000	62	7.073,7	0,1551	0,1739	0,1625		0,87021	0,22248	-0,00621
5	45480000	57	6.174,9	0,1522	0,2105	0,1405		0,86722	0,20656	-0,06248
6	46035000	31	6.448,6	0,1777	0,2304	0,1614		0,84179	0,23362	-0,09241
Região				0,1731	0,2092	0,1596	μ_Θ	0,84920	0,23541	-0,06051
							σ_Θ	0,02898	0,02042	0,06337

(1) Média da amostra de vazões máximas anuais

Portanto, a distribuição *a priori* regional será uma Normal multivariada, com as seguintes características:

$$\Theta = \begin{bmatrix} \beta \\ \alpha \\ k \end{bmatrix} \qquad \mu_\Theta = \begin{bmatrix} 1 \\ 1 \\ -0{,}06 \end{bmatrix} \qquad \sigma_\Theta = \begin{bmatrix} \infty \\ \infty \\ 0{,}0634 \end{bmatrix} \qquad R_\Theta = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.1}$$

As seguintes constatações podem ser feitas a respeito dessa distribuição:

1. O número de postos fluviométricos utilizados para a análise regional de freqüência é bastante reduzido, e, possivelmente, apresentam correlação espacial.

2. Foram observadas discrepâncias entre os dados fluviométricos oficiais e as informações disponíveis no relatório da Comissão Interministerial de Estudos para Controle das Enchentes do Rio São Francisco (1980), especialmente para os eventos correspondentes ao ramo superior das curvas-chave. Além disso, particularmente as séries de vazões máximas anuais de Manga (44500000) e Carinhanha (45298000) parecem apresentar inconsistências, conduzindo a valores inesperados do parâmetro de forma k.

3. Os valores encontrados para a média regional e para o desvio-padrão de k resultam numa distribuição Normal que restringe demasiadamente a faixa de variação do referido parâmetro. Em outras palavras, o nível de informação contido na distribuição *a priori* regional está maior que o esperado, o que, no contexto da análise bayesiana de freqüência, é incompatível com o caráter subjetivo da informação *a priori*.

Por isso, a distribuição *a priori* regional, formulada anteriormente, não parece ser a mais adequada para expressar o comportamento *a priori* do parâmetro de forma da GEV, e, portanto, não será empregada no presente estudo de caso.

5.4.2 Distribuição *a priori* geofísica

Martins e Stedinger (2000) demonstraram que o método do máximo de verossimilhança para estimação dos parâmetros da GEV, baseado em amostras de pequeno tamanho, pode resultar em valores absurdos do parâmetro de forma k. Esses autores empregaram a abordagem bayesiana para restringir as estimativas de k a uma faixa de valores estatisticamente e/ou fisicamente plausíveis. A distribuição *a priori* proposta, dita geofísica por produzir valores do parâmetro k compatíveis com a experiência mundial para fenômenos geofísicos (tais como precipitações intensas e vazões de enchentes), é uma Beta, com as seguintes características:

$$\pi(k) = \frac{(0{,}5+k)^{p-1} \cdot (0{,}5-k)^{q-1}}{B(p,q)} \qquad -0{,}5 \le k \le 0{,}5 \tag{5.2}$$

onde:

$$B(p,q) = \frac{\Gamma(p) \cdot \Gamma(q)}{\Gamma(p+q)}$$

com $p = 6$ e $q = 9$. Essa distribuição tem $E[k] = -0{,}10$ e $Var[k] = (0{,}122)^2$. Vale destacar que, para valores fora do intervalo $-0{,}4 \le k \le 0{,}2$, a distribuição Beta representada pela equação (5.2) apresenta valores praticamente nulos (veja Figura 5.9).

Conforme mostrado também na Figura 5.9, uma boa aproximação da distribuição proposta por Martins e Stedinger (2000) é obtida por uma distribuição Normal, com média $\mu = -0{,}10$ e variância $\varphi = (0{,}122)^2$. Assim, é possível expressar a distribuição *a priori* geofísica como uma Normal multivariada, com as seguintes características:

$$\Theta = \begin{bmatrix} \beta \\ \alpha \\ k \end{bmatrix} \qquad \mu_\Theta = \begin{bmatrix} 1 \\ 1 \\ -0{,}10 \end{bmatrix} \qquad \sigma_\Theta = \begin{bmatrix} \infty \\ \infty \\ 0{,}122 \end{bmatrix} \qquad R_\Theta = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.3}$$

Essa distribuição *a priori* geofísica será empregada no presente estudo de caso.

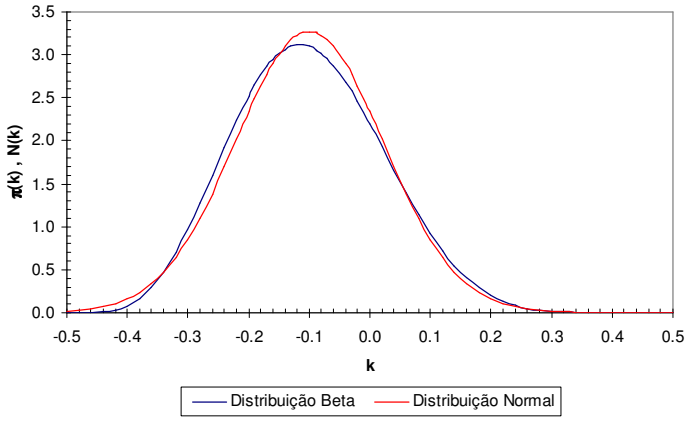


FIGURA 5.9: Distribuição *a priori* geofísica – aproximação da distribuição Beta pela Normal.

5.5 Análise de freqüência de vazões máximas anuais no rio São Francisco em São Francisco

Uma vez coletados os dados não sistemáticos e formulada a distribuição *a priori* informativa, é possível agregar essas informações suplementares à análise de freqüência de vazões máximas anuais no rio São Francisco em São Francisco. Para isso, diferentes cenários foram definidos, permitindo não só comparar as abordagens clássica e bayesiana, mas também avaliar o ganho proveniente da incorporação de cheias históricas. De acordo com o exposto no item 5.4, a distribuição de probabilidades adotada é a GEV.

5.5.1 Caracterização dos diferentes cenários

Seis cenários distintos serão analisados nesse estudo de caso, a saber:

- *Cenário 1*:

  Análise de freqüência tradicional, utilizando apenas a série de vazões máximas anuais do posto fluviométrico de São Francisco (dados sistemáticos). Os parâmetros do modelo distributivo são estimados pelo método do máximo de verossimilhança.

- *Cenário 2*:

  Análise de freqüência tradicional, utilizando a série de vazões máximas anuais do posto fluviométrico de São Francisco (dados sistemáticos) e o conjunto de cheias históricas apresentado no item 5.3, que constitui uma amostra binomial censurada (dados não sistemáticos). Os parâmetros do modelo distributivo são, mais uma vez, estimados pelo método do máximo de verossimilhança.

- *Cenário 3*:

  Análise bayesiana de freqüência, utilizando apenas os dados sistemáticos. Considerando que, num primeiro momento, não se dispõe de informações sobre os parâmetros, uma distribuição *a priori* não informativa é adotada. O *software* FLIKE é empregado para a determinação das distribuições da probabilidade esperada (ProE – distribuição bayesiana) e dos parâmetros esperados (ParE – estimação pontual bayesiana).

- *Cenário 4*:

  Análise bayesiana de freqüência, utilizando apenas os dados sistemáticos. A distribuição *a priori* geofísica (distribuição *a priori* informativa), descrita no item 5.4.2, é adotada, e o *software* FLIKE é novamente empregado.

- *Cenário 5*:

  Análise bayesiana de freqüência, utilizando os dados sistemáticos e não sistemáticos. Tal como no cenário 3, uma distribuição *a priori* não informativa é adotada, e o *software* FLIKE é empregado.

- *Cenário 6*:

  Análise bayesiana de freqüência, utilizando os dados sistemáticos e não sistemáticos. Tal como no cenário 4, a distribuição *a priori* geofísica é adotada, e o *software* FLIKE é empregado.

5.5.2 Apresentação e discussão dos resultados

Nas Tabelas 5.5, 5.6 e 5.7, são apresentados os resultados da análise de freqüência de vazões máximas anuais no rio São Francisco em São Francisco, para os seis cenários descritos anteriormente.

TABELA 5.5: Resultados da análise de freqüência de vazões máximas anuais no rio São Francisco em São Francisco, usando a distribuição GEV, para os cenários 1 e 2.

AF clássica (MVS)				
Tempo de retorno (anos)	Dados sistemáticos		Dados sistemáticos + Cheias históricas	
	Vazão (m³/s)	AIC (1) (% da vazão)	Vazão (m³/s)	AIC (1) (% da vazão)
1.1	4489	17.0%	4486	16.0%
2	6828	14.4%	6893	15.1%
5	9172	17.0%	9647	16.8%
10	10843	21.3%	11826	19.6%
20	12540	27.7%	14229	24.7%
50	14885	38.5%	17869	34.3%
100	16757	47.9%	21053	43.2%
200	18728	58.1%	24673	53.2%
500	21498	72.5%	30244	67.8%
1000	23727	84.1%	35148	79.5%
2000	26081	96.0%	40742	91.9%
5000	29394	112.5%	49360	108.9%
10000	32063	125.3%	56952	122.3%

(1) Amplitude do intervalo de confiança

TABELA 5.6: Resultados da análise de freqüência de vazões máximas anuais no rio São Francisco em São Francisco, usando a distribuição GEV, para os cenários 3 e 4.

AF bayesiana – Dados sistemáticos						
Tempo de retorno (anos)	Distribuição *a priori* não informativa			Distribuição *a priori* informativa		
	Distribuição ParE	Distribuição bayesiana (ProE)		Distribuição ParE	Distribuição bayesiana (ProE)	
	Vazão (m³/s)	Vazão (m³/s)	AIC (1) (% da vazão)	Vazão (m³/s)	Vazão (m³/s)	AIC (1) (% da vazão)
1.1	4449	4449	17.9%	4454	4454	17.6%
2	6847	6847	14.9%	6841	6841	14.6%
5	9287	9298	18.2%	9267	9277	17.8%
10	11047	11074	23.5%	11014	11045	21.9%
20	12852	12911	31.1%	12806	12866	27.4%
50	15371	15596	43.3%	15304	15489	36.3%
100	17403	17928	54.3%	17318	17698	44.1%
200	19561	20608	65.7%	19455	20137	52.5%
500	22626	25288	79.6%	22487	24136	63.7%
1000	25117	29440	90.3%	24950	27740	71.9%

(1) Amplitude do intervalo de confiança

TABELA 5.7: Resultados da análise de freqüência de vazões máximas anuais no rio São Francisco em São Francisco, usando a distribuição GEV, para os cenários 5 e 6.

AF bayesiana – Dados sistemáticos + Cheias históricas						
Tempo de retorno (anos)	Distribuição *a priori* não informativa			Distribuição *a priori* informativa		
	Distribuição ParE	Distribuição bayesiana (ProE)		Distribuição ParE	Distribuição bayesiana (ProE)	
	Vazão (m³/s)	Vazão (m³/s)	AIC (1) (% da vazão)	Vazão (m³/s)	Vazão (m³/s)	AIC (1) (% da vazão)
1.1	4445	4445	17.0%	4449	4449	17.6%
2	6902	6902	15.3%	6945	6945	15.0%
5	9714	9732	17.0%	9718	9735	16.8%
10	11937	11968	20.1%	11859	11890	19.2%
20	14390	14440	25.4%	14178	14229	22.9%
50	18107	18259	36.0%	17619	17748	30.2%
100	21357	21695	45.6%	20566	20819	37.1%
200	25054	25709	56.2%	23860	24306	45.0%
500	30742	32463	70.5%	28825	29919	56.1%
1000	35751	38910	81.3%	33109	35035	64.9%

(1) Amplitude do intervalo de confiança

As Figuras 5.10 e 5.11 mostram as curvas de quantis correspondentes aos cenários 1 e 2. Para os cenários 3 e 4, as curvas de quantis são apresentadas nas Figuras 5.12 e 5.13, enquanto as respectivas funções de verossimilhança, distribuições *a priori* e distribuições *a posteriori* são plotadas nas Figuras 5.14 e 5.15. Para os cenários 5 e 6, tais informações são mostradas nas Figuras 5.16, 5.17, 5.18 e 5.19.

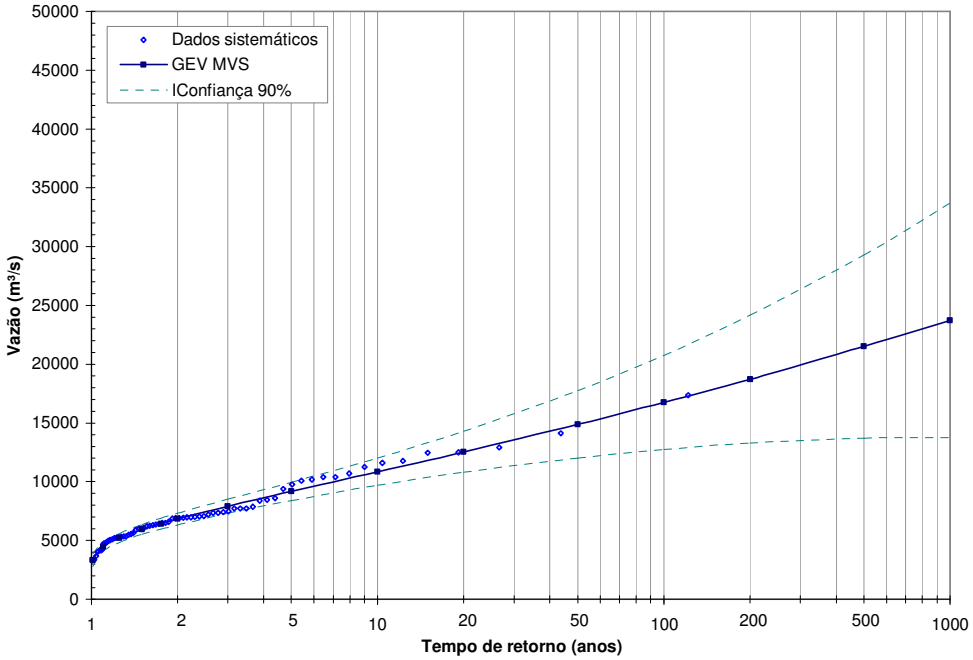


FIGURA 5.10: Análise de freqüência tradicional de vazões máximas anuais no rio São Francisco em São Francisco, usando a distribuição GEV (cenário 1).

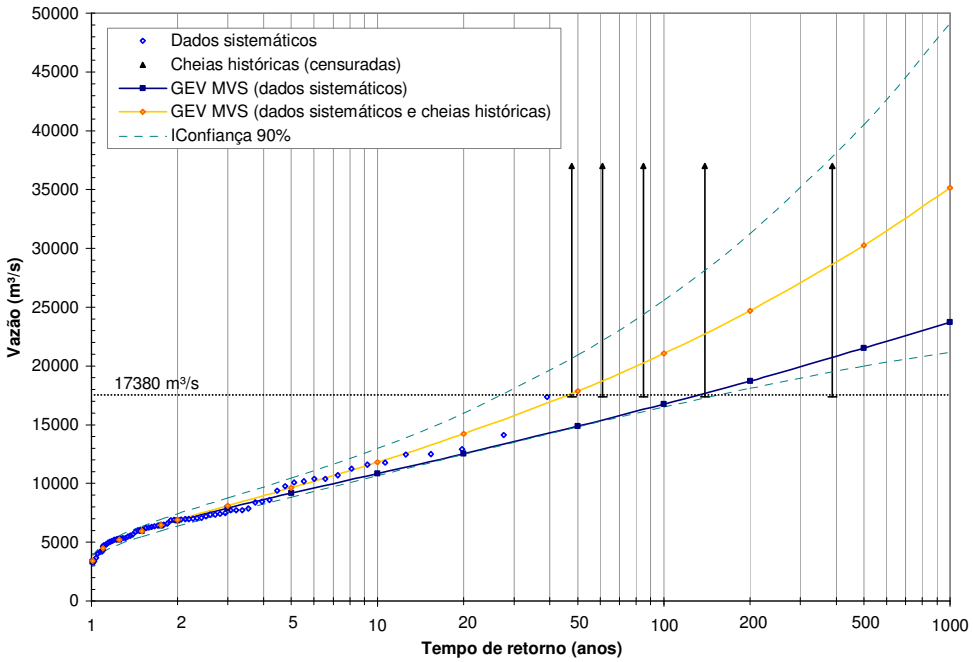


FIGURA 5.11: Análise de freqüência tradicional de vazões máximas anuais no rio São Francisco em São Francisco, com cheias históricas, usando a distribuição GEV (cenário 2).

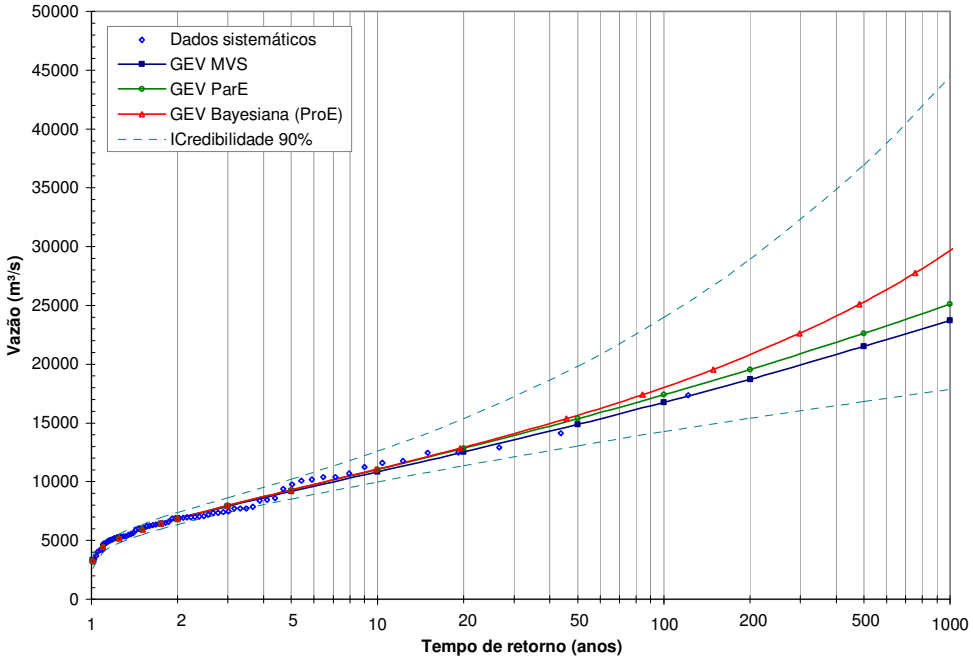


FIGURA 5.12: Análise bayesiana de freqüência de vazões máximas anuais no rio São Francisco em São Francisco, com distribuição *a priori* não informativa, usando a distribuição GEV (cenário 3).

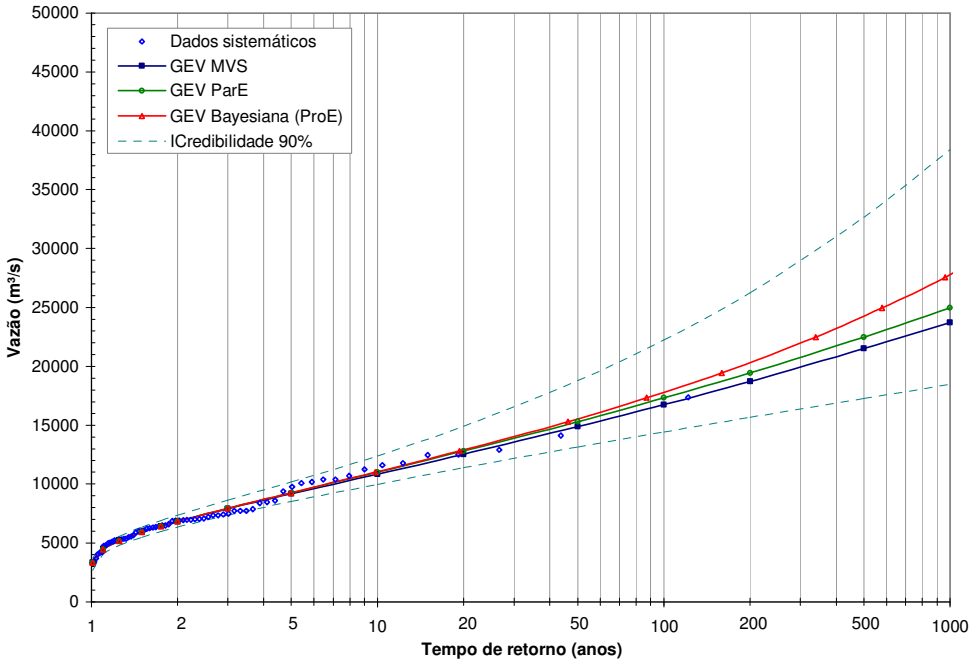


FIGURA 5.13: Análise bayesiana de freqüência de vazões máximas anuais no rio São Francisco em São Francisco, com distribuição *a priori* informativa, usando a distribuição GEV (cenário 4).

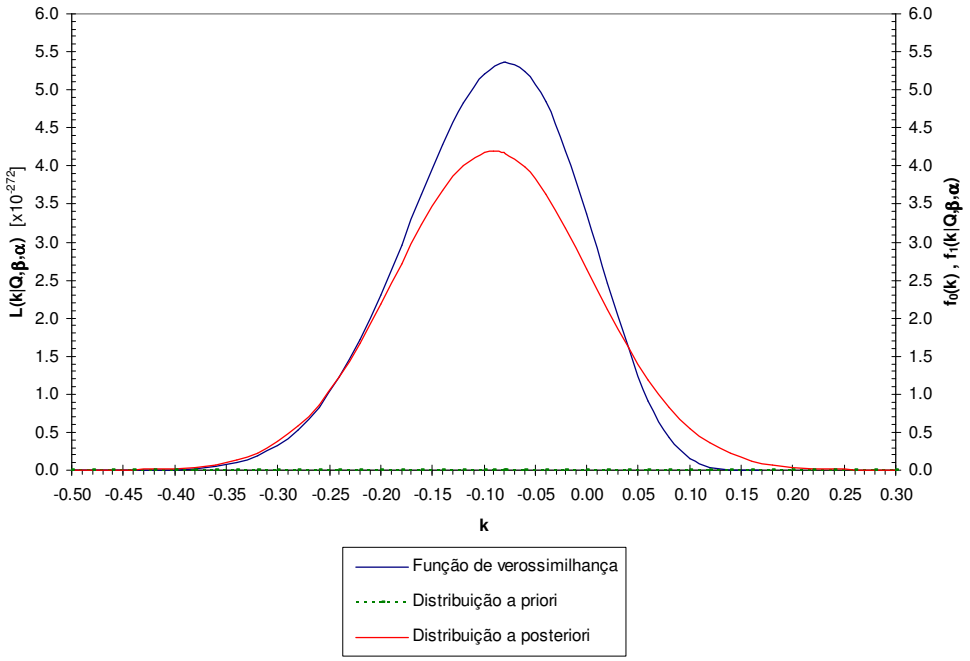


FIGURA 5.14: Função de verossimilhança, distribuição a priori e distribuição a posteriori para o cenário 3.

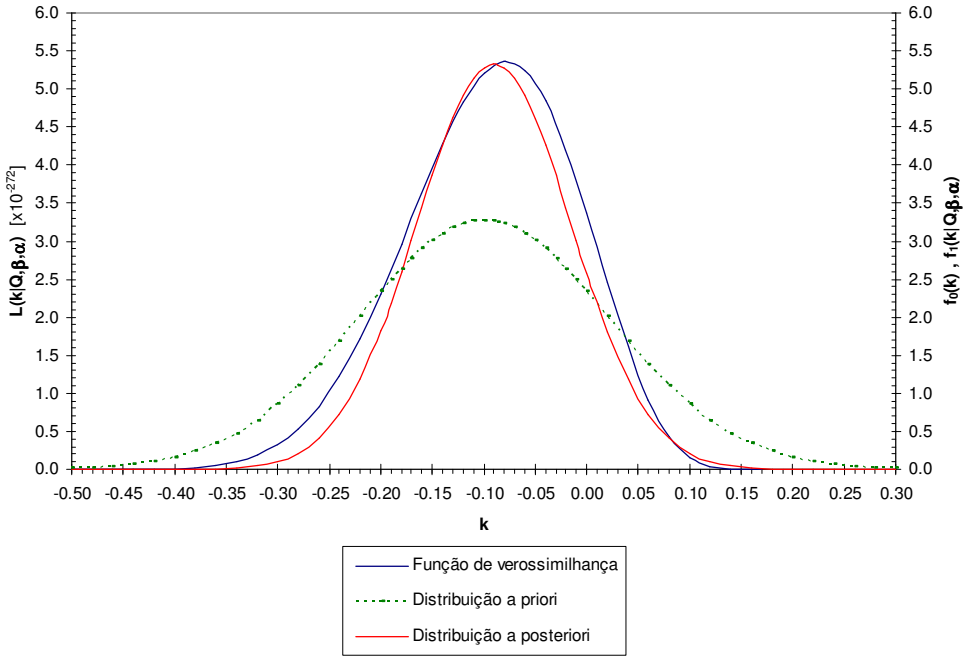


FIGURA 5.15: Função de verossimilhança, distribuição *a priori* e distribuição *a posteriori* para o cenário 4.

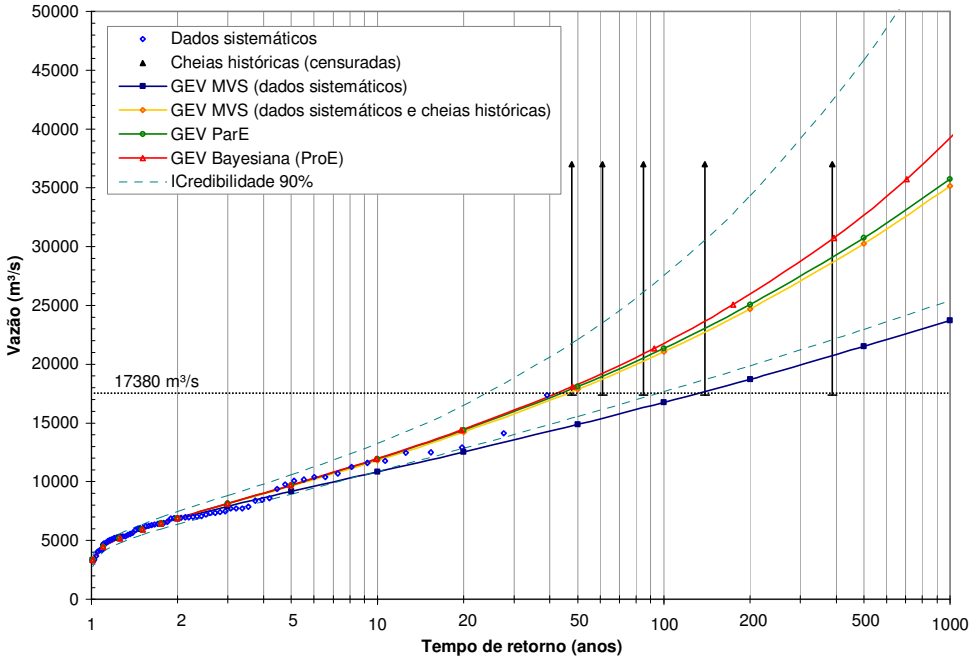


FIGURA 5.16: Análise bayesiana de freqüência de vazões máximas anuais no rio São Francisco em São Francisco, com cheias históricas e distribuição *a priori* não informativa, usando a distribuição GEV (cenário 5).

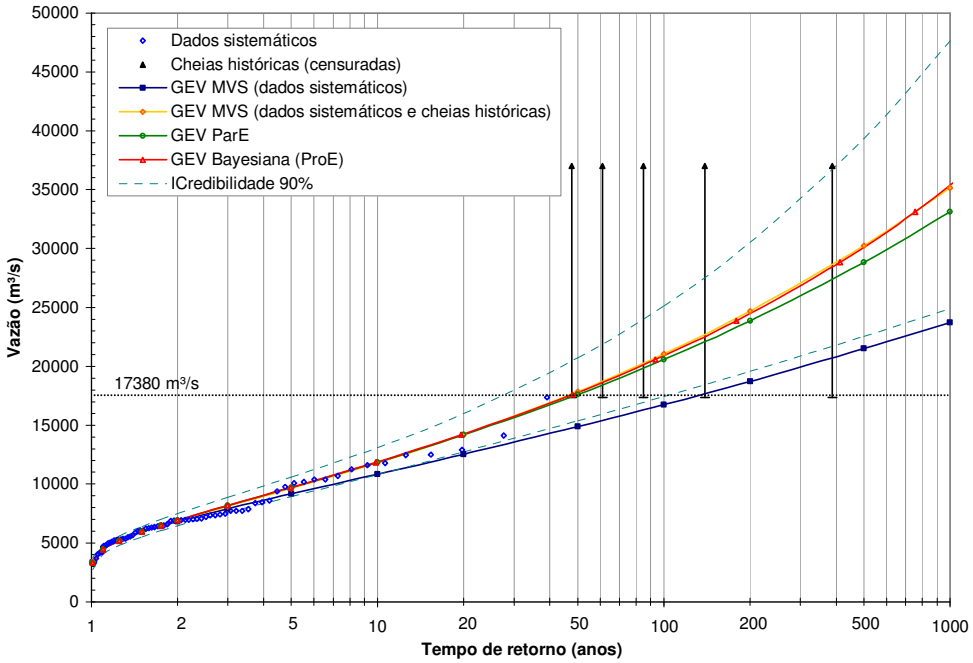


FIGURA 5.17: Análise bayesiana de freqüência de vazões máximas anuais no rio São Francisco em São Francisco, com cheias históricas e distribuição *a priori* informativa, usando a distribuição GEV (cenário 6).

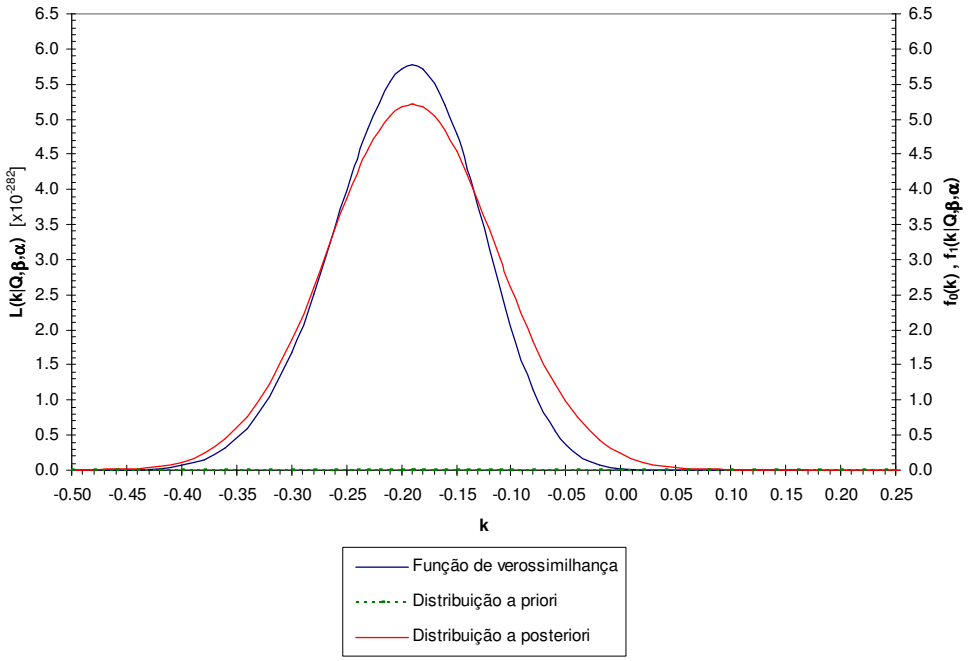


FIGURA 5.18: Função de verossimilhança, distribuição *a priori* e distribuição *a posteriori* para o cenário 5.

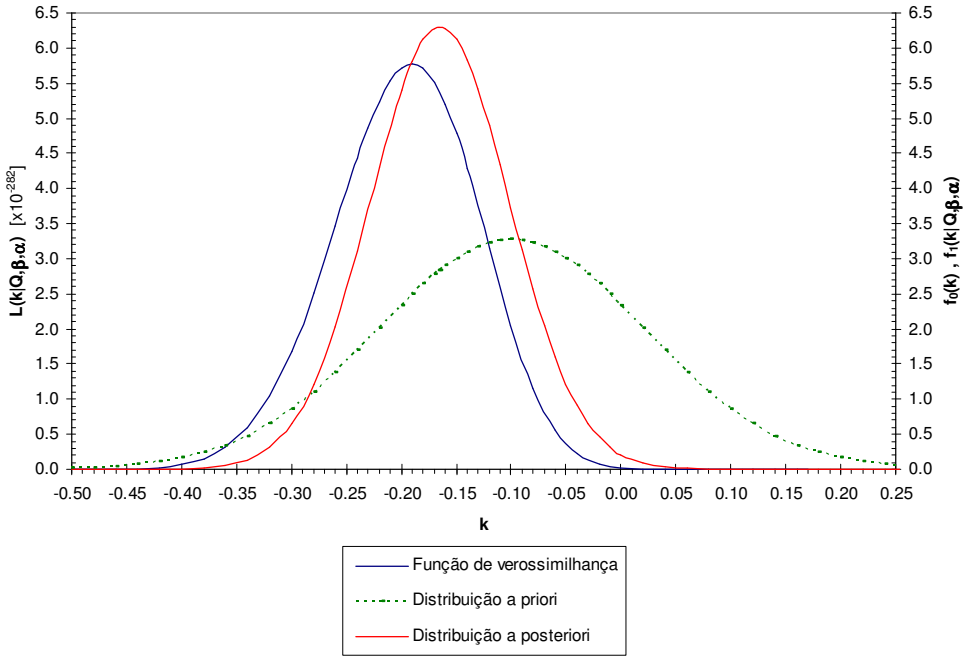


FIGURA 5.19: Função de verossimilhança, distribuição *a priori* e distribuição *a posteriori* para o cenário 6.

Apresentados os resultados desse estudo de caso, alguns aspectos podem ser comentados. Para um mesmo conjunto de dados, o uso da abordagem bayesiana, com uma distribuição *a priori* não informativa, conduz a distribuições mais conservadoras que aquelas obtidas pelo método clássico de análise de freqüência, isto é, com quantis mais elevados para um mesmo período de retorno. No entanto, à medida que aumenta o tamanho da amostra, o que acontece, por exemplo, quando cheias históricas são incorporadas à análise, esses procedimentos apresentam resultados cada vez mais similares. Essas observações são válidas tanto para a distribuição da probabilidade esperada, que representa, de fato, a distribuição bayesiana, independente de estimativas paramétricas, como para a distribuição dos parâmetros esperados. Essa última se aproxima mais da curva de quantis da análise de freqüência tradicional, pois corresponde ao mesmo modelo probabilístico, porém usando os estimadores bayesianos dos parâmetros.

Já as distribuições obtidas empregando-se a abordagem bayesiana, com uma distribuição *a priori* informativa, são resultantes de uma ponderação entre a informação avaliada *a priori* sobre os parâmetros e a informação contida na função de verossimilhança, proveniente da amostra. Sendo assim, o conjunto de informações utilizadas difere daquele disponível para a análise de freqüência tradicional, e, por isso, não se considera prudente comparar as posições relativas das respectivas curvas de quantis.

Além disso, os resultados obtidos demonstram que, para um mesmo quantil, a amplitude do intervalo de confiança sofre apenas uma pequena redução quando a abordagem bayesiana, com uma distribuição *a priori* não informativa, é usada na análise de freqüência. Como discutido no Capítulo 4, nesse contexto, a vantagem principal se refere à obtenção de uma distribuição marginal de probabilidades, livre das estimativas dos parâmetros. A incorporação de informações adicionais, entretanto, tais como as cheias históricas e a distribuição *a priori* informativa, reduz, de forma significativa, a amplitude do intervalo de confiança dos quantis, expressando um menor grau de incerteza em relação ao seu valor mais provável.

Outra observação importante se refere à mudança da posição de plotagem dos elementos da amostra, quando são utilizados dados não sistemáticos. Essa mudança, discutida no item 3.6, oferece condições interessantes para a interpretação de possíveis eventos atípicos (*outliers*). De fato, se eles pertencerem à mesma população da qual foram extraídos os demais eventos, sua posição de plotagem incoerente poderá ser corrigida por uma amostra mais longa. Caso

contrário, mesmo inserido em um contexto apropriado de longo termo, seu comportamento distinto será evidente.

A Figura 5.20 mostra, para os cenários correspondentes à análise bayesiana de freqüência, a aproximação, provida pela distribuição Normal multivariada $I(\Theta)$, da distribuição *a posteriori* dos parâmetros de escala e forma da GEV, condicionada ao parâmetro de posição. Pode-se perceber que, para os quatro casos, a aproximação é muito boa. No entanto, ao se estender o período de observações por meio da incorporação das cheias históricas, verifica-se que a distribuição *a posteriori* dos parâmetros se aproxima ainda mais da distribuição Normal multivariada. Para os casos em que se considera a distribuição *a priori* informativa, que, no *software* FLIKE, é representada por uma distribuição Normal multivariada, a aproximação é praticamente perfeita, já que a distribuição *a posteriori* resulta de uma combinação entre a distribuição *a priori* e a função de verossimilhança. Os mesmos argumentos são válidos para as outras duas distribuições condicionais *a posteriori* dos parâmetros da GEV: posição e escala, dado o parâmetro de forma (veja Figura 5.21), e posição e forma, dado o parâmetro de escala (veja Figura 5.22).

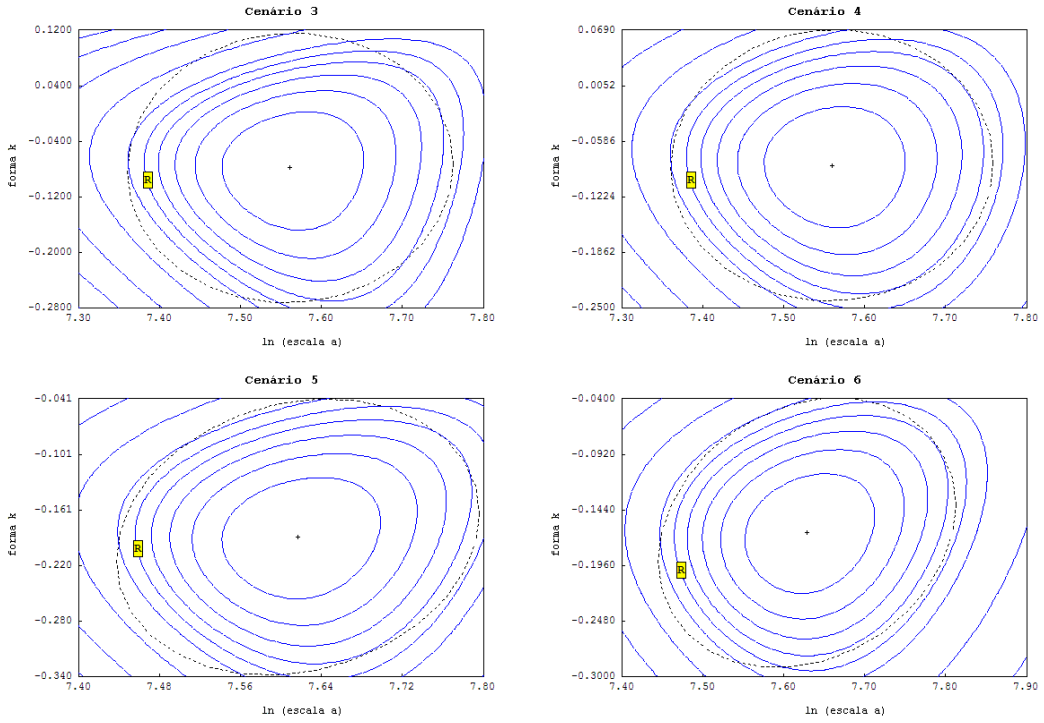


FIGURA 5.20: Superfícies da distribuição condicional *a posteriori* dos parâmetros de escala e forma da GEV e regiões de 90% de probabilidade da aproximação Normal multivariada, para os cenários 3, 4, 5 e 6.

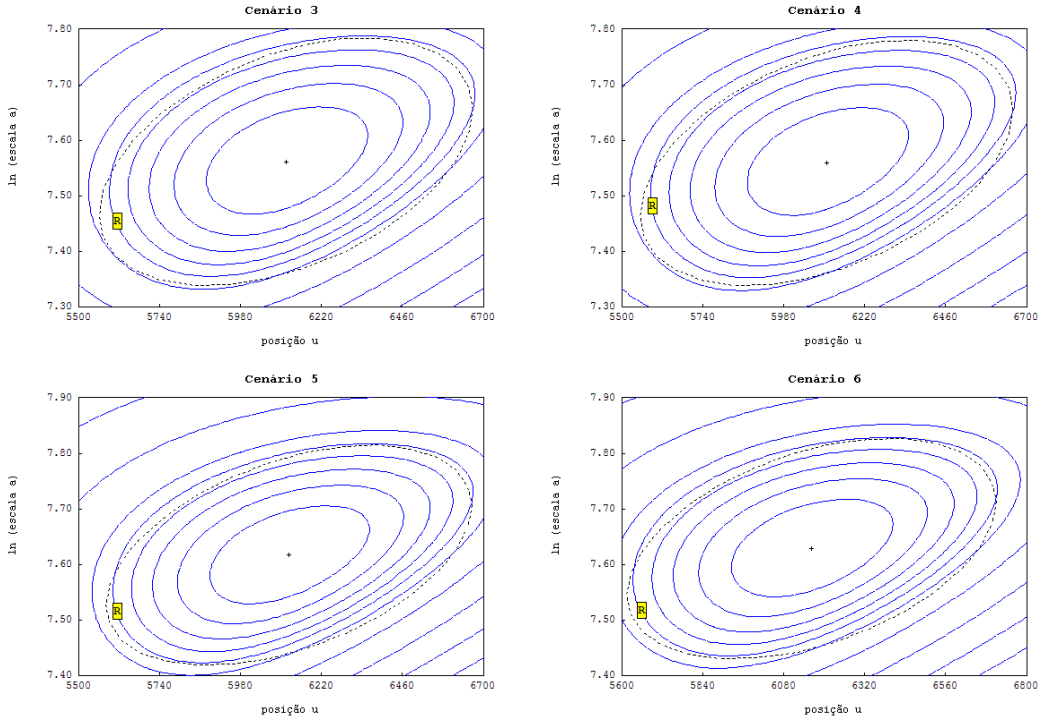


FIGURA 5.21: Superfícies da distribuição condicional a posteriori dos parâmetros de posição e escala da GEV e regiões de 90% de probabilidade da aproximação Normal multivariada, para os cenários 3, 4, 5 e 6.

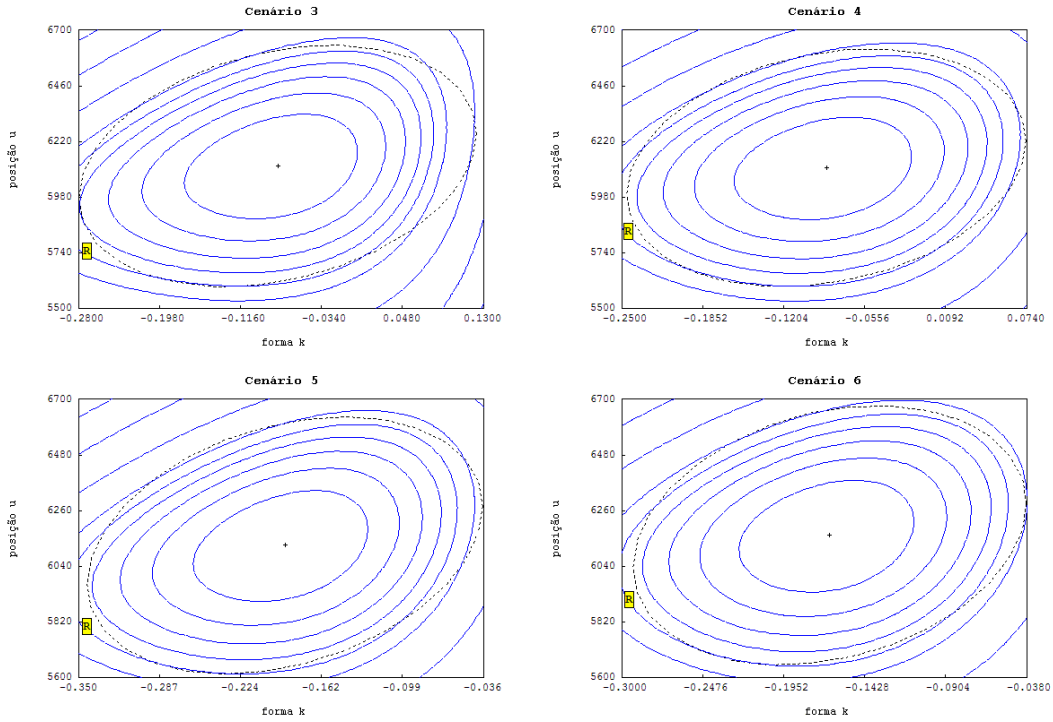


FIGURA 5.22: Superfícies da distribuição condicional *a posteriori* dos parâmetros de posição e forma da GEV e regiões de 90% de probabilidade da aproximação Normal multivariada, para os cenários 3, 4, 5 e 6.

6 CONCLUSÕES E RECOMENDAÇÕES

Nessa pesquisa, foi realizada a análise bayesiana de freqüência de vazões máximas anuais no rio São Francisco em São Francisco, incorporando dados não sistemáticos de cheias. O *software* FLIKE (KUCZERA, 1999) foi utilizado para esse fim. A coleta de informações sobre cheias históricas na bacia hidrográfica do São Francisco permitiu construir uma amostra binomial censurada de vazões máximas anuais, constituída por dados sistemáticos e não sistemáticos. Com relação à abordagem bayesiana, foram consideradas duas possibilidades: o uso de uma distribuição *a priori* não informativa e da distribuição *a priori* geofísica, proposta por Martins e Stedinger (2000).

As seguintes considerações podem ser feitas a respeito da metodologia empregada nesse estudo:

- Não só na análise bayesiana de freqüência, mas também quando são utilizados os métodos clássicos de estimação dos parâmetros (análise de freqüência tradicional), a incorporação dos dados não sistemáticos tem o potencial de produzir estimativas mais confiáveis dos quantis de interesse, uma vez que as inferências são baseadas em uma maior quantidade de informação. Assim, recomenda-se uma busca mais abrangente por dados não sistemáticos de cheias, o que pode prover informações adicionais valiosas a serem incorporadas na análise. Nesse contexto, as investigações paleohidrológicas, desde que viáveis, seriam de grande importância, servindo inclusive para confirmar hipóteses relativas às cheias históricas. No entanto, cuidados especiais devem ser tomados quando se utilizam dados não sistemáticos, já que informações inadequadas podem, ao invés de melhorar, piorar a qualidade das estimativas.
- Enquanto na análise de freqüência tradicional busca-se encontrar os estimadores mais apropriados dos parâmetros do modelo distributivo, a abordagem bayesiana oferece ferramentas para a formulação de uma distribuição marginal de probabilidades, que leva em conta todos os valores possíveis dos parâmetros, chamada distribuição bayesiana. Esse procedimento permite lidar, de forma coerente, com a incerteza decorrente do fato de que os verdadeiros valores populacionais dos parâmetros são desconhecidos.
- O *software* FLIKE permite realizar a análise bayesiana de freqüência utilizando duas abordagens distintas. A primeira, baseada na teoria da distribuição bayesiana, resulta na

distribuição da probabilidade esperada (ProE). A segunda, que usa o procedimento de estimação pontual bayesiana, fornece a distribuição dos parâmetros esperados (ParE). Apesar de se verificar, para um mesmo conjunto de dados, uma diferença entre essas duas distribuições, principalmente na cauda superior, observa-se que um aumento da quantidade de informação incorporada à análise de freqüência (seja pelo aumento do tamanho da amostra de dados sistemáticos ou pela utilização de dados não sistemáticos de cheias) faz com que as referidas distribuições apresentem comportamentos mais parecidos. Além disso, pode-se dizer que, quanto mais informação relevante for utilizada na estimação dos parâmetros do modelo distributivo, menor será a diferença entre os estimadores clássicos (obtidos, por exemplo, pelo método do máximo de verossimilhança) e os estimadores bayesianos. Dessa forma, assintoticamente, a distribuição clássica e a distribuição dos parâmetros esperados (ParE) tendem a coincidir.

- Embora, no contexto da análise bayesiana de freqüência, não haja restrições quanto ao emprego de uma distribuição *a priori* não informativa, o uso de informações subjetivas avaliadas *a priori*, desde que coerentes com o fenômeno geofísico em questão, pode resultar em estimativas mais confiáveis *a posteriori*.

Conclui-se, portanto, que a utilização dos dados não sistemáticos e da estatística bayesiana na análise de freqüência de vazões de enchentes permite estender o tamanho da amostra e levar em conta a incerteza estatística envolvida no processo de inferência, conduzindo a estimativas mais confiáveis dos quantis de interesse e de suas respectivas probabilidades de excedência.

Finalmente, acredita-se que essa pesquisa tenha demonstrado o potencial de se criar e manter um registro histórico detalhado dos eventos mais importantes ocorridos em uma bacia hidrográfica, de modo que a informação não se perca com o tempo e possa ser utilizada em inferências futuras. Esse processo deve envolver, inclusive, a investigação e materialização de marcas de cheias, bem como estudos de depósitos de sedimentos.

Vale lembrar que esse trabalho constituiu um passo inicial na revisão dos conhecimentos científicos disponíveis e na aplicação, a uma bacia hidrográfica brasileira, dos procedimentos de análise bayesiana de freqüência de cheias, incorporando dados não sistemáticos. Dessa forma, espera-se ter contribuído com o avanço científico e dar suporte a novas pesquisas nessa área de conhecimento.

REFERÊNCIAS

BAKER, V. R. Paleoflood hydrology and extraordinary flood events. *Journal of Hydrology*, 96, p. 79-99, 1987.

BAYLISS, A. C.; REED, D. W. *The use of historical data in flood frequency estimation*. Centre for Ecology and Hydrology, Natural Environment Research Council, Wallingford, 2001, 87 p.

BENITO, G.; LANG, M.; BARRIENDOS, M.; LLASAT, M. C.; FRANCÉS, F.; OUARDA, T.; THORNDYCRAFT, V. R.; ENZEL, Y.; BARDOSSY, A.; COEUR, D.; BOBÉE, B. Use of systematic, paleoflood and historical data for the improvement of flood risk estimation. Review of scientific methods. *Natural Hazards*, 31, p. 623-643, 2004.

BENJAMIN, J. R.; CORNELL, C. A. *Probability, statistics and decision for civil engineers*. McGraw-Hill, New York, 1970.

BOBÉE, B.; RASMUSSEN, P. Recent advances in flood frequency analysis. U.S. National Report to IUGG, 1991-1994, *Geophysics*, 33(suppl.), 1995.

BRADLEY, A. A.; POTTER, K. W. Flood frequency analysis of simulated flows. *Water Resources Research*, 28(9), p. 2375-2385, 1992.

BURN, D. H. Cluster analysis as applied to regional flood frequency. *Journal of Water Resources Planning and Management*, 115, p. 567-582, 1989.

CÂNDIDO, M. O. SEAF – Um protótipo de um sistema especialista para análise de freqüência local de eventos hidrológicos máximos anuais. Dissertação de Mestrado, Programa de Pós-graduação em Saneamento, Meio Ambiente e Recursos Hídricos – UFMG, Belo Horizonte, 2003.

CLARKE, R. T. *Statistical modelling in hydrology*. J. Wiley and Sons, Chichester, Inglaterra, 1994.

COHN, T. A.; LANE, W. L.; BAIER, W. G. An algorithm for computing moments-based flood quantile estimates when historical flood information is available. *Water Resources Research*, 33(9), p. 2089-2096, 1997.

COHN, T. A.; LANE, W. L.; STEDINGER, J. R. Confidence intervals for Expected Moments Algorithm flood quantile estimates. *Water Resources Research*, 37(6), p. 1695-1706, 2001.

COHN, T. A.; STEDINGER, J. R. Use of historical information in a maximum-likelihood framework. *Journal of Hydrology*, 96, p. 215-223, 1987.

COMISSÃO INTERMINISTERIAL DE ESTUDOS PARA CONTROLE DAS ENCHENTES DO RIO SÃO FRANCISCO. Relatório de avaliação do impacto da cheia de 1979 na bacia do rio São Francisco. Brasília, 1980.

CUNNANE, C. Unbiased plotting positions, a review. *Journal of Hydrology*, 37, p. 205-222, 1978.

DALRYMPLE, T. *Flood frequency analysis, Manual of hydrology, Part 3: Flood-flow techniques*. Geological Survey Water Supply Paper 1543-A, U.S. Government Printing Office, Washington DC, 1960, 80 p.

DAVIS, D. R.; DUCKSTEIN, L.; KRZYSZTOFOWICZ, R. The work of hydrologic data for nonoptimal decision making. *Water Resources Research*, 15(6), p. 1733-1742, 1979.

DAVIS, D. R.; KISIEL, C. S.; DUCKSTEIN, L. Bayesian decision theory applied to design in hydrology. *Water Resources Research*, 8(1), p. 33-41, 1972.

DAVIS, E. G.; NAGHETTINI, M. C. *Estudo de chuvas intensas no estado do Rio de Janeiro*. Companhia de Pesquisa de Recursos Minerais, Belo Horizonte, 2001, 140 p.

DUBAND, D. Pour une meilleure prise en compte de l'information hydrométéorologique historique, sur les crues importantes des bassins supérieurs de certaines rivières à risque. In: CONGRÈS DE LA SHF, 23ème JOURNÉE DE L'HYDRAULIQUE, Nimes, França, p. 137-144, 1994.

ENGLAND Jr., J. F. *At-site flood frequency estimation with historical/paleohydrologic data*. U.S. Department of the Interior, Bureau of Reclamation, Denver, 1999.

ESCHWEGE, W. L. von. *Brasil, novo mundo*. Tradução: Domício de Figueiredo Murta, Centro de Estudos Históricos e Culturais da Fundação João Pinheiro, Belo Horizonte, 1996, 275 p.

FRANCÉS, F.; SALAS, J. D.; BOES, D. C. Flood frequency analysis with systematic and historical or paleoflood data based on the two-parameter general extreme value models. *Water Resources Research*, 30(6), p. 1653-1664, 1994.

GLASER, R. Data and methods of climatological evaluation in historical climatology. *Historical Social Research*, 21, p. 56-88, 1996.

GREENWOOD, J. A.; LANDWEHR, J. M.; MATALAS, N. C.; WALLIS, J. R. Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form. *Water Resources Research*, American Geophysical Union, 15(5), p. 1049-1054, 1979.

GUO, S. L.; CUNNANE, C. Evaluation of the usefulness of historical and paleological floods in quantile estimation. *Journal of Hydrology*, 129, p. 245-262, 1991.

GUTTMAN, N. B. The use of L-moments in the determination of regional precipitation climates. *Journal of Climate*, 6, p. 2309-2325, 1993.

HARTIGAN, J. A. *Clustering algorithm*. Wiley, New York, 1975, *apud* Statsoft Inc., Electronic Statistics Textbook, Statsoft, Tulsa, EUA (http://www.statsoft.com/textbook/stathome.html), 1997.

HEC. HEC-RAS River Analysis System – User's Manual, Version 3.1.1. U.S. Army Corps of Engineers, Hydrologic Engineering Center, Davis, EUA, 2003.

HIRSCH, R. M.; STEDINGER, J. R. Plotting positions for historical floods and their precision. *Water Resources Research*, 23(4), p. 715-727, 1987.

HOSKING, J. R. M. The theory of probability weighted moments. *IBM Research Report*, RC 12210, IBM Research Division, Yorktown Heights, New York, 1986.

HOSKING, J. R. M. L-moments: analysis and estimation of distributions using linear combination of order statistics. *Journal of the Royal Statistical Society*, Series B, 52, p. 105-124, 1990.

HOSKING, J. R. M.; WALLIS, J. R. Paleoflood hydrology and flood frequency analysis. *Water Resources Research*, 22(4), p. 543-550, 1986a.

HOSKING, J. R. M.; WALLIS, J. R. The value of historical data in flood frequency analysis. *Water Resources Research*, 22(11), p. 1606-1612, 1986b.

HOSKING, J. R. M.; WALLIS, J. R. Some statistics useful in regional frequency analysis. *Water Resources Research*, 29(1), p. 271-281, 1993.

HOSKING, J. R. M.; WALLIS, J. R. Correction to "some statistics useful in regional frequency analysis". *Water Resources Research*, 31(1), p. 251, 1995.

HOSKING, J. R. M.; WALLIS, J. R. *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press, Cambridge, Reino Unido, 1997, 224 p.

HOUSE, P. K.; WEBB, R. H.; BAKER, V. R.; LEVISH, D. R. *Ancient floods, modern hazards. Principles and applications of paleoflood hydrology*. Water Science and Application 5, American Geophysical Union, Washington DC, 2001, 385 p.

INSTITUTION OF ENGINEERS AUSTRALIA. *Australian rainfall and runoff: a guide to flood estimation, Vol. 1*. Institution of Engineers Australia, Canberra, 1987, 374 p.

KIRBY, W. H. *Instructions for peak flow file*. Technical Report 79-1336-1, U.S. Geological Survey Open File Report, 1981.

KITE, G. W. *Frequency and risk analysis in hydrology*. Water Resources Publications, Fort Collins, 1977.

KOTTEGODA, N. T.; ROSSO, R. *Statistics, probability and reliability for civil and environmental engineers*. McGraw-Hill, New York, 1997, 735 p.

KUCZERA, G. Correlated rating curve error in flood frequency inference. *Water Resources Research*, 32(7), p. 2119-2127, 1996.

KUCZERA, G. Comprehensive at-site flood frequency analysis using Monte Carlo Bayesian inference. *Water Resources Research*, 35(5), p. 1551-1557, 1999.

LANE, W. L.; COHN, T. A. Expected Moments Algorithm for flood frequency analysis. In: NORTH AMERICAN WATER AND ENVIRONMENT CONGRESS' 96, Anaheim, California, EUA, 1996, 6 p.

LEE, P. M. *Bayesian statistics: an introduction*. Halsted Press, Grã-Bretanha, 1989, 294 p.

MARTINS, E. S.; STEDINGER, J. R. Generalized maximum likelihood generalized extreme value quantile estimators for hydrologic data. *Water Resources Research*, 36(3), p. 737-744, 2000.

MARTINS, E. S.; STEDINGER, J. R. Generalized maximum likelihood Pareto-Poisson estimators for partial duration series. *Water Resources Research*, 37(10), p. 2551-2557, 2001a.

MARTINS, E. S.; STEDINGER, J. R. Historical information in a generalized maximum likelihood framework with partial duration and annual maximum series. *Water Resources Research*, 37(10), p. 2559-2567, 2001b.

NAGHETTINI, M. C.; PINTO, E. J. A. *Hidrologia estatística: boletim técnico*. Companhia de Pesquisa de Recursos Minerais, Belo Horizonte, no prelo.

NAULET, R. Utilisation de l'information des crues historiques pour une meilleure prédétermination du risque d'inondation. Application au bassin de l'Ardèche à Vallon Pont-d'Arc et St-Martin d'Ardèche. Thèse UJF, PhD INRS-ETE, Grenoble, França, 2002.

NERC. *Flood studies report, Vol. 1: Hidrological studies*. National Environment Research Council, Londres, Reino Unido, 1975.

NRC. *Estimating probabilities of extreme floods*. National Research Council, National Academy Press, Washington, 1987, 141 p.

O'CONNELL, D. R. H.; OSTENAA, D. A.; LEVISH, D. R.; KLINGER, R. E. Bayesian flood frequency analysis with paleohydrologic bound data. *Water Resources Research*, 38(5), p. 16.1-16.13, 2002.

OUARDA, T. B. M. J.; RASMUSSEN, P. F.; BOBÉE, B.; BERNIER, J. Utilisation de l'information historique en analyse hydrologique fréquentielle. *Revue des Sciences de l'Eau* (spécial), p. 41-49, 1998.

PARENT, E.; BERNIER, J. Encoding prior experts judgments to improve risk analysis of extreme hydrological events via POT modeling. *Journal of Hydrology*, 283, p. 1-18, 2003.

PEARSON, C. P. Regional flood frequency for small New Zealand basins 2: flood frequency groups. *Journal of Hydrology* (Nova Zelândia), 30, p. 53-64, 1991.

PINTO, E. J. A.; NAGHETTINI, M. C. Definição de regiões homogêneas e regionalização de freqüência das precipitações diárias máximas anuais da bacia do alto rio São Francisco. Anais... 13º SIMPÓSIO BRASILEIRO DE RECURSOS HÍDRICOS (CD-ROM), Belo Horizonte, 1999.

PIZARRO, J. S. A. *Memórias históricas do Rio de Janeiro, Vol. VIII: Memórias das províncias anexas à jurisdição do vice-rei do estado do Brasil*. Imprensa Nacional, Rio de Janeiro, 1948.

POTTER, K. W. Research on flood frequency analysis: 1983-1986. *Geophysics*, 26(3), p. 113-118, 1987.

PRESCOTT, P.; WALDEN, A. T. Maximum likelihood estimation of the three-parameter generalized extreme value distribution from censored samples. *J. Statist. Comput. Simul.*, 16, p. 241-250, 1983.

RAO, A. R.; HAMED, K. H. *Flood frequency analysis*. CRC Press, Boca Raton, Florida, 2000.

SAINT-HILAIRE, A. de. *Viagem pelas províncias do Rio de Janeiro e Minas Gerais*. Tradução: Vivaldi Moreira, Editora Itatiaia, Belo Horizonte, 1975, 378 p.

SALAS, J. D.; WOLD, E. E.; JARRETT, R. D. *Determination of flood characteristics using systematic, historical and paleoflood data*. In: Coping with Floods, Kluwer, Dordrecht, p. 111-134, 1994.

SCHAEFER, M. C. Regional analysis of precipitation annual maxima in Washington state. *Water Resources Research*, 26(1), p. 119-131, 1990.

STEDINGER, J. R. Design events with specified flood risk. *Water Resources Research*, 19(2), p. 511-522, 1983.

STEDINGER, J. R.; COHN, T. A. Flood frequency analysis with historical and paleoflood information. *Water Resources Research*, 22(5), p. 785-793, 1986.

STEDINGER, J. R.; VOGEL, R. M.; FOUFOULA-GEORGIOU, E. *Frequency analysis of extreme events*. Chapter 18. In: Handbook of Hydrology, McGraw-Hill, New York, p. 18.1-18.66, 1993.

SUTCLIFFE, J. V. The use of historical records in flood frequency analysis. *Journal of Hydrology*, 96, p. 159-171, 1987.

THORNDYCRAFT, V. R.; BENITO, G.; BARRIENDOS, M.; LLASAT, M. C. *Paleofloods, historical data and climatic variability: applications in flood risk assessment*. Centro de Ciencias Medioambientales – CSIC, Madrid, 2003, 372 p.

TODOROVIC, P.; ZELENHASIC, E. A stochastic model for flood analysis. *Water Resources Research*, 6(6), p. 411-424, 1970.

USWRC. *Guidelines for determining flood flow frequency*. United States Water Resources Council, Hydrology Committee, Bulletin 17 B (revised), U.S. Government Printing Office, Washington DC, 1982.

VICENS, G. J.; RODRIGUEZ-ITURBE, I.; SCHAAKE Jr., J. C. A Bayesian framework for the use of regional information in hydrology. *Water Resources Research*, 11(3), p. 405-414, 1975.

VOGEL, R. M.; FENNESSEY, N. M. L-moment diagrams should replace product moment diagrams. *Water Resources Research*, 29(6), p. 1745-1752, 1993.

WARD, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, p. 236, 1963, *apud* Statsoft Inc., Electronic Statistics Textbook, Statsoft, Tulsa, EUA (http://www.statsoft.com/textbook/stathome.html), 1997.

WILTSHIRE, S. E. Grouping basins for regional flood frequency analysis. *Hydrological Sciences Journal*, 30(1), p. 151-159, 1985.

WILTSHIRE, S. E. Identification of homogeneous regions for flood frequency analysis. *Journal of Hydrology*, 84, p. 287-302, 1986.

WOOD, E. F.; RODRIGUEZ-ITURBE, I.; SCHAAKE Jr., J. C. *The methodology of Bayesian inference and decision making applied to extreme hydrologic events*. Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1974, 296 p.

WOOD, E. F.; RODRIGUEZ-ITURBE, I. Bayesian inference and decision making for extreme hydrologic events. *Water Resources Research*, 11(4), p. 533-542, 1975a.

WOOD, E. F.; RODRIGUEZ-ITURBE, I. A Bayesian approach to analyzing uncertainty among flow frequency models. *Water Resources Research*, 11(6), p. 839-843, 1975b.

WOOD, E. F. Analyzing hydrologic uncertainty and its impact upon decision making in water resources. *Advances in Water Resources*, 1(5), p. 299-305, 1978.

A DISTRIBUIÇÕES DE PROBABILIDADES MAIS UTILIZADAS NA ANÁLISE DE FREQÜÊNCIA DE VARIÁVEIS HIDROLÓGICAS

Nesse anexo, são apresentadas as características mais importantes das principais distribuições de probabilidades utilizadas em hidrologia, bem como os seus estimadores paramétricos, calculados pelo método dos momentos (MOM), pelo método do máximo de verossimilhança (MVS) e pelo método dos momentos-L (MML). Essa descrição foi retirada de Naghettini e Pinto (no prelo).

A.1 Distribuição Normal (NOR)

Notação:	$X \sim N(\mu,\sigma)$
Parâmetros:	μ e σ
Função densidade de probabilidade:	$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad -\infty < x < +\infty$
Função acumulada de probabilidade:	$\Phi\left(\frac{x-\mu}{\sigma}\right) \qquad \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt$
Função de quantis:	Não tem forma analítica explícita
Média:	$E[X] = \mu$
Variância:	$Var[X] = \sigma^2$
Coeficiente de assimetria:	$\gamma = 0$
Curtose:	$\kappa = 3$

Estimação dos parâmetros:

- Método MOM:

  $\hat{\mu}_X = \bar{x}$

  $\hat{\sigma}_X = s_X$

- Método MVS:

  $\hat{\mu}_X = \bar{x}$

  $\hat{\sigma}_X = s_X$

- Método MML:

  $\hat{\mu}_X = l_1$

  $\hat{\sigma}_X = \sqrt{\pi} l_2$

A.2 Distribuição Exponencial (EXP)

Notação:	$X \sim E(\theta)$
Parâmetros:	θ
Função densidade de probabilidade:	$f_X(x) = \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right) \qquad x \geq 0$
Função acumulada de probabilidade:	$F_X(x) = 1 - \exp\left(-\frac{x}{\theta}\right)$
Função de quantis:	$x(F) = -\theta \ln(1-F)$
Média:	$E[X] = \theta$
Variância:	$Var[X] = \theta^2$
Coeficiente de assimetria:	$\gamma = 2$
Curtose:	$\kappa = 9$

Estimação dos parâmetros:

- Método MOM:

  $\hat{\theta} = \bar{x}$

- Método MVS:

  $\hat{\theta} = \bar{x}$

- Método MML:

  $\hat{\theta} = l_1$

A.3 Distribuição Generalizada de Valores Extremos (GEV)

Notação:	$X \sim GEV(\alpha, \beta, k)$
Parâmetros:	α , β e k
Função densidade de probabilidade:	$f_X(x) = \frac{1}{\alpha}\left[1 - k\left(\frac{x-\beta}{\alpha}\right)\right]^{1/k-1} \exp\left\{-\left[1 - k\left(\frac{x-\beta}{\alpha}\right)\right]^{1/k}\right\}$ se $k \neq 0$ Amplitude: $-\infty < x \leq \beta + \alpha/k$ se $k > 0$ (Tipo III ou Weibull) $\beta + \alpha/k \leq x < +\infty$ se $k < 0$ (Tipo II ou Fréchet) $f_X(x) = \frac{1}{\alpha}\exp\left[-\frac{x-\beta}{\alpha} - \exp\left(-\frac{x-\beta}{\alpha}\right)\right]$ se $k = 0$ Amplitude: $-\infty < x < +\infty$ (Tipo I ou Gumbel)

Função acumulada de probabilidade:	$F_X(x) = \exp\left\{-\left[1-k\left(\frac{x-\beta}{\alpha}\right)\right]^{1/k}\right\}$ se $k \neq 0$ $F_X(x) = \exp\left[-\exp\left(-\frac{x-\beta}{\alpha}\right)\right]$ se $k = 0$
Função de quantis:	$x(F) = \beta + \frac{\alpha}{k}\{1-[-\ln(F)]^k\}$ se $k \neq 0$ $x(F) = \beta - \alpha \ln[-\ln(F)]$ se $k = 0$
Média:	$E[X] = \beta + \frac{\alpha}{k}[1-\Gamma(1+k)]$
Variância:	$Var[X] = \left(\frac{\alpha}{k}\right)^2 [\Gamma(1+2k) - \Gamma^2(1+k)]$
Coeficiente de assimetria:	$\gamma = \langle \text{sinal de } k \rangle \frac{-\Gamma(1+3k) + 3\Gamma(1+k)\Gamma(1+2k) - 2\Gamma^3(1+k)}{[\Gamma(1+2k) - \Gamma^2(1+k)]^{3/2}}$

Caso especial: distribuição Gumbel $(k = 0)$

Estimação dos parâmetros:

- Método MOM:

  Alternativa 1:

  Resolver para k a equação do coeficiente de assimetria, substituindo γ pelo seu valor amostral g_X. A solução é iterativa, pelo método de Newton.

  Alternativa 2:

  Para coeficientes de assimetria amostrais $1{,}1396 < g_X < 10$:

  $$\hat{k} = 0{,}2858221 - 0{,}357983 g_X + 0{,}116659 g_X^2 - 0{,}022725 g_X^3 + \\ + 0{,}002604 g_X^4 - 0{,}000161 g_X^5 + 0{,}000004 g_X^6$$

  Para coeficientes de assimetria amostrais $-2 < g_X < 1{,}1396$:

  $$\hat{k} = 0{,}277648 - 0{,}322016 g_X + 0{,}060278 g_X^2 + 0{,}016759 g_X^3 - \\ - 0{,}005873 g_X^4 - 0{,}00244 g_X^5 - 0{,}00005 g_X^6$$

  Para coeficientes de assimetria amostrais $-10 < g_X < 0$:

  $$\hat{k} = -0{,}50405 - 0{,}00861 g_X + 0{,}015497 g_X^2 + 0{,}005613 g_X^3 + \\ + 0{,}00087 g_X^4 + 0{,}000065 g_X^5$$

  Em seguida, $\hat{\alpha} = \frac{s_X \hat{k}}{\sqrt{\Gamma(1+2\hat{k}) - \Gamma^2(1+\hat{k})}}$ e $\hat{\beta} = \bar{x} - \frac{\hat{\alpha}}{\hat{k}}[1-\Gamma(1+\hat{k})]$.

- Método MVS:

  Os estimadores $\hat{\alpha}$, $\hat{\beta}$ e $\hat{k}$ são as soluções simultâneas (obtidas pelo método de Newton) do seguinte sistema de equações:

$$\frac{1}{\alpha}\left[\sum_{i=1}^{N}\exp(-y_i - ky_i) - (1-k)\sum_{i=1}^{N}\exp(ky_i)\right] = 0$$

$$\frac{1}{k\alpha}\left[\sum_{i=1}^{N}\exp(-y_i - ky_i) - (1-k)\sum_{i=1}^{N}\exp(ky_i) + N - \sum_{i=1}^{N}\exp(-y_i)\right] = 0$$

$$\frac{1}{k^2}\left[\sum_{i=1}^{N}\exp(-y_i - ky_i) - (1-k)\sum_{i=1}^{N}\exp(ky_i) + N - \sum_{i=1}^{N}\exp(-y_i)\right] +$$
$$+\frac{1}{k}\left[-\sum_{i=1}^{N}y_i + \sum_{i=1}^{N}y_i\exp(y_i) + N\right] = 0$$

  onde $y_i = -\frac{1}{k}\ln\left[1 - k\left(\frac{x_i - \beta}{\alpha}\right)\right]$

  A solução desse sistema é complexa; sugere-se a referência Prescott e Walden (1983) para algoritmo de resolução.

- Método MML:

$$\hat{k} = 7{,}8590C + 2{,}9554C^2, \text{ onde } C = \frac{2}{3+t_3} - \frac{\ln 2}{\ln 3}$$

$$\hat{\alpha} = \frac{l_2\hat{k}}{(1-2^{-\hat{k}})\Gamma(1+\hat{k})}$$

$$\hat{\beta} = l_1 - \frac{\hat{\alpha}}{\hat{k}}[1-\Gamma(1+\hat{k})]$$

A.4 Distribuição Gumbel (GUM)

Notação:	$X \sim GumMax(\alpha, \beta)$
Parâmetros:	α e β
Função densidade de probabilidade:	$f_X(x) = \frac{1}{\alpha}\exp\left[-\frac{x-\beta}{\alpha} - \exp\left(-\frac{x-\beta}{\alpha}\right)\right]$ Amplitude: $-\infty < x < +\infty$
Função acumulada de probabilidade:	$F_X(x) = \exp\left[-\exp\left(-\frac{x-\beta}{\alpha}\right)\right]$
Função de quantis:	$x(F) = \beta - \alpha\ln[-\ln(F)]$

Média:	$E[X] = \beta + 0{,}5772\alpha$
Variância:	$Var[X] = \frac{\pi^2 \alpha^2}{6}$
Coeficiente de assimetria:	$\gamma = 1{,}1396$
Curtose:	$\kappa = 5{,}4$

Estimação dos parâmetros:

- Método MOM:

  $\hat{\alpha} = 0{,}7797 s_X$

  $\hat{\beta} = \bar{x} - 0{,}45 s_X$

- Método MVS:

  Os estimadores $\hat{\alpha}$ e $\hat{\beta}$ são as soluções do seguinte sistema de equações:

  $$\frac{\partial}{\partial \alpha} \ln[L(\alpha, \beta)] = -\frac{N}{\alpha} + \frac{1}{\alpha^2} \sum_{i=1}^{N} (x_i - \beta) - \frac{1}{\alpha^2} \sum_{i=1}^{N} (x_i - \beta) \exp\left(-\frac{x_i - \beta}{\alpha} \right) = 0$$

  $$\frac{\partial}{\partial \beta} \ln[L(\alpha, \beta)] = \frac{N}{\alpha} - \frac{1}{\alpha} \sum_{i=1}^{N} \exp\left(-\frac{x_i - \beta}{\alpha} \right) = 0$$

  Manipulando ambas as equações, chega-se a:

  $$F(\alpha) = \sum_{i=1}^{N} x_i \exp\left(-\frac{x_i}{\alpha} \right) - \frac{1}{N} \sum_{i=1}^{N} (x_i - \alpha) \sum_{i=1}^{N} \exp\left(-\frac{x_i}{\alpha} \right) = 0$$

  cuja solução, pelo método de Newton, fornece $\hat{\alpha}$.

  Em seguida, $\hat{\beta} = \hat{\alpha} \ln\left[N \Big/ \sum_{i=1}^{N} \exp(-x_i / \hat{\alpha}) \right]$.

- Método MML:

  $\hat{\alpha} = \frac{l_2}{\ln 2}$

  $\hat{\beta} = l_1 - 0{,}5772\hat{\alpha}$

A.5 Distribuição Log-Normal 2 parâmetros (LN2)

A distribuição Log-Normal 2 parâmetros da variável *X* refere-se à distribuição Normal da variável transformada $Y = \ln(X)$.

Notação:	$X \sim LN(\mu_Y, \sigma_Y)$
Parâmetros:	μ_Y e σ_Y, com $Y = \ln(X)$
Função densidade de probabilidade:	$f_X(x) = \frac{1}{x\sigma_Y\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{\ln(x)-\mu_Y}{\sigma_Y}\right]^2\right\} \qquad x > 0$
Função acumulada de probabilidade:	$\Phi\left(\frac{\ln(x)-\mu_Y}{\sigma_Y}\right) \qquad \Phi(y) = \int_{-\infty}^{y} \frac{1}{\sigma_Y\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt$
Função de quantis:	Não tem forma analítica explícita
Média:	$E[X] = \mu_X = \exp\left[\mu_Y + \frac{\sigma_Y^2}{2}\right]$
Variância:	$Var[X] = \sigma_X^2 = \mu_X^2[\exp(\sigma_Y^2) - 1]$
Coeficiente de variação:	$CV_X = \sqrt{\exp(\sigma_Y^2) - 1}$
Coeficiente de assimetria:	$\gamma = 3\,CV_X + CV_X^3$
Curtose:	$\kappa = 3 + (e^{\sigma_Y^2} - 1)(e^{3\sigma_Y^2} + 3e^{2\sigma_Y^2} + 6e^{\sigma_Y^2} + 6)$

Estimação dos parâmetros:

- Método MOM:

  $\hat{\sigma}_Y = \sqrt{\ln(CV_X^2 + 1)}$

  $\hat{\mu}_Y = \ln(\bar{x}) - \frac{\hat{\sigma}_Y^2}{2}$

- Método MVS:

  $\hat{\mu}_Y = \bar{y}$

  $\hat{\sigma}_Y = s_Y$

- Método MML:

  $\hat{\sigma}_Y = 2 \cdot erf^{-1}(t)$

  $\hat{\mu}_Y = \ln(l_1) - \frac{\hat{\sigma}_Y^2}{2}$

  onde $erf(w) = \frac{2}{\sqrt{\pi}} \int_0^w e^{-u^2} du$. A inversa $erf^{-1}(t)$ é igual a $u/\sqrt{2}$, com u representando a variável Normal padrão correspondente $\Phi\left(\frac{t+1}{2}\right)$.

A.6 Distribuição Pearson III (PE3)

Notação:	$X \sim PE3(\alpha, \beta, \xi)$
Parâmetros:	α, β e ξ
Função densidade de probabilidade:	$f_X(x) = \frac{1}{\alpha\Gamma(\beta)}\left(\frac{x-\xi}{\alpha}\right)^{\beta-1} \exp\left(-\frac{x-\xi}{\alpha}\right) \qquad \xi \le x < +\infty$
Função acumulada de probabilidade:	$F_X(x) = G\left(\beta, \frac{x-\xi}{\alpha}\right) \Big/ \Gamma(\beta) \qquad G(\beta, x) = \int_0^x t^{\beta-1} e^{-t} dt$
Função de quantis:	Não tem forma analítica explícita
Média:	$E[X] = \alpha\beta + \xi$
Variância:	$Var[X] = \alpha^2 \beta$
Coeficiente de assimetria:	$\gamma = \frac{2}{\sqrt{\beta}}$
Curtose:	$\kappa = 3 + \frac{6}{\sqrt{\beta}}$

Caso especial: distribuição Exponencial $(\beta = 1)$

Estimação dos parâmetros:

- Método MOM:

$$\hat{\beta} = \left(\frac{2}{g_X}\right)^2$$

$$\hat{\alpha} = \sqrt{\frac{s_X^2}{\hat{\beta}}}$$

$$\hat{\xi} = \bar{x} - \sqrt{s_X^2 \hat{\beta}}$$

- Método MVS:

  Os estimadores $\hat{\alpha}$, $\hat{\beta}$ e $\hat{\xi}$ são as soluções (obtidas pelo método de Newton) do seguinte sistema de equações:

$$\sum_{i=1}^{N} (x_i - \xi) = N\alpha\beta$$

$$N\Psi(\beta) = \sum_{i=1}^{N} \ln[(x_i - \xi)/\alpha]$$

$$N = \alpha(\beta - 1)\sum_{i=1}^{N} \frac{1}{x_i - \xi}$$

onde $\Psi(\beta) = \frac{\Gamma'(\beta)}{\Gamma(\beta)} \cong \ln\beta - \frac{1}{2\beta} - \frac{1}{12\beta^2} + \frac{1}{120\beta^4} - \frac{1}{252\beta^6} + \frac{1}{240\beta^8} - \frac{1}{132\beta^{10}}$

- Método MML:

  Para $t_3 \geq 1/3$ e com $t_m = 1 - t_3$, $\hat{\beta} = \frac{0{,}36067 t_m - 0{,}5967 t_m^2 + 0{,}25361 t_m^3}{1 - 2{,}78861 t_m + 2{,}56096 t_m^2 - 0{,}77045 t_m^3}$

  Para $t_3 < 1/3$ e com $t_m = 3\pi t_3^2$, $\hat{\beta} = \frac{1 + 0{,}2906 t_m}{t_m + 0{,}1882 t_m^2 + 0{,}0442 t_m^3}$

  $\hat{\alpha} = l_2 \sqrt{\pi} \frac{\Gamma(\hat{\beta})}{\Gamma(\hat{\beta} + 0{,}5)}$

  $\hat{\xi} = l_1 - \hat{\alpha}\hat{\beta}$

A.7 Distribuição Log-Pearson III (LP3)

A distribuição Log-Pearson III da variável *X* refere-se à distribuição Pearson III da variável transformada $Z = \ln(X)$.

Estimação dos parâmetros:

- Método MOM:

  Lembrando que $\mu'_r = \frac{\exp(\xi r)}{(1 - r\alpha)^\beta}$ são estimados por m'_r, os estimadores $\hat{\alpha}$, $\hat{\beta}$ e $\hat{\xi}$ são as soluções de:

  $\ln(m'_1) = \xi - \beta \ln(1 - \alpha)$

  $\ln(m'_2) = 2\xi - \beta \ln(1 - 2\alpha)$

  $\ln(m'_3) = 3\xi - \beta \ln(1 - 3\alpha)$

  Para a solução desse sistema, Kite (1977) sugere:

  seja $B = \frac{\ln(m'_3) - 3\ln(m'_1)}{\ln(m'_2) - 2\ln(m'_1)}$, $A = \frac{1}{\alpha} - 3$ e $C = \frac{1}{B - 3}$

  para $3{,}5 < B < 6{,}0$; $A = -0{,}23019 + 1{,}65262 C + 0{,}20911 C^2 - 0{,}04557 C^3$

  para $3{,}0 < B \leq 3{,}5$; $A = -0{,}47157 + 1{,}99955 C$

  em seguida, $\hat{\alpha} = \frac{1}{A + 3}$, $\hat{\beta} = \frac{\ln(m'_2) - 2\ln(m'_1)}{\ln(1 - \hat{\alpha})^2 - \ln(1 - 2\hat{\alpha})}$ e $\hat{\xi} = \ln(m'_1) + \hat{\beta}\ln(1 - \hat{\alpha})$.

- Método MVS:

  Os estimadores $\hat{\alpha}$, $\hat{\beta}$ e $\hat{\xi}$ são as soluções (obtidas pelo método de Newton) do seguinte sistema de equações:

  $$\sum_{i=1}^{N}(\ln x_i - \xi) = N\alpha\beta$$

  $$N\Psi(\beta) = \sum_{i=1}^{N}\ln[(\ln x_i - \xi)/\alpha]$$

  $$N = \alpha(\beta - 1)\sum_{i=1}^{N}\frac{1}{\ln x_i - \xi}$$

  onde $\Psi(\beta) = \frac{\Gamma'(\beta)}{\Gamma(\beta)} \cong \ln\beta - \frac{1}{2\beta} - \frac{1}{12\beta^2} + \frac{1}{120\beta^4} - \frac{1}{252\beta^6} + \frac{1}{240\beta^8} - \frac{1}{132\beta^{10}}$

- Método MML:

  As estimativas pelo método dos momentos-L podem ser obtidas por procedimento idêntico ao apresentado para a distribuição Pearson III, com a transformação $z_i = \ln(x_i)$.

B DISTRIBUIÇÃO BAYESIANA DE PROBABILIDADES DE UMA VARIÁVEL ALEATÓRIA MODELADA PELA DISTRIBUIÇÃO NORMAL

Esse anexo, adaptado de Wood *et al.* (1974), apêndice A, apresenta a demonstração de que a distribuição bayesiana de probabilidades de uma variável aleatória Q, modelada pela distribuição Normal, com média μ e variância $\varphi = \sigma^2$ desconhecidas, é *t* de Student.

Conforme mostrado na equação (4.16), tem-se:

$$\tilde{f}(q) = \int_{\mu}\int_{\varphi} f(q \mid \mu,\varphi)\cdot f_1(\mu,\varphi \mid Q)\, d\varphi\, d\mu$$

$$\tilde{f}(q) = \int_{\mu}\int_{\varphi} f(q \mid \mu,\varphi)\cdot f_1(\mu \mid \varphi,Q)\cdot f_1(\varphi \mid Q)\, d\varphi\, d\mu \qquad \text{(B.1)}$$

Para maior simplicidade, no decorrer da demonstração será omitido o índice "1" nos termos referentes à distribuição *a posteriori* dos parâmetros μ e φ. Mudando a ordem de integração na equação (B.1), tem-se:

$$\tilde{f}(q) = \int_{\varphi}\left[\int_{\mu} f(q \mid \mu,\varphi)\cdot f(\mu \mid \varphi,Q)\, d\mu\right]\cdot f(\varphi \mid Q)\, d\varphi \qquad \text{(B.2)}$$

Resolvendo para $\tilde{f}(q \mid \varphi) = \int_{\mu} f(q \mid \mu,\varphi)\cdot f(\mu \mid \varphi,Q)\, d\mu$

e substituindo por $f(q \mid \mu,\varphi) = N[\mu,\varphi]$ e $f(\mu \mid \varphi,Q) = N\left[m,\frac{\varphi}{n}\right]$, tem-se:

$$\tilde{f}(q \mid \varphi) = \int_{\mu}\frac{1}{\sqrt{2\pi}}\varphi^{-1/2}\exp\left[-\frac{1}{2\varphi}(q-\mu)^2\right]\cdot\frac{1}{\sqrt{2\pi}}\left(\frac{\varphi}{n}\right)^{-1/2}\exp\left[-\frac{n}{2\varphi}(\mu-m)^2\right]d\mu$$

$$= \frac{1}{\sqrt{2\pi}}n^{1/2}\varphi^{-1/2}\int_{\mu}\frac{1}{\sqrt{2\pi}}\varphi^{-1/2}\exp\left\{-\frac{1}{2\varphi}\underbrace{\left[(q-\mu)^2+n(\mu-m)^2\right]}_{A}\right\}d\mu$$

$$= \frac{1}{\sqrt{2\pi}}n^{1/2}\varphi^{-1/2}\int_{\mu}\frac{1}{\sqrt{2\pi}}\varphi^{-1/2}\exp\left\{-\frac{1}{2\varphi}\underbrace{\left[\frac{n}{n+1}(q-m)^2+(n+1)\left(\mu-\frac{(q+nm)}{n+1}\right)^2\right]}_{B}\right\}d\mu \qquad \text{(B.3)}$$

Nesse momento, considera-se oportuno provar que $B = A$:

$$
\begin{aligned}
B &= \frac{n}{n+1}(q-m)^2 + (n+1)\left(\mu - \frac{(q+nm)}{n+1}\right)^2 \\
&= \frac{nq^2 - 2nqm + nm^2}{n+1} + (n+1)\left(\mu^2 - \frac{2\mu(q+nm)}{n+1} + \frac{(q+nm)^2}{(n+1)^2}\right) \\
&= \frac{nq^2 - 2nqm + nm^2}{n+1} + (n+1)\mu^2 - 2\mu(q+nm) + \frac{q^2 + 2nqm + n^2m^2}{n+1} \\
&= \frac{nq^2 + q^2 + n^2m^2 + nm^2}{n+1} + n\mu^2 + \mu^2 - 2\mu q - 2n\mu m \\
&= \frac{(n+1)q^2 + (n+1)nm^2}{n+1} + n\mu^2 + \mu^2 - 2\mu q - 2n\mu m \\
&= q^2 - 2\mu q + \mu^2 + n\mu^2 - 2n\mu m + nm^2 \\
&= (q-\mu)^2 + n(\mu - m)^2 = A
\end{aligned}
$$

Resolvendo a equação (B.3):

$$
\begin{aligned}
\tilde{f}(q \mid \varphi) &= \left(\frac{n+1}{n+1}\right)^{1/2} \cdot \frac{1}{\sqrt{2\pi}} n^{1/2} \varphi^{-1/2} \cdot \\
&\qquad \cdot \int_{\mu} \frac{1}{\sqrt{2\pi}} \varphi^{-1/2} \exp\left\{-\frac{1}{2\varphi}\left[\frac{n}{n+1}(q-m)^2 + (n+1)\left(\mu - \frac{(q+nm)}{n+1}\right)^2\right]\right\} d\mu \\
&= \frac{1}{\sqrt{2\pi}}\left(\frac{n}{n+1}\right)^{1/2} \varphi^{-1/2} \exp\left[-\frac{1}{2\varphi} \cdot \frac{n}{n+1}(q-m)^2\right] \cdot \\
&\qquad \cdot \int_{\mu} \frac{1}{\sqrt{2\pi}}\left(\frac{\varphi}{n+1}\right)^{-1/2} \exp\left[-\frac{(n+1)}{2\varphi}\left(\mu - \frac{(q+nm)}{n+1}\right)^2\right] d\mu \\
&= \frac{1}{\sqrt{2\pi}}\left(\frac{\varphi}{r}\right)^{-1/2} \exp\left[-\frac{r}{2\varphi}(q-m)^2\right] \cdot \int_{\mu} f_N\left(\frac{(q+nm)}{n+1}, \frac{\varphi}{n+1}\right) d\mu \quad \rightarrow 1
\end{aligned}
$$

$$
\tilde{f}(q \mid \varphi) = \frac{1}{\sqrt{2\pi}}\left(\frac{\varphi}{r}\right)^{-1/2} \exp\left[-\frac{r}{2\varphi}(q-m)^2\right] = f_N\left(m, \frac{\varphi}{r}\right) \qquad \text{(B.4)}
$$

onde: $r = \dfrac{n}{n+1}$

Levando (B.4) em (B.2) e sabendo-se que $\varphi \sim s^2 \nu \chi_\nu^{-2}$, tem-se:

$$
\tilde{f}(q) = \int_{\varphi} \tilde{f}(q \mid \varphi) \cdot f(\varphi \mid Q)\, d\varphi
$$

$$
\tilde{f}(q) = \int_{\varphi} f_N(m, \varphi / r) \cdot f_{\chi^{-2}}(s^2, \nu)\, d\varphi \qquad \text{(B.5)}
$$

Resolvendo a equação (B.5):

$$\tilde{f}(q) = \int_{\varphi} \frac{1}{\sqrt{2\pi}} \left(\frac{\varphi}{r}\right)^{-1/2} \exp\left[-\frac{r}{2\varphi}(q-m)^2\right] \cdot \frac{(\frac{1}{2}s^2\nu)^{\nu/2}}{\Gamma(\nu/2)} \varphi^{-(\nu/2)-1} \exp\left[-\frac{1}{2\varphi}s^2\nu\right] d\varphi \qquad \text{(B.6)}$$

Substituindo $\nu' = \nu + 1$ e $s'^2 = \frac{r(q-m)^2 + s^2\nu}{\nu'}$ na equação (B.6), tem-se:

$$\begin{aligned}
\tilde{f}(q) &= \frac{(\frac{1}{2}s'^2\nu')^{\nu'/2}}{(\frac{1}{2}s'^2\nu')^{\nu'/2}} \cdot \frac{\Gamma(\nu'/2)}{\Gamma(\nu'/2)} \cdot \int_{\varphi} \frac{r^{1/2}}{\sqrt{2\pi}} \cdot \frac{(\frac{1}{2}s^2\nu)^{\nu/2}}{\Gamma(\nu/2)} \varphi^{-(\nu'/2)-1} \exp\left[-\frac{1}{2\varphi}s'^2\nu'\right] d\varphi \\
&= \frac{r^{1/2}}{\sqrt{2\pi}} \cdot \frac{(\frac{1}{2}s^2\nu)^{\nu/2}}{\Gamma(\nu/2)} \cdot \frac{\Gamma(\nu'/2)}{(\frac{1}{2}s'^2\nu')^{\nu'/2}} \cdot \int_{\varphi} \frac{(\frac{1}{2}s'^2\nu')^{\nu'/2}}{\Gamma(\nu'/2)} \varphi^{-(\nu'/2)-1} \exp\left[-\frac{1}{2\varphi}s'^2\nu'\right] d\varphi \\
&= \frac{r^{1/2}}{\sqrt{2\pi}} \cdot \frac{(\frac{1}{2}s^2\nu)^{\nu/2}}{\Gamma(\nu/2)} \cdot \frac{\Gamma(\nu/2 + \frac{1}{2})}{\{\frac{1}{2}[r(q-m)^2 + s^2\nu]\}^{(\nu+1)/2}} \cdot \int_{\varphi} f_{\chi^{-2}}(s'^2, \nu')\, d\varphi \\
&= \frac{1}{B(\nu/2, \frac{1}{2})} r^{1/2} (s^2\nu)^{\nu/2} \left[\frac{s^2\nu}{s^2\nu} \cdot [r(q-m)^2 + s^2\nu]\right]^{-(\nu+1)/2} \\
&= \frac{1}{B(\nu/2, \frac{1}{2})} r^{1/2} (s^2\nu)^{\nu/2} (s^2\nu)^{-(\nu+1)/2} \left[\frac{r}{s^2\nu}(q-m)^2 + 1\right]^{-(\nu+1)/2}
\end{aligned}$$

$$\tilde{f}(q) = B(\nu/2, \frac{1}{2})^{-1} \cdot \left[1 + \frac{(q-m)^2}{(s^2\nu)/r}\right]^{-(\nu+1)/2} \cdot \left(\frac{s^2\nu}{r}\right)^{-1/2} \qquad \text{(B.7)}$$

onde: $B(\nu/2, \frac{1}{2}) = \frac{\pi^{1/2}\Gamma(\nu/2)}{\Gamma(\nu/2 + \frac{1}{2})}$ e $r = \frac{n}{n+1}$

Portanto, a distribuição $\tilde{f}(q)$, mostrada na equação (B.7), é *t* de Student, com momentos:

$$E[q] = m$$

$$VAR[q] = \frac{s^2}{r} \cdot \frac{\nu}{(\nu-2)} = s^2 \cdot \frac{(n+1)}{n} \cdot \frac{\nu}{(\nu-2)}$$

Em conseqüência, a variável aleatória Q será distribuída conforme a equação (B.8):

$$\frac{q-m}{s/\sqrt{n/(n+1)}} \sim t_{\nu} \qquad \text{(B.8)}$$

C FONTES DE INFORMAÇÕES HISTÓRICAS SOBRE EVENTOS HIDROLÓGICOS E HIDROMETEOROLÓGICOS

Durante o desenvolvimento dessa pesquisa, as principais fontes de informações históricas sobre eventos hidrológicos e hidrometeorológicos foram o Arquivo Público Mineiro e a Hemeroteca Histórica. Embora outros órgãos, tais como o Centro de Estudos Históricos e Culturais da Fundação João Pinheiro (CEHC/FJP) e o Instituto de Geociências Aplicadas (IGA/MG) também sejam importantes fontes de informação, os dois primeiros constituem o ponto de partida para investigações futuras mais profundas.

Um breve histórico do Arquivo Público Mineiro e da Hemeroteca Pública é apresentado a seguir. As informações foram extraídas do site oficial da Secretaria de Estado da Cultura de MG (www.cultura.mg.gov.br). Algumas características dos documentos encontrados também são descritas.

C.1 Arquivo Público Mineiro

O Arquivo Público Mineiro tem por finalidade executar a gestão, o recolhimento, a guarda, a preservação e o acesso ao acervo arquivístico da Administração Pública Estadual e aos documentos privados de interesse público. Criado em 11 de Julho de 1895, pela Lei n. 126, o Arquivo Público Mineiro foi instituído, em Ouro Preto, como repartição destinada a receber e conservar todos os documentos concernentes ao direito público, à legislação, à administração, à história, à geografia e às manifestações do movimento científico, literário e artístico do estado de Minas Gerais.

Instalado, inicialmente, na residência de seu fundador e primeiro diretor, José Pedro Xavier da Veiga, o Arquivo teve como acervo original documentos recolhidos de diversas repartições do Estado e de particulares. Sua transferência definitiva para Belo Horizonte ocorreu somente em fins de 1901, passando a denominar-se Diretoria de Arquivo e Estatística, como 5ª Seção da Secretaria do Interior. Em anos posteriores, a instituição esteve subordinada às Secretarias de Educação e de Governo e, a partir de 1984, passou a integrar a estrutura da Secretaria de Estado da Cultura.

O Arquivo Público Mineiro está instalado, desde 1938, na antiga sede da Prefeitura de Belo Horizonte, edificação destinada originalmente à residência do Secretário das Finanças. Em Março de 1975, incorporou o prédio contíguo (Rua Aimorés, n. 1.450), cuja função é abrigar o acervo, a administração e o atendimento ao público. O prédio, construído em 1897 e remanescente do núcleo original de Belo Horizonte, foi projetado na Seção de Arquitetura da Comissão Construtora da Nova Capital, e encontra-se hoje tombado por decreto estadual.

O acervo do Arquivo Público Mineiro é constituído por aproximadamente 1.400 metros lineares de documentos produzidos e acumulados por órgãos da Administração Pública de Minas Gerais e diversos arquivos privados, abrangendo os séculos XVIII, XIX e parte do século XX. Além de documentos manuscritos e impressos, reúne mapas, plantas, fotografias, gravuras, filmes, livros, folhetos e periódicos. A Biblioteca, instituída com a criação do Arquivo, completa o acervo documental custodiado. Possui entre seus títulos uma valiosa coleção de obras raras e preciosas que remonta ao século XVI.

A instituição é ainda responsável pela publicação da Revista do Arquivo Público Mineiro, que apresenta assuntos ligados à História de Minas Gerais e do Brasil, transcrições de documentos do acervo, instrumentos de pesquisa (como inventários, catálogos, repertórios e bibliografias), biografias e monografias diversas.

Dentre os vários documentos de interesse para a presente pesquisa encontrados no Arquivo Público Mineiro, estão os relatórios da Comissão Geográfica e Geológica do Estado, produzidos no final do século XIX. A Comissão era vinculada à então Secretaria de Agricultura, Comércio e Obras Públicas, e seus relatórios apresentam diversas informações resultantes da exploração, àquela época, das principais regiões do Estado. As informações se referem, especialmente, às características geográficas, geológicas e climáticas dos locais explorados.

As obras dos naturalistas Auguste de Saint-Hilaire e José de Souza Azevedo Pizarro também fazem parte do acervo bibliográfico do Arquivo Público Mineiro. Elas apresentam um relato muito interessante das viagens realizadas por esses autores a várias regiões do Brasil, até meados do século XIX.

C.2 Hemeroteca Histórica

A Hemeroteca Histórica de Minas Gerais, criada pela Lei n. 12.221, de 01 de Julho de 1996, integra a estrutura da Biblioteca Pública Estadual e hoje funciona em prédio próprio, na Avenida Assis Chateaubriand, n. 167, em Belo Horizonte.

O seu valioso acervo, composto por jornais, revistas e periódicos raros, registra a história de Minas Gerais desde meados do século XIX, relatando, sob um viés jornalístico, os mais diversos acontecimentos daquela época, sejam eles políticos, sociais, econômicos, históricos, científicos, ou relacionados à ocorrência de fenômenos naturais marcantes. As Figuras C.1 e C.2 mostram alguns periódicos encontrados na Hemeroteca Histórica. Pode-se observar que, apesar de bastante antigos, os documentos encontram-se em bom estado de conservação.

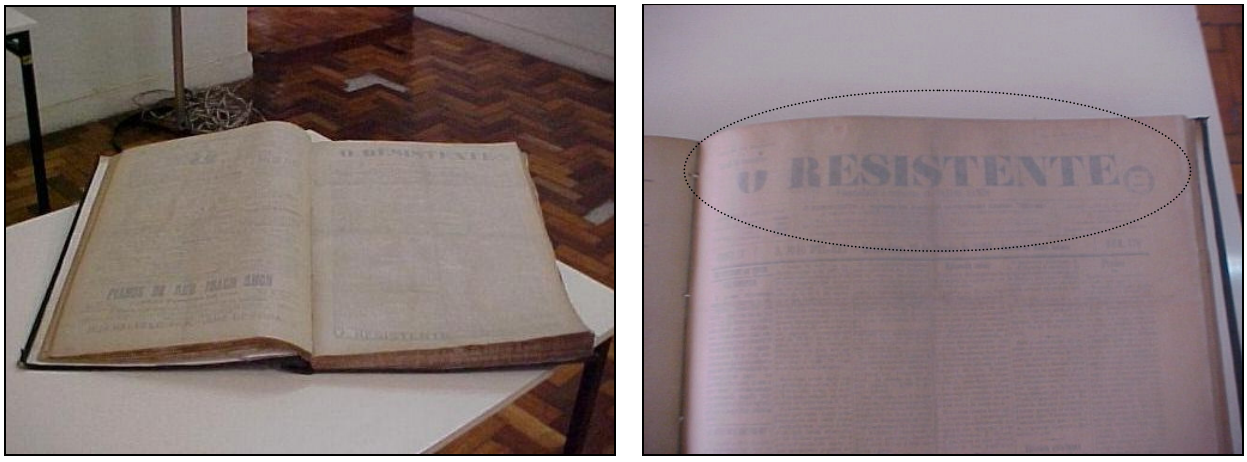


FIGURA C.1: Jornal *O Resistente* – São João del Rei, final do século XIX.

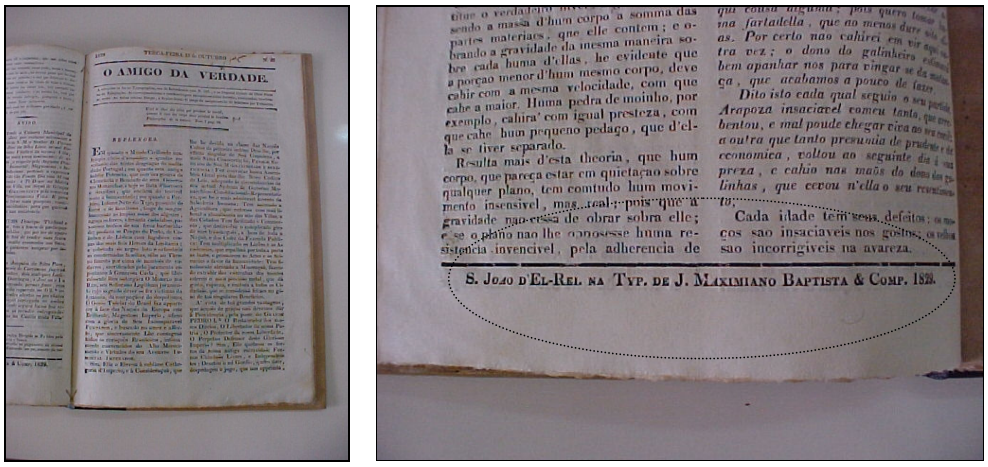


FIGURA C.2: Jornal *O Amigo da Verdade* – São João del Rei, 1829.

D ANÁLISE REGIONAL DE FREQÜÊNCIA DE VARIÁVEIS HIDROLÓGICAS: SÍNTESE DA METODOLOGIA DOS MOMENTOS-L

O texto que se segue apresenta uma síntese da metodologia dos momentos-L para a análise regional de freqüência de variáveis hidrológicas e foi extraído de Davis e Naghettini (2001).

D.1 Princípios do método index-flood ("cheia-índice")

O termo *index-flood* ("cheia-índice") foi introduzido por Dalrymple (1960), em um contexto de regionalização de vazões de cheia. Trata-se de um expediente para adimensionalizar quaisquer dados obtidos em pontos distintos de uma região considerada homogênea, com a finalidade de utilizá-los como um conjunto amostral único. Apesar de possuir referências a enchentes, o método e o termo *index-flood* têm uso consagrado em estudos de regionalização de freqüência de qualquer tipo de variável.

Seja o caso de se regionalizar as freqüências de uma variável genérica *X*, cuja variabilidade espaço-temporal foi amostrada em *N* locais ou postos de observação, situados em uma certa região geográfica. As observações, tomadas nos postos indexados por *i*, formam amostras de tamanho variável n_i e são denotadas por $X_{i,j}$, $i = 1,2,...,N$, $j = 1,2,...,n_i$. Se F $(0 < F < 1)$ representa a distribuição de freqüência da variável *X* no posto *i*, então a função de quantis nesse local é simbolizada por $X_i(F)$. A hipótese básica do método *index-flood* é a de que os postos formam uma região homogênea, ou seja, que as distribuições de freqüência nos *N* pontos são idênticas, a menos de um fator de escala local denominado *index-flood*. Formalmente:

$$X_i(F) = \mu_i \cdot x(F) \qquad i = 1,2,...,N \tag{D.1}$$

onde μ_i denota o *index-flood* do local *i* e $x(F)$ representa a curva regional de quantis adimensionais, algumas vezes denominada curva regional de crescimento, comum a todos os postos. O fator de escala μ_i pode ser estimado por qualquer medida de posição ou tendência central da amostra de observações $(X_{i,1}, X_{i,2},..., X_{i,n_i})$. Por conveniência matemática, Hosking e Wallis (1997) utilizam como estimador do *index-flood* a média aritmética das observações no posto *i*, ou seja, $\hat{\mu}_i = \overline{X}_i$.

Os dados adimensionais padronizados $x_{i,j} = X_{i,j} / \hat{\mu}_i$, $i = 1,2,...,N$, $j = 1,2,...,n_i$, formam a base para se estimar a curva regional de quantis adimensionais $x(F)$. A forma de $x(F)$ é supostamente conhecida a menos dos *p* parâmetros $\theta_1, \theta_2, ..., \theta_p$, os quais são próprios da distribuição *F* e, em geral, funções das características populacionais de posição, dispersão e assimetria.

Hosking e Wallis (1997) propõem que os parâmetros da curva regional de quantis adimensionais, agora denotada por $x(F; \theta_1, \theta_2, ..., \theta_p)$, sejam os resultados da ponderação dos parâmetros locais $\hat{\theta}_k^{(i)}$, $k = 1,2,..., p$, estimados separadamente para cada posto *i*. Portanto, a estimativa do parâmetro regional θ_k^R é dada por:

$$\hat{\theta}_k^R = \frac{\sum_{i=1}^{N} n_i \cdot \hat{\theta}_k^{(i)}}{\sum_{i=1}^{N} n_i} \tag{D.2}$$

Com essas estimativas em $x(F)$, pode-se obter a estimativa da curva regional de quantis adimensionais $\hat{x}(F) = x(F; \hat{\theta}_1^R, \hat{\theta}_2^R, ..., \hat{\theta}_p^R)$. Inversamente, as estimativas dos quantis para o posto *i* podem ser obtidas pelo produto de $\hat{x}(F)$ por $\hat{\mu}_i$, ou seja:

$$\hat{X}_i(F) = \hat{\mu}_i \cdot \hat{x}(F) \tag{D.3}$$

As premissas inerentes ao método *index-flood* são:

a) as observações em um posto qualquer são identicamente distribuídas;

b) as observações em um posto qualquer não apresentam dependência estatística serial;

c) as observações em diferentes postos são estatisticamente independentes;

d) as distribuições de freqüência em diferentes postos são idênticas, a menos de um fator de escala; e

e) a forma matemática da curva regional de quantis adimensionalizados foi corretamente especificada.

Segundo Hosking e Wallis (1997), as premissas (a) e (b) são plausíveis para diversos tipos de variáveis, principalmente aquelas relacionadas a totais ou máximos anuais. Entretanto, é improvável que as três últimas premissas possam ser empiricamente verificadas por dados hidrológicos, meteorológicos ou ambientais. Sabe-se, por exemplo, que precipitações frontais ou estiagens severas são eventos que afetam extensas áreas. Como essas áreas podem conter vários postos de observação da variável em questão, é provável que as amostras, coletadas em pontos distintos, apresentem um grau de correlação significativo. Ainda segundo Hosking e Wallis (1997), na prática, as premissas (d) e (e) jamais são verificadas com exatidão. Apesar dessas restrições, esses autores sugerem que as premissas do método *index-flood* podem ser razoavelmente aproximadas tanto pela escolha criteriosa dos postos componentes de uma região, como também pela seleção apropriada de uma distribuição de freqüência que apresente consistência com os dados amostrais.

D.2 Etapas da análise regional de freqüência

A metodologia para análise regional de freqüência, proposta por Hosking e Wallis (1997), fundamenta-se tanto nos princípios enunciados no item D.1, como também em algumas estatísticas construídas a partir dos chamados momentos-L, descritos no Capítulo 3 dessa dissertação. Essas estatísticas constituem instrumentos valiosos para diminuir o grau de subjetividade presente nas quatro etapas usuais da análise regional de freqüência. Essas etapas encontram-se sumarizadas a seguir e serão detalhadas no próximos itens.

Etapa 1: Análise regional de consistência de dados

Essa etapa se refere à detecção e à eliminação de erros grosseiros e/ou sistemáticos eventualmente existentes nas amostras individuais dos vários postos de observação. Além das técnicas usuais de análise de consistência, como as curvas de dupla acumulação, por exemplo, Hosking e Wallis (1997) sugerem o uso de uma estatística auxiliar, denominada medida de discordância, a qual fundamenta-se na comparação das características estatísticas do conjunto de postos com aquelas apresentadas pela amostra individual em questão.

Etapa 2: Identificação de regiões homogêneas

Uma região homogênea consiste de um agrupamento de postos de observação cujas curvas de quantis adimensionalizados podem ser aproximadas por uma única curva

regional. Para se determinar a correta divisão dos postos em regiões homogêneas, Hosking e Wallis (1997) sugerem o emprego da técnica de análise de *clusters*. De acordo com essa técnica, os postos são agrupados em regiões consonantes com a variabilidade espacial de algumas características locais, as quais foram selecionadas entre aquelas que supostamente podem influir sobre a variável a ser regionalizada. Depois dos postos terem sido convenientemente agrupados em regiões, Hosking e Wallis (1997) sugerem o cálculo da medida de heterogeneidade para testar a correção dos agrupamentos efetuados. Essa medida baseia-se na comparação da variabilidade grupal das características estatísticas dos postos de observação com a variabilidade esperada dessas mesmas características em uma região homogênea.

Etapa 3: Seleção da função regional de distribuição de probabilidades

Depois dos erros grosseiros e/ou sistemáticos terem sido eliminados das amostras individuais, e das regiões homogêneas terem sido identificadas, a etapa seguinte é a correta prescrição do modelo probabilístico, representado por $x(F)$ na equação (D.1). Para a seleção da função regional de distribuição de probabilidades entre os diversos modelos candidatos, Hosking e Wallis (1997) sugerem o emprego do teste da medida de aderência. Esse teste é construído de forma que se possa comparar algumas características estatísticas regionais com aquelas que se espera obter de uma amostra aleatória simples, retirada de uma população cujas propriedades distributivas são as mesmas do modelo candidato.

Etapa 4: Estimação dos parâmetros e quantis da função regional de distribuição de probabilidades

Identificado o modelo probabilístico regional, denotado por $\hat{x}(F) = x(F;\hat{\theta}_1^R,\hat{\theta}_2^R,...,\hat{\theta}_p^R)$, os parâmetros locais $\hat{\theta}_k^{(i)}$, $k = 1,2,...,p$, são estimados separadamente para cada posto *i* e, em seguida, ponderados conforme a equação (D.2) para produzir a curva regional de quantis adimensionais. Hosking e Wallis (1997) sugerem a utilização dos chamados momentos-L para a estimação de parâmetros e quantis da função regional de distribuição de probabilidades.

Hosking e Wallis (1997) codificaram um conjunto de rotinas em linguagem Fortran-77 para automatização das quatro etapas da metodologia proposta para análise regional de freqüência.

Esse conjunto de rotinas encontra-se disponibilizado ao público no repositório de programas Statlib, acessível via Internet através da URL http://lib.stat.cmu.edu/general/lmoments.

D.3 Análise regional de consistência de dados

A primeira etapa da análise regional de freqüência de variáveis aleatórias é certificar-se: (1) que os dados coletados em qualquer um dos postos de observação estão isentos de erros grosseiros, e (2) que todos os dados individuais provêm de uma mesma distribuição de freqüências.

No caso de dados hidrológicos ou hidrometeorológicos, os erros grosseiros devem-se principalmente à leitura, transcrição ou processamento incorretos, e são muito freqüentes nas leituras linimétricas e pluviométricas, nas quais a intervenção humana é mais presente e, em conseqüência, a probabilidade de erro é maior. Em alguns casos, a identificação e eliminação dos erros grosseiros presentes nas séries hidrológicas/hidrometeorológicas não são tarefas de fácil execução.

Quando são alteradas as circunstâncias (localização, regime, equipamento de medição) sob as quais os dados são coletados, as séries hidrológicas/hidrometeorológicas podem vir a apresentar tendências e não-estacionariedade. Nesses casos, a distribuição de freqüência dos dados coletados passa a não ser constante no tempo e a série hidrológica/hidrometorológica, como uma amostra única, não pode ser considerada homogênea e nem utilizada para a inferência estatística. São exemplos pertinentes: (a) a relocação de um posto pluviométrico para local com características de vento muito diferentes daquelas apresentadas na instalação de origem, (b) a alteração do regime hidrológico causada pela implantação de reservatório de acumulação a montante de um posto fluviométrico, e (c) a utilização de equipamentos não aferidos, defeituosos ou incompatíveis com a sistemática padrão de coleta de dados primários.

As técnicas mais usuais para a identificação de erros e heterogeneidades nas séries hidrológicas/hidrometeorológicas são:

- comparação de cotagramas e/ou fluviogramas de postos fluviométricos próximos;
- comparação dos totais mensais de precipitação entre postos pluviométricos próximos ou entre um posto e a média de postos vizinhos;

- curvas de dupla acumulação de séries mensais/anuais do posto em questão e do "padrão regional", esse tomado como a média de vários postos das proximidades; e
- testes estatísticos convencionais para verificação de independência, homogeneidade e pontos atípicos (Spearman, Mann-Whitney, Grubbs-Beck, entre outros).

Além dessas técnicas de uso corrente em hidrologia, Hosking e Wallis (1997) sugerem também a comparação entre os quocientes de momentos-L amostrais, calculados para os diferentes postos de observação. Segundo esses autores, os quocientes de momentos-L amostrais são capazes de refletir erros, pontos atípicos e heterogeneidades eventualmente presentes em uma série de observações. Isso pode ser efetuado através de uma estatística-sumário, a qual representa uma medida da discordância entre os quocientes de momentos-L amostrais de um dado local e a média dos quocientes de momentos-L dos vários postos da região.

D.3.1 A medida de discordância

Em um grupo de amostras, a medida de discordância tem por objetivo identificar aquelas que apresentam características estatísticas muito discrepantes das grupais. Essa medida é expressa como uma estatística única, envolvendo as estimativas dos quocientes de momentos-L CV-L (τ), Assimetria-L (τ_3) e Curtose-L (τ_4). Em um espaço tridimensional de variação desses quocientes de momentos-L, a idéia é assinalar como discordantes as amostras cujos valores $(\hat{\tau}, \hat{\tau}_3, \hat{\tau}_4)$, representados por um ponto no espaço, se afastam "demasiadamente" do núcleo de concentração das amostras do grupo.

Para melhor visualização do significado dessa estatística, considere o plano definido pelos limites de variação das estimativas do CV-L e da Assimetria-L para os diversos postos de observação de uma região geográfica (Figura D.1). Nessa figura, as médias grupais encontram-se no ponto assinalado pelo símbolo +, em torno do qual constroem-se elipses concêntricas cujos eixos maiores e menores são funções da matriz de covariância amostral dos quocientes de momentos-L. Os pontos considerados discordantes são aqueles que encontram-se fora da área definida pela elipse mais externa.

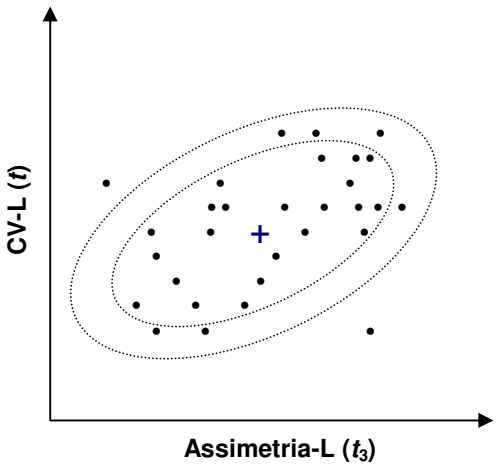


FIGURA D.1: Descrição esquemática da medida de discordância Di.

Os quocientes de momentos-L de um local *i*, a saber CV-L, Assimetria-L e Curtose-L, são considerados como um ponto em um espaço tridimensional. Em termos formais, considere que u_i representa um vetor (3x1) contendo esses quocientes de momentos-L, dado por:

$$u_i = (t^i \; t_3^{\;i} \; t_4^{\;i})^T \tag{D.4}$$

onde t, t_3 e t_4 denotam respectivamente CV-L, Assimetria-L e Curtose-L, e o símbolo *T* indica matriz transposta. Seja $\overline{u}$ um vetor (3x1) contendo a média grupal ou regional dos quocientes de momentos-L, tomada como a média aritmética simples de u_i para todos os postos estudados, ou seja:

$$\overline{u} = \frac{\sum_{i=1}^{N} u_i}{N} = (t^R \; t_3^{\;R} \; t_4^{\;R})^T \tag{D.5}$$

onde *N* representa o número de postos de observação do grupo ou região *R* em questão. Dada a matriz de covariância amostral *S*, definida por:

$$S = (N-1)^{-1} \sum_{i=1}^{N} (u_i - \overline{u})(u_i - \overline{u})^T \tag{D.6}$$

Hosking e Wallis (1995) definem a medida de discordância Di, para o local *i*, pela expressão:

$$Di = \frac{N}{3(N-1)}(u_i - \overline{u})^T S^{-1}(u_i - \overline{u}) \qquad \text{(D.7)}$$

Em trabalhos anteriores, Hosking e Wallis (1993) sugeriram o valor limite $Di = 3$ como critério para decidir se a amostra é discordante das características grupais. Por exemplo, quando uma certa amostra produz $Di > 3$, isso significa que ela pode conter erros grosseiros e/ou sistemáticos, ou mesmo pontos atípicos, que a tornam discordantes ou discrepantes das demais amostras do grupo. Posteriormente, Hosking e Wallis (1995) apresentaram novos valores críticos para Di, para grupos ou regiões com menos de 15 postos de observação. Esses valores críticos são listados na Tabela D.1.

TABELA D.1: Valores críticos da medida de discordância Di.

Número de postos da região	$Di_{crít.}$
5	1,333
6	1,648
7	1,917
8	2,140
9	2,329
10	2,491
11	2,632
12	2,757
13	2,869
14	2,971
≥ 15	3

Fonte: HOSKING e WALLIS, 1995.

De acordo com Hosking e Wallis (1995), para grupos ou regiões com número muito reduzido de postos de observação, a estatística Di não é informativa. Por exemplo, para $N < 3$, a matriz de covariância S é singular, e o valor de Di não pode ser calculado. Para $N = 4$, $Di = 1$, e para $N = 5$ ou $N = 6$, os valores de Di, como indicado na Tabela D.1, são bastante próximos do limite algébrico dessa estatística, definido por $Di \leq (N-1)/3$. Em conseqüência, os autores sugerem o uso da medida de discordância Di somente para $N > 7$.

Hosking e Wallis (1997) fazem as seguintes recomendações para o uso da medida de discordância Di:

a) A análise regional de consistência de dados inicia-se com o cálculo das estatísticas Di individuais de todos os postos de uma grande região geográfica, sem considerações

preliminares relativas à homogeneidade regional. Aqueles postos assinalados como discordantes devem ser submetidos a cuidadosa análise individual (testes estatísticos, curva de dupla acumulação, comparação com postos vizinhos), visando à identificação/eliminação de eventuais inconsistências em seus dados.

b) Em seguida, quando a homogeneidade regional (veja item D.4) já houver sido definida, as medidas de discordância devem ser recalculadas, desta feita com os postos devidamente agrupados em suas respectivas regiões homogêneas. Se um certo posto se apresentar discordante em uma região, deve ser considerada a possibilidade de sua transferência para outra.

c) Ao longo de toda a análise regional de consistência de dados, deve-se levar em conta que os quocientes de momentos-L amostrais podem apresentar diferenças naturalmente possíveis, mesmo entre postos similares do ponto de vista dos processos físicos em questão. Hosking e Wallis (1997) exemplificam que um evento extremo, porém localizado, pode ter afetado somente alguns postos em uma região. Entretanto, se é provável que um evento como este possa afetar qualquer posto da região, então a providência mais sensata seria a de tratar todo o grupo de postos como uma única região homogênea, mesmo que alguns possam apresentar medidas de discordância superiores aos valores limites estabelecidos.

D.4 Identificação e delimitação de regiões homogêneas

Das etapas que compõem a análise regional de freqüência de variáveis aleatórias, a identificação e delimitação de regiões homogêneas é considerada a mais difícil e mais sujeita a subjetividades. Uma região é homogênea se existem evidências suficientes de que as diferentes amostras do grupo possuem a mesma distribuição de freqüência, a menos, é claro, do fator de escala local. Potter (1987) considera que essa etapa é crucial por exigir do analista e da metodologia empregada a capacidade para discernir se observações anômalas, eventualmente existentes em uma ou mais amostras do grupo, se devem a diferenças populacionais em relação ao modelo probabilístico proposto ou a meras flutuações amostrais. Embora diversas técnicas tenham sido propostas para a identificação e delimitação de regiões homogêneas, nenhuma delas constitui um critério estritamente objetivo ou uma solução consensual para o problema. De fato, Bobée e Rasmussen (1995) reconhecem que a análise regional de freqüência e, em particular, a delimitação de regiões homogêneas, são construídas

com base em premissas difíceis de serem tratadas com rigor matemático. Esses autores concluem enfatizando que esse fato deve ser visto como um desafio a ser vencido por futuras investigações pertinentes à área de análise de freqüência.

Uma primeira fonte de controvérsias quanto à correta abordagem para a identificação de regiões homogêneas diz respeito ao tipo de dado local a ser utilizado. Faz-se distinção entre estatísticas locais e características locais. As estatísticas locais referem-se, por exemplo, a estimadores das medidas de dispersão e assimetria, tais como CV-L e Assimetria-L calculados diretamente a partir das amostras de dados submetidos à análise regional de freqüência. Por outro lado, as características locais são, em princípio, quantidades previamente conhecidas e não dedutíveis ou estimadas a partir das amostras pontuais. Como exemplos de características locais para o caso de variáveis hidrológicas ou hidrometeorológicas, podem ser citadas a latitude, a longitude, a altitude e outras propriedades físicas relacionadas a um certo local específico. Podem ser incluídas também outras características indiretamente relacionadas à amostra, tais como a altura média de precipitação anual, o mês mais freqüente de ocorrência de cheias ou o volume médio anual do escoamento-base. Alguns autores, nominalmente Wiltshire (1986), Burn (1989) e Pearson (1991), propuseram técnicas que fazem uso somente das estatísticas locais para definir regiões homogêneas de vazões de enchentes na Inglaterra, Estados Unidos e Nova Zelândia, respectivamente. Ao contrário, Hosking e Wallis (1997) recomendam que a identificação de regiões homogêneas se faça em duas etapas consecutivas: a primeira, consistindo de uma delimitação preliminar baseada unicamente nas características locais, e a segunda, consistindo de um teste estatístico, construído com base somente nas estatísticas locais, cujo objetivo é o de verificação dos resultados preliminarmente obtidos.

De fato, dentro da construção proposta por Hosking e Wallis (1997), tratar-se-ia de um raciocínio circular usar os mesmos dados tanto para identificar as regiões como para testar a sua correção. Além, evidentemente, de agregar novas informações independentes, o processo de identificação de regiões homogêneas em duas etapas, tal como recomendado por Hosking e Wallis (1997), é reforçado por outros argumentos. Considere, por exemplo, o caso em que uma estatística, como o CV-L local, é empregada como critério único para agrupar as amostras e identificar regiões homogêneas. Nesse contexto, existirá sempre uma tendência de agrupar aquelas amostras com valores atípicos (*outliers*) altos (conseqüentemente, com elevadas estimativas locais de CV-L), muito embora esses *outliers* possam dever-se a meras flutuações de uma amostra, as quais podem não estar presentes em locais vizinhos.

A identificação de regiões homogêneas em duas etapas, proposta por Hosking e Wallis (1997), encontra-se sintetizada a seguir. Inicialmente, são apresentados alguns dos métodos existentes para a identificação preliminar de regiões homogêneas, seguidos de uma descrição mais detalhada da técnica de *clusters*. Na seqüência, apresenta-se um teste estatístico, materializado pela medida de heterogeneidade, e construído com base nos quocientes de momentos-L amostrais.

D.4.1 Identificação preliminar de regiões homogêneas: métodos existentes

De acordo com Hosking e Wallis (1997), os diversos métodos e técnicas de agrupamento de locais similares em regiões homogêneas podem ser categorizados como se segue:

- *Conveniência geográfica*

  Dentro dessa categoria, encontram-se todas as experiências de identificação de regiões homogêneas que se baseiam no agrupamento subjetivo e/ou conveniente dos postos de observação, geralmente contíguos, em áreas administrativas ou em zonas previamente definidas segundo limites arbitrários. Dentre os inúmeros trabalhos que fizeram uso da conveniência geográfica, podem ser citadas as regionalizações de vazões de enchentes das Ilhas Britânicas (NERC, 1975) e da Austrália (Institution of Engineers Australia, 1987).

- *Agrupamento subjetivo*

  Nessa categoria, a delimitação subjetiva das regiões homogêneas é feita por agrupamento dos postos de observação em conformidade com a similaridade de algumas características locais, tais como classificação climática, relevo ou conformação das isoietas anuais. Schaefer (1990), por exemplo, utilizou alturas similares de precipitação anual para delimitar regiões homogêneas de chuvas máximas anuais no estado americano de Washington. Da mesma forma, Pinto e Naghettini (1999) utilizaram de modo combinado as conformações de relevo, clima e isoietas anuais para a delimitação preliminar de regiões homogêneas para alturas diárias de chuvas máximas anuais na bacia do Alto Rio São Francisco. Embora um grau considerável de subjetividade esteja presente nessas experiências, os seus resultados podem ser objetivamente verificados através do teste estatístico da medida de heterogeneidade, a ser descrito no item D.4.3.

- *Agrupamento objetivo*

  Nesse caso, as regiões são formadas pelo agrupamento dos postos de observação em um ou mais conjuntos, de modo que uma dada estatística não exceda um valor limiar previamente selecionado. Esse valor limiar é arbitrado de forma a minimizar critérios variados de heterogeneidade. Por exemplo, Wiltshire (1985) utilizou como critério a razão de verossimilhança e, posteriormente, Wiltshire (1986) e Pearson (1991) empregaram as variabilidades intra-grupos de estatísticas locais como os coeficientes de variação e assimetria. Na seqüência, os grupos são subdivididos iterativamente até que se satisfaça o critério de homogeneidade proposto. Hosking e Wallis (1997) apontam como uma desvantagem dessa técnica o fato de que as iterações sucessivas de reagrupamento dos postos de observação nem sempre conduzem a uma solução final otimizada. Apontam também para o fato de que as estatísticas intra-grupos utilizadas podem ser influenciadas, em grau indeterminado, pela eventual existência de dependência estatística entre as amostras consideradas.

- *Análise de clusters*

  Trata-se de um método usual de análise estatística multivariada, no qual associa-se a cada posto um vetor de dados contendo as características e/ou estatísticas locais. Em seguida, os postos são agrupados e reagrupados de forma que seja possível identificar a maior ou menor similaridade entre os seus vetores de dados. Hosking e Wallis (1997) citam diversos estudos (BURN, 1989; GUTTMAN, 1993; entre outros), nos quais a análise de *clusters* foi empregada com sucesso para a regionalização de freqüências de precipitação, vazões de enchentes e outras variáveis. Esses autores consideram a análise de *clusters* como o método mais prático, porém ainda sujeito a subjetividades, para a identificação preliminar de regiões homogêneas. Por constituir um método preferencial, apresenta-se no item D.4.2 uma descrição da técnica de análise de *clusters* e recomendações para o seu emprego na identificação preliminar de regiões homogêneas.

D.4.2 Análise de *clusters*: noções

Essencialmente, a análise de *clusters* é a aglomeração seqüencial de indivíduos em grupos cada vez maiores, de acordo com algum critério, distância ou medida de dissimilaridade. Um indivíduo pode ter diversos atributos ou características, as quais são quantificadas e representadas pelo vetor de dados locais $(Z_1, Z_2, ..., Z_p)$. As medidas ou distâncias de

dissimilaridade entre dois indivíduos devem ser representativas da variação mútua das características locais em um espaço *p*-dimensional. A medida mais usada é a *distância Euclidiana generalizada*, que é simplesmente a distância geométrica tomada em um espaço de *p* dimensões. Por exemplo, a distância Euclidiana entre dois indivíduos *i* e *j* é dada por:

$$d_{i,j} = \sqrt{\sum_{k=1}^{p} (Z_{i,k} - Z_{j,k})^2} \qquad \text{(D.8)}$$

Para efeito de entendimento da lógica inerente à análise de *clusters*, tomemos um de seus métodos de aglomeração mais simples, que é conhecido como o do "vizinho mais próximo". A aglomeração em *clusters* inicia-se pelo cálculo das distâncias *d* entre um certo indivíduo e todos os outros do grupo, para cada um deles. Inicialmente, existem tantos grupos quanto numerosos forem os indivíduos. O primeiro *cluster* se forma com o par de indivíduos mais próximos (ou de menor distância Euclidiana); se a distância para outros indivíduos for a mesma da anterior, estes também farão parte do *cluster*. Em seguida, forma-se o *cluster* seguinte com o par (ou grupo, ou *cluster*) de menor distância Euclidiana, e assim sucessivamente, até que, ao final, todos os indivíduos estejam aglomerados. Considere o exemplo hipotético da Figura D.2, no qual 10 indivíduos, assinalados em abscissas, tiveram calculadas suas distâncias Euclidianas, em ordenadas, de acordo com um certo número de atributos.

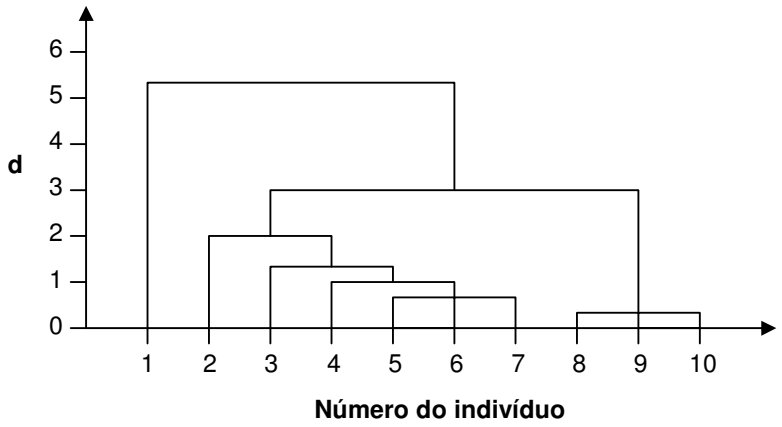


FIGURA D.2: Dendograma hipotético – 10 indivíduos.
Fonte: adaptado de Kottegoda e Rosso, 1997.

Se forem considerados somente dois *clusters*, o primeiro seria formado pelo indivíduo 1 e o segundo pelos nove indivíduos restantes. Na seqüência, o segundo *cluster* poderia ser dividido

em dois: um formado pelos indivíduos 8, 9 e 10, enquanto o outro seria formado pelos indivíduos restantes; dessa forma, teríamos um total de três *clusters*. Se agora seis *clusters* são necessários, então os indivíduos 1 a 4 formariam quatro *clusters* e os seis indivíduos restantes se agrupariam tal como apresentado no dendograma da Figura D.2. Dessa maneira, é possível ler em ordenadas a distância em que os indivíduos se aglomeram para formar um *cluster* e pode-se, através das distintas ramificações do dendograma, interpretar a estrutura de similaridade dos dados.

Inicialmente, quando cada indivíduo constitui o seu próprio *cluster*, as distâncias entre indivíduos são definidas por *d*, tal como calculado pela equação (D.8). Entretanto, a partir do momento em que vários indivíduos formam um ou mais *clusters*, põe-se a questão de como serão determinadas as distâncias de dissimilaridade entre esses novos *clusters*. Em outras palavras, faz-se necessária uma regra de aglomeração para definir quando dois *clusters* são suficientemente similares para se juntarem. Uma das várias possibilidades para se definir essa regra foi exemplificada na Figura D.2; nesse caso, usou-se o critério do "vizinho mais próximo", segundo o qual, a distância entre dois *clusters* é determinada pela distância entre os seus dois respectivos indivíduos que mais se aproximam. Esse critério pode conduzir à formação de extensos *clusters*, que se aglomeram meramente porque contêm indivíduos próximos.

Um método alternativo e muito utilizado como regra de aglomeração é o descrito por Ward (1963). Em linhas gerais, o método de Ward emprega a análise de variância para determinar as distâncias entre *clusters* e aglomerá-los de forma a minimizar a soma dos quadrados de quaisquer pares de *clusters* hipotéticos, a cada iteração. O método de Ward é considerado eficiente e, em geral, tende a produzir *clusters* pouco extensos e de igual número de indivíduos.

Outro método muito empregado é o devido a Hartigan (1975) e conhecido como o das *k*-médias (*k-means clustering*). O princípio desse método é o de que o analista pode, *a priori*, ter indícios ou hipóteses relativas ao número "correto" de *clusters* a ser considerado. Dessa forma, o método das *k*-médias irá produzir *k clusters*, os quais deverão ser os mais distintos entre si. Para fazê-lo, o método começa com a formação de *k clusters* iniciais, cujos membros são escolhidos aleatoriamente entre os indivíduos a serem agrupados. Em seguida, os indivíduos são movidos iterativamente de um *cluster* para outro de forma a (1) minimizar a

variabilidade intra-*cluster* e (2) maximizar a variabilidade entre os *clusters*. Essa lógica é análoga a uma análise de variância ao revés, no sentido de que, ao testar a hipótese nula de que as médias grupais são diferentes entre si, a análise de variância confronta a variabilidade entre grupos com a variabilidade intra-grupos. Em geral, os resultados do método das *k*-médias devem ser examinados de forma a se avaliar quão distintas são as médias dos *k* *clusters* obtidos.

Quando aplicada à identificação preliminar de regiões homogêneas para estudos regionais de freqüência de variáveis hidrológicas/hidrometeorológicas, a análise de *clusters* requer algumas considerações específicas. Hosking e Wallis (1997) recomendam atenção para os seguintes pontos:

1) Muitos algoritmos para a aglomeração em *clusters* utilizam o recíproco da distância Euclidiana como medida de similaridade. Nesse caso, é usual padronizar os elementos do vetor das características, dividindo-os pela sua amplitude ou desvio-padrão, de forma que passem a ter aproximadamente a mesma variabilidade. Essa padronização implica em atribuir ponderações iguais às diferentes características locais, o que pode ocultar a maior ou menor influência relativa de uma delas na forma da curva regional de freqüência. Pode-se compensar essa deficiência pela atribuição direta de diferentes ponderações às características locais consideradas.

2) Os métodos como o das *k*-médias requerem a definição do número de *clusters* a se considerar, muito embora, objetivamente, não se conheça *a priori* o número "correto" de *clusters*. Na prática, deve-se buscar um equilíbrio entre regiões demasiadamente grandes ou demasiadamente pequenas, com muitos ou poucos postos de observação. Para as metodologias de análise regional de freqüência que utilizam o princípio do *index-flood*, existe muito pouca vantagem em se empregar regiões muito extensas. Segundo Hosking e Wallis (1997), ganha-se muito pouca precisão nas estimativas de quantis ao se usar mais de 20 postos em uma região. Portanto, não há razão premente para se agrupar regiões extensas cujas estimativas das distribuições de freqüência são similares.

3) Os resultados da análise de *clusters* devem ser considerados como preliminares. Em geral, são necessários ajustes, muitas vezes subjetivos, cuja finalidade é tornar fisicamente coerente a delimitação das regiões, assim como reduzir a medida de heterogeneidade, a ser

descrita no item D.4.3. Os ajustes mencionados podem ser obtidos pelas seguintes providências:

- mover um ou mais postos de uma região para outra;
- desconsiderar ou remover um ou mais postos;
- subdividir uma região;
- abandonar uma região e realocar os seus postos para outras regiões;
- combinar uma região com outra ou outras;
- combinar duas ou mais regiões e redefini-las;
- obter mais dados e redefinir as regiões.

D.4.3 A medida de heterogeneidade regional

Em uma região homogênea, todos os indivíduos possuem os mesmos quocientes de momentos-L populacionais. Entretanto, as suas estimativas, ou seja, os quocientes de momentos-L calculados a partir das amostras, serão diferentes devido às flutuações amostrais. Portanto, é natural questionar se a dispersão dos quocientes de momentos-L amostrais, calculados para um certo conjunto de postos, é maior do que aquela que se esperaria encontrar em uma região homogênea. Essencialmente, é essa a lógica empregada para a construção da medida de heterogeneidade regional.

Pode-se visualizar o significado da medida de heterogeneidade através de diagramas de quocientes de momentos-L, como o da Figura D.3. Embora também se possa usar outras estatísticas, no exemplo hipotético desta figura encontram-se grafados, de um lado, o CV-L e a Assimetria-L amostrais, e do outro os coeficientes correspondentes, tal como obtidos a partir de simulações de amostras de mesmo tamanho das originais, localizadas, por hipótese, em uma região homogênea. Em diagramas como esses, uma região possivelmente heterogênea apresentaria, por exemplo, CV-L's amostrais muito mais dispersos do que aqueles obtidos por simulação. Em termos quantitativos, essa idéia básica pode ser traduzida pelo cálculo da diferença relativa centrada entre as dispersões observada e simulada, ou seja, pela razão $\frac{(\text{dispersão observada}) - (\text{média das dispersões simuladas})}{(\text{desvio - padrão das dispersões simuladas})}$.

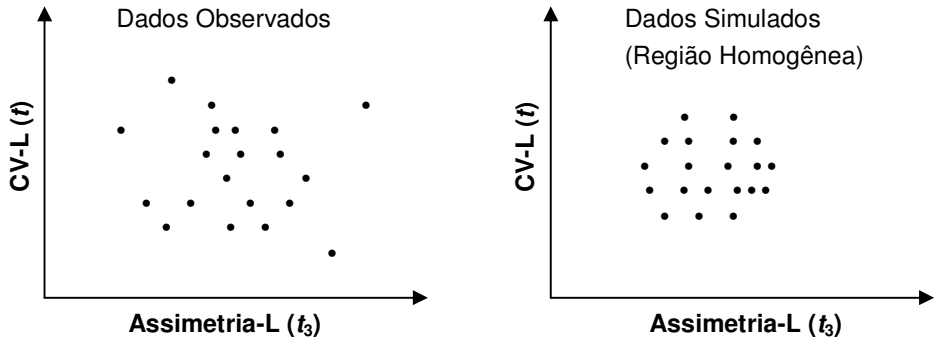


FIGURA D.3: Descrição esquemática do significado de heterogeneidade regional.

Para tornar possível o cálculo das estatísticas simuladas para a região homogênea, é necessário especificar uma função de distribuição de probabilidades para a população de onde serão extraídas as amostras. Hosking e Wallis (1997) recomendam o emprego da distribuição Kapa de 4 parâmetros, e justificam que essa recomendação se deve à preocupação de não assumir *a priori* nenhum comprometimento com distribuições de 2 ou 3 parâmetros. A distribuição Kapa (definida pelos parâmetros ξ, α, k e h) inclui, como casos particulares, as distribuições Logística, Generalizada de Valores Extremos e Generalizada de Pareto, sendo, portanto, teoricamente capaz de modelar o comportamento de variáveis hidrológicas e hidrometeorológicas. Uma definição formal da distribuição Kapa pode ser encontrada em Hosking e Wallis (1997).

Considere que uma dada região contenha *N* postos de observação, cada um deles indexado por *i*, com amostra de tamanho n_i e quocientes de momentos-L amostrais representados por t^i, t_3^i e t_4^i. Considere também que t^R, t_3^R e t_4^R denotam, respectivamente, as médias regionais dos quocientes CV-L, Assimetria-L e Curtose-L, ponderados, de forma análoga à especificada pela equação (D.2), pelos tamanhos das amostras individuais. Hosking e Wallis (1997) recomendam que a medida de heterogeneidade, denotada por *H*, se baseie preferencialmente no cálculo da dispersão de *t*, ou seja, do CV-L, para as regiões proposta e simulada. Inicialmente, efetua-se o cálculo do desvio-padrão ponderado *V* dos CV-L's das amostras observadas, através da seguinte expressão:

$$V = \sqrt{\frac{\sum_{i=1}^{N} n_i \cdot (t^i - t^R)^2}{\sum_{i=1}^{N} n_i}} \qquad \text{(D.9)}$$

Os parâmetros da população Kapa são então estimados de forma a reproduzir os quocientes de momentos-L regionais $(1, t^R, t_3^R, t_4^R)$. Com os parâmetros populacionais, são simuladas N_{SIM} regiões homogêneas, sem correlações cruzada e serial, contendo *N* amostras individuais, cada uma com n_i valores da variável normalizada. Na seqüência, as estatísticas V_j, com $j = 1,2,...,N_{SIM}$, são calculadas para todas as regiões homogêneas através da equação (D.9). A sugestão é que se faça o número de simulações N_{SIM} igual a 500.

A média aritmética das estatísticas V_j, obtidas por simulação, fornecerá a dispersão média esperada para a região homogênea:

$$\mu_V = \frac{\sum_{j=1}^{N_{SIM}} V_j}{N_{SIM}} \tag{D.10}$$

A medida de heterogeneidade *H* compara a dispersão observada com a simulada:

$$H = \frac{V - \mu_V}{\sigma_V} \tag{D.11}$$

onde σ_V é o desvio-padrão dos N_{SIM} valores da medida de dispersão V_j, ou seja:

$$\sigma_V = \sqrt{\frac{\sum_{j=1}^{N_{SIM}} (V_j - \mu_V)^2}{N_{SIM} - 1}} \tag{D.12}$$

De acordo com o teste de significância, proposto por Hosking e Wallis (1997), se $H < 1$, considera-se a região como "aceitavelmente homogênea", se $1 \le H < 2$, a região é vista como "possivelmente heterogênea" e, finalmente, se $H \ge 2$, a região deve ser classificada como "definitivamente heterogênea". Conforme mencionado anteriormente, alguns ajustes subjetivos, como a remoção ou o reagrupamento de postos de uma ou mais regiões, podem se tornar necessários para fazer com que a medida de heterogeneidade se adeqüe aos limites propostos. Entretanto, é possível que, em alguns casos, a heterogeneidade aparente se deva à presença de um pequeno número de postos "atípicos" na região. Uma alternativa é reagrupá-los em outra região, na qual sejam "mais típicos", muito embora não exista nenhuma razão

física evidente para que esse pequeno grupo de postos tenha comportamento distinto do restante dos postos da região de origem. Hosking e Wallis (1997) argumentam que, nesses casos, as razões de natureza física devem ter precedência sobre as de natureza estatística, e recomendam a alternativa de manter o grupo de postos "atípicos" na região originalmente proposta. Continuam a argumentação tomando, como exemplo, a situação em que uma certa combinação de eventos meteorológicos extremos possa ocorrer em qualquer ponto de uma região, mas que, de fato, tenha sido registrada somente em alguns de seus postos, durante o período disponível de observações. Os verdadeiros benefícios potenciais da regionalização poderiam ser atingidos em situações como a exemplificada, na qual o conhecimento dos mecanismos físicos associados à ocorrência de eventos extremos permite agrupar todos os postos em uma única região homogênea. Para esse exemplo, os dados locais encontram-se indevidamente influenciados pela presença ou ausência de eventos raros, e a curva regional de freqüência, construída como a média das curvas individuais, constitui certamente o melhor instrumento para se estimar os riscos de futuras ocorrências dessa natureza.

A medida de heterogeneidade é construída como um teste de significância da hipótese nula de que a região é homogênea. Entretanto, Hosking e Wallis (1997) argumentam que não se deve interpretá-lo rigorosamente como tal, porque um teste de homogeneidade exata só seria válido sob as premissas de que os dados não possuem correlações cruzada e/ou serial e que a função Kapa representa a verdadeira distribuição regional. Mesmo se fosse possível construir um rigoroso teste de significância, ele teria utilidade duvidosa, pois, na prática, mesmo uma região moderadamente heterogênea pode produzir melhores estimativas de quantis do que aquelas produzidas pela análise exclusiva de dados locais.

Os critérios $H = 1$ e $H = 2$, embora arbitrários, representam indicadores úteis. Se a medida de heterogeneidade fosse interpretada como um teste de significância, e supondo que a estatística V possuísse uma distribuição Normal, o critério de rejeição da hipótese nula de homogeneidade, ao nível $\alpha = 10\%$, seria $H = 1{,}28$. Nesse contexto, o critério arbitrário $H = 1$ pode parecer muito rigoroso; entretanto, conforme argumentação anterior, não se quer interpretar a medida H como um teste de significância exato. A partir de resultados de simulação, Hosking e Wallis (1997) demonstraram que, em média, $H \approx 1$ para uma região suficientemente heterogênea, na qual as estimativas de quantis são 20 a 40% menos precisas que as obtidas para uma região homogênea. Assim sendo, o limite $H = 1$ é visto como o ponto a partir do qual a redefinição da região pode apresentar vantagens. Analogamente, o

limite $H = 2$ é visto como o ponto a partir do qual a redefinição da região é definitivamente vantajosa.

Em alguns casos, *H* pode apresentar valores negativos. Eles indicam que há menos dispersão entre os valores amostrais de CV-L do que se esperaria de uma região homogênea, com distribuições individuais de freqüência independentes. A causa mais provável para esses valores negativos é a presença de correlação positiva entre os dados dos diferentes postos. Se valores muito negativos, como $H < -2$, são observados durante a regionalização, isso pode ser uma indicação de que há muita correlação cruzada entre as distribuições individuais de freqüência ou de que há uma regularidade excessiva dos valores amostrais de CV-L. Para esses casos, Hosking e Wallis (1997) recomendam reexaminar os dados de forma mais cuidadosa.

D.5 Seleção da distribuição regional de freqüência

Existem diversas famílias de distribuições de probabilidades que podem ser consideradas candidatas a modelar um conjunto de dados regionais. As argumentações apresentadas no Capítulo 3, a respeito da prescrição do modelo probabilístico, também são válidas para a seleção da distribuição regional de freqüência. Além dessas argumentações, Hosking e Wallis (1997) consideram uma escolha natural tomar como base para um teste de aderência das distribuições candidatas ao conjunto de dados amostrais, as médias regionais de estatísticas de momentos-L, como, por exemplo, a Assimetria-L e a Curtose-L, e então compará-las às características teóricas das diferentes distribuições candidatas. Essa é a idéia básica da medida de aderência *Z*, descrita a seguir.

D.5.1 A medida de aderência

Em uma região homogênea, os quocientes de momentos-L individuais flutuam em torno de suas médias regionais. Na maioria dos casos, as distribuições de probabilidades candidatas a modelar o comportamento da variável em estudo possuem parâmetros de posição e escala que reproduzem a média e o CV-L regionais. Portanto, a aderência de uma certa distribuição aos dados regionais deve se basear necessariamente em momentos-L de ordem superior. Hosking e Wallis (1997) consideram suficientes a Assimetria-L e a Curtose-L. Logo, pode-se julgar a aderência pelo grau com que uma certa distribuição aproxima as médias regionais de Assimetria-L e Curtose-L. Por exemplo, suponha que a distribuição candidata seja a

Generalizada de Valores Extremos (GEV), de três parâmetros. Quando ajustada aos dados da região pelo método dos momentos-L, essa distribuição irá reproduzir a média regional de Assimetria-L. Pode-se julgar o grau de ajuste, portanto, pela diferença entre a Curtose-L da distribuição τ_4^{GEV} e a média regional correspondente t_4^R, tal como esquematizado na Figura D.4. Entretanto, essa diferença deve levar em conta a variabilidade amostral de t_4^R. Essa variabilidade pode ser quantificada através de σ_4, ou seja, o desvio-padrão de t_4^R, o qual é obtido por simulação de um grande número de regiões homogêneas, todas extraídas de uma população modelada pela GEV, contendo os mesmos indivíduos e tamanhos de amostras dos dados observados. Nesse caso, portanto, a medida de aderência da distribuição GEV pode ser calculada como $Z^{GEV} = (t_4^R - \tau_4^{GEV})/\sigma_4$.

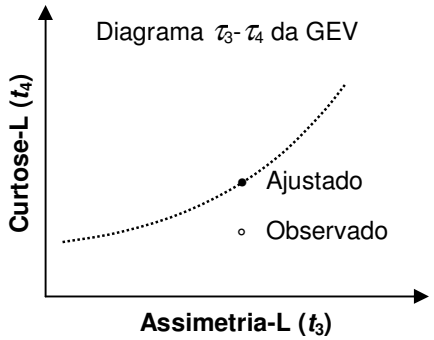


FIGURA D.4: Descrição esquemática da medida de aderência Z.

Hosking e Wallis (1997) reportam as seguintes dificuldades relacionadas ao procedimento de cálculo da medida de aderência tal como anteriormente descrito:

- Para obter os valores corretos de σ_4, é necessário um conjunto de simulações específico para cada distribuição candidata. Entretanto, na prática, Hosking e Wallis (1997) consideram que é suficiente supor que σ_4 tem o mesmo valor para todas as distribuições candidatas de três parâmetros. Justificam afirmando que, como todas as distribuições ajustadas têm a mesma Assimetria-L, é razoável supor que elas também se assemelham com relação a outras características. Assim sendo, também é razoável supor que uma distribuição Kapa de quatro parâmetros, ajustada aos dados regionais, terá um valor de σ_4 próximo ao das distribuições candidatas. Portanto, σ_4 pode ser obtido a partir da simulação de um grande número de regiões homogêneas extraídas de uma população

Kapa. Para esse objetivo, podem ser empregadas as mesmas simulações usadas no cálculo da medida de heterogeneidade, conforme descrito no item D.4.3.

- As estatísticas aqui mencionadas pressupõem a inexistência de qualquer viés no cálculo dos momentos-L amostrais. Hosking e Wallis (1997) observam que essa suposição é válida para t_3, mas não para t_4, sob as condições de amostras de pequeno tamanho $(n_i \leq 20)$ ou de populações de grande assimetria $(\tau_3 \geq 0{,}4)$. A solução desse problema é feita por uma correção de viés para t_4. Essa correção, denotada por B_4, pode ser calculada através dos mesmos resultados de simulação usados para se calcular σ_4.

- A medida de aderência *Z* se refere a distribuições candidatas de três parâmetros. Embora seja possível construir um procedimento semelhante para as distribuições de dois parâmetros, elas possuem valores populacionais fixos de τ_3 e τ_4, e, em conseqüência, tornam problemática a estimação de σ_4. Apesar de terem sugerido algumas adaptações plausíveis, Hosking e Wallis (1997) desaconselham o uso da medida de aderência para distribuições de apenas dois parâmetros.

Considere que uma dada região contenha *N* postos de observação, cada um deles indexado por *i*, com amostra de tamanho n_i e quocientes de momentos-L amostrais representados por t^i, t_3^i e t_4^i. Considere também que t^R, t_3^R e t_4^R denotam, respectivamente, as médias regionais dos quocientes CV-L, Assimetria-L e Curtose-L, ponderados, de forma análoga à especificada pela equação (D.2), pelos tamanhos das amostras individuais. Dado o conjunto de distribuições candidatas de três parâmetros proposto por Hosking e Wallis (1997), a saber: Logística Generalizada (GLO), Generalizada de Valores Extremos (GEV), Generalizada de Pareto (GPA), Log-Normal 3 parâmetros (LN3) e Pearson III (PE3), cada distribuição deverá ter seus parâmetros ajustados ao grupo de quocientes de momentos-L regionais $(1, t^R, t_3^R, t_4^R)$. Denota-se por τ_4^{DIST} a Curtose-L da distribuição ajustada, onde *DIST* poderá ser qualquer uma das distribuições candidatas.

Na seqüência, deve-se ajustar a distribuição Kapa ao grupo de quocientes de momentos-L regionais e proceder à simulação de um grande número N_{SIM} de regiões homogêneas, cada qual tendo a Kapa como distribuição de freqüência. Essa simulação deverá ser efetuada

exatamente como descrito para o cálculo da medida de heterogeneidade regional (veja item D.4.3). Em seguida, calculam-se as médias regionais t_3^m e t_4^m da Assimetria-L e Curtose-L da *m*-ésima região simulada. O viés de t_4^R é dado por:

$$B_4 = \frac{\sum_{m=1}^{N_{SIM}} (t_4^m - t_4^R)}{N_{SIM}} \tag{D.13}$$

O desvio-padrão de t_4^R é dado pela expressão:

$$\sigma_4 = \sqrt{\frac{\sum_{m=1}^{N_{SIM}} (t_4^m - t_4^R)^2 - N_{SIM} \cdot B_4^2}{N_{SIM} - 1}} \tag{D.14}$$

A medida de aderência *Z* de cada distribuição candidata pode ser calculada pela equação:

$$Z^{DIST} = \frac{\tau_4^{DIST} - t_4^R + B_4}{\sigma_4} \tag{D.15}$$

A hipótese de um ajuste adequado é tão mais verdadeira quanto mais próxima de zero for a medida de aderência. Hosking e Wallis (1997) sugerem como critério razoável o limite $|Z^{DIST}| \leq 1{,}64$.

A estatística *Z* é especificada sob a forma de um teste de significância e, segundo Hosking e Wallis (1997), possui uma distribuição que se aproxima da Normal padrão, sob as premissas de que a região é perfeitamente homogênea e de que não há correlação cruzada entre os seus indivíduos. Se a distribuição de *Z* é de fato Normal, o critério $|Z^{DIST}| \leq 1{,}64$ corresponde à aceitação da hipótese de que os dados provêm da distribuição candidata, com um nível de confiança de 90%. Entretanto, as premissas necessárias para se aproximar a distribuição de *Z* pela Normal padrão, dificilmente são completamente satisfeitas na prática. Assim sendo, o critério $|Z^{DIST}| \leq 1{,}64$ é simplesmente um indicador de boa aderência e não uma estatística de teste formal. Hosking e Wallis (1997) relatam que o critério $|Z^{DIST}| \leq 1{,}64$ é particularmente inconsistente se os dados apresentarem correlação serial e/ou correlação cruzada. Tanto uma quanto a outra tendem a fazer aumentar a variabilidade de t_4^R; como não há correlação para as

regiões simuladas de população Kapa, a estimativa de σ_4 resulta ser excessivamente pequena, e a estatística *Z* excessivamente grande, conduzindo a uma falsa indicação de falta de aderência.

Quando, ao se aplicar o teste da medida de aderência a uma região homogênea, resultar que várias distribuições são consideradas aceitas, Hosking e Wallis (1997) recomendam o exame das curvas de quantis adimensionais. Se elas fornecem resultados aproximadamente iguais, qualquer das distribuições aceitas pode ser selecionada. Entretanto, se os resultados diferem significativamente, a escolha deve tender para o modelo probabilístico que apresentar maior robustez. Nesses casos, ao invés de um modelo de três parâmetros, recomenda-se a seleção da distribuição Kapa de quatro parâmetros ou da Wakeby de cinco parâmetros, as quais são mais robustas à incorreta especificação da curva regional de freqüência. A mesma recomendação se aplica aos casos em que nenhuma das distribuições de três parâmetros atendeu ao critério $|Z^{DIST}| \leq 1{,}64$, ou aos casos de regiões "possivelmente heterogêneas" ou "definitivamente heterogêneas".

Além da verificação da medida de aderência *Z*, recomenda-se grafar as médias regionais da Assimetria-L e Curtose-L (t_3^R, t_4^R) em um diagrama de quocientes de momentos-L. Hosking e Wallis (1993) sugerem que, se o ponto (t_3^R, t_4^R) se localizar acima da curva da distribuição Logística Generalizada, nenhuma distribuição de dois ou três parâmetros se ajustará aos dados, devendo, possivelmente, adotar-se uma distribuição Kapa de quatro parâmetros ou Wakeby de cinco parâmetros. Finalmente, ao se analisar uma grande área geográfica, sujeita a divisão em várias regiões homogêneas, a especificação da distribuição de freqüência de uma região pode afetar a das outras. Se uma determinada distribuição se ajusta bem aos dados da maioria das regiões, é de bom senso utilizá-la para todas, muito embora ela possa não ser a distribuição que particularmente melhor se ajusta aos dados de uma ou de algumas das regiões.

D.6 Estimação dos parâmetros e quantis da distribuição regional de freqüência

Depois que os dados dos diferentes postos da área em estudo foram submetidos às etapas descritas nos itens D.3, D.4 e D.5, tem-se como resultado a divisão da área em regiões aproximadamente homogêneas, cujos indivíduos possuem distribuições de freqüência

idênticas, a menos de um fator de escala local, representadas por uma única distribuição de probabilidades regional, selecionada entre diversas funções candidatas. Essa relação entre as distribuições de freqüência dos diversos locais representa a própria justificativa para a análise regional de freqüência, permitindo a obtenção de melhores estimativas de parâmetros e quantis a partir da combinação de dados espacialmente disseminados.

Diversos métodos podem ser utilizados para se ajustar uma distribuição de probabilidades aos dados de uma região homogênea. Para descrevê-los, considere inicialmente uma variável aleatória *X*, cuja variabilidade foi amostrada em *N* locais ou postos de observação, situados em uma região homogênea. As observações, tomadas nos postos indexados por *i*, formam amostras de tamanho variável n_i e são denotadas por $X_{i,j}$, $i = 1,2,...,N$, $j = 1,2,...,n_i$. Se *F* ($0 < F < 1$) representa a distribuição de freqüência da variável *X* no posto *i*, então, a função de quantis nesse local é simbolizada por $X_i(F)$. Em uma região homogênea, as distribuições de freqüência nos *N* pontos distintos são idênticas, a menos de um fator de escala local μ_i, o *index-flood*, ou seja:

$$X_i(F) = \mu_i \cdot x(F) \qquad i = 1,2,...,N \tag{D.16}$$

Se $\hat{\mu}_i$ denota a estimativa do fator de escala no local *i*, pode-se representar os dados adimensionais padronizados por $x_{i,j} = X_{i,j} / \hat{\mu}_i$, $i = 1,2,...,N$, $j = 1,2,...,n_i$.

O método mais simples e antigo para se combinar os dados locais, com o objetivo de se estimar os parâmetros e quantis da distribuição regional, é conhecido como o da *estação-ano*. Esse método simplesmente agrupa todos os dados adimensionais padronizados em uma única amostra, considerada aleatória simples, que é, em seguida, usada para se ajustar a distribuição regional. Hosking e Wallis (1997) consideram que, na atualidade, esse método é raramente empregado principalmente porque não é correto tratar os dados adimensionais padronizados como uma amostra aleatória simples, ou seja, uma realização de variáveis aleatórias independentes e igualmente distribuídas. De fato, como os fatores de escala locais $\hat{\mu}_i$ são, em geral, estimativas obtidas a partir de amostras de diferentes tamanhos, os dados adimensionais padronizados dos diversos postos considerados não serão igualmente distribuídos. Em outro extremo, encontra-se o método de estimação através do *máximo da função de verossimilhança*, tal como aplicado aos *N* fatores de escala locais μ_i e aos *p* parâmetros de

$x(F;\theta_1,\theta_2,...,\theta_p)$, contidos na equação (D.16). O modelo estatístico procura encontrar, em geral de forma iterativa, as $N+p$ soluções de um sistema de $N+p$ equações que visam maximizar a função de verossimilhança. Esse método pode ser usado também para situações em que os fatores de escala são considerados parâmetros dependentes de informações covariadas, ou seja, $\mu_i = h(z_i, \omega)$, onde z_i representa um vetor de características ou informações covariadas no local i, h uma função matemática convenientemente escolhida e ω um vetor dos parâmetros a serem estimados.

O método *index-flood* utiliza as estatísticas características dos dados locais para obter as estimativas regionais, ponderando-as através da equação:

$$\hat{\lambda}_k^R = \frac{\sum_{i=1}^{N} n_i \cdot \hat{\lambda}_k^{(i)}}{\sum_{i=1}^{N} n_i} \tag{D.17}$$

onde $\hat{\lambda}_k^R$ denota a estimativa regional e $\hat{\lambda}_k^{(i)}$, $k = 1,2,...,p$, representam as estatísticas locais. Se essas estatísticas locais se baseiam nos quocientes de momentos-L, Hosking e Wallis (1997) definem a metodologia de estimação como a do *algoritmo dos momentos-L regionais*. Apesar de reconhecerem não haver nenhuma superioridade teórica da metodologia proposta em relação à do máximo de verossimilhança, esses autores justificam o seu emprego pela maior simplicidade de cálculo. O algoritmo dos momentos-L regionais será descrito no item seguinte, tomando como premissa a inexistência de correlação cruzada entre as observações dos diferentes indivíduos de uma região homogênea, ou de correlação serial entre as observações de um dado local.

D.6.1 O algoritmo dos momentos-L regionais

O objetivo é ajustar uma única distribuição de freqüência aos dados adimensionais padronizados, observados em diferentes locais de uma região considerada aproximadamente homogênea. O ajuste é efetuado através do método dos momentos-L, o qual consiste em igualar os momentos-L populacionais da distribuição em questão aos respectivos momentos-L amostrais. Convenientemente, os quocientes de momentos-L locais são ponderados pelos respectivos tamanhos das amostra, de forma a produzir as estimativas regionais dos

quocientes de momentos-L, as quais são, em seguida, empregadas para a inferência estatística. Se o *index-flood* é representado pela média da distribuição local de freqüência, cuja estimativa é dada pela média amostral dos dados individuais, então a média dos dados adimensionais padronizados, bem como da ponderação regional, é 1. Isso faz com que os quocientes de momentos-L amostrais t e t_r, para $r \geq 3$, sejam os mesmos, não importando se foram calculados a partir dos dados originais $X_{i,j}$ ou dos dados adimensionais padronizados $x_{i,j}$.

Considere que uma dada região contenha *N* postos de observação, cada um deles indexado por *i*, com amostra de tamanho n_i e quocientes de momentos-L amostrais representados por t^i, t_3^i, $t_4^i, \ldots$. Considere também que t^R, t_3^R, $t_4^R, \ldots$ denotam as médias regionais dos quocientes de momentos-L, ponderados, de forma análoga à especificada pela equação (D.17), pelos tamanhos das amostras individuais. Conforme justificativa anterior, a média regional é 1, ou seja, $l_1^R = 1$.

Efetua-se o ajuste da distribuição regional igualando-se os seus quocientes de momentos-L populacionais λ_1, τ, τ_3, $\tau_4, \ldots$ às médias regionais 1, t^R, t_3^R, $t_4^R, \ldots$. Se *F*, ou seja, a distribuição a ser ajustada, é definida por *p* parâmetros θ_k, $k = 1,2,\ldots,p$, resultará um sistema de *p* equações e *p* incógnitas, cujas soluções serão as estimativas $\hat{\theta}_k$, $k = 1,2,\ldots,p$, as quais permitirão estimar a curva regional de quantis adimensionais $\hat{x}(F) = x(F;\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_p)$. Assim, as estimativas dos quantis para o posto *i* serão obtidas pelo produto de $\hat{x}(F)$ por $\hat{\mu}_i$, ou seja:

$$\hat{X}_i(F) = l_1^i \cdot \hat{x}(F) \qquad \text{(D.18)}$$

E SÉRIES DE VAZÕES MÁXIMAS ANUAIS DAS ESTAÇÕES FLUVIOMÉTRICAS UTILIZADAS

TABELA E.1: Séries de vazões máximas anuais (valores em m³/s).

ESTAÇÃO		44200000	44290002	44500000	45298000	45480000	46035000
N		68	30	64	62	57	31
ANO HIDROLÓGICO	31/32				5632.6		
	32/33				7210.0		
	33/34				6762.0		
	34/35	6476.0		6564.0	6987.4		
	35/36	4221.0					
	36/37	6980.0		6791.0			
	37/38	8392.0		7944.0	8405.0		
	38/39	10180.0		8874.0	9061.5		
	39/40	6252.0		6302.0	6724.1		
	40/41	6336.0		6110.0	6345.0	5018.0	
	41/42	6910.0		6703.0	6770.0		
	42/43	11585.0		9924.0	10144.5		
	43/44	12463.0		9560.0		7785.0	
	44/45	10080.0		8944.0	9524.0	8366.0	
	45/46	12507.0		10344.0	10878.0	9923.0	
	46/47	7736.0		7455.0	8102.1	6990.0	
	47/48	7416.0		6879.0	7640.5	6436.0	
	48/49	12913.0		10800.0	11555.8		
	49/50	5912.0		5762.0	6241.0	5161.0	
	50/51	6966.0		6146.0	6693.5	5568.0	
	51/52	9380.0		8468.0	9096.7	7830.0	
	52/53	4787.0		4657.0	4959.4	4644.0	
	53/54	6070.0		5599.0	5861.4	5330.0	
	54/55	6490.0		5657.0	6167.6	5239.0	
	55/56	7864.0		6879.0	7286.5	6380.0	
	56/57	8584.0		7798.0	8235.7	7200.0	
	57/58	5338.0		5309.0	5422.0	4836.0	
	58/59	5488.0		5113.0	5216.0	4680.0	
	59/60	6994.0		6715.0	7334.2	6480.0	
	60/61	7050.0		6703.0	7238.8	6310.0	
	61/62	5050.0		4976.0	5243.2	4823.0	
	62/63	10380.0		8874.0		8094.0	
	63/64	6042.0		6038.0	7049.8	6170.0	
	64/65	6588.0		5954.0	6770.0	6016.0	
	65/66	7720.0		7785.0	7787.6	6810.0	
	66/67	6210.0			6035.6	5291.0	
	67/68	7720.0		7191.0	8405.0	7290.0	
	68/69	4975.0		4703.0	4892.8	4488.0	
	69/70	7092.0		6426.0	7270.8	6465.0	6889.8
	70/71	3390.0		3320.0	3488.0		4009.1
	71/72	6434.0		5564.0	6021.0	5200.0	5633.3
	72/73	6364.0	6074.0	5762.0	6137.8	5442.0	5759.7
	73/74	5662.0	5474.0	5344.0	5703.6	5278.0	
	74/75	5350.0	5175.0	4919.0	5039.0	4560.0	

ESTAÇÃO		44200000	44290002	44500000	45298000	45480000	46035000
N		68	30	64	62	57	31
ANO HIDROLÓGICO	75/76						
	76/77	6896.0	6425.0	5930.0	6405.0	5764.0	6184.4
	77/78	7232.0	6762.0	6098.0	6556.2	5659.0	6171.4
	78/79	17380.0	[1] 15345.0	12240.0	13178.3	12400.0	12220.4
	79/80	10400.0	9279.0		9452.0	9178.0	9186.2
	80/81	6882.0	6371.0	6062.0	6495.4	6018.0	6434.0
	81/82	10720.0	9834.0	9042.0	9506.0	8566.0	8238.2
	82/83	11787.0	10521.0	9616.0	9388.0	9226.0	9050.6
	83/84	7496.0	7240.0	6980.0	7128.0	6760.0	6971.4
	84/85	9780.0	9038.0	8720.0	9096.7	8128.0	7949.9
	85/86	8456.0	8045.0	7917.0	8490.0	7570.0	7623.1
	86/87	4037.0	3830.0	3871.0	3860.3	3793.0	4008.1
	87/88	5550.0	5520.0	5495.0	5876.0	5668.0	5900.0
	88/89	4175.0	4028.0	4009.0		4230.0	4438.9
	89/90	11248.0	9945.0	9294.0		8976.0	8707.4
	90/91	6952.0	6533.0	6254.0	6525.8	6145.0	6197.5
	91/92	14130.0	12427.0	11168.0	12298.6		11571.5
	92/93	7372.0	7073.0	7020.0	7494.6	6747.0	7012.4
	93/94	7372.0	6884.0	6715.0		6294.0	6553.5
	94/95	3360.0	3253.0			3395.0	3618.9
	95/96	4750.0	4672.3	4543.0		4443.0	4718.7
	96/97	6014.0	5698.0	5506.4	5905.0	5307.0	5595.5
	97/98	5312.5	5016.3	4782.4	5025.7	4479.0	4730.5
	98/99	5125.0	4941.0	4714.0	4985.8	4587.0	4922.4
	99/00	5200.0	5040.8	4907.8	5259.0	4868.2	5174.6
	00/01	3692.0	3733.3	3706.7	3980.0	3997.7	4379.1
	01/02	5275.0	5094.5	4896.4	5020.1	4635.7	4825.0
	02/03	5250.0	5094.5	5113.0	5305.7	5029.7	5230.0
Máximo		17380.0	15345.0	12240.0	13178.3	12400.0	12220.4
Média		7385.2	6812.2	6710.3	7073.7	6174.9	6448.6
DP [2]		2729.0	2737.6	1941.1	1995.4	1726.2	2091.2
CV [3]		0.37	0.40	0.29	0.28	0.28	0.32
Assimetria		1.261	1.377	0.757	0.908	1.160	1.150

(1) Quantil resultante da regressão linear apresentada na Figura E.1
(2) Desvio padrão
(3) Coeficiente de variação

A Figura E.1 mostra o resultado da regressão linear realizada entre as amostras de vazões máximas anuais de São Francisco (44200000) e de Pedras de Maria da Cruz (44290002). Pode-se perceber que a descarga fluvial em Pedras de Maria da Cruz é altamente dependente daquela observada em São Francisco, situada cerca de 70 km a montante. Isso pode ser explicado pelos seguintes fatos, observados no trecho do rio São Francisco localizado entre as referidas estações fluviométricas:

- Ausência de afluentes importantes;
- Comprimento e área de drenagem incremental relativamente pequenos;
- Região com baixa vazão específica (expressa em m³/s·km²).

Sendo assim, a referida regressão linear pode ser considerada adequada para o preenchimento de falhas na série de vazões máximas anuais observada em Pedras de Maria da Cruz.

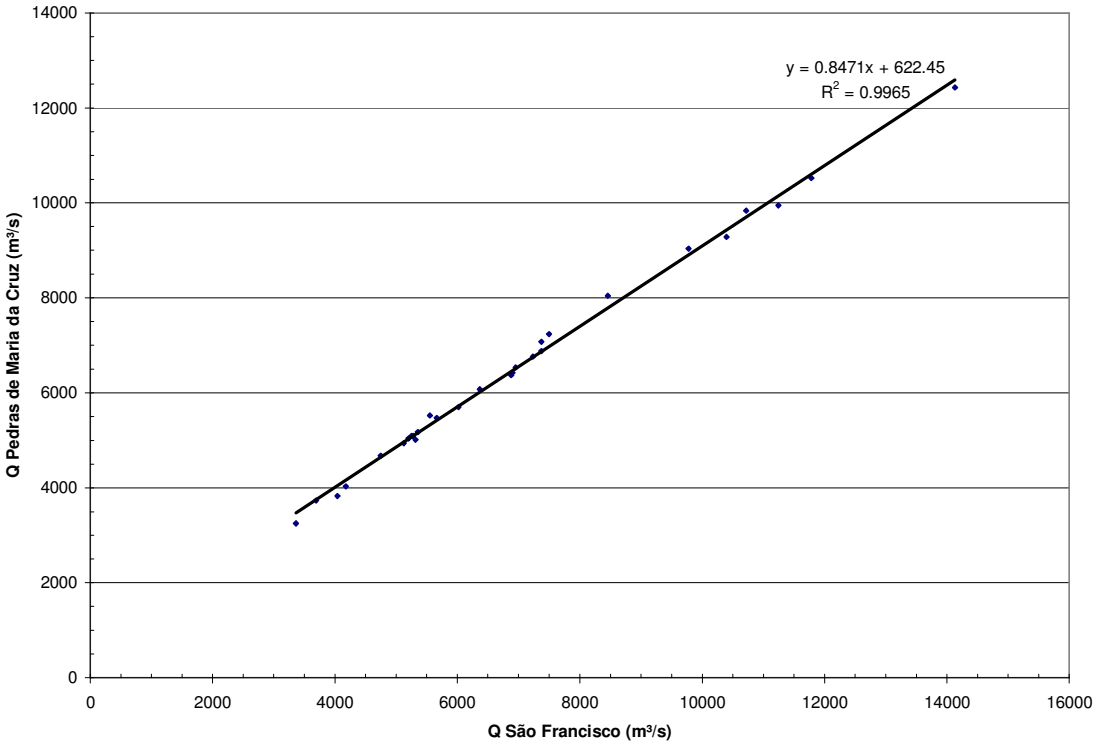


FIGURA E.1: Regressão linear entre as amostras de vazões máximas anuais de São Francisco (44200000) e de Pedras de Maria da Cruz (44290002).