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

# SEAF – UM PROTÓTIPO DE UM SISTEMA ESPECIALISTA PARA ANÁLISE DE FREQÜÊNCIA LOCAL DE EVENTOS HIDROLÓGICOS MÁXIMOS ANUAIS

Márcio de Oliveira Cândido

Belo Horizonte
Escola de Engenharia
Abril/2003

Márcio de Oliveira Cândido

# SEAF – UM PROTÓTIPO DE UM SISTEMA ESPECIALISTA PARA ANÁLISE DE FREQÜÊNCIA LOCAL DE EVENTOS HIDROLÓGICOS MÁXIMOS ANUAIS

Dissertação apresentada ao Programa de Pós-Graduação em Saneamento, Meio Ambiente e Recursos Hídricos dos Departamentos de Engenharia Sanitária e Ambiental e Engenharia Hidráulica 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: Recursos Hídricos

Orientador: Prof. Mauro da Cunha Naghettini

Belo Horizonte

Escola de Engenharia da UFMG

Abril/2003

CÂNDIDO, Márcio de Oliveira

C217s SEAF – um protótipo de um sistema especialista para análise de freqüência local de eventos hidrológicos máximos anuais / Márcio de Oliveira Cândido. – Belo Horizonte: EEUFMG, 2003.
174p.

Orientador: Mauro da Cunha Naghettini
Dissertação (mestrado) Universidade Federal de Minas Gerais, Escola de Engenharia.

1. Hidrologia estatística. 2. Sistema especialista. 3. Análise de freqüência. I. NAGHETTINI, Mauro da Cunha. II. Universidade Federal de Minas Gerais, Escola de Engenharia. III. Título.

CDU: 556.047

*"Uma inteligência humana é a capacidade de resolver problemas ou criar produtos que são importantes num determinado ambiente cultural ou comunidade".*

GARDNER; em sua Teoria das Inteligência Múltiplas

*"Inteligência Artificial é o estudo das faculdades mentais com o uso de modelos computacionais"*

*CHARNIAC e MCDERMOTT*

## RESUMO

As amostras usualmente curtas de observações hidrológicas, bem como as grandes incertezas envolvidas na estimação de parâmetros e quantis, tornam subjetiva a tarefa de selecionar uma função de distribuição de probabilidades para a análise de freqüência local de máximos anuais. Os testes de aderência convencionais não são nem suficientemente potentes e nem discriminatórios para prover a sustentação objetiva a este processo de tomada de decisão, podendo levar um hidrólogo inexperiente a escolhas inapropriadas. Na prática, os especialistas em análise de freqüência de variáveis hidrológicas empregam seu conhecimento e experiência passada para construir um conjunto de regras heurísticas e de procedimentos *ad hoc*, o qual provê os fundamentos necessários para justificar a escolha de uma distribuição específica, entre os seguintes modelos mais usuais: Normal, Log-Normal de 2 parâmetros, Gumbel, Generalizada de Valores Extremos, Exponencial, Generalizada de Pareto, Pearson III e Log-Pearson III. Entretanto, diferentes especialistas podem chegar a diferentes conclusões. Essa dissertação descreve a experiência de se empregar a tecnologia de inteligência artificial e elementos de lógica difusa na construção de um sistema especialista computacional que emula os princípios de raciocínio de um especialista humano ao selecionar uma distribuição de probabilidades para proceder à análise de freqüência local de variáveis hidrológicas. Um conjunto de regras contemporâneas e fundamentadas em conhecimento foi implementado em FuzzyCLIPS, uma versão da ferramenta CLIPS, desenvolvida pela NASA para a construção de sistemas especialistas; a expectativa é que estas regras forneçam diretrizes razoavelmente similares àquelas empregadas por um especialista humano ao selecionar uma distribuição de probabilidades para uso na análise de freqüência hidrológica. O sistema computacional, aqui chamado SEAF, foi escrito em linguagem Delphi e encontra-se em *interface* com o FuzzyCLIPS. Em suma, o SEAF primeiramente extrai a informação numérica dos dados amostrais, analisando-a, na seqüência, à luz do conjunto interno de regras heurísticas fundamentadas em conhecimento, transformando-a, finalmente, em declarações lingüísticas de decisão. Os momentos-L e os seus quocientes são empregados para sumarizar a variabilidade presente nos dados amostrais. Então, o processo indutivo monotônico de raciocínio usa os momentos-L amostrais para decidir sobre a plausibilidade das distribuições candidatas de 2

e 3 parâmetros, com base em informação análoga aos diagramas de momentos-L, bem como em propriedades da assimetria-L e curtose-L obtidas por simulação de Monte Carlo. O raciocínio adotado no SEAF prossegue com o teste de aderência de Filliben, o qual serve para atribuir um novo grau de confiança, independente do anterior, para a distribuição em análise; ressalte-se aqui que, durante o processo, o nível de confiança previamente atribuído a cada uma das distribuições candidatas não é alterado, o que faz com que o modo de raciocínio possa ser considerado indutivo e monotônico. Finalmente, a combinação matemática de todos os níveis de confiança já atribuídos e do número de parâmetros estimados fornece o critério de parcimônia estatística para discriminar entre distribuições da mesma família. As distribuições que permaneceram ao longo do processo são, então, classificadas de acordo com seus respectivos níveis médios de confiança e as declarações de decisão são formuladas. Este sistema foi aplicado a 20 amostras relativamente longas de alturas diárias de chuva e vazões médias diárias máximas anuais observadas em estações pluviométricas e fluviométricas localizadas na região sudeste do Brasil. Com o objetivo de verificar o desempenho do sistema, as mesmas amostras foram submetidas a um painel de especialistas em análise de freqüência. A despeito das complexidades inerentes à análise de freqüência local de eventos máximos anuais de variáveis hidrológicas, é uma conclusão dessa dissertação que um sistema protótipo, como o SEAF, desempenha-se em um nível especialista e pode vir a fornecer substancial auxílio a uma pessoa não-especialista na escolha de uma distribuição de probabilidades apropriada, entre um conjunto de possíveis modelos candidatos.

## ABSTRACT

The usually short samples of hydrological observations recorded at a site and the large uncertainties involved in parameter and quantile estimation make the task of selecting a probability distribution function for frequency analysis of annual maxima a subjective matter. Conventional goodness-of-fit tests are neither powerful enough nor discriminatory to provide the necessary objective backing to such a decision-making process and may lead a novice hydrologist to improper choices. In practice, experts in frequency analysis of hydrological variables employ their knowledge and past experience to construct a set of heuristic rules and rather *ad hoc* procedures, which provide the *rationale* for justifying the choice of a specific probability distribution function, among the following most often used candidate models: Normal, 2-parameter Log-Normal, Gumbel, Generalized Extreme Value, Exponential, Generalized Pareto, Pearson III and Log-Pearson III. However, different experts may reach to different conclusions. This dissertation provides a description of an experience of employing the technology of artificial intelligence and fuzzy-logic theory to build a computer expert system that emulates the reasoning principles of a human expert in selecting a probability distribution for at-site hydrologic frequency analysis. A set of contemporary knowledge-based rules has been implemented into FuzzyCLIPS, a version of NASA's CLIPS tool for building expert systems; these rules are expected to provide guidelines reasonably similar to those employed by a human expert for selecting a probability distribution for hydrologic frequency analysis. The computer system, called SEAF, is written in Delphi language and is interfaced with FuzzyCLIPS. Briefly, SEAF first takes the numerical information abstracted from a data sample, then analyze it under the light of the built-in knowledge-based set of heuristic rules, and, finally, transform it into decision statements. L-moments and L-moment ratios are employed to summarize the data sample variability. Then, the monotonically inductive reasoning process first uses the sample L-moments to decide on plausible 2-parameter and 3-parameter candidate parametric forms, on a basis analogous to L-moment diagrams and on Monte Carlo-derived properties of L-skewness and L-kurtosis. The reasoning process, as adopted in SEAF, proceeds with Filliben's test, which serves to assign a new confidence level, independently of the previously assigned levels, to the distribution in focus; it is worthwhile to say that, at

a given point of the process, the previous confidence levels are not changed, which makes SEAF's reasoning process a monotonically inductive one. Finally, a mathematical combination of these confidence levels and the number of estimated parameters provides the criterion of statistical parsimony to discriminate among distributions of the same family. The remaining distributions are then classified according to their respective mean overall confidence levels and the decision statements are formulated. Such a system has been applied to 20 relatively large samples of daily rainfall and streamflow annual maxima, recorded at gauging stations located in the Brazilian southeast. In order to check the system performance, the same samples have been submitted to a panel of experts in frequency analysis. Despite the complexities inherent to at-site frequency analysis of hydrological annual maxima, it is a conclusion of this dissertation that a prototype system, such as SEAF, performs at an expert level and may provide a powerful tool to help an inexpert to make a reasonable choice among a number of possible probability distribution models.

## DEDICATÓRIA

Para  minhas filhas Carolina e Ana Clara.

## AGRADECIMENTOS

Gostaria de agradecer a Deus e todas as pessoas que contribuíram para o desenvolvimento deste trabalho, em especial:

ao Prof. Mauro da Cunha Naghettini, meu orientador, pelo seu incentivo, dedicação e envolvimento em todas etapas e linhas desta dissertação, sem os quais não teria concluído;

aos membros anônimos do Painel de Especialistas, por contribuírem significativamente para o desenvolvimento deste trabalho;

à minha filha Carol, pela fundamental ajuda durante a condução dos experimentos;

ao Prof. Rafael Palmier, que, sempre através da expressão *"fala garoto!"*, demostrou o seu entusiasmo e amizade;

à minha mãe, pai, irmã, filhas e Nice, pelo companheirismo, incentivo e cobrança;

às amigas Letícia e Beth, pelos puxões de orelha;

ao Éber, Alice e Margarida, pelas suas contribuições;

aos meus amigos Marcelo Jorge e Virgínia, pelo apoio e amizade.

**ÍNDICE**

## LISTA DE FIGURAS

## LISTA DE TABELAS

## LISTA DE NOTAÇÕES

| | |
|---|---|
| $\bar{x}$ | Média aritmética |
| $\hat{C}_v$ | Coeficiente de variação amostral |
| $P_r^*(p)$ | Polinômios ortogonais |
| $\hat{g}$ ou g | Coeficiente de assimetria amostrtal |
| $\hat{s}^2$ ou $s^2$ | Variância amostral |
| 2P | Dois parâmetros |
| Ã | Conjunto difuso A |
| CF[h,e] | Fator de crença |
| CLIPS | C Language Integrated Production System |
| $C_v$ | Coeficiente de variação |
| E(x) | Valor esperado de X |
| EXP | Distribuição Exponencial |
| FLOPS | Fuzzy Logic Production System |
| $F_x$(x) ou $F$(x) | Função de distribuição de probabilidades acumuladas de X |
| $f_x$(x) ou $f$(x) | Função densidade de probabilidades de X |
| GEV | Distribuição generalizada de valores extremos |
| GPA | Distribuição generalizada de Pareto |
| GUM | Distribuição de Gumbel |
| H | Assimetria-L padronizada |
| IA | Inteligência artificial |
| iid | independentes e igualmente distribuídas |
| $K_n$ | Estatística modificada de Grubbs e Beck |
| $L(\Phi_1, \ldots, \Phi_\kappa)$ | Função de verossimilhança |
| LNR | Distribuição Lognormal de dois parâmetros |
| LP3 | Distribuição Log-Pearson III |

| | |
|---|---|
| $l_r$ | Momentos-L amostrais de ordem r |
| MC[h,e] | Medida de crença |
| MD[h,e] | Medida de descrença |
| $M_{p,r,s}$ | Momentos ponderados de probabilidades |
| $m_r$ | Momentos amostrais centrais de ordem r |
| n | Tamanho da amostra |
| NOR | Distribuição Normal |
| NRC | National Research Council |
| $\hat{k}$ ou k | Curtose amostral |
| $P(X \leq x)$ | Probabilidade da variável X não exceder a x |
| P[h,e] | Probabilidade de ocorrência de “h” condicionada a ocorrência de “e” |
| P[h] | Probabilidade de ocorrência de “h” |
| PE3 | Distribuição Pearson III |
| PMP | Precipitação máxima provável |
| RC | Representação do conhecimento |
| REQM | Raiz quadrada do erro quadrado médio |
| $R_{mim}$ | Valor de referência para o coeficiente de correlação de Filliben |
| SE | Sistema especialista |
| SEAF | Sistema Especialista para Análise de Freqüência |
| $S_{ln}$ | Desvio padrão dos logaritmos dos dados amostrais |
| $t_r$ | Razões-L amostrais de ordem r |
| USWRC | United States Water Resources Council |
| Var(x) | Variância de X |
| X | Variável aleatória contínua |
| $x(P)$ | Função dos quantis de X |

| | |
|---|---|
| $X_{alto}$ | Estatística de Grubbs e Beck para valores de outliers altos |
| $X_{baixo}$ | Estatística de Grubbs e Beck para valores de outliers baixos |
| $X_{ln}$ | Média aritmética dos logaritmos dos dados amostrais |
| Z | Curtose-L padronizada |
| $\Phi_1, \Phi_2, \ldots, \Phi_\kappa$ | Parâmetros populacionais da função de distribuição de probabilidades |
| $\Gamma(.)$ | Função gama |
| $\alpha$ | Parâmetro de escala |
| $\alpha_r$ | Momentos ponderados de probabilidades $M_{1,0,r}$ |
| $\beta_r$ | Momentos ponderados de probabilidades $M_{1,r,0}$ |
| $\phi_1, \phi_2, \ldots, \phi_\kappa$ | Estimativa dos parâmetros populacionais da função de distribuição de probabilidades |
| $\gamma$ | Coeficiente de assimetria populacional |
| $\kappa$ | Parâmetro de forma |
| $\kappa$ | Curtose populacional |
| $\lambda_r$ | Momentos-L de ordem r |
| $\mu$ | Média aritmética |
| $\mu'_r$ | Momentos populacionais de ordem r |
| $\mu_A(x)$ | Função de pertinência de *x* em Ã |
| $\mu_r$ | Momentos centrais populacionais de ordem r |
| $\sigma$ | Desvio padrão populacional |
| $\tau_r$ | Razões-L de ordem r |
| $\xi$ | Parâmetro de posição |

# CAPÍTULO I

# INTRODUÇÃO, OBJETIVOS E ORGANIZAÇÃO DA DISSERTAÇÃO

## I.1 – INTRODUÇÃO

A análise e a determinação das vazões de enchentes representam problemas correntes da Engenharia de Recursos Hídricos. Ao longo da história, não é difícil constatar a atração natural que as planícies e os vales fluviais exercem sobre as civilizações, uma vez que as áreas ribeirinhas a elas proporcionam condições favoráveis para a agricultura e vias de transporte, além de fácil acesso aos recursos hídricos, elementos indispensáveis ao desenvolvimento agrícola, industrial e urbano. Entretanto, os benefícios econômicos e sociais advindos da ocupação e uso das planícies marginais aos cursos d'água podem, com certa freqüência, ser ofuscados pelos efeitos negativos dos desastres provocados por enchentes, como perda de vidas e prejuízos econômicos às propriedades ribeirinhas. Do ponto de vista geomorfológico, não há surpresa alguma no fato dos rios ocasionalmente se re-apropriarem de suas próprias construções permanentes e dinâmicas, que são seus respectivos vales e planícies. Entretanto, é surpreendente constatar que as comunidades humanas por vezes ignoram o fato que habitar ou usar as planícies de inundação significa co-habitar com o risco de cheias.

A redução do risco de cheias e a mitigação de seus efeitos danosos podem ser proporcionadas por intervenções diversas no sistema fluvial natural, entre as quais cita-se a construção de reservatórios, o erguimento de diques de proteção, as estratégias de planejamento mais racional da ocupação das planícies de inundação, além de medidas de proteção das habitações e benfeitorias ribeirinhas. Da mesma forma, a análise e determinação de vazões de enchentes é parcela crucial do projeto e operação de estruturas hidráulicas destinadas ao aproveitamento de recursos hídricos, uma vez que a segurança destas estruturas depende fundamentalmente dessa informação. Em qualquer dessas situações, compete aos engenheiros e hidrólogos estimar características relevantes das enchentes como as precipitações a elas associadas, as vazões de pico, o volume e a duração dos hidrogramas, as áreas inundáveis, bem como seus valores críticos ou de projeto e/ou de operação. Existem metodologias diversas para a análise e

determinação destas características das enchentes, algumas das quais são construídas em bases puramente determinísticas, enquanto outras procuram associar a magnitude das variáveis à freqüência de sua superação. A preferência por uma entre estas metodologias depende de diversos fatores, tais como o objetivo do estudo, o porte das estruturas, a existência, a abundância e a qualidade dos registros plúvio-fluviométricos, entre outros.

Em diversas das circunstâncias mencionadas, é prática corrente da Engenharia de Recursos Hídricos que as variáveis hidrológicas sejam vistas como variáveis aleatórias, o que lhes confere suscetibilidade de serem analisadas sob o ponto de vista da teoria de probabilidades e da estatística matemática. Nesse sentido, a análise de freqüência de variáveis hidrológicas, aqui brevemente referida como a quantificação do número esperado de ocorrências de um evento de certa magnitude, representa uma das principais aplicações da teoria de probabilidades e da estatística matemática no campo da Engenharia de Recursos Hídricos. Os métodos de análise de freqüência buscam extrair inferências quanto à probabilidade com que uma variável irá igualar ou superar um certo valor ou quantil, a partir de um conjunto amostral de ocorrências daquela variável. Se as ocorrências referem-se a observações tomadas unicamente em um ponto específico do espaço geográfico (e.g. : uma estação fluviométrica, em uma dada bacia hidrográfica), a análise de freqüência é dita local. Contrariamente, se um número maior de observações da variável em questão, tomadas em diferentes pontos de uma certa região, forem empregadas conjuntamente para a inferência estatística, a análise de freqüência é dita regional.

A análise de freqüência local de variáveis hidrológicas dispõe de um conjunto de técnicas de inferência estatística e de modelos probabilísticos, os quais têm sido objetos freqüentes de investigação, visando principalmente a obtenção de estimativas cada vez mais eficientes e confiáveis. Entretanto, a inexistência de amostras suficientemente longas impõe um limite superior ao grau de sofisticação estatística a ser empregado na análise de freqüência local. Nesse sentido, a análise regional de freqüência representa uma alternativa que procura compensar a insuficiente caracterização temporal do comportamento de eventos extremos por uma coerente caracterização espacial da variável hidrológica em questão. Embora seja fácil constatar a preferência pelos fortes argumentos apontados pelos métodos da análise regional [e.g.: Hosking e Wallis (1997), Potter (1987) e Bobée e Rasmussen (1995)], é um fato que inúmeras decisões

envolvendo eventos hidrológicos extremos sejam tomadas com base somente na análise de freqüência local dos registros das variáveis associadas. Esse fato certamente está vinculado à inexistência ou à pouca abundância de registros de variáveis hidrológicas em uma dada região geográfica, ou à falta de conhecimento dos métodos de análise regional, ou mesmo à forma expedita com que se formulam soluções de engenharia para os problemas de mitigação e controle de eventos hidrológicos extremos. Não obstante tal fato e por ser intuitivamente perceptível que a análise de freqüência local de variáveis hidrológicas permanecerá ainda por muito tempo como um método de uso corrente da Engenharia, focaliza-se na presente dissertação as técnicas e dificuldades inerentes a esse tipo de análise e, em particular, a seleção da função de distribuição de probabilidades a ser empregada.

A análise de freqüência local de variáveis hidrológicas consiste no ajuste de uma função de distribuição de probabilidades a uma amostra de observações máximas anuais da variável em questão, as quais devem ser consideradas, por princípio, como aleatórias, independentes e homogêneas; as chamadas séries de duração parcial também podem ser usadas, embora sua utilização seja muito menos freqüente e não sejam aqui consideradas. Uma vez escolhida a função de distribuição de probabilidades dentre os diversos modelos probabilísticos disponíveis, estimam-se seus parâmetros e quantis de interesse para a variável em estudo. No contexto da análise de freqüência local de eventos máximos anuais, as funções de distribuição de probabilidades de uso mais comum em hidrologia podem ser agrupadas em (a) modelos de 2 parâmetros, entre os quais podem ser citados os de Gumbel, Log-Normal 2P, Gama e Exponencial 2P e (b) modelos de mais de 2 parâmetros, como os prescritos pelas distribuições Generalizada de Valores Extremos (GEV), Pearson III, Log-Pearson III e Wakeby, entre outras. A despeito da disponibilidade de um amplo conjunto de modelos probabilísticos, não há, entre os hidrólogos e especialistas da área, qualquer consenso quanto à prescrição de uma única função de distribuição de probabilidades que seja considerada adequada à análise de freqüência de variáveis hidrológicas. De fato, a inexistência de leis dedutivas de seleção de uma distribuição de probabilidades para a análise de freqüência de variáveis hidrológicas faz com que esse seja um procedimento *ad hoc*, baseando-se principalmente na aderência do modelo prescrito a uma amostra de observações da variável em questão.

Na prática, as amostras, geralmente curtas, de observações máximas anuais das variáveis hidrológicas tornam difícil e subjetiva a escolha de uma distribuição de probabilidades com base, principalmente, em sua aderência a um conjunto de dados. Com amostras de tamanhos típicos entre 20 e 60, é impossível afirmar categoricamente que uma certa distribuição de probabilidades, considerada aderente aos dados, irá representar o verdadeiro comportamento populacional. Os métodos convencionais de inferência estatística produzem estimativas pouco confiáveis de parâmetros e quantis, devido principalmente à grande incerteza imposta pelas amostras de pequeno tamanho. Por sua vez, os testes estatísticos de aderência, como o do Qui-Quadrado e o de Kolmogorov-Smirnov, são pouco potentes e incapazes de discriminar entre modelos probabilísticos, resultando que mais de uma distribuição pode ser considerada aderente a uma dada amostra.

Estas dificuldades fazem com que a seleção judiciosa de uma certa função de distribuição de probabilidades seja uma tarefa de especialistas, que geralmente a desempenham à luz de um conjunto de *regras heurísticas* formuladas de acordo com o conhecimento acumulado ao longo de anos de experiência e estudo. A abordagem heurística limita os caminhos a seguir, selecionando aqueles considerados melhores e reduzindo uma tarefa complexa a um conjunto de operações de julgamento. Por exemplo, um determinado especialista pode sugerir algumas distribuições candidatas simplesmente com base no exame do coeficiente de assimetria amostral e, na seqüência, sugerir a adoção daquela que produzir a melhor aderência em papel de probabilidades. Em geral, essas regras heurísticas facilitam o trabalho do especialista, porém, se usadas indiscriminadamente, podem conduzir a resultados tendenciosos. Como exemplo disso, suponha que um evento, sabidamente muito raro em uma dada amostra, produza um coeficiente de assimetria extremamente alto. Nesse caso, se a seleção das distribuições candidatas se der com base somente no coeficiente de assimetria amostral, correr-se-ia o risco de se prescrever modelos ‘adequadamente’ assimétricos, porém incorretos. Esse exemplo vem mostrar que, longe de ser uma tarefa balizada por um conjunto sistemático de regras, a seleção judiciosa de uma distribuição de probabilidades é um procedimento multi-critério, de caráter heurístico e inexato, o qual é passível de ser realizado por analistas especializados, com experiência e conhecimentos específicos.

Apesar de numerosos programas computacionais terem sido desenvolvidos como ferramentas auxiliares para a análise de freqüência local de eventos máximos anuais, eles não são capazes de orientar o usuário sobre a escolha da função de distribuição de probabilidades ou do conjunto de distribuições mais adequadas para a amostra em questão. Os resultados apresentados são dados numéricos, os quais podem levar um hidrólogo inexperiente a encontrar dificuldades em decidir qual distribuição será adotada em sua análise, haja vista que duas ou mais distribuições podem ser aceitas pelos testes de hipóteses adotados.

Dependendo da distribuição populacional escolhida, o quantil estimado para a variável característica de projeto ou de decisão pode inviabilizar economicamente o empreendimento ou sujeitá-lo a um maior risco de falha. Dentro deste cenário, a escolha de uma distribuição de probabilidade para um conjunto de dados é importante e não pode ser realizada por um simples e único algoritmo. Ela requer a agregação de análises objetivas e subjetivas, as quais podem conduzir a resultados diferentes dependendo do padrão de raciocínio adotado por cada especialista.

De um modo geral, sempre que a solução de um problema envolve a combinação de critérios subjetivos, a utilização dos chamados Sistemas Especialistas pode ser uma alternativa interessante para a padronização e informatização do processo de análise e tomada de decisão. Um sistema especialista é um programa de computador projetado e desenvolvido para atender a uma aplicação determinada do conhecimento humano. Ele é capaz de auxiliar a tomada de uma decisão, apoiada em conhecimento justificado a partir de uma base de informações, tal qual um especialista de uma área específica do conhecimento. No caso presente, essa base de informações pode ser construída a partir da reunião de conhecimentos de profissionais qualificados e experientes em análise de freqüência local de variáveis hidrológicas e compor uma lógica para a tomada de decisão sob condições subjetivas. A presente dissertação constitui um relato da experiência de se conceber, implementar e testar um protótipo de um sistema especialista de auxílio à tomada de decisão, tendo como base um conjunto definido de regras heurísticas que reflitam as tendências e os resultados da pesquisa recente em análise de freqüência local de variáveis hidrológicas.

## I.2 - OBJETIVOS

O objetivo geral dessa pesquisa é a concepção e o desenvolvimento de um protótipo de um sistema especialista integrado para análise de freqüência local de eventos hidrológicos máximos anuais, utilizando a tecnologia de inteligência artificial. Esse sistema foi concebido tendo como principal requisito a capacidade de emular um certo padrão de raciocínio, adotado por especialistas da área, durante a condução da análise de freqüência local de eventos hidrológicos máximos anuais.

Especificamente o objetivo principal é o de organizar e implementar um sistema de raciocínio que seja capaz de selecionar uma única (ou um conjunto pequeno de modelos distributivos) entre as distribuições candidatas Normal, Log-Normal 2P, Gumbel, Exponencial, Pearson III, Log-Pearson III, Generalizada de Valores Extremos (GEV) e Generalizada de Pareto (GPA) para modelar a distribuição populacional de dados amostrais de máximos hidrológicos anuais. Para cumprir tal objetivo, o protótipo do sistema foi projetado de modo a ter duas componentes distintas:

- A primeira, desenvolvida em linguagem Delphi, é responsável pelos cálculos matemáticos e estatísticos envolvidos em uma típica análise de freqüência de máximos hidrológicos anuais, a saber: estatísticas descritivas amostrais, testes de hipóteses, testes de aderência, estimação dos parâmetros e intervalos de confiança;
- A segunda, desenvolvida em linguagem CLIPS (*C Language Integrated Production System*), é responsável pela análise dos dados e emissão de uma opinião justificada sobre a opção escolhida, utilizando os conceitos de conjuntos difusos para classificar as distribuições.

Os objetivos específicos da dissertação são:

- Testar o protótipo do sistema desenvolvido, aplicando-o a um conjunto de 20 amostras de máximos hidrológicos anuais, com tamanhos amostrais suficientes para se proceder à análise de freqüência local;
- Avaliar o sistema desenvolvido por meio da comparação de seus resultados àqueles obtidos por um painel de reconhecidos especialistas em análise de

freqüência local de eventos hidrológicos máximos anuais, com base nas mesmas 20 amostras mencionadas.

## I.3 – ORGANIZAÇÃO DA DISSERTAÇÃO

Esta dissertação está organizada da seguinte forma:

- Capítulo I - INTRODUÇÃO, OBJETIVO E ORGANIZAÇÃO: aqui, é feita uma introdução à importância do tema, bem como o objetivo e organização do conteúdo da presente dissertação;

- Capítulo II - ANÁLISE DE FREQÜÊNCIA DE EVENTOS HIDROLÓGICOS MÁXIMOS ANUAIS: neste capítulo, é feita uma revisão bibliográfica do processo de análise de freqüência local de eventos hidrológicos máximos anuais;

- Capítulo III - INTELIGÊNCIA ARTIFICIAL – SISTEMAS ESPECIALISTAS: onde são abordados os principais tópicos de inteligência artificial, mais especificamente, aqueles ligados à construção de sistemas especialistas de auxílio à decisão;

- Capítulo IV – SEAF - SISTEMA ESPECIALISTA PARA ANÁLISE DE FREQUÊNCIA LOCAL DE EVENTOS HIDROLÓGICOS MÁXIMOS ANUAIS: neste capítulo, são descritas todas as regras heurísticas implementadas no protótipo do sistema especialista, aqui referenciado pelo acrônimo SEAF (Sistema Especialista para Análise de Freqüência), e onde também se exemplifica suas principais funções e *interfaces*;

- Capítulo V - ANÁLISE DE DESEMPENHO DO SISTEMA ESPECIALISTA – SEAF: aqui é feita uma análise do desempenho do sistema SEAF, por meio de comparação entre os resultados obtidos pelo painel de quatro especialistas e os apresentados pelo sistema SEAF para 20 amostras de dados de alturas de precipitação e vazões médias diárias máximas anuais;

- Capítulo VI - CONCLUSÕES E RECOMENDAÇÕES: neste capítulo são apresentadas as principais conclusões obtidas na concepção, implementação e teste do sistema SEAF e recomendações para investigações pertinentes futuras;

- REFERÊNCIAS BIBLIOGRÁFICAS: aqui estão listadas todas as publicações referenciadas no texto da dissertação;

- ANEXO A1 - ALGUMAS DISTRIBUIÇÕES DE PROBABILIDADES UTILIZADAS EM ANÁLISE DE FREQÜÊNCIA: este anexo apresenta as propriedades das distribuições de probabilidades mais usadas na análise de frequência de eventos hidrológicos máximos anuais;

- ANEXO A2 - TESTES NÃO PARAMÉTRICOS: aqui estão detalhados dois testes não paramétricos, sendo um para identificação da presença de tendência (Mann-Kendall) e o outro para identificação da presença de dependência serial (Kendall);

- ANEXO A3 - TESTES PARA VERIFICAÇÃO DE ADERÊNCIA: neste anexo estão descritos quatro testes para verificação de ajuste dos dados a uma distribuição de probabilidades; são eles os testes do Qui-Quadrado, Kolmogorov-Smirnov, Filliben e o de Momentos-L amostrais versus teóricos;

- ANEXO A4 - MOMENTOS-L: aqui foram abordados alguns conceitos referentes à teoria dos momentos-L;

- ANEXO A5 – AMOSTRAS DE DADOS HIDROLÓGICOS: estão apresentadas neste anexo os dados referentes às 20 amostras de alturas de precipitação e de descargas médias diárias máximas anuais utilizadas para a análise do desempenho do sistema SEAF.

# CAPITULO II

## ANÁLISE DE FREQÜÊNCIA DE EVENTOS HIDROLÓGICOS MÁXIMOS ANUAIS

### II.1 – INTRODUÇÃO

A Hidrologia trata dos complexos processos naturais de armazenamento e transporte encontrados no ciclo da água. Os engenheiros hidrólogos aliam seus conhecimentos àqueles provindos da ciência hidrológica e os empregam em suas atividades de planejamento, projeto e operação de estruturas de aproveitamento de recursos hídricos e em estratégias de controle ou mitigação de enchentes e estiagens severas. Por exemplo, os projetos de obras de engenharia, tais como pontes, barragens, vertedores, galerias de drenagem, entre outras, necessitam, de forma geral, do conhecimento da variabilidade, bem como da estimativa de variáveis hidrológicas. Entre estas, as mais comuns são a vazão de pico e o volume de cheia para uma dada duração, cujos valores característicos representam informações cruciais para o dimensionamento das estruturas de segurança pertinentes ao empreendimento.

Os principais processos do ciclo hidrológico, tais como a precipitação, a evaporação, a infiltração e o escoamento em rios, possuem considerável variabilidade espaço-temporal e uma complexa interdependência, fatos que dificultam a análise quantitativa e qualitativa dos mesmos. Em geral, os problemas hidrológicos podem ser analisados e visualizados por meio de modelos, os quais podem se estender desde uma simples relação empírica, ou mesmo um aparato analógico, até uma complexa combinação de equações ou relações matemáticas, as quais podem ter forte fundamentação física, e/ou estatística e/ou empírica. Geralmente, os chamados modelos hidrológicos são classificados como determinísticos, paramétricos e estocásticos.

Nos modelos determinísticos, todas as equações que governam o fenômeno hidrológico em questão podem ser determinadas e resolvidas, sendo o resultado, decorrente da realização do processo, o mesmo para um determinado conjunto condicional de valores de entrada. Em contraposição direta, os modelos estocásticos possuem resultados previsíveis apenas em termos estatísticos, ou seja, as repetições de um dado conjunto de entrada produzem resultados distintos. Os modelos paramétricos

são formados por um conjunto de relações matemáticas e/ou empíricas que possuem parâmetros, os quais devem ser estimados a partir de dados experimentais. Eles podem ser considerados determinísticos no sentido que irão produzir sempre o mesmo resultado para as mesmas condições de entrada. Por outro lado, eles podem ser considerados como estocásticos, no sentido de que as estimativas dos parâmetros irão mudar caso a amostra de dados experimentais for alterada.

Em termos simplistas, poder-se-ia dizer que se todos os fatores causais pudessem ser definidos e medidos com a precisão necessária e se todas as relações de interdependência entre eles fossem conhecidas e determinadas, os processos hidrometeorológicos seriam classificados como determinísticos. Entretanto, o estágio atual das observações sistemáticas e do conhecimento humano sobre tais processos não permite que eles sejam tratados desta forma, em todas as situações; por exemplo, o número e a magnitude das enchentes de um rio, durante um certo período do ano, podem somente ser descritos como processos do tipo estocástico. Nesse caso, trata-se de um processo aleatório porque nem todos os fatores causais e/ou influentes na formação local de cheias, bem como suas interdependências nas escalas espacial e temporal, podem ser determinados. De fato, as distribuições espacial e temporal da precipitação, a velocidade e a direção de deslocamento da tormenta sobre a bacia, as variações temporais e espaciais das perdas por intercepção, evaporação e infiltração, bem como as condições antecedentes de armazenamento da umidade do solo, fazem parte do grande número de fatores interdependentes que podem causar cheias ou influir em sua formação e intensificação.

A teoria de probabilidades e a estatística matemática apresentam um conjunto de metodologias a serem utilizadas para a identificação e modelação da aleatoriedade presente em fenômenos dessa natureza. Dentre estas, a análise de freqüência de variáveis hidrológicas tem grande destaque, pois pode ser usada como ferramenta para a solução de problemas que envolvam a inferência quanto a estimativas de alguns de seus valores característicos, e conseqüente tomada de decisão. De fato, a análise de freqüência é, sem dúvida, o emprego mais comum e mais antigo da teoria de probabilidades e da estatística matemática em hidrologia. Apesar disso, e devido não só à falta de consenso entre os pesquisadores quanto aos diferentes métodos empregados, como também à grande

incerteza presente em suas estimativas, a análise de freqüência de variáveis hidrológicas continua sendo objeto de um grande volume de investigações recentes, como verificado por Potter (1987) e Bobée e Rasmussen (1995).

Quanto à sua abordagem no espaço, a análise de freqüência pode ser classificada como local ou regional. No primeiro caso, é utilizada uma série de registros hidrométricos ou hidrometeorológicos de uma dada estação de observação para definir os chamados quantis de interesse, ou sejam os valores da variável em questão, associados a certas probabilidades de excedência. Nesta abordagem, somente os dados hidrométricos ou hidrometeorológicos do local em estudo são utilizados durante a análise. Uma segunda abordagem, que leva em conta certas características regionais e a maior possibilidade de flutuação das observações pontuais em torno dos verdadeiros valores populacionais, constitui a chamada análise de freqüência regional de variáveis hidrológicas. Neste caso, os dados coletados em vários postos hidrométricos ou hidrometeorológicos podem ser agrupados e analisados em subconjuntos, os quais estão associados a certas similaridades fisiográficas ou climáticas de uma área geográfica, permitindo, assim, a transferência de informações de um local para outro, sob a premissa de semelhança hidrológica ou hidrometeorológica. Portanto, o princípio da análise regional baseia-se na similaridade espacial de algumas funções, variáveis e parâmetros que permitem a sua transferência de um local a outro, dentro de um contexto geográfico. Além da estimativa de quantis em locais desprovidos de observações ou com amostras muito pequenas, a análise de freqüência regional pode ser também utilizada na estimação mais confiável de parâmetros e quantis de distribuições de probabilidades em relação às estimativas realizadas somente pela análise de freqüência dos registros locais.

Em relação aos valores constituintes da série hidrológica ou hidrometeorológica, a análise de freqüência pode ser conduzida a partir das chamadas séries de duração anual ou parcial. As séries de duração anual são construídas pela seleção dos máximos (ou mínimos) registrados nas estações hidrométricas ou hidrometeorológicas, para cada ano hidrológico; ou seja, destas séries, consta apenas um valor de máximo (ou mínimo) por ano. Em contraposição, as séries de duração parcial compreendem somente os

registros de magnitudes superiores (ou inferiores) a um determinado limiar de referência, independentemente de suas respectivas datas de ocorrência.

Tal como observado por Bobée e Rasmussen (1995), pode-se verificar uma tendência recente bastante clara de uso preferencial dos métodos da análise regional, em particular aqueles descritos por Hosking e Wallis (1997). Por outro lado, existe grande controvérsia quanto ao emprego preferencial das séries de duração parcial em relação às séries de máximos anuais; por exemplo, enquanto alguns preconizam as primeiras [e.g.: Smith (1984) e Naghettini et al. (1996)], outros, como Cox et al. (2002), demonstram que, sob certas circunstâncias, as séries de máximos anuais produzem quantis de variância relativamente menor. Não obstante essas observações e tal como registrado no Capítulo I da presente dissertação, a pouca abundância espacial de estações hidrométricas ou hidrometeorológicas em diversos países, o Brasil entre eles incluído, e a ainda relativamente incompleta compreensão das vantagens dos métodos regionais, fazem com que a análise de freqüência local de eventos hidrológicos máximos anuais continue a ter, no meio técnico, uma difusão muito maior do que as abordagens alternativas mencionadas. Tendo em vista tal fato, bem como o objetivo principal desta dissertação, que é o de conceber e implementar um sistema especialista para auxiliar um engenheiro hidrólogo na análise de freqüência local de eventos hidrológicos máximos anuais, o presente capítulo restringe-se apenas aos tópicos associados ao referido objetivo. Desse modo, não serão aqui contemplados os aspectos relacionados à análise de freqüência regional, bem como aqueles ligados à utilização de séries hidrológicas de duração parcial, sendo o leitor remetido às referências anteriormente citadas para os detalhes pertinentes. Entretanto, é oportuno vislumbrar a futura possível extensão dos objetivos e métodos da presente dissertação às abordagens de análise de freqüência regional e/ou séries de duração parcial.

## II.2 – FUNDAMENTOS

Visando a correta contextualização da terminologia e dos métodos da análise de freqüência de variáveis hidrológicas ao longo da presente dissertação, faz-se aqui uma

breve formalização dos conceitos e da notação empregados. Seja X uma variável aleatória contínua, cuja função de distribuição de probabilidades acumuladas é dada por

$$F_x(x) = P(X \leq x) \tag{II.1}$$

A função densidade de probabilidades, denotada por $f_x$(x), é definida como a derivada primeira de $F_X(x)$ em relação a $X$, enquanto $x(p)$ representa a função dos quantis de $X$, tal que a probabilidade da variável não exceder o valor $x(p)$ é igual a $p$.

O valor esperado ou esperança matemática da variável aleatória $X$, denotado por $E(X)$, é um operador definido por

$$E(x) = \int_{-\infty}^{\infty} x f(x) dx \tag{II.2}$$

Considerando a transformação $p=F(x)$, pode-se reescrever a equação II.2 da seguinte forma

$$E(x) = \int_{0}^{1} x(p) dp \tag{II.3}$$

Da mesma forma, a função de variável aleatória $g(x)$ é também uma variável aleatória e sua esperança matemática é dada por

$$E[g(x)] = \int_{-\infty}^{\infty} g(x) f_x(x) dx = \int_{0}^{1} g[x(p)] dp \tag{II.4}$$

A variância de $X$, simbolizada por var($x$), representa uma medida da dispersão dos valores de X em torno do valor central $E(x)$ e é definida pela seguinte expressão:

$$\text{var}(x) = E\{[X - E(x)]^2\} = E(x^2) - [E(X)]^2 \tag{II.5}$$

A distribuição da variável aleatória X é completamente conhecida se o conjunto de parâmetros $\Phi_1, \Phi_2, \ldots, \Phi_k$, associado à definição das funções $f_x$ (x;$\Phi_1$,$\Phi_2$,... ,$\Phi_k$) ou $x(p; \Phi_1, \Phi_2, \ldots, \Phi_k)$ for conhecido. A maioria das funções de distribuição de probabilidades requer a definição dos parâmetros de posição e de escala. O parâmetro de posição $\xi$ é o número real que satisfaz

$$x(p; x, \Phi_2, \ldots, \Phi_k) = x + x(p; 0, \Phi_2, \ldots, \Phi_k) \tag{II.6}$$

O parâmetro de escala $\alpha$ de uma distribuição, cujo parâmetro de posição é $\xi$ , é dito de escala se

$$x(p;x,a,\Phi_3,...,\Phi_k) = x + a.x(p;0,1,\Phi_3,...,\Phi_k) \quad \text{(II.7)}$$

Os parâmetros de uma distribuição devem ser estimados a partir de uma amostra de dados observados. O estimador de um certo parâmetro $\Phi$ é representado por $\phi$, o qual, por ser uma função dos dados amostrais, é também uma variável aleatória.

As características das distribuições de probabilidades podem ser sumarizadas pelos momentos populacionais, os quais podem ser calculados pela seguinte equação:

$$m_r' = \int_{-\infty}^{\infty} x^r f_x(x)dx \quad \text{(II.8)}$$

onde *r* é *r*-ésimo momento.

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

$$m = E(x) \quad \text{(II.9)}$$

O *r*-ésimo momento central é definido como o *r*-ésimo momento sobre a média, μ, de uma distribuição de probabilidades e é dado por:

$$m_r = \int_{-\infty}^{\infty} (x-m)^r f_x(x)dx \quad \text{(II.10)}$$

Em decorrência da equação II.10, os momentos centrais de ordem igual ou superior a 2 podem ser calculados como valores esperados das *r*-ésimas potências dos desvios da variável em relação ao centro da distribuição μ. Em termos formais,

$$m_r = E(x-m)^r; r = 2,3,... \quad \text{(II.11)}$$

Alguns momentos centrais de particular interesse são os de ordem 2, 3 e 4. O momento central de ordem 2 é por definição a variância de *X*, geralmente simbolizada por var(*x*) ou $\sigma^2$. As quantidades que podem ser deduzidas do momento central de ordem 2 são o desvio padrão $\sigma$ e o coeficiente de variação $C_v$, as quais são definidas por:

$$s = \sqrt{m_2} = \sqrt{s^2} \quad \text{(II.12)}$$

$$C_v = \frac{s}{m} \quad \text{(II.13)}$$

Para $r > 2$, é usual descreverem-se as características das funções de distribuição através de razões adimensionais do tipo $m_r / (m_2)^{r/2}$, entre as quais se destacam o coeficiente de assimetria:

$$g = \frac{m_3}{(m_2)^{\frac{3}{2}}} \tag{II.14}$$

e a curtose:

$$k = \frac{m_4}{(m_2)^2} \quad ou \quad k = \frac{m_4}{(m_2)^2} - 3 \tag{II.15}$$

## II.3 – ETAPAS DA ANÁLISE DE FREQÜÊNCIA LOCAL DE MÁXIMOS ANUAIS

A análise convencional de freqüência de realizações de uma variável aleatória, da qual se conhece uma amostra e a distribuição de probabilidades da população de onde a amostra foi retirada, consiste em estimar os parâmetros populacionais a partir dos dados observados e, em seguida, estimar os quantis para a probabilidade desejada. No caso de eventos máximos (e/ou mínimos) de variáveis hidrológicas, a distribuição de probabilidades da população não é conhecida e tem-se somente uma amostra de dados observados. Esse fato complicador leva à proposição de modelos probabilísticos, funções paramétricas de probabilidade, os quais, em função de suas características de assimetria e da eventual existência de limites superiores (e/ou inferiores) no domínio de definição da variável aleatória, se atribuem propriedades de modelarem os fenômenos hidrológicos. Muitas distribuições têm sido propostas para a modelação estatística dos valores máximos anuais de variáveis hidrológicas ou hidrometeorológicas, mas não há uma distribuição específica consensual que seja capaz de, sob quaisquer condições, descrever o comportamento da variável em foco. Portanto, em uma análise típica, cabe ao especialista selecionar dentre as diversas distribuições candidatas, aquela que parece mais apropriada à modelação dos dados amostrais. Em geral, os procedimentos típicos de uma análise de freqüência local são os seguintes:

- verificação dos dados amostrais;
- escolha da distribuição de probabilidades;

- estimativa dos parâmetros das distribuições; e
- identificação e tratamento de pontos atípicos ou *outliers*.

As seções que se seguem apresentam uma breve discussão sobre cada uma das etapas descritas acima, enquanto o anexo A1 apresenta, em detalhes, algumas das principais distribuições de probabilidades utilizadas na modelação de eventos máximos de variáveis hidrológicas e hidrometeorológicas.

### II.3.1 – Verificação dos dados amostrais

A qualidade e a aplicabilidade da análise de freqüência dependem diretamente dos dados utilizados para estimação de seus parâmetros. Desse modo, é um fato reconhecido que, por mais sofisticado que seja, a qualidade de um modelo estocástico jamais superará a dos dados disponíveis para a estimação de seus parâmetros. Nesse sentido, cabe ao especialista julgar a qualidade dos registros hidrológicos disponíveis para dar prosseguimento à analise de freqüência.

É um pressuposto da análise de freqüência convencional que a amostra de dados disponível seja uma entre um número infinito de outras amostras possíveis que também poderiam ser sorteadas, todas com igual chance, de uma população. Também são pressupostos da análise de freqüência convencional que os dados hidrológicos devem satisfazer as condições de independência, estacionariedade e representatividade. De modo sintético, pode-se dizer que os eventos são considerados independentes quando não há correlação entre os valores da série. No caso de vazões máximas anuais, a independência significa a inexistência de correlação entre o registro de um dado ano e os registros posterior e anterior, considerados todos os anos disponíveis. Por outro lado, uma série de máximos hidrológicos é dita estacionária quando não ocorrem modificações nas características estatísticas de sua série ao longo do tempo. A análise de freqüência de séries hidrológicas não estacionárias e, por conseguinte, a estimação de parâmetros e quantis com tendências ou variações temporais são objetos de investigação muito recentes [e.g: Cox et al. (2002) e Clarke (2002)] e não serão aqui considerados. Em termos da análise de freqüência convencional, dados não estacionários devem ser analisados em sub-séries homogêneas ou ajustados de modo a corrigir as

heterogeneidades encontradas. As causas principais de possíveis heterogeneidades em uma série hidrológica ou hidrometeorológica são: a relocação das estações de observação, a construção de barragens a montante, a urbanização ou o desmatamento das bacias, as eventuais modificações do leito fluvial, a ocorrência de cheias catastróficas, além, evidentemente, de mudanças climáticas.

A confiabilidade das estimativas dos parâmetros de uma dada distribuição de probabilidade está intrinsecamente ligada ao tamanho da amostra e à sua representatividade. Os dados da amostra devem ser representativos da variabilidade inerente ao processo natural ou experimento em foco. Em se tratando de variáveis hidrológicas ou hidrometeorológicas, uma amostra, obtida ao longo de um período predominantemente seco (ou úmido), irá certamente distorcer os resultados da análise, produzindo, em conseqüência, estimativas tendenciosas dos parâmetros populacionais. Por outro lado, uma amostra de dados possui propriedades estatísticas apenas similares às da população; elas serão idênticas se e somente se toda a população tiver sido amostrada. Yevjevich (1972) resume a questão afirmando que tanto a presença de erros sistemáticos em uma amostra, os quais podem ser provenientes de problemas de processamento e medição, de heterogeneidades e falta de representatividade, quanto os erros aleatórios, esses inerentes às naturais flutuações amostrais em torno de valores populacionais, podem produzir grandes incertezas quanto às estimativas de parâmetros estatísticos a partir de amostras de tamanho relativamente pequeno. De qualquer modo, é um pressuposto básico dos métodos de inferência estatística a inexistência de erros sistemáticos, atribuindo somente às flutuações amostrais as diferenças entre estimativas e valores populacionais.

Benson (1960) utilizou uma série sintética de 1000 anos de vazões máximas e demonstrou que para se estimar uma cheia de 50 anos são necessárias amostras de pelo menos 39 anos, para que as estimativas ficassem na faixa de 24% do valor correto em 95% dos casos. Caso a confiança de acerto decresça para 80%, o período mínimo de dados necessário seria de 15 anos. É freqüente encontrar na literatura referências à consideração de que, a partir de uma série de máximos anuais de $n$ valores pode-se estimar, com alguma confiabilidade, quantis com tempos de retorno de até $2n$. Watt et al. (1988), editores do guia "Hydrology of Floods in Canada - A Guide to Planning and

Design", elaborado para o Conselho Nacional de Pesquisas do Canadá, relacionam o tamanho da amostra ao tipo de abordagem a ser tomada pela análise de freqüência de vazões máximas. Neste guia, a análise de freqüência local de vazões máximas anuais é recomendada apenas para as amostras com mais de 10 anos de dados e para estimativas de quantis com tempos de retorno no máximo inferiores a quatro vezes o tamanho da série. Por outro lado, Tucci (2002) opina que " ... *na realidade, não é o número de anos, mas a representatividade dos anos da série utilizada para amostrar a estatística da variável, que permite uma boa estimativa dos parâmetros da população".* Apesar de existirem outras formas de avaliar qualitativamente a aplicabilidade da análise de freqüência, conforme indicado por Tucci (2002), não se pode negar a importância do tamanho da amostra como uma boa forma de avaliação qualitativa dos estimadores amostrais e quantis, uma vez que a variância de todos eles é inversamente proporcional ao tamanho da amostra.

Testes estatísticos paramétricos e não paramétricos podem ser usados como ferramentas auxiliares na identificação da presença de dependência e heterogeneidade serial. Os testes paramétricos são fundamentados em suposições distributivas mais severas do que as exigidas por testes não paramétricos similares. Geralmente, em sua formulação, os testes paramétricos são baseados na suposição de uma distribuição de probabilidades específica para os dados amostrais. Em contraposição, os testes não paramétricos, também chamados de "testes livres de distribuição", não exigem a premissa de uma distribuição de probabilidade específica e têm suas estatísticas de decisão construídas com base em características indiretas dos dados originais, tais como sinal em relação a um valor de referência ou número de ordem dos dados classificados, entre outras. Portanto, tendo como motivação não assumir *a priori* compromissos com as características distributivas populacionais durante a etapa de verificação de dados amostrais, é claramente recomendável o uso de testes não paramétricos para a identificação da eventual presença de heterogeneidade e dependência serial na amostra. A esse respeito, o anexo A2 apresenta a descrição detalhada de dois testes estatísticos não paramétricos, nominalmente os de Kendall e Mann-Kendall, para a identificação da presença de dependência serial e heterogeneidade em uma amostra, respectivamente. Estes testes representam o ponto de partida de uma típica análise de freqüência de eventos máximos anuais de variáveis hidrológicas ou hidrometeorológicas. Cabe

esclarecer, entretanto, que, embora os testes estatísticos sejam válidos para pequenas amostras e sob situações diversas, eles devem ser vistos apenas como indicadores, pois não constituem por si argumentos suficientemente fortes para se abandonar uma amostra caso indiquem, por exemplo, a presença de dependência serial entre seus dados. Nestes casos, deve-se procurar por uma evidência física que justifique o resultado do teste.

Ainda na etapa de verificação inicial de dados, deve-se lembrar que alguns cuidados devem ser tomados durante a seleção dos eventos de modo a assegurar a independência serial da amostra. Em regiões com sazonalidade muito acentuada, a seleção de eventos para compor uma dada amostra deve ser feita de forma diferenciada para vazões máximas e mínimas anuais; por exemplo, no estado de Minas Gerais, como de resto em grande parte da região sudeste do Brasil, a estação chuvosa vai de Outubro a Março, com grande possibilidade de ocorrência de eventos máximos em Dezembro. Neste caso, as vazões máximas anuais devem ser individualizadas por ano hidrológico, o qual corresponde a um período fixo de 12 meses, a começar no início do período chuvoso (Outubro) e terminar no final da estação seca (Setembro). Mesmo em regiões com sazonalidade não tão evidente como o sudeste brasileiro, tais como o sul de Santa Catarina e grande parte do Rio Grande do Sul, o ano hidrológico de Maio a Abril deve ser empregado para a seleção de eventos. Por outro lado, no caso da seleção da amostra de vazões mínimas anuais, a abordagem anterior merece restrições, já que uma estiagem prolongada pode fazer com que valores dependentes sejam escolhidos. Neste caso, os períodos anuais devem ser limitados pelos meses mais chuvosos.

### II.3.2 – Escolha da distribuição de probabilidades

Existe um conjunto não muito extenso de funções de distribuição de probabilidades que podem ser empregadas para a modelação de eventos máximos anuais de variáveis hidrológicas e hidrometeorológicas. Dentro desse conjunto, pode-se distinguir as distribuições oriundas da teoria clássica de valores, quais sejam, as distribuições Gumbel, Fréchet, Weibull e a Generalizada de Valores Extremos (GEV), e aquelas ditas não-extremais, entre as quais as de maior uso são: as distribuições Exponencial e sua forma mais geral que é a Generalizada de Pareto, Pearson III, Log-

Pearson III e Log-Normal 2p. Embora a adequação destas distribuições candidatas dependa de critérios variados, incluindo alguns de caráter subjetivo, talvez o atributo mais desejável seja a capacidade dessas distribuições de reproduzir algumas características amostrais relevantes. Apresentam-se, a seguir, as principais considerações, algumas delas adaptadas do trabalho de Davis e Naghettini (2001), a levar em conta quando da seleção de um modelo probabilístico local.

No que concerne a distribuições limitadas à direita, é um fato que algumas quantidades físicas possuem limites superiores inerentemente definidos; é o caso, por exemplo, da concentração de oxigênio dissolvido em um corpo d'água, limitado fisicamente em um valor entre 9 a 10 mg/l, a depender da temperatura ambiente. Outras quantidades podem possuir um limite superior; entretanto, esse limite não é conhecido *a priori*, fato decorrente da insuficiente compreensão e/ou quantificação de todos os processos físicos causais envolvidos. A esse respeito, é bastante conhecida a controvérsia quanto à existência da Precipitação Máxima Provável (PMP), originalmente formulada como um limite superior de produção de precipitação pelo ar atmosférico; se de fato existe a PMP, a determinação desse limite superior fica comprometida pela insuficiente quantificação da variabilidade espaço-temporal das variáveis que lhe dão origem. Entretanto, pode-se conjeturar que seria fisicamente impossível a ocorrência de uma vazão, digamos de 100.000 $m^3/s$, em uma pequena bacia hidrográfica, por exemplo, da ordem de 100 $km^2$ de área de drenagem. Por essa razão, alguns pesquisadores, como Boughton (1980) e Laursen (1983), recomendam que somente distribuições limitadas superiormente devem ser usadas para modelar variáveis com essas características. Hosking & Wallis (1997) consideram errônea essa recomendação e sustentam que, se o objetivo da análise de freqüência é o de estimar o quantil de tempo de retorno de 100 anos, é irrelevante considerar como "fisicamente impossível" a ocorrência do quantil de 100.000 anos. Acrescentam que impor um limite superior ao modelo probabilístico pode comprometer a obtenção de boas estimativas de quantis para os tempos de retorno que realmente interessam. Hosking & Wallis (1997) concluem afirmando que, ao se empregar uma distribuição ilimitada superiormente, as premissas implícitas são (i) que o limite superior não é conhecido e nem pode ser estimado com a precisão necessária e (ii) 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 um limite superior. Evidentemente, quando existem evidências empíricas 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.

O chamado "peso" da cauda superior de uma função 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 é proporcional às probabilidades de excedência associadas a quantis elevados e é reflexo da intensidade com que a função densidade $f(x)$ decresce quando $x$ tende para valores muito elevados. Os pesos das caudas superiores de algumas das principais funções de distribuição de probabilidades encontram-se relativizados na Tabela II.1.

Tabela II.1 – Pesos das caudas superiores de algumas distribuições de probabilidade •.

| Cauda Superior | Forma de $f(x)$ para valores elevado de $x$ | Distribuição |
|---|---|---|
| Pesada | $x^{-A}$ | Generalizada de valores extremos, generalizada de Pareto e Logística generalizada com parâmetro de forma $k<0$. |
| ↑ | $x^{-A\ln x}$ | Lognormal com assimetria positiva. |
| | $\exp(-x^{A})$ $0<A<1$ | Weibull com parâmetro de forma λ<1. |
| | $x^{A}\exp(-Bx)$ | Pearson tipo III com assimetria positiva. |
| | $\exp(-x)$ | Exponencial, Gumbel. |
| ↓ | $\exp(-x^{A}), A>1$ | Weibull com parâmetro de forma λ<1. |
| Leve | Limite superior | Generalizada de valores extremos, generalizada de Pareto e Logística generalizada com parâmetro de forma $k>0$, Lognormal e Pearson tipo III com assimetria negativa. |

•$A$ e $B$ representam constantes positivas. (adap. de *Hosking & Wallis*, 1997, p. 75)

Para a maioria das aplicações envolvendo variáveis hidrológicas/hidrometeorológicas, a correta prescrição da cauda superior de uma distribuição de probabilidades é de importância fundamental e, em muitos casos,

representa a motivação primeira da análise de freqüência. Entretanto, os tamanhos das amostras disponíveis para essas aplicações são invariavelmente insuficientes para se determinar, com exatidão, a forma da cauda superior do modelo probabilístico. Segundo Hosking & Wallis (1997), não havendo razões suficientes para se recomendar o emprego exclusivo de somente um tipo de cauda superior, é aconselhável utilizar um grande conjunto de distribuições candidatas cujos pesos de suas caudas superiores se estendam por um amplo espectro.

Considerações semelhantes às anteriores se aplicam à cauda inferior: é necessário utilizar um conjunto razoável de distribuições candidatas cujos pesos de suas caudas inferiores se estendam por um amplo espectro. Entretanto, se o interesse do estudo encontra-se centrado em se prescrever a melhor aproximação da cauda superior, a forma da cauda inferior é irrelevante. Em alguns casos, conforme enfatizado no relatório *Estimating Probabilities of Extreme Floods, Methods and Recommended Research* do National Research Council (*NRC*, 1987), a presença de *outliers* baixos em uma dada amostra pode inclusive vir a comprometer a correta estimação das características da cauda superior.

Considerações semelhantes às do limite superior também se aplicam ao limite inferior. Contudo, diferentemente do limite superior, o inferior é, em geral, conhecido ou pode ser igualado a zero; algumas distribuições, como a Generalizada de Pareto, permitem com facilidade o ajuste do parâmetro de posição, quando se conhece ou se prescreve o limite inferior. Hosking & Wallis (1997) ressaltam, entretanto, que, em diversos casos, a prescrição de limite inferior nulo é inútil e que melhores resultados podem ser obtidos sem nenhuma prescrição *a priori*. Exemplificam afirmando que os totais anuais de precipitação em regiões úmidas, apesar de números positivos, são muito superiores a zero; para esse exemplo, uma distribuição de probabilidades realista deve ter um limite inferior muito maior do que zero.

As distribuições oriundas da teoria clássica de valores extremos (Gumbe1, 1958), como os modelos Gumbel, Fréchet e Weibull, são as únicas para as quais existem justificativas teóricas para seu emprego na modelação de valores máximos (ou mínimos) de dados empíricos. Por exemplo, o modelo de valores extremos do tipo I para máximos (EV1 ou Gumbel) é a distribuição assintótica do maior valor de uma seqüência ilimitada de variáveis aleatórias independentes e igualmente distribuidas (*iid*),

e possui uma cauda superior do tipo exponencial. Analogamente, a distribuição do tipo II para máximos (EV2 ou Fréchet) relaciona-se a variáveis *iid* com cauda superior do tipo polinomial, enquanto a distribuição do tipo III (EV3 ou Weibull) refere-se a variáveis *iid* que possuem um limite superior finito. Sob as premissas da teoria de valores extremos, por exemplo, a distribuição de probabilidades das vazões médias diárias máximas anuais de uma certa bacia hidrográfica depende da distribuição inicial única dos valores diários considerados independentes. A maior objeção ao uso das distribuições oriundas da teoria de valores extremos em hidrologia refere-se à premissa de variáveis iniciais *iid*, a qual muito dificilmente é satisfeita por variáveis hidrológicas ou hidrometeorológicas. A esse respeito, transcreve-se o seguinte comentário escrito por Perichi & Rodríguez-Iturbe (1985, p. 515) :

"*Presumir que duas vazões médias diárias, observadas digamos no dia 15 de maio e em 20 de Dezembro, são variáveis aleatórias identicamente distribuídas é uma clara violação da realidade hidrológica. Essa premissa 'regulariza' as distribuições históricas iniciais afirmando não só que elas são do mesmo tipo, mas também que elas possuem os mesmos parâmetros (e.g. média e variância) para qualquer dia do ano. Sob essa premissa, não se pode admitir o fato que se uma mesma vazão média diária foi observada em dois dias diferentes, é mais provável que aquele que possui a maior variância produzirá cheias maiores do que aquele de menor variância. A realidade hidrológica é que a combinação da média e da variância de um dado mês faz com que alguns meses do ano sejam mais suscetíveis à ocorrência de cheias do que outros.*"

Além dessas considerações, a seqüência de variáveis hidrológicas/hidrometeorológicas, amostradas em intervalos horários ou diários ao longo de um ano, apresenta correlação serial significativa e não pode ser considerada suficientemente grande em termos assintóticos.

O fato que variáveis hidrológicas e hidrometeorológicas dificilmente satisfazem as premissas da teoria clássica de valores extremos vem justificar o uso de distribuições não-extremais, tais como a Lognormal, na análise local de freqüência de eventos máximos anuais. Chow (1954) apresenta a seguinte justificativa para o emprego da distribuição Lognormal: os fatores causais de várias variáveis hidrológicas agem de

forma multiplicativa, ao invés de aditiva, e a soma dos logaritmos desses fatores, em consequência do *teorema central limite* da teoria de probabilidades, tende a ser normalmente distribuída. Stedinger et al. (1993) afirmam que algumas variáveis, como a diluição, por exemplo, podem resultar do produto de fatores causais. Entretanto, para o caso de enchentes ou precipitações máximas, a interpretação dessa ação multiplicativa não é evidente.

As objeções anteriores referem-se às justificativas teóricas inerentes à distribuição Lognormal, bem como às distribuições oriundas da teoria clássica de valores extremos, porém, não têm o objetivo de excluí-las do elenco de distribuições candidatas à modelação de variáveis hidrológicas e hidrometeorológicas. No contexto da análise de freqüência local de variáveis hidrológicas, elas devem ser consideradas candidatas como quaisquer outras distribuições e devem ser discriminadas de acordo com outros critérios, tais como suas *medidas de aderência* aos dados amostrais.

Com relação ao número de parâmetros desconhecidos de uma distribuição de probabilidades, Hosking & Wallis (1997) afirmam que as distribuições de dois parâmetros produzem estimativas precisas de quantis quando as características distributivas populacionais a elas se assemelham. Entretanto, quando isso não ocorre, podem-se produzir estimativas tendenciosas dos quantis. A busca de um modelo probabilístico mais geral e flexível levou as agências do governo norte-americano a preconizarem o uso da distribuição Log-Pearson do tipo III para a análise local de freqüência de cheias máximas anuais em projetos com participação federal. O modelo Log-Pearson III é uma distribuição de três parâmetros, resultante da transformação logarítmica de variáveis aleatórias distribuídas de acordo com Gama ou Pearson do tipo III. Embora os seus três parâmetros confiram flexibilidade de forma a essa distribuição, a sua estimação, com base exclusiva em dados locais, é uma fonte de controvérsias. Bobée (1975) reporta situações em que a simples alteração do método de inferência estatística faz com que o parâmetro de forma dessa distribuição passe de negativo a positivo, o que a torna limitada superiormente ou inferiormente de acordo com o sinal do parâmetro. São essas características indesejáveis da distribuição Log-Pearson do tipo III que levaram, por exemplo, Reich (1977) a argumentar contra a sua utilização na análise de freqüência local de vazões máximas anuais. No contexto da análise regional, Hosking & Wallis (1997) observam que, obedecido o preceito da *parcimônia*

*estatística*, recomenda-se o uso de distribuições de mais de dois parâmetros por produzirem estimativas menos tendenciosas dos quantis nas caudas superior e inferior. No contexto da análise local, entretanto, resta apenas o preceito da parcimônia de parâmetros na especificação da função de distribuição de probabilidades.

As considerações anteriores, revelando a inexistência de leis dedutivas para a seleção de uma distribuição de probabilidades ou de uma família de distribuições para a análise de freqüência de eventos hidrológicos máximos anuais, remetem o analista a critérios variados e de algum modo subjetivos, entre os quais aqueles relacionados à capacidade descritiva dos modelos propostos. Alguns especialistas utilizam, como um possível critério de escolha, a comparação entre o coeficiente de assimetria amostral e o valor de assimetria teórico esperado para uma determinada distribuição de probabilidade. Por exemplo, enquanto estimativas amostrais do coeficiente de assimetria amostral próximas de zero podem sugerir a distribuição Normal como candidata à modelação estatística, amostras com assimetrias próximas a 1,14 indicam a prescrição de uma distribuição de Gumbel. A utilização deste critério está sujeita à precisão da estimativa do coeficiente de assimetria, a qual cresce com o aumento do tamanho da amostra, e serve apenas como um indicador de ajuste, tornando necessário o emprego de outros critérios, tais como indicadores de aderência, para selecionar uma distribuição probabilidades apropriada.

Apesar de ser um procedimento subjetivo, o exame visual do ajuste entre as distribuições de probabilidades candidatas e os dados observados pode ser útil na seleção da distribuição de probabilidades apropriada. Para isto, os dados observados são ordenados de forma decrescente, para análise de máximos, e plotados em um papel de probabilidade específico para cada distribuição. A tendência linear dos pontos plotados em papel de probabilidade apropriado é um indício de que a amostra pode ter sido extraída daquela população; por exemplo, uma tendência linear em um papel de probabilidade normal é uma evidência que os dados amostrais possam ter sido sorteados de uma distribuição normal. No caso de distribuições de 3 parâmetros, o exame visual ainda pode ser realizado nos papéis de probabilidade mais comuns, tais como exponencial ou normal; entretanto, neste caso, serão observadas tendências curvilíneas e não mais lineares. Embora útil, o exame visual dos dados é adequado para amostras de grandes tamanhos, uma vez que amostras pequenas são muito mais sensíveis à presença

de erros amostrais ou de imprecisão na estimação da posição de plotagem, os quais podem tornar a análise visual pouco informativa ou, mesmo, pouco confiável.

Testes de hipótese de aderência, tais como o do Qui-Quadrado, o de Kolmogorov-Smirnov, o de Filliben e o de comparação de quocientes de momentos-L, entre outros, podem ser utilizados durante o processo de seleção da função de distribuição de probabilidades. As estatísticas destes testes objetivam encontrar evidências que possam aceitar ou rejeitar a plausibilidade da distribuição de probabilidades testada; apresenta-se no Anexo A3 uma descrição detalhada dos principais testes de aderência, a saber: Qui-Quadrado, Kolmogorov-Smirnov, Filliben e o de comparação de quocientes de momentos-L. Segundo Kite (1977), os dois primeiros são os testes de aderência mais empregados na prática da análise de freqüência local de variáveis hidrológicas, enquanto Stedinger et al. (1993) sugerem alguma preferência pelo teste de Filliben. A verificação de aderência por comparação de quocientes de momentos-L apresenta vantagens em relação a seus predecessores, as quais são intrínsecas à utilização de momentos não convencionais; por razões de adequação à seqüência lógica da presente dissertação, essas vantagens serão abordadas em outro item desse capítulo.

De qualquer modo, entre os vários testes de hipótese de aderência, não há um teste estatístico que seja suficientemente potente para as amostras de pequeno tamanho comumente encontradas em hidrologia. De fato, tais testes não são sensíveis aos valores amostrais das caudas superior e inferior, o que é particularmente verdadeiro para as séries históricas típicas de pequeno tamanho. No Brasil, por exemplo, estas séries variam de 20 a 60 anos de dados. Junte-se a esse fato a incerteza inevitável quanto às estimativas dos parâmetros e quantis das distribuições sob teste. Além de tudo isso, resta dizer que os testes de aderência não discriminam as distribuições, ou seja, as verificações devem ser usadas para cada distribuição separadamente e não se prestam a selecionar uma ou um grupo de distribuições, entre as candidatas, com base nos respectivos valores das estatísticas de teste. Desse modo, a escolha da distribuição continua a cargo do especialista envolvido na análise, o qual utiliza outros critérios subjetivos para fazê-lo.

As dificuldades mencionadas fazem com que a seleção judiciosa de uma certa função de distribuição de probabilidades seja uma tarefa de especialistas, que geralmente a desempenham à luz de um conjunto de *regras heurísticas* formuladas de acordo com o conhecimento acumulado ao longo de anos de experiência e estudo. A abordagem heurística limita os caminhos a seguir, selecionando aqueles considerados melhores e reduzindo uma tarefa complexa a um conjunto de operações de julgamento. Em resumo, cabe ao especialista utilizar uma combinação de critérios objetivos e subjetivos para selecionar a distribuição que melhor se ajusta aos dados amostrais. O alto grau de subjetividade envolvido no processo de escolha de uma distribuição de probabilidades pode levar a soluções diferenciadas, para um mesmo conjunto de dados, dependendo dos critérios utilizados pelo especialista envolvido na análise. Cabe lembrar, entretanto, que o pequeno tamanho das amostras, normalmente encontradas na prática, torna impossível comprovar que a distribuição de probabilidades selecionada é superior às outras candidatas e que, de fato, ela representa a verdadeira distribuição populacional.

### II.3.3 – Estimativa dos parâmetros das distribuições

Uma função de distribuição de probabilidades é definida por um certo número de parâmetros, os quais a descrevem integralmente. Sabendo-se que uma variável aleatória segue uma determinada distribuição de probabilidades, e sob a premissa de que se tem em mãos uma amostra aleatória simples das realizações da variável em questão, é necessário estimar os valores numéricos dos parâmetros característicos da distribuição a partir dos dados amostrais da variável aleatória. Uma vez que as estimativas dos parâmetros populacionais das distribuições de probabilidades estão relacionadas com a amostra de dados observados, os parâmetros estimados são também variáveis aleatórias com média, variância e distribuição de probabilidades próprias.

O estimador $\phi$ do parâmetro $\Phi$ da distribuição de probabilidades ajustada pode ou não apresentar as seguintes propriedades:

- **Não-enviesamento:** um estimador $\phi$ de uma parâmetro $\Phi$ é dito não-enviesado se $E[\phi]=\Phi$. O viés é definido como $E[\phi]-\Phi$. O fato de um

estimador ser não-enviesado não garante que ele seja igual ao parâmetro populacional da distribuição.

- **Consistência:** um estimador $\phi$ de um parâmetro $\Phi$ é dito consistente se a probabilidade de $\phi$ diferir de $\Phi$ por mais que um determinado valor constante $\varepsilon$, para valores pequenos de $\varepsilon$, aproxima-se de zero quando o tamanho da amostra tende a infinito.

- **Eficiência:** um estimador $\phi$ é dito ser o mais eficiente se ele for não-enviesado e se a sua variância for menor que a variância de qualquer outro estimador de $\Phi$. Define-se como eficiência relativa de $\phi_1$, em relação a $\phi_2$, a razão de suas respectivas variâncias.

- **Suficiência:** um estimador $\phi$ é dito suficiente se ele usa todas as informações relevantes a $\Phi$ presentes na amostra.

A qualidade do estimador depende de quanto a sua estimativa desvia-se do verdadeiro valor $\Phi$. Esse desvio pode ser decomposto em um viés e uma variabilidade. O viés, conforme descrito acima, representa o desvio sistemático positivo ou negativo, enquanto que a variabilidade diz respeito aos desvios aleatórios em relação ao valor populacional de $\Phi$. Essa variabilidade pode ser quantificada pela variância do estimador, simbolizada por var[$\phi$]. Outra medida que combine o viés e a variabilidade do estimador é dada pela raiz quadrada do erro quadrático médio (*REQM*) definida por

$$REQM(\Phi) = \sqrt{E(f-\Phi)^2} = \sqrt{[viés(f)]^2 + \mathrm{var}(f)} \qquad \text{(II.16)}$$

Para estimativas com base em amostras de tamanho *n*, o viés e a variância de $\Phi$ são assintoticamente proporcionais ao inverso de *n*. Consequentemente, REQM é inversamente proporcional a $n^{½}$. Como essas quantidades possuem as unidades do parâmetro a ser estimado, Hosking e Wallis (1997) sugerem as razões var($\phi$)/$\Phi$ e *REQM*($\phi$)/$\Phi$, respectivamente o viés e o *REQM* relativos, como medidas mais convenientes e representativas.

Padrões mais formais sobre as quatro propriedades dos estimadores, acima descritas, bem como a descrição de processos para determinar se um estimador possui

ou não estas propriedades, podem ser encontradas em Lindgren (1968), Freund (1962) e Mood et al. (1974).

Existem várias metodologias para a estimação dos parâmetros populacionais a partir dos dados amostrais. As mais empregadas são as seguintes:

- Método dos Momentos;
- Método do Máximo de Verossimilhança; e
- Métodos dos Momentos-L.

**II.3.3.1 – Método dos Momentos**

O método dos momentos é o mais simples e mais utilizado para estimação dos parâmetros de uma distribuição de probabilidades a partir dos dados amostrais. Ele consiste simplesmente em igualar os momentos amostrais aos populacionais, sendo que o resultado dessa operação produzirá os estimadores dos parâmetros da distribuição de probabilidades em questão.

Os momentos populacionais podem ser estimados por quantidades similares aos momentos populacionais, calculadas a partir de uma amostra de tamanho n. O estimador natural de μ é a média aritmética ou o momento amostral de 1ª ordem,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{II.17}$$

enquanto os momentos amostrais centrais de ordem *r*, dados por

$$m_r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^r}{n} \tag{II.18}$$

são estimadores enviesados dos momentos populacionais de mesma ordem. Entretanto, os momentos amostrais $m_r$ podem ser corrigidos para produzirem estimadores sem viés. Por exemplo, as seguintes quantidades são, respectivamente, os estimadores sem viés da variância e dos coeficientes de variação, assimetria e cutose:

$$\hat{s}^2 = s^2 = \frac{n}{n-1} m_2 \tag{II.19}$$

$$\hat{C}_v = \frac{s}{\bar{x}} \tag{II.20}$$

$$\hat{g} = g = \frac{n^2}{(n-1)(n-2)} \frac{m_3}{s^3} \qquad \text{(II.21)}$$

$$\hat{k} = k = \frac{n^2}{s^4 (n-2)(n-3)} \left[ \left( \frac{n+1}{n-1} \right) m_4 - 3 m_2^{\,2} \right] + 3 \qquad \text{(II.22)}$$

Como anteriormente assinalado, o método dos momentos consiste em igualar os momentos amostrais aos populacionais. Formalmente, sejam $\{X_1, X_2, X_3,...,X_n\}$ os dados observados constituintes de uma amostra aleatória simples retirada de uma população de uma variável aleatória distribuída conforme a densidade $f_x$ (x; $\Phi_1$, $\Phi_2$, ..., $\Phi_k$) de *k* parâmetros. Se $\mu_j$ e $m_j$ representam os momentos populacionais e amostrais, respectivamente, então o sistema de equações do método resume-se a:

$$\mu_j (\Phi_1, \Phi_2, ..., \Phi_k) = m_j \quad ; \quad j=1,2,..., k \qquad \text{(II.23)}$$

As soluções desse sistema de *k* equações e *k* incógnitas serão as estimativas dos parâmetros $\Phi_1$, $\Phi_2$, ..., $\Phi_k$ da distribuição de probabilidades, pelo método dos momentos. Remete-se o leitor a Rao e Hamed (2000) para uma detalhada exposição de resultados de estimativas pelo método dos momentos para as principais distribuições de probabilidade empregadas em hidrologia.

### II.3.3.2 – Método do Máximo de Verossimilhança

Para grandes amostras, esse método geralmente produz os melhores estimadores dos parâmetros de uma distribuição de probabilidades. Consiste basicamente em maximizar uma função dos parâmetros da distribuição, conhecida como função de verossimilhança. O equacionamento para a condição de máximo resulta em um sistema de igual número de equações e incógnitas, cuja solução produz os estimadores de máxima verossimilhança.

Formalmente, sejam $\{X_1, X_2, X_3,...,X_n\}$ os dados constituintes de uma amostra aleatória simples retirada de uma população de uma variável aleatória distribuída conforme a densidade

$f_x$ (x; $\Phi_1$, $\Phi_2$, ..., $\Phi_k$) (II.24)

onde k é o número de parâmetros da distribuição de probabilidades e $\Phi_1$, $\Phi_2$, ..., $\Phi_k$ representam os parâmetros da distribuição de probabilidades. Como os valores observados $X_1$, $X_2$, $X_3$,...,$X_n$ são independentes entre si, eles possuem uma distribuição de probabilidade conjunta dada por

$$f_y\,(X_1,...,X_n;\Phi_1,...,\Phi_k) = f_x\,(X_1;\Phi_1,...,\Phi_k)\,f_x\,(X_2;\Phi_1,...,\Phi_k)\,...\,f_x\,(X_n;\Phi_1,...,\Phi_k) \quad \text{(II.25)}$$

a qual pode ser também representada da seguinte forma:

$$L(\Phi_1,\Phi_2,...,\Phi_k) = \prod_{i=1}^{n} f_X\left(X_i \middle| \Phi_1,\Phi_2,...,\Phi_k\right) \quad \text{(II.26)}$$

Este produto é chamado de função de verossimilhança e é proporcional à probabilidade de que a amostra tenha, de fato, sido sorteada de uma população distribuída conforme a equação II.24; $L(\Phi_1,\Phi_2,...,\Phi_k)$ é função dos parâmetros $\Phi_j$, exclusivamente. Maximizar a função de verossimilhança é o mesmo que maximizar a probabilidade de sorteio da amostra. A busca da condição de máximo para esta função resulta no seguinte sistema de k equações e k incógnitas:

$$\frac{\partial L\left(\Phi_1,\Phi_2,...,\Phi_k\right)}{\partial \Phi_j} = 0\,;\ \text{com j variando de 1 a k.} \quad \text{(II.27)}$$

As soluções desse sistema de equações, representadas por $\phi_1$, $\phi_2$,..., $\phi_k$, são os chamados estimadores de verossimilhança para a distribuição de probabilidade adotada. É comum empregar a função logaritmo de verossimilhança Ln[ L($\Phi$)], em substituição à função de verossimilhança propriamente dita, a fim de facilitar a construção do sistema de equações acima. Essa transformação justifica-se pelo fato da função logaritmo ser contínua, monótona e crescente; em outras palavras, maximizar o logaritmo da função é o mesmo que maximizar a função.

Em termos assintóticos, é possível demonstrar que os estimadores de máximo de verossimilhança não possuem viés e são superiores aos estimadores obtidos por meio do método dos momentos; em amostras finitas e pequenas, esta superioridade nem sempre é evidente. Uma possível dificuldade de aplicação do método de estimação pelo

máximo de verossimilhança refere-se à solução do sistema de equações II.27, o qual, às vezes, resulta ser não linear, com funções implícitas bastante complexas. Os livros de autoria de Clarke (1994) e de Rao e Hamed (2000) são excelentes referências para consulta e detalhamento do método do máximo de verossimilhança.

**II.3.3.3 – Método dos Momentos-L**

Os momentos-L são derivados dos "momentos ponderados por probabilidades", ou simplemente MPP's, os quais foram introduzidos na literatura científica por Greenwood et al. (1979). Os MPP's de uma variável aleatória X, descrita pela função de probabilidades acumulada $F_X(x)$, são quantidades definidas pela equação

$$M_{p,r,s} = E\left\{X^{p}\left[F_x(x)\right]^{r}\left[1-F_x(x)\right]^{s}\right\} \qquad \text{(II.28)}$$

e os MPP's $\alpha_r=M_{1,0,r}$ e $\beta_r=M_{1,r,0}$ representam casos especiais de relevância particular para a inferência estatística.

Diversos autores, como Landwehr et al. (1979) e Hosking & Wallis (1990), utilizaram os MPP's $\alpha_r$ e $\beta_r$ como base para a estimação de parâmetros de distribuições de probabilidades. Hosking & Wallis (1997) ponderaram, entretanto, que $\alpha_r$ e $\beta_r$ são de interpretação difícil, em termos das medidas de escala e forma de uma distribuição de probabilidades, e sugerem, alternativamente, certas combinações lineares de $\alpha_r$ e $\beta_r$, chamadas de Momentos-L. Detalhes formais sobre a teoria dos momentos-L podem ser encontradas no anexo-A4.

A utilização dos momentos-L para a estimação dos parâmetros das distribuições de probabilidades é análoga à dos momentos amostrais convencionais e consiste em se obter as estimativas dos parâmetros igualando-se os primeiros *k* momentos-L amostrais aos seus correspondentes populacionais. Esse procedimento irá resultar em um sistema de *k* equações e *k* incógnitas, cujas soluções serão as estimativas pelo método dos momentos-L.

A principal vantagem de se utilizar os momentos-L, em oposição ao uso dos momentos convencionais, está no fato de que as estimativas, resultantes dos momentos amostrais convencionais, além de apresentarem maior viés e maior variância para amostras de tamanho inferior a 100 (Vogel e Fennessey, 1993), envolvem potências

sucessivas dos desvios dos dados em relação ao valor central. Em conseqüência, amostras tipicamente pequenas, como as de variáveis hidrológicas, tendem a produzir estimativas não confiáveis, particularmente para as funções de momentos de ordem superior como assimetria e a curtose.

Hosking & Wallis (1997) mostram que os estimadores de parâmetros e quantis, obtidos por momentos-L para as distribuições mais comumentes utilizadas, são assintoticamente distribuídos como uma distribuição Normal, a partir da qual podem ser calculados erros padrões das estimativas e intervalos de confiança. Além disso, mostram que, para amostras de tamanho pequeno a moderado, o método dos momentos-L é geralmente mais eficiente do que o da máxima verossimilhança.

As vantagens potenciais do método de estimação pelos momentos-L têm resultado em um crescimento significativo do número e diversidade de suas aplicações em análise de freqüência de variáveis hidrológicas e hidrometeorológicas [e.g.: Muhara (2001), Davis e Naghettini (2001)]. Em outra vertente, os quocientes de momentos-L, descritos nos Anexos III e IV e representativos de quantidades padronizadas análogas aos coeficientes de variação, assimetria e curtose, têm encontrado aplicação na discriminação entre distribuições de probabilidade empregadas na análise de freqüência de variáveis hidrológicas. Em particular, Vogel e Fennessey (1993) advogam categoricamente o uso preferencial de diagramas de quocientes de momentos-L (ver Anexo III) para a discriminação entre várias hipóteses distributivas, enquanto Peel et al. (2001) sustentam sua utilização na análise de freqüência regional, em conjunto com testes de heterogeneidade. Zvi e Azmon (1997) aplicaram os diagramas de quocientes de momentos-L, em um primeiro estágio discriminatório, e o teste de aderência de Anderson-Darling, em um segundo estágio confirmatório, a 68 estações hidrométricas de Israel e concluíram que o uso conjunto de tais medidas, de fato, contribuiu para reduzir o grau de subjetividade na seleção da distribuição de probabilidades mais apropriada. Esses argumentos parecem ser suficientemente fortes para indicar o uso preferencial do método de estimação dos momentos-L na análise de freqüência, tanto local quanto regional, de variáveis hidrológicas e hidrometeorológicas.

### II.3.4 – Identificação e tratamento de *outlier*

Hawkins (1980) apresenta uma definição, algo intuitiva, de um *outlier*, ou ponto atípico, como uma observação que se desvia tanto dos outros dados observados que chega a despertar a suspeita de que não tenha sido gerada pelo mesmo fenômeno que determinou as outras ocorrências. Em seu trabalho, ele sugere a existência de três mecanismos geradores de *outliers*. Primeiramente, um *outlier* pode ser simplesmente um evento raro, produzido pela mesma população que deu origem aos outros dados. Um segundo mecanismo é o representado pela possibilidade de que o *outlier* tenha sido produzido por uma população diferente da que originou os outros dados. Finalmente, um *outlier* pode ser resultado de um erro de observação.

Em um guia para análise de vazões de projeto, preparado pela Institution of Engineers (1977), uma observação é identificada como *outlier* se ela localiza-se fora dos limites dos intervalos a 95% de confiança, em torno da distribuição de probabilidades ajustada. Quanto maior o afastamento de uma observação dos limites de confiança, maior é a evidência de que esta observação é um *outlier*. Uma vez detectada como atípica, a decisão de manter a observação depende do grau de certeza de que ela seja, de fato, um *outlier*, bem como da convicção sobre as razões de sua atipicidade; embora esta prescrição seja conceitualmente simples e lógica, a decisão de manter ou excluir a observação é de difícil execução prática. Uma descrição mais detalhada sobre esse tratamento específico da questão de *outliers* pode ser encontrada em Institution of Engineers (1977).

O United States Water Resources Council (USWRC, 1976) recomenda o uso da estatística modificada de Grubbs e Beck (1972) para a identificação de *outliers*. De acordo com esta proposta, os limites superior e inferior para a detecção de *outliers* altos e baixos, respectivamente, são dados pelas seguintes estatísticas:

$$X_{alto} = X_{\ln} + K_n S_{\ln} \tag{II.29}$$

$$X_{baixo} = X_{\ln} - K_n S_{\ln} \tag{II.30}$$

onde

$X_{\ln}$ representa a média dos logaritmos dos dados amostrais;
$S_{\ln}$ representa o desvio padrão dos logaritmos dos dados amostrais; e

$K_n$ denota a estatística modificada de Grubbs e Beck (1972).

Pilon et al. (1985) apresenta o seguinte estimador polinomial da estatística de Grubbs e Beck, para um nível de significância de 10%:

$$k_n = -3{,}62201 + 6{,}28446n^{\frac{1}{4}} - 2{,}49835n^{\frac{1}{2}} + 0{,}491436n^{\frac{3}{4}} - 0{,}37911n \qquad \text{(II.31)}$$

O United States Water Resources Council (USWRC, 1976) recomenda que os valores de vazão de pico, detectados como *outliers* altos, sejam comparados com as informações dos históricos de vazões obtidas em locais próximos ao estudado. Caso as informações históricas não sejam suficientes para permitir um ajuste da posição de plotagem do ponto atípico, a curva de freqüência final deverá ser obtida com a presença do *outlier*, sem qualquer alteração. Por outro lado, se houver informação histórica suficiente para ajustar a posição de plotagem do *outlier*, ela deverá ser convenientemente modificada antes de se proceder ao ajuste da curva de frequência. Quanto aos *outliers* baixos, e em se tratando de análise de freqüência de máximos, o USWRC recomenda a exclusão destes pontos da amostra.

Os dois métodos anteriormente descritos podem ser considerados como técnicas objetivas de identificação de *outliers* e apresentam alguns pontos importantes. O método que usa os intervalos a 95% de confiança, sugerido pelo Institution of Engineers (1977), pela própria maneira com que foi construído, depende da distribuição de probabilidades ajustada, ou seja, um determinado valor pode ser ou não considerado como *outlier* dependendo da distribuição de probabilidades ajustada. No caso da estatística modificada de Grubbs e Beck, conforme constatado por Chow (1988), o teste para identificação de *outliers* altos é satisfatório, sendo, contudo, ineficiente para a detecção de *outliers* baixos.

Além das técnicas objetivas descritas acima, critérios subjetivos podem ser empregados para a identificação de *outliers.* Por exemplo, pode-se considerar uma observação como um *outlier* se a sua retirada da amostra provoca uma alteração muito importante, ou até mesmo mudança de sinal, do valor de estimativa do coeficiente de assimetria amostral. Em qualquer circunstância, cabe lembrar a oportunidade e o pragmatismo da recomendação de Chow (1988), segundo a qual, caso não haja evidência clara de que o *outlier* detectado tenha sido produzido por uma população

diferente daquela que produziu os outros dados amostrais, ou mesmo que ele seja o resultado de um notório erro de medida ou observação, o ponto atípico deve ser tratado somente como um evento raro e não deve ser removido da amostra.

## II.4 - COMENTÁRIOS

Do exposto, é possível concluir que existem dois problemas fundamentais a serem solucionados durante a análise de freqüência de eventos hidrológicos máximos anuais. O primeiro, e talvez o que imprime o maior grau de incerteza e subjetividade à questão, refere-se à escolha da distribuição de probabilidades para modelar as observações amostrais. O segundo diz respeito à estimação dos parâmetros e quantis da distribuição escolhida, a qual envolve a aplicação de métodos estatísticos que venham produzir estimadores não-enviesados e eficientes dos parâmetros populacionais de uma dada distribuição que será, então, utilizada na modelação dos dados amostrais. Neste caso, escolhido o método para estimar os parâmetros de uma distribuição de probabilidades, os estimadores dependerão apenas dos registros da variável aleatória em questão. Neste ponto em particular, consideradas as ponderações do item II.3.3, podemos dizer que o método dos momentos-L configura-se como uma potente metodologia destinada à estimação de parâmetros e quantis de uma dada função de distribuição de probabilidades.

Por outro lado, a seleção de uma função particular de distribuição de probabilidades para modelar eventos hidrológicos máximos anuais é altamente subjetiva, em decorrência, principalmente, do fato que os testes de hipótese de aderência não são sensíveis a alterações da cauda superior e não estabelecem graus de prioridade entre as distribuições aprovadas. Freqüentemente, o emprego dos testes de hipótese de aderência conduz o analista apenas à rejeição de hipóteses distributivas obviamente rejeitáveis ou à constatação da não evidência de rejeição de um modelo paramétrico particular, restando ao especialista exercer suas preferências pessoais, consolidadas por sua experiência e conhecimentos específicos.

Uma tentativa pioneira para promover uma metodologia uniforme e consistente para a condução da análise de freqüência de vazões máximas anuais foi desenvolvida pelo United States Water Resources Council, em 1967, e apresentada no boletim de

número 15 intitulado “A Uniform Technique for Determining Flood Flow Frequencies” (USWRC, 1967). Neste boletim, o problema da seleção de uma distribuição de probabilidades para modelar os dados de vazões máximas foi contornado com a adoção do modelo paramétrico Log Pearson-III como padrão em todo território americano.

Em inúmeros outros países, bem como no Brasil, não há consenso quanto a se arbitrar um dado modelo distributivo como sendo mais apropriado para a análise de freqüência de eventos máximos anuais de variáveis hidrológicas e hidrometeorológicas. No caso brasileiro, uma pequena exceção é feita para a recomendação da Eletrobrás contida em suas “Diretrizes para Estudos e Projetos de Pequenas Centrais Hidrelétricas”, disponível pela URL http://www.eletrobras.gov.br. A recomendação, automatizada pelo programa computacional “QMAXIMAS”, disponível por meio de acesso à URL http://www.eletrobras.gov.br/downloads/programas/pch, consiste na seleção entre os modelos distributivos de Gumbel ou Exponencial, com base na maior ou menor proximidade do coeficiente de assimetria amostral aos populacionais 1,14 ou 2, respectivamente.

Em geral, sempre que um alto grau de subjetividade esteja presente em uma certa análise que envolva o conhecimento humano sobre determinado assunto, pode-se fazer uso das chamadas técnicas de inteligência artificial como ferramentas alternativas de padronização do problema e auxílio à decisão. O capítulo-III descreve brevemente as principais técnicas de inteligência artificial pertinentes ao problema em foco, algumas das quais foram utilizadas na implementação do sistema especialista de seleção de distribuições de probabilidades apropriadas à análise de freqüência de eventos hidrológicos máximos anuais.

# CAPÍTULO-III

## INTELIGÊNCIA ARTIFICIAL – SISTEMAS ESPECIALISTAS

O texto deste capítulo é, basicamente, uma compilação de vários artigos extraídos do portal da *Associação Americana de Inteligência Artificial*, URL http://www.aaai.org/Pathfinder/html/history.html#good, sendo que algumas das referências apresentadas aqui não foram diretamente consultadas.

### III.1 – INTRODUÇÃO

De acordo com diversos artigos do portal eletrônico da Associação Americana de Inteligência Artificial as correntes de pensamento que se formaram em torno da implementação da Inteligência Artificial (IA) já estavam em gestação desde os anos 30. No entanto, a expressão "Inteligência Artificial" teve sua primeira menção explícita apenas em 1956.

Ainda segundo artigos do portal eletrônico mencionado, durante todo o período de evolução das técnicas de IA duas principais linhas de pesquisa para a construção de sistemas inteligentes foram desenvolvidas: a linha conexionista e a linha simbólica. A linha conexionista, cujo objetivo é o de emular a inteligência humana através da simulação dos componentes do cérebro, ou seja de seus neurônios e de suas interligações, foi formalizada inicialmente em 1943, quando o neuropsicólogo McCulloch e o lógico Pitts propuseram um primeiro modelo matemático para um neurônio. Durante um longo período, essa linha de pesquisa não foi muito ativa, uma vez que envolve a resolução de um grande sistema de equações matemáticas complexas. Com a criação dos microprocessadores, pequenos e baratos, tornou-se praticável a implementação de máquinas de conexão compostas de milhares de microprocessadores, o que, aliado à solução de alguns problemas teóricos importantes, deu um novo impulso às pesquisas na área. O modelo conexionista deu origem à área de redes neuronais artificiais.

Por outro lado, a linha simbólica segue a tradição lógica e teve em McCarthy e Newell seus principais defensores. Sua história pode ser dividida em três épocas

distintas: clássica, romântica e moderna. Durante a fase clássica, que perdurou de 1956 a 1970, a pesquisa em manipulação de símbolos se concentrou no desenvolvimento de formalismos gerais capazes de resolver qualquer tipo de problema. Estes esforços iniciais ajudaram a estabelecer os fundamentos teóricos dos sistemas de símbolos e forneceram à área da IA uma série de técnicas de programação voltadas à manipulação simbólica, como, por exemplo, as técnicas de busca heurística. Os sistemas gerais desenvolvidos nesta época obtiveram resultados interessantes, por vezes até impressionantes, mas apenas em domínios simplificados, onde o objetivo era principalmente a demonstração da técnica utilizada e não a solução de um problema real.

Na fase romântica, que perdurou durante a década de setenta, a IA estava praticamente restrita ao ambiente acadêmico. Nesta época, os objetivos da pesquisa eram, principalmente, a construção de teorias e o desenvolvimento de programas que as verificassem para alguns poucos exemplos. A característica mais importante desta fase é uma maior exigência de formalização matemática, com critérios acadêmicos mais estritos de julgamento de trabalhos em IA. A partir de 1980, o programa computacional, em si, passou a ser a parte menos importante; a análise formal da metodologia, incluindo a capacidade de decisão, completude e complexidade, além de uma semântica bem fundada, passou a ser o ponto fundamental. O aparecimento dos primeiros Sistemas Especialistas (SE), nesta fase, possibilitou, através da tecnologia de IA, o desenvolvimento de sistemas com desempenho intelectual equivalente ao de um ser humano, abrindo, assim, perspectivas de aplicações comerciais e industriais.

Na época moderna, entre 1980 e 1990, a tecnologia dos SE’s disseminou-se rapidamente e foi responsável por mais um dos episódios ligados a promessas não cumpridas pela IA, ou seja, o sucesso dos primeiros SE’s chamou a atenção dos empresários que, então, iniciaram a busca por um produto comercializável que utilizasse esta tecnologia. No entanto, um SE não era um produto; um produto, na visão dos empresários, não deveria ser um sistema específico para um dado problema, mas algo que fosse implementado uma única vez e vendido em 100.000 unidades. Em decorrência desse fato, começaram a surgir ferramentas lógicas e computacionais para a construção de sistemas especialistas, ou seja as chamadas ‘*shells*’. Com isso, foi

colocada no mercado uma grande quantidade de *shells* que prometiam solucionar o problema de construção de SE's. A conseqüência foi uma grande insatisfação por parte dos usuários, pois, apesar de uma ferramenta de programação adequada ajudar muito a construir um sistema complexo, saber o que programar continuava sendo o ponto mais importante.

Se as *shells* deveriam ser vendidas como produtos de IA, então, em algum lugar, deveria haver IA; o lugar escolhido foi o chamado 'motor de inferência' que passou, assim, a ser considerado como sinônimo de IA. Isto levou à ilusão de que para se construir um SE, bastaria comprar uma *shell*. Entretanto, a verdade é que a IA em um SE encontra-se basicamente na forma como é representado o conhecimento sobre o domínio, ou seja, na tentativa de entender o comportamento inteligente a ser modelado ou, em outros termos, o comportamento do especialista humano ao empreender a resolução de um problema. Uma outra conseqüência desta visão distorcida das *shells* foi a pouca ênfase dada inicialmente à aquisição de conhecimento, certamente a parte mais difícil do desenvolvimento de um SE.

Entre os diversos benefícios associados ao desenvolvimento de SE's, podem ser citados: a distribuição de conhecimento especializado, a memória institucional, a flexibilidade no fornecimento de serviços (consultas médicas, jurídicas, técnicas, etc), a facilidade na operação de equipamentos, a maior confiabilidade de operação, a possibilidade de tratar situações a partir de conhecimentos incompletos ou incertos e o treinamento de pessoas não especializadas, entre outros.

Atualmente, as principais áreas de pesquisa em IA são: sistemas especialistas, aprendizagem, representação de conhecimento, aquisição de conhecimento, tratamento de informação imperfeita, visão computacional, robótica, controle inteligente, inteligência artificial distribuída, modelagem cognitiva, arquiteturas para sistemas inteligentes, linguagem natural e *interfaces* inteligentes. Além das linhas conexionista e simbólica, observa-se hoje o crescimento de uma nova linha de pesquisa em IA, baseada na observação de mecanismos evolutivos encontrados na natureza, tais como a auto-organização e o comportamento adaptativo. Nesta linha, os modelos mais conhecidos são os chamados algoritmos genéticos.

A gradativa mudança das metas da IA, desde o sonho de se construir uma inteligência artificial de caráter geral, comparável à do ser humano, até os bem mais modestos objetivos atuais de tornar os computadores mais úteis através de ferramentas que auxiliam as atividades intelectuais de seres humanos, coloca a IA como representativa de uma atividade que praticamente caracteriza a espécie humana, ou seja a capacidade de utilizar representações externas, seja na forma de linguagem, seja através de outros meios.

## III.2 – SISTEMA ESPECIALISTA

Sistema Especialista (SE) é um programa de computador projetado e desenvolvido para atender a uma aplicação determinada do conhecimento humano. Ele é capaz de tomar uma decisão, apoiado em conhecimento justificado, a partir de uma base de informações, tal como o faria um especialista humano de uma área específica do conhecimento. Essa base de informações pode ser construída a partir da reunião de conhecimentos de profissionais experientes e compor uma lógica para a tomada de decisão em condições subjetivas.

Durante a construção de um SE, a questão fundamental é como representar o conhecimento humano dentro do computador. Pode-se dizer que um bom método de Representação do Conhecimento (RC) mantém o significado semântico da informação implementada. No campo dos SE, a representação do conhecimento implica em encontrar uma forma sistemática de codificar a experiência de um especialista na área de domínio desejada. Contudo, é um engano supor que essa representação seja igual à codificação do conhecimento em alguma linguagem de programação; ela implica em organização do conhecimento, tornando-o acessível e de fácil aplicação via mecanismos aproximadamente naturais.

Winston (1992) define a RC como *"... o conjunto de convenções sintáticas e semânticas que tornam possível a descrição de coisas"*. Em IA, "coisas" normalmente significam o estado de algum problema; por exemplo, os objetos avaliados, suas propriedades, e qualquer relacionamento existente entre eles.

Basicamente, os métodos de RC's possuem duas características importantes: (i) força expressiva para descrição do conhecimento e (ii) procedimento computacional

para inferir o conhecimento codificado. Tais características são importantes durante o processo de seleção de um método de RC apropriado para modelar um problema específico. Os métodos de RC's com grande força de expressão necessitam, geralmente, de um procedimento de raciocínio complicado durante o processo inferência, aumentando, assim, as dificuldades computacionais do sistema. Já os métodos com menos força expressiva podem ser manipulados mais facilmente.

A determinação do tipo de método de RC a ser utilizado depende, principalmente, da natureza do conhecimento envolvido e do tipo de processo de raciocínio requerido para resolver esse problema, tornando necessária a distinção entre as classes do conhecimento e dos tipos de procedimentos de raciocínio empregados.

### III.2.1 – Classificação do conhecimento

Usualmente o conhecimento é definido como uma acumulação de fatos e idéias sobre o universo, podendo envolver uma relação causa-efeito; por exemplo, é conhecimento saber que a transformação logarítmica reduz o valor da assimetria de amostras assimétricas positivas e aumenta o valor da assimetria de amostras assimétricas negativas. Ele pode responder também a um determinado fato; por exemplo, é um fato saber que a vazão máxima anual é geralmente usada no processo de análise de freqüência. Barr e Feigenbaum (1981) classificam o conhecimento nos seguintes tipos:

- Objetos – referem-se ao conhecimento sobre o mundo, fatos ou idéias. Exemplos: saber que o coeficiente de assimetria da distribuição Gumbel é 1,14 e que o coeficiente de assimetria da distribuição normal é nulo;

- Eventos – referem-se ao conhecimento de uma seqüência de fatos, indexados no tempo, ou uma reação causa-efeito de uma ação. Exemplos: saber que o rio São Francisco inundou a cidade de Pirapora em 1979 e que uma determinada estação fluviométrica foi relocada no ano de 1981;

- Modo de Execução – refere-se ao conhecimento de como realizar uma tarefa. Exemplos: saber como escrever um relatório ou conduzir uma análise de freqüência;

- Meta-conhecimento – refere-se ao conhecimento geral sobre o assunto abordado. Exemplo: sabe-se que uma amostra submetida à análise de freqüência de vazões deve satisfazer às premissas de aleatoriedade, independência serial e homogeneidade.

### III.2.2 – Utilização do conhecimento

A força do conhecimento não é obvia até que ele seja utilizado para resolver alguma tarefa; são necessidades a compreensão da linguagem natural, o reconhecimento e a distinção de objetos. Neste sentido, o conhecimento em si é estático, enquanto que o processo de manipulação do conhecimento é dinâmico. Por exemplo, digamos que o sistema saiba que a assimetria da distribuição Normal é zero (conhecimento estático) e a seguinte pergunta é feita ao sistema: 'uma amostra com assimetria igual a zero pode vir de uma distribuição Normal?' e o sistema responde que 'sim'. A forma de manipulação do conhecimento para obtenção dessa resposta é classificada como conhecimento dinâmico.

O processo exato de raciocínio para alcançar esta conclusão em um cérebro humano é complexo e não é compreendido por completo. Representar este processo de raciocínio em um computador é um problema da IA. Abaixo, encontram-se listados os tipos de raciocínios estudados em IA:

- Raciocínio formal – envolve a manipulação sintática das estruturas de dados para deduzir novas verdades, conforme um conjunto predefinido de regras de inferência (lógica matemática);

- Raciocínio procedimental – envolve o estabelecimento de procedimentos para resolver problemas, como, por exemplo, os procedimentos para calcular a média, a variância e a assimetria amostral;

- Raciocínio monotônico – é um processo de raciocínio que produz novos conhecimentos a partir de velhos conhecimentos e descobre novas verdades, com base em verdades já conhecidas (e.g.: prova de teoremas);

- Raciocínio aproximado – envolve procedimentos para combinar informações com certo grau de incerteza para a tomada de decisões, usando medidas quantitativas e qualitativas (e.g.: diagnóstico médico).

### III.2.3 – Arquitetura de um Sistema Especialista

Geralmente os SE's possuem três componentes importantes: a base de regras, a memória de trabalho e um motor de inferência. A base de regras e a memória de trabalho formam a chamada base de conhecimento do SE, lugar onde estão armazenadas as informações do sistema, ou sejam os fatos e as regras pertencentes a uma determinada área de domínio. O motor de inferência é o mecanismo de controle do sistema que avalia e aplica as regras de acordo com as informações da memória de trabalho.

A chave para o desempenho de um SE está no conhecimento armazenado em suas regras e em sua memória de trabalho. Este conhecimento deve ser obtido junto a um especialista humano do domínio e representado de acordo com regras formais ou o formalismo definido para a codificação de regras no SE em questão. Isto divide um SE em duas partes: a ferramenta de programação que define o formato do conhecimento da memória de trabalho e das regras, além dos aspectos operacionais de sua utilização, e o conhecimento do domínio propriamente dito.

Atualmente, devido a esta separação, os SE's são desenvolvidos em geral a partir de *shells*; como anteriormente mencionado, as *shells* são ferramentas que suportam todas as funcionalidades de um SE, restando ao programador apenas codificar o conhecimento especializado de acordo com a linguagem de representação do conhecimento disponível. A existência das *shells* facilitou bastante a implementação de SE's e foi um dos fatores responsáveis por sua disseminação.

### III.2.4 - Aquisição de conhecimento

A parte mais sensível no desenvolvimento de um SE é, certamente, a aquisição de conhecimento. Esta não pode limitar-se à adição de novos elementos à base de conhecimentos; é necessário integrar o novo conhecimento àquele já disponível, por meio da definição de relações entre os elementos que constituem o novo conhecimento e os elementos já armazenados na base. Dois tipos de mecanismos para a definição de tais relações foram propostos: ligar os elementos de conhecimento diretamente através de ponteiros ou reunir diversos elementos relacionados em grupos.

Outro ponto importante na aquisição de conhecimento é o tratamento de incoerências. Dependendo da forma como o novo conhecimento é adquirido, pode haver erros de aquisição. Estes erros podem resultar da própria natureza do conhecimento, como em dados obtidos através de sensores sujeitos a ruído, ou podem ser gerados pela interface humana existente entre o mundo real e o sistema de representação. Algumas técnicas foram desenvolvidas para evitar erros de aquisição, como a especificação de regras de aquisição, onde o tipo de conhecimento esperado é definido. Estas técnicas são comuns aos sistemas de representação de conhecimento e aos sistemas de gerenciamento de bancos de dados. Por outro lado, uma base de conhecimento pode ser examinada periodicamente com a finalidade de detectar incoerências eventualmente introduzidas no processo de aquisição. Este método é limitado pelo fato de que linguagens de representação razoavelmente expressivas não contam com procedimentos conhecidos e completos de verificação. Finalmente, deve-se observar que a adequação do formalismo de representação ao tipo de conhecimento do mundo real a ser representado é fundamental para a eficiência do processo de aquisição.

### III.2.5 – Sistema Especialista x Programa Convencional

Um programa convencional é baseado em algum algoritmo, o qual, passo a passo, chega a uma resposta após um tempo específico de processamento. Ele é projetado para processar um volume de dados, de maneira repetitiva ou não, e terminar emitindo um resultado final ao término de sua execução.

Um Sistema Especialista é baseado em busca heurística e trabalha com problemas para os quais não existe uma solução convencional algoritmizada disponível ou, se existe, ela é demasiadamente demorada. Seu processo de busca normalmente conduz rapidamente à solução, podendo inclusive conduzir a uma solução distorcida dentro de determinadas situações, as quais são justificadas pelo sistema, ou mesmo não chegar a nenhuma conclusão.

Os SE's processam conhecimento e não dados, sendo o conhecimento armazenado dentro da base de conhecimento do sistema e os dados ajustados contra a base. Todo processamento é feito sobre o conhecimento, não existindo, portanto, o processamento de dados.

### III.2.6 – Sistemas Especialistas x Peritos Reais

Embora os SE's e Peritos Reais possam, em alguns casos, desempenhar tarefas idênticas, suas características intrínsecas podem divergir. Mesmo com algumas vantagens evidentes da aplicação dos SE's, eles não podem substituir os peritos em todas as situações, uma vez que apresentam limitações próprias. Uma delas é a ausência de criatividade; um perito real pode reorganizar informações e usá-las para sintetizar novos conhecimentos. Pode também manusear eventos inesperados usando a imaginação ou novas abordagens, inclusive o raciocínio por analogia baseado em outra área de domínio completamente diferente. Os SE's, por sua vez, trabalham sem inspiração rotineiramente e possuem um enfoque restrito à área do conhecimento implementada. Contudo, em situações comuns de uma determinada área do conhecimento, a opção pela utilização de um SE, geralmente apresenta alto grau de desempenho e agilidade nas análises. Além disso, os resultados encontrados são consistentes, imparciais e justificados tecnicamente, uma vez que o SE utiliza sempre as mesmas regras e ponderações na condução das análises subjetivas. Por outro lado, o custo da hora trabalhada de um Perito é geralmente alto comparado ao valor da hora trabalhada de um técnico não tão especializado, o que somente se justifica em situações atípicas.

## III.3 – REPRESENTAÇÃO DO CONHECIMENTO

A parte mais importante no projeto de um SE é a escolha do método de representação de conhecimento. A linguagem associada ao método escolhido deve ser suficientemente expressiva, porém não mais do que suficientemente, para permitir a representação do conhecimento a respeito do domínio escolhido de maneira completa e eficiente. Problemas de eficiência, facilidade de uso e a necessidade de expressar conhecimento incerto e incompleto levaram ao desenvolvimento de diversos tipos de formalismos de representação de conhecimento; descrições detalhadas sobre os vários métodos de representação do conhecimento podem ser encontradas em Barr e Feigeinbaum (1981) e Kobsa (1984).

Segundo Chow (1988), se o conhecimento é composto por regras heurísticas, ele pode ser expresso como uma associação empírica entre premissas e conclusões de fatos. Neste caso, as regras de produção são apropriadas para representar o conhecimento abordado. Caso o conhecimento envolva conceitos sutis, os quais necessitem de uma estrutura abstrata de representação e a implementação de componentes representativos, as redes semânticas ou quadros podem ser utilizados.

A seguir estão apresentados alguns dos formalismos de representação de conhecimento mais utilizados.

### III.3.1 – Regras de produção

As regras de produção, par condição-ação, foram inicialmente propostas por Post (1943) e introduzidas nas pesquisas de AI por Newell e Simon (1972). Este tipo de método de representação é usualmente referido como uma superfície de representação do conhecimento, uma vez que esse esquema é orientado mais de forma sintática do que semântica; seu objetivo não está concentrado no significado semântico do conhecimento. Os processos de inferência nesse tipo de formalismo podem ser facilmente mecanizados e automatizados conforme o seguinte exemplo de regra de produção:

Se "a assimetria amostral é nula"
Então "a amostra é normalmente distribuída"

A parte condicional da regra ("Se") deve ser satisfeita para colocá-la aplicável ao sistema, sendo a regra executada e a ação indicada na segunda parte da regra ("Então") concluída. Neste caso, dado que "a assimetria amostral é nula", o sistema concluirá que "a amostra é normalmente distribuída". Contudo, se a parcela correspondente à ação da regra for modificada ou se for incluída uma nova regra, conforme abaixo exemplificado, um novo resultado será obtido pelo sistema.

Se "a assimetria amostral é nula"
Então "a amostra pode ser descrita pela distribuícão Lognormal"

Agora, dado que "a assimetria amostral é zero", o sistema concluirá que "a amostra pode ser descrita pela distribuícão Lognormal".

Os sistemas baseados em regras de produção não são aptos a compreender o significado semântico de suas conclusões ou, mesmo, se elas estão corretas ou distorcidas. Entretanto, as regras de produção são um tipo de formalismo conveniente para expressar o conhecimento de especialistas em domínio específico; quando devidamente projetadas, representam bem o domínio dependente do conhecimento e podem ser adicionadas, excluídas ou modificadas facilmente. A maior vantagem deste formalismo está em sua facilidade de atualização da base de conhecimento.

### III.3.2 – Redes semânticas

Rede semântica é um nome utilizado para definir um conjunto heterogêneo de sistemas, onde a única característica comum a todos estes sistemas é a notação utilizada. Ela consiste em um conjunto de nós conectados por um conjunto de arcos. Os nós em geral representam objetos e os arcos representam as relações binárias entre esses objetos. Os nós podem também ser utilizados para representar predicados, classes,

palavras de uma linguagem, entre outras possíveis interpretações, dependendo do sistema de redes semânticas em questão.

A herança de propriedades através de caminhos formados por arcos é uma das características mais importantes do método, permitindo que as propriedades de um nó sejam especificadas apenas uma vez, sendo herdadas por todos os conceitos derivados, o que implica numa economia substancial de memória. Os algoritmos de herança utilizados em redes semânticas na forma de árvores são bastante simples e muito eficientes, mas se heranças múltiplas forem permitidas, especialmente na presença de arcos que definam exceções, o problema da determinação de caminhos de herança se torna bastante complexo. Mais grave ainda é o fato de que na presença de herança múltipla e exceções, a intuição sobre o que é uma política de herança coerente passa a ser discutível, levando diferentes sistemas a implementarem diferentes políticas de herança.

Além da herança de propriedades, um outro mecanismo de inferência utilizado em redes semânticas é a correspondência de um fragmento de rede em relação a uma rede dada. A especificação da semântica deste mecanismo é ainda mais complexa que a do mecanismo de herança, por depender da escolha do significado dos arcos da rede.

### III.3.3 – Quadros

Os quadros e sua variação, quais sejam os roteiros, foram introduzidos para permitir a expressão das estruturas internas dos objetos, mantendo a possibilidade de representar herança de propriedades como as redes semânticas.

Em geral, um quadro consiste em um conjunto de atributos que, através de seus valores, descrevem as características do objeto representado pelo quadro. Os valores atribuídos a estes atributos podem ser outros quadros, criando uma rede de dependências entre os quadros. Os quadros são também organizados em uma hierarquia de especialização, criando uma outra dimensão de dependência entre eles. Os atributos também apresentam propriedades, que dizem respeito ao tipo de valores e às restrições

de número que podem ser associados a cada atributo. Essas propriedades são chamadas de facetas.

Da mesma maneira que as redes semânticas, os sistemas baseados no método de quadros não constituem um conjunto homogêneo; no entanto, algumas idéias fundamentais são compartilhadas por estes sistemas. Uma dessas idéias é o conceito de herança de propriedades, o que permite a especificação de propriedades de uma classe de objetos através da declaração de que esta classe é uma subclasse de outra que goza da propriedade em questão.

Outra idéia comum aos sistemas baseados em quadros é o raciocínio guiado por expectativas. Um quadro contém atributos, os quais podem ter valores típicos ou valores "a priori", os chamados valores de exceção, em analogia à terminologia inglesa '*default values*'. Ao tentar estabelecer uma instância a um quadro para que corresponda a uma situação dada, o processo de raciocínio deve tentar preencher os valores dos atributos do quadro com as informações disponíveis na descrição da situação. Saber o que procurar para completar a informação necessária pode ser um fator fundamental na eficiência do processo de reconhecimento de uma situação complexa, dentro do processo de raciocínio.

## III.4 – MECANISMO DE INFERÊNCIA

O motor de inferência é a parte de um SE responsável pelo controle das atividades desenvolvidas pelo sistema, as quais ocorrem em ciclos divididos em três fases:

- Correspondência de dados, onde as regras que satisfazem a descrição da situação atual são selecionadas;
- Resolução de conflitos, onde as regras que serão realmente executadas são escolhidas dentre aquelas que foram selecionadas na primeira fase, sendo também ordenadas.
- Ação, a qual corresponde à execução propriamente dita das regras.

As principais características do motor de inferência disponível em uma *shell* dizem respeito às seguintes funcionalidades: modo de raciocínio, estratégia de busca, resolução de conflito e representação de incerteza.

### III.4.1 - Modo de raciocínio

Existem basicamente dois modos de raciocínio aplicáveis a regras de produção: encadeamento progressivo ou encadeamento a frente (do inglês, “forward chaining”) e encadeamento regressivo ou encadeamento para trás (do inglês, “backward chaining”). No encadeamento progressivo, também chamado de encadeamento dirigido por dados, a parte esquerda da regra é comparada com a descrição da situação atual, contida na memória de trabalho. As regras que satisfazem a esta descrição têm sua parte direita executada, o que, em geral, significa a introdução de novos fatos na memória de trabalho.

No encadeamento regressivo, também chamado de encadeamento dirigido por objetivos, o comportamento do sistema é controlado por uma lista de objetivos. Um objetivo pode ser satisfeito diretamente por um elemento da memória de trabalho ou podem existir regras que permitam inferir algum dos objetivos correntes, isto é, aqueles que contenham uma descrição deste objetivo em suas partes direitas. As regras que satisfazem esta condição têm as instâncias correspondentes às suas partes esquerdas adicionadas à lista de objetivos correntes. Caso uma dessas regras tenha todas as suas condições satisfeitas diretamente pela memória de trabalho, o objetivo em sua parte direita é também adicionado à memória de trabalho. Um objetivo que não possa ser satisfeito diretamente pela memória de trabalho, nem inferido através de uma regra, é abandonado. Quando o objetivo inicial é satisfeito, ou não há mais objetivos, o processamento termina.

Em geral, o tipo de encadeamento é definido de acordo com o tipo de problema a ser resolvido. Problemas de planejamento, projeto e classificação tipicamente utilizam encadeamento progressivo, enquanto problemas de diagnóstico, onde existem apenas algumas saídas possíveis, com um grande número de estados iniciais, utilizam encadeamento regressivo.

Em geral, as '*shells*' adotam apenas um modo de raciocínio; no entanto, existem alguns que permitem ambos os modos, mas de maneira independente, e ainda outros que permitem um encadeamento misto, onde os encadeamentos progressivo e regressivo se alternam de acordo com o desenvolvimento da solução do problema e com a disponibilidade de dados.

Uma característica importante do modo de raciocínio refere-se à existência ou inexistência do atributo da monotonicidade do método de inferência. Sistemas monotônicos não permitem a revisão de fatos, isto é, uma vez um fato declarado verdadeiro, ele não pode mais tornar-se falso. Sistemas não monotônicos, por outro lado, permitem a alteração dinâmica dos fatos. O preço desta capacidade é a necessidade de um mecanismo de revisão de crenças, pois uma vez que um fato, antes verdadeiro, torna-se falso, todas as conclusões baseadas neste fato também devem se tornar falsas.

### III.4.2 - Estratégia de busca

Uma vez definido o tipo de encadeamento, o motor de inferência necessita ainda de uma estratégia de busca para guiar a pesquisa na memória de trabalho e na base de regras. Este tipo de problema é conhecido como busca no espaço de estados. Este tópico foi um dos primeiros estudados em IA, no contexto de solução de problemas do tipo quebra-cabeças e jogos por computador, tais como os jogos de damas e xadrez.

### III.4.3 - Resolução de conflito

Ao terminar o processo de busca, o motor de inferência dispõe de um conjunto de regras que satisfazem à situação atual do problema, o chamado conjunto de conflito. Se esse conjunto for vazio, a execução é terminada; caso contrário, é necessário escolher quais regras serão realmente executadas e em que ordem. Os métodos de resolução de conflito mais utilizados ordenam as regras de acordo com os seguintes critérios: prioridades atribuídas estaticamente; características da estrutura das regras como complexidade, simplicidade e especificidade; características dos dados associados

às regras como o tempo decorrido desde sua obtenção, sua confiabilidade ou seu grau de importância; como última opção, tem-se também a seleção ao acaso.

Em geral, a utilização de um desses critérios é insuficiente para resolver os conflitos. Neste caso, a '*shell*' pode combinar mais de um método na forma de método primário, secundário etc. As melhores '*shells*' dispõem de diversos métodos de resolução de conflito e permitem ao usuário a especificação de quais métodos utilizar e em que ordem.

### III.4.4 – Representação da incerteza

No mundo real há situações onde as informações disponíveis apresentam algum grau de incerteza: imprecisão, conflito, ignorância parcial etc. A forma humana de pensar e raciocinar manipula constantemente estas informações durante o processo de resolução de problemas. Bonissone e Tong (1985) indicaram que a presença da incerteza em SE's pode ser associada a causas como: (i) incerteza relativa à realidade da informação, (ii) herança da imprecisão da linguagem de representação da regra, (iii) decorrência da inferência baseada em informações incompletas, e (iv) a agregação de diferentes fontes de conhecimentos ou especialistas.

O tratamento da incerteza é uma ativa área de pesquisa em SE's, pois os domínios adequados à implementação de SE's se caracterizam exatamente por não serem modelados por nenhuma teoria geral, o que implica em descrições incompletas, inexatas ou incertas. Diversos métodos foram propostos para tratar este problema; o método Bayesiano de inferência, o dos fatores de certeza, a teoria dos conjuntos difusos ou nebulosos, a teoria de probabilidades subjetivas e a teoria de possibilidades.

De maneira geral, estes métodos atribuem aos fatos e regras uma medida numérica que representa de alguma forma a "confiança" do especialista. Os métodos utilizados não são necessariamente coerentes uns com os outros e cada método adapta-se melhor a determinados tipos de problemas. Diversas '*shells*' dispõem de mais de um método de tratamento de incerteza, deixando ao usuário a escolha do mais adequado ao seu problema. Uma característica freqüente desses métodos é a existência de um limite

mínimo para a medida de incerteza, abaixo do qual o fato ou regra é desconsiderado. Este limite pode, em geral, ser fixado pelo usuário.

A seguir serão introduzidos alguns modelos numéricos mais conhecidos para a representação da incerteza, a saber: o das probabilidades subjetivas, o dos fatores de certeza e o da lógica difusa.

### III.4.4.1 – Probabilidades subjetivas

É comum a associação da definição de probabilidade ao conceito de freqüência relativa de ocorrências de um certo evento de atributos particulares. Este conceito ou interpretação da probabilidade, geralmente, não é adequado para descrever a incerteza presente em problemas do mundo real. Pode-se associar a uma certa assertiva ou a um certo padrão um valor de probabilidade para indicar o grau de incerteza sobre a veracidade ou não de tal assertiva, ou seja, o valor de probabilidade somente reflete a crença sobre o padrão e não com que freqüência este padrão revela-se verdadeiro. A probabilidade utilizada em SE's para descrever a incerteza, geralmente, é intuitiva e subjetiva.

Quando o método de probabilidade subjetiva é utilizado em sistemas baseados em regras de produção, a probabilidade subjetiva da regra corresponde à probabilidade condicional de ocorrência da parte "então" dado que a parte "se" da regra seja satisfeita.

### III.4.4.2 – Fatores de certeza

Shortliffe e Buchanan (1975) argumentaram que o grau de incerteza, atribuído pelo especialista ao padrão lógico da regra de produção, não poderia ser associado ao conceito de probabilidade para todos os problemas reais, uma vez que este valor é uma medida da convicção pessoal do especialista, refletindo o seu nível de aceitação do padrão lógico. Baseados na teoria da confirmação, eles propuseram duas novas medidas para representar a incerteza: medida de crença (MC[h,e]) e medida de descrença (MD[h,e]). A MC[h,e] é um número que varia entre 0 e 1 e mede o aumento da crença

de uma hipótese “h” baseada na evidência “e”. Similarmente, a MD[h,e] é um número que varia entre 0 e 1, e mede o aumento da descrença da hipótese “h” baseada na evidência “e”. A fim de representar o grau certeza que uma hipótese “h” ocorra dada uma evidência “e”, Shortliffe e Buchanan (1975) combinaram o MC[h,e] e o MD[h,e] em um número chamado fator de certeza (CF[h,e]), como:

$$CF[h,e] = MC[h,e] – MD[h,e]$$

O fator de certeza é um número que varia entre –1 e 1, sendo seus valores positivos indicativos da existência de mais razões para acreditar na hipótese “h” do que em desacreditar. Um valor negativo de CF[h,e] indica que há evidências a favor da negação da hipótese.

Shortliffe e Buchanan (1975) chegaram à conclusão de que o conceito intuitivo de CF[h,e] era o mais natural para a representação da incerteza, uma vez que freqüentes problemas eram encontrados pelos especialistas para expressar em termos quantitativos os valores da probabilidade de ocorrência da hipótese “h” (P[h]) e a probabilidade da hipótese “h” condicionada à evidência “e” (P[h|e]); o valor de CF[h,e] combina ambos os conhecimentos sobre P[h] e P[h|e].

### III.4.4.3 – Lógica difusa

A lógica difusa ou teoria dos conjuntos difusos foi inicialmente proposta por Zadeh (1965) como uma generalização do conceito da teoria clássica dos conjuntos. Na abordagem clássica, cada elemento do conjunto tem um valor de pertinência “$\mu_A$” que vale 0 e 1, indicando a sua pertinência ou não pertinência, respectivamente. Um conjunto difuso permite vários graus de pertinência, para os elementos do conjunto, que podem ser representados pela função de pertinência $\mu_A$, definida no intervalo $\mu_A$=[0,1].

O conceito de dualidade, estabelecendo que algo pode e deve coexistir com seu oposto, faz a lógica difusa parecer natural ou até mesmo inevitável. A lógica clássica trata com valores “verdade” das afirmações, classificando-as como verdadeiras ou falsas. Contudo, muitas das experiências humanas não podem ser consideradas

simplesmente como dicotômicas tais como verdadeiras ou falsas, sim ou não, preto ou branco.

Na lógica difusa, um conjunto Ã definido no universo de discurso *X* é caracterizado por uma função de pertinência $\mu_A$, a qual mapeia os elementos de *X* para o intervalo [0,1]. Desta forma, a função de pertinência $\mu_A(x)$ representa o grau de possibilidade que o elemento *x* venha a pertencer ao conjunto Ã, isto é, quanto é possível ao elemento *x* pertencer ao conjunto Ã.

Considere o seguinte padrão difuso: "a assimetria amostral tende a zero". Ninguém pode definir objetivamente um valor limiar a partir do qual a assimetria amostral possa ser considerada como zero. Intuitivamente, se a assimetria amostral for igual a 0,001, afirma-se, com alto grau de confiança, que a assimetria tende a zero e neste caso $\mu_A=1$. Mas se a assimetria estimada é igual a 0,3, este nível de confiança diminui, decrescendo à medida que o valor da estimativa se afasta de zero, $\mu_A<1$. Nesse exemplo, o padrão difuso pode ser representado pela seguinte função de pertinência:

$$m_A = \begin{cases} 0 \rightarrow g \leq -4 \\ \frac{1}{4}g + 1 \rightarrow -4 < g \leq 0 \\ -\frac{1}{4}g + 1 \rightarrow 0 < g \leq 4 \\ 0 \rightarrow g > 4 \end{cases}$$

## III.5 – COMENTÁRIOS

Os Sistemas Especialistas são concebidos para reproduzir o comportamento de especialistas humanos na resolução de problemas do mundo real, mas o domínio destes problemas é altamente restrito. Segundo um artigo de Feigenbaum (1990), disponível na URL http://www.kurzweilai.net/meme/frame.html?main=/articles/art0098.html, os primeiros Sistemas Especialistas que obtiveram sucesso em seu objetivo foram os sistemas DENDRAL (acrônimo para *'DENDRitic ALgorithm*) e MYCIN . O sistema DENDRAL é capaz de inferir a estrutura molecular de compostos desconhecidos a partir de dados espectrais de massa e de resposta magnética nuclear. O sistema MYCIN

auxilia médicos na escolha de uma terapia por antibióticos para pacientes com bacteremia, meningite e cistite infecciosa, em ambiente hospitalar.

Desde então, muitos projetos de Sistemas Especialistas foram desenvolvidos para resolver problemas em muitos domínios diferentes, incluindo: agricultura, química, sistemas de computadores, eletrônica, engenharia, geologia, gerenciamento de informações, direito, matemática, medicina, aplicações militares, física, controle de processos e tecnologia espacial.

Na área de análise de freqüência de eventos máximos anuais de variáveis hidrológicas, Chow (1988) propôs a utilização de um Sistema Especialista baseado na comparação dos valores dos coeficientes de assimetria amostrais e populacionais, bem como o coeficiente de correlação entre as posições de plotagem empíricas e teóricas, tal como no teste de Filliben, e na teoria dos conjuntos difusos para auxiliar na seleção da função de distribuição de probabilidades. No SE desenvolvido por Chow (1988), o mecanismo de inferência foi implementado dentro de um sistema de produção chamado FLOPS – *'Fuzzy Logic Production System'* e o universo de modelos paramétricos analisados pelo sistema compreende as distribuições Normal, Lognormal 2p, Gumbel, Generalizada de Valores Extremos (GEV) e Log-Pearson III.

De modo resumido, o sistema de raciocínio proposto por Chow (1988) contém as seguintes regras seqüenciais:

- Atribuir a cada uma das distribuições analisadas um nível de confiança preliminar de acordo com o valor numérico do critério de informação desenvolvido por Akaike (1974), o qual estabelece uma medida da parcimônia do modelo em análise;
- Verificar o grau de pertinência do coeficiente de assimetria amostral no conjunto difuso zero, o qual é modelado por uma função de pertinência em forma de sino, semelhante àquela da função densidade de probabilidades Normal. Caso o grau de pertinência for maior que 0,5, atribui-se à distribuição normal um nível de confiança igual ao grau de pertinência encontrado.

- Verificar o grau de pertinência do coeficiente de assimetria amostral dos logaritmos das observações no conjunto difuso zero, o qual é também modelado por uma função de pertinência do tipo sino. Se o grau de pertinência for maior que 0,5, atribui-se à distribuição Log-Normal 2p um nível de confiança igual ao grau de pertinência encontrado.
- Verificar o grau de pertinência do coeficiente de assimetria amostral no conjunto difuso 1,14, o qual é modelado por uma função de pertinência do tipo sino. Se o grau de pertinência for maior que 0,5, atribui-se à distribuição Gumbel um nível de confiança igual ao grau de pertinência encontrado.
- Verificar para cada distribuição analisada o grau de pertinência do coeficiente de correlação da posição de plotagem no conjunto difuso específico para cada distribuição, este modelado por uma função de pertinência crescente em forma de S, relativamente à função de distribuição de probabilidades analisada e ao tamanho da amostra. Se o grau de pertinência for maior que 0,5, atribui-se à distribuição analisada um nível de confiança igual ao grau de pertinência encontrado.
- Rejeitar a distribuição GEV como candidata, caso ocorra um dos seguintes fatos: (i) uma distribuição de dois parâmetros já tenha sido selecionada ou (ii) o limite máximo teórico da distribuição GEV, caso o coeficiente de forma seja positivo, seja significativamente menor que o valor máximo amostral.
- Rejeitar a distribuição Log-Pearson III como candidata, caso ocorra um dos seguintes fatos: (i) uma distribuição de dois parâmetros já tenha sido selecionada ou (ii) o limite máximo teórico da distribuição Log_Pearson III, caso a assimteria seja negativa, seja significativamente menor que o máximo amostral.
- Verificada a presença de *outlier* nos dados amostrais, a análise deverá ser refeita sem a presença do ponto atípico; o objetivo desta análise é verificar se a presença do *outlier* afetou a seleção de uma distribuição de probabilidades de dois parâmetros.

- Selecionar a distribuição de probabilidades que obtiver maior nível de significância.

Estas regras privilegiam a seleção de distribuições de probabilidades de dois parâmetros, uma vez que existem dois critérios diferentes para a atribuição do nível de confiança às distribuições de dois parâmetros, sendo o valor do nível de confiança adotado, para a distribuição analisada, o maior encontrado entre os dois critérios. As distribuições de três parâmetros analisadas possuem apenas um critério de seleção e a aceitação anterior de qualquer distribuição de probabilidades de dois parâmetros selecionada é um motivo para a sua rejeição.

Neste sistema, os estimadores das estatísticas descritivas e dos parâmetros das distribuições analisadas são calculados pelo método dos momentos convencionais. Em conseqüência, pequenas amostras tendem a produzir estimativas não confiáveis, particularmente para as funções de momentos de ordem superior a 2, tais como os coeficientes de assimetria e curtose. Tendo em vista as principais conclusões do Capítulo II da presente dissertação, relativas aos métodos de estimação mais usuais, é possível afirmar que o uso do método dos momentos convencionais pode conduzir à escolha errada da distribuição de probabilidades populacional. Além desses pontos, o sistema de raciocínio implementado por Chow (1988) permite que sejam aceitas distribuições limitadas na direção dos máximos, fato que pode causar um truncamento da variável hidrológica na direção do máximo, provocado por erros amostrais ou pelo tamanho reduzido da amostra de dados utilizada. A esse respeito, é conveniente considerar as discussões do item II.3.2 dessa dissertação.

O Capítulo IV, a seguir, propõe e descreve um protótipo de um sistema especialista, denominado SEAF (Sistema Especialista de Análise de Freqüência), para a seleção de uma distribuição de probabilidades para análise de freqüência local de eventos hidrológicos máximos anuais diários, o qual apresenta, por um lado, algumas similaridades com o desenvolvido por Chow (1988) e, por outro, acrescenta um sistema de raciocínio mais coerente com resultados recentes da pesquisa científica sobre variáveis hidrológicas.

Dentre as inovações implementadas no sistema SEAF podem ser destacadas:

- A ampliação do universo de modelos paramétricos analisados pelo sistema, compreendendo as distribuições Normal, Lognormal 2p, Gumbel, Exponencial, Pearson III, Log-Pearson III, Generalizada de Valores Extremos e Generalizada de Pareto;
- A utilização das estatísticas descritivas amostrais dos momentos-L para a estimação dos parâmetros das distribuições, uma vez que tais estimadores são menos tendenciosos do que aqueles obtidos pelo método dos momentos convencionais, conforme descrito no Capítulo II.
- A utilização de um mesmo conjunto de critérios para a classificação das distribuições analisadas, onde o nível de confiança atribuído a cada distribuição é a média dos valores encontrados em cada um dos critérios;
- A implementação de regras de produção fundamentadas nas estatísticas amostrais dos quocientes de momentos-L para a classificação das distribuições;
- A criação de um critério para a verificação da parcimônia entre a utilização de uma distribuição de três ou de dois parâmetros, pertencentes a mesma família;
- A exclusão das distribuições limitadas à direita;
- A criação de critérios restritivos as distribuições limitadas à esquerda.

# CAPÍTULO-IV

# SEAF – UM SISTEMA ESPECIALISTA PARA ANÁLISE DE FREQÜÊNCIA LOCAL DE EVENTOS HIDROLÓGICOS MÁXIMOS ANUAIS

## IV.1 – INTRODUÇÃO

Os argumentos desenvolvidos no Capítulo II da presente dissertação certamente permitem afirmar que a decisão sobre qual distribuição de probabilidade selecionar, para proceder à análise de freqüência local de eventos hidrológicos máximos anuais, é baseada na combinação de critérios objetivos e subjetivos, não havendo um procedimento padrão para cumprir tal tarefa. Diferentes especialistas podem selecionar diferentes distribuições de probabilidades para modelar a mesma amostra de dados, com base em diferentes conjuntos de regras. A teoria de probabilidades não oferece regras dedutivas para a escolha de uma distribuição particular ou de uma família de distribuições. Além disto, conforme demonstra a experiência de Chow (1988), relatada no Capítulo III, é uma tarefa muito complexa reunir regras heurísticas e, em seguida, escrever e implementar um algoritmo capaz de emular o processo de raciocínio adotado por um especialista durante a escolha da distribuição para a análise de freqüência local de eventos hidrológicos máximos anuais.

O protótipo do Sistema Especialista de Análise de Freqüência (SEAF), a ser descrito nesse capítulo, tem como objetivo implementar alguns critérios de seleção da distribuição de probabilidades, utilizando estatísticas amostrais baseadas nos momentos-L, a fim de padronizar o processo de escolha de uma distribuição de probabilidades na análise de freqüência local de eventos hidrológicos máximos anuais. Vale ressaltar que não é objetivo deste trabalho identificar a distribuição populacional dos dados, mas selecionar, entre um grupo de distribuições candidatas, aquela (ou aquelas) distribuição (distribuições) de probabilidades mais apropriada (s) para modelar os dados amostrais sob análise.

O conjunto de distribuições candidatas, do qual se extraem as possíveis escolhas do sistema, é formado pelas distribuições Normal, Log-Normal de 2 parâmetros ou Log-Normal 2p, Gumbel, Generalizada de Valores Extremos (GEV), Exponencial, Generalizada de Pareto, Pearson III e Log-Pearson III; com exceção da distribuição

Normal, que aqui é empregada como paradigma para algumas decisões, o conjunto de distribuições candidatas agrupa aquelas mais utilizadas em análise de freqüência local de eventos hidrológicos máximos anuais.

Antes de discutir as regras heurísticas, para a seleção de uma distribuição de probabilidades, entre aquelas que foram implementadas no Sistema Especialista SEAF, vale ressaltar que elas representam aproximações de um certo processo de raciocínio que foi adotado neste trabalho. Dado o grau de subjetividade, presente em tais regras e inerente à problemática em si, elas poderão ser modificadas e/ou revistas no futuro. Por essa razão, o programa SEAF é aqui caracterizado como um protótipo de um sistema especialista para a análise de freqüência local de eventos hidrológicos máximos anuais.

Em resumo, as regras heurísticas implementadas têm o objetivo de selecionar, entre as distribuições candidatas, aquela que venha a apresentar a maior confiança de que seja capaz de representar a distribuição populacional dos dados. Essas regras têm como linhas gerais as seguintes etapas seqüenciais:

- Análise comparativa entre os quocientes (ou razões) de momentos-L amostrais e teóricos, para cada uma das distribuições de probabilidades analisadas.
- Avaliação do coeficiente de correlação entre as posições de plotagem empíricas e teóricas, para cada uma das distribuições de probabilidades analisadas, por meio do teste de Filliben.
- Atribuição de um nível de confiança às distribuições, com base nas duas etapas anteriores.
- Busca de razões para se rejeitar alguma das distribuições anteriormente selecionadas.
- Verificação da parcimônia entre distribuições de uma mesma família.

Com um certo conjunto de regras heurísticas definido previamente, pode parecer natural e simples para um especialista selecionar uma distribuição de probabilidades apropriada. Entretanto, não é trivial codificar esse conjunto de regras heurísticas em um computador e usá-lo de forma sistemática. Em primeiro lugar, as regras heurísticas são formuladas com base em informação numérica, a qual está sujeita a erros de

amostragem. A esse respeito, são típicas algumas questões tais como: quão próxima de zero deve ser a assimetria para que possa ser considerada nula? ou quão elevado deve ser o coeficiente de correlação entre quantis empíricos e ajustados para que uma dada distribuição seja aceita de acordo com o teste de Filliben? Portanto, alguns critérios quantitativos são necessárias para que se possa expressar a informação numérica em termos qualitativos. Em segundo lugar, as regras heurísticas são difusas por si mesmas e podem ser alteradas a partir de novas informações. Por exemplo, suponha que se detecte um ponto atípico em um conjunto de observações e que o coeficiente de assimetria seja reduzido de 0,4 para 0,01 quando esse evento é removido da amostra. Nesse caso, talvez um certo especialista possa sugerir a seleção da distribuição Normal, mesmo sabendo que a assimetria dos dados originais seja maior do que, digamos, a distância de dois erros padrões de estimativa da assimetria nula da distribuição Normal. Pode-se ver, portanto, que embora a informação numérica seja essencial ao modo de inferência, as técnicas para interpretá-la e usá-la representam a parte mais crítica de um sistema especialista para a análise de freqüência local de variáveis hidrológicas. De fato, converter a informação numérica incerta ou incompleta em declarações linguísticas, implica associar a essas um certo grau de confiança. Assim, o sistema SEAF foi projetado e desenvolvido em dois módulos. O primeiro é constituído por um programa de computador que extrai toda a informação relevante de uma amostra e produz as medidas quantitativas necessárias. O segundo módulo, em interface com primeiro, interpreta as medidas quantitativas, associando-as a números difusos, de modo a convertê-las em declarações lingüísticas, associadas a um certo grau de confiança, e, em seguida, processar a seleção de uma ou mais distribuições de probabilidades.

O modo de inferência, ou seja, o segundo módulo do sistema, possui algumas características apropriadas ao problema em questão. As regras são agrupadas em blocos, os quais representam as diferentes etapas de um processo de raciocínio eminentemente indutivo. Cada bloco contém as regras que irão guiar a conversão da informação numérica em declarações linguísticas. Uma composição das declarações linguísticas obtidas em cada bloco permite que o sistema classifique as distribuições e apresente aquela que, segundo a base de conhecimento implementada, represente melhor a distribuição populacional dos dados. O sistema de raciocínio foi construído de forma monotônica difusa, ou seja, permite-se que o grau de confiança de uma certa distribuição nunca diminua e só aumente ao longo dos blocos do processo indutivo.

Dentre as ferramentas disponíveis para a construção de sistemas especialistas com números difusos, distingue-se o *software* FuzzyCLIPS, desenvolvido pelo *Integrated Reasoning Group do Institute for Information Technology do National Research Council* do Canadá (http://ai.iit.nrc.ca/IR_public/fuzzy/fuzzyClips), por ser de domínio público, bem como por suas características de portabilidade e fácil interfaceamento. O software FuzzyCLIPS é uma extensão do conhecido sistema de domínio público CLIPS - *C Language Integrated Production System*, descrito em Giarratano (1998), e acha-se em constante desenvolvimento por parte do *National Research Council* do Canadá. As considerações acima tornam o *software* FuzzyCLIPS uma escolha natural para o desenvolvimento de um sistema especialista capaz de representar e manipular fatos e regras de caráter difuso.

## IV.2 – INFORMAÇÕES NUMÉRICAS

O primeiro módulo do sistema SEAF, escrito em linguagem Delphi 4.0, extrai toda informação numérica de uma amostra e produz as medidas quantitativas necessárias ao processo de raciocínio, as quais podem ser sumarizadas a seguir:

- estatísticas descritivas amostrais;
- momentos-L e razões de momentos-L;
- testes não paramétricos;
- estimativas de parâmetros para cada uma das distribuições analisadas;
- intervalos bilaterais a 90% de confiança de $\tau_3$ (assimetria-L), para cada uma das distribuições de 2 parâmetros analisadas;
- intervalos bilaterais a 90% de confiança de $\tau_4$ (curtose-L), para cada uma das distribuições de 3 parâmetros analisadas,
- coeficiente de correlação do teste de Filliben e seu respectivo intervalo unilateral a 90% de confiança, para cada uma das distribuições analisadas,
- comparação dos limites amostrais e eventuais limites teóricos para cada uma das distribuições analisadas.

### IV.2.1 – Estatísticas Descritivas Amostrais

O primeiro passo do sistema é calcular as seguintes estatísticas descritivas amostrais: valor mínimo, valor médio, valor máximo, desvio padrão e coeficiente de assimetria convencional dos registros de dados. Em seguida, o programa recalcula estas mesmas estatísticas para os logaritmos dos dados amostrais.

### IV.2.2 – Momentos-L e Razões-L

Após o cálculo das estatísticas descritivas amostrais convencionais, o sistema estima os 4 primeiros momentos-L ($\lambda_1, \lambda_2, \lambda_3, \lambda_4$) e calcula as razões-L ($\tau_3$ e $\tau_4$), conforme descrito no item A4.4 do Anexo A4. Estes mesmos valores são também calculados para os logaritmos dos dados amostrais.

### IV.2.3 – Testes Não Paramétricos

Objetivando avaliar a adequação da amostra em relação às premissas de base da análise de freqüência, o sistema testa a presença de correlação serial, heterogeneidade e *outliers* entre os dados amostrais. A presença de correlação serial é verificada pelo teste do coeficiente de correlação de Kendall, enquanto a heterogeneidade é identificada através do teste de Mann-Kendall. Ambos os testes estão descritos em detalhes no Anexo A2.

A presença de *outliers* é verificada de duas formas distintas pelo sistema: a primeira, objetiva, é através da aplicação do teste de Grubbs e Beck aos dados amostrais, conforme descrito no item II.3.4 do Capítulo II. A segunda, subjetiva, é feita pela comparação entre a assimetria da amostra original e a assimetria da amostra sem um de seus extremos. Conforme Chow (1988), uma observação é identificada como *outlier* pelo sistema caso a sua retirada da amostra provoque uma alteração na assimetria maior que um desvio padrão das estimativas. Segundo a mesma referência, o desvio padrão de amostragem do coeficiente de assimetria pode ser aproximado por:

$$s = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \qquad \text{(IV.1)}$$

onde:

$\sigma$ é o desvio padrão da assimetria; e

*n* é o tamanho da amostra.

Sendo assim, se a retirada do máximo amostral alterar, em mais que um desvio padrão, a estimativa do coeficiente de assimetria convencional da amostra, o sistema identificará uma possível presença de um *outlier* alto. Da mesma forma, o sistema avalia a presença de *outlier* baixo com a retirada do mínimo amostral.

Vale ressaltar que os testes não paramétricos executados pelo sistema são apenas de caráter informativo e não afetam o processo de escolha e classificação das distribuições. O sistema apenas informa o usuário quanto à possível presença de correlação serial, heterogeneidade e *outliers*. A partir daí, cabe ao usuário decidir sobre a continuidade ou descontinuidade da análise em curso.

### IV.2.4 – Estimação dos Parâmetros das Distribuições

Nesta etapa da análise, o sistema estima os parâmetros das distribuições Normal, Log-Normal 2p, Gumbel, Exponencial, Pearson III, Log-Pearson III, Generalizada de Valores Extremos e Generalizada de Pareto pelo método dos momentos-L. Aqui, ressalta-se novamente que a escolha por este método de estimação deu-se pelas seguintes justificativas:

- As estimativas baseadas nos momentos amostrais convencionais envolvem potências sucessivas dos desvios dos dados em relação ao valor central. Em conseqüência, pequenas amostras ou aquelas com importantes erros de amostragem tendem a produzir estimativas não confiáveis, particularmente para as funções de momentos de ordem superior como assimetria e a curtose;

- Os estimadores de máxima verossimilhança são não-enviesados assintoticamente quando o tamanho da amostra tende a infinito, o que definitivamente não representa um argumento a favor de seu emprego em amostras de pequeno tamanho, como, de fato, são as amostras de eventos máximos anuais de variáveis hidrológicas. Além disso, algumas vezes a solução dos sistemas de equações que maximizam a função de

verossimilhança apresenta falta de convergência e grande complexidade numérica.

### IV.2.5 – Intervalo de Confiança de $t_3$ (assimetria-L)

Objetivando avaliar o coeficiente de assimetria-L amostral, para cada uma das distribuições de 2 parâmetros analisadas, o sistema calcula a estatística *H*, definida por

$$H = \frac{\left(t_3^{amostra} - \bar{t}_3\right)}{s_3} \qquad \text{(IV.2)}$$

onde:

$t_3^{amostra}$ é a assimetria-L dos dados amostrais;

$\bar{t}_3$ é o valor teórico populacional da assimetria-L; e

$s_3$ é o desvio padrão populacional da assimetria-L.

O valor de $\bar{t}_3$ é fixo para cada uma das distribuições de 2 parâmetros e independe do tamanho da amostra, enquanto $s_3$ é função da distribuição analisada e depende do tamanho da amostra. O valor de $s_3$ é calculado por simulação, realizada internamente pelo sistema, para as distribuições Normal, Log-Normal 2p, Gumbel e Exponencial.

O processo de cálculo de $s_3$, para cada uma das distribuições de 2 parâmetros analisadas, é descrito a seguir:

- São geradas pelo método de simulação de Monte Carlo, em função da distribuição de probabilidades e dos parâmetros estimados, 500 amostras de mesmo tamanho que a amostra analisada.
- Para cada uma das 500 amostras geradas é calculado o valor de $\tau_3$ amostral.
- O valor de $s_3$ é calculado a partir dos 500 valores de $\tau_3$ obtidos.

Supondo que se possa aproximar a distribuição dos valores de *H* à distribuição Normal Padrão, analogamente ao apresentado por Hosking e Wallis (1997) para a

estatística $Z = \frac{\left(t_4^{amostra} - \overline{t_4}\right)}{s_4}$ no contexto de análise de freqüência regional, o sistema calcula os intervalos de confiança bilaterais a 90%, para cada uma das distribuições de 2 parâmetros analisadas.

### IV.2.6 – Intervalo de Confiança de $t_4$ (curtose-L)

Objetivando avaliar a curtose-L amostral, para cada uma das distribuições de 3 parâmetros analisadas, o sistema calcula a estatística Z, definida por

$$Z = \frac{\left(t_4^{amostra} - \overline{t}_4\right)}{s_4} \qquad \text{(IV.3)}$$

onde:

$t_4^{amostra}$ é a curtose-L dos dados amostrais;

$\overline{t_4}$ é o valor teórico populacional da curtose-L; e

$s_4$ é o desvio padrão populacional da curtose-L.

Hosking e Wallis (1997) apresentam uma aproximação polinomial de $\overline{t}_4$, como função de $\tau_3$, para algumas distribuições de 3 parâmetros, a qual é utilizada pelo sistema para o calculo de $\overline{t}_4$ em função de $\tau_3$ amostral.

Em seguida, o sistema calcula o valor de $s_4$ para as distribuições Pearson III, Log-Pearson III, Generalizada de Valores Extremos e Generalizada de Pareto. A seguir são descritas as etapas do processo de cálculo de $s_4$ para cada uma das distribuições de 3 parâmetros analisadas:

- São geradas pelo método de simulação de Monte Carlo, em função da distribuição de probabilidades e dos parâmetros estimados, 500 amostras de mesmo tamanho que a amostra analisada.
- Para cada uma das 500 amostras geradas é calculado o valor de $\tau_4$ amostral.
- O valor de $s_4$ é calculado a partir dos 500 valores de $\tau_4$ obtidos.

Segundo Hosking e Wallis (1997), a estatística Z é distribuída aproximadamente como uma variável Normal Padrão. Com base neste resultado, o sistema calcula os intervalos de confiança bilaterais, a 90%, para cada uma das distribuições de 3 parâmetros analisadas.

### IV.2.7 – Intervalo de Confiança do Coeficiente de Correlação de Filliben

O sistema calcula o coeficiente de correlação de Filliben e o respectivo valor limiar de referência para cada uma das distribuições analisadas. O processo de cálculo do valor de referência pode ser descrito da seguinte forma:

- São geradas pelo método de simulação de Monte Carlo, em função da distribuição de probabilidades e dos parâmetros estimados, 500 amostras de mesmo tamanho que a amostra analisada.
- Para cada uma das 500 amostras geradas, são executadas as seguintes etapas:
  - Cálculo das estimativas dos seguintes momentos-L e razões de momentos-L: $\lambda_1$, $\lambda_2$, $\tau_3$ e $\tau_4$.
  - Ajuste da distribuição de probabilidades analisada aos momentos-L e razões de momentos-L para cada amostra gerada.
  - Cálculo do coeficiente de correlação de Filliben para cada amostra gerada.
- Os 500 valores do coeficiente de correlação de Filliben são ordenados de forma crescente e o valor de referência $R_{min}$ é definido como sendo aquele que é igualado ou superado 90% das vezes pelo conjunto dos coeficientes de correlação estimados.

Após determinados os valores de referência, o sistema calcula os intervalos de confiança unilaterais, a 90%, do coeficiente de correlação de Filliben para cada uma das distribuições analisadas.

### IV.2.8 – Comparação entre os Limites Mínimos Amostrais e Populacionais

Para cada uma das distribuições limitadas na direção do mínimo, o sistema compara o seu o limite mínimo com o menor valor amostrado. Caso o menor valor amostrado seja inferior a 90% do valor mínimo da distribuição ajustada, o módulo de cálculo envia uma mensagem ao módulo de interpretação informando que espaço amostral da distribuição ajustada não contempla o menor valor da amostra.

Além disso, o sistema verifica se a distribuição ajustada é limitada na direção dos máximos. Caso afirmativo, ele envia uma mensagem informando tal limitação.

## IV.3 – INTERPRETAÇÃO E ANÁLISE DAS INFORMAÇÕES NUMÉRICAS

O primeiro passo para a interpretação e posterior análise dos dados em um sistema especialista é transformar medidas quantitativas em declarações linguísticas. Para isto, o procedimento adotado foi criar padrões de comportamento para as estatísticas analisadas e compará-los aos seus correspondentes valores amostrais.

Basicamente, o sistema de raciocínio implementado utiliza dez estatísticas amostrais para atribuir níveis de confiança e, em seguida, classificar as distribuições. São elas: o coeficiente de assimetria-L, o coeficiente de curtose-L e as estimativas dos coeficientes de correlação de Filliben correspondentes às distribuições Normal, Lognormal 2p, Gumbel, Exponencial, Pearson III, Log-Pearson III, Generalizada de Valores Extremos (GEV) e Generalizada de Pareto (GPA).

Uma vez que existem erros de estimação das estatísticas amostrais, também existirão graus de incerteza intrínsecos nas declarações linguísticas fundamentadas nestas estatísticas. Além disto, os padrões de comportamento de tais medidas são descritos de forma apenas aproximada. Sendo assim, optou-se pela utilização de um sistema difuso de regras para modelar o processo de raciocínio. O *software* utilizado para implementação da base de conhecimento foi o FuzzyCLIPS, o qual é específico para a construção sistemas difusos de regras.

Os padrões de comportamentos das dez estatísticas analisadas são definidos com base em intervalos de confiança calculados para cada estatística e são modelados através

de variáveis linguísticas representadas por um ou vários conjuntos difusos. A seguir estão listadas as variáveis linguísticas utilizadas na análise.

- *"coeficiente de assimetria-L"*: representada através de quatro conjuntos difusos: *"Normal"*, *"Lognormal"*, *"Gumbel"* e *"Exponencial"*, os quais são formados pelas estimativas do coeficiente de assimetria-L, geradas respectivamente pelas distribuições Normal, Lognormal, Gumbel e Exponencial ajustadas;

- *"curtose-L"*: representada através de quatro conjuntos difusos: *"Pearson III"*, *"Log-Pearson III"*, *"GEV"* e *"GPA"*, os quais são formados pelas estimativas do coeficiente de curtose-L, geradas respectivamente pelas distribuições Pearson III, Log-Pearson III, Generalizada de Valores Extremos e Generalizada de Pareto ajustadas;

- *"Filliben-1"*: representada através do conjunto difuso *"Normal"*, o qual é formado pelos valores do coeficiente de correlação de Filliben gerados pela distribuição Normal ajustada.

- *"Filliben-2"*: representada através do conjunto difuso *"Lognormal"*, o qual é formado pelos valores do coeficiente de correlação de Filliben gerados pela distribuição Lognormal ajustada.

- *"Filliben-3"*: representada através do conjunto difuso *"Gumbel"*, o qual é formado pelos valores do coeficiente de correlação de Filliben gerados pela distribuição Gumbel ajustada.

- *"Filliben-4"*: representada através do conjunto difuso *"Exponencial"*, o qual é formado pelos valores do coeficiente de correlação de Filliben gerados pela distribuição Exponencial ajustada.

- *"Filliben-5"*: representada através do conjunto difuso *"Pearson III"*, o qual é formado pelos valores do coeficiente de correlação de Filliben gerados pela distribuição Pearson III ajustada.

- *"Filliben-6"*: representada através do conjunto difuso *"Log-Pearson III"*, o qual é formado pelos valores do coeficiente de correlação de Filliben gerados pela distribuição Log-Pearson III ajustada.

- *"Filliben-7"*: representada através do conjunto difuso *"GEV"*, o qual é formado pelos valores do coeficiente de correlação de Filliben gerados pela distribuição generalizada de valores extremos ajustada.

- *"Filliben-8"*: representada através do conjunto difuso *"GPA"*, o qual é formado pelos valores do coeficiente de correlação de Filliben gerados pela distribuição generalizada de Pareto ajustada.

Em termos formais um conjunto difuso $\tilde{A}$ é definido como:

$$\tilde{A} = \{[x, m_A(x)] \mid x \in X\} \tag{IV.4}$$

onde
$X$ é o universo onde os elementos $x$ estão definidos; e
$m_A(x)$ é a função de pertinência de $x$ em $\tilde{A}$.

É usual descrever as funções de pertinência utilizando uma relação padronizada de representação. No FuzzyCLIPS, elas podem assumir três formas distintas:

- forma *S*

$$\begin{aligned} S(x,a,c) &= 0, && x \le a, \quad x \in X \\ S(x,a,c) &= 2\left(\frac{x-a}{c-a}\right)^2, && a < x \le \frac{a+c}{2} \\ S(x,a,c) &= 1-2\left(\frac{c-x}{c-a}\right)^2, && \frac{a+c}{2} < x \le c \\ S(x,a,c) &= 1, && x > c, \quad x \in X \end{aligned} \tag{IV.5}$$

- forma *Z*

$$Z(x,a,c) = 1 - S(x,a,c) \tag{IV.6}$$

- forma Sino

$$\Pi(x,b,d) = S(x,b-d,b), \quad x \le b$$
$$\Pi(x,b,d) = Z(x,b,b+d) \quad x > b \qquad \text{(IV.7)}$$

As figuras IV.1, IV.2 e IV.3 ilustram as formas destas funções.

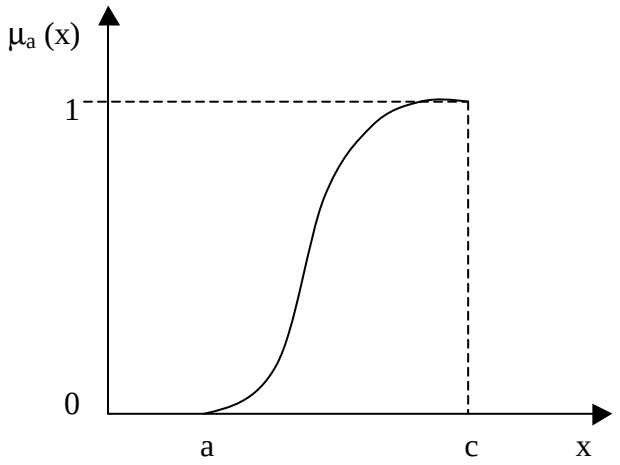



Figura IV.1: Função de pertinência do tipo *S*.

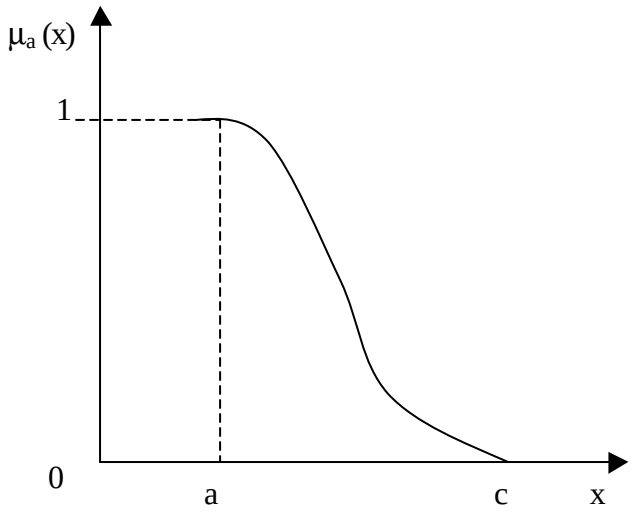



Figura IV.2: Função de pertinência do tipo *Z*.

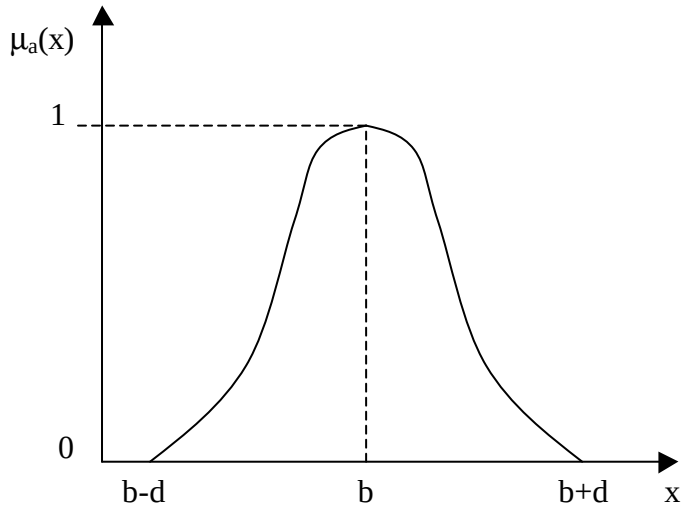



Figura IV.3: Função de pertinência do tipo Sino.

O módulo de interpretação do sistema define os vários conjuntos difusos representantes das variáveis linguísticas utilizadas para modelar o padrão de comportamento das estatísticas analisadas. Para isto, são definidos o universo *X* e a função de pertinência $m_A(x)$ de cada conjunto.

O sistema SEAF utiliza a forma *Sino,* ilustrada na Figura IV.3, para modelar os graus de pertinência dos elementos dos conjuntos difusos representativos das variáveis linguísticas *"coeficiente de assimetria-L"* e *"curtose-L"*; esta decisão é justificada pela expectativa, intuitivamente aceitável, de que as estimativas dos coeficientes de assimetria-L e curtose-L se distribuam simetricamente, com pequena dispersão, em torno de seu valor teórico populacional. No caso dos conjuntos difusos representativos das variáveis linguísticas *"Filliben-1"*, *"Filliben-2"*, ..., *"Filiben-8"* , o sistema utiliza a forma *S,* ilustrada na Figura IV.1, para modelar os graus de pertinência de seus elementos, uma vez que estimativas do coeficiente de correlação de Filliben próximas de 1 indicariam uma boa aderência ao modelo paramétrico em teste.

O universo *X* dos elementos de cada conjunto difuso é calculado em função dos limites dos intervalos de confiança correspondentes, a 90%, bem como da forma utilizada para modelar os graus de pertinência dos elementos do conjunto. Optou-se aqui por calcular os parâmetros da função de pertinência de cada conjunto difuso, pela

atribuição de um nível de confiança de no mínimo 0,5 aos elementos de *X* que, de fato, estejam inseridos nos limites de seu respectivo intervalo de confiança estatístico. A seguir são apresentados dois exemplos de como o sistema define os parâmetros das funções de pertinência.

Exemplo 1

Suponha que, para definir o conjunto difuso *"Normal"* da variável linguística *"coeficiente de assimetria-L"*, o sistema utilize o intervalo de confiança $-0{,}05 \leq t_3^{Normal} \leq 0{,}05$, construído com base nos valores de coeficientes de assimetria-L gerados pela distribuição Normal ajustada, com 90% de confiança estatística.

Os parâmetros da função de pertinência *Sino,* os quais definem o conjunto difuso *"Normal"*, são calculados da seguinte forma pelo sistema:

Primeiramente, o sistema cria os pares ordenados (-0,05; 0,5) e (0,05; 0,5), atribuindo aos limites do intervalo de confiança um grau de pertinência igual a 0,5.

Em seguida, ele substitui os pares ordenados na equação IV.7 criando um sistema de duas equações e duas incógnitas, conforme apresentado em IV.8

$$\begin{cases} 2\left(\dfrac{-0{,}05-(b-d)}{b-(b-d)}\right)^2 = 0{,}5 \\ 1-2\left(\dfrac{0{,}05-b}{b+d-b}\right)^2 = 0{,}5 \end{cases} \qquad \text{(IV.8)}$$

A solução do sistema determina os parâmetros da função de pertinência para o conjunto difuso *"Normal"* da variável linguística *"coeficiente de assimetria-L"*, a qual está representada de forma analítica pela equação IV.9 e graficamente na figura IV.4. Observa-se que todos os valores dentro dos limites do intervalo de confiança possuem graus de pertinência superior a 0,5.

$$\begin{cases} 2\left(\dfrac{t_3+0{,}1}{0{,}1}\right)^2, & -0{,}1 \le t_3 < -0{,}05 \\ 1-2\left(\dfrac{t_3}{0{,}1}\right)^2, & -0{,}05 \le t_3 \le 0{,}05 \\ 2\left(\dfrac{0{,}1-t_3}{0{,}1}\right)^2, & 0{,}05 < t_3 \le 0{,}1 \\ 0, & t_3 < -0{,}1 \quad e \quad t_3 > 0{,}1 \end{cases} \qquad \text{(IV.9)}$$

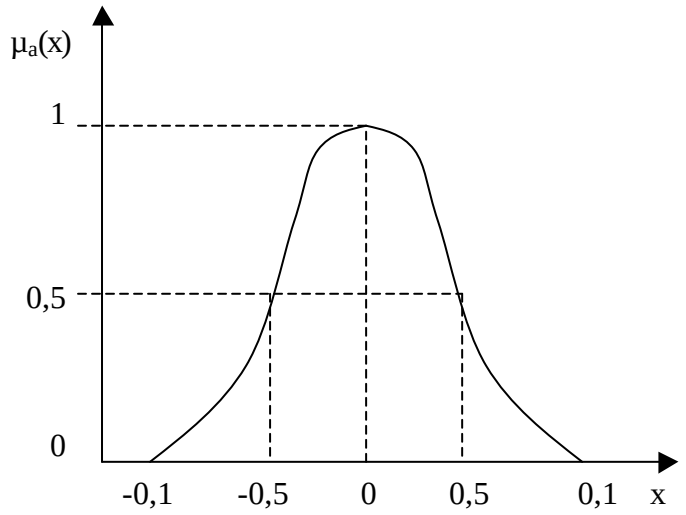



Figura IV.4: Função de pertinência do conjunto difuso *"Normal"* da variável linguística *"coeficiente de assimetria"*

Exemplo 2

Suponha que, para definir o conjunto difuso *"Normal"* da variável linguística *"Fillben-1"*, o sistema utilize o intervalo de confiança $0{,}98 \le R^{Normal} \le 1{,}00$, construído com base nos valores de coeficientes de correlação de Filliben gerados pela distribuição Normal ajustada, com 90% de confiança estatística.

Os parâmetros da função de pertinência *S*, os quais definem o conjunto difuso *"Normal"*, são calculados da seguinte forma pelo sistema:

Primeiramente, o sistema cria o par ordenado (0,98; 0,5), atribuindo ao limite inferior do intervalo de confiança um grau de pertinência igual a 0,5.

Em seguida, substitui o par ordenado na equação IV.5, obtendo:

$$2\left(\frac{0{,}98-a}{c-a}\right)^2 = 0{,}5 \qquad \text{(IV.10)}$$

Como o limite superior do conjunto difuso *"Normal"* é, por definição, igual a 1 ($c$=1), o valor do parâmetro $a$ é determinado pela resolução da equação IV.10. A solução desta equação determina os parâmetros da função de pertinência para o conjunto difuso *"Normal"* da variável linguística *"Filliben-1"*, a qual está representada de forma analítica pela equação IV.11 e graficamente pela Figura IV.5. Observa-se que todos os valores dentro dos limites do intervalo de confiança estatístico possuem graus de pertinência superior a 0,5.

$$\begin{cases} 0, & x \le 0{,}96 \\ 2\left(\dfrac{x-0{,}96}{0{,}04}\right)^2, & 0{,}96 < x \le 0{,}98 \\ 1-2\left(\dfrac{1-x}{0{,}04}\right)^2, & 0{,}98 < x \le 1{,}00 \\ 1, & x > 1{,}00 \end{cases} \qquad \text{(IV.11)}$$

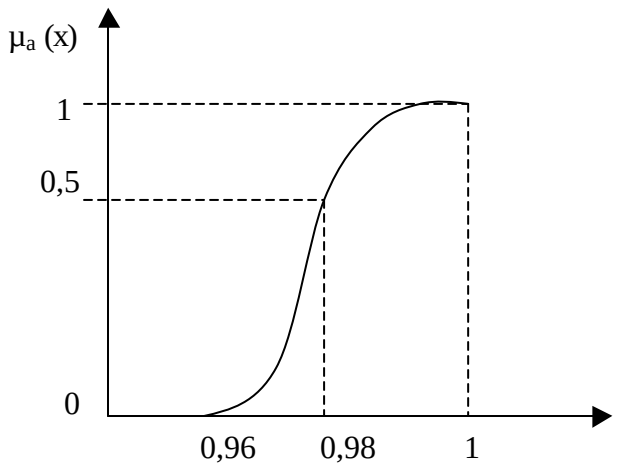



Figura IV.5: Função de pertinência do conjunto difuso *"Normal"* da variável linguística *"Filiben-1"*

Após serem determinados todos os conjuntos difusos pertencentes às variáveis linguísticas, o sistema inicia o processo de conversão das informações numéricas em declarações linguísticas. Nesta etapa, o sistema compara os valores das estatísticas amostrais analisadas com suas respectivas variáveis linguísticas, ou seja, ele verifica a pertinência do valor numérico da estatística em cada conjunto difuso pertencente à variável linguística. Para cada um dos conjuntos difusos, para os quais o grau de pertinência da estatística analisada for igual ou maior que 0,5, o sistema faz uma declaração linguística específica, associando um nível de confiança idêntico ao valor do grau de pertinência da estatística no conjunto. O exemplo 3, a seguir, ilustra como este processo é feito pelo sistema.

Exemplo 3

Suponha que uma amostra apresente um coeficiente de assimetria-L igual a 0,03 e um coeficiente de correlação de Filliben igual a 0,99, para um ajuste à distribuição Normal.

O sistema define o grau de pertinência do coeficiente de assimetria-L amostral em todos os conjuntos difusos que representem a variável linguística *"coeficiente de assimetria-L"*. No caso do conjunto difuso definido no Exemplo-1, equação IV.9, o grau de pertinência desta estatística ao conjunto difuso *"Normal"* é igual a 0,820. Como o grau de pertinência é maior que 0,5, o sistema faz a seguinte declaração: *"coeficiente de assimetria-L Normal, com 0,820 de nível de confiança"*. Isto significa que com base no coeficiente de assimetria-L, o sistema acredita que a amostra possa ter sido extraída de uma população Normal, com nível de confiança de 0,820.

A determinação do grau de pertinência do coeficiente de Filliben, para o ajuste à distribuição Normal, é feita pela verificação do grau de pertinência desta estatística no conjunto difuso *"Normal"* da variável linguística *"Filliben-1"*. No caso do conjunto difuso, definido no Exemplo-2, formalizado pela equação IV.11, o grau de pertinência desta estatística ao conjunto difuso *"Normal"* é igual a 0,875. Novamente, como o grau de pertinência é maior que 0,5, o sistema faz a seguinte declaração: *"coeficiente de correlação Filliben Normal, com 0,875 de nível de confiança"*. Isto significa que com base no teste de aderência de Filliben para a distribuição Normal, o sistema acredita que a amostra possa ter sido extraída de uma população Normal, com nível de confiança de 0,875.

Uma vez que existem 16 conjuntos difusos definidos, o sistema poderá fazer até 16 declarações linguísticas a partir das estatísticas amostrais, sendo quatro referentes aos coeficientes de assimetria, quatro aos de curtose e oito relativos aos coeficientes de correlação de Filliben. A seguir, estão listadas as possíveis declarações linguísticas obtidas pelo sistema em função das estatísticas amostrais.

- *"coeficiente de assimetria-L Normal, com XXX de nível de confiança".*
- *"coeficiente de assimetria-L Lognormal, com XXX de nível de confiança".*
- *"coeficiente de assimetria-L Gumbel, com XXX de nível de confiança".*
- *"coeficiente de assimetria-L Exponecial, com XXX de nível de confiança".*
- *"curtose-L Pearson III, com XXX de nível de confiança"*
- *"curtose-L Log-Pearson III, com XXX de nível de confiança"*
- *"curtose-L GEV, com XXX de nível de confiança"*
- *"curtose-L GPA, com XXX de nível de confiança"*
- *"coeficiente de correlação de Filliben Normal, com XXX de nível de confiança".*
- *"coeficiente de correlação de Filliben Lognormal, com XXX de nível de confiança".*
- *"coeficiente de correlação de Filliben Gumbel, com XXX de nível de confiança".*
- *"coeficiente de correlação de Filliben Exponencial, com XXX de nível de confiança".*
- *"coeficiente de correlação de Filliben Pearson III, com XXX de nível de confiança".*
- *"coeficiente de correlação de Filliben Log-Pearson III, com XXX de nível de confiança".*
- *"coeficiente de correlação de Filliben GEV, com XXX de nível de confiança".*
- *"coeficiente de correlação de Filliben GPA, com XXX de nível de confiança".*

As declarações linguísticas obtidas são, então, agrupadas de acordo com as distribuições de probabilidades correspondentes. Com base nestas declarações, o sistema atribui um nível de confiança a cada distribuição analisada. O nível de confiança é calculado como a média aritmética dos níveis de confiança de suas

declarações correspondentes. Conforme pode ser visto na listagem acima, o sistema pode apresentar até duas declarações linguísticas para cada distribuição analisada. No caso de existir apenas uma declaração linguística, para uma determinada distribuição, o sistema atribui a ela um nível de confiança igual a média entre o valor do nível de confiança de sua declaração e o valor 0,5. Isto permite a penalização da distribuição por não ter sido bem sucedida em um dos testes de aderência. Caso não haja nenhuma declaração linguística para uma distribuição, ela é descartada da análise. O Exemplo 4, a seguir, ilustra o funcionamento do sistema.

Exemplo 4

Suponha que o sistema faça as seguintes declarações linguísticas:

- *"coeficiente de assimetria-L Gumbel, com 0,900 de nível de confiança"*.
- *"curtose-L GEV, com 0,85 de nível de confiança"*
- *"coeficiente de correlação de Filliben GEV, com 0,780 de nível de confiança"*.

O sistema faz as seguintes declarações:

- Distribuição Normal, com 0,700 de nível de confiança
- Distribuição GEV, com 0,815 de nível de confiança

O próximo passo da análise é a verificação se há alguma razão para se rejeitar uma dada distribuição de probabilidades anteriormente classificada. Para isto, o sistema SEAF compara os limites de aplicação de cada distribuição ajustada com os limites mínimo e máximo amostrados e descarta a sua utilização quando:

- a distribuição for limitada na direção do máximo; e
- o mínimo amostrado for menor que 90% do valor do limite mínimo da distribuição ajustada.

Antes de apresentar as distribuições selecionadas, o sistema verifica o grau de parcimônia entre as distribuições de mesma família, através da comparação de seus respectivos níveis de confiança. Caso uma distribuição de 2 parâmetros tenha um nível de confiança maior do que o de uma distribuição de 3 parâmetros da mesma família, o sistema descarta a distribuição de 3 parâmetros, uma vez que a incerteza na estimação

de mais um parâmetro não contribuiu para melhorar o nível de confiança desta distribuição em relação à sua parente de 2 parâmetros.

As distribuições foram agrupas em quatro famílias, a saber: Família 1, composta pela Normal e Pearson III, Família 2, composta pela Log-Normal 2p e Log-Pearson III, Família 3, composta pela Gumbel e GEV, e Família 4, composta pela Exponencial e GPA.

Finalmente, o sistema apresenta as distribuições selecionadas com os seus respectivos níveis de confiança. O sistema SEAF recomenda o emprego da distribuição que apresentar o maior nível de confiança.

## IV.4 – PROGRAMA SEAF

O programa SEAF foi desenvolvido para trabalhar em ambiente Windows 95, 98 e NT. A arquitetura do sistema é dividida em duas partes: uma, responsável pela interface gráfica com o usuário e também pelos cálculos matemáticos, desenvolvida no aplicativo Delphi 4.0 Professional, da Borland. A outra, responsável por armazenar a base de conhecimento do sistema, bem como fazer as interpretações e análise, foi desenvolvida em linguagem FuzzyCLIPS, mantida e distribuída pelo *Integrated Reasoning Group do Institute for Information Technology do National Research Council* do Canadá. Os dois módulos constituintes do sistema são interligados entre si e foram projetados para serem acionados automaticamente pelo programa sem que seja necessária uma intervenção do usuário.

O sistema SEAF encontra-se disponível para *download* de um arquivo de instalação a partir da URL http://www.ehr.ufmg.br. O processo de instalação é automático e possibilita ao usuário apenas definir o diretório de instalação do programa.

A utilização do programa é simples, sendo necessário que o usuário crie um arquivo texto, em um aplicativo de sua preferência, contendo os registros de eventos máximos anuais da variável hidrológica em estudo para um dado local. O arquivo de registros deve conter apenas um valor amostrado por linha e a digitação deve respeitar a ordem cronológica dos registros. Deve ser usado o ponto como símbolo para identificação de decimal, tal como exemplificado na Figura-IV.6.


Figura-IV.6: Exemplo de um arquivo de entrada de dados

Para iniciar o programa o usuário deve clicar com o botão esquerdo do *mouse* na opção “Iniciar”, seguido de “Programas” e, finalmente, “SEAF”. A tela inicial do programa é, então, aberta e tem a forma geral apresentada na Figura IV.7.

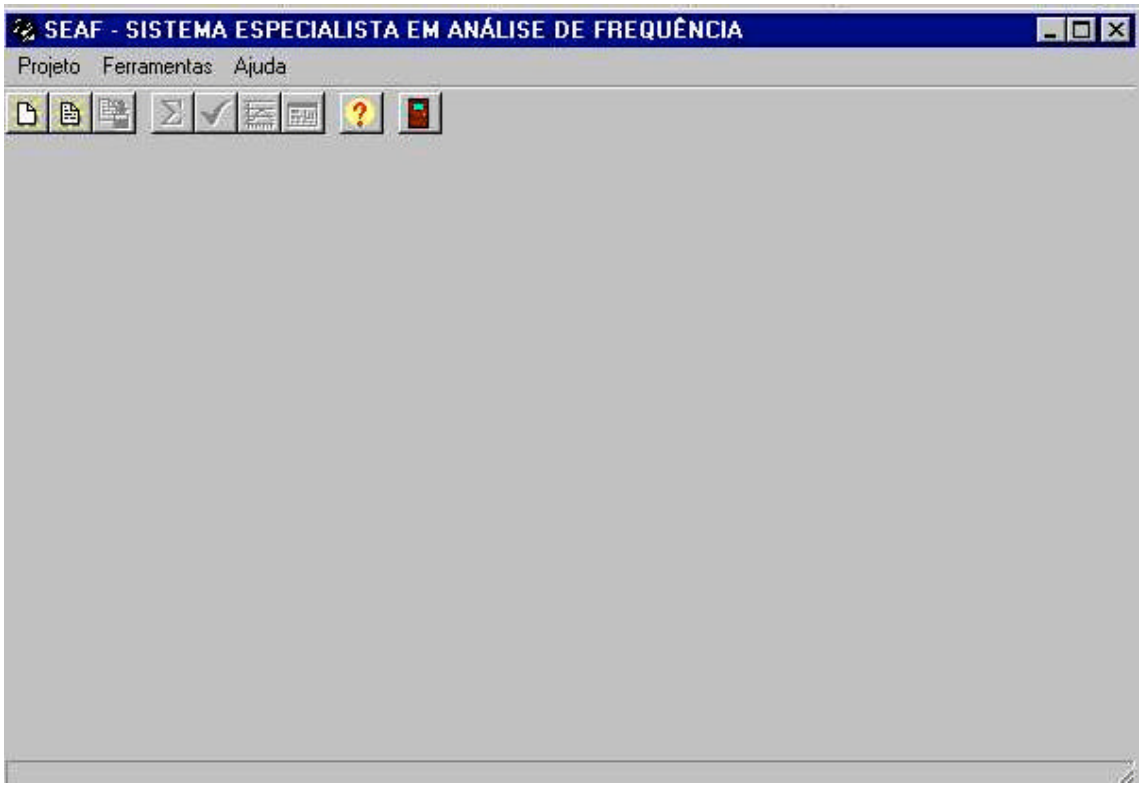



Figura-IV.7: Janela principal do programa SEAF

Após inicializado, o programa SEAF solicita ao usuário a decisão de optar por fazer uma nova análise ou visualizar uma análise anterior.

Para executar uma nova análise, o usuário deve escolher a opção “Projeto” , em seguida, “Novo”, ilustrada na Figura IV.8.


Figura-IV.8:Criando um novo projeto.

Em seguida, o sistema apresentará uma janela para que sejam fornecidos: um título para o projeto em curso, a identificação da estação de registros hidrológicos ou hidrometeorológicos em análise, um comentário e o nome do arquivo que contém os dados, tal como ilustrado na Figura IV.9. Note que para entrar com o nome do arquivo, o usuário deverá clicar no botão localizado ao lado da caixa de texto nomeada “Arquivo de dados”.

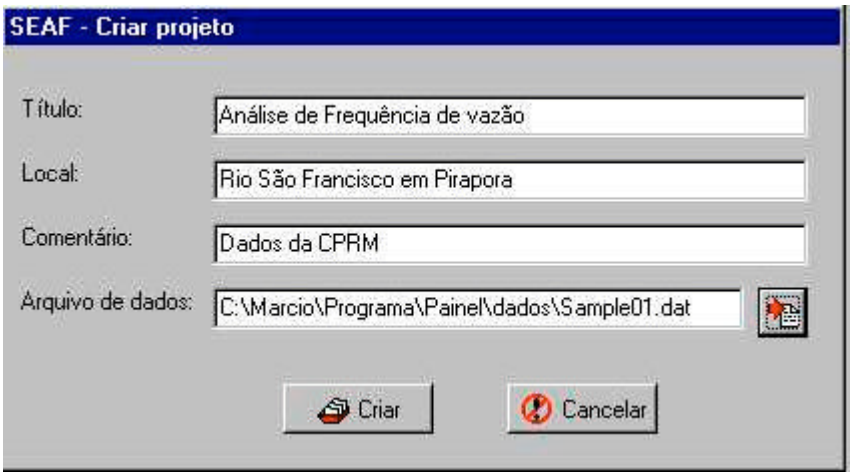



Figura-IV.9: Janela “Criar projeto”

Terminada a entrada dos dados, o usuário deverá clicar no botão “Criar” para que o sistema inicie o processo de análise. Isto poderá levar alguns minutos dependendo das características do computador utilizado.

Na seqüência, o sistema apresentará algumas estatísticas amostrais e informações auxiliares, tais como histogramas e cronologia dos eventos.

A primeira janela, mostrada na Figura IV.10, apresenta, de um lado, as estatísticas descritivas convencionais, tais como valores máximo, mínimo e médio, desvio-padrão e coeficiente de assimetria, tanto para as observações originais como para seus logaritmos, além dos dois primeiros momentos-L e os quocientes de momentos-L $t_3$ e $t_4$. Do outro lado, a janela permite uma primeira idéia da forma da distribuição empírica, por meio da visualização do histograma dos dados amostrais.

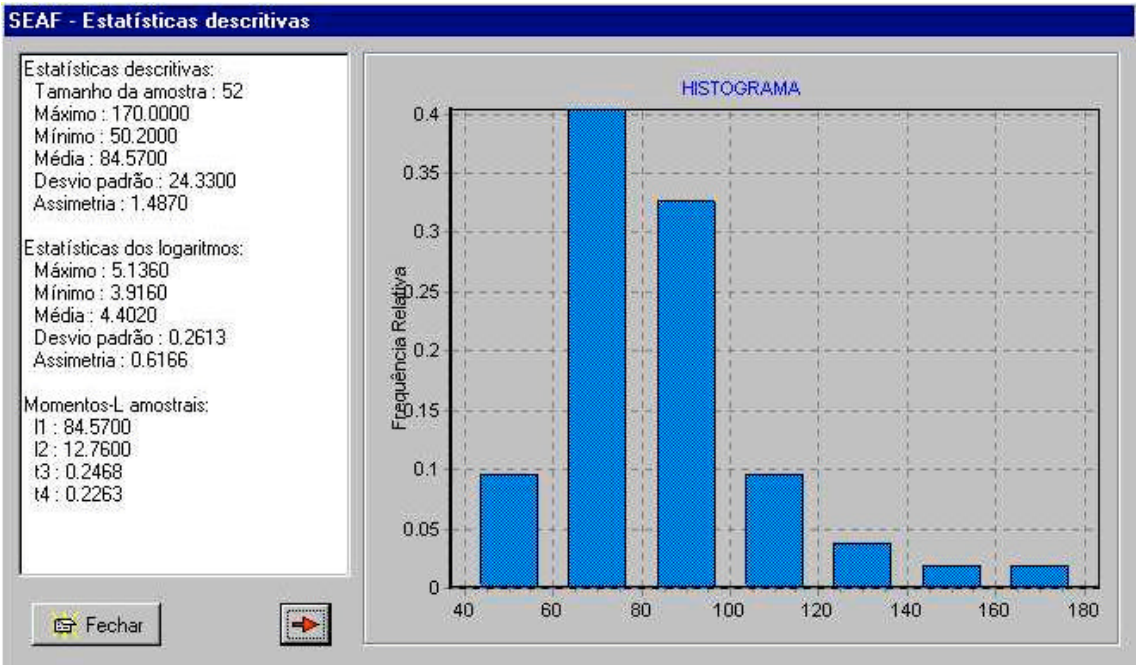



Figura-IV.10: Janela “Estatísticas descritivas”

Para avançar, o usuário deve clicar na seta direcionada para a direita. A segunda janela, denominada “Teste não paramétricos”, apresenta, de um lado, os resultados dos testes não paramétricos de independência, homogeneidade, verificando a eventual presença de *outliers*. De outro lado, apresenta um gráfico com a cronologia dos eventos, do primeiro ao enésimo ano da amostra.

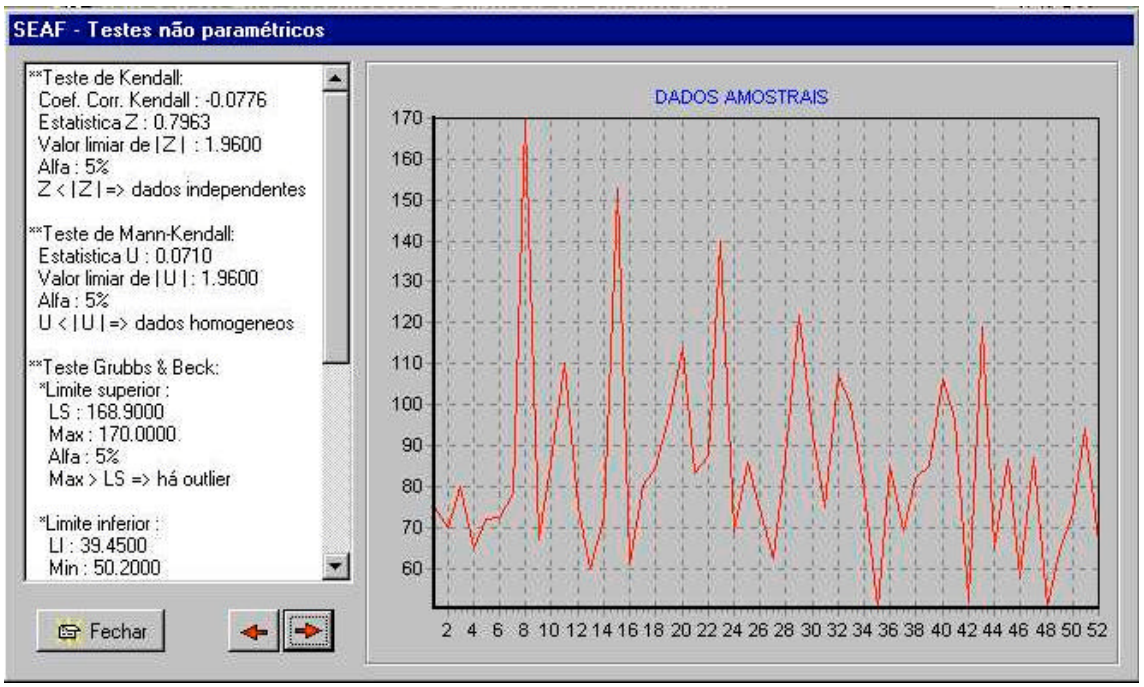



Figura-IV.11: Janela “Testes não paramétricos”.

Se o usuário desejar rever a janela “Estatísticas descritivas”, basta clicar na seta direcionada para a esquerda; caso contrário, a seta direcionada para a direita deve ser escolhida. Prosseguindo a análise, o sistema apresenta a janela “Estimação dos parâmetros”, exemplificada na Figura IV.12. Nesta janela, são apresentados os parâmetros estimados pelo método dos momentos-L para as oito distribuições analisadas pelo SEAF; as primeiras quatro são as de 2 parâmetros, seguidas pelas outras de 3 parâmetros. A janela também permite a visualização gráfica do ajuste de cada distribuição; as posições de plotagem aqui empregadas são aquelas consideradas de menor viés para cada distribuição específica, tal como recomendação de Stedinger et al. (1993). Se o usuário desejar visualizar o ajuste de uma outra distribuição, ele deverá clicar com o botão direito do *mouse* em qualquer ponto sobre o gráfico. Em conseqüência, aparecerá uma lista com todas as opções disponíveis, na qual ele poderá selecionar a distribuição desejada.

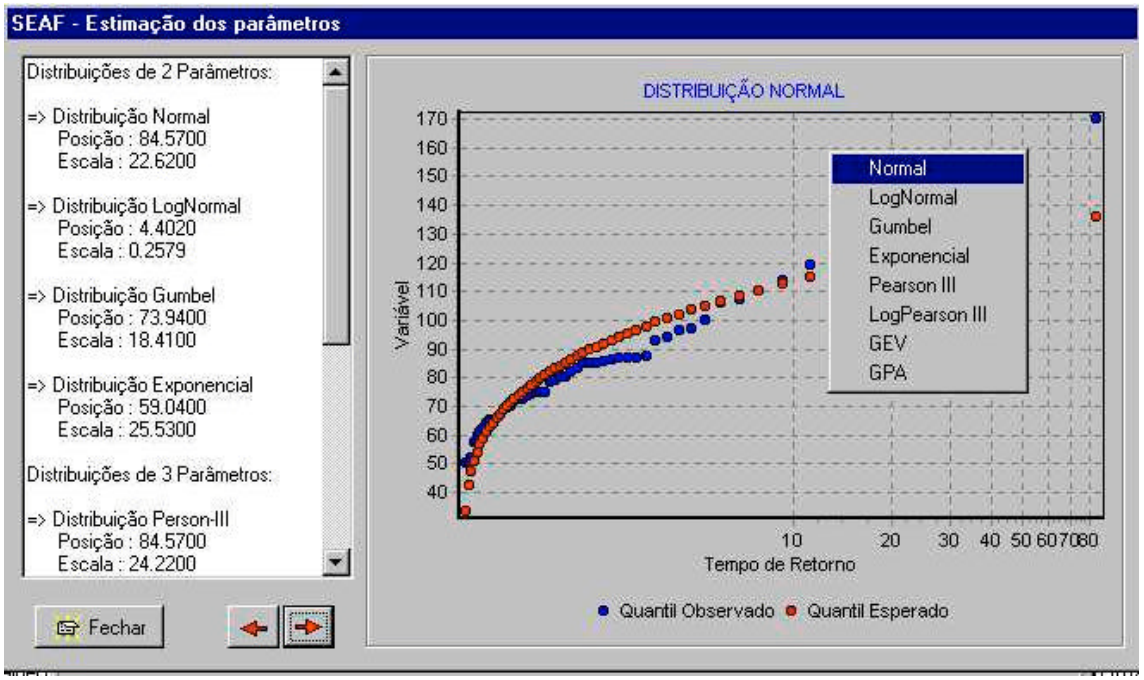


Figura-IV.12: Janela “Estimação dos parâmetros”.

Avançando, chega-se à janela “Memória de cálculo”, ilustrada na Figura IV.13. Nesta janela, o sistema SEAF apresenta um resumo dos procedimentos já efetuados sob a forma de texto de memória de cálculo. Clicando no botão “Salvar” da janela, o usuário poderá armazenar estas informações em um arquivo do tipo texto, de denominação de sua escolha.

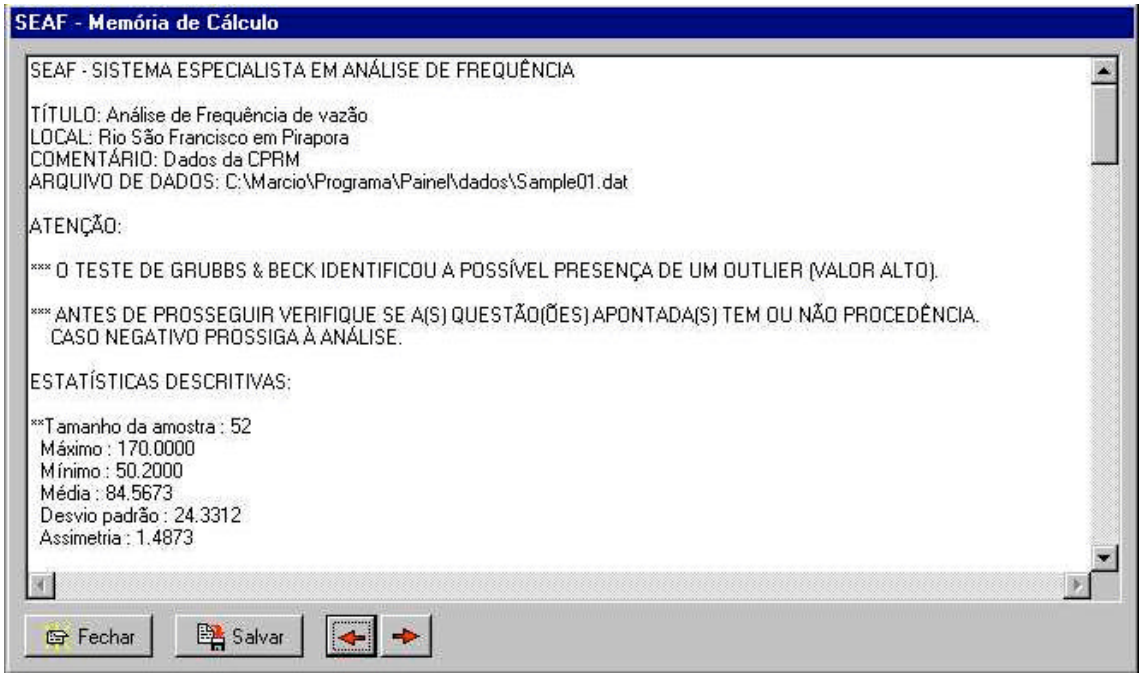


Figura-IV.13: Janela “Memória de cálculo”

Agora que o primeiro módulo do sistema já extraiu da amostra e armazenou toda a informação, o programa SEAF aciona o modo de inferência, no qual foram construídos todos os blocos de regras que emulam as etapas de um processo de raciocínio de um especialista. Clicando-se na seta direcionada para a direita, o usuário aciona o segundo módulo do sistema, constituído pelo software FuzzyCLIPS com as características apropriadas ao problema em questão, o qual, em interface com o primeiro módulo, interpreta as medidas quantitativas, associando-as a números difusos, de modo a convertê-las em declarações lingüísticas, associadas a um certo grau de confiança. Finalmente, o sistema SEAF lista as distribuições recomendadas e as classifica de acordo com seus respectivos graus de confiança. Essa etapa do sistema SEAF encontra-se exemplificada na Figura IV.14.

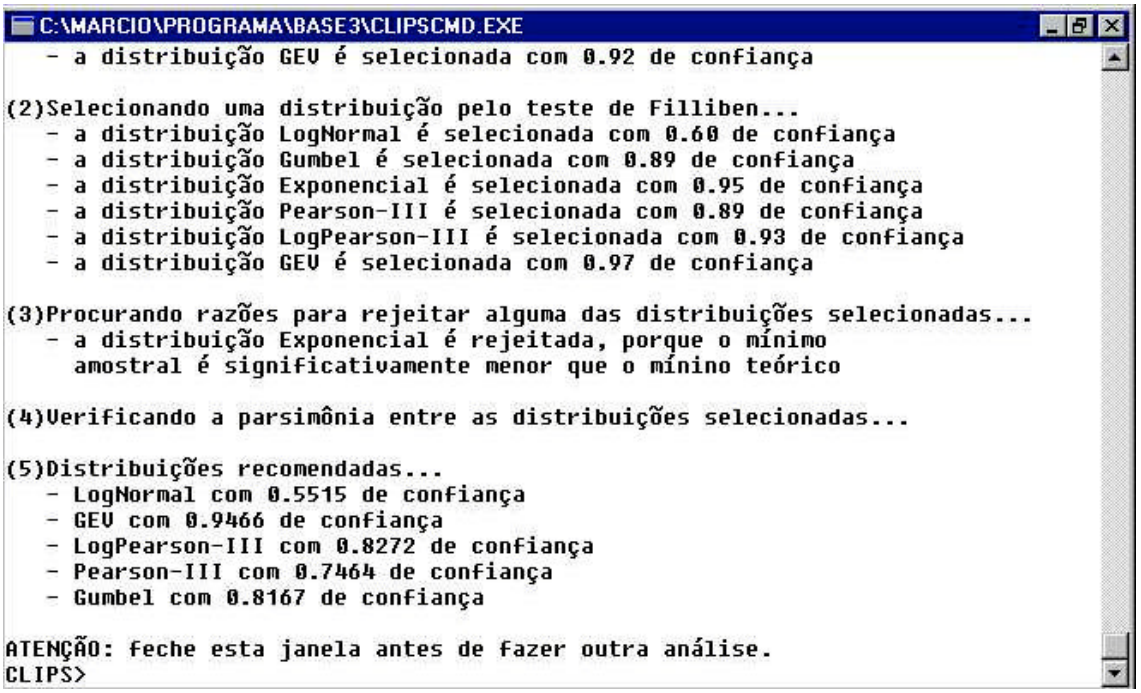



Figura-IV.14: Janela “Resultado da análise”.

Para que outra análise possa ser efetuada pelo sistema, o usuário deverá fechar a janela “Resultado da análise”.

Terminada uma certa análise, o usuário poderá gravar todos os seus resultados em um arquivo com extensão “*.prj”, para consultas posteriores e documentação. Para isso, o usuário deve entrar na opção “Projeto” e selecionar “Salvar” na lista de opções

da janela principal. Este procedimento é util, uma vez que o processo de análise pode demorar alguns minutos, dependendo do computador utilizado. Finalmente, para visualizar uma análise já armazenada, o usuário deve escolher a opção “Projeto” e, em seguida, “Abrir” na lista de opções da janela principal do programa.

O sistema SEAF foi aplicado a 20 amostras de eventos máximos anuais registrados em estações pluviométricas e fluviométricas, quase todas localizadas no estado de Minas Gerais; a escolha dessas estações teve como principal razão os tamanhos de suas respectivas amostras, os quais foram considerados suficientes para se proceder à análise de freqüência local. Essa aplicação visou avaliar se o protótipo do sistema SEAF, de fato, se comporta como especialista. Para isso, os resultados dessa aplicação foram comparados àqueles obtidos por um painel de especialistas humanos, os quais tiveram acesso e analisaram as mesmas amostras. A descrição desse experimento e as conclusões sobre a análise de desempenho do sistema SEAF são os principais objetos do Capítulo V a seguir.

# CAPÍTULO V

# ANÁLISE DE DESEMPENHO DO SISTEMA ESPECIALISTA – SEAF

## V.1 - INTRODUÇÃO

Objetivando avaliar o desempenho do sistema SEAF foram selecionadas 20 amostras de máximos anuais de alturas diárias de precipitação e vazões médias diárias para serem submetidas à análise pelo sistema. Estas amostras foram obtidas a partir dos registros diários observados em estações hidrológicas, pertencentes à Agência Nacional de Águas e localizadas na região sudeste do Brasil.

O critério utilizado para a seleção das estações foi o de ter em mãos amostras com mais de 35 anos de registros contínuos de observações diárias, cujos dados já estivessem consistidos e classificados como de boa qualidade. A justificativa para tal critério é a constatação de que amostras de tamanhos muito pequenos ou compostas por dados de má qualidade poderiam conduzir a maiores erros na estimação dos parâmetros amostrais e, consequentemente, induzir o sistema à seleção de uma distribuição de probabilidades imprópria. No caso das estações fluviométricas, observou-se também o critério de vazões naturais, ou seja, a inexistência de regularização significativa a montante. Todas as amostras selecionadas estão apresentadas no Anexo A5.

A grande maioria das estações selecionadas está contida no Estado de Minas Gerais, particularmente nas bacias de drenagem dos rios São Francisco e Doce. Apenas duas estações pluviométricas estão localizadas no Estado do Rio de Janeiro, a saber: Ponte do Souza e Conservatória. As Tabelas V.1 e V.2 apresentam uma lista das estações pluviométicas e fluviométricas utilizadas na verificação do sistema, com suas respectivas coordenadas geográficas.

Em termos gerais, o tamanho médio das amostras selecionadas é de 49 anos de observações, com um mínimo de 39 e um máximo de 59 anos. As estações pluviométricas apresentam valores de precipitação diária máxima anual entre 32,8 e 210,8mm e estão localizadas em uma faixa de altitude que varia entre 448 a 1073m. Já as estações fluviométricas apresentam valores de vazão média diária máxima anual entre 5,6 e 2.772,0$m^3/s$, com áreas de drenagens compreendidas entre 272 a 13.087$km^2$.

Tanto as vazões como as precipitações máximas anuais foram extraídas com base no ano hidrológico, o qual, no sudeste brasileiro, estende-se de outubro a setembro. A Tabela V.3 apresenta um resumo característico das amostras selecionadas.

Tabela V.1: Estações pluviométricas utilizadas na verificação do sistema

| **Código** | **Nome** | **Coordenadas** | | **Altitude** |
|---|---|---|---|---|
| | | **Latitude** | **Longitude** | **m** |
| 01544012 | São Francisco | 15º 56' 58"S | 44º 52' 05"W | 448 |
| 01645000 | São Romão | 16º 22' 18"S | 45º 04' 58"W | 472 |
| 01943000 | Mineração Moro Velho | 19º 45' 58"S | 43º 51' 00"W | 770 |
| 01944004 | Ponte Nova do Paraopeba | 19º 57' 20"S | 44º 18' 24"W | 721 |
| 01944007 | Fazenda Escola Forestal | 19º 52' 47"S | 44º 25' 18"W | 745 |
| 02044012 | Ibirité | 20º 02' 34"S | 44º 02' 36"W | 1073 |
| 02045005 | Lamounier | 20º 28' 20"S | 45º 02' 10"W | 738 |
| 02244038 | Ponte do Souza | 22º 16' 15"S | 44º 23' 26"W | 918 |
| 01943009 | Vespasiano | 19º 41' 14"S | 43º 55' 15"W | 676 |
| 02243004 | Conservatória | 22º 17' 16"S | 43º 55' 44"W | 530 |

Tabela V.2: Estações fluviométricas utilizadas na verificação do sistema

| **Código** | **Nome** | **Rio** | **Coordenadas** | | **Área** |
|---|---|---|---|---|---|
| | | | **Latitude** | **Longitude** | **$m^2$** |
| 40025000 | Vargem Bonita | São Francisco | 20º 19' 43"S | 46º 21' 58"W | 299 |
| 40050000 | Iguatama | São Francisco | 20º 10' 12"S | 45º 42' 57"W | 4846 |
| 40100000 | Porto das Andorinhas | São Francisco | 19º 16' 50"S | 45º 16' 06"W | 13087 |
| 40680000 | Entre Rio de Minas | Brumado | 20º 39' 37"S | 44º 04' 19"W | 469 |
| 41250000 | Vespasiano | da Mata | 19º 41' 14"S | 43º 55' 14"W | 676 |
| 40800001 | Ponte Nova do Paraopeba | Paraopeba | 19º 56' 57"S | 44º 18' 19"W | 5663 |
| 56028000 | Piranga | Piranga | 20º 41' 17"S | 43º 18' 02"W | 1395 |
| 56075000 | Porto Firme | Piranga | 20º 40' 13"S | 43º 05' 30"W | 4251 |
| 56415000 | Rio Casca | Casca | 20º 13' 34"S | 42º 39' 00"W | 2036 |
| 56500000 | Abre Campo | Santana | 20º 17' 56"S | 42º 28' 41"W | 272 |

Tabela V.3: Caracterização das amostras

| Código | Nº de Registros | Período | Mínimo | Máximo | Média | Desvio | Coeficiente de Assimetria |
|---|---|---|---|---|---|---|---|
| 01544012 | 52 | 1939 - 1999 | 50,2 | 170,0 | 84,6 | 24,3 | 1,5 |
| 01645000 | 41 | 1953 - 1999 | 51,0 | 140,6 | 82,6 | 22,0 | 0,7 |
| 01943000 | 46 | 1941 - 1998 | 51,0 | 144,0 | 88,0 | 25,0 | 0,6 |
| 01944004 | 54 | 1942 - 1998 | 36,0 | 159,4 | 81,7 | 22,5 | 0,8 |
| 01944007 | 42 | 1942 - 1999 | 53,3 | 166,0 | 82,2 | 24,6 | 1,8 |
| 02044012 | 41 | 1945 - 1999 | 46,1 | 196,2 | 90,1 | 31,5 | 1,4 |
| 02045005 | 53 | 1942 - 1998 | 53,8 | 170,0 | 91,1 | 28,2 | 1,0 |
| 02244038 | 50 | 1942 - 1996 | 62,0 | 200,4 | 93,1 | 28,1 | 1,7 |
| 01943009 | 39 | 1941 - 1998 | 32,8 | 210,8 | 80,1 | 29,6 | 2,5 |
| 02243004 | 48 | 1945 - 1996 | 45,8 | 143,9 | 75,2 | 20,9 | 1,3 |
| 40025000 | 43 | 1940 - 1999 | 18,6 | 295,0 | 88,8 | 58,1 | 1,9 |
| 40050000 | 57 | 1936 - 1999 | 140,0 | 1227,0 | 494,4 | 189,3 | 1,4 |
| 40100000 | 41 | 1959 - 1999 | 234,0 | 2771,0 | 879,6 | 400,5 | 3,0 |
| 40680000 | 59 | 1939 - 1999 | 36,5 | 434,0 | 84,2 | 58,5 | 4,2 |
| 41250000 | 47 | 1940 - 1995 | 12,9 | 246,0 | 72,4 | 43,9 | 1,8 |
| 40800001 | 59 | 1939 - 1999 | 226,0 | 1070,0 | 512,5 | 189,8 | 0,8 |
| 56028000 | 54 | 1938 - 1998 | 74,7 | 727,0 | 226,7 | 145,9 | 1,7 |
| 56075000 | 58 | 1939 - 1999 | 183,0 | 1150,0 | 357,9 | 163,5 | 2,4 |
| 56415000 | 57 | 1931 - 1999 | 63,2 | 350,0 | 155,7 | 61,1 | 1,2 |
| 56500000 | 47 | 1940 - 1997 | 5,6 | 249,0 | 42,7 | 37,2 | 4,0 |

## V.2 – ANÁLISES DAS AMOSTRAS

Os momentos-L e razões-L das amostras selecionadas estão listados na Tabela A5.21 do Anexo-A5. Verifica-se que o valor médio de $\tau_3$ para as amostras selecionadas é 0,278, e varia de 0,122 a 0,456. Conforme apresentado na Figura-V.1, há uma variação aproximadamente uniforme dos valores amostrais de $\tau_3$ em sua amplitude de amostragem, o que indica a seleção de amostras não tendenciosas.

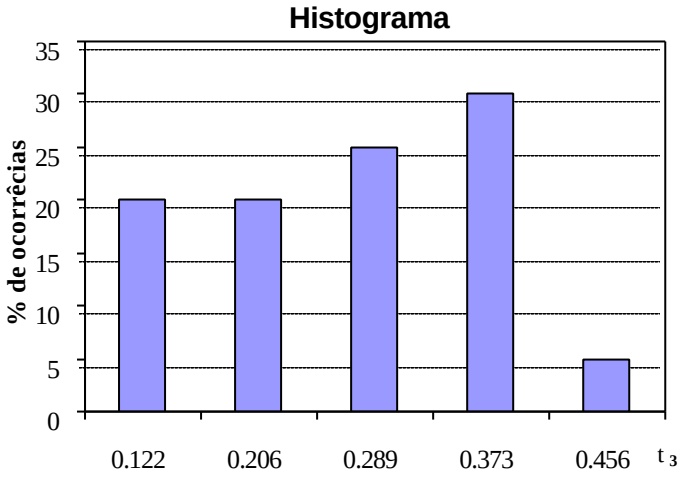



Figura V.1: Histograma de freqüências relativas de $\tau_3$.

A transformação logarítmica dos dados amostrados provoca uma redução no valor médio de $\tau_3$ de 0,278 para 0,086. Contudo as amostras continuam apresentado uma distribuição aproximadamente uniforme de $\tau_3$ variando de -0,031 a 0,226, conforme pode ser visualizado na Figura-V.2. É interessante notar que as amostras originais não apresentam valores negativos de $\tau_3$. Entretanto, após a transformação logarítmica dos dados, as estações 01944004, 41250000, 40800001 e 56500000 apresentaram valores de negativos de $\tau_3$.

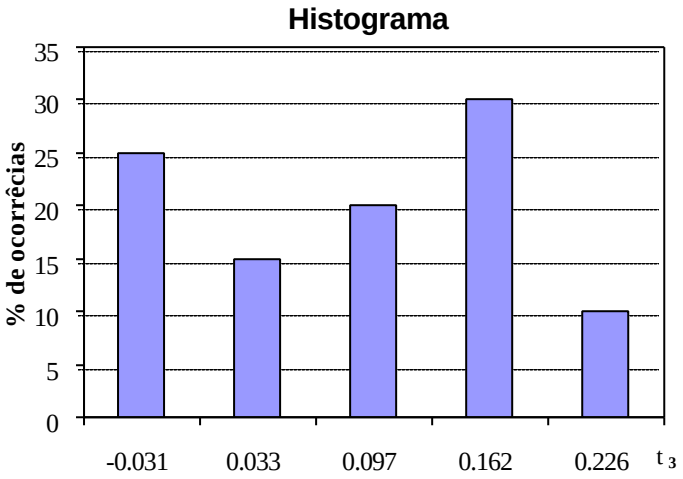



Figura V.2: Histograma de freqüências relativas de $\tau_3$ para os valores dos logaritmos dos dados amostrais.

Os parâmetros das oito distribuições analisadas foram estimados com base nos momentos-L amostrais e estão apresentados nas Tabelas de A5.22 a A5.24 do Anexo-A5.

Os ajustes das amostras à distribuição Exponencial resultaram em limites mínimos teóricos superiores aos valores mínimos contidos em todas as amostras analisadas, sendo que para 15 estações (01544012, 01645000, 01943000, 01944004, 02044012, 02045005, 01943009, 02243004, 40025000, 40050000, 40100000, 41250000, 40800001, 56415000, 56500000) eles foram significativamente maiores que os seus respectivos mínimos amostrais. Isto diminui o grau de confiança que o sistema SEAF tem sobre a possibilidade da distribuição Exponencial representar adequadamente a distribuição destas 15 amostras, uma vez que elas apresentam valores amostrais significativamente fora dos limites de valores das distribuições Exponenciais ajustadas.

No caso da distribuição Pearson III, os limites mínimos estimados das distribuições ajustadas são superiores aos valores mínimos amostrados em 15 amostras, sendo que em 9 delas (02044012, 01943009, 40025000, 40050000, 40100000, 40680000, 41250000, 56028000, 56500000) eles foram significativamente maiores que

os seus respectivos mínimos amostrais. Tal fato também diminui o grau de confiança do sistema SEAF sobre a escolha da distribuição Pearson III para modelar os dados destas estações.

As amostras 01944004, 41250000, 40800001 e 56500000, quando ajustadas à distribuição Log-Pearson III, apresentaram-se limitadas à direita, o que, mais uma vez, diminui o grau de confiança do sistema SEAF sobre a escolha da distribuição Log-Pearson III para modelar os dados destas amostras. Este mesmo fato ocorreu para as amostras 01645000, 01943000, 01944004 e 40800001, quando ajustadas à distribuição Generalizada de Valores Extremos.

Os parâmetros de forma estimados para a distribuição Generalizada de Pareto são positivos em 13 casos, a saber: 01544012, 01645000, 01943000, 01944004, 02044012, 02045005, 01943009, 02243004, 40050000, 41250000, 40800001, 56075000, 56415000; tal ocorrência impõe limites superiores às distribuições ajustadas. Além disto, nas seguintes 13 amostras: 01544012, 01944004, 02044012, 01943009, 02243004, 40025000, 40100000, 40680000, 40050000, 41250000, 40800001, 56415000, e 56500000, os limites à esquerda da distribuição GPA foram significativamente superiores aos correspondentes mínimos amostrais.

Apenas em três amostras (01944007, 02244038, 56028000), os ajustes das distribuições candidatas aos dados não tiveram ressalvas em relação aos seus limites estimados.

É possível verificar a ocorrência de *outliers* nas seguintes 5 amostras: 01943009, 40100000, 406800000, 56075000 e 56500000, por meio da análise gráfica dos ajustes apresentados no Anexo-A5

## V.3 – ANÁLISE DAS AMOSTRAS POR MEIO DO SISTEMA SEAF

As 20 amostras selecionadas foram submetidas à análise pelo sistema SEAF e os resultados estão apresentados a seguir:

Identificação da presença de *outliers*

O sistema SEAF identificou a presença de *outliers* em 11 das 20 amostras analisadas, ver Tabela-V.4. Verifica-se que os testes utilizados pelo sistema para a identificação da presença de *outliers* mostraram-se eficientes, mas conservadores quando avaliados isoladamente. Contudo, considerando-se os casos de resultados consensuais, em ambos os testes, o número de acertos foi maior em relação ao número de amostras identificadas com suspeitas de *outliers*, quando comparados com o exame visual realizado pelo autor do presente trabalho.

Tabela-V.4: Resultados do testes para verificação da presença de *outliers*

| **Amostras** | **Avaliação da presença de outlier** | | |
|---|---|---|---|
| | **Grubbs e Beck** | **Assimetria** | **Visual** |
| 01544012 | * | | |
| 01944004 | | * | |
| 01944007 | * | | |
| 02244038 | * | * | |
| 01943000 | * | * | * |
| 40050000 | | * | |
| 40100000 | * | * | * |
| 40680000 | * | * | * |
| 41250000 | | * | |
| 56075000 | * | * | * |
| 56500000 | * | * | * |

Distribuições de probabilidades aprovadas em testes estatísticos

As distribuições aprovadas nos testes estatísticos, utilizados pelo sistema SEAF, estão listadas na Tabela V.5. Observa-se que, em média, 5 das 8 distribuições analisadas podem representar os dados amostrais, com base apenas nestes testes. Além disto, com exceção da distribuição Normal, todas as distribuições analisadas foram aprovadas no mínimo em 12 das 20 amostras selecionadas, sendo que a distribuição Generalizada de Valores Extremos ajustou-se em todas elas. As distribuições de 3 parâmetros foram

aprovadas em um número maior de amostras do que as distribuições de dois parâmetros, uma vez que elas possuem um grau de liberdade a mais, o que permite que elas se ajustem melhor aos dados amostrados devido à adição do parâmetro de forma em sua modelagem. Por outro lado, a adição de mais um parâmetro à modelagem pode provocar um acréscimo nas incertezas dos parâmetros populacionais estimados. Além disso, erros de amostragem podem induzir erroneamente a seleção de modelos com maior número de parâmetros. Com isto, vê-se a importância da adoção de outros critérios, além dos testes de ajuste tradicionais, para a seleção e classificação das distribuições de probabilidades analisadas, uma vez que tais testes não são capazes de fazer qualquer distinção entre as distribuições aprovadas. Um exemplo disto, está apresentado na Tabela V.5 onde foram aprovadas mais de uma distribuição de probabilidades para modelagem dos dados de cada amostra.

Tabela-V.5: Resultados dos testes estatísticos para as amostras selecionadas

| **Amostra** | **Distribuições aprovadas** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | NOR | LNR | GUM | EXP | PE3 | LP3 | GEV | GPA |
| 01544012 | | * | * | * | * | * | * | |
| 01645000 | | * | * | * | * | * | * | * |
| 01943000 | * | * | * | | * | * | * | * |
| 01944004 | | * | * | | * | * | * | |
| 01944007 | | | | * | * | * | * | * |
| 02044012 | | * | * | * | * | * | * | * |
| 02045005 | | * | * | * | * | * | * | * |
| 02244038 | | | * | * | * | * | * | * |
| 01943009 | | * | | * | * | * | * | |
| 02243004 | | * | * | * | * | * | * | * |
| 40025000 | | | | * | * | * | * | * |
| 40050000 | | * | * | * | * | * | * | |
| 40100000 | | * | | * | | | * | |
| 40680000 | | | | | | * | * | * |
| 41250000 | | * | * | * | * | * | * | * |
| 40800001 | | * | * | * | * | * | * | * |
| 56028000 | | * | | * | * | * | * | * |
| 56075000 | | | | * | * | * | * | * |
| 56415000 | | * | * | * | * | * | * | * |
| 56500000 | | * | | * | | | * | |
| **TOTAL** | **1** | **15** | **12** | **17** | **17** | **18** | **20** | **14** |

Nota: NOR – Distribuição Normal, LNR – Distribuição Lognormal, GUM – Distribuição de Gumbel, EXP – Distribuição Exponencial, PE3 – Distribuição Pearson III, LP3 – Distribuição Log-Pearson III, GEV – Distribuição Generalizada de Valores Extremos, GPA – Distribuição Generalizada de Pareto.

Distribuições rejeitadas

O sistema SEAF busca evidências para rejeitar algumas das distribuições aprovadas em seus testes de ajuste. Os critérios adotados para isto estão baseados nos limites teóricos de cada uma das distribuições ajustadas e na verificação da parcimônia entre distribuições de mesma família, conforme apresentado no Capítulo IV.

Em conformidade com a Tabela-V.6, observa-se que, em média, 2 distribuições satisfizeram os critérios adotados para sua exclusão da lista de distribuições classificadas, sendo o critério de verificação dos limites amostrais responsável por 84,4% dos casos de rejeição. Vale ressaltar que as distribuições Exponencial e Generalizada de Pareto representam, juntas, 57,8% de todos os casos de rejeição.

Tabela-V.6: Distribuições rejeitadas pelo sistema

| **Amostra** | **CRITÉRIO** | | | | |
|---|---|---|---|---|---|
| | **Limites superior e/ou inferior** | | | | **Parcimônia** |
| 01544012 | EXP | | | | |
| 01645000 | EXP | GPA | GEV | | LP3 |
| 01943000 | GPA | GEV | | | LP3 |
| 01944004 | GEV | LP3 | | | |
| 01944007 | | | | | GPA |
| 02044012 | GPA | PE3 | EXP | | |
| 02045005 | GPA | EXP | | | GEV |
| 02244038 | | | | | GPA |
| 01943009 | PE3 | EXP | | | |
| 02243004 | GPA | EXP | | | |
| 40025000 | GPA | EXP | PE3 | | |
| 40050000 | EXP | LP3 | PE3 | | GEV |
| 40100000 | EXP | | | | |
| 40680000 | GPA | | | | |
| 41250000 | GPA | LP3 | PE3 | EXP | |
| 40800001 | EXP | GPA | GEV | LP3 | |
| 56028000 | PE3 | | | | |
| 56075000 | GPA | | | | |
| 56415000 | GPA | EXP | | | LP3 |
| 56500000 | EXP | | | | |

Nota: NOR – Distribuição Normal, LNR – Distribuição Lognormal, GUM – Distribuição de Gumbel, EXP – Distribuição Exponencial, PE3 – Distribuição Pearson III, LP3 – Distribuição Log-Pearson III, GEV – Distribuição Generalizada de Valores Extremos, GPA – Distribuição Generalizada de Pareto.

Distribuições classificadas pelo sistema

Os resultados finais das análises realizadas pelo sistema SEAF estão apresentados na Tabela V.7, em ordem decrescente de níveis médios de confiança atribuídos pelo SEAF a cada distribuição não rejeitada anteriormente, de acordo com as regras descritas no Capítulo 4 dessa dissertação. Na Tabela V.7, a primeira coluna após a de identificação da amostra, representa as distribuições com maiores níveis de confiança associados durante as análises realizadas pelo sistema, ou sejam, as escolhas primeiras do sistema.

Tabela-V.7: Distribuições classificadas pelo sistema para as amostras selecionadas

| **Amostra** | **Distribuições classificadas** | | | | |
|---|---|---|---|---|---|
| | **1ª** | **2ª** | **3ª** | **4ª** | **5ª** |
| 01544012 | GEV | LP3 | GUM | PE3 | LNR |
| 01645000 | GUM | PE3 | LNR | | |
| 01943000 | LNR | GUM | PE3 | NOR | |
| 01944004 | GUM | LNR | PE3 | | |
| 01944007 | GEV | EXP | LP3 | PE3 | |
| 02044012 | GEV | LP3 | GUM | LNR | |
| 02045005 | PE3 | GUM | LP3 | LNR | |
| 02244038 | EXP | PE3 | GEV | LP3 | GUM |
| 01943009 | GEV | LP3 | LNR | | |
| 02243004 | GEV | LP3 | PE3 | GUM | LNR |
| 40025000 | GEV | LP3 | | | |
| 40050000 | GUM | LNR | | | |
| 40100000 | GEV | LNR | | | |
| 40680000 | GEV | LP3 | | | |
| 41250000 | GEV | LNR | GUM | | |
| 40800001 | GUM | LNR | PE3 | | |
| 56028000 | GPA | EXP | GEV | LP3 | LNR |
| 56075000 | GEV | LP3 | EXP | PE3 | |
| 56415000 | GEV | LNR | GUM | PE3 | |
| 56500000 | LNR | GEV | | | |

Nota: NOR – Distribuição Normal, LNR – Distribuição Lognormal, GUM – Distribuição de Gumbel, EXP – Distribuição Exponencial, PE3 – Distribuição Pearson III, LP3 – Distribuição Log-Pearson III, GEV – Distribuição Generalizada de Valores Extremos, GPA – Distribuição Generalizada de Pareto.

## V.4 – AVALIAÇÃO DO DESEMPENHO DO SISTEMA SEAF

Selecionar uma distribuição de probabilidades não é um problema completamente objetivo, para o qual se tem uma solução claramente correta ou totalmente errada. Freqüentemente, é impossível demonstrar de forma objetiva, de um lado, que o modo de raciocínio de um especialista humano é correto ou, de outro, que a base de conhecimentos implementada em um sistema computacional está errada. Conforme pode ser visto na Tabela V.5, de modo geral, para 5 das 8 distribuições analisadas não houve evidências estatísticas que as excluíssem da análise. Para dar continuidade ao processo de classificação, as distribuições aprovadas nos testes estatísticos foram submetidas a uma análise subjetiva, a qual envolve um conjunto de critérios pessoais, os quais podem conduzir a diferentes escolhas, quando comparadas àquelas obtidas por outros especialistas.

A fim de avaliar a performance do sistema SEAF foram realizados dois experimentos. No primeiro, as mesmas 20 amostras de alturas diárias de chuva e vazões médias diárias máximas anuais observadas em estações pluviométricas e fluviométricas localizadas na região sudeste do Brasil, que haviam sido objeto de análise pelo sistema SEAF, foram também submetidas à análise de freqüência por parte de um painel de peritos. Em um segundo experimento, aplicou-se o sistema SEAF a amostras de tamanho típico retiradas de populações sintéticas correspondentes às distribuições candidatas.

Primeiro Experimento: Painel de especialistas humanos

O primeiro experimento para avaliar a performance do sistema SEAF consistiu na formação inicial de um painel composto por 12 profissionais especialistas em análise de freqüência. A cada membro do painel foram enviados, por correio eletrônico, as 20 amostras selecionadas, os respectivos gráficos de quantis classificados versus probabilidades empíricas, as principais estatísticas amostrais calculadas por momentos convencionais e momentos-L, bem como uma carta contendo as orientações necessárias à condução do experimento. Além disto, foi enviada uma descrição da versão preliminar das regras heurísticas do sistema; a esse respeito, salientou-se na carta que caberia aos

peritos decidir sobre a leitura e a emissão de eventuais comentários ou avaliações sobre as regras heurísticas.

Em função das exigências relativas ao tempo de execução deste trabalho, não foi possível esperar o envio das análises de todos os especialistas, os quais, por motivos pessoais, não puderam concluir ou até mesmo realizar as suas análises, sendo que apenas 4 dos 12 especialistas enviaram suas respostas.

Tal fato não prejudicou as avaliações realizadas neste experimento, uma vez que seu objetivo não foi avaliar a performance individual do sistema em relação a cada especialista, e sim o seu comportamento nos casos de maior consenso entre os especialistas. Além disto, conforme será descrito a seguir, pode se somar o fato de que os processos de análises utilizados por cada um dos membros do painel foram diferentes, o que dá aos casos de consenso uma maior confiabilidade, uma vez que um mesmo resultado foi obtido por processos analíticos distintos.

Em sua formação final o painel de especialistas ficou composto por: dois estatísticos, um engenheiro-hidrólogo e um engenheiro professor universitário, todos com conhecimento específico e experiência na área de análise de freqüência de variáveis hidrológicas e/ou geofísicas.

Cada membro do painel analisou o conjunto de amostras e selecionou, para cada amostra, uma distribuição de probabilidades com base em suas regras heurísticas pessoais; 3 dos 4 peritos escreveram breves justificativas a respeito de suas escolhas. Os resultados enviados por cada especialista estão apresentados na Tabela V.8. Conforme norma de conduta pré-estabelecida, os membros do painel de especialistas não foram nominalmente identificados.

De acordo com os resultados, para nenhuma amostra ocorreu uma escolha consensual entre os todos especialistas. Trata-se de um fato já esperado, uma vez que a formação do painel é composta por especialistas com diferentes experiências em análise de freqüência, bem como com diferentes expectativas e visões em relação aos produtos da análise em si, tendo cada um conduzido a escolha do modelo apropriado de forma diferenciada.

O especialista-1 fez sua análise com o auxílio do aplicativo *"GenStat"*, no qual a medida da qualidade de ajuste é expressa em termos de uma 'estatística de desvio' ou *"Deviance statistic"*. Quanto menor o valor desta estatística, tanto melhor é o seu ajuste. O critério utilizado pelo especialista-1 foi objetivamente escolher a distribuição que tivesse a menor *"Deviance statistic"*. O universo de distribuições utilizadas na análise feita pelo aplicativo *"GenStat"* não foi o mesmo daquele considerado pelo SEAF. Para efeito da presente avaliação, somente foram consideradas as distribuições analisadas pelo *"GenStat"* que também fazem parte do conjunto de distribuições analisadas pelo SEAF.

O especialista-2 não acredita que seja possível identificar um único modelo como a distribuição populacional de uma série, cujo tamanho amostral seja da ordem de 50 anos de dados. Por outro lado, ele preconiza o uso de informações regionais para a identificação da distribuição apropriada à série em questão. Em sua análise, ele presumiu não haver diferenças climáticas ou geomorfológicas entre as estações que pudessem proporcionar significativa heterogeneidade regional do ponto de vista da freqüência de ocorrência das variáveis em estudo. Em conseqüência, os seus resultados foram obtidos através de análise de freqüência regional, considerando separadamente os dados de precipitação e vazão, ambos sob a premissa de uma única região homogênea. A metodologia utilizada pelo especialista-2 é aquela apresentada em Hosking e Wallis (1997).

O especialista-3 utilizou critérios subjetivos de análise similares aos implementadas pelo sistema SEAF. Seus resultados estão fundamentados basicamente na análise visual das distribuições ajustadas em papel de probabilidades e das razões-L amostrais grafadas no diagrama de momentos-L. Além disto, ele utiliza os testes de Filliben e Komolgorov-Smirnov para verificar a aderência da distribuição ajustada à empírica.

O especialista-4 conduziu o processo de análise de modo semelhante ao empregado pelo especialista-3, contudo dando maior ênfase ao ajuste dos dados à cauda superior das distribuições.

Os resultados obtidos foram agrupados em dois grupos para avaliação da performance do sistema. O primeiro grupo, onde estão reunidas as amostras com duas ou mais soluções coincidentes entre os especialistas, está apresentado na coluna de título "≥2" da Tabela V.8.

Tabela-V.8: Resultados enviados pelos membros do painel de especialistas.

| Amostra | Especialista | | | | nº de Coincidências | |
|---|---|---|---|---|---|---|
| | 1º | 2º | 3º | 4º | ≥ 2 | = 3 |
| 01544012 | LNR | GEV | GEV | EXP | GEV | |
| 01645000 | LNR | GEV | GUM | GUM | GUM | |
| 01943000 | LNR | GEV | GUM | LNR | LNR | |
| 01944004 | LNR | GEV | GUM | EXP | | |
| 01944007 | LNR | GEV | GEV | EXP | GEV | |
| 02044012 | LP3 | GEV | GEV | EXP | GEV | |
| 02045005 | LP3 | GEV | GUM | LP3 | LP3 | |
| 02244038 | LP3 | GEV | GEV | LP3 | | |
| 01943009 | LNR | GEV | GEV | GEV | GEV | GEV |
| 02243004 | LNR | GEV | GEV | EXP | GEV | |
| 40025000 | LP3 | GEV | GPA | GEV | GEV | |
| 40050000 | LNR | GEV | GEV | EXP | GEV | |
| 40100000 | LNR | GEV | GEV | LP3 | GEV | |
| 40680000 | LP3 | GEV | GEV | LP3 | | |
| 41250000 | LP3 | GEV | GEV | GEV | GEV | GEV |
| 40800001 | LNR | GEV | GUM | GUM | GUM | |
| 56028000 | LP3 | GEV | GPA | LP3 | LP3 | |
| 56075000 | LNR | GEV | GEV | GEV | GEV | GEV |
| 56415000 | LNR | GEV | GEV | LP3 | GEV | |
| 56500000 | LNR | GEV | GEV | GEV | GEV | GEV |

Nota: NOR – Distribuição Normal, LNR – Distribuição Lognormal, GUM – Distribuição de Gumbel, EXP – Distribuição Exponencial, PE3 – Distribuição Pearson III, LP3 – Distribuição Log-Pearson III, GEV – Distribuição Generalizada de Valores Extremos, GPA – Distribuição Generalizada de Pareto.

Comparando os resultados deste grupo a seus correspondentes obtidos pelo sistema SEAF, verifica-se que estes valores são em 76,5% dos casos coincidentes com a primeira escolha do sistema. Considerando as duas primeiras escolhas do SEAF este valor aumenta para 82,4% chegando a 94,1%, quando todas as distribuições classificadas estiverem sendo consideradas. Já o segundo grupo, formado pelas amostras com três soluções coincidentes entre os especialistas, está apresentado na coluna de título "=3" da Tabela V.8. Comparando os resultados deste grupo aos obtidos pelo

SEAF, os valores, acima apresentados para o primeiro grupo, passam para 75%, 100% e 100%, respectivamente. Um resumo desta mesma avaliação com os resultados individuais dos especialistas está apresentado na Tabela V.9.

Tabela-V.9: Comparação entre os resultados do painel e do SEAF

| **Critério de escolha** | **Especialista** | | | | **nº de Coincidências** | |
|---|---|---|---|---|---|---|
| | **1º** | **2º** | **3º** | **4º** | **≥ 2** | **= 3** |
| 1ª | 10,0 | 55,0 | 70,0 | 35,0 | 76,5 | 75,0 |
| 1ª ou 2ª | 50,0 | 60,0 | 85,0 | 50,0 | 82,4 | 100,0 |
| entre as Classificadas | 85,0 | 70,0 | 90,0 | 65,0 | 94,1 | 100,0 |

Segundo Experimento: Simulação de amostras sintéticas

No segundo experimento foram geradas, pelo método de *"Monte Carlo"*, amostras sintéticas com base nos momentos-L e razões-L correspondentes às 20 amostras selecionadas. Para cada conjunto de momentos-L e razões-L amostrais, foram geradas 8 séries sintéticas correspondentes às formas paramétricas das distribuições candidatas, exceto nos casos onde a distribuição ajustada fosse limitada na cauda superior. No total, foram geradas 139 séries sintéticas com distribuições populacionais conhecidas, as quais foram submetidas ao sistema SEAF com o intuito de avaliar o número de acertos do sistema e se há qualquer tendência em suas análises. Os resultados por este experimento estão apresentados na Tabela-V.10.

Comparando os resultados obtidos pelo sistema SEAF, relativos às séries sintéticas, verifica-se que em 53,2% dos casos o sistema identifica, na primeira ou na segunda escolha, a distribuição que gerou a série sintética analisada. Considerando todas as distribuições classificadas pelo sistema, este número aumenta para 89,2% dos casos. Observa-se também que, em 73,4% dos casos, a distribuição que gerou a série sintética ou sua parente da mesma família estão classificadas, pelo sistema, como a primeira ou a segunda escolha.

Vale ressaltar que as amostras extraídas das séries sintéticas modeladas pela distribuição Generalizada de Pareto tiveram em sua maioria o parâmetro de forma positivo em seus ajustes à distribuição Generalizada de Pareto. Tal fato é condição para

que haja, neste modelo, um limite máximo à direita, o que explica a baixíssima performance do sistema ao modelo, principalmente em relação a primeira escolha, uma vez que o sistema SEAF foi construído para rejeitar distribuições limitadas na direção dos máximos.

À exceção da distribuição Generalizada de Pareto, verifica-se que acima de 80,0% dos casos o modelo populacional das amostras geradas está presente na lista de distribuições classificadas pelo sistema SEAF, o que mostra uma certa coerência dos resultados apresentados.

Tabela-V.10: Resultados obtidos pelo SEAF na avaliação das séries sintéticas

| **Amostras sintéticas** | **Porcentagem de acertos** | | | | |
|---|---|---|---|---|---|
| | **Distribuição** | | | **Família** | |
| | **1ª** | **2ª** | **Classificadas** | **1ª** | **2ª** |
| NOR | 50,0 | 70,0 | 95,0 | 90,0 | 100,0 |
| LNR | 25,0 | 65,0 | 95,0 | 30,0 | 75,0 |
| GUM | 40,0 | 55,0 | 95,0 | 70,0 | 80,0 |
| EXP | 20,0 | 30,0 | 95,0 | 35,0 | 60,0 |
| PE3 | 15,0 | 45,0 | 90,0 | 20,0 | 50,0 |
| LP3 | 6,3 | 31,3 | 81,3 | 25,0 | 50,0 |
| GEV | 56,3 | 75,0 | 81,3 | 62,5 | 93,8 |
| GPA | 0,0 | 57,1 | 57,1 | 28,6 | 85,7 |
| **TOTAL** | **28,8** | **53,2** | **89,2** | **46,8** | **73,4** |

Nota: NOR – Distribuição Normal, LNR – Distribuição Lognormal, GUM – Distribuição de Gumbel, EXP – Distribuição Exponencial, PE3 – Distribuição Pearson III, LP3 – Distribuição Log-Pearson III, GEV – Distribuição Generalizada de Valores Extremos, GPA – Distribuição Generalizada de Pareto.

Avaliando a primeira escolha do sistema, verifica-se que o nível de acerto está abaixo dos 50%, a exceção das amostras sintéticas geradas pelas distribuições Normal e GEV, as quais apresentaram níveis de acerto de 50% e 56,3%, respectivamente. Conforme os resultados apresentados na Tabela V.11, observa-se que não há uma tendência ou preferência por um modelo de distribuição de probabilidades específico.

Tabela V.11: Primeira escolha do SEAF para as amostras sintéticas

| Amostras sintéticas | 1ª escolha do SEAF | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **NOR** | **LNR** | **GUM** | **EXP** | **PE3** | **LP3** | **GEV** | **GPA** |
| **NOR** | **50,0** | 5,0 | 5,0 | 0,0 | 40,0 | 0,0 | 0,0 | 0,0 |
| **LNR** | 5,0 | **25,0** | 10,0 | 0,0 | 40,0 | 5,0 | 15,0 | 0,0 |
| **GUM** | 0,0 | 25,0 | **40,0** | 0,0 | 5,0 | 0,0 | 30,0 | 0,0 |
| **EXP** | 0,0 | 0,0 | 5,0 | **20,0** | 35,0 | 5,0 | 20,0 | 15,0 |
| **PE3** | 5,0 | 5,0 | 30,0 | 15,0 | **15,0** | 0,0 | 20,0 | 10,0 |
| **LP3** | 0,0 | 18,8 | 12,5 | 6,3 | 12,5 | **6,3** | 37,5 | 6,3 |
| **GEV** | 0,0 | 12,5 | 6,3 | 0,0 | 12,5 | 12,5 | **56,3** | 0,0 |
| **GPA** | 0,0 | 0,0 | 14,3 | 28,6 | 0,0 | 14,3 | 42,9 | **0,0** |
| **TOTAL** | **8,6** | **12,2** | **15,8** | **7,2** | **22,3** | **4,3** | **25,2** | **4,3** |

Nota: NOR – Distribuição Normal, LNR – Distribuição Lognormal, GUM – Distribuição de Gumbel, EXP – Distribuição Exponencial, PE3 – Distribuição Pearson III, LP3 – Distribuição Log-Pearson III, GEV – Distribuição Generalizada de Valores Extremos, GPA – Distribuição Generalizada de Pareto.

## V.5 – DISCUSSÃO DOS RESULTADOS

De acordo com os dois experimentos realizados neste trabalho, é possível afirmar que o desempenho do sistema SEAF foi satisfatório, podendo ser comparado àquele de um especialista em análise de freqüência de variáveis hidrológicas. Nesse sentido, o SEAF apresenta-se como uma ferramenta bastante útil de auxílio à escolha de uma distribuição de probabilidade entre diversos modelos candidatos. Conforme apresentado na Tabela V.5, em média, 5 das 8 distribuições candidatas podem estatisticamente representar as amostras selecionadas. Entretanto, o ato de escolher uma entre delas envolve a aplicação de critérios subjetivos, os quais dependem do especialista envolvido. Isto justifica a total falta de consenso entre os membros do painel. Evidentemente, o sistema não dispensa a capacitação técnica e o discernimento próprio do ser humano no processo de seleção de uma distribuição de probabilidades. Entretanto, os resultados mostram que o protótipo SEAF demonstrou sua utilidade como ferramenta de auxílio à decisão e instrumento de valor para o melhor entendimento dos métodos de análise de freqüência por parte de estudantes e hidrólogos não-especialistas.

# CAPÍTULO-VI

# CONCLUSÕES E RECOMENDAÇÕES

## IV.1 – CONCLUSÕES

Na presente dissertação, foi desenvolvido um protótipo de um sistema especialista para auxiliar na condução da análise de freqüência local de eventos máximos anuais de variáveis hidrológicas e hidrometeorológicas, particularmente no que se refere à escolha do modelo paramétrico mais apropriado. O sistema, denominado SEAF (Sistema Especialista de Análise de Freqüência), é instrumentado para analisar a veracidade das premissas necessárias para a análise de freqüência convencional, a saber: as hipóteses de homogeneidade e independência serial, possui meios para identificar a presença de *outliers* na amostra, bem como extrair das informações numéricas calculadas as justificativas necessárias para selecionar um ou um pequeno número de modelos paramétricos apropriados, dentre o conjunto formado pelas seguintes distribuições de probabilidades: Normal, Log-Normal de 2 parâmetros, Gumbel, Generalizada de Valores Extremos, Exponencial, Generalizada de Pareto, Pearson III e Log-Pearson III.

De modo resumido, o programa SEAF primeiramente extrai a informação numérica dos dados amostrais, analisando-a, na seqüência, à luz do conjunto interno de regras heurísticas fundamentadas em conhecimento, transformando-a, finalmente, em declarações lingüísticas de decisão. Nesse sentido, o programa SEAF, em interface com a s*hell* FuzzyCLIPS, emprega a tecnologia de inteligência artificial e elementos de lógica difusa na emulação dos princípios de raciocínio de um especialista humano ao selecionar uma distribuição de probabilidades para proceder à análise de freqüência local de variáveis hidrológicas. No julgamento da plausibilidade de uma dada distribuição, o sistema baseia-se em um conjunto de regras fundamentadas em conhecimento contemporâneo que constituem diretrizes razoavelmente similares àquelas empregadas por um especialista humano ao selecionar uma distribuição de probabilidades para uso na análise de freqüência hidrológica.

Para descrever a variabilidade presente nos dados amostrais, bem como para inferir quanto à estimação de parâmetros e às características distributivas de forma, foram empregados o método dos momentos-L e os quocientes de momentos-L. O emprego dessas estatísticas encontrou justificativa em argumentos sobre sua relativa superioridade, argumentos estes que têm sido enfocados na literatura especializada recente e foram objetos de revisão no Capítulo II e Anexo IV dessa dissertação. No SEAF, a plausibilidade das distribuições candidatas de 2 e 3 parâmetros, quantificada por meio da atribuição de um nível de confiança, é feita com base, primeiramente, nas respectivas propriedades da assimetria-L e curtose-L obtidas por simulação de Monte Carlo. O raciocínio adotado no SEAF prossegue com o teste de aderência de Filliben, o qual serve para atribuir um novo grau de confiança, independente do anterior, para a distribuição em análise; ressalte-se aqui que, durante o processo, o nível de confiança previamente atribuído a cada uma das distribuições candidatas não é alterado, o que faz com que o modo de raciocínio possa ser considerado indutivo e monotônico. Finalmente, a combinação matemática de todos os níveis de confiança já atribuídos e do número de parâmetros estimados fornece o critério de parcimônia estatística para discriminar entre distribuições da mesma família. Tal como descrito no Capítulo IV, o SEAF classifica as distribuições que não foram descartadas durante o decurso da análise por ordem decrescente de seus respectivos níveis de confiança médios.

Tal como descrito no Capítulo V, o sistema teve seu desempenho analisado por meio de dois experimentos. No primeiro, o SEAF foi aplicado a 20 amostras relativamente longas de alturas diárias de chuva e vazões médias diárias máximas anuais observadas em estações pluviométricas e fluviométricas localizadas na região sudeste do Brasil, as quais foram também submetidas à análise de freqüência por parte de um painel de peritos. Em um segundo experimento, aplicou-se o sistema SEAF a amostras de tamanho típico retiradas de populações sintéticas correspondentes às distribuições candidatas. Conforme demonstram os resultados apresentados no Capítulo V, em ambos os experimentos o sistema SEAF desempenhou-se satisfatoriamente. Não obstante a falta de consenso entre especialistas e as complexidades inerentes à análise de freqüência local de eventos máximos anuais de variáveis hidrológicas, é uma conclusão geral desse trabalho que o sistema SEAF, quando utilizado em amostras de tamanho não muito inferior, digamos, a 30 anos, é capaz de

fornecer importantes diretrizes a hidrólogos não-especialistas, sobre a escolha de uma distribuição de probabilidades apropriada, entre um conjunto de possíveis modelos candidatos. Cabe lembrar, entretanto, que as regras heurísticas para a seleção de uma distribuição de probabilidades, implementadas no sistema SEAF, representam aproximações de um certo padrão de raciocínio que reflete as convicções e as preferências do autor deste trabalho; dada a subjetividade, sempre presente em tais regras e inerente à problemática em si, outras convicções e preferências podem ser implementadas em outros sistemas semelhantes. De qualquer modo, o presente trabalho vem demonstrar, por meio do protótipo SEAF, a possibilidade de construção de um sistema de auxílio à decisão para a análise de freqüência local de eventos hidrológicos máximos anuais, com benefícios prováveis no que concerne a correta prescrição de variáveis hidrológicas características para o projeto e a operação de estruturas de aproveitamento de recursos hídricos, assim como a mitigação dos danos devidos às cheias.

## VI.2 – RECOMENDAÇÕES

Os resultados da presente dissertação representam um primeiro passo para a construção de um sistema especialista completo para análise de freqüência de variáveis hidrológicas. Um futuro sistema completo poderia contemplar decisões interdependentes sobre:

(i) se a análise deve ser local ou regional, com base no cotejo entre a quantidade de informações disponíveis e o tempo de retorno até o qual os quantis devem ser estimados;

(ii) se a análise deve se processar com base em séries anuais ou parciais e, nesse último caso, para qual valor limiar de referência as excedências deverão ser consideradas;

(iii) quais outros modelos poderiam pertencer ao elenco de distribuições candidatas;

(iv) o aperfeiçoamento de metodologias para identificação de *ouliers* e mecanismos de definição de ações para sua permanência ou retirada da amostra;

(v) a inclusão de informações históricas úteis para a redefinição de probabilidades empíricas com menor grau de incerteza;

(vi) uma vez que há divergência entre os padrões de raciocínios adotados entre os especialistas, implementar um sistema que agrupe vários padrões de análises e que dê ao usuário a liberdade em decidir qual o padrão de raciocínio será utilizado em suas análises;

(vii) o desenvolvimento de critérios mais objetivos para a definição das formas das funções de pertinência usadas para atribuição de níveis de confiança;

(viii) a flexibilização à escolha do usuário, ou mesmo a definição, dos limites inferiores acima dos quais a pertinência é atribuída;

(ix) uma melhor análise dos efeitos da adoção do critério de monotonicidade ao padrão de raciocínio do sistema;

(x) o desenvolvimento de uma *interface* mais informativa sobre a extensão e conseqüências de cada uma das decisões que estão sendo tomadas pelo sistema, de modo a garantir maior envolvimento e maior elaboração da análise em curso, por parte do usuário, entre outras;

(xi) as maiores dificuldades encontradas durante a condução deste trabalho foram relativas à maneira como cada especialista interpretou as orientações dadas na carta de encaminhamento e a obtenção de suas respostas. Tal fato pode estar ligado a vários fatores intrínsecos à própria natureza das consultas realizadas. Sendo assim, é importante em trabalhos desta natureza que o painel de especialistas seja composto por pessoas de fácil acesso, para que quaisquer dúvidas, de ambas as partes envolvidas (membros do painel e coordenador das atividades), sejam esclarecidas em tempo hábil. Vale lembrar que o grau de comprometimento de cada membro do painel com a execução do trabalho é um fator fundamental e determinante ao bom andamento de todo o processo e que as suas respostas formam uma base de comparação e avaliação dos resultados do sistema. Portanto, recomenda-se que no desenvolvimento de trabalhos similares sejam realizadas entrevistas com os membros do painel, para que se forneça um melhor esclarecimento de quais são os objetivos do trabalho e como se pretende alcançá-los.

## REFERÊNCIAS BIBLIOGRÁFICAS


AKAIKE, H., A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, v. 19, n. 6, p. 716-721, 1974.

BARR, A. e E. A. FEIGENBAUM, *The handbook of artificial intelligence*, HeuristTech Press, William Kanfmann Inc., Los Altos, Califórnia, 1981.

BENSON, M. A., *Evolution of the methods for evaluating the occurrence of floods*, USGS Water Resources Paper 1580-A, 1960.

BOBÉE, B., The log Pearson type 3 distribution and its application in hydrology. *Water Resources Research*, v.11, n.5, p. 681-689, 1975.

BOBÉE, B. e P. RASMUSSEN, Recent advances in flood frequency analysis. U.S. National Report to IUGG, 1991-1994, R*ev. Geophysics*, v. 33 Suppl. (http://earth.agu.org/revgeophys/bobee01/bobee01.htm), 1995.

BONISSONE, P. P. e R. M. TONG, Editorial: reasonings with uncertainty in expert systems, *International Journal of Man-Machine Studies*, n. 22, p. 241-250, 1985.

BOUGHTON, W.C. *A frequency distribution for annual floods.* Water Resources Research, v. 16, p. 347-354, 1980.

CHOW, K. C. A., *Towards a knowledge-based expert system for flood frequency analysis*, Tese de Doutoramento (PhD), Queen's University, Kingston, Ontario, Canadá, 1988.

CHOW, V. T., The log probability law and its engineering applications. *Proceedings ASCE*, V. 80(536), p. 1-25, 1954.

CLARKE, R. T., *Statistical Modelling in Hydrology*, J. Wiley and Sons, Chichester, Inglaterra, 1994.

CLARKE, R. T., Estimating trends in data from the Weibull and a generalized extreme value distribution, *Water Resources Research*, v. 38, n. 6, p. 25.1-25.10, 2002.

COX D. R., V. S. ISHAM e P. J. NORTHROP, *Floods: some probabilistic and statistical approaches*, Research Report 224, University College London, Londres, 2002.

DAVIS, E. G. e M. C. NAGHETTINI, *Estudo de Chuvas Intensas*, Projeto Rio de Janeiro, CPRM, Belo Horizonte, 2001.

FEIGENBAUM, E., The Age of Intelligent Machines: Knowledge Processing--From File Servers to Knowledge Servers, 1990. (disponível pela URL http://www.kurzweilai.net/frame.html?main=/articles/art0098.html)

FILLIBEN, J.J., The probability plot correlation coefficient test for normality, *Technometrics*, v. 17, n. 1, p. 111-117, 1975.

FREUND, J. E., *Mathematical Statistics*, Prentice Hall, Englewood Cliffs, New Jersey, 1962.

GIARRATANO, J. C. *CLIPS user's guide*: version 6.10, Boston: PWS Publishing Company, 1998.

GREENWOOD, J.A., J. M., LANDWEHR, N. C., MATALAS, e J. R. WALLIS, Probability weighted moments: definition and relation to parameters expressible in inverse form. *Water Resources Research*, v. 15, n. 5, p.1049-1054, 1979.

GRUBBS, F. E. e G. BECK, Extension of sample sizes and percentage points for significance tests of outlying observations, *Technometrics*, v. 14, n. 4, p. 847-854, 1972.

GUMBEL, E. J., *Statistics of Extremes.* Columbia University Press, New York, 1958.

HAAN, C. T., *Statistical Methods in Hydrology*, The Iowa State University Press, Ames, Iowa, 1977.

HAWKINS, D. M., *Identification of outliers*. Monographs on Applied Probability and Statistics, Chapman and Hall, New York, 1980.

HOSKING, J. R. M. e J. R WALLIS, Parameter and quantile estimation for the generalized Pareto distribution, *Technometrics*, V. 29, p. 339-349, 1987.

HOSKING, J. R. M. e J. R WALLIS, *Regional Frequency Analysis - An Approach Based on L-Moments*, Cambridge University Press, Cambridge, Reino Unido, 1997.

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*, v. 52, p. 105-124, 1990.

HOSKING, J. R. M., Some theoretical results concerning L-moments. *IBM Research Report*, RC 14492, IBM Research Division, Yorktown Heights, NY, EUA, 1989.

INSTITUTION OF ENGINEERS, *Australian rainfall and runoff: flood analysis and design*, Institution of Engineers Australia, Sydney, Austrália, 1977.

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

KOBSA, A., Knowledge representation: a survey of its mechanisms, a sketch of its semantics, *Cybernetics and Systems*, n. 15, p. 41-89, 1984.

LANDWEHR, J. M., N. C. MATALAS e J. R. WALLIS, Probability-weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. *Water Resources Research*, v. 15, n. 5, p. 1055-1046, 1979.

LAURSEN, E. M., Comment on "Paleohydrology of southwestern Texas" by R. C. Kochel, V. R. Baker e P. C. Patton. *Water Resources Research*, v.19, p.1339, 1983.

LINDGREN, B. W., *Statistical Theory*, Collier-MacMillan, Toronto, Ontario, Canadá, 1968.

MOOD, A. M., F. A. GRAYBILL e D. C. BOES, *Introduction to the theory of statistics*, McGraw-Hill, New York, 1974.

MUHARA, G., Selection of flood frequency model in Tanzania using L-moments and the region of influence approach, in *Second WARFSA/WaterNet Symposium: Integrated Water Resources Management - Theory, Practice, Cases*, Cape Town, África do Sul, 2001.

NAGHETTINI, M., K. W. POTTER e T. ILLANGASEKARE, Estimating the upper-tail of flood-peak frequency distributions using hydrometeorological information. *Water Resources Research*, v.32, n.6, p.1729-1740, 1996.

NEWELL, A. e H. A. SIMON, *Human problem solving*, Prentice-Hall, Englewood Cliffs, New Jersey, 1972.

NRC, *Estimating Probabilities of Extreme Floods*. National Research Council, National Academy Press, Washington, 1987.

PEEL, M. C., Q. J. WANG, R. M. VOGEL e T. A. MCMAHON, The utility of L-moment ratio diagrams for selecting a regional probability distribution, *Hydrological Sciences Journal*, v. 46, n. 1, p. 147-155, 2001.

PERICHI, L. R. e I. RODRÍGUEZ-ITURBE, On the statistical analysis of floods, in *A Celebration of Statistics*, ed. A C. Atkinson e S. E. Fienberg, Springer-Verlag, New York, p. 511-541, 1985.

PILON, P. J., *Consolidated Frequency Analysis – CFA, User Manual for Version 1*, Environment Canada, Ottawa, 1985.

POST, E., Formal reduction of the general combinatorial problem, *American Journal of Mathematics*, n. 65, p. 197-268, 1943.

POTTER, K. W., Research on flood frequency analysis : 1983-1986. *Rev. Geophys.* , V. 26, n. 3, p. 113-118, 1987.

RAO A. R. e K. H. HAMED, *Flood Frequency Analysis*, CRC Press, Boca Raton, Flórida, 2000.

REICH, B. M., Lysenkoism in U. S. flood determinations, *AGU Surface Runoff Committee – Session on flood frequency methods*, San Francisco, CA, 13 pp., 1977.

SHORTLIFFE, E. H. e B. G. BUCHANAN, A model of inexact reasoning in medicine, *Mathematicam Biosciences*, n. 23, p. 351-379, 1975.

SMITH, R.L., Threshold methods for sample extremes, in *Statistical Extremes and Applications*, editor J. Tiago de Oliveira, p. 621-638, Reidel Dordrecht, Holanda, 1984.

STEDINGER, J. R., R. M. VOGEL e E. FOUFOULA-GEORGIOU, Frequency analysis of extreme events, Chapter 18 in *Handbook of Hydrology*, ed. D. R. Maidment, McGraw-Hill, New York, 18.1-18.66, 1993.

TUCCI, C. E. M., *Regionalização de Vazões*, Editora da Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.

U. S. WATER RESOURCES COUNCIL (USWRC), A Uniform technique for determining flood flow frequency – Bulletin 15, Hydrology Committee, USWRC, Washington, 1967.

U. S. WATER RESOURCES COUNCIL (USWRC), *Guidelines for determining flood flow frequency – Bulletin 17*, USWRC Hydrology Committee, Washington, 1976.

VOGEL, R. M. e N. M. FENNESSEY, L-moment diagrams should replace product moment diagrams, Water Resources Research, v. 29, n. 6, p. 1745-1752, 1993.

WATT, W. E., K. W. LATHEM, C. R. NEILL, T.L. RICHARDS e J. ROUSSELLE, The hydrology of floods in Canada: A guide to planning and design, National Research Council of Canada, 1988.

WINSTON, P. H., *Artificial Intelligence*, Terceira Edição, Addison Wesley, Reading, Massachusetts, 1992.

YEVJEVICH, V, *Stochastic Processes in Hydrology*, Water Resources Publications, Fort Collins, Colorado, 1972.

ZADEH, L. A., Fuzzy sets, *Information and Control*, n. 8, p. 338-353, 1965.

ZVI, A. B. e B. AZMON, Joint use of L-moment diagram and goodness-of-fit test: a case study of diverse series, *Journal of Hydrology*, 198, p. 245-259, 1997.

## ANEXO – A1

## ALGUMAS DISTRIBUIÇÕES DE PROBABILIDADES UTILIZADAS EM ANÁLISE DE FREQÜÊNCIA DE VARIÁVEIS HIDROLÓGICAS

Para cada uma das distribuições de probabilidades listadas neste anexo, são fornecidos: a função densidade de probabilidades *f*(x), a função de distribuição de probabilidades acumulada *F*(x), a função para cálculo do quantil x(*F*), a expressão que fornece os valores dos momentos-L em função dos parâmetros da distribuição e a função que fornece os parâmetros da distribuição em função dos momentos-L.

### A1.1 – DISTRIBUIÇÃO NORMAL*

#### A1.1.1 - Definição

Parâmetros (2): μ (posição), σ (escala).

Amplitude de x: $-\infty < x < \infty$.

$$f(x) = s^{-1} f\left(\frac{x-m}{s}\right) \tag{A1.1}$$

$$F(x) = \Phi\left(\frac{x-m}{s}\right) \tag{A1.2}$$

onde

$$f(x) = \frac{e^{-\frac{1}{2}x^2}}{(2p)^{\frac{1}{2}}}, \qquad \Phi(x) = \int_{-\infty}^{x} f(t)dt \tag{A1.3}$$

x(*F*) não possui forma analítica explícita.

#### A1.1.2 – Momentos-L

$\lambda_1 = \mu$ (A1.4)

$\lambda_2 = 0{,}5642\sigma$ (A1.5)

$\tau_3 = 0$ (A1.6)

$\tau_4 = 0{,}1226$ (A1.7)

### A1.1.3 – Parâmetros

$$\mu = \lambda_1 \quad \text{(A1.8)}$$

$$\sigma = 1{,}7725\lambda_2 \quad \text{(A1.9)}$$

## A1.2 – DISTRIBUIÇÃO DE GUMBEL*

### A1.2.1 - Definição

Parâmetros (2): $\xi$ (posição), $\alpha$ (escala).

Amplitude de x: $-\infty < x < \infty$.

$$f(x) = a^{-1} \exp\left[-(x-x)/a\right] \exp\left\{-\exp\left[-(x-x)/a\right]\right\} \quad \text{(A1.10)}$$

$$F(x) = \exp\left\{-\exp\left[-(x-x)/a\right]\right\} \quad \text{(A1.11)}$$

$$x(F) = x - a \ln\left[-\ln(F)\right] \quad \text{(A1.12)}$$

### A1.2.2 – Momentos-L

$$\lambda_1 = \xi + \alpha\gamma \quad \text{(A1.13)}$$

$$\lambda_2 = \alpha \ln(2) \quad \text{(A1.14)}$$

$$\tau_3 = 0{,}1699 \quad \text{(A1.15)}$$

$$\tau_4 = 0{,}1504 \quad \text{(A1.16)}$$

### A1.2.3 – Parâmetros

$$\alpha = \lambda_2 / \ln(2) \quad \text{(A1.17)}$$

$$\xi = \lambda_1 - \gamma\alpha \quad \text{(A1.18)}$$

## A1.3 – DISTRIBUIÇÃO EXPONENCIAL*

### A1.3.1 - Definição

Parâmetros (2): $\xi$ (posição; limite inferior de X), $\alpha$ (escala).

Amplitude de x: $\xi \leq x < \infty$.

$$f(x) = a^{-1} \exp\left[-(x-x)/a\right] \quad \text{(A1.19)}$$

$$F(x) = 1 - \exp\left[-(x-x)/a\right] \quad \text{(A1.20)}$$

$$x(F) = x - a \ln(1-F) \quad \text{(A1.21)}$$

### A1.3.2 – Momentos-L

$$\lambda_1 = \xi + \alpha \quad \text{(A1.22)}$$

$$\lambda_2 = \alpha / 2 \quad \text{(A1.23)}$$

$$\tau_3 = 0{,}3333... \quad \text{(A1.24)}$$

$$\tau_4 = 0{,}1666... \quad \text{(A1.25)}$$

### A1.3.3 – Parâmetros

Caso o valor de $\xi$ seja conhecido, $\alpha = \lambda_1 - \xi$ e os estimadores pelos momentos-L, momentos convencionais e de máximo verossimilhança são idênticos. Se $\xi$ não for conhecido, os parâmetros são dados por:

$$\alpha = 2\,\lambda_2 \quad \text{(A1.26)}$$

$$\xi = \lambda_1 - \alpha \quad \text{(A1.27)}$$

## A1.4 – DISTRIBUIÇÃO GENERALIZADA DE VALORES EXTREMOS*

### A1.4.1 - Definição

Parâmetros (3): $\xi$ (posição), $\alpha$ (escala), $\kappa$ (forma)

Amplitude de x: $-\infty < x \le \xi + \alpha/\kappa$ se $\kappa > 0$; $-\infty < x < \infty$ se $\kappa = 0$; $\xi + \alpha/\kappa \le x < \infty$ se $\kappa < 0$.

$$f(x) = a^{-1} \exp\left[-(1-k)y - \exp(-y)\right] \quad \text{(A1.28)}$$

$$F(x) = \exp\left[-\exp(-y)\right] \quad \text{(A1.29)}$$

onde

$$y = \begin{cases} -k^{-1} \ln\left[1 - k(x - x)/a\right], & k \neq 0 \\ (x - x)/a, & k = 0 \end{cases} \quad \text{(A1.30)}$$

$$x(F) = \begin{cases} x + a\left\{1 - \left[-\ln(F)\right]^k\right\}/k, & k \neq 0 \\ x - a\,\ln\left[-\ln(F)\right], & k = 0 \end{cases} \quad \text{(A1.31)}$$

Casos especiais: $\kappa = 0$, distribuição de Gumbel; $\kappa = 1$, inverso da distribuição exponencial, onde $[1 - F(-x)]$ é a função de distribuição de probabilidades acumulada de uma distribuição exponencial.

As distribuições de probabilidades de valores extremos máximos são freqüentemente classificadas segundo três tipos de função de distribuição acumuladas:

Tipo-I ou Gumbel: $F(x) = \exp[\exp(-x)]$, $-\infty < x < \infty$, (A1.32)

Tipo-II ou Fréchet: $F(x) = \exp(-x^{-\delta})$, $0 \le x < \infty$, (A1.33)

Tipo-III ou Weibull: $F(x) = \exp(-|x|^{\delta})$, $-\infty < x \le 0$, (A1.34)

A distribuição Generalizada de Valores Extremos engloba os três tipos de distribuições de valores extremos, descritos acima, em conformidade com o sinal de seu parâmetro de forma: $\kappa=0$, Tipo-I; $\kappa<0$, Tipo-II ; $\kappa>0$, Tipo-III.

### A1.4.2 – Momentos-L

Os momentos-L são definidos para valores de $\kappa>-1$.

$\lambda_1 = \xi + \alpha[1 - \Gamma(1 + \kappa)] / \kappa$ (A1.35)

$\lambda_2 = \alpha(1 - 2^{-\kappa})\Gamma(1 + \kappa) / \kappa$ (A1.36)

$\tau_3 = 2(1 - 3^{-\kappa}) / (1 - 2^{-\kappa}) - 3$ (A1.37)

$\tau_4 = [5(1 - 4^{-\kappa}) - 10(1 - 3^{-\kappa}) + 6(1 - 2^{-\kappa})] / (1 - 2^{-\kappa})$ (A1.38)

Onde $\Gamma( . )$ representa a função gama

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t}\, dt \qquad \text{(A1.39)}$$

### A1.4.3 – Parâmetros

Não há uma solução analítica para o cálculo do parâmetro de forma ($\kappa$) e a sua estimativa é obtida pela solução númerica da equação (A1.37).

Os outros parâmetros são dados pelas seguintes equações:

$$a = \frac{l_2 k}{(1 - 2^{-k})\Gamma(1 + k)} \qquad \text{(A1.40)}$$

$$x = l_1 - a[1 - \Gamma(1 + k)] / k \qquad \text{(A1.41)}$$

## A1.5 – DISTRIBUIÇÃO GENERALIZADA DE PARETO*

### A1.5.1 - Definição

Parâmetros (3): $\xi$ (posição), $\alpha$ (escala), $\kappa$ (forma)

Amplitude de x: $\xi \leq x \leq \xi+\alpha/\kappa$ se $\kappa>0$; $\xi \leq x < 8$ se $\kappa\leq 0$.

$$f(x) = a^{-1} \exp[-(1-k)y] \tag{A1.42}$$

$$F(x) = 1 - \exp(-y) \tag{A1.43}$$

onde

$$y = \begin{cases} -k^{-1} \ln[1-k(x-x)/a], & k \neq 0 \\ (x-x)/a, & k = 0 \end{cases} \tag{A1.44}$$

$$x(F) = \begin{cases} x + a\left[1-(1-F)^{-k}\right]/k, & k \neq 0 \\ x - a \ln(1-F), & k = 0 \end{cases} \tag{A1.45}$$

Casos especiais: $\kappa = 0$, distribuição Exponencial e $\kappa = 1$, distribuição Uniforme no intervalo $\xi \leq x \leq \xi + \alpha$.

### A1.5.2 – Momentos-L

Os momentos-L são definidos para valores de $\kappa>-1$.

$$\lambda_1 = \xi + \alpha/(1+\kappa) \tag{A1.46}$$

$$\lambda_2 = \alpha / [(1+\kappa)(2+\kappa)] \tag{A1.47}$$

$$\tau_3 = (1-\kappa) / (3+\kappa) \tag{A1.48}$$

$$\tau_4 = (1-\kappa)(2-\kappa) / [(3+\kappa)(4+\kappa)] \tag{A1.49}$$

### A1.5.3 – Parâmetros

Caso o parâmetro $\xi$ seja conhecido, os parâmetros $\alpha$ e $\kappa$ são dados por:

$$\kappa = (\lambda_1 - \xi)/\lambda_2 - 2 \tag{A1.50}$$

$$\alpha = (1+\kappa)(\lambda_1 - \xi) \tag{A1.51}$$

Caso o parâmetro $\xi$ seja desconhecido, os parâmetros são dados por:

$\kappa = (1 - 3\tau_3) / (1 + \tau_3)$ (A1.52)

$\alpha = (1 + \kappa)(2 + \kappa)\lambda_2$ (A1.53)

$\xi = \lambda_1 - (2 + \kappa)\lambda_2$ (A1.54)

## A1.6 – DISTRIBUIÇÃO PEARSON III*

### A1.6.1 - Definição

Parâmetros (3): $\mu$ (posição), $\sigma$ (escala), $\gamma$ (forma)

Para $\gamma \neq 0$, faça $\alpha = 4 / \gamma^2$, $\beta = 0{,}5\,\sigma\,|\gamma|$ e $\xi = \mu - 2\sigma / \gamma$.

Amplitude:

- Se $\gamma > 0$, $\xi \leq x \leq \infty$ e

$$f(x) = \frac{(x-x)^{a-1} e^{-(x-x)/b}}{b^a \Gamma(a)} \quad \text{(A1.55)}$$

$$F(x) = G\left(a, \frac{x-x}{b}\right) / \Gamma(a) \quad \text{(A1.56)}$$

- Se $\gamma = 0$, $-\infty \leq x \leq \infty$ e

$$f(x) = f\left(\frac{x-m}{s}\right), \qquad F(x) = \Phi\left(\frac{x-m}{s}\right) \quad \text{(A1.57)}$$

onde as funções $\phi(\,.\,)$ e $\Phi(\,.\,)$ são definidas na equação (A1.3).

- Se $\gamma < 0$, $-\infty \leq x \leq ?$ e

$$f(x) = \frac{(x-x)^{a-1} e^{-(x-x)/b}}{b^a \Gamma(a)} \quad \text{(A1.58)}$$

$$F(x) = 1 - G\left(a, \frac{x-x}{b}\right) / \Gamma(a) \quad \text{(A1.59)}$$

onde $\Gamma(\,.\,)$ é a função gama definida pela equação (A1.39) e $G(\alpha,x)$ é a função gama incompleta.

$$G(a, x) = \int_0^x t^{a-1} e^{-t} dt \quad \text{(A1.60)}$$

Casos especiais: $\gamma = 2$, distribuição Exponencial; $\gamma = 1$, distribuição Normal; $\gamma = -2$, inverso da distribuição Exponencial

Não há uma forma analítica para x(*F*).

**A1.6.2 – Momentos-L**

Caso sejam usados os parâmetros padronizados, as expressões que relacionam os momentos-L aos parâmetros da distribuição Pearson III tornam-se mais simples. Os resultados apresentados aqui assumem uma parametrização padrão com $\gamma > 0$. Os resultados correspondentes a $\gamma<0$ são obtidos pela mudança dos sinais de $\lambda_1$, $\tau_3$ e $\xi$ onde eles ocorrerem nas expressões (A1.61) – (A1.68).

Os momentos-L são definidos para todos os valores de $\alpha$, $0 < \alpha < \infty$.

$$\lambda_1 = \xi + \alpha\,\beta \qquad \text{(A1.61)}$$

$$\lambda_2 = \pi^{-0,5}\,\beta\,\Gamma(\alpha + 0,5) / \Gamma(\alpha) \qquad \text{(A1.62)}$$

$$\tau_3 = 6\,I_{1/3}(\alpha, 2\alpha) - 3 \qquad \text{(A1.63)}$$

Onde $I_x(p, q)$ denota a razão da função beta incompleta.

$$I_x(p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\int_0^x t^{p-1}(1-t)^{q-1}dt \qquad \text{(A1.64)}$$

Não existe uma simples expressão para $\tau_4$. Uma função de aproximação racional pode ser utilizada como uma aproximação de $\tau_3$ e $\tau_4$, em função de $\alpha$. As seguintes aproximações tem precisão de$10^{-6}$.

Se $\alpha \geq 1$,

$$t_3 \approx a^{-0,5}\,\frac{A_0 + A_1 a^{-1} + A_2 a^{-2} + A_3 a^{-3}}{1 + B_1 a^{-1} + B_2 a^{-2}} \qquad \text{(A1.65)}$$

$$t_4 \approx \frac{C_0 + C_1 a^{-1} + C_2 a^{-2} + C_3 a^{-3}}{1 + D_1 a^{-1} + D_2 a^{-2}} \qquad \text{(A1.66)}$$

Se $\alpha < 1$,

$$t_3 \approx \frac{1 + E_1 a^1 + E_2 a^2 + E_3 a^3}{1 + F_1 a^1 + F_2 a^2 + F_3 a^3} \quad \text{(A1.67)}$$

$$t_4 \approx \frac{1 + G_1 a^1 + G_2 a^2 + G_3 a^3}{1 + H_1 a^1 + H_2 a^2 + H_3 a^3} \quad \text{(A1.68)}$$

Tabela A1.1: Coeficientes das funções de aproximação de $\tau_3$ e $\tau_4$.

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| A | 3,2573501 x $10^{-1}$ | 1,6869150 x $10^{-1}$ | 7,8327243 x $10^{-2}$ | -2,9120539 x $10^{-3}$ |
| B | - | 4,6697102 x $10^{-1}$ | 2,4255406 x $10^{-1}$ | - |
| C | 1,2260172 x $10^{-1}$ | 5,3730130 x $10^{-2}$ | 4,3384378 x $10^{-2}$ | 1.1101277 x $10^{-2}$ |
| D | - | 1,8324466 x $10^{-1}$ | 2,0166036 x $10^{-1}$ | - |
| E | - | 2,3807576 | 1,5931792 | 1,1618371 x $10^{-1}$ |
| F | - | 5,1533299 | 7,1425260 | 1,9745056 |
| G | - | 2.1235833 | 4,1670213 | 3,1925299 |
| H | - | 9,0551443 | 2,6649995 x $10^{-1}$ | 2,6193668 x $10^{-1}$ |

Fonte: Hosking e Wallis (1997).

### A1.6.3 – Parâmetros

Para estimar o valor de $\alpha$, a equação (A1.63) deve ser resolvida para $\alpha$, substituindo o valor de $\tau_3$ pelo $|\tau_3|$. As seguintes expressões de aproximação de $\alpha$ tem precisão relativa maior que 5 x $10^{-5}$ para todos os valores de $\alpha$.

Se $0 < |\tau_3| < 1/3$, faça $z = 3\pi\tau_3^2$ e use

$$a \approx \frac{1 + 0{,}2906z}{z + 0{,}1882z^2 + 0{,}0442z^3} \quad \text{(A1.69)}$$

Se $1/3 \leq |\tau_3| < 1$, faça $z = 1 - |\tau_3|$ e use

$$a \approx \frac{0{,}36067z - 0{,}59567z^2 + 0{,}25361z^3}{1 - 2{,}78861z + 2{,}56096z^2 - 0{,}77045z^3} \quad \text{(A1.70)}$$

Encontrado $\alpha$, os outros parâmetros são dados por:

$$g = 2a^{-1/2} sign(t_3) \quad \text{(A1.71)}$$

$$s = l_2 p^{1/2} a^{1/2} \Gamma(a) / \Gamma(a + 0{,}5) \quad \text{(A1.72)}$$

$$m = l_1 \quad \text{(A1.73)}$$

## A1.7 – DISTRIBUIÇÃO LOG-NORMAL DE 2 PARÂMETROS*

A distribuição Log-Normal de 2 parâmetros da variável X refere-se à distribuição Normal da variável transformada Y=ln(X) ou Y=log(X).

## A1.8 – DISTRIBUIÇÃO LOG-PEARSON III*

A distribuição Log-Pearson III da variável X refere-se à distribuição Pearson III da variável transformada Y=ln(X) ou Y=log(X).

* Remete-se o leitor à referência Rao e Hamed (2000) ("Flood Frequency Analysis") para detalhes completos sobre as distribuições aqui listadas e outras distribuições, incluindo estimação de parâmetros, quantis e intervalos de confiança pelos métodos dos momentos convencionais, do máximo de verossimilhança e dos momentos-L.

## ANEXO – A2

## TESTES NÃO PARAMÉTRICOS

### A2.1 – TESTE DE MANN-KENDALL

O teste de Mann-Kendall é utilizado para a identificação de heterogeneidade ou tendência em uma série temporal, sem especificar se a tendência é ou não linear.

Seja uma série de dados anuais $X_t$, t=1,....,N. Compare cada valor de $X_{t'}$, t'=1,...,N-1, com todos os subsequentes valores de $X_t$, t=t'+1,...,N, e crie uma nova série de dados $Z_k$, conforme:

$$\begin{aligned} Z_k &= 1 \quad \text{if } X_t > X_{t'} \\ Z_k &= 0 \quad \text{if } X_t = X_{t'} \\ Z_k &= -1 \quad \text{if } X_t < X_{t'} \end{aligned} \tag{A2.1}$$

onde

$$k = \frac{(t'-1)(2N-t')}{2} + (t-t') \tag{A2.2}$$

A estatística do teste de Mann-Kendall é dada pela soma de todos os valores da nova série $Z_k$,, ou seja

$$S = \sum_{t'=1}^{N-1} \sum_{t=t'+1}^{N} Z_k \tag{A2.3}$$

Esta estatística representa o número de diferenças positivas menos o número de diferenças negativas, para todas as diferenças consideradas.

Para N>40, o teste prossegue a partir da estatística

$$u_c = \frac{S+m}{\sqrt{V(S)}} \tag{A2.4}$$

$$V(S) = \frac{1}{18}\left[ N(N-1)(2N+5) - \sum_{i=1}^{n} e_i (e_i - 1)(2e_i + 5) \right] \tag{A2.5}$$

onde:

$$m=1 \quad \text{if } S<0$$

$$m=-1 \quad \text{if } S>0 \qquad \text{(A2.6)}$$

- n é o número de grupos de valores repetidos.
- $e_i$ é o número de dados repetidos em cada grupo i.

A estatística $u_c$ é assumida como zero, se S=0. A hipótese de tendência crescente ou decrescente não pode ser rejeitada, a um nível de significância $\alpha$ adotado, se $|u_c| > u_{1-\alpha/2}$, onde $u_{1-\alpha/2}$ é o quantil correspondente à probabilidade $1-\alpha/2$ da distribuição normal padrão. Kendall indica que este teste pode ser utilizado na identificação de tendência em amostras de pequeno tamanho, com aproximadamente 10 valores, desde que a amostra não contenha muitos valores repetidos.

## A2.2 – TESTE DO COEFICIENTE DE CORRELAÇÃO DE KENDALL

Esse teste não-paramétrico é útil para a identificação de eventual dependência serial em uma dada amostra ou entre duas amostras. Uma medida geral e eficiente da medida de correlação entre duas variáveis é o coeficiente de correlação de Kendall, geralmente denotado por $\tau$. Ele é baseado no ordenamento dos dados, o que o torna resistente ao efeito da presença de *outliers* na amostra e a desvios de uma relação não linear entre os dados.

A seguir, são apresentados os passos para o cálculo do coeficiente de correlação de Kendall, assim como o respectivo teste de hipótese para verificação de dependência entre os dados amostrais. Aqui, a hipótese nula, $H_0$, refere-se à suposição de que a distribuição de y não varia em função dos valores de x, sendo y considerada a variável dependente; no caso de uma única amostra, os valores de y correspondem aos valores de x defasados de um intervalo de tempo.

- Ordene os n pares $(x_1,y_1)$, $(x_2,y_2)$, ... , $(x_n,y_n)$ indexados de acordo com a magnitude dos valores de x, tal que $x_1 \leq x_2 \leq ... \leq x_n$ e $y_i$ é o valor da variável dependente correspondente a $x_i$;

- Examine todos os n(n-1)/2 pares ordenados dos valores de $y_i$. Faça P ser o número de casos para os quais $y_i > y_j$ (i>j) e faça M ser o número de casos para os quais $y_i < y_j$ (i>j);

- Calcule a estatística de teste S=P-M;

- Para n>10, o teste é conduzido usando a aproximação normal para a estatística S. A variável normal padrão Z é calculada por:

$$Z = \begin{cases} \dfrac{S-1}{\sqrt{Var(S)}} \Leftrightarrow S > 0 \\ 0 \Leftrightarrow S = 0 \\ \dfrac{S+1}{\sqrt{Var(S)}} \Leftrightarrow S < 0 \end{cases} \qquad \text{(A2.7)}$$

onde

$$Var(S) = n(n-1)(2n+5)/18. \qquad \text{(A2.8)}$$

- A hipótese nula é rejeitada, a um nível de significância $\alpha$, se $|Z| > Z_{(1-\alpha/2)}$, onde $Z_{(1-\alpha/2)}$ é o valor da variável normal padrão com probabilidade de excedência igual a $\alpha/2$. Caso existam valores repetidos de x e/ou de y na amostra de dados a fórmula para o cálculo da Var(S) deve ser modificada para:

$$Var(S) = \frac{n(n-1)(2n+5) - \sum_{i=1}^{n} t_i i(i-1)(2i+5)}{18} \qquad \text{(A2.9)}$$

onde

$t_i$ é o número de grupos com i valores repetidos.

- O coeficiente de correlação de Kendall $\tau$ é definido como:

$$t = \frac{S}{n(n-1)/2} \qquad \text{(A2.10)}$$

Da mesma forma que os outros coeficientes de correlação, o valor de $\tau$ varia entre -1 e 1. Sinal negativo do coeficiente $\tau$ indica relacionamento decrescente entre as variáveis. Já a magnitude indica o grau de relacionamento, ou seja, valores absolutos de t próximos a 1 indicam um forte relacionamento entre as variáveis.

## ANEXO – A3

## TESTES PARA VERIFICAÇÃO DE ADERÊNCIA

### A3.1 – TESTE DO QUI-QUADRADO

O teste do Qui-Quadrado é o mais comum para verificação da aderência ou qualidade do ajuste dos dados amostrais a uma distribuição de probabilidades específica. Seu fundamento está na comparação entre o número real de observações e o número teórico esperado de observações, dentro de um intervalo de classe. O número esperado de observações é calculado pelo produto da freqüência relativa do respectivo intervalo de classe, obtida pela função de probabilidades teórica, pelo número total de observações. O valor da estatística do teste é calculado pela seguinte formula:

$$c_c^2 = \sum_{j=1}^{k} \left(O_j - E_j\right)^2 / E_j \qquad \text{(A3.1)}$$

onde

k é o número de intervalos de classe;

$O_j$ é o número real de observações no j-ésimo intervalo de classe;

$E_j$ é o número teórico esperado de observações (de acordo com a distribuição de probabilidade testada) no j-ésimo intervalo de classe.

A estatística $\chi^2_c$ tem uma distribuição Qui-Quadrado com k-p-1 graus de liberdade, onde p é o número de parâmetros estimados para a distribuição de probabilidades testada. A hipótese de que os dados tenham sido retirados de uma distribuição especifica é rejeitada se

$$\chi^2_c > \chi^2_{1-\alpha,\, k-p-1} \qquad \text{(A3.2)}$$

onde

$\alpha$ é o nível de significância do teste.

O teste Qui-Quadrado é muito sensível nas caudas da distribuição e uma simples alteração do número ou largura dos intervalos de classe pode alterar o resultado do teste. Portanto, recomenda-se combinar os intervalos classe de forma que o número esperado em cada classe nunca seja inferior a 5, ou dividir os intervalos de forma que o número esperado de observações seja igual em qualquer intervalo de classe.

## A3.2 – TESTE DE KOLMOGOROV-SMIRNOV

O teste de Kolmogorov-Smirnov é baseado na comparação da distância máxima entre a função de distribuição acumulada teórica e a distribuição empírica de probabilidades. O teste é conduzido da seguinte maneira:

- Seja $P_x(x)$ a função de distribuição acumulada teórica a ser testada, sob a hipótese $H_0$.

- Seja $S_n(x)$ a representação da distribuição empírica de probabilidades, baseada nas n observações. Para qualquer observação x, tem-se

  $$S_n(x) = k/n \qquad \text{(A3.3)}$$

  onde

  k é o número de observações menores ou iguais a x encontradas na amostra.

- Determine o máximo desvio absoluto entre $P_X(x)$ e $S_n(x)$, ou seja

  $$D = \max |P_x(x) - S_n(x)| \qquad \text{(A3.4)}$$

- Se, para o nível de significância adotado, o valor observado da estatística D for maior ou igual ao valor $D_{crit,}$ Tabela-2.1, a hipótese $H_0$ é rejeitada.

  $$D = D_{crit} \qquad \text{(A3.5)}$$

Tabela-A3.1: Valores críticos para o teste de aderência de Komogorov-Smirnov

| Tamanho da Amostra | Nível de Significância | | |
|---|---|---|---|
| | **0,10** | **0,05** | **0,01** |
| 10 | 0,368 | 0.409 | 0,486 |
| 15 | 0,304 | 0,338 | 0,404 |
| 20 | 0,264 | 0,294 | 0,352 |
| 25 | 0,240 | 0,264 | 0,320 |
| 30 | 0,220 | 0,242 | 0,290 |
| 35 | 0,210 | 0,230 | 0,270 |
| 40 | 0,193 | 0,210 | 0,250 |
| 50 | 0,173 | 0,190 | 0,230 |
| 60 | 0,158 | 0,170 | 0,210 |
| 70 | 0,146 | 0,160 | 0,190 |
| 90 | 0,129 | 0,140 | 0,172 |
| 100 | 0,122 | 0,140 | 0,163 |

Fonte: Haan (1979), p 348.

## A3.3 – TESTE DE FILLIBEN

O teste de Filliben usa como estatística o coeficiente de correlação r entre as observações ordenadas $x_{(i)}$ e os correspondentes quantis ajustados $w_i=P^{-1}(1-q_i)$, onde para cada $x_i$ é determinada uma posição de plotagem $q_i$ correspondente. O teste é conduzido da seguinte forma:

- Ordene os valores observados em ordem decrescente: $x_1>x_2>x_3>...>x_{n-1}>x_n$.
- Calcule as posições de plotagem $q_i$, para cada valor $x_i$, em função da distribuição de probabilidades testada, conforme recomendação da Tabela A3.2, adaptada de Stedinger et al. (1992).

Tabela-A3.2: Posição de plotagem para algumas distribuições de probabilidade.

| Distribuição | Posição de Plotagem | |
|---|---|---|
| | **Nome** | **Formula** |
| Normal | Blom | (i - 0,375)/(n + 0,250) |
| Lognormal 2 parâmetros | Blom | (i - 0,375)/(n + 0,250) |
| Gumbel | Gringorten | (i – 0,44)/(n + 0,12) |
| Exponencial | Gringorten | (i – 0,44)/(n + 0,12) |
| Pearson III | Blom | (i - 0,375)/(n + 0,250) |
| Log Pearson III | Blom | (i - 0,375)/(n + 0,250) |
| GEV | Cunname | (i – 0,40)/(n + 0,20) |
| GPA | Weibull | i/(n+1) |

- Calcule os quantis teóricos $w_i$ correspondestes:

  $w_i=P^{-1}(1-q_i)$ (A3.6)

  onde

  $P^{-1}$ é a inversa da função de distribuição de probabilidades testada; e

  $q_i$ é a posição de plotagem correspondente ao valor observado $x_i$.

- Calcule o coeficiente de correlação r entre $x_i$ e $w_i$:

$$r = \frac{\sum \left(x_i - \bar{x}\right)\left(w_i - \bar{w}\right)}{\sqrt{\sum \left(x_i - \bar{x}\right)^2 \sum \left(w_i - \bar{w}\right)^2}} \qquad \text{(A3.7)}$$

- Se, para o nível de significância adotado, o valor observado do coeficiente de correlação r for menor do que o valor de $r_{crit}$, a hipótese $H_0$ é rejeitada. Ou seja, neste caso, há evidências para se rejeitar a distribuição testada como modelo da distribuição populacional dos dados. As tabelas A3.3 e A3.4 apresentam os valores de $r_{crit}$ para as distribuições Normal e Gumbel, respectivamente.

Tabela-A3.3: Valores críticos mínimos para o teste de Filliben (Distribuição Normal)

| **Tamanho da Amostra** | **Nível de significância** | | |
|---|---|---|---|
| | **0,10** | **0,05** | **0,01** |
| 10 | 0,9347 | 0,9180 | 0,8804 |
| 15 | 0,9506 | 0,9383 | 0,9110 |
| 20 | 0,9600 | 0,9503 | 0,9290 |
| 30 | 0,9707 | 0,9639 | 0,9490 |
| 40 | 0,9767 | 0,9715 | 0,9597 |
| 50 | 0,9807 | 0,9764 | 0,9664 |
| 60 | 0,9835 | 0,9799 | 0,9710 |
| 75 | 0,9865 | 0,9835 | 0,9757 |
| 100 | 0,9893 | 0,9870 | 0,9812 |
| 300 | 0,99602 | 0,99525 | 0,99354 |
| 1000 | 0,99854 | 0,99824 | 0,99755 |

Fonte: Stedinger et al. (1993), p 18.28

Tabela-A3.4: Valores críticos mínimos para o teste de Filliben (Distribuição Gumbel)

| Tamanho da Amostra | Nível de significância | | |
|---|---|---|---|
| | **0,10** | **0,05** | **0,01** |
| 10 | 0,9260 | 0,9084 | 0,8630 |
| 20 | 0,9517 | 0,9390 | 0,9060 |
| 30 | 0,9622 | 0,9526 | 0,9191 |
| 40 | 0,9689 | 0,9594 | 0,9286 |
| 50 | 0,9729 | 0,9646 | 0,9389 |
| 60 | 0,9760 | 0,9685 | 0,9367 |
| 70 | 0,9787 | 0,9720 | 0,9506 |
| 80 | 0,9804 | 0,9747 | 0,9525 |
| 100 | 0,9831 | 0,9779 | 0,9596 |
| 300 | 0,9925 | 0,9902 | 0,9819 |
| 1000 | 0,99708 | 0,99622 | 0,99334 |

Fonte: Stedinger et al. (1993), p 18.28

## A3.4 – MOMENTOS-L AMOSTRAIS x TEÓRICOS

Hosking e Wallis (1997) apresentaram uma metodologia para escolha da distribuição de probabilidades regional dos dados baseada na comparação entre os momentos-L regionais e os teóricos, para cada uma das distribuições de probabilidades candidatas a modelo paramétrico regional. Apesar desta metodologia ter sido desenvolvida para a seleção de uma distribuição probabilidades no contexto da análise de freqüência regional, ela pode ser adaptada para aplicações em análise de freqüência local, pela substituição dos momentos-L regionais pelos momentos-L de uma única amostra. Uma revisão mais detalhada sobre as principais propriedades dos momentos-L pode ser encontrada no Anexo A4 da presente dissertação. Tal como adaptado à análise de freqüência local, o teste é conduzido da seguinte forma:

- Calcule os momentos-L e razões-L amostrais: $l_1^A$, $t^A$, $t_3^A$ e $t_4^A$, os quais representam respectivamente a média-L, o coeficiente de variação-L, a assimetria-L e o curtose-L;

- Ajuste a distribuição de probabilidades testada aos momentos-L e razões-L calculados a partir da amostra;

- Simule N amostras de mesmo tamanho daquele da amostra analisada pelo método de Monte Carlo, com base na distribuição ajustada;

- Calcule os momentos-L e razões-L: $l_1$, t, $t_3$, $t_4$, para cada uma das N amostras simuladas;

- Calcule o desvio padrão ($\sigma_4$) de $t_4$ em função dos N valores calculados;

- Assintoticamente, a razão-L teórica $\tau_4$ é distribuída normalmente com média $\tau_4^m$ e devio padrão $\sigma_4$, onde

  - $\tau_4^m$ é o valor teórico esperado para uma dada distribuição;
  - $\sigma_4$ é o desvio padrão calculado em função das N simulações;

  Calcule a variável normal reduzida Z:

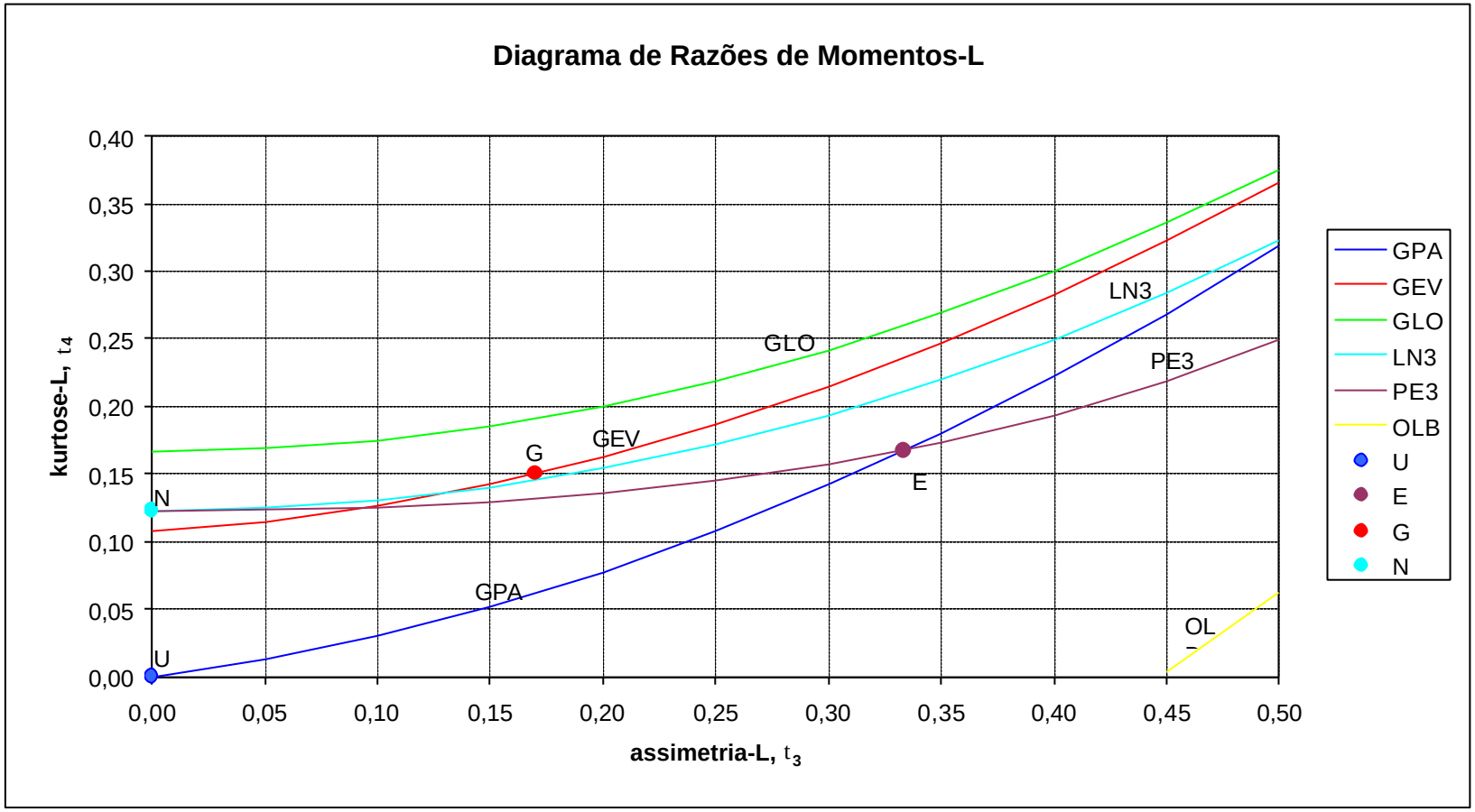
$$Z = \frac{\left(\tau_4^A - \tau_4^m\right)}{\sigma_4} \qquad \text{(A3.8)}$$

- Se, para o nível de significância de 90%, o valor | Z | > 1,64, a hipótese $H_0$ deve ser rejeitada. Em outras palavras, há evidências para se rejeitar a distribuição testada como modelo da distribuição populacional dos dados.

Em seu trabalho Hosking e Wallis (1997) admitiram que 500 simulações são suficientes para o cálculo do desvio padrão das razões-L.

Uma variação deste teste é o exame visual da plotagem do par ordenado ($t_3^A$, $t_4^A$) no Diagrama de Razões de Momentos-L. Neste caso, as razões-L amostrais são comparadas aos valores teóricos correspondentes a cada distribuição. A Figura- A3.1 apresenta o Diagrama de Razões de Momentos-L.

Figura –A3.1: Diagrama de Razões de Momentos-L




Nota: Distribuições de 2 e 3 parâmetros estão representadas por pontos e linhas respectivamente. E – Exponencial, G –Gumbel, N – Normal, U – Uniforme, GLO – Generalizada Logistica, GEV – Generalizada de Valores Extremos, GPA – Generalizada de Pareto, LN3 – Lognormal, PE3 – Pearson tipo III e OLB é o limite inferior da função $\tau_3$ x $\tau_4$.

## ANEXO – A4

## MOMENTOS-L

Esse anexo apresenta os fundamentos da teoria dos momentos-L e algumas de suas propriedades que foram utilizadas na presente dissertação. O texto que se segue foi transcrito e adaptado ao presente contexto de partes do trabalho de Davis e Naghettini (2001), com a devida permissão destes autores.

### A4.1 - CONCEITOS BÁSICOS

Além de dependentes de *n*, as estimativas com base em momentos amostrais convencionais envolvem potências sucessivas dos desvios dos dados em relação ao valor central. Em conseqüência, pequenas amostras tendem a produzir estimativas não confiáveis, particularmente para as funções de momentos de ordem superior como a assimetria e a curtose. Os momentos-L, a serem abordados a seguir, compõem um sistema de medidas estatísticas mais confiáveis para a descrição das principais características das distribuições de probabilidades.

Os momentos-L são derivados dos "momentos ponderados por probabilidades", ou simplesmente MPP's, os quais foram introduzidos na literatura científica por Greenwood et al. (1979). Os MPP's de uma variável aleatória *X*, variável essa descrita pela função de probabilidades acumuladas $F_X(x)$, são as quantidades definidas por

$$M_{p,r,s} = \mathbf{E}\left\{X^{\,p}\left[F_X(x)\right]^r\left[1-F_X(x)\right]^s\right\} \qquad \text{(A4.1)}$$

Os MPP's $\alpha_r = M_{1,0,r}$ e $\beta_r = M_{1,r,0}$ representam casos especiais de relevância particular para a inferência estatística. Com efeito, considerando-se uma distribuição cuja função de quantis seja dada por $x(p)$, pode-se combinar as equações II.4 e A4.1 para expressar $\alpha_r$ e $\beta_r$ da seguinte forma :

$$\alpha_r = \int_0^1 x(p)\,(1-p)^r\,dp \quad , \quad \beta_r = \int_0^1 x(p)\;p^r\,dp \tag{A4.2}$$

Contrastando as equações acima com a definição de momentos convencionais, ou seja, $\mathbf{E}(X) = \int_0^1 [x(p)]^r\,dp$, observa-se que esses implicam em potências sucessivamente crescentes da função de quantis $x(p)$, enquanto que $\alpha_r$ e $\beta_r$ implicam em potências sucessivamente crescentes de $p$ ou $(1-p)$; dessa forma, os MPP's $\alpha_r$ e $\beta_r$ podem ser vistos como integrais de $x(p)$, ponderadas pelos polinômios $p^r$ ou $(1-p)^r$.

Diversos autores, como Landwehr et al. (1979) e Hosking e Wallis (1987), utilizaram os MPP's $\alpha_r$ e $\beta_r$ como base para a estimação de parâmetros de distribuições de probabilidades. Hosking e Wallis (1997) ponderam, entretanto, que $\alpha_r$ e $\beta_r$ são de interpretação difícil, em termos das medidas de escala e forma de uma distribuição de probabilidades, e sugerem, para esse efeito, certas combinações lineares de $\alpha_r$ e $\beta_r$. Ainda segundo Hosking e Wallis (1997), essas combinações advém da ponderação das integrais de $x(p)$ por um conjunto de polinômios ortogonais, denotados por $P_r^*(p), r = 0,1,2,...,$ definidos pelas seguintes condições :

(i) $P_r^*(p)$ é um polinômio de grau $r$ em $p$.

(ii) $P_r^*(1) = 1$

(iii) $\int_0^1 P_r^*(p)\;P_s^*(p)\,dp = 0$ , para $r \neq s$ (condição de ortogonalidade)

Essas condições definem os polinômios de Legendre, devidamente modificados para a condição de ortogonalidade no intervalo $0 \leq p \leq 1$ e não $-1 \leq p \leq 1$, como em sua formulação original. Formalmente, esses polinômios são dados por

$$P_r^*(p) = \sum_{k=0}^{r} l_{r,k}^*\,p^k \tag{A4.3}$$

onde $l_{r,k}^* = (-1)^{r-k}\binom{r}{k}\binom{r+k}{k} = \dfrac{(-1)^{r-k}(r+k)!}{(k!)^2(r-k)!}$.

De posse das definições acima, os momentos-L de uma variável aleatória $X$ podem ser agora conceituados como sendo as quantidades

$$\lambda_r = \int_0^1 x(p)\, P_{r-1}^*(p)\, dp \qquad \text{(A4.4)}$$

Em termos dos MPP's, os momentos-L são dados por

$$\lambda_{r+1} = (-1)^r \sum_{k=0}^{r} l_{r,k}^*\, \alpha_k = \sum_{k=0}^{r} l_{r,k}^*\, \beta_k \qquad \text{(A4.5)}$$

Os primeiros quatro momentos-L são, portanto,

$$\lambda_1 = \alpha_0 = \beta_0 \text{ (\textit{média ou momento-L de posição})} \qquad \text{(A4.6)}$$

$$\lambda_2 = \alpha_0 - 2\alpha_1 = 2\beta_1 - \beta_0 \text{ (\textit{momento-L de escala})} \qquad \text{(A4.7)}$$

$$\lambda_3 = \alpha_0 - 6\alpha_1 + 6\alpha_2 = 6\beta_2 - 6\beta_1 + \beta_0 \qquad \text{(A4.8)}$$

$$\lambda_4 = \alpha_0 - 12\alpha_1 + 30\alpha_2 - 20\alpha_3 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 \qquad \text{(A4.9)}$$

Em termos de medidas de forma das distribuições, torna-se mais conveniente que os momentos-L sejam expressos em quantidades adimensionais. Essas são representadas pelos quocientes de momentos-L, dados por:

$$\tau_r = \frac{\lambda_r}{\lambda_2},\ r = 3,4,\ldots \qquad \text{(A4.10)}$$

Dessa forma, $\tau_3$ e $\tau_4$ são, respectivamente, as medidas de assimetria e curtose, independentes da escala da distribuição de probabilidades. Pode-se definir, também em termos de momentos-L, uma medida análoga ao coeficiente de variação, qual seja:

$$CV - L = \tau = \frac{\lambda_2}{\lambda_1} \qquad \text{(A4.11)}$$

## A4.2 – MOMENTOS-L E ESTATÍSTICAS DE ORDEM

Os momentos-L podem ser expressos como combinações lineares das estatísticas de ordem de uma amostra. Para esse efeito, considere uma amostra de tamanho $n$,

disposta em ordem crescente $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$, de forma que a $k$-ésima menor observação, ou estatística de ordem $k$, seja denotada por $X_{k:n}$. De forma consistente com a equação A4.4, os momentos-L da distribuição de probabilidades da qual a amostra foi retirada são dados por

$$\lambda_r = r^{-1} \sum_{j=0}^{r-1} (-1)^j \binom{r-1}{j} \mathbf{E}(X_{r-j:r}) \tag{A4.12}$$

onde a esperança matemática **E**(.) de uma estatística de ordem $r$ é o operador definido por

$$\mathbf{E}(X_{r:n}) = \frac{n!}{(r-1)!\,(n-r)!} \int_0^1 x(p)\, p^{r-1} (1-p)^{n-r}\, dp \tag{A4.13}$$

Dessa forma, os quatro primeiros momentos-L podem ter as seguintes expressões :

$$\lambda_1 = \mathbf{E}(X_{1:1}) \tag{A4.14}$$

$$\lambda_2 = \frac{1}{2} \mathbf{E}(X_{2:2} - X_{1:2}) \tag{A4.15}$$

$$\lambda_3 = \frac{1}{3} \mathbf{E}(X_{3:3} - 2X_{2:3} + X_{1:3}) \tag{A4.16}$$

$$\lambda_4 = \frac{1}{4} \mathbf{E}(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}) \tag{A4.17}$$

## A4.3 – PROPRIEDADES DOS MOMENTOS-L

Hosking (1989, 1990) apresenta as provas matemáticas para as seguintes propriedades dos momentos-L :

- Existência : se a média de uma distribuição existe, então todos os momentos-L existem.
- Singularidade : se a média de uma distribuição existe, então os momentos-L a definem singularmente.
- Valores Limites :

$-\infty \le \lambda_1 \le \infty$.

$\lambda_2 \ge 0$

se a distribuição é definida somente para $X \ge 0 \Rightarrow 0 \le \tau \le 1$.

$|\tau_r| < 1$ para $r \ge 3$.

$\frac{1}{4}(5\tau_3^2 - 1) \le \tau_4 \le 1$.

se a distribuição é definida somente para $X \ge 0 \Rightarrow 2\tau - 1 \le \tau_3 \le 1$.

- Transformações Lineares :

Se $X$ e $Y = aX + b$ são duas variáveis aleatórias de momentos-L $\lambda_r$ e $\lambda_r^*$, respectivamente, então são válidas as seguintes relações :

$\lambda_1^* = a\lambda_1 + b$;

$\lambda_2^* = |a|\lambda_2$ ; e

- Simetria : se $X$ é uma variável aleatória, descrita por uma distribuição de probabilidades simétrica, então todos os quocientes de momentos-L de ordem impar ($\tau_r$, $r$=3,5, ...) serão nulos.

## A4.4 – MOMENTOS-L AMOSTRAIS

A estimação dos MPP's e momentos-L, a partir de uma amostra finita de tamanho $n$, inicia-se com a ordenação de seus elementos constituintes em ordem crescente, ou seja $x_{1:n} \le x_{2:n} \le \ldots \le x_{n:n}$. Um estimador não-enviesado do MPP $\beta_r$ pode ser escrito como

$$b_r = \hat{\beta}_r = \frac{1}{n}\sum_{j=r+1}^{n} \frac{(j-1)(j-2)\ldots(j-r)}{(n-1)(n-2)\ldots(n-r)} x_{j:n} \qquad \text{(A4.18)}$$

Dessa forma, os estimadores de $\beta_r, r \le 2$, são dados por

$$b_0 = \frac{1}{n}\sum_{j=1}^{n} x_{j:n} \tag{A4.19}$$

$$b_1 = \frac{1}{n}\sum_{j=2}^{n} \frac{(j-1)}{(n-1)} x_{j:n} \tag{A4.20}$$

$$b_2 = \frac{1}{n}\sum_{j=3}^{n} \frac{(j-1)(j-2)}{(n-1)(n-2)} x_{j:n} \tag{A4.21}$$

Analogamente às equações A4.6 a A4.9, os estimadores não-enviesados de $\lambda_r$ são os momentos-L amostrais, esses definidos pelas seguintes expressões :

$$\ell_1 = b_0 \tag{A4.22}$$

$$\ell_2 = 2b_1 - b_0 \tag{A4.23}$$

$$\ell_3 = 6b_2 - 6b_1 + b_0 \tag{A4.24}$$

$$\ell_4 = 20b_3 - 30b_2 + 12b_1 - b_0 \tag{A4.25}$$

$$\ell_{r+1} = \sum_{k=0}^{r} l_{r,k}^{*} b_k \; ; \quad r = 0,1,\ldots,n-1 \tag{A4.26}$$

Na equação A4.26, os coeficientes $l_{r,k}^{*}$ são definidos tal como na equação A4.3. Da mesma forma, os quocientes de momentos-L amostrais são dados por

$$t_r = \frac{\ell_r}{\ell_2} \; ; \quad r \geq 3 \tag{A4.27}$$

enquanto o *CV*-L amostral calcula-se através de

$$t = \frac{\ell_2}{\ell_1} \tag{A4.28}$$

Os estimadores de $\tau_r$, fornecidos pelas equações A4.27 e A4.28, são muito pouco enviesados quando calculados para amostras de tamanho moderado a grande. Hosking (1990, p. 116) utilizou a teoria assintótica para calcular o viés para amostras grandes; para a distribuição Gumbel, por exemplo, o viés assintótico de $t_3$ é $0{,}19n^{-1}$ ,

enquanto o de $t_4$, para a distribuição Normal, é $0{,}03n^{-1}$, onde $n$ representa o tamanho da amostra. Para amostras de pequeno tamanho, o viés pode ser avaliado por simulação. Segundo Hosking e Wallis (1997, p. 28) e para uma gama variada de distribuições, o viés de $t$ pode ser considerado desprezível para $n \geq 20$. Ainda segundo esses autores, mesmo em se tratando de amostras de tamanho em torno de 20, o viés de $t_3$ e o viés de $t_4$ são considerados relativamente pequenos e definitivamente menores do que os produzidos por estimadores convencionais de assimetria e curtose.

## ANEXO-A5

## DADOS HIDROLÓGICOS

### A5.1 – AMOSTRAS

Tabela A5.1: Dados de precipitação diária máxima anual da estação 01544012

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 19/01/39 | 75,0 | 02/01/52 | 72,4 | 31/10/70 | 62,2 | 05/01/85 | 106,2 |
| 06/03/40 | 70,0 | 17/03/53 | 153,0 | 19/03/72 | 87,0 | 06/11/85 | 96,1 |
| 18/01/41 | 80,0 | 13/02/56 | 61,0 | 25/12/72 | 122,0 | 24/12/86 | 52,3 |
| 09/11/41 | 65,0 | 24/02/57 | 80,4 | 28/03/74 | 93,0 | 10/12/87 | 118,9 |
| 05/11/42 | 72,0 | 21/03/58 | 84,7 | 15/11/75 | 75,0 | 06/03/89 | 65,0 |
| 21/11/43 | 72,2 | 30/01/59 | 97,0 | 25/01/77 | 107,0 | 13/12/91 | 86,6 |
| 18/04/45 | 78,2 | 16/01/62 | 114,0 | 06/10/77 | 100,0 | 12/02/93 | 57,6 |
| 27/12/45 | 170,0 | 14/12/62 | 83,4 | 25/01/79 | 79,0 | 09/04/94 | 87,0 |
| 25/01/47 | 67,0 | 16/01/64 | 87,6 | 07/04/80 | 50,2 | 23/04/95 | 50,8 |
| 14/02/48 | 85,6 | 21/01/65 | 140,0 | 01/01/81 | 85,0 | 18/11/95 | 64,0 |
| 25/12/48 | 110,0 | 21/10/66 | 69,0 | 23/03/82 | 69,2 | 01/01/97 | 73,5 |
| 18/10/49 | 75,0 | 01/02/68 | 86,0 | 13/02/83 | 82,0 | 20/01/98 | 94,0 |
| 21/01/51 | 60,0 | 04/12/69 | 74,4 | 29/11/83 | 85,0 | 02/03/99 | 66,0 |

Tabela A5.2: Dados de precipitação diária máxima anual da estação 01645000

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 11/03/53 | 90,0 | 11/01/66 | 100,0 | 02/01/81 | 83,6 | 15/01/92 | 65,0 |
| 17/12/53 | 55,0 | 26/12/67 | 109,0 | 19/01/82 | 85,0 | 29/11/92 | 69,9 |
| 21/11/54 | 105,0 | 03/12/68 | 70,0 | 24/01/83 | 95,4 | 12/12/93 | 77,1 |
| 20/12/55 | 90,0 | 29/01/70 | 66,0 | 11/11/83 | 112,8 | 16/12/94 | 87,0 |
| 02/01/59 | 75,0 | 03/10/70 | 77,0 | 19/03/85 | 115,4 | 29/11/95 | 68,6 |
| 30/12/59 | 92,0 | 22/11/72 | 81,0 | 02/01/86 | 63,0 | 02/03/97 | 137,1 |
| 06/03/61 | 53,0 | 09/03/74 | 58,4 | 12/12/86 | 52,0 | 20/01/98 | 67,0 |
| 22/01/62 | 68,0 | 20/12/74 | 58,4 | 10/11/87 | 140,6 | 04/03/99 | 89,8 |
| 05/12/62 | 77,0 | 25/02/76 | 92,2 | 11/12/88 | 52,8 | | |
| 17/01/64 | 70,0 | 15/01/79 | 93,5 | 14/12/89 | 80,2 | | |
| 07/02/65 | 51,0 | 17/01/80 | 107,2 | 23/03/91 | 103,6 | | |

Tabela A5.3: Dados de precipitação diária máxima anual da estação 01943000

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 01/12/41 | 59,2 | 21/01/55 | 80,8 | 25/12/69 | 104,6 | 03/03/86 | 100,0 |
| 12/03/43 | 144,0 | 26/12/56 | 81,3 | 09/11/70 | 55,1 | 01/01/88 | 76,2 |
| 08/01/45 | 71,1 | 28/11/57 | 90,4 | 25/12/71 | 74,4 | 21/11/88 | 51,0 |
| 30/10/45 | 60,9 | 08/10/58 | 56,1 | 10/03/73 | 78,3 | 13/12/89 | 75,0 |
| 19/02/47 | 79,2 | 29/02/60 | 84,8 | 06/02/75 | 72,0 | 07/03/94 | 89,7 |
| 29/12/47 | 103,6 | 26/01/61 | 87,4 | 01/11/75 | 52,0 | 23/12/94 | 96,8 |
| 27/01/49 | 126,0 | 20/12/62 | 73,9 | 13/01/78 | 119,1 | 14/12/95 | 124,0 |
| 05/02/50 | 54,6 | 13/01/64 | 88,4 | 09/02/79 | 130,0 | 04/01/97 | 141,0 |
| 28/03/51 | 96,0 | 09/01/65 | 76,7 | 11/11/81 | 140,4 | 09/12/97 | 57,0 |
| 13/01/52 | 68,1 | 13/01/66 | 76,7 | 03/01/83 | 99,1 | 30/11/98 | 107,0 |
| 24/10/52 | 100,8 | 29/02/68 | 61,7 | 28/11/83 | 70,3 | | |
| 20/11/53 | 101,1 | 23/01/69 | 104,4 | 09/01/85 | 110,0 | | |

Tabela A5.4: Dados de precipitação diária máxima anual da estação 01944004

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 17/04/42 | 68,8 | 27/10/58 | 36,0 | 12/03/73 | 81,3 | 11/03/87 | 109,0 |
| 01/04/45 | 67,3 | 16/10/59 | 64,2 | 14/01/74 | 85,3 | 17/03/88 | 88,0 |
| 20/02/47 | 70,2 | 04/01/61 | 83,4 | 06/02/75 | 58,4 | 07/03/89 | 99,6 |
| 12/03/48 | 113,2 | 10/01/62 | 64,2 | 12/10/75 | 66,3 | 13/11/89 | 74,0 |
| 08/02/49 | 79,2 | 15/12/62 | 76,4 | 18/11/76 | 91,3 | 28/01/91 | 94,0 |
| 18/12/49 | 61,2 | 13/01/64 | 159,4 | 13/01/78 | 72,8 | 04/02/92 | 99,2 |
| 16/12/50 | 66,4 | 19/01/65 | 62,1 | 05/02/79 | 100,0 | 12/12/92 | 101,6 |
| 26/11/51 | 65,1 | 31/01/66 | 78,3 | 26/01/80 | 78,4 | 07/03/94 | 76,6 |
| 20/03/53 | 115,0 | 10/01/67 | 74,3 | 01/12/80 | 61,8 | 23/12/94 | 84,8 |
| 13/02/54 | 67,3 | 20/01/68 | 41,0 | 11/11/81 | 83,4 | 14/12/95 | 114,4 |
| 24/01/55 | 102,2 | 03/12/68 | 101,6 | 11/02/83 | 93,4 | 14/12/97 | 95,8 |
| 17/02/56 | 54,4 | 11/01/70 | 85,6 | 19/10/83 | 99,0 | 04/11/98 | 65,4 |
| 12/01/57 | 69,3 | 09/11/70 | 51,4 | 21/11/84 | 133,0 | | |
| 09/12/57 | 54,3 | 26/12/71 | 70,3 | 04/03/86 | 101,0 | | |

Tabela A5.5: Dados de precipitação diária máxima anual da estação 01944007

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 18/04/42 | 72,0 | 13/01/64 | 153,2 | 21/12/74 | 68,0 | 21/12/89 | 105,1 |
| 16/03/43 | 84,3 | 11/02/65 | 64,4 | 29/03/76 | 104,6 | 04/02/92 | 78,8 |
| 05/02/44 | 78,2 | 06/12/65 | 79,8 | 23/01/77 | 79,2 | 11/12/92 | 115,2 |
| 27/12/51 | 58,8 | 26/12/66 | 70,2 | 13/01/78 | 86,4 | 06/03/94 | 74,0 |
| 04/12/55 | 65,6 | 23/12/67 | 69,6 | 25/11/78 | 91,0 | 24/12/94 | 84,4 |
| 11/04/57 | 60,4 | 16/11/68 | 60,0 | 10/01/81 | 78,0 | 14/12/95 | 86,6 |
| 11/02/58 | 71,5 | 15/11/69 | 59,8 | 23/11/83 | 69,0 | 21/11/96 | 121,2 |
| 06/03/59 | 61,6 | 09/03/71 | 82,6 | 04/01/86 | 60,0 | 30/11/97 | 65,8 |
| 29/02/60 | 130,4 | 25/12/71 | 79,0 | 11/03/87 | 86,0 | 02/03/99 | 62,0 |
| 04/01/61 | 61,8 | 10/03/73 | 94,0 | 20/12/87 | 166,0 | | |
| 25/01/62 | 80,0 | 23/12/73 | 80,0 | 15/12/88 | 53,3 | | |

Tabela A5.6: Dados de precipitação diária máxima anual da estação 02044012

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 17/12/45 | 62,5 | 14/01/64 | 100,0 | 26/11/79 | 144,4 | 24/01/92 | 71,6 |
| 29/11/46 | 65,0 | 07/11/64 | 86,8 | 07/02/81 | 94,6 | 12/12/92 | 149,6 |
| 19/02/48 | 66,4 | 19/12/66 | 77,4 | 28/12/81 | 96,2 | 07/03/94 | 75,0 |
| 22/03/50 | 62,5 | 29/02/68 | 59,2 | 20/01/83 | 90,4 | 09/02/95 | 134,9 |
| 19/11/50 | 71,4 | 25/12/69 | 104,0 | 05/04/84 | 107,8 | 14/12/95 | 196,2 |
| 25/10/52 | 64,0 | 05/10/70 | 57,0 | 13/03/85 | 117,4 | 03/01/97 | 148,5 |
| 15/12/56 | 64,2 | 25/11/71 | 79,3 | 18/02/86 | 61,2 | 23/11/97 | 66,6 |
| 09/11/57 | 107,4 | 26/03/74 | 66,2 | 17/11/87 | 75,4 | 03/03/99 | 67,5 |
| 29/02/60 | 74,0 | 07/12/74 | 46,1 | 23/02/89 | 79,8 | | |
| 05/01/62 | 87,8 | 19/01/77 | 117,0 | 13/12/89 | 85,4 | | |
| 24/12/62 | 80,8 | 18/09/79 | 105,0 | 28/01/91 | 129,4 | | |

Tabela A5.7: Dados de precipitação diária máxima anual da estação 02045005

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 09/01/42 | 62,2 | 23/12/55 | 170,0 | 19/03/74 | 57,0 | 24/11/88 | 60,6 |
| 20/01/43 | 116,4 | 27/03/57 | 80,2 | 19/12/74 | 67,0 | 12/12/89 | 115,0 |
| 25/01/44 | 79,2 | 19/12/57 | 93,6 | 23/03/76 | 82,6 | 05/01/91 | 57,0 |
| 02/12/44 | 78,2 | 23/01/59 | 70,2 | 12/11/76 | 94,6 | 24/01/92 | 117,4 |
| 15/03/46 | 64,8 | 17/01/60 | 68,4 | 25/11/77 | 76,2 | 05/11/92 | 98,4 |
| 25/01/47 | 75,2 | 25/01/61 | 83,2 | 20/01/79 | 104,0 | 15/05/94 | 116,0 |
| 29/12/47 | 74,8 | 30/11/62 | 68,6 | 23/12/79 | 86,2 | 03/02/95 | 70,0 |
| 09/12/48 | 166,8 | 13/01/64 | 125,2 | 03/01/82 | 117,4 | 05/02/96 | 63,8 |
| 12/12/49 | 109,2 | 18/02/65 | 98,0 | 31/05/83 | 139,0 | 22/11/96 | 96,4 |
| 16/11/50 | 125,0 | 14/12/67 | 97,2 | 11/12/83 | 83,4 | 21/10/97 | 65,1 |
| 03/02/52 | 68,6 | 23/01/69 | 155,0 | 27/01/85 | 116,8 | 06/12/98 | 127,6 |
| 07/11/52 | 78,6 | 14/11/69 | 100,2 | 27/12/85 | 53,8 | | |
| 15/12/53 | 93,0 | 09/11/70 | 56,0 | 25/12/86 | 81,6 | | |
| 29/12/54 | 58,2 | 27/12/71 | 74,2 | 04/03/88 | 92,4 | | |

Tabela A5.8: Dados de precipitação diária máxima anual da estação 02244038

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 01/12/42 | 83,6 | 02/01/56 | 79,6 | 17/11/69 | 75,4 | 21/12/82 | 117,0 |
| 19/02/44 | 80,2 | 14/01/57 | 105,0 | 25/02/71 | 68,4 | 07/12/83 | 69,2 |
| 16/01/45 | 72,3 | 20/01/58 | 78,4 | 23/12/71 | 66,6 | 18/03/86 | 71,0 |
| 30/12/45 | 106,7 | 12/01/59 | 100,2 | 14/11/72 | 88,0 | 01/02/88 | 87,7 |
| 26/01/47 | 200,4 | 05/03/60 | 139,6 | 26/12/73 | 77,2 | 14/03/89 | 104,2 |
| 15/03/48 | 98,3 | 29/11/60 | 164,4 | 02/04/75 | 98,9 | 20/12/89 | 75,0 |
| 15/10/48 | 80,5 | 22/02/63 | 65,2 | 03/01/76 | 116,4 | 17/02/91 | 120,4 |
| 06/01/50 | 127,5 | 04/01/64 | 78,0 | 19/11/76 | 62,0 | 17/01/92 | 67,6 |
| 13/01/51 | 81,4 | 23/12/64 | 139,4 | 09/01/78 | 115,2 | 05/11/92 | 78,5 |
| 29/02/52 | 101,4 | 13/01/66 | 140,4 | 21/02/79 | 79,4 | 01/02/94 | 77,2 |
| 20/12/52 | 75,5 | 19/03/67 | 97,4 | 02/11/79 | 71,0 | 03/03/96 | 69,2 |
| 19/10/53 | 79,4 | 17/03/68 | 77,2 | 09/01/81 | 110,2 | | |
| 11/12/54 | 67,6 | 24/01/69 | 98,0 | 08/12/81 | 71,0 | | |

Tabela A5.9: Dados de precipitação diária máxima anual da estação 01943009

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 29/12/41 | 62,6 | 13/12/58 | 54,6 | 02/11/71 | 65,2 | 14/12/86 | 60,0 |
| 29/12/42 | 61,0 | 01/03/60 | 86,0 | 11/03/73 | 128,9 | 10/12/87 | 68,5 |
| 24/03/44 | 86,0 | 26/01/61 | 70,0 | 07/12/74 | 61,2 | 10/02/91 | 60,8 |
| 10/01/45 | 108,6 | 17/11/61 | 75,8 | 04/07/76 | 32,8 | 12/11/91 | 62,0 |
| 26/12/45 | 54,4 | 19/02/65 | 79,0 | 14/02/78 | 210,8 | 12/12/92 | 88,0 |
| 29/11/46 | 90,0 | 11/03/66 | 84,0 | 01/02/79 | 102,6 | 07/03/94 | 69,3 |
| 16/02/48 | 79,0 | 25/12/66 | 66,0 | 25/12/79 | 77,9 | 15/03/95 | 68,0 |
| 12/12/51 | 81,4 | 05/01/68 | 54,0 | 25/01/82 | 98,2 | 27/12/95 | 126,8 |
| 12/01/57 | 83,4 | 27/10/69 | 88,2 | 10/01/85 | 114,0 | 10/01/98 | 84,2 |
| 05/12/57 | 63,0 | 09/11/70 | 55,4 | 13/05/86 | 62,5 | | |

Tabela A5.10: Dados de precipitação diária máxima anual da estação 02243004

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 30/12/45 | 66,8 | 27/10/57 | 60,7 | 01/12/69 | 47,4 | 07/12/81 | 67,2 |
| 18/02/47 | 143,9 | 11/01/59 | 70,0 | 26/03/71 | 65,4 | 01/01/83 | 77,3 |
| 01/02/48 | 60,2 | 09/04/60 | 80,7 | 26/01/72 | 63,4 | 15/11/83 | 81,2 |
| 13/11/48 | 64,8 | 12/02/61 | 131,0 | 05/03/73 | 87,7 | 06/01/86 | 101,4 |
| 06/01/50 | 78,3 | 15/01/62 | 97,0 | 30/10/73 | 59,2 | 08/02/88 | 104,0 |
| 18/01/51 | 102,4 | 28/10/62 | 63,3 | 13/01/75 | 95,3 | 03/02/89 | 72,0 |
| 08/01/52 | 57,5 | 27/11/63 | 60,3 | 25/11/75 | 77,2 | 11/11/89 | 82,0 |
| 25/03/53 | 65,5 | 13/12/64 | 52,1 | 18/01/77 | 64,2 | 12/03/91 | 59,1 |
| 28/12/53 | 54,2 | 26/03/66 | 72,2 | 22/11/77 | 96,4 | 25/04/92 | 57,1 |
| 09/12/54 | 45,8 | 18/03/67 | 81,3 | 14/02/79 | 112,2 | 25/04/93 | 106,2 |
| 16/02/56 | 50,5 | 22/12/67 | 77,3 | 03/12/79 | 68,2 | 14/01/94 | 67,2 |
| 22/10/56 | 60,4 | 24/01/69 | 70,2 | 12/11/80 | 63,3 | 12/09/96 | 70,0 |

Tabela A5.11: Dados de vazões médias diárias máximas anuais da estação 40025000

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 10/01/40 | 255,0 | 09/01/63 | 68,2 | 16/01/78 | 41,6 | 06/01/89 | 49,2 |
| 21/11/40 | 81,3 | 15/02/64 | 71,0 | 23/02/79 | 43,5 | 04/01/90 | 58,8 |
| 22/12/41 | 117,0 | 03/12/68 | 295,0 | 25/12/79 | 38,8 | 28/03/91 | 156,0 |
| 24/01/43 | 134,0 | 13/11/69 | 65,4 | 14/01/81 | 38,1 | 23/01/92 | 156,0 |
| 03/03/52 | 207,0 | 27/02/71 | 18,6 | 17/12/81 | 79,7 | 24/02/93 | 65,4 |
| 27/02/55 | 67,9 | 04/12/71 | 52,6 | 03/02/83 | 84,5 | 04/01/94 | 64,6 |
| 18/12/55 | 68,3 | 31/03/73 | 47,4 | 28/12/83 | 71,4 | 23/12/94 | 99,1 |
| 14/02/57 | 142,0 | 14/03/74 | 67,6 | 30/01/85 | 52,6 | 12/01/96 | 60,0 |
| 23/01/58 | 160,0 | 27/02/75 | 52,8 | 15/01/86 | 141,0 | 14/02/98 | 79,0 |
| 13/02/61 | 69,6 | 28/11/75 | 52,8 | 30/12/86 | 121,0 | 12/03/99 | 44,5 |
| 23/03/62 | 61,8 | 30/01/77 | 69,5 | 11/02/88 | 48,0 | | |

Tabela A5.12: Dados de vazões médias diárias máximas anuais da estação 40050000

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 30/03/36 | 356,0 | 18/02/54 | 271,0 | 03/04/73 | 416,0 | 11/02/88 | 520,0 |
| 20/01/37 | 547,0 | 01/03/55 | 252,0 | 09/01/74 | 397,0 | 09/01/89 | 309,0 |
| 02/01/38 | 706,0 | 31/12/55 | 369,0 | 04/01/75 | 436,0 | 06/01/90 | 721,0 |
| 12/01/40 | 515,0 | 29/12/56 | 532,0 | 28/12/75 | 390,0 | 03/04/91 | 982,0 |
| 05/01/41 | 363,0 | 27/01/58 | 305,0 | 02/02/77 | 437,0 | 06/02/92 | 1227,0 |
| 22/02/42 | 324,0 | 01/02/59 | 420,0 | 18/01/78 | 378,0 | 26/02/93 | 533,0 |
| 08/01/43 | 683,0 | 24/02/64 | 306,0 | 30/01/79 | 486,0 | 07/01/94 | 613,0 |
| 10/03/44 | 243,0 | 28/02/65 | 831,0 | 28/01/80 | 552,0 | 13/02/95 | 605,0 |
| 30/01/47 | 743,0 | 23/01/66 | 600,0 | 18/01/81 | 312,0 | 03/01/96 | 349,0 |
| 17/02/48 | 445,0 | 21/01/67 | 475,0 | 20/12/81 | 555,0 | 12/01/97 | 645,0 |
| 01/02/49 | 486,0 | 29/12/67 | 501,0 | 11/02/83 | 640,0 | 19/12/97 | 358,0 |
| 30/01/50 | 455,0 | 06/12/68 | 401,0 | 30/12/83 | 447,0 | 02/03/99 | 376,0 |
| 30/03/51 | 571,0 | 16/11/69 | 403,0 | 03/02/85 | 513,0 | | |
| 14/03/52 | 540,0 | 28/02/71 | 140,0 | 17/01/86 | 598,0 | | |
| 28/03/53 | 346,0 | 11/12/71 | 397,0 | 01/01/87 | 858,0 | | |

Tabela A5.13: Dados de vazões médias diárias máximas anuais da estação 40100000

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 01/02/59 | 602,0 | 17/11/69 | 653,0 | 20/01/81 | 621,0 | 12/02/92 | 1853,0 |
| 25/01/60 | 782,0 | 28/02/71 | 234,0 | 29/01/82 | 1091,0 | 19/02/93 | 712,0 |
| 12/01/61 | 859,0 | 27/12/71 | 687,0 | 13/02/83 | 1342,0 | 11/01/94 | 869,0 |
| 29/01/62 | 676,0 | 06/04/73 | 653,0 | 01/01/84 | 864,0 | 21/02/95 | 765,0 |
| 29/12/62 | 1071,0 | 09/01/74 | 691,0 | 06/02/85 | 1122,0 | 04/01/96 | 760,0 |
| 27/01/64 | 688,0 | 06/01/75 | 751,0 | 17/01/86 | 729,0 | 10/01/97 | 2771,0 |
| 08/03/65 | 1281,0 | 02/12/75 | 586,0 | 08/01/87 | 785,0 | 20/02/98 | 656,0 |
| 23/01/66 | 1016,0 | 04/02/77 | 840,0 | 18/02/88 | 778,0 | 07/03/99 | 843,0 |
| 18/01/67 | 867,0 | 20/01/78 | 755,0 | 11/01/89 | 471,0 | | |
| 01/03/68 | 890,0 | 01/02/79 | 987,0 | 08/01/90 | 767,0 | | |
| 28/01/69 | 766,0 | 31/01/80 | 1117,0 | 12/04/91 | 811,0 | | |

Tabela A5.14: Dados de vazões médias diárias máximas anuais da estação 40680000

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 11/02/39 | 93,9 | 01/04/54 | 69,4 | 09/11/70 | 58,3 | 14/01/86 | 56,6 |
| 18/03/40 | 57,1 | 16/11/54 | 55,5 | 26/12/71 | 141,0 | 26/12/86 | 106,0 |
| 24/12/40 | 140,0 | 03/03/56 | 95,4 | 05/02/73 | 71,5 | 26/12/87 | 47,3 |
| 12/12/41 | 61,0 | 28/12/56 | 92,3 | 16/01/74 | 65,1 | 15/02/89 | 48,2 |
| 29/12/42 | 116,0 | 26/11/57 | 44,7 | 03/01/75 | 54,5 | 02/01/90 | 40,6 |
| 22/12/43 | 62,9 | 07/03/59 | 38,7 | 28/11/75 | 36,5 | 20/01/91 | 214,0 |
| 20/01/45 | 60,5 | 05/02/60 | 39,1 | 30/01/77 | 71,3 | 24/12/91 | 212,0 |
| 04/01/46 | 49,3 | 15/02/61 | 150,0 | 16/01/78 | 96,3 | 13/12/92 | 49,9 |
| 16/03/47 | 90,6 | 24/02/64 | 79,7 | 28/01/79 | 74,5 | 15/01/94 | 69,4 |
| 31/12/47 | 69,6 | 20/12/64 | 58,3 | 16/01/80 | 49,5 | 24/12/94 | 63,0 |
| 09/02/49 | 82,0 | 13/01/66 | 82,5 | 14/12/80 | 92,3 | 12/01/96 | 97,9 |
| 28/01/50 | 69,0 | 20/02/67 | 106,0 | 04/01/82 | 92,5 | 03/01/97 | 434,0 |
| 17/12/50 | 87,6 | 27/11/67 | 73,5 | 14/02/83 | 86,0 | 09/01/98 | 57,7 |
| 24/12/51 | 57,8 | 06/12/68 | 46,8 | 11/12/83 | 71,7 | 08/03/99 | 61,6 |
| 06/04/53 | 57,7 | 03/11/69 | 63,3 | 28/01/85 | 95,9 | | |

Tabela A5.15: Dados de vazões médias diárias máximas anuais da estação 41250000

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 19/03/40 | 36,1 | 23/12/52 | 53,8 | 05/12/71 | 48,6 | 12/12/83 | 92,5 |
| 16/11/40 | 43,5 | 19/02/54 | 56,4 | 12/03/73 | 44,3 | 27/01/85 | 155,0 |
| 11/01/42 | 43,3 | 26/01/55 | 64,5 | 01/02/74 | 38,1 | 15/01/86 | 36,4 |
| 22/12/43 | 47,6 | 29/12/55 | 65,6 | 07/02/75 | 54,6 | 14/12/86 | 27,8 |
| 29/03/45 | 48,2 | 17/12/56 | 105,0 | 26/11/75 | 18,2 | 08/03/89 | 42,2 |
| 23/12/45 | 48,2 | 20/12/57 | 66,1 | 24/01/77 | 145,0 | 21/12/89 | 103,0 |
| 17/03/47 | 47,3 | 10/01/59 | 12,9 | 15/02/78 | 89,8 | 18/01/91 | 87,8 |
| 31/12/47 | 41,0 | 19/11/59 | 32,4 | 06/02/79 | 246,0 | 06/02/92 | 158,0 |
| 28/01/49 | 74,6 | 28/01/61 | 89,4 | 28/12/79 | 78,6 | 12/12/92 | 69,0 |
| 24/12/49 | 69,7 | 23/01/69 | 55,8 | 09/12/80 | 72,3 | 23/01/94 | 101,0 |
| 26/01/51 | 71,6 | 26/12/69 | 73,9 | 25/01/82 | 129,0 | 16/03/95 | 78,4 |
| 05/02/52 | 57,2 | 20/11/70 | 26,4 | 02/03/83 | 156,0 | | |

Tabela A5.16: Dados de vazões médias diárias máximas anuais da estação 40800001

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 12/02/39 | 574,0 | 02/04/54 | 283,0 | 25/01/70 | 328,0 | 29/01/85 | 1070,0 |
| 18/03/40 | 414,0 | 01/01/56 | 360,0 | 10/11/70 | 226,0 | 15/01/86 | 438,0 |
| 25/12/40 | 473,0 | 29/12/56 | 677,0 | 26/12/71 | 498,0 | 11/02/88 | 567,0 |
| 12/01/42 | 451,0 | 13/02/58 | 243,0 | 28/12/72 | 398,0 | 15/03/89 | 245,0 |
| 13/03/43 | 659,0 | 06/03/59 | 393,0 | 25/12/73 | 424,0 | 22/12/89 | 488,0 |
| 23/12/43 | 400,0 | 25/01/60 | 418,0 | 04/01/75 | 335,0 | 29/01/91 | 937,0 |
| 22/01/45 | 341,0 | 15/02/61 | 779,0 | 29/11/75 | 261,0 | 25/01/92 | 860,0 |
| 23/12/45 | 309,0 | 23/01/62 | 321,0 | 31/01/77 | 553,0 | 06/01/93 | 432,0 |
| 17/03/47 | 482,0 | 25/12/62 | 501,0 | 15/01/78 | 740,0 | 13/01/94 | 555,0 |
| 01/01/48 | 509,0 | 24/02/64 | 695,0 | 06/02/79 | 812,0 | 25/12/94 | 641,0 |
| 29/01/49 | 758,0 | 20/02/65 | 538,0 | 19/01/80 | 555,0 | 27/12/95 | 670,0 |
| 27/01/50 | 347,0 | 15/01/66 | 670,0 | 15/12/80 | 436,0 | 03/01/97 | 904,0 |
| 30/03/51 | 694,0 | 14/02/67 | 450,0 | 05/01/82 | 622,0 | 18/02/98 | 295,0 |
| 04/02/52 | 400,0 | 29/11/67 | 434,0 | 11/02/83 | 702,0 | 12/03/99 | 397,0 |
| 14/12/52 | 270,0 | 05/12/68 | 432,0 | 12/12/83 | 574,0 | | |

Tabela A5.17: Dados de vazões médias diárias máximas anuais da estação 56028000

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 31/12/38 | 139,0 | 07/12/53 | 83,1 | 26/12/71 | 285,0 | 15/01/86 | 178,0 |
| 24/01/40 | 124,0 | 25/01/55 | 94,3 | 06/02/73 | 247,0 | 24/12/86 | 148,0 |
| 16/11/40 | 274,0 | 07/01/59 | 89,1 | 31/01/74 | 115,0 | 10/02/88 | 290,0 |
| 24/12/41 | 123,0 | 22/03/60 | 74,7 | 02/02/75 | 96,3 | 18/01/91 | 302,0 |
| 31/12/42 | 397,0 | 15/02/61 | 673,0 | 29/11/75 | 156,0 | 25/01/92 | 469,0 |
| 22/12/43 | 285,0 | 22/12/62 | 166,0 | 31/01/77 | 181,0 | 06/11/92 | 133,0 |
| 14/01/45 | 177,0 | 27/02/64 | 109,0 | 14/01/78 | 479,0 | 13/01/94 | 282,0 |
| 17/12/45 | 151,0 | 20/02/65 | 245,0 | 01/02/79 | 521,0 | 25/12/94 | 105,0 |
| 26/01/47 | 331,0 | 13/01/66 | 479,0 | 17/01/80 | 284,0 | 10/01/96 | 105,3 |
| 18/02/48 | 216,0 | 28/01/67 | 180,0 | 27/01/81 | 127,0 | 03/01/97 | 727,0 |
| 10/02/50 | 126,0 | 28/11/67 | 144,0 | 18/12/81 | 200,0 | 14/12/97 | 123,9 |
| 29/03/51 | 386,0 | 04/12/68 | 155,0 | 06/01/83 | 189,0 | 27/11/98 | 115,8 |
| 13/03/52 | 217,0 | 22/01/70 | 115,0 | 15/10/83 | 162,0 | | |
| 13/12/52 | 157,0 | 10/11/70 | 170,0 | 28/01/85 | 339,0 | | |

Tabela A5.18: Dados de vazões médias diárias máximas anuais da estação 560750000

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 26/01/39 | 315,0 | 07/12/53 | 250,0 | 05/12/68 | 291,0 | 12/12/83 | 337,0 |
| 18/03/40 | 266,0 | 25/01/55 | 226,0 | 26/12/69 | 212,0 | 29/01/85 | 513,0 |
| 17/11/40 | 436,0 | 30/12/55 | 256,0 | 10/11/70 | 191,0 | 15/01/86 | 287,0 |
| 26/01/42 | 315,0 | 28/12/56 | 438,0 | 29/02/72 | 284,0 | 01/01/87 | 302,0 |
| 30/12/42 | 494,0 | 12/12/57 | 217,0 | 19/01/73 | 351,0 | 11/02/88 | 411,0 |
| 23/12/43 | 481,0 | 07/01/59 | 305,0 | 10/01/74 | 221,0 | 19/01/91 | 593,0 |
| 14/01/45 | 356,0 | 08/03/60 | 212,0 | 04/01/75 | 183,0 | 25/01/92 | 679,0 |
| 18/12/45 | 242,0 | 16/02/61 | 710,0 | 30/11/75 | 284,0 | 06/11/92 | 286,0 |
| 28/01/47 | 307,0 | 11/02/62 | 393,0 | 07/12/76 | 240,0 | 13/01/94 | 314,0 |
| 18/02/48 | 351,0 | 25/12/62 | 366,0 | 15/01/78 | 412,0 | 25/12/94 | 218,0 |
| 28/01/49 | 511,0 | 24/02/64 | 342,0 | 24/01/79 | 451,0 | 04/01/97 | 1150,0 |
| 28/01/50 | 254,0 | 20/02/65 | 407,0 | 20/01/80 | 330,0 | 22/01/98 | 219,7 |
| 31/03/51 | 507,0 | 14/01/66 | 680,0 | 13/01/81 | 194,0 | 08/01/99 | 204,3 |
| 14/02/52 | 361,0 | 28/01/67 | 338,0 | 05/01/82 | 344,0 | | |
| 12/11/52 | 257,0 | 28/11/67 | 265,0 | 03/01/83 | 397,0 | | |

Tabela A5.19: Dados de vazões médias diárias máximas anuais da estação 56415000

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 31/01/31 | 222,0 | 31/12/46 | 169,0 | 04/12/70 | 86,3 | 12/03/87 | 122,0 |
| 07/01/32 | 142,0 | 31/12/47 | 191,0 | 25/11/71 | 156,0 | 09/12/87 | 130,0 |
| 03/01/33 | 229,0 | 20/04/58 | 109,0 | 25/03/73 | 99,7 | 22/12/89 | 144,0 |
| 01/01/34 | 192,0 | 10/01/59 | 66,0 | 10/12/73 | 81,3 | 23/03/91 | 128,0 |
| 12/02/35 | 192,0 | 10/03/60 | 138,0 | 03/01/75 | 155,0 | 05/01/92 | 153,0 |
| 12/12/35 | 79,2 | 15/02/61 | 283,0 | 16/12/75 | 128,0 | 21/11/92 | 124,0 |
| 05/02/37 | 258,0 | 11/02/62 | 164,0 | 30/01/77 | 122,0 | 06/01/94 | 111,0 |
| 25/01/39 | 299,0 | 20/12/62 | 193,0 | 15/01/78 | 106,0 | 24/12/94 | 130,0 |
| 20/03/40 | 63,2 | 31/03/64 | 73,7 | 28/01/80 | 193,0 | 26/12/95 | 111,0 |
| 17/11/40 | 156,0 | 31/12/64 | 122,0 | 03/12/80 | 152,0 | 02/01/97 | 178,0 |
| 27/01/42 | 172,0 | 14/01/66 | 311,0 | 15/03/82 | 196,0 | 12/02/98 | 91,4 |
| 31/12/42 | 253,0 | 15/02/67 | 115,0 | 03/01/83 | 121,0 | 07/01/99 | 128,0 |
| 24/12/43 | 193,0 | 05/01/68 | 130,0 | 11/12/83 | 141,0 | | |
| 22/01/45 | 124,0 | 25/01/69 | 144,0 | 28/01/85 | 350,0 | | |
| 03/01/46 | 122,0 | 03/12/69 | 161,0 | 08/01/86 | 169,0 | | |

Tabela A5.20: Dados de vazões médias diárias máximas anuais da estação 56500000

| Data | Valor | Data | Valor | Data | Valor | Data | Valor |
|---|---|---|---|---|---|---|---|
| 29/11/40 | 33,3 | 13/12/52 | 47,5 | 16/02/67 | 42,2 | 19/02/79 | 249,0 |
| 26/01/42 | 25,1 | 31/12/53 | 7,3 | 29/02/68 | 46,5 | 27/01/80 | 70,2 |
| 02/01/43 | 38,5 | 19/11/54 | 12,0 | 06/12/68 | 28,9 | 12/12/80 | 21,5 |
| 23/12/43 | 29,1 | 30/12/55 | 19,5 | 27/01/70 | 31,5 | 15/02/90 | 43,4 |
| 13/01/45 | 22,1 | 29/12/56 | 52,0 | 18/12/70 | 45,1 | 25/01/91 | 20,0 |
| 31/12/45 | 21,8 | 24/12/57 | 36,9 | 25/11/71 | 85,0 | 05/01/92 | 12,2 |
| 31/12/46 | 84,3 | 08/01/59 | 12,0 | 27/12/72 | 43,0 | 12/12/92 | 38,1 |
| 30/12/47 | 38,5 | 06/02/60 | 33,6 | 27/10/73 | 26,2 | 04/12/94 | 5,6 |
| 25/02/49 | 65,5 | 20/12/62 | 38,0 | 02/01/75 | 61,0 | 01/01/96 | 41,3 |
| 24/12/49 | 27,1 | 26/02/64 | 27,7 | 15/12/75 | 39,0 | 04/01/97 | 98,0 |
| 18/12/50 | 29,5 | 20/12/64 | 38,0 | 22/11/76 | 33,4 | 27/10/97 | 21,5 |
| 26/01/52 | 47,5 | 14/01/66 | 89,1 | 12/01/78 | 27,4 | | |

## A5.2: ESTATÍSTICAS AMOSTRAIS

Tabela A5.21: Momentos-L e razões-L das amostras analisadas

| Código | Dados Originais | | | | Logaritmos dos Dados | | | |
|---|---|---|---|---|---|---|---|---|
| | $\lambda_1$ | $\lambda_2$ | $\tau_3$ | $\tau_4$ | $\lambda_1$ | $\lambda_2$ | $\tau_3$ | $\tau_4$ |
| 01544012 | 84,567 | 12,764 | 0,247 | 0,226 | 4,402 | 0,146 | 0,106 | 0,183 |
| 01645000 | 82,551 | 12,428 | 0,135 | 0,110 | 4,380 | 0,151 | 0,018 | 0,087 |
| 01943000 | 88,048 | 14,228 | 0,129 | 0,094 | 4,439 | 0,163 | 0,007 | 0,079 |
| 01944004 | 81,665 | 12,395 | 0,122 | 0,149 | 4,366 | 0,154 | -0,021 | 0,146 |
| 01944007 | 82,186 | 12,417 | 0,339 | 0,240 | 4,373 | 0,141 | 0,211 | 0,171 |
| 02044012 | 90,144 | 16,860 | 0,286 | 0,145 | 4,450 | 0,179 | 0,148 | 0,093 |
| 02045005 | 91,123 | 15,585 | 0,210 | 0,109 | 4,469 | 0,168 | 0,085 | 0,071 |
| 02244038 | 93,086 | 14,420 | 0,336 | 0,164 | 4,497 | 0,145 | 0,226 | 0,099 |
| 01943009 | 80,105 | 14,246 | 0,306 | 0,275 | 4,332 | 0,169 | 0,116 | 0,209 |
| 02243004 | 75,229 | 11,146 | 0,264 | 0,180 | 4,287 | 0,142 | 0,143 | 0,145 |
| 40025000 | 88,781 | 28,528 | 0,402 | 0,253 | 4,325 | 0,305 | 0,138 | 0,179 |
| 40050000 | 494,368 | 100,365 | 0,203 | 0,201 | 6,137 | 0,205 | 0,000 | 0,176 |
| 40100000 | 879,561 | 174,924 | 0,372 | 0,406 | 6,707 | 0,186 | 0,102 | 0,347 |
| 40680000 | 84,158 | 23,165 | 0,456 | 0,372 | 4,309 | 0,237 | 0,188 | 0,204 |
| 41250000 | 72,385 | 22,476 | 0,297 | 0,225 | 4,125 | 0,319 | -0,001 | 0,182 |
| 40800001 | 512,509 | 106,404 | 0,162 | 0,114 | 6,173 | 0,212 | -0,003 | 0,097 |
| 56028000 | 226,676 | 74,404 | 0,364 | 0,181 | 5,261 | 0,315 | 0,134 | 0,085 |
| 56075000 | 357,879 | 79,616 | 0,328 | 0,230 | 5,804 | 0,208 | 0,137 | 0,134 |
| 56415000 | 155,663 | 32,819 | 0,222 | 0,201 | 4,978 | 0,211 | 0,027 | 0,175 |
| 56500000 | 42,679 | 15,393 | 0,390 | 0,373 | 3,528 | 0,360 | -0,031 | 0,259 |

## A5.3 – PARÂMETROS ESTIMADOS PARA AS DISTRIBUIÇÕES AJUSTADAS

Tabela A5.22: Parâmetros estimados para as distribuições Normal, Lognormal 2p, Gumbel e Exponencial

| Código | Normal | | Lognormal | | Gumbel | | Exponencial | |
|---|---|---|---|---|---|---|---|---|
| | m | s | $m_{LN}$ | $s_{LN}$ | x | a | x | a |
| 01544012 | 84,57 | 22,62 | 4,40 | 0,26 | 73,94 | 18,41 | 59,04 | 25,53 |
| 01645000 | 82,55 | 22,03 | 4,38 | 0,27 | 72,20 | 17,93 | 57,70 | 24,86 |
| 01943000 | 88,05 | 25,22 | 4,44 | 0,29 | 76,20 | 20,53 | 59,59 | 28,46 |
| 01944004 | 81,66 | 21,97 | 4,37 | 0,27 | 71,34 | 17,88 | 56,88 | 24,79 |
| 01944007 | 82,19 | 22,01 | 4,37 | 0,25 | 71,85 | 17,91 | 57,35 | 24,83 |
| 02044012 | 90,14 | 29,88 | 4,45 | 0,32 | 76,10 | 24,32 | 56,42 | 33,72 |
| 02045005 | 91,12 | 27,62 | 4,47 | 0,30 | 78,14 | 22,48 | 59,95 | 31,17 |
| 02244038 | 93,09 | 25,56 | 4,50 | 0,26 | 81,08 | 20,80 | 64,25 | 28,84 |
| 01943009 | 80,11 | 25,25 | 4,33 | 0,30 | 68,24 | 20,55 | 51,61 | 28,49 |
| 02243004 | 75,23 | 19,76 | 4,29 | 0,25 | 65,95 | 16,08 | 52,94 | 22,29 |
| 40025000 | 88,78 | 50,57 | 4,33 | 0,54 | 65,02 | 41,16 | 31,72 | 57,06 |
| 40050000 | 494,37 | 177,89 | 6,14 | 0,36 | 410,79 | 144,80 | 293,64 | 200,73 |
| 40100000 | 879,56 | 310,05 | 6,71 | 0,33 | 733,89 | 252,36 | 529,71 | 349,85 |
| 40680000 | 84,16 | 41,06 | 4,31 | 0,42 | 64,87 | 33,42 | 37,83 | 46,33 |
| 41250000 | 72,39 | 39,84 | 4,13 | 0,56 | 53,67 | 32,43 | 27,43 | 44,95 |
| 40800001 | 512,51 | 188,60 | 6,17 | 0,38 | 423,90 | 153,51 | 299,70 | 212,81 |
| 56028000 | 226,68 | 131,88 | 5,26 | 0,56 | 164,72 | 107,34 | 77,87 | 148,81 |
| 56075000 | 357,88 | 141,12 | 5,80 | 0,37 | 291,58 | 114,86 | 198,65 | 159,23 |
| 56415000 | 155,66 | 58,17 | 4,98 | 0,37 | 128,33 | 47,35 | 90,03 | 65,64 |
| 56500000 | 42,68 | 27,28 | 3,53 | 0,64 | 29,86 | 22,21 | 11,89 | 30,79 |

Tabela A5.23: Parâmetros estimados para as distribuições Pearson III e Log-Pearson III

| Código | Pearson III | | | Log-Pearson III | | |
|---|---|---|---|---|---|---|
| | m | s | g | $m_{LN}$ | $s_{LN}$ | $g_{LN}$ |
| 01544012 | 84,57 | 24,22 | 1,49 | 4,40 | 0,26 | 0,65 |
| 01645000 | 82,55 | 22,50 | 0,82 | 4,38 | 0,27 | 0,11 |
| 01943000 | 88,05 | 25,71 | 0,78 | 4,44 | 0,29 | 0,04 |
| 01944004 | 81,66 | 22,35 | 0,75 | 4,37 | 0,27 | -0,13 |
| 01944007 | 82,19 | 24,92 | 2,03 | 4,37 | 0,26 | 1,27 |
| 02044012 | 90,14 | 32,71 | 1,72 | 4,45 | 0,33 | 0,90 |
| 02045005 | 91,12 | 29,04 | 1,27 | 4,47 | 0,30 | 0,52 |
| 02244038 | 93,09 | 28,90 | 2,02 | 4,50 | 0,27 | 1,36 |
| 01943009 | 80,11 | 27,98 | 1,83 | 4,33 | 0,30 | 0,71 |
| 02243004 | 75,23 | 21,35 | 1,59 | 4,29 | 0,26 | 0,87 |
| 40025000 | 88,78 | 60,04 | 2,42 | 4,33 | 0,55 | 0,84 |
| 40050000 | 494,37 | 186,41 | 1,23 | 6,14 | 0,36 | 0,00 |
| 40100000 | 879,56 | 359,78 | 2,24 | 6,71 | 0,33 | 0,62 |
| 40680000 | 84,16 | 51,08 | 2,77 | 4,31 | 0,44 | 1,14 |
| 41250000 | 72,39 | 43,89 | 1,78 | 4,13 | 0,56 | -0,01 |
| 40800001 | 512,51 | 194,34 | 0,98 | 6,17 | 0,38 | -0,02 |
| 56028000 | 226,68 | 152,12 | 2,19 | 5,26 | 0,57 | 0,82 |
| 56075000 | 357,88 | 158,61 | 1,97 | 5,80 | 0,38 | 0,84 |
| 56415000 | 155,66 | 61,50 | 1,34 | 4,98 | 0,37 | 0,17 |
| 56500000 | 42,68 | 32,10 | 2,35 | 3,53 | 0,64 | -0,19 |

Tabela A5.24: Parâmetros estimados para as distribuições GEV e GPA

| Código | GEV | | | GPA | | |
|---|---|---|---|---|---|---|
| | x | a | k | x | a | k |
| 01544012 | 73,03 | 16,34 | -0,12 | 56,38 | 34,06 | 0,21 |
| 01645000 | 72,66 | 18,81 | 0,05 | 51,20 | 47,76 | 0,52 |
| 01943000 | 76,83 | 21,72 | 0,07 | 51,86 | 55,87 | 0,54 |
| 01944004 | 71,98 | 19,08 | 0,08 | 49,89 | 49,69 | 0,56 |
| 01944007 | 70,13 | 13,46 | -0,25 | 57,50 | 24,40 | -0,01 |
| 02044012 | 74,38 | 20,16 | -0,17 | 54,57 | 39,49 | 0,11 |
| 02045005 | 77,54 | 21,18 | -0,06 | 55,17 | 46,97 | 0,31 |
| 02244038 | 79,11 | 15,69 | -0,24 | 64,34 | 28,55 | -0,01 |
| 01943009 | 66,59 | 16,44 | -0,20 | 50,71 | 31,27 | 0,06 |
| 02243004 | 65,00 | 13,86 | -0,14 | 51,11 | 28,08 | 0,16 |
| 40025000 | 60,09 | 27,08 | -0,33 | 35,91 | 45,12 | -0,15 |
| 40050000 | 407,54 | 137,87 | -0,05 | 260,99 | 309,31 | 0,33 |
| 40100000 | 706,32 | 176,97 | -0,29 | 544,61 | 306,42 | -0,09 |
| 40680000 | 60,33 | 19,39 | -0,40 | 43,69 | 30,23 | -0,25 |
| 41250000 | 51,20 | 26,37 | -0,19 | 25,53 | 50,82 | 0,08 |
| 40800001 | 424,83 | 155,37 | 0,01 | 252,46 | 375,50 | 0,44 |
| 56028000 | 153,32 | 76,54 | -0,28 | 82,93 | 133,97 | -0,07 |
| 56075000 | 281,15 | 88,14 | -0,23 | 197,61 | 162,35 | 0,01 |
| 56415000 | 126,70 | 43,76 | -0,08 | 81,04 | 95,04 | 0,27 |
| 56500000 | 27,29 | 14,98 | -0,32 | 13,79 | 25,34 | -0,12 |

## A5.4 – GRÁFICOS

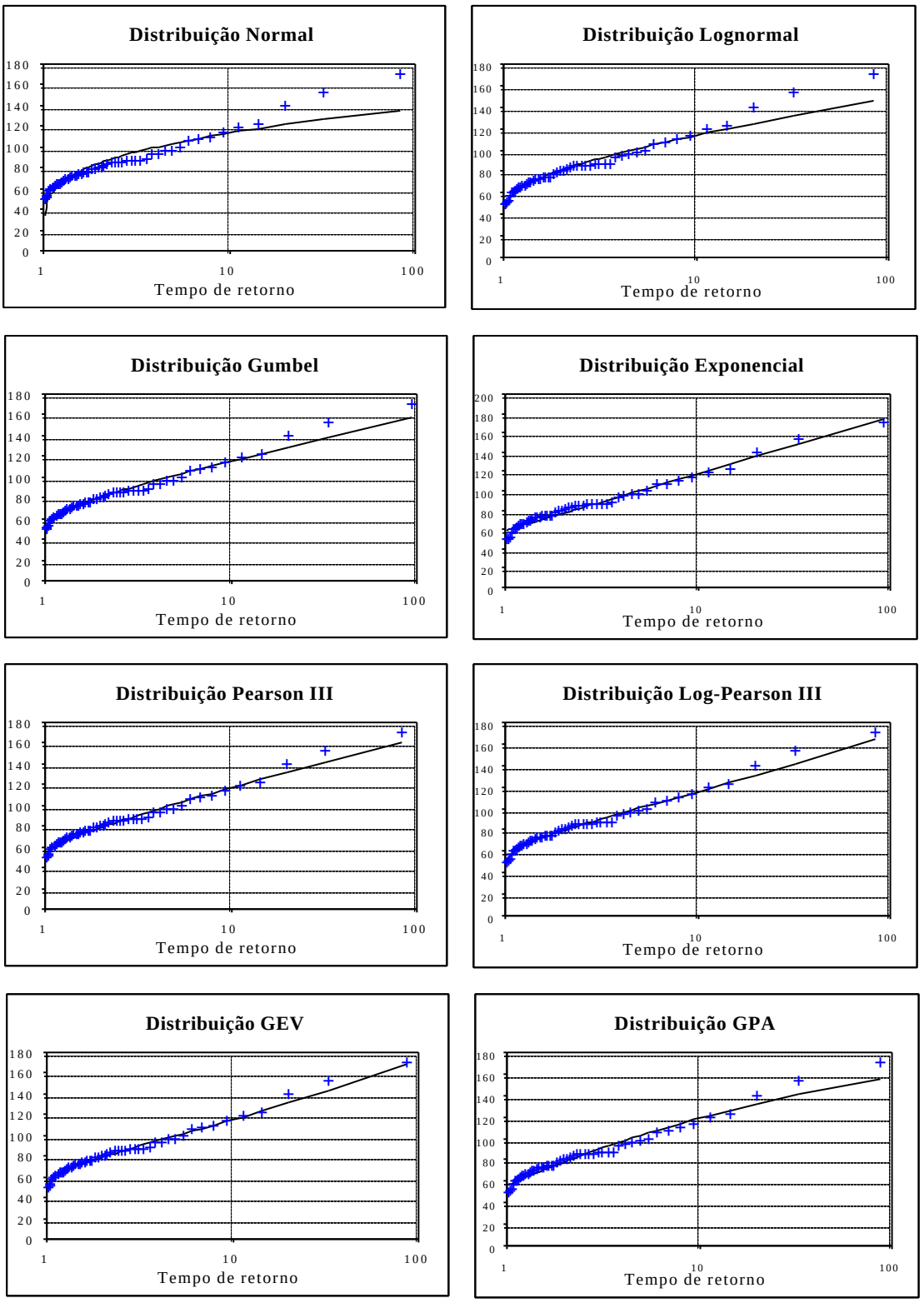



Figura A5.1: Ajuste visual dos dados de precipitação da estação 01544012

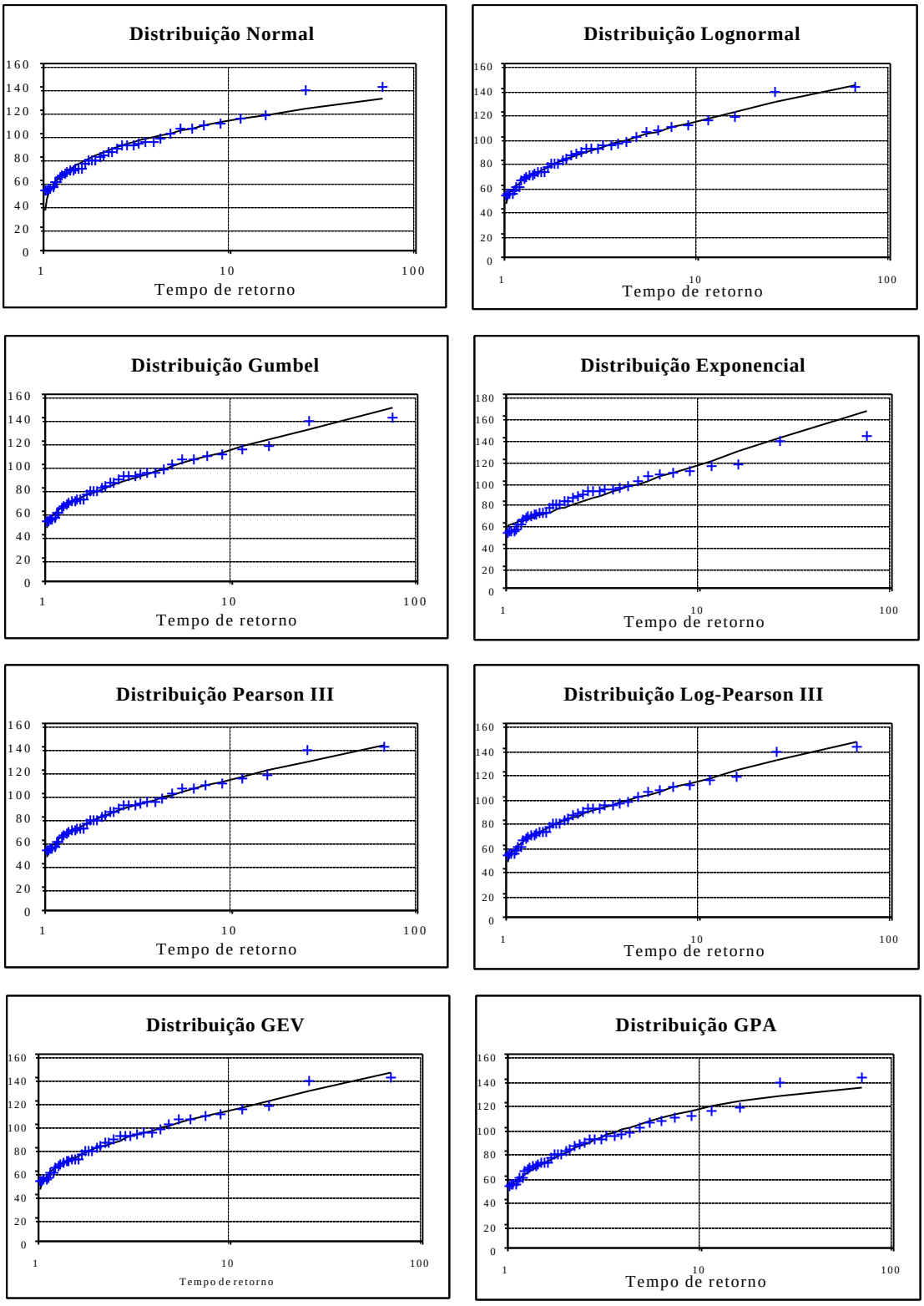



Figura A5.2: Ajuste visual dos dados de precipitação da estação 01645000

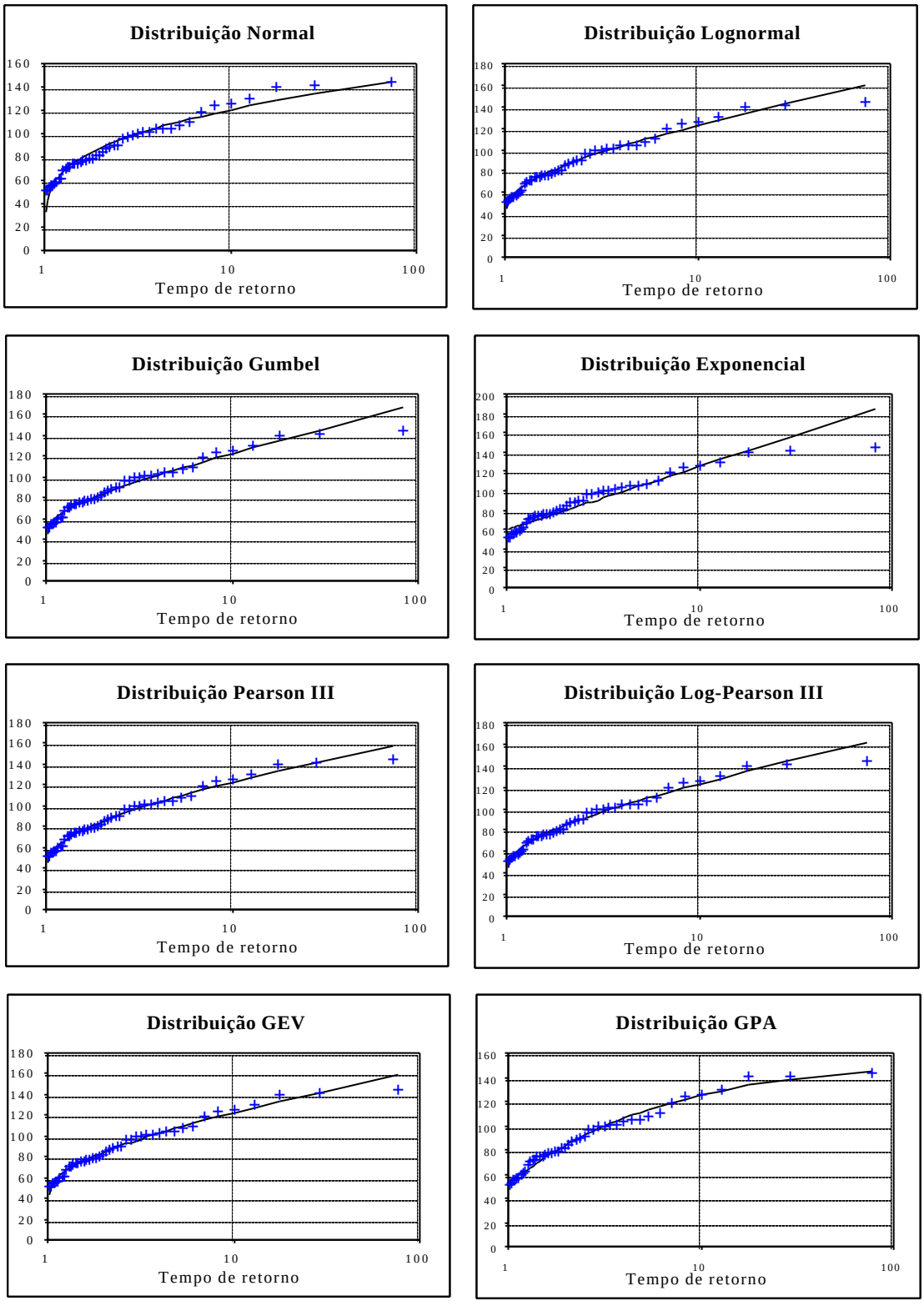



Figura A5.3: Ajuste visual dos dados de precipitação da estação 01943000

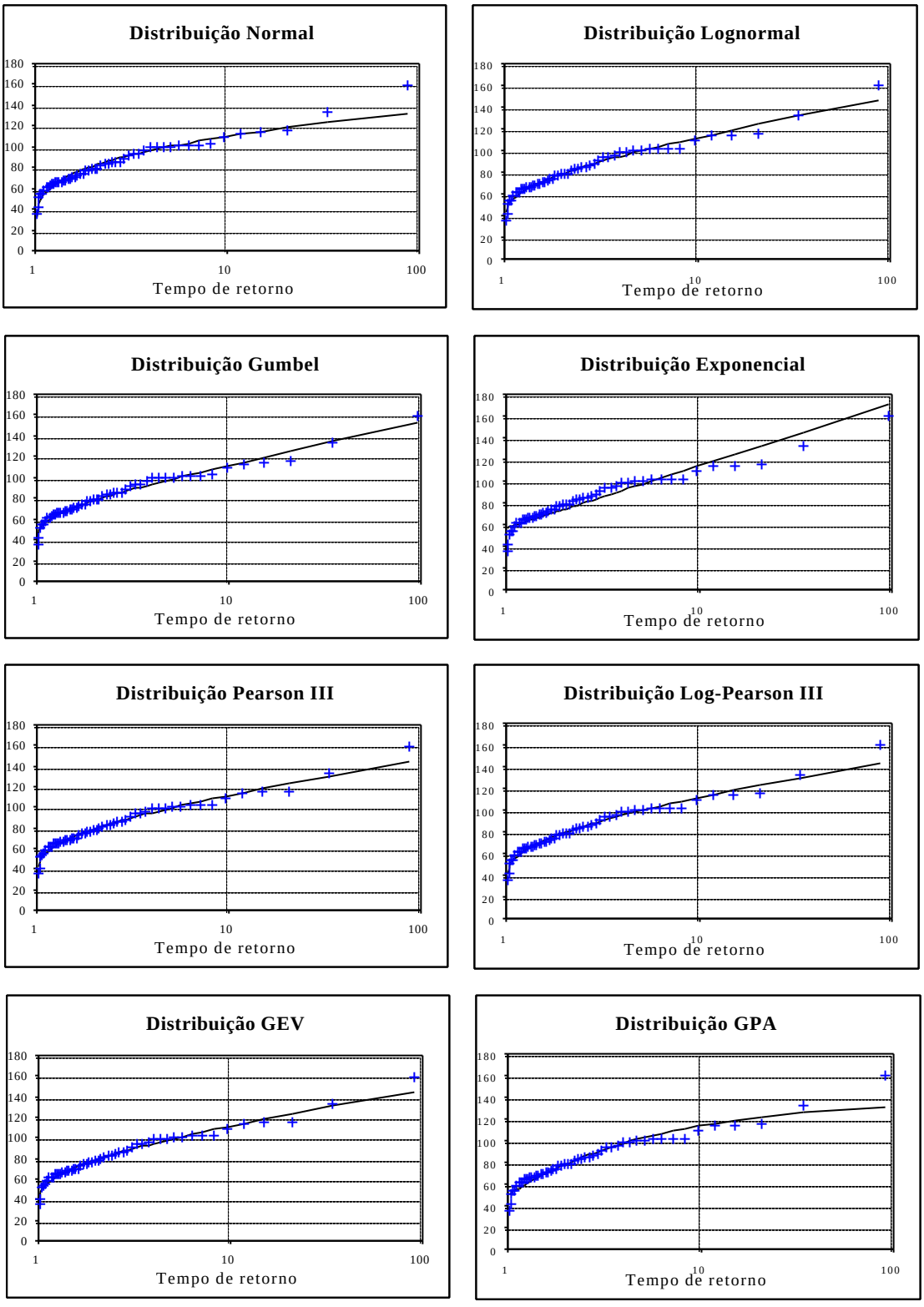



Figura A5.4: Ajuste visual dos dados de precipitação da estação 01944004

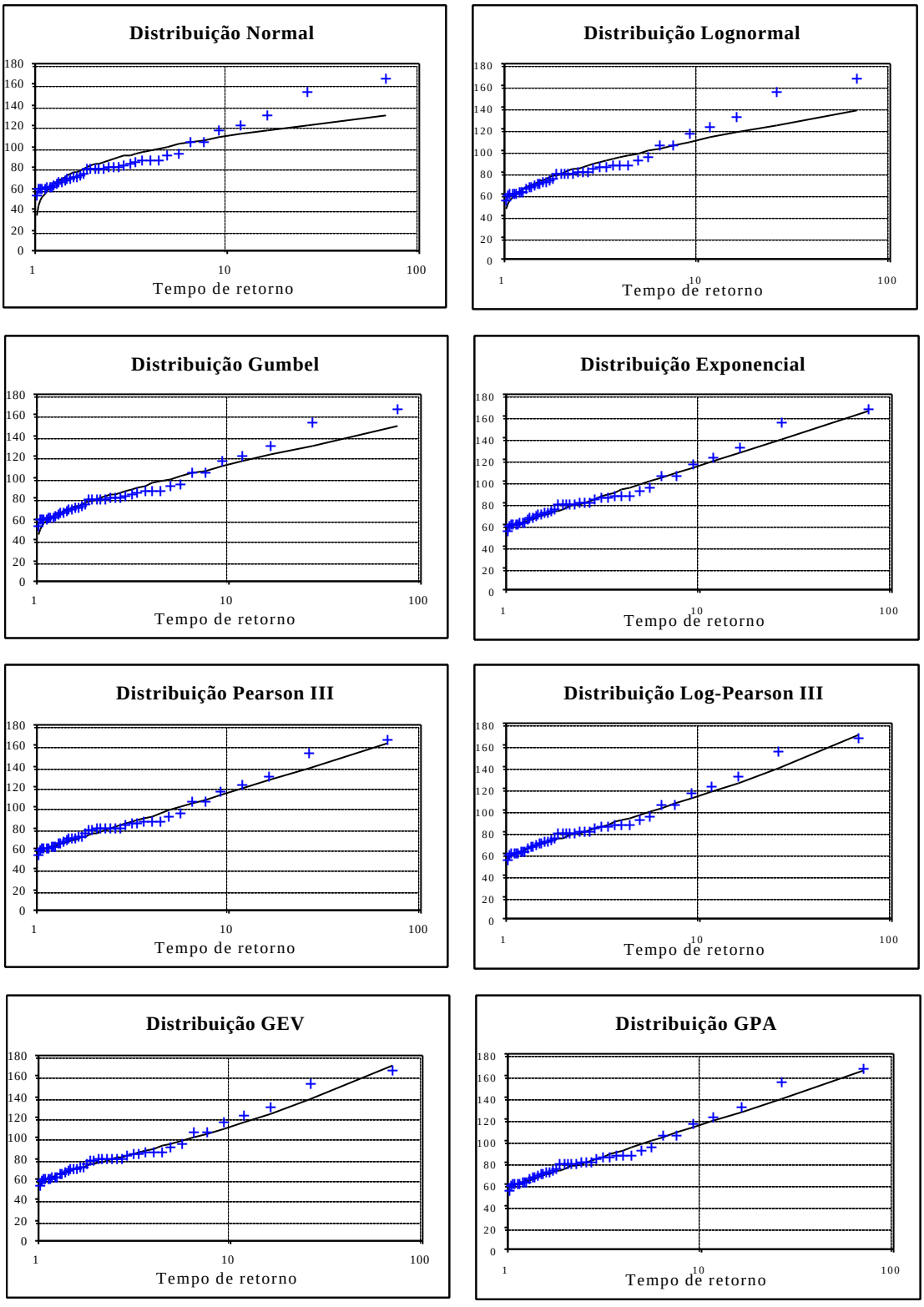



Figura A5.5: Ajuste visual dos dados de precipitação da estação 01944007

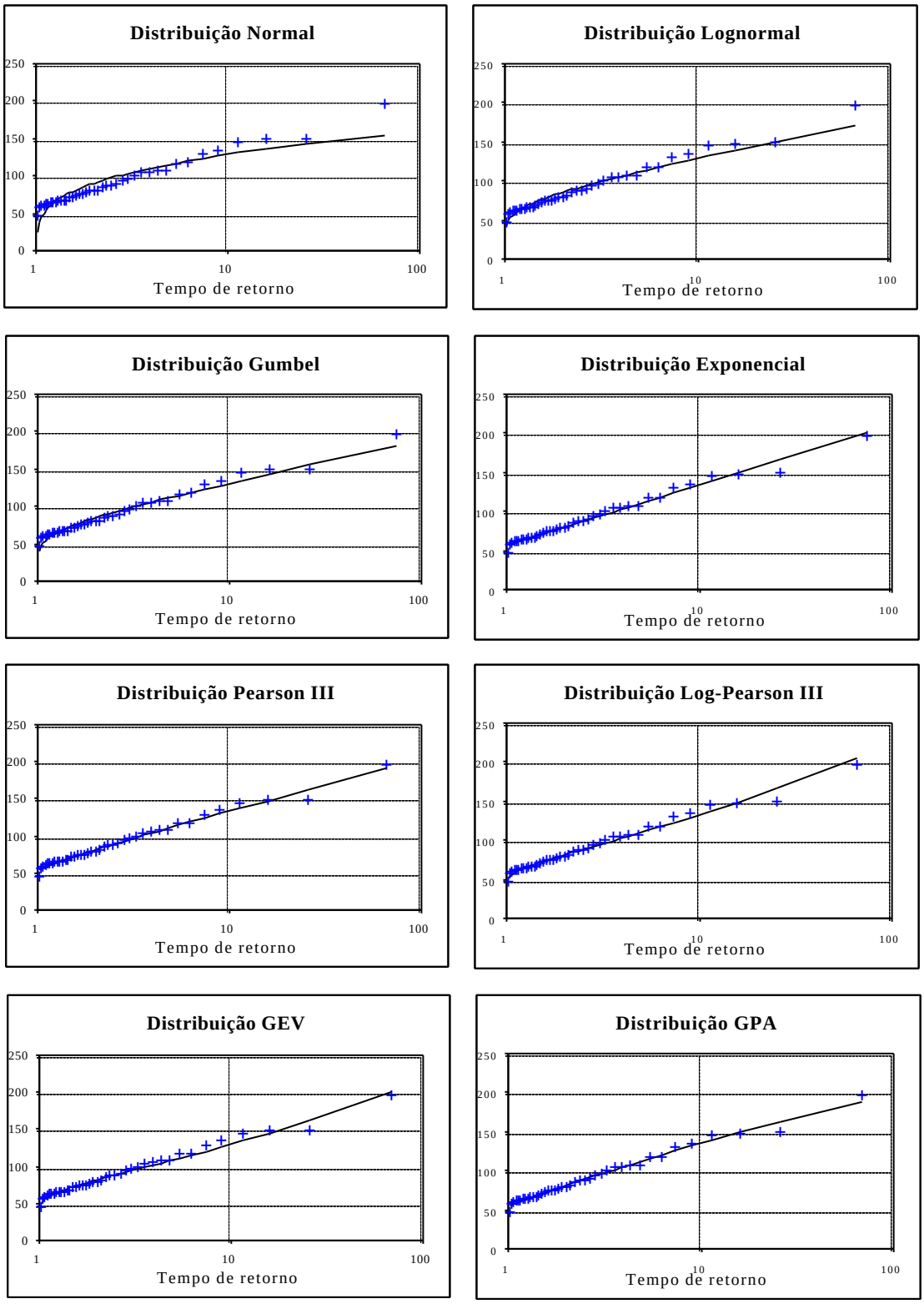



Figura A5.6: Ajuste visual dos dados de precipitação da estação 02044012

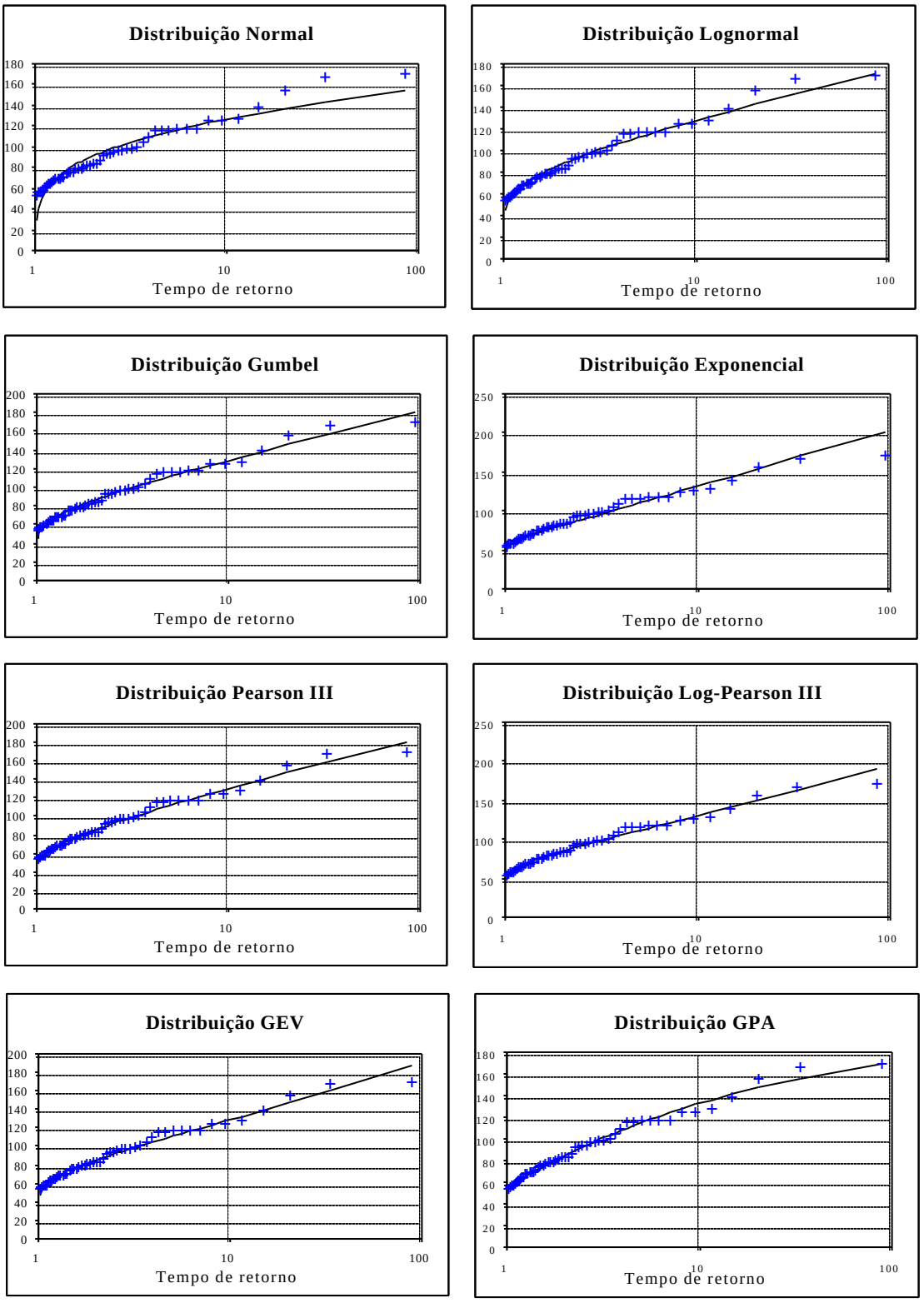



Figura A5.7: Ajuste visual dos dados de precipitação da estação 02045005

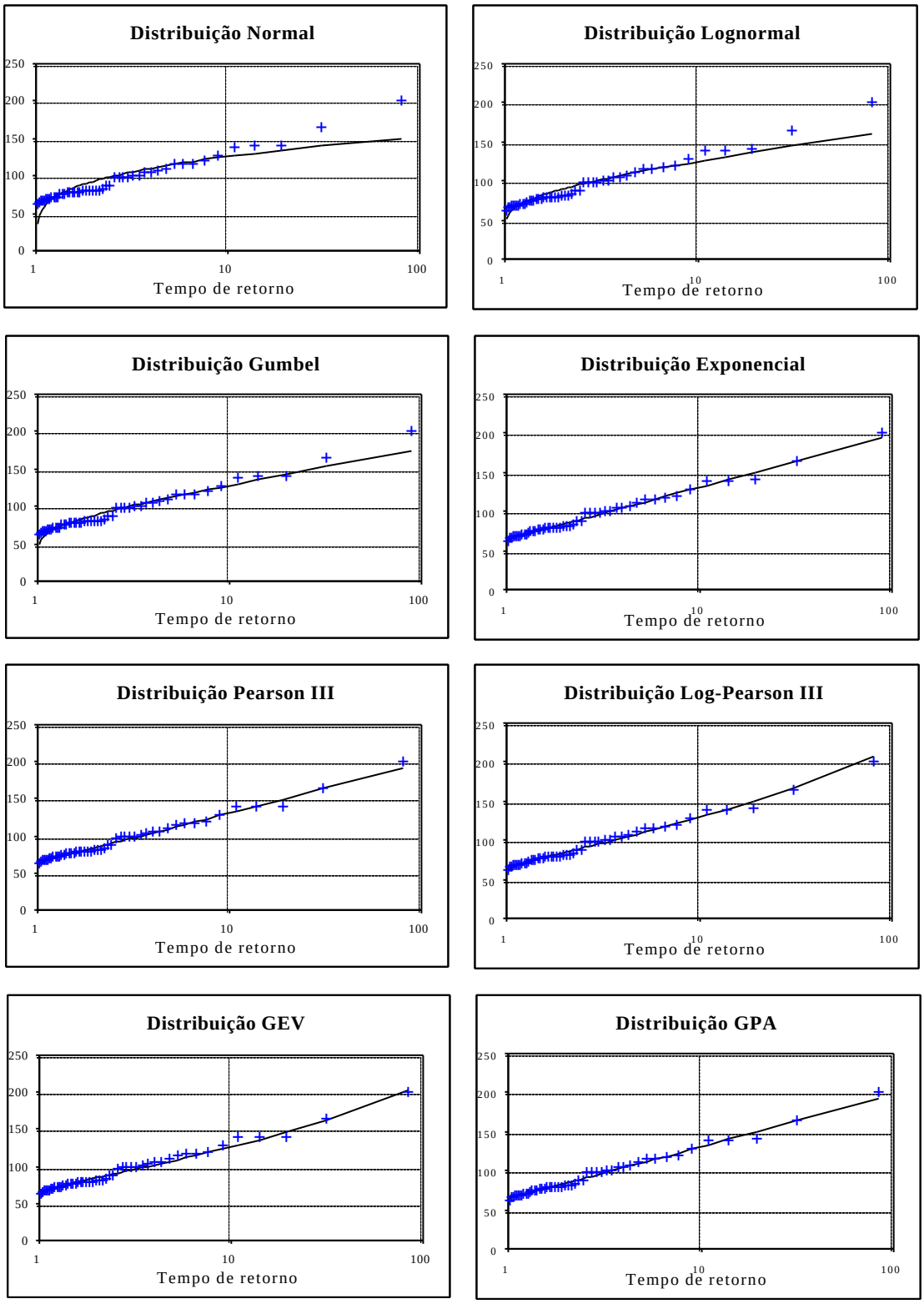



Figura A5.8: Ajuste visual dos dados de precipitação da estação 02244038

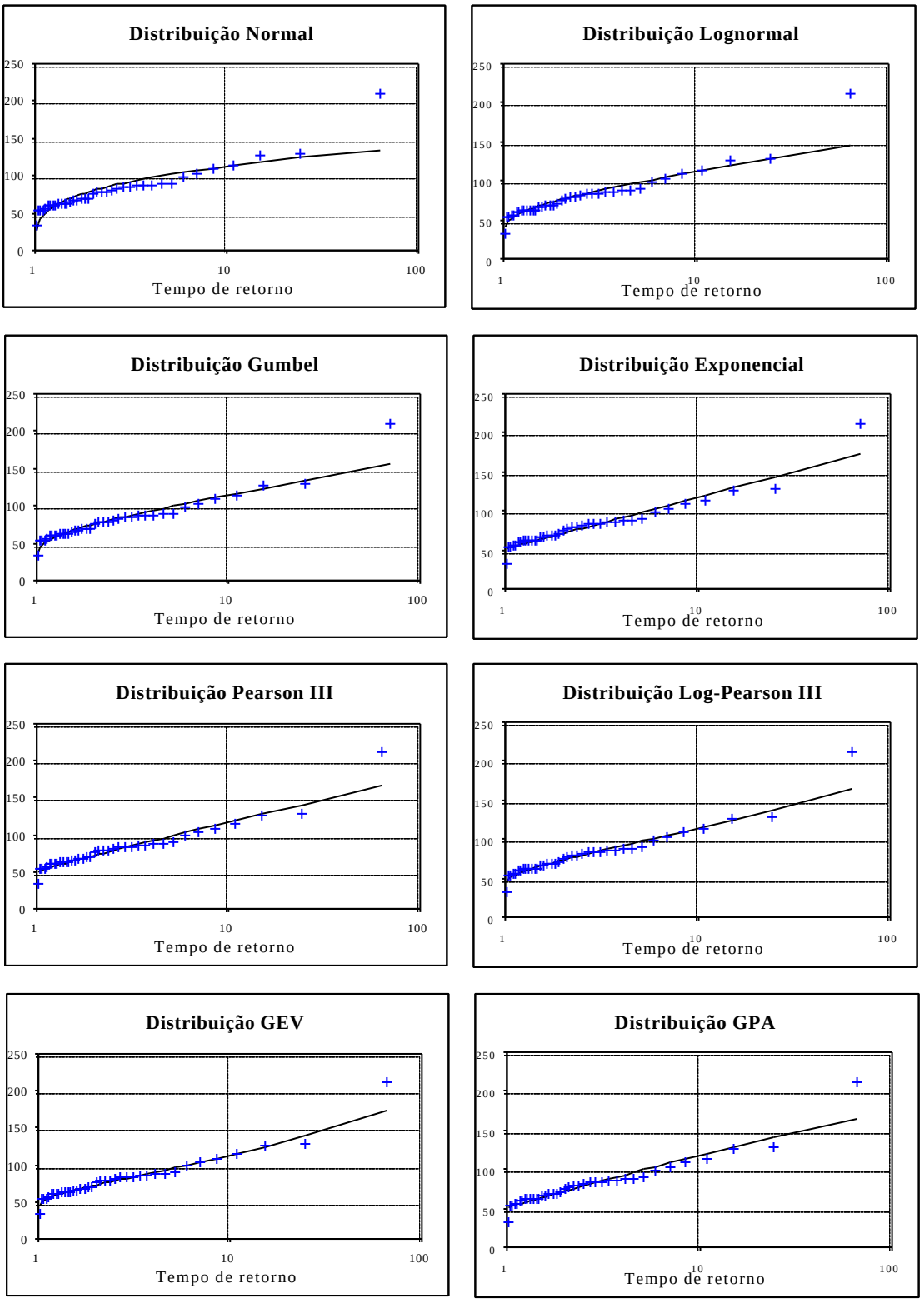



Figura A5.9: Ajuste visual dos dados de precipitação da estação 01943009

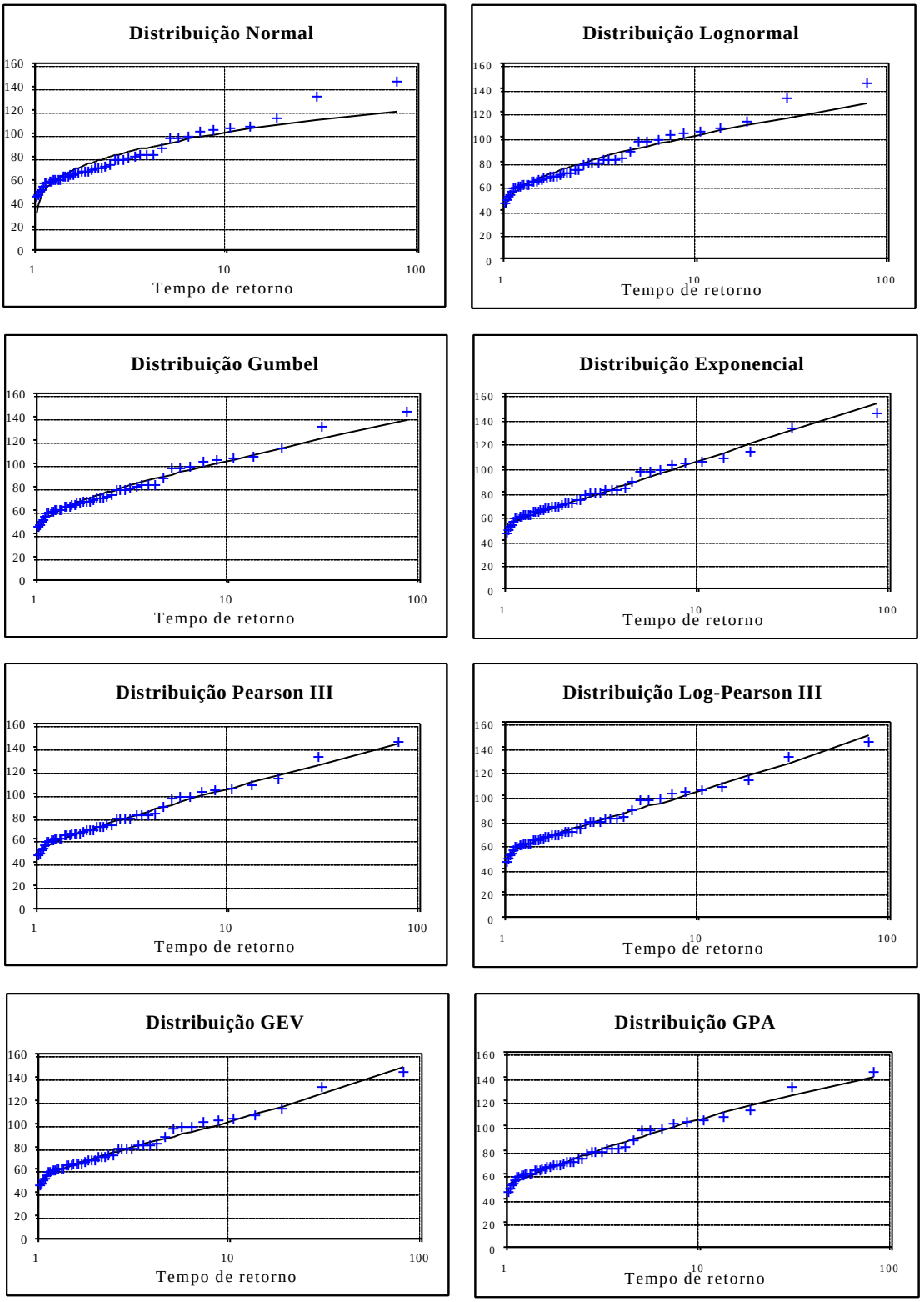



Figura A5.10: Ajuste visual dos dados de precipitação da estação 02243004

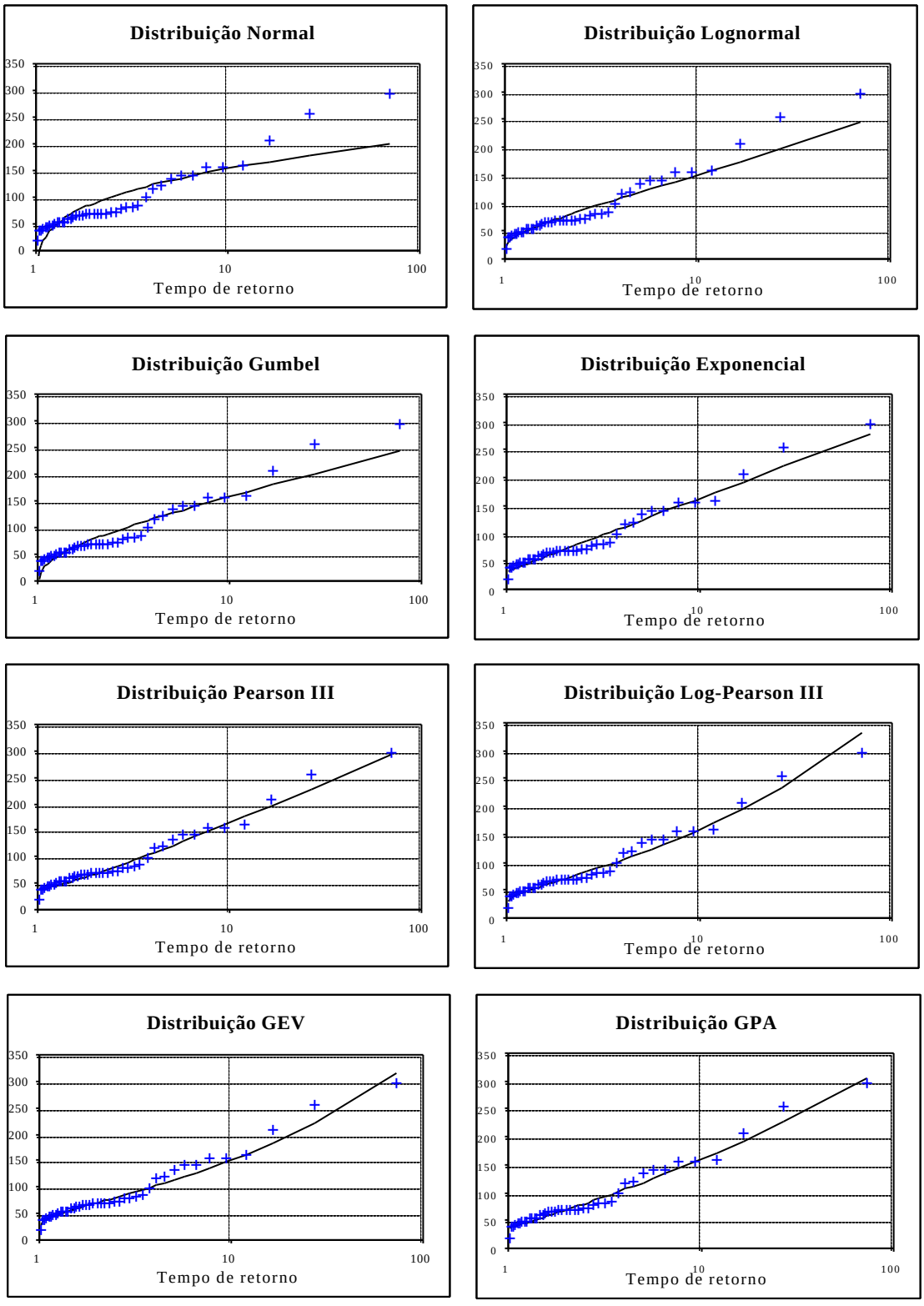



Figura A5.11: Ajuste visual dos dados de vazão da estação 40025000

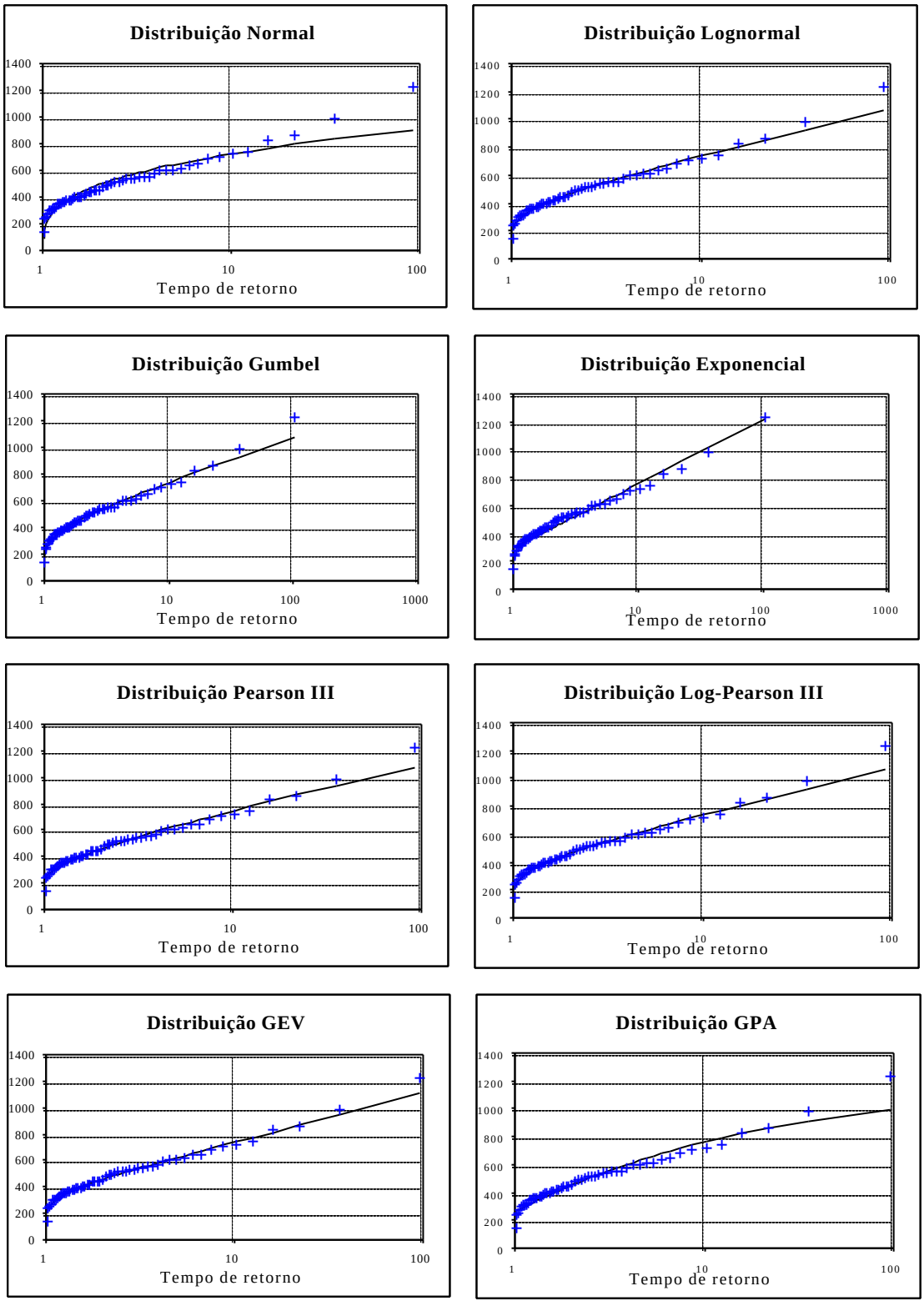



Figura A5.12: Ajuste visual dos dados de vazão da estação 40050000

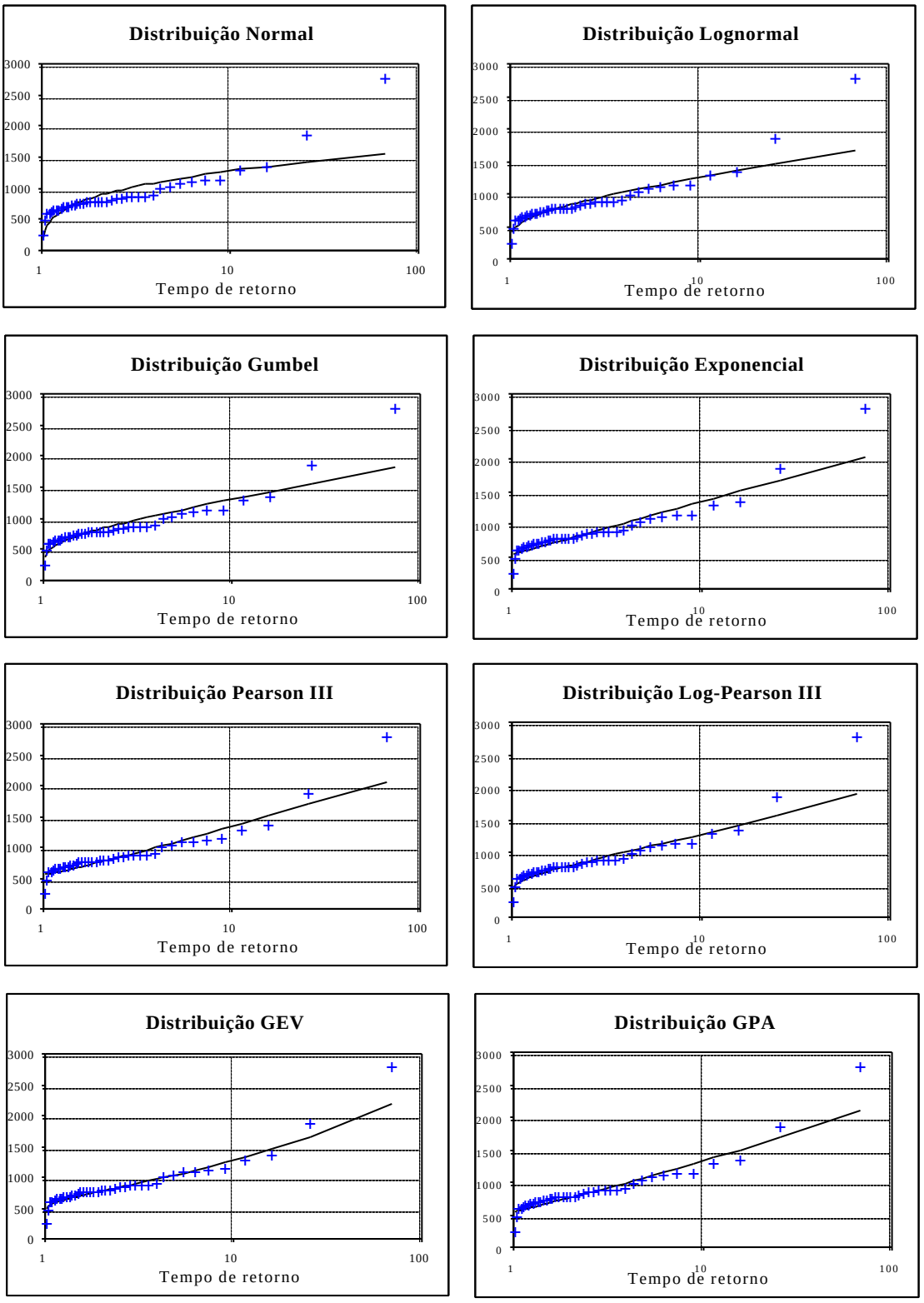



Figura A5.13: Ajuste visual dos dados de vazão da estação 40100000

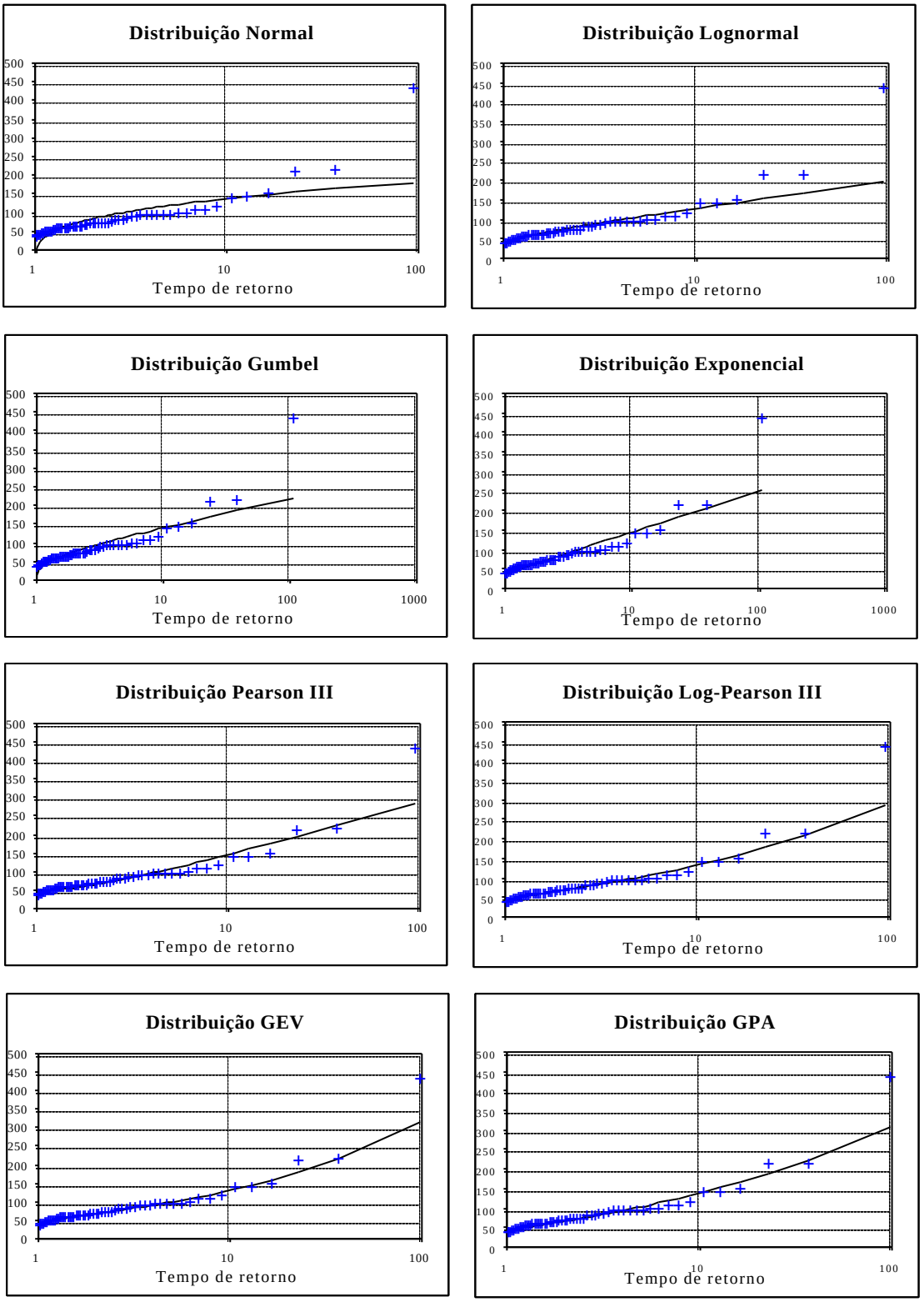



Figura A5.14: Ajuste visual dos dados de vazão da estação 40680000

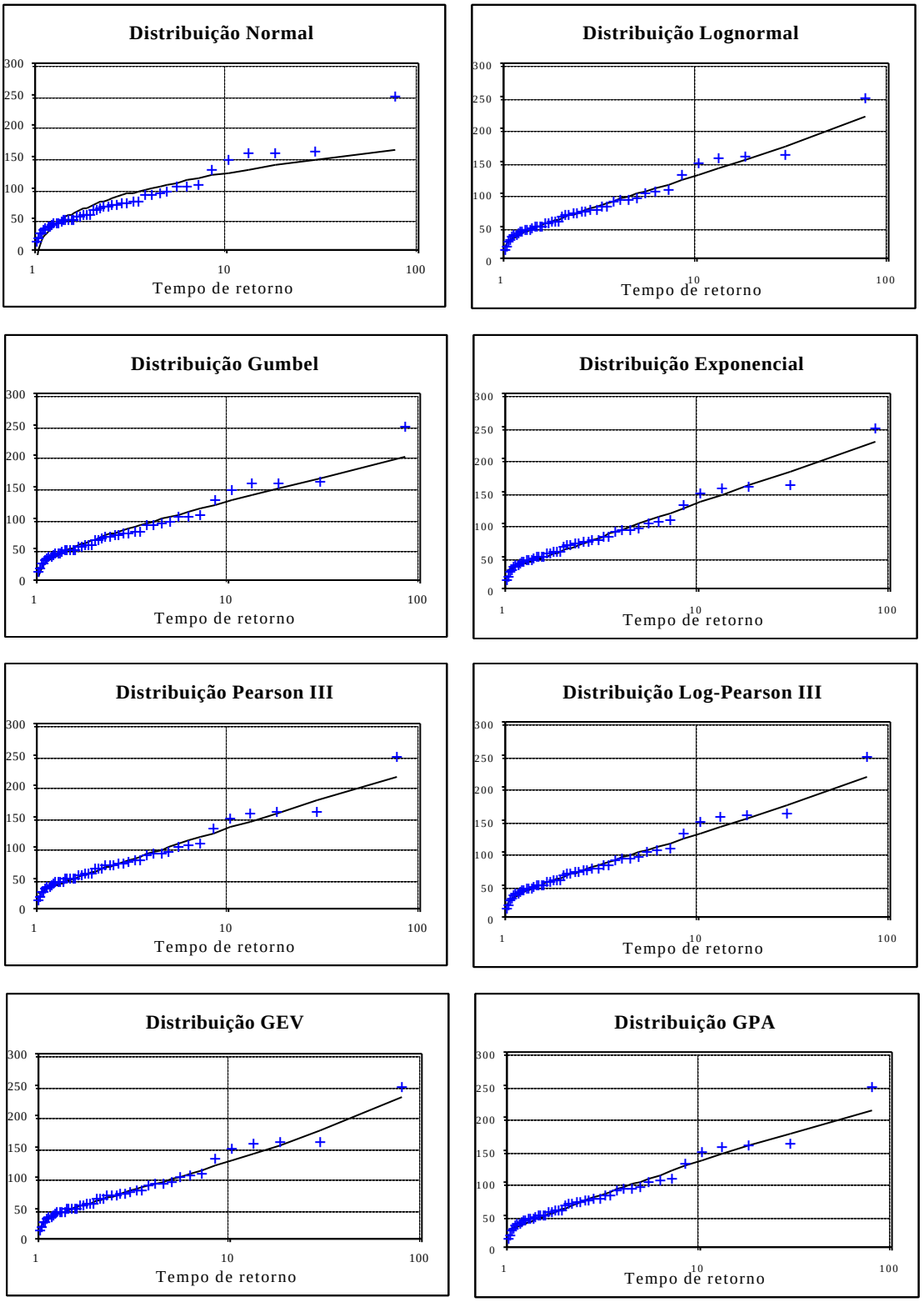



Figura A5.15: Ajuste visual dos dados de vazão da estação 41250000

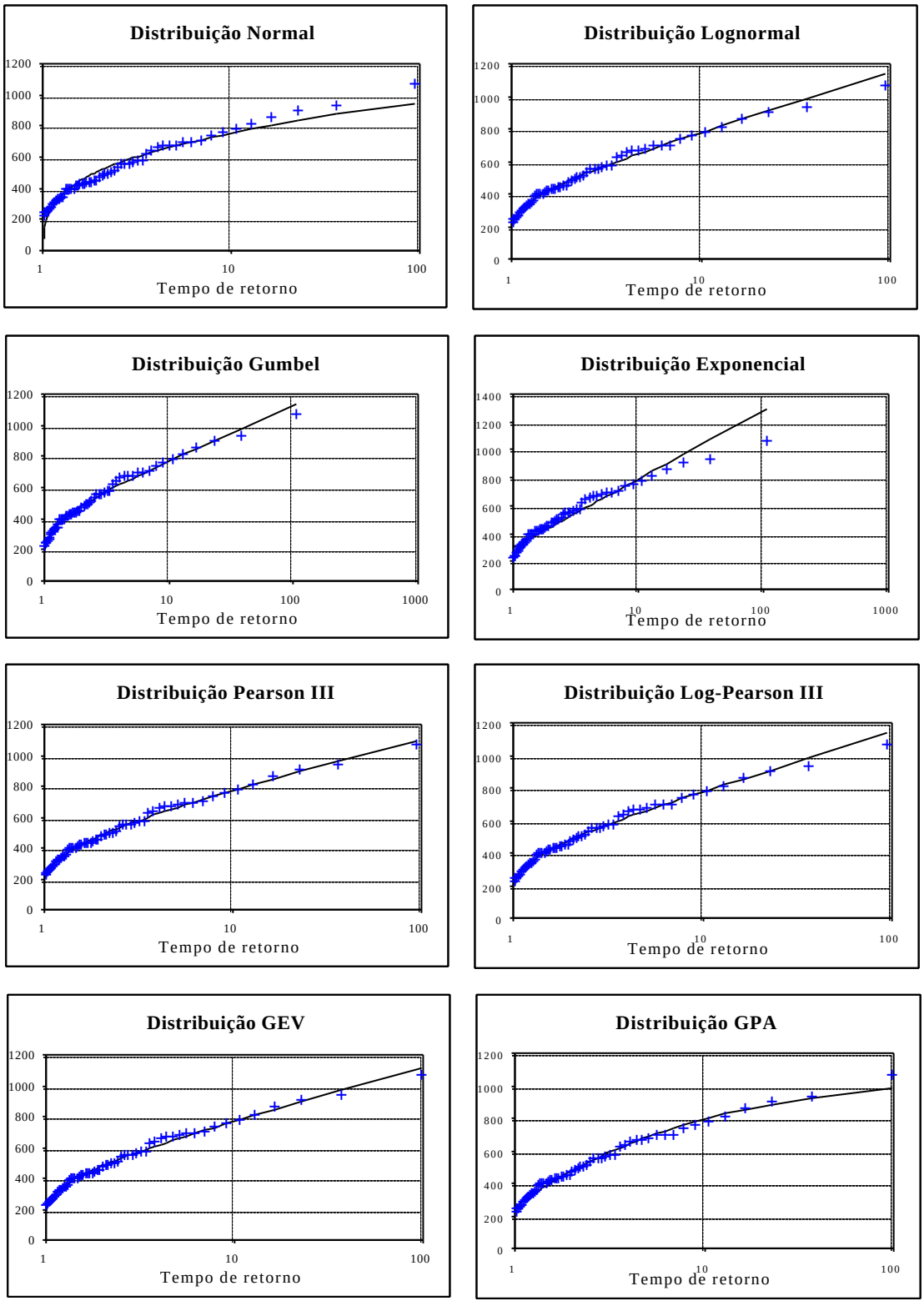



Figura A5.16: Ajuste visual dos dados de vazão da estação 40800001

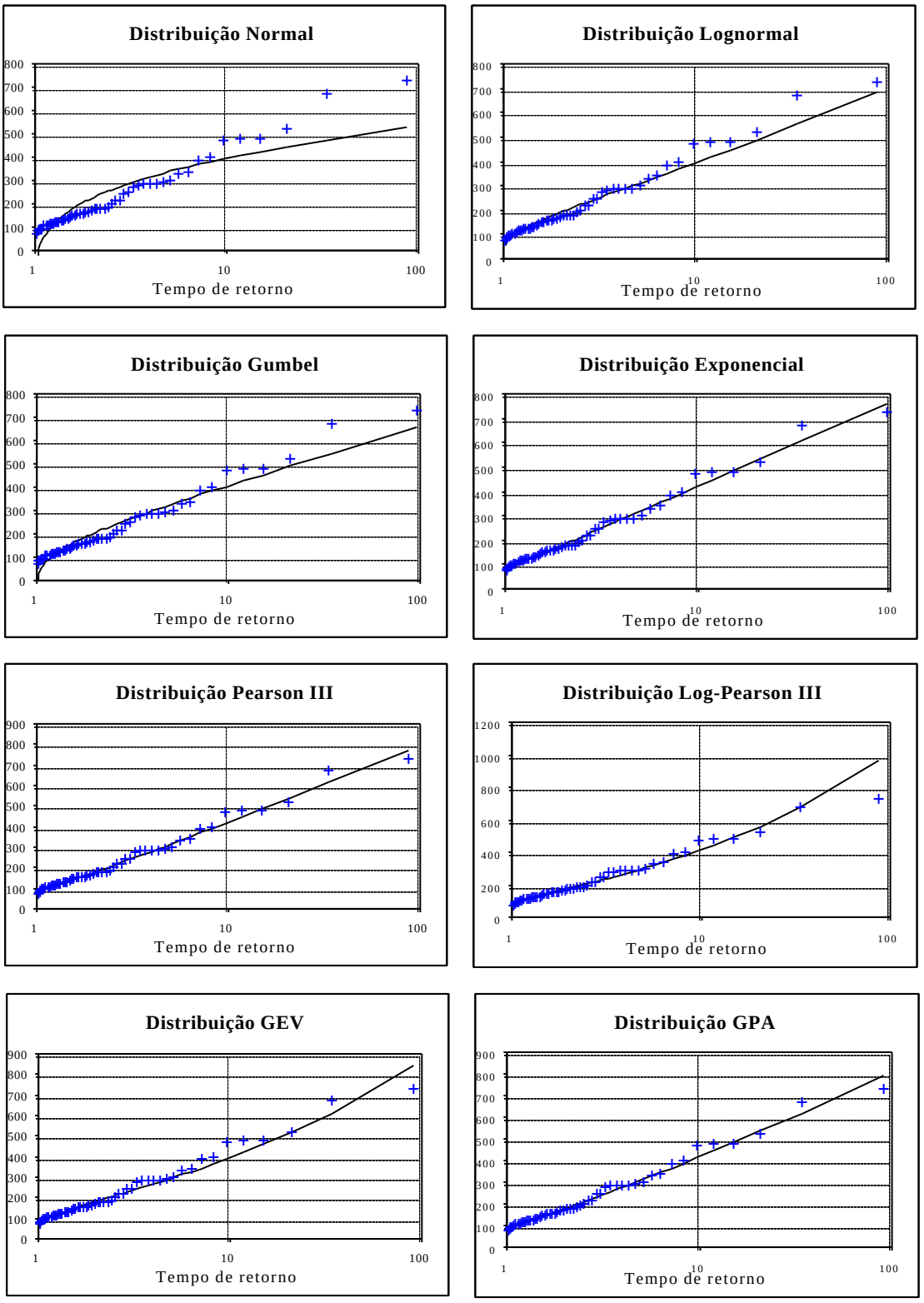



Figura A5.17: Ajuste visual dos dados de vazão da estação 56028000

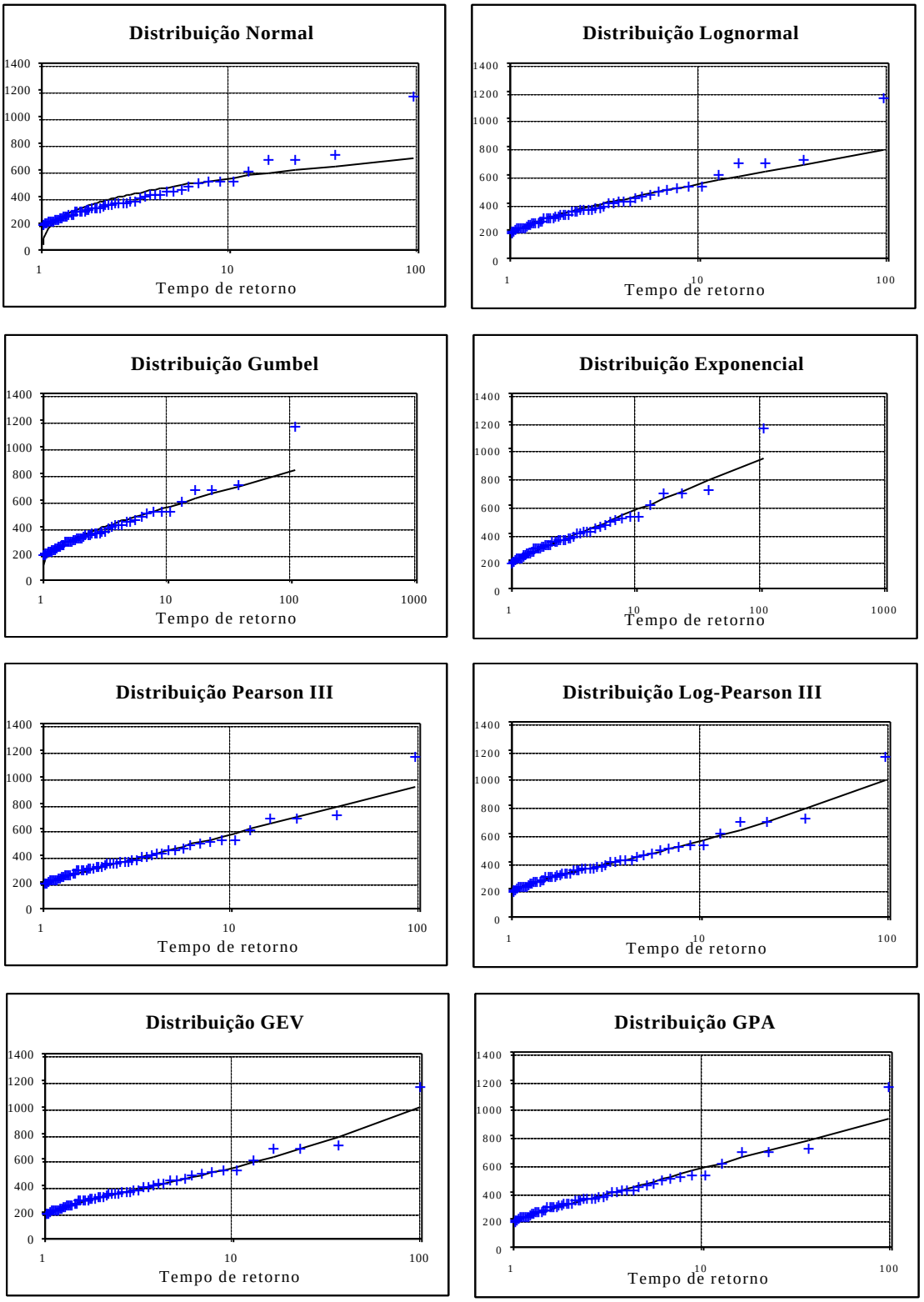



Figura A5.18: Ajuste visual dos dados de vazão da estação 56075000

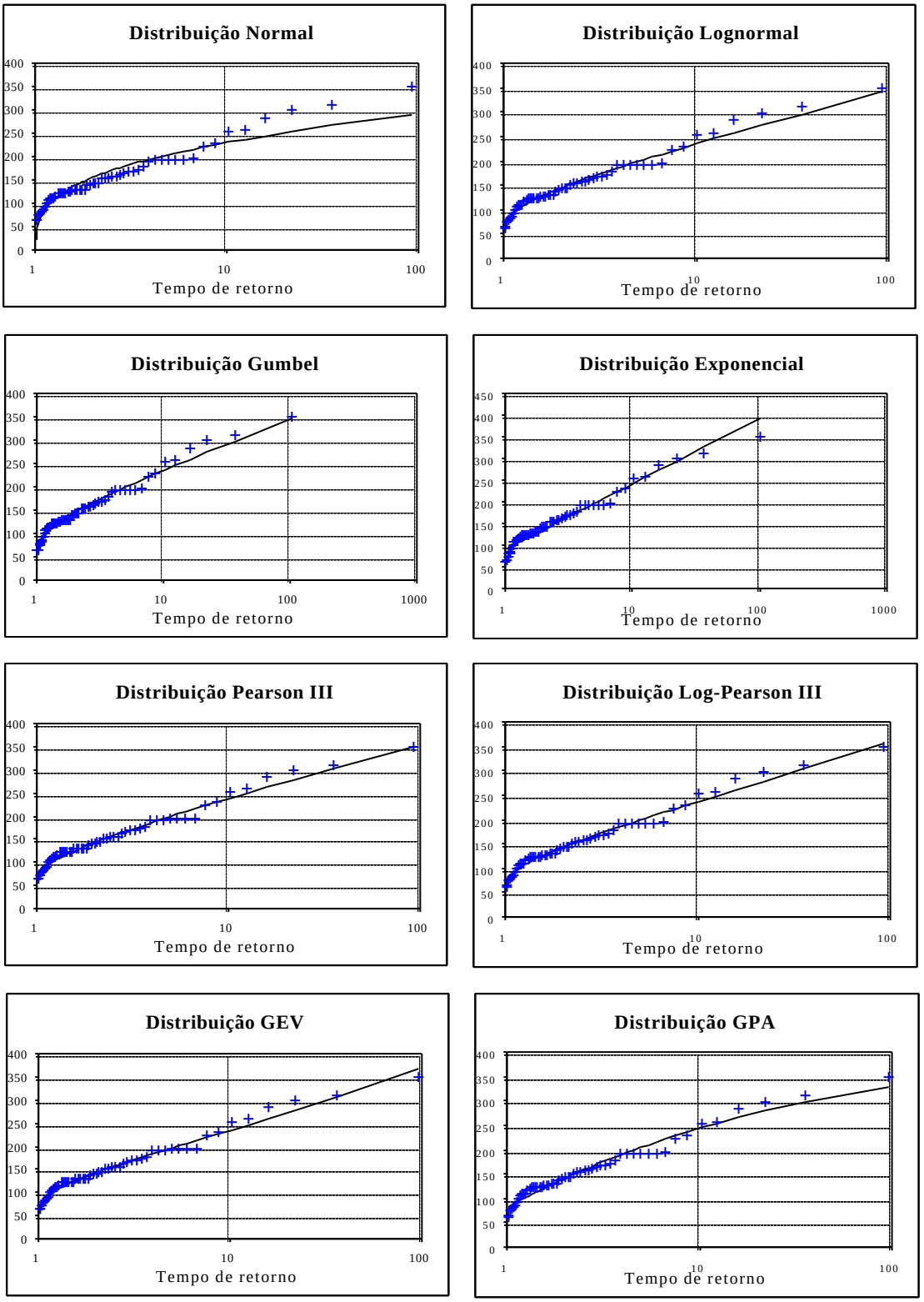



Figura A5.19: Ajuste visual dos dados de vazão da estação 56415000

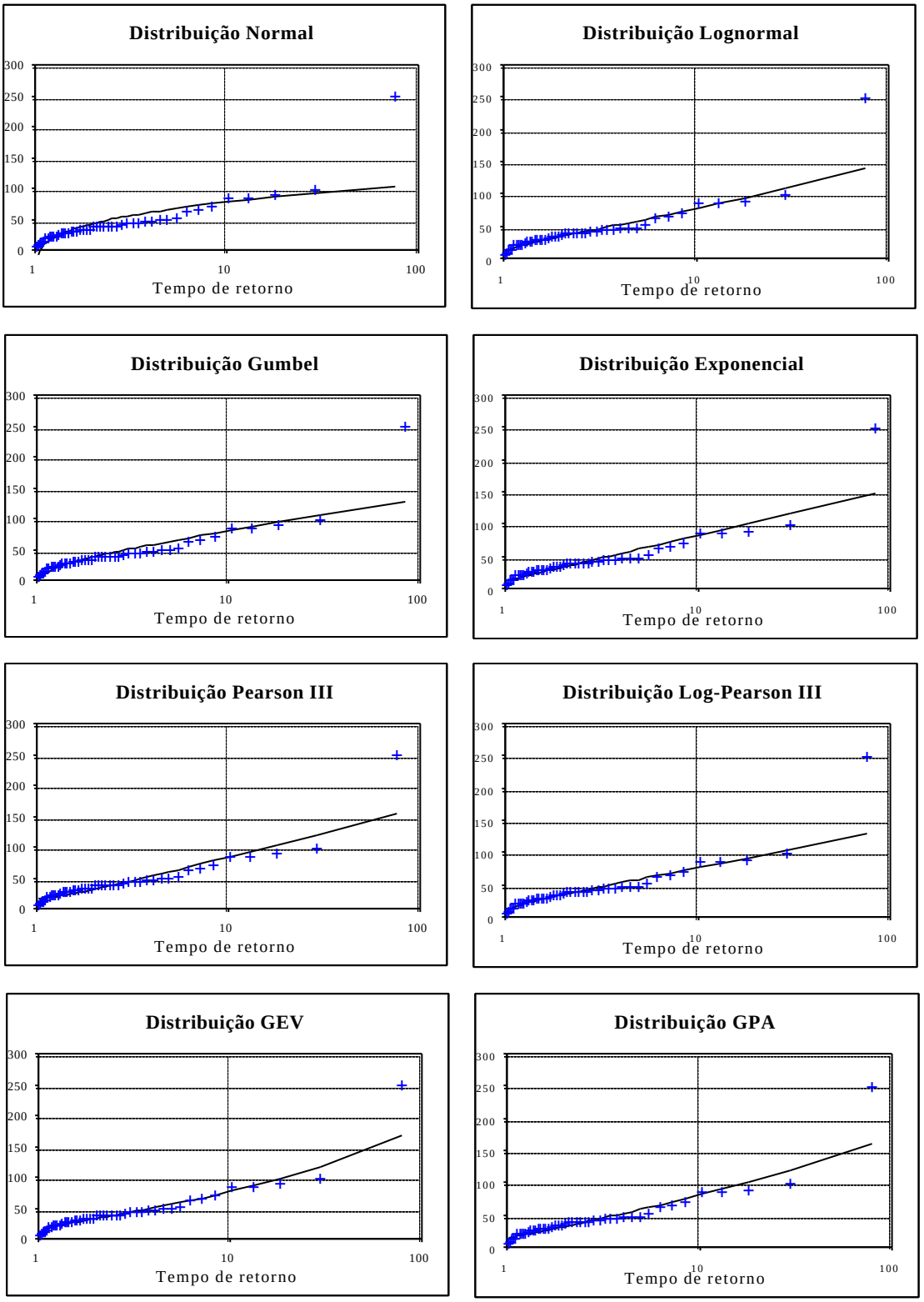



Figura A5.20: Ajuste visual dos dados de vazão da estação 56500000