UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE PESQUISA HIDRÁULICAS

MODELOS LINEARES GENERALIZADOS EM SIMULAÇÃO HIDROLÓGICA

EDUARDO SÁVIO PASSOS RODRIGUES MARTINS

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Recursos Hídricos e Saneamento Ambiental da Universidade Federal do Rio Grande do Sul como requisito parcial para a obtenção do título de Mestre em Engenharia.

Porto Alegre, Agosto de 1993

APRESENTAÇÃO

Este trabalho foi desenvolvido no Programa de Pós-Graduação de Recursos Hídricos e Saneamento Ambiental do Instituto de Pesquisas Hidráulicas da Universidade Federal do Rio Grande do Sul, sob a orientação do Prof. Robin Thomas Clarke.

Agradeço ao Prof. Robin Clarke pela orientação e cooperação dedicadas ao longo do desenvolvimento deste trabalho, sem as quais não seria garantida a qualidade final do mesmo.

À equipe da biblioteca, em especial Sras. Jussara Silva e Jussara Barbieri, sempre prestativas e pacientes ao uso persistente das facilidades da biblioteca.

Aos colegas do IPH, em particular Luis Carlos Brusa, Juan Carlos Bertoni e Josete de Fátima de Sá, pelo companheirismo, apoio e incentivo prestados no decorrer do curso.

À Mário de Castro Andrade Filho pela cooperação e intercâmbio de informações, além de ter possibilitado o uso do acervo do sistema de bibliotecas da UFRJ e IMPA.

À minha esposa Vládia pela paciência e dedicação durante a execução do trabalho, além da revisão e sugestões à forma final da dissertação.

Não posso deixar de reconhecer também o suporte do IPH que representou uma economia de tempo, facilitando muito o desenvolvimento do presente trabalho.

RESUMO

O projeto e operação de sistemas de recursos hídricos usualmente requerem simulação de sequências de vazão, a fim de que a frequência com a qual o sistema falha possa ser estimada para determinação das demandas hídricas. Um modelo de simulação freqüentemente empregado para sequências de vazões mensais é o modelo Thomas-Fiering, baseado em regressão linear. A aplicação deste modelo supõe a distribuição Normal dos desvios e a homogeneidade da variância. A falha destas suposições pode ser, algumas vezes, retificada por uma transformação Box-Cox da variável resposta o que pode prover benefícios adicionais de aproximações à normalidade e à aditividade. Nos anos recentes, uma família mais ampla de modelos (Modelos Lineares Generalizados), a qual inclui regressão linear como um caso particular, tem sido desenvolvida; esta dissertação explora o uso de MLGs em simulação de seqüências de vazão.

Em MLGs, a suposição de normalidade e homogeneidade de variância são relaxadas pela possibilidade de escolha de outra distribuição que a Normal e por permitir a aditividade dos efeitos sistemáticos na escala transformada, sendo a escolha da escala (função de ligação) independente da escolha da distribuição. Assim, um modelo adequado para modelagem de vazões mensais possivelmente seria o MLG log-Gama ou simplesmente Gama, o que provê um padrão de dispersão adequado das séries geradas a partir deste.

Embora as abordagens acima possam ser adotadas, freqüentemente é de interesse permitir explicitamente a heterogeneidade da variância (heterocedasticidade) na análise. A dificuldade é encontrar e ajustar um modelo satisfatório para variância. Logo, outra alternativa possível ao modelo Thomas-Fiering é o MLG Normal para a média das vazões mensais juntamente com o MLG log-Gama para a variância das mesmas.

Como às vezes torna-se necessária a geração de seqüências simultâneas de vazões mensais de vários postos, visando preservar a estrutura correlacional cruzada entre vazões mensais para cada par de postos, é proposto aqui uma abordagem multivariada para os modelos acima. Além disto, também é estudado o problema da geração de vazões mensais em rios intermitentes, sugerindo-se uma modificação no método de geração para levar em consideração não só a variabilidade das vazões mensais, mas também de sua ocorrência.

ABSTRACT

The design and operation of water resource systems usually requires simulation of river flow sequences, in order that the frequency can be estimated with which the system fails to meet water demands. A common simulation model for monthly flow sequences is the Thomas-Fiering model, based in linear regression. The application of this model assumes Normal distribution of deviations and homogeneous variance. Any failure in these assumptions may, sometimes, be corrected by a Box-Cox transformation of the response variable, which provides additional benefits of approximations to normality and additivity. In recent years, a much broader family of models (Generalized Linear Models) which include linear regression as a particular case, has been developed; this thesis explores the use of GLMs for simulating sequences of river flows.

In GLMs, the assumptions of normality and homogeneous variance are relaxed by the possibility of choosing a distribution different from the Normal one, and allowing additivity of systematic effects in the transformed scale, and the choice of distribution. Thus, an adequate model for the modelling of monthly flows might be the log-Gamma or simply Gamma GLM, which supplies an adequate dispersion pattern of the series generated from this model.

Although the approaches shown above may be used, frequently it is useful explicity to allow heterogeneity of variance (heterocedascity) in analysis. The difficult lies in finding and adjusting a satisfactory model for variance. Therefore, another possible alternative to the Thomas-Fiering model is the Normal GLM for mean monthly flows, together with log-Gamma GLM for their variance.

Since it is sometimes necessary to generate simultaneous sequences of monthly flows from several stations, to preserve the cross correlation structure between monthly flows for each pair of stations, a multivariate approach to the models above is proposed here. The problem of monthly flow generation in intermittent rivers is also studied, and a modification in the generation method is suggested to take into account not only the variability of monthly flows, but also their occurrence.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

LISTAS DE SÍMBOLOS

MLG	modelo linear generalizado
$\underline{y}$	vetor da variável resposta
$\underline{Y}$	vetor de variável aleatória, população de $\underline{y}$
$\underline{\mu}$	vetor de médias ou valores ajustados ($\underline{\mu}$, $\underline{\mu}_k$, ...)
$\underline{\eta}$	preditor linear ($\underline{\eta}$, $\underline{\eta}_k$, ...)
$\underline{x}_i$	i-ésima observação do vetor de covariáveis
$\underline{X}$	matriz de covariáveis
$\underline{W}$	matriz diagonal de pesos iterativos ($\underline{W}$ e $\underline{W}_k$)
$\underline{H}$	matriz de projeção ou 'Hat'
$\underline{I}$	matriz identidade de uma dada ordem
$\underline{z}$	variável dependente ajustada ($\underline{z}$ e $\underline{z}_k$)
$\underline{\beta}$	vetor de parâmetros do MLG ($\underline{\beta}$, $\underline{\beta}_0$, $\underline{\beta}_1$ e $\underline{\beta}_k$)
$\underline{b}$	estimativa de $\underline{\beta}$
$\underline{\beta}_{max}$	vetor de parâmetros do modelo completo
$\underline{b}_{max}$	estimativa de $\underline{\beta}_{max}$
$\underline{\mu}^D$	valores ajustados para o modelo da dispersão
$\underline{\eta}^D$	preditor linear para o modelo da dispersão
$\underline{\gamma}$	vetor de parâmetros do modelo da dispersão
$\underline{z}_i$	i-ésima observação da matriz de covariância para o modelo de dispersão
ϕ	parâmetro de escala do MLG (ϕ, ϕ_1)
θ	parâmetro natural ou canônico
N	número de observações
p	número de parâmetros (p, p_1, p_2, ..., p_N)
$f_Y(.;.)$	função densidade de probabilidade ($f_Y(\underline{y};\theta,\phi)$, $f_Y(\underline{y};\underline{\beta})$)
a(.), b(.), c(.)	funções que determinam $f_Y(.;.)$
$\mathbb{L}(.;.)$	função verossimilhança ($\mathbb{L}(\underline{\beta};\underline{y})$, $\mathbb{L}(\underline{b};\underline{y})$ e $\mathbb{L}(\underline{b}_{max};\underline{y})$)
$l(.;.)$	função log-verossimilhança

$pl(.)$	*profile log likelihood*
λ_s	estatística razão de verossimilhança
D	*deviance* (D, D_0, D_1, ..., D_N)
d_i	contribuição da i-ésima observação a D
ΔD	diferença entre *deviances*
σ^2	variância (σ^2, σ_i^2 e σ_t^2)
k	parâmetro de forma da distribuição Gama (k, k_t)
χ^2	estatística generalizada Pearson
χ^2_{N-p}	valor crítico da distribuição quadrado (χ^2_{p-q})
$F_{p-q,N-p}$	valor crítico da distribuição F
M	modelos (M_1, M_2, ..., M_N)
H_0	hipótese nula
H_1	hipótese alternativa
r	resíduo (r_i)
r_P	resíduo Pearson
$\bar{r}_P$	resíduo Pearson padronizado
r_D	resíduo *deviance*
$\bar{r}_D$	resíduo *deviance* padronizado
r_A	resíduo Anscombe
r_G	resíduo entre $\bar{r}_P$ e $\bar{r}_D$
G_i	redução aproximada em D/ϕ quando o i-ésimo caso é deletado
A1, A2, A3	aproximações de um passo para mudanças na *deviance* quando um simples caso é deletado
C	estatística Cook
h_i	i-ésimo elemento da diagonal principal da matriz de de projeção $\underline{H}$ conhecido como *leverage*
v_i	quantidade diagnóstica
$v_{(i)}$	quantidade diagnóstica ordenada
q_i	quantis
s_i	valor exato esperado da i-ésima estatística ordenada para a distribuição Normal

$y^{(\lambda)}$	variável resposta transformada com a família Box-Cox de transformações
λ	parâmetro da transformação Box-Cox (λ, λ_t, λ_y)
$\hat{\lambda}$	valor estimado para λ (λ^1_*)
λ_L	parâmetro da função de ligação potência (λ^2_*)
Q^i_t	i-ésima observação da vazão no mês t
$(Q^i_t)^{\lambda_t}$	i-ésima observação da vazão transformada no mês t
ξ	variável Normal com média 0 e variância apropriada
U_1, U_2, W	variável unifomemente distribuida entre 0 e 1
V_1, V_2	variável unifomemente distribuida entre -1 e 1
X_1, X_2	variável Normal com média 0 e variância 1, N(0,1)
$N(\mu,\sigma^2)$	variável Normal com média μ e variância σ^2
$\mathcal{G}(\alpha,\beta)$	variável Gama com parâmetro de forma α e de escala β
$\mathcal{B}e(a,b)$	variável beta com parâmetros a e b
$\mathcal{B}(N,p)$	variável binomial em N experimentos com proporção de sucessos p
J(t)	estado da cadeia de Markov no mês t
N_{ij}	freqüências de transição
N_i	Número de dias em que J(t-1) = i
P_{ij}	probabilidades de transição
μ_{ij}	valor ajustado para P_{ij}
U_t	variável unifomemente distribuida entre 0 e 1 gerada no mês t
P_t	probabilidade de não ocorrer vazões nulas
$Q_{t-1}[I]$	vazão no ano I, mês t-1
π	3.1415...
Σ	somatório (Σ, $\sum_{k=1}^{m}$)
ln(.)	função logaritmica
exp(.)	função exponencial
sen(.)	função seno
cos(.)	função cosseno

$\Psi(.)$	função digma (Psi)
$\Gamma(.)$	função Gama
$\Phi(.)$	função Normal acumulado
$g(.)$	função de ligação
$h(.)$	inversa de $g(.)$
$V(.)$	função de variância
$A(.)$	função normalizadora do resíduo Anscombe, r_A
$t(.)$	função beta incompleta
$E[.]$	esperança de uma variável aleatória
$var[.]$	variância de uma variável aleatória
$cov[.,.]$	covariância entre duas variáveis aleatórias
Mediana[.]	mediana de uma variável aleatória
P	probabilidade ($P[J(t)=1]$, P_t, ...)

CAPÍTULO I A IMPORTÂNCIA DA SIMULAÇÃO ESTATÍSTICA NO PLANEJAMENTO E OPERAÇÃO DE RECURSOS HÍDRICOS

Um problema chave em projetos de aproveitamento hídrico ou operação destes é a consideração da variabilidade hidrológica. Abordagens quantitativas de tal variabilidade geralmente requerem um tratamento estocástico das afluências, já que as mesmas são funções da precipitação e outros processos os quais, no atual nível do conhecimento, são considerados aleatórios no tempo e espaço.

Além da variabilidade do processo, qualquer tentativa de medir tais variáveis envolvem erros que, em si, têm também componentes aleatórios. Pode-se notar que o conceito de probabilidade apresenta-se aqui de duas formas, intrinsicamente relacionadas, seja para descrever a referida variabilidade hidrológica, seja para representar a incerteza do conhecimento.

Geralmente são utilizados modelos de simulação/ otimização na elaboração de projetos de aproveitamento hídrico. As seqüências históricas de dados incluem poucos, ou até mesmo não incluem, eventos extremos que são importantes na avaliação da freqüência de falhas de um sistema sendo planejado. Portanto, procura-se um modelo que descreva as características aleatórias das seqüências históricas e usa-se este modelo, ou modelos, para simular o comportamento do sistema visando obter uma boa estimativa dos riscos envolvidos no projeto. Em outras palavras, a seqüência histórica pode prover apenas uma resposta do sistema, sendo pouco provável que a mesma repita-se no futuro o que leva a uma resposta futura do sistema não representativa.

Para ter-se idéia dos riscos envolvidos pode-se gerar várias seqüências de vazões, ou melhor, realizações do processo estocástico natural gerador da seqüência histórica, utilizando um modelo estocástico adequado. Uma vez obtidas, estas seqüências, não equiprováveis, podem ser utilizadas amplamente em conjunção com os referidos modelos de simulação

para obtenção do valor esperado da resposta e sua variabilidade. Com isto pode-se avaliar os riscos envolvidos na alternativa em estudo, bem como, comparar várias alternativas de projetos e estratégias.

Por exemplo, suponha que pretende-se planejar ou operar algum sistema de aproveitamento hídrico em um local com uma série histórica de N anos. Esta série é utilizada por um modelo de simulação/otimização visando a obtenção do volume necessário V^* para regularizar uma descarga Q^*. Utilizando-se um modelo estocástico adequado obtem-se, a partir da série histórica, outras possíveis realizações do processo estocástico natural distintas entre si, digamos m, com extensão igual ao horizonte de planejamento. Destas obtem-se, empregando-se o mesmo modelo de simulação/otimização já referido, a amostra aleatória V_j^* (j=1,...,m; m - número de séries geradas), representando o volume necessário para regularizar Q^* da série gerada j, ou melhor, cada série está associada a um tamanho de reservatório. Portanto, é possível inferir a distribuição probabilística de V^* a partir da amostra aleatória V_j^* (j=1,...,m+1), e, por conseguinte, pode-se definir o volume do reservatório associado a uma probabilidade α de atendimento à descarga Q^*, V_α^*.

O problema crucial de séries sintéticas é que as mesmas devem assemelhar-se à seqüência histórica em termos estatísticos (mesmo padrão de dispersão e outros) e hidrológicos (persistência e outros), ou seja, características de influência no sistema considerado.

Objetivos

Pretende-se utilizar modelos lineares generalizados (MLGs) em simulação hidrológica, procurando não só adaptar modelos já conhecidos a esta técnica, mas também solucionar algumas das deficiências dos mesmos. Mais especificamente estamos interessados em :

(i) modificação do modelo Thomas-Fiering, formulando-o como um modelo linear generalizado, para modelagem de vazões ao nível mensal, verificando se as estatísticas anuais são preservadas;

(ii) utilização de técnicas diagnósticas, apropriadas a MLGs (i), visando a identificação de *outliers* e medição da influência de simples casos;

(iii) uso de técnicas estatísticas eficientes na modelação de vazões em vários postos, ou seja, modelagem multivariada também relacionada ao item (i);

(iv) adaptação de (i) para modelagem de séries de vazões em rios intermitentes;

(v) finalmente, a modelagem da variância de vazões mensais, junto a modelagem de suas médias (i), o que provê um útil e sensível diagnóstico para o modelo da média de vazões.

No que diz respeito ao objetivo (iii), pretende-se introduzir técnicas estatísticas modernas de modo a dar um maior rigor no uso de modelos multivariados. Este rigor refere-se ao questionamento da forma do modelo, do uso da distribuição Normal na geração dos componentes aleatórios do mesmo, assim como da necessidade de incluir o mesmo número de termos em cada posto.

CAPÍTULO II REVISÃO BIBLIOGRÁFICA

II.1 - TRANSFERÊNCIA DE INFORMAÇÕES

Um dos problemas da geração de seqüências sintéticas de vazão é que nem sempre as seqüências históricas possuem igual extensão. Se isto ocorre, é necessário utilizar alguma técnica, por exemplo análise de regressão, com o objetivo de prover estimativas para registros de modo a obtermos seqüências de igual extensão, utilizando-se, assim, melhor a informação de tais registros. Contudo, estas estimativas podem levar a tendenciosidades nos parâmetros do modelo utilizado. Uma análise das vantagens e desvantagens do emprego destas técnicas é apresentada a seguir. Antes de prosseguir na discussão de tais técnicas, faz-se necessário definir alguns termos aqui utilizados.

Entende-se aqui por transferência de informações o preenchimento de falhas e a extensão de registros curtos. No que diz respeito à extensão de registros curtos, tem-se que observar que esta objetiva obter vetores de variáveis de mesmo comprimento, aplicada quando a análise realizada é de regressão múltipla ou multivariada. Esta é uma etapa de preparação dos dados para aproveitar toda informação disponível.

Na análise de registros hidrológicos, tais como precipitação e vazão, a primeira necessidade que surge é o preenchimento de falhas e a extensão dos registros curtos quando necessário.

Várias abordagens têm sido sugeridas na literatura para transferir informações. Algumas destas são clássicas em engenharia hidrológica, como citam LINSLEY, KOHLER e PAULHUS (1988) o método da razão normal e o método da interpolação da distância ponderada.

Além dos métodos clássicos existem os métodos estatísticos, entre os quais regressão linear simples, múltipla e modelos de séries temporais. A ênfase nestes métodos tem sido concentrada, pelo menos em textos

hidrológicos, no uso da regressão linear simples como em MATALAS e JACOBS (1964) e no uso da regressão múltipla como em GILROY (1970). Pouca ou nenhuma atenção tem sido dada no uso de técnicas multivariadas que, quando muito mencionadas, serão aqui empregadas para a transferência de informações, seja para preenchimento de falhas ou extensão de registros curtos.

BEALE e LITTLE (1975) apresentam os resultados computacionais para alguns métodos alternativos que analisam dados multivariados com falhas aleatórias. Eles recomendam para o caso multinormal um algoritmo devido a ORCHARD e WOODBURY (1972) apud BEALE e LITTLE (1975) que fornece estimadores de máxima verossimilhança. Eles também fornecem outro algoritmo, uma forma iterativa do método de BUCK (1960) apud BEALE e LITTLE (1975), que não assume a multinormalidade ou qualquer outra distribuição multivariada pertencente à família exponencial. Esta é a grande vantagem deste esquema que se torna proibitivo quando existem muitas falhas.

DEMPSTER, LAIRD e RUBIN (1977) apresentam uma abordagem geral para o cálculo das estimativas de máxima verossimilhança de dados incompletos. Esta técnica, chamada de algoritmo EM, consiste de um cálculo iterativo envolvendo dois passos, denominados passos de predição e estimação por JOHNSON e WICHERN (1992). A estimativa de máxima verossimilhança é possível ser obtida devido à suposição que a amostra segue distribuição conhecida pertencente à família exponencial. O algoritmo predição-estimação foi desenvolvido supondo também que as observações estão faltando aleatoriamente. Se esta suposição não é satisfeita, então o tratamento das observações que faltam, conforme o algoritmo EM, pode introduzir tendenciosidades sérias nos procedimentos de estimação. Outra desvantagem do algoritmo EM, além da necessidade de conhecer a distribuição seguida pela amostra, é que este algoritmo converge muito lentamente.

AITKIN et alli (1989) apresentam e discutem três métodos aproximados disponíveis e de uso comum em pacotes estatísticos, como no pacote GLIM 3.77 (1985). O primeiro simplesmente omite todas as observações que possuam falha em qualquer variável, utilizando apenas os registros completos. Esta abordagem pode resultar em uma redução drástica no número

de observações e também em uma tendenciosidade séria nas estimativas dos parâmetros. O segundo método estima média e variância de cada variável utilizando todas as observações completas de cada variável e a covariância entre cada par de variáveis das observações completas de ambas. Este método também não evita a tendenciosidade na estimativa dos parâmetros e pode resultar em uma matriz de covariância para as variáveis explanatórias que não seja positiva definida. E, finalmente, o último método discutido é como já visto em BEALE e LITTLE (1975) e DEMPSTER, LAIRD e RUBIN (1977), ou seja, métodos que preenchem falhas com valores imputados de modo a obter-se uma matriz completa dos dados.

O procedimento aqui adotado é o empregado por GENSTAT 5 (1989). Este é um procedimento similar ao de ORCHARD e WOODBURY (1972) apud BEALE e LITTLE (1975) e, apesar de não estar explícito, é razoável supor que o mesmo também não faz qualquer suposição sobre a distribuição da amostra, apenas exigindo que as falhas sejam aleatórias. O algoritmo estima as falhas para unidades em um conjunto de dados multivariados, usando uma técnica de regressão iterativa.

II.2 - MODELOS DE GERAÇÃO DE VAZÃO

As tentativas de gerar séries sintéticas de vazões não são recentes, remontam do início do século. Uma das primeiras é devida a HAZEN (1914) apud HURST (1951) que estudou a descarga de treze rios americanos, utilizando-se principalmente de papel probabilístico. O autor, para geração, utilizou-se de extrapolação da distribuição de freqüências normal e, também, da combinação de registros de vários rios.

Posteriormente, SUDLER (1927) apud HURST (1951) associou a cada registro de uma série anual observada a um cartão, sendo depois embaralhados para obter-se seqüências distintas de vazão. Apesar dos inconvenientes da metodologia, foi um grande avanço para época.

HURST (1951) examinou um grande número de seqüências de vazão, bem como seqüências de outros fenômenos naturais, mostrando que para uma seqüência de comprimento N, desvio padrão S e diferenças acumuladas da

média R desta seqüência, tem-se que $(R/S) \sim N^h$, onde estimativas de h têm média 0.7 e desvio padrão 0.08. Para eventos independentes, normalmente distribuidos com N grande, h=1/2. A tendência de que para seqüências hidrológicas obtenha-se estimativas de h maiores do que 1/2 e menores do que 1 tem sido referida na literatura como fenômeno Hurst. Esta medida de persistência, h, representou um grande avanço na compreensão do processo estocástico natural.

Além do estimador original de Hurst, denotado por K, vários outros estimadores têm sido propostos e utilizados em hidrologia estocástica, tais como o estimador H de MANDELBROLT e WALLIS (1969a) e WALLIS e MATALAS (1970) e o estimador H_n de SALAS e BOES (1974). Todos estes estimadores são transientes, ou seja, dependem do tamanho da série temporal N, e quando $N \rightarrow \infty$, todos convergem ao valor limite 1/2 como é mostrado em SALAS et alli (1979).

Na literatura três explicações têm sido dadas ao fenômeno Hurst, como cita O'CONNELL (1971): distribuição marginal, transitoriedade e correlação serial. As três hipóteses estão ligadas a suposição presente em HURST (1951) de h=1/2, ou seja, normalidade, N grande e independência serial. A primeira pode ser descartada com base nos estudos de MATALAS e HUZZEN (1967) e MANDELBROT e WALLIS (1969c). A segunda, intrinsicamente relacionada com a última, sugere que a seqüência observada não é 'suficientemente longa' para permitir que h atinja o seu valor assintótico de 1/2. WALLIS e MATALAS (1970) concluem que a dependência serial, refletindo a persistência de longo prazo, é uma explicação do fenômeno Hurst, enquanto assimetria não. No que diz respeito à assimetria, esta não é uma questão amplamente aceita.

A necessidade de preservar o fenômeno Hurst está diretamente ligada à sua relação com armazenamento de longo período. Preservar esta característica na geração significa obter séries sintéticas cujo valor seja idêntico ao obtido da série histórica. Como o fenômeno Hurst é transiente, tem-se que definir a extensão da série gerada. Neste sentido tem-se três posições: FIERING (1967) sugere que este período deve ser igual à vida útil do projeto, MANDELBROT e WALLIS (1968) afirmam que o mesmo deve ser bem

maior que esta vida útil e, finalmente, WALLIS e MATALAS (1971a) consideram mais razoável obter esta semelhança entre seqüências históricas e sintéticas com extensões comparáveis.

WALLIS e MATALAS (1971a) afirmam que para gerar seqüências de vazão que se assemelhem à seqüência histórica, em termos de h, é necessário utilizar modelos que não pertençam ao domínio Browniano de atração. Modelos que pertençam a este domínio levam a seqüências que se caracterizam por terem valores de h=1/2, não podendo também reproduzir a estrutura do correlograma das seqüências históricas. Fora deste domínio de atração MANDELBROT (1965) e MANDELBROT e VAN NESS (1968) apud KLEMEŠ (1974) destacam que é possível encontrar uma classe de processo com memória infinita capazes de explicar o fenômeno Hurst, denominada por eles *fractional Brownian noises* (fBn's). Uma interpretação hidrológica desta classe de modelos é fornecida por MANDELBROT e WALLIS (1968) e sua estrutura explorada em detalhes com introdução de aproximações discretas dos mesmos para facilitar sua simulação em MANDELBROT e WALLIS (1969a,b).

Além desta classe, outros modelos podem reproduzir o fenômeno Hurst, em particular certas versões de processos ARIMA, como em O'CONNELL (1971) apud KLEMEŠ (1974), e modelos formados pelas somas de processos *Broken line* como em RODRIGUEZ-ITURBE, MEJIA e DAWDY (1972). Estes últimos podem ser encarados como aproximação de fBn's como mostra MANDELBROT (1972), indicando que os fBn's parecem representar um processo mais geral capazes de exibir o fenômeno Hurst.

Para seqüências pouco longas, WALLIS e MATALAS (1970) mostraram que mudanças na correlação serial de retardo 1, ρ_1, podem levar a estimativas de $h \neq 1/2$. Assim, a estrutura Markoviana, ou de outras famílias de modelos de curta memória, pode preservar um estado transiente com $h \neq 1/2$, mas, em geral, não pode preservar simultaneamente ρ_1, como usualmente estimado, e um valor de h para um limite diferente de 1/2.

A consideração do coeficiente Hurst h oferece, conforme WALLIS e MATALAS (1971b), uma nova visão na interpretação dos correlogramas observados de seqüências de vazão. Os valores observados de h, maiores do

que 1/2, mostram que as estruturas dos correlogramas apresentam sintomas de persistência de longo período. Assim, tais correlogramas podem ser úteis indicadores da persistência de longo período enquanto testes formais possam fornecer resultados contrários.

A maioria das pesquisas anteriores têm associado o fenômeno Hurst, como já mencionado, com o armazenamento de longo período, mas KLEMEŠ (1974) mostra que este fenômeno não é necessariamente um indicador de memória infinita de um processo. O mesmo pode ser causado também pela não estacionariedade na média, por passeios aleatórios com uma barreira absorvente que surgem freqüentemente em sistemas de armazenamento natural, bem como devido a outras causas. O autor não questiona a importância operacional dos modelos fBn's e suas aproximações, mas inferências sobre características físicas de um processo, baseadas nestes modelos, podem não somente ser imprecisas, mas também rudemente mal tratadas.

Um dos primeiros modelos para geração de vazão mensal foi proposto por THOMAS E FIERING (1962) apud RAUDKIVI (1979). Neste método os N anos de registros são transformados em doze registros, um para cada mês, sendo aplicadas doze regressões lineares da vazão do mês em sua antecedente. O modelo assume uma persistência de retardo 1, ou melhor, de natureza Markoviana.

Várias modificações do modelo Thomas-Fiering foram realizadas na tentativa de eliminar algumas das deficiências do modelo. A transformação das vazões em seus logaritmos têm sido utilizada como forma de eliminar o inconveniente da geração de vazões negativas. Com o uso desta transformação, estaríamos preservando os parâmetros estatísticos dos valores logaritimizados de vazões e não das vazões.

Uma variante do modelo original é proposta por THOMAS e FIERING (1963) apud MAŢALAS (1967), em que as vazões seguem a distribuição gama com um coeficiente de assimetria γ. A estrutura deste modelo é similar ao original, a não ser pelo componente aleatório do modelo que segue a distribuição gama ao invés da normal.

Outra transformação utilizada, devida a MATALAS (1967), propõe a utilização de $Y = \ln(Q-a)$, onde a é o limite inferior das vazões, obtido a partir da série histórica. Com esta formulação são preservados o parâmetros estatísticos (média, desvio padrão, coeficiente de assimetria e correlação serial) das vazões e evitada a geração de vazões negativas, embora apareçam alguns problemas.

Os modelos de THOMAS e FIERING (1963) e MATALAS (1967) possuem o inconveniente de utilizarem o coeficiente de assimetria como parâmetro da distribuição. Este coeficiente é estimado fazendo-se uso de momentos estatísticos de ordem superior, o que, em função da extensão da série histórica de vazões, promovem uma instabilidade muito grande no seu valor.

Um aprimoramento do modelo de Thomas e Fiering foi sugerido por HARMS e CAMPBELL (1967) cujas características principais são a distribuição normal das vazões anuais, a distribuição log-normal das vazões mensais e a preservação da correlação entre vazões anuais e também entre as mensais.

CLARKE (1973) afirma que existem, pelo menos, duas formas de generalização do modelo Thomas-Fiering :

(i) pela inclusão de termos de maior retardo no lado esquerdo das equações que definem o modelo, tornando-se, assim, equações de regressão múltipla;

(ii) pela inclusão de outras variáveis (tais como vazões de outros postos fluviométricos, precipitação, e outras) no lado esquerdo das equações que definem o modelo.

A natureza de estudos hidrológicos é comumente multivariada, como sugere (ii). Sendo assim, seria necessário considerar conjuntamente dados de vários postos fluviométricos, ou ainda, incluir dados de postos pluviométricos como em LANNA (1971). Neste sentido MATALAS (1967) propõe um modelo para gerar seqüências sintéticas multivariadas que se assemelhem à seqüências históricas multivariadas em termos da média, desvio padrão, assimetria, correlação serial de retardo um e correlação cruzada de retardo

zero. Posteriormente, uma extensão deste modelo multivariado para casos de múltiplo retardo é proposta por PEGRAM e JAMES (1972).

Um dos grandes problemas da utilização de modelos de geração de vazão em rios intermitentes é a modelagem de vazões nulas. Neste sentido, uma das primeiras contribuições para a modelagem em rios de região semi-árida é devida a CLARKE (1973). No método proposto não se considera a correlação entre meses com vazão não nula e meses de vazão nula, gerando para estes casos uma variável *aleatória* normal com média e variância iguais às do respectivo mês quando este é o primeiro mês de ocorrência de vazão no ano. FILHO (1978) relaxa esta condição do esquema original de CLARKE (1973), não exigindo que o mês seja o primeiro de ocorrência de vazão no ano, obtendo melhores resultados para rios intermitentes no semi-árido nordestino. Esta metodologia foi empregada por SARMENTO (1989) na geração de vazões mensais para rios do semi-árido cearense, sendo apresentados resultados comparativos com modelos baseados na geração de vazões mensais a partir de vazões anuais, previamente geradas com um modelo apropriado.

Outras abordagens para o manuseio de vazões nulas são sugeridas por BEARD (1973) e LEE (1975) apud SRIKANTHAN e McMAHON (1980). LEE (1975) apud SRIKANTHAN e McMAHON (1980), como CLARKE (1973), assume a ausência de correlação entre meses de vazão e meses com vazão nula, mas também considera uma relação de regressão entre o volume total escoado e sua duração.

Com exceção do modelo sugerido por HARMS e CAMPBELL (1967), os modelos, até agora mencionados, para geração de séries sintéticas de vazão mensal são elaborados de modo a preservarem somente estatíticas mensais, não sendo assegurado que as estatísticas anuais sejam preservadas. Logo, modelos, conhecidos como modelos de desagregação, têm sido desenvolvidos de modo a garantir a preservação de estatísticas em mais de um nível de agregação.

Uma das principais contribuições ao uso de técnicas de desagregação é devida a VALÊNCIA e SCHAAKE (1973), que desagregando vazões anuais, previamente geradas, em vazões semestrais, mensais, semanais e

diárias, garante que as estatísticas nestes diferentes níveis sejam preservadas. Este modelo também pode ser empregado a seqüências multivariadas anuais, geradas por qualquer modelo de geração de vazões anuais. A maior deficiência deste método é que as estatísticas são preservadas somente dentro do ano, não sendo garantido que, por exemplo, seja preservada a covariância entre dezembro de um ano e janeiro do ano seguinte.

MEJIA e ROUSSELLE (1976), visando suprir a deficiência do método de Valência e Schaake, propuseram uma modificação no referido método pela inclusão de um termo extra que representa a covariância entre as vazões de um ano e o subseqüente. Contudo, uma inconsistência na formulação de Mejia e Rousselle foi descoberta por LANE (1982) apud STEDINGER e VOGEL (1984), sendo esta, posteriormente, amplamente discutida por STEDINGER e VOGEL (1984).

Tanto o método de Valência e Schaake e o de Mejia-Rousselle têm um inconveniente de possuirem um número excessivo de parâmetros. Visando reduzir o tamanho das matrizes envolvidas e, assim, o número de parâmetros a serem calculados BRAS e RODRIGUEZ-ITURBE (1985) sugerem que a desagregação seja realizada por etapa.

Seis procedimentos para geração de vazões mensais são comparados por SRIKANTHAN e McMAHON (1980), aplicado a 8 rios intermitentes australianos. Os resultados sugerem que o método dos fragmentos é o mais recomendado, embora vários outros, dos procedimentos analisados, apresentam resultados satisfatórios.

II.3 - MODELOS LINEARES GENERALIZADOS

A teoria de modelos lineares generalizados (MLGs) foi formulada explicitamente por NELDER e WEDDERBURN (1972) e examinada cuidadosamente por McCULLAGH e NELDER (1989). Esta teoria tem fornecido um conceito útil e unificado para a identificação da estrutura de dados que seguem uma distribuição da família exponencial (veja pág. 29), permitindo uma

abordagem flexível ao ajuste de um modelo. Esta família com um parâmetro desconhecido tem propriedades interessantes para estimação e outros problemas de inferência, possuindo como membros as distribuições Normal, Binomial, Gama e Poisson. A estrutura linear destes modelos, combinada com a distribuição do erro dentro da família exponencial resulta na sua forma simples para a função de verossimilhança.

Os MLGs incluem, como afirma CORDEIRO (1986), uma grande variedade de modelos usuais que são casos particulares para análise de dados univariados. O autor cita entre estes:

(a) modelo clássico de regressão com erro normal;

(b) modelo log-linear aplicado à análise de tabelas de contigência;

(c) modelo logístico aplicado à análise de tabelas multidimensionais de proporções;

(d) modelo *probit* para estudo de proporções;

(e) modelo de análise de variância com efeitos aleatórios;

(f) modelo estrutural para dados com distribuição gama;

(g) análise de regressão não-simétrica;

(h) polinômios inversos;

(i) modelos de testes de vida.

Além destes, CORDEIRO (1986) afirma que outros podem ser definidos no contexto de MLGs, como por exemplo modelos BOX-COX (1964) e alguns de

séries temporais. Posteriormente discute-se algumas abordagens dos modelos BOX-COX (1964) no contexto de MLGs, e, referente aos modelos de séries temporais, pode-se citar o trabalho de JøRGENSEN (1989) que discute processos estocásticos como extensões de MLGs.

A importância da teoria de MLGs reside não só no fato de englobarmos um grande número de modelos, unificando-os em uma mesma teoria, mas extende-se também à unificação do método de inferência, ou análise de *deviance* (definida na pág. 31) como é conhecido, que generaliza a, já conhecida, análise de variância dos modelos normais.

As estimativas de máxima verossimilhança dos parâmetros podem ser obtidas, iterativamente, por mínimos quadrados, método este utilizado pelo sistema GLIM, AITKIN et alli (1989). Devido a relação com mínimos quadrados, diagnósticos análogos aos de regressão linear são disponíveis, embora apareçam complicações.

O uso de MLGs tornou-se muito comum nos últimos anos, surgindo a necessidade de estudar resíduos apropriados para propósitos de diagnósticos. Neste sentido COX e SNELL (1968) dão o primeiro tratamento sistemático da noção de resíduos generalizados, posteriormente pode-se citar o trabalho de PIERCE e SCHAFER (1986).

No trabalho de PIERCE e SCHAFER (1986) é discutido, principalmente, o uso de resíduos baseados na *deviance* para :

(a) identificar observações individuais pobremente ajustadas;

(b) exame dos efeitos potenciais de novas covariáveis ou efeitos não-lineares daquelas que já estão no modelo;

(c) testes de *goodness-of-fit*;

(d) diagnósticos de influência de pontos.

Outro aspecto importante na definição de um modelo linear generalizado é a escolha da função de ligação (veja pág. 29). PREGIBON (1980) propõe procedimentos analíticos para exame da adequacidade da ligação assumida no ajuste de um modelo linear generalizado. Testes e técnicas de estimação são fornecidas através da expansão e linearização do modelo.

Em alguns casos esta função de ligação pode não ser conhecida exatamente, mas pode-se assumir pertencer àlguma forma paramétrica geral como mostram SCALLAN, GILCHRIST e GREEN (1984). Os autores sugerem uma maneira de se ajustar MLGs com tais funções de ligação. Poder-se-ia chamá-la de função de ligação Box-Cox.

AITKIN et alli (1989) fornecem uma alternativa a esta metodologia para determinar o parâmetro da função de ligação λ. É uma técnica iterativa que determina, para cada valor de λ, os parâmetros do modelo e o parâmetro de dispersão. Uma vez determinados estes, substitui-se os mesmos na função de verossimilhança do modelo, sendo o valor de λ o que maximiza esta função.

Outra abordagem, também utilizando a função Box-Cox, é sugerida por JøRGENSEN (1984). A diferença para abordagem acima é que o problema, aqui resultante, é de mínimos quadrados não-linear pois o parâmetro de Box-Cox é determinado juntamente com os parâmetros do modelo. Comparações deste modelo Box-Cox com a apresentada por AITKIN et alli (1989) sugerem uma boa concordância dos resultados.

II.4 - TÉCNICAS DIAGNÓSTICAS E TRANSFORMAÇÕES

Uma área que tem merecido muita atenção em estatística aplicada é a transformação de variáveis resposta para ajustar modelos lineares aos dados. Este problema foi abordado primeiramente por BOX-COX (1964) que propõem a estimativa de uma transformação indexada por um ou dois parâmetros, cujos objetivos seriam prover homogeneidade da variância,

aditividade e normalidade, embora nem sempre seja possível a obtenção simultânea dos mesmos. A necessidade ou não de realizar-se uma transformação na variável resposta revela-se pela análise dos resíduos de regressão e plotagem normal dos mesmos.

JOHN e DRAPER (1980) alertam que a melhor transformação de uma dada família nem sempre é satisfatória, pois pode-se ter escolhido uma família inadequada que não levará a resultados úteis. Pode ser, por exemplo, que os dados já pertençam a uma distribuição simétrica não normal, não necessitando corrigir a assimetria cuja transformação de potência tenta remover.

EFRON (1982) discute aspectos fundamentais sobre a teoria de transformações, tais como a existência de uma transformação monótona que promove a normalização de uma dada variável aleatória, e considera a relação entre normalidade e estabilização da variância. O autor deixa claro que os cálculos desenvolvidos não são conclusivos, apenas pretendem alertar contra o uso indiscriminado da normalidade como critério para uma transformação de sucesso.

A análise dos dados transformados é discutida por HINKLEY e RUNGER (1984) tendo como ponto chave o significado dos parâmetros do modelo linear. Os autores esclarecem que este significado tem sentido apenas com escalas específicas, levando a concluir que inferências sobre parâmetros de modelos devem referir-se a escalas já especificadas, não permitindo a seleção das mesmas com a utilização dos dados. RUBIN, na discussão do mesmo artigo, afirma que apesar de ser geralmente apropriado tratar a escala escolhida para definir estimandos como fixa, é geralmente inapropriado considerar fixa a escala para normalizar os dados.

Métodos diagnósticos são utilizados para verificação das suposições feitas na modelagem e também para localizar características não comuns dos dados que podem levar a conclusões errôneas. A literatura atual, como será visto a seguir, enfoca principalmente a detecção de observações influentes. COOK e WEISBERG (1986) e ATKINSON (1987) fornecem uma revisão do assunto.

Diagnósticos para verificação das suposições na modelagem não têm recebido a mesma atenção, embora sendo de igual importância. Como exceção pode-se citar COOK e WEISBERG (1983) que fornecem técnicas diagnósticas para acessar à validade da suposição usual de homocedasticidade.

ATKINSON (1981,1982) descreve o uso de plotagens diagnósticas para observações influentes ou *outliers* e plotagens para avaliarem o efeito de observações individuais na transformação estimada. Além disto ATKINSON (1982) estuda uma estimativa rápida para o parâmetro da transformação, e, finalmente, extensões destas ferramentas diagnósticas são utilizadas para testar funções de ligação em um modelo linear generalizado.

O autor apresenta um exemplo que leva a escolha entre tranformação da resposta, modelo Lognormal, e um modelo linear generalizado log-Gama. Para isto emprega uma forma gráfica de teste para hipóteses separadas.

Diagnósticos de regressão para modelos lineares clássicos estão bem estabelecidos na literatura. Como exemplo pode-se citar COOK (1977,1979) que considera o problema de detectar observações que são influentes em termos de seus efeitos na estimativa dos parâmetros. Muitos destes diagnósticos usam estatísticas que medem os efeitos de deleção de simples pontos dos dados. Estas estatísticas exploram a relação algébrica exata entre o ajuste de mínimos quadrados ao modelo linear de n pontos, e o ajuste a n-1 após a deleção de um simples ponto.

Embora o problema de localização e teste de um simples *outlier* já tenha sido amplamente resolvido, até mesmo em dados estruturados como os de regressão múltipla, múltiplos *outliers* apresentam dificuldades adicionais como afirmam HAWKINS, BRADU e KASS (1984). Esta dificuldade que *outliers* múltiplos podem causar, até mesmo em simples amostras aleatórias, foi reconhecida a muito tempo por PEARSON e SEKAR (1936) apud HAWKINS, BRADU e KASS (1984). Isto deve-se ao fato de que alguns *outliers* podem parecer *inliers*, bem como o inverso, podendo ocorrer, individualmente ou

simultaneamente, em amostras contendo múltiplos *outliers* e tornando-se até mesmo mais marcantes e intratáveis em conjuntos de dados estruturados.

Embora a estrutura geral de procedimentos diagnósticos para MLGs seja similar àquela dos modelos de regressão linear, existem algumas complicações. A estimativa de máxima verossimilhança da maioria dos MLGs requer métodos iterativos. A referida estimativa para n-1 pontos não pode ser obtida como função explícita dos resultados do ajuste a n pontos. PREGIBON (1981) deriva aproximações de um passo para as mudanças na estimativa de verossimilhança e da *deviance* do modelo quando um simples ponto é deletado. Além disso, o mesmo discute alguns métodos diagnósticos que usam estas aproximações.

As aproximações, acima referidas, são exploradas por WILLIAMS (1987) para as mudanças na *deviance* de um modelo linear generalizado quando um simples ponto é deletado. Estas aproximações sugerem que um conjunto particular de resíduos podem ser usados, não somente para identificação de *outliers* e exame de suposições de distribuições, mas também para calcular medidas de influência de simples observações em várias inferências que podem ser extraidas do modelo ajustado, usando a estatística razão de verossimilhança.

II.5 - HETEROCEDASTICIDADE

A suposição de mesma variância dos erros para todos os níveis das covariáveis é fundamental tanto para justificar a escolha da função dos erros a ser minimizada, quanto na derivação das propriedades amostrais dos estimadores. Em outras palavras, a homogeneidade da variância é uma das suposições padrões da análise de regressão Normal. A falha desta suposição pode ser, quase sempre, retificada pela transformação Box-Cox da variável resposta, como já comentado anteriormente, ou transformação das covariáveis, ou de ambas, o que provê benefícios adicionais de aproximações à aditividade e normalidade. Contudo, como isto nem sempre é possível, AITKIN (1987) sugere que às vezes é de interesse permitir explicitamente a heterogeneidade da variância na análise. O autor modela a heterogeneidade

da variância em análise de regressão Normal, utilisando um modelo de regressão log-linear para a variância.

SMYTH (1989) mostra como a estrutura dos MLGs pode tornar-se mais geral ainda pela consideração da estrutura da média e dispersão separadamente. O autor, aqui, não trata apenas do modelo Normal, mas também Gama e Gaussiana Inversa, sugerindo uma generalização para outras distribuições da família exponencial pelo uso de funções de quasi-verossimilhança.

MAK (1992) trata da modelagem de variância com o uso de modelos de regressão linear, utilizando uma função suave das variáveis regressores para a mesma. Para estimação dos parâmetros do modelo da dispersão o autor utiliza o método de mínimos quadrados ponderados iterativos.

O uso das equações de estimação de WEDDERBURN (1974) para os parâmetros do modelo da média são amplamentes aceitas. Para MLGs estas são equações de máxima verossimilhança ou de quasi-verossimilhança. Por outro lado, para estimação dos parâmetros do modelo de dispersão existe alguma controvérsia, como citam NELDER e LEE (1992). Os autores comparam funções de verossimilhança, quasi-verossimilhança e pseudo-verossimilhança, concluindo que a função máxima verossimilhança estendida é superior em termos erro quadrado médio padronizado.

II.6 - GERAÇÃO DE VARIÁVEIS ALEATÓRIAS NORMAL, GAMA E BINOMIAL

A simulação da variabilidade hidrológica naturalmente requer um mecanismo para gerar sequências de eventos, no caso vazões, onde cada seqüência obedece uma determinada lei probabilística. Esta lei probabilística pode tomar várias formas, em especial estamos interessados em vazões independentes e identicamente distribuidas conforme a distribuição Binomial, Normal e Gama.

Vários métodos que utilizam a transformação Box-Müller para geração de variáveis *aleatórias* Normais são discutidos em termos práticos e

teóricos por GOLDER e SETTLE (1976). Entre estes métodos pode-se citar: o convencional, permuta de Chay, gerador desordenado, Neave e o das duas seqüências. Os métodos apresentados diferem basicamente na maneira que são geradas as variáveis *aleatórias* Uniformes. A comparação entre os métodos é feita a partir do ajuste de distribuições teóricas. Os resultados da simulação de Monte Carlo sugerem que o método das duas seqüências é computacionalmente eficiente e não sofre qualquer influência de desvantagens inerentes ao método de geração das variáveis Uniformes. Este método também é indicado por FISHMAN (1973).

ATKINSON e PEARCE (1976) apresenta dez métodos para geração de variáveis *aleatórias* Normais, sendo divididos em métodos exatos e aproximados. Pela simplicidade e velocidade o autor conclui que o método Polar de Marsaglia é o melhor tanto para geradores lentos ou rápidos. Este método é uma modificação do Box-Müller, já mencionado, pela substituição das funções trigonométricas, sendo mais simples e apreciavelmente mais rápido.

FISHMAN (1973) apresenta dois métodos muito simples para geração de variáveis *aleatórias* Gama, o primeiro com parâmetro de forma inteiro e o outro real, devido a JONHK (1964) apud BOBÉE e ASHKAR (1991), para $\alpha > 0$. Este último e outros quatro métodos para geração de variáveis *aleatórias* Gama são comparados por ATKINSON e PEARCE (1976), obtendo como o mais rápido, exceto para alguns valores pequenos do parâmetro de forma, um método devido a FORSYTHE. Embora, em geral, seja o mais rápido, isto se faz às custas de extensos programas para armazenamentos de tabelas de constantes.

CHENG (1977) sugere um método de rejeição para gerar variáveis Gama exatas com parâmetro de forma α, onde $\alpha > 1$. Este método é comparado com alguns dos métodos já citados em termos de velocidade e simplicidade. Para parâmetros de forma até 1.5 o método de FISHMAN é sensivelmente mais rápido, ocorrendo uma inversão para valores superiores. CHENG e FEAST (1979) sugerem um algoritmo, também para $\alpha > 1$, muito mais rápido do que o de CHENG (1977) independente do valor de α.

Em um estudo sobre o uso da família Gama e distribuições derivadas em hidrologia, BOBÉE e ASHKAR (1991) sugerem a utilização do método de JONHK para $\alpha \leq 5$, enquanto que para $\alpha > 5$ a transformação Wilson-Hilferty pode ser empregada.

Para a geração de variáveis Binomiais, FISHMAN (1973) apresenta três métodos: o primeiro baseado na geração de variáveis Bernoulli, sendo muito simples e rápido para um número moderado de experimentos; o segundo baseado no método da transformação inversa, também rápido para um número moderado de experimentos; e, finalmente, outro método baseado em resultado assintótico que emprega a geração de variáveis N(0,1), fazendo que o tempo de geração seja bastante reduzido quando o número de experimentos é grande.

CAPÍTULO III METODOLOGIA

III.1 MODELOS LINEARES GENERALIZADOS

III.1.1 Introdução

A especificação de um MLG compreende (i) a definição de uma distribuição de probabilidade, membro da família exponencial de distribuições, para variável resposta, (ii) um conjunto de variáveis independentes, ou melhor covariáveis, descrevendo a estrutura linear do modelo e, finalmente, (iii) uma função que especifica a ligação entre a média da variável resposta e, a já mencionada, estrutura linear do modelo. Identifica-se, então, duas componentes nestes modelos, uma aleatória e outra sistemática.

Entre as distribuições da família exponencial pode-se citar: Normal, Poisson, Binomial, Gama, Normal Inversa e Binomial Negativa. No contexto de simulação hidrológica, três delas são de interesse especial: a Normal, a Gama e a Binomial. As duas primeiras utilizadas na modelagem das quantidades, enquanto a última na modelagem da ocorrência de variáveis hidrológicas.

Admite-se que o vetor de observações $\underline{y}=[y_1, y_2, \ldots, y_N]^T$ é uma realização da população, aqui identificada pelo vetor de variáveis aleatórias $\underline{Y}=[Y_1, Y_2, \ldots, Y_N]^T$, independentemente distribuidas (família exponencial), com o vetor de médias $\underline{\mu}=E[\underline{Y}]=[\mu_1=E[Y_1], \mu_2=E[Y_2], \ldots, \mu_N=E[Y_N]]^T$ constituindo a parte sistemática do modelo e, ainda, vetor de variâncias $var[\underline{Y}]=[\phi.V(\mu_1), \phi.V(\mu_2), \ldots, \phi.V(\mu_N)]^T$, sendo $V(.)$ uma função denominada função variância e ϕ um parâmetro de dispersão.

Assume-se ainda a existência de covariáveis $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_p$ com valores conhecidos tais que

$$\eta_i = g(\mu_i) = \underline{x}_i^T . \underline{\beta}, \qquad (3.1)$$

onde g(.) é uma função diferenciável e monótona, conhecida como função de ligação, fazendo, como já mencionado, o elo entre o preditor linear η_i (estrutura linear) e o valor esperado de Y_i, μ_i. Portanto, as médias são, em geral, não lineares no conjunto de parâmetros da estrutura linear. Na tabela 3.1 apresenta-se algumas funções de ligação.

Tabela 3.1 Funções de ligação

Nome	g(.)
identidade	$\eta = \mu$
log	$\eta = \ln[\mu]$
log log complementar	$\eta = \ln[-\ln[1-\mu]]$
logit	$\eta = \ln[\mu/(1-\mu)]$
Normal acumulada inversa	$\eta = \Phi^{-1}(\mu)$ *
potência	$\eta = \mu^{\lambda}$

$$* \ \Phi\left[\frac{x-\mu}{\sigma}\right] = \frac{1}{\sigma.\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{(s-\mu)^2}{2.\sigma^2}\right] ds$$

A componente aleatória de um MLG considera que Y_i tem função densidade de probabilidade, segundo notação de McCULLAGH, NELDER (1989), na família exponencial

$$f_Y(y;\theta,\phi) = \exp\left\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\right\}, \qquad (3.2)$$

para algumas funções especificadas a(.), b(.) e c(.), e o parâmetro ϕ. Este, denominado parâmetro de escala ou dispersão, é suposto conhecido ou estimado iterativamente durante o ajuste do modelo, enquanto que o parâmetro θ, denominado parâmetro natural ou canônico, é desconhecido. As

funções a(.), b(.) e c(.), bem como a função variância, $V(.) = \frac{d\mu}{d\theta}$, estão apresentados na tabela 3.2. Considerações a respeito do parâmetro de escala ϕ serão abordadas no item III.1.6.

Tabela 3.2 Características das distribuições Normal, Binomial e Gama

Caract.	Normal	Binomial	Gama
y	$(-\infty,\infty)$	$\frac{0(1)n}{n}$	$(0,\infty)$
a(.)	$\phi = \sigma^2$	$1/n$	$\phi = k^{-1}$
b(.)	$(1/2).\theta^2$	$\ln(1+e^{\theta})$	$-\ln(-\theta)$
c(.)	$-\frac{1}{2}(\frac{y^2}{\phi} + \ln(2\pi\phi))$	$\ln\left[\binom{n}{n.y}\right]$	$(k-1).\ln(y.k)+\ln(k)-\ln(\Gamma(k))$
μ=E[Y]	θ	$e^{\theta}/(1+e^{\theta})$	$-1/\theta$
V(.)	1	$\mu(1-\mu)$	μ^2

* FONTE McCULLAGH, NELDER (1989)

Nesta nova formulação o modelo linear clássico pode ser estruturado da seguinte forma:

(i) componente aleatória: a variável aleatória $\underline{Y}$ segue a distribuição Normal com variância constante σ^2 e $E[\underline{Y}] = \underline{\mu}$;

(ii) componente sistemática: preditor linear η formado pelas covariáveis $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_p$

$$\eta_i = \underline{x}_i^T.\underline{\beta} \ ;$$

(iii) função de ligação: identidade, $\underline{\mu} = \underline{\eta}$.

III.1.2 Deviance e χ^2

Ao ajustar um modelo aos dados, uma questão que surge é quão discrepantes os valores ajustados são dos observados, ou seja, está-se interessado na adequacidade do modelo na descrição dos dados. Para acessar esta adequacidade pode-se comparar, como definido por NELDER, WEDDERBURN (1972), a verossimilhança do modelo sob investigação com a verossimilhança do modelo completo ou saturado. Este último tem a mesma estrutura (componente aleatória, função de ligação) do primeiro, diferindo apenas pelo número de parâmetros que é igual ao número de observações, fornecendo, por isto mesmo, uma descrição completa dos dados (a não ser, talvez, pela distribuição assumida), embora não cumpra um dos objetivos da modelagem, sumarizar os dados. Tem-se, portanto, juntamente com estes dois modelos:

(i) modelo completo ou saturado como já definido;
(ii) modelo já testado, para o qual já foi procedida a análise de *deviance*;
(iii) modelo a ser testado ou sob investigação, para o qual procura-se acessar sua adequacidade pela análise de *deviance*;
(iv) e, finalmente, o modelo nulo, constituido de um parâmetro apenas, este representando um μ para todos os y's. Conseqüentemente, este modelo atribui toda a variação entre os y's à componente aleatória.

A função de verossimilhança, aqui denotada por $\mathbb{L}(\underline{\beta};\underline{y})$, é algebricamente igual a função densidade de probabilidade $f(\underline{y};\underline{\beta})$[1], embora mude a ênfase das variáveis aleatórias $\underline{y}$, com $\underline{\beta}$ fixo, para os parâmetros $\underline{\beta}$ com $\underline{y}$ fixo. Dentre todos os valores possíveis de $\underline{\beta}$, o estimador de máxima verossimilhança de $\underline{\beta}$ é aquele que maximiza a função de verossimilhança, ou melhor, $\mathbb{L}(\underline{b};\underline{y}) \geq \mathbb{L}(\underline{\beta};\underline{y})$ para qualquer que seja β.

Logo, sejam $\mathbb{L}(\underline{b}_{max};\underline{y})$ e $\mathbb{L}(\underline{b};\underline{y})$ as funções de verossimilhança

---

[1] Por questão de simplicidade, trabalha-se aqui com a notação $f_Y(y;\underline{\beta})$ em vez de $f_Y(y;\theta,\phi)$ como em (3.2).

avaliadas nas estimativas de máxima verossimilhança $\underline{b}_{max}$ e $\underline{b}$ para os modelos completo e o em investigação, respectivamente. Quanto mais próximo (distante) $\mathbb{L}(\underline{b};\underline{y})$ esteja de $\mathbb{L}(\underline{b}_{max};\underline{y})$, mais (menos) adequadamente o modelo sob investigação descreve os dados. Isto sugere o uso da estatística razão de verossimilhança

$$\lambda_s = \frac{\mathbb{L}(\underline{b}_{max};\underline{y})}{\mathbb{L}(\underline{b};\underline{y})},$$

ou, equivalentemente, da estatística diferença entre as funções de log-verossimilhança $\ln(\lambda_s) = l(\underline{b}_{max};\underline{y}) - l(\underline{b};\underline{y})$. NELDER, WEDDERBURN (1972) definiram como estatística razão de log-verossimilhança, ou *deviance* como foi chamada,

$$D = 2.\ln(\lambda) = 2.[l(\underline{b}_{max};\underline{y}) - l(\underline{b};\underline{y})]. \qquad (3.3)$$

Assim, quanto melhor (pior) ajustado aos dados, com $l(\underline{b};\underline{y})$ mais próxima (distante) a $l(\underline{b}_{max};\underline{y})$, tem-se uma menor (maior) *deviance*. Na tabela 3.3 apresenta-se a *deviance* para as distribuição Normal, Gama e Binomial, devendo-se ressaltar que para distribuição Normal a mesma está relacionada com a soma dos quadrados dos desvios.

A *deviance* é uma medida de distância entre os valores ajustados e os observados, sendo sempre maior ou igual a zero, e o somatório de todas suas componentes mede a discrepância total entre o modelo corrente e o saturado.

Tabela 3.3 *Deviance*

Distribuição	Função*
Normal	$\phi.\Sigma_i(y_i-\hat{\mu}_i)^2$
Binomial	$2.\phi.\Sigma_i\{[y_i.\ln(y_i/\hat{\mu}_i]+(n-y_i).\ln[(n-y_i)/(n-\hat{\mu}_i)]\}$
Gama	$2.\phi.\Sigma_i[-\ln(y_i/\hat{\mu}_i)+(y_i-\hat{\mu}_i)/\hat{\mu}_i]$

* ϕ - PARÂMETRO DE ESCALA

Quanto mais covariáveis entrem na componente sistemática, menor a *deviance* D, tornando-se igual a zero para o modelo completo. Entretanto, um grande número de covariáveis, embora reduza a *deviance*, implica em um grau de complexidade na interpretação do modelo, devendo buscar-se modelos parcimoniosos, ou em outras palavras, modelos simples com desvios moderados, situados entre os modelos mais complexos e os que não descrevem adequadamente os dados.

Logo, torna-se necessário testar a adequacidade do modelo sob investigação, e para isto é preciso comparar a *deviance* e os graus de liberdade com alguma distribuição de probabilidade que sirva de referência. Apresenta-se, então, um problema: qual a distribuição da *deviance* ? Para responder a esta pergunta, DOBSON (1990) reescreve (3.3) da seguinte forma

$$\begin{aligned} D = 2.\{&[l(\underline{b}_{max};\underline{y}) - l(\underline{\beta}_{max};\underline{y})] \\ &- [l(\underline{b};\underline{y}) - l(\underline{\beta};\underline{y})] \\ &+ [l(\underline{\beta}_{max};\underline{y}) - l(\underline{\beta};\underline{y})]\}, \end{aligned} \qquad (3.4)$$

onde $\underline{\beta}$ e $\underline{\beta}_{max}$ são os parâmetros dos modelos sob investigação e completo, respectivamente, enquanto $\underline{b}$ e $\underline{b}_{max}$ suas estimativas. O autor mostra que o primeiro termo do lado direito de (3.4) segue a distribuição χ^2_N e,

similarmente, o segundo segue χ^2_p , enquanto que o último termo é uma constante positiva que será próxima de zero se o modelo sob investigação descreve adequadamente os dados. Logo, se as variáveis aleatórias definidas pelos dois primeiros termos forem independentes e o terceiro termo for próximo a zero temos que

$$D \sim \chi^2_{N-P}. \tag{3.5}$$

O resultado (3.5) é exato para o modelo Normal e aproximado para os demais modelos, devendo-se ressaltar que quando o modelo sob investigação não é adequado, a *deviance* não segue χ^2_{N-P}, sequer assintoticamente como afirma CORDEIRO (1986).

A estatística generalizada Pearson χ^2_p também pode ser utilizada como medida de adequação, sendo estimada por

$$\chi^2_p = \sum_i \frac{(y_i - \mu_i)^2}{V(\mu_i)}, \tag{3.6}$$

onde V(.) é a função variância, V(.), já apresentada na tabela 3.2. Apesar de mais fácil interpretação, χ^2_p não é aditivo para modelos encaixados[2] o que não ocorre com a função *deviance*. Em outras palavras, modelos com um menor número de parâmetros apresentam uma menor *deviance* o que não se verifica necessariamente para estatística χ^2_p.

---

2 Modelos encaixados são modelos que diferem um do outro pelo número p de covariáveis $\underline{x}_j$, j=1,...,p, na expressão $\eta_i = \underline{x}_i^T.\underline{\beta}$ e, por conseguinte, no número de parâmetros p. O vetor $\underline{x}_i^T$ sendo formado pelas i-ésimas observações das covariáveis $\underline{x}_j$, j=1,...,p, da mesma forma que em (3.1).

III.1.3 Análise de deviance

A análise de *deviance* ou ANODEV, como denomina CORDEIRO (1986), consiste em uma tabela de diferenças de *deviances* para uma seqüência de modelos encaixados. Nada mais é que uma generalização da análise de variância ou ANOVA, visando, como esta, obter os efeitos da inclusão de cada covariável e além disto suas interações. Surge porém uma dificuldade, em geral os termos de um MLG são não ortogonais[3] o que dificulta a interpretação da ANODEV.

Dada uma seqüência de modelos encaixados M_1, M_2, ..., M_N com a mesma estrutura (componente aleatória e função de ligação), diferindo apenas pelo número de parâmetros, tais que

$$p_1 < p_2 < \ldots < p_N, \tag{3.7}$$

conseqüentemente as *deviances* obedecem as desigualdades

$$D_1 > D_2 > \ldots > D_N. \tag{3.8}$$

Devido a esta propriedade da função *deviance* podemos comparar modelos encaixados utilizando a mesma, buscando não só obter modelos mais parcimoniosos, mas também avaliar a influência de algumas covariáveis na variação dos dados diante a presença de outras já incluidas.

Para comparar modelos encaixados considerar-se-á duas hipóteses: seja a hipótese nula H_0, representando um modelo mais simples, e a hipótese altenativa H_1, representando um modelo mais complexo. Temos então

---

3 Ortogonalidade pode ser assim interpretada: Considere o preditor linear $\eta=[(\beta_o.x_o+\ldots+\beta_p.x_p) + (\beta_{p+1}.x_{p+1}+\ldots+\beta_{p+q}.x_{p+q})]$ e denote por A e B os grupos de parâmetros $(\beta_o,\ldots,\beta_p)$ e $(\beta_{p+1},\ldots,\beta_{p+q})$, respectivamente. Se a redução na *deviance*, causada pelo ajuste de B quando A já foi ajustado, é igual a redução na *deviance*, causada pelo ajuste de A quando B já foi ajustado, A e B são ditos ortogonais.

$$H_o : \underline{\beta} = \underline{\beta}_o = \begin{bmatrix} \beta_1 \\ . \\ . \\ . \\ \beta_q \end{bmatrix}$$

e

$$H_1 : \underline{\beta} = \underline{\beta}_1 = \begin{bmatrix} \beta_1 \\ . \\ . \\ . \\ \beta_p \end{bmatrix},$$

onde $q < p < N$.

Pode-se testar H_0 contra H_1 usando a diferença de *deviances* do seguinte modo:

$$\begin{aligned} \Delta D &= D_o - D_1 \\ &= 2.[l(\underline{b}_{max};y) - l(\underline{b}_o;y)] - 2.[l(\underline{b}_{max};y) - l(\underline{b}_1;y)] \\ &= 2.[l(\underline{b}_1;y) - l(\underline{b}_o;y)], \end{aligned} \tag{3.9}$$

com $D_o \sim \chi^2_{N-q}$ e $D_1 \sim \chi^2_{N-p}$ se ambos modelos descrevem adequadamente os dados. Utilizando (3.5) temos que $\Delta D \sim \chi^2_{p-q}$, o que possibilita a comparação de ΔD com o valor crítico da distribuição qui-quadrado, χ^2_{p-q}. Logo, se o valor de $\Delta D > \chi^2_{p-q}$, para um determinado nível de significância α, não se pode ignorar os efeitos das (p-q) covariáveis, devendo-se rejeitar H_0 em favor de H_1, ou seja, $\underline{\beta}_1$ prover uma descrição melhor dos dados. Caso contrário, se $\Delta D \leq \chi^2_{p-q}$ deve-se escolher o modelo correspondente a H_o porque é o mais simples, ou mais parcimonioso.

Para eliminar ϕ nas expressões da *deviance* (tabela 3.3), pode-se utilizar a razão

$$F = \frac{\frac{D_o - D_1}{p - q}}{\frac{D_1}{N - p}}.$$

Este valor calculado é, então, comparado com o tabelado da distribuição $F_{p-q,N-p}$. Se $F \leq F_{p-q,N-p}$ os dados são consistentes com H_o e F segue, pelo menos aproximadamente, a distribuição $F_{p-q,N-p}$. Caso contrário, se $F > F_{p-q,N-p}$ H_o tem que ser rejeitada.

III.1.4 Resíduos

Resíduos são amplamente úteis para acessar a adequacidade do modelo e muito importantes para detectar observações influentes e *outliers*, bem como para o exame de suposições distribucionais. Enfim, o uso de resíduos facilita o exame de aspectos específicos do modelo sob investigação.

A discrepância entre o valor observado y_i e o ajustado μ_i é medida pelo resíduo r_i que pode ser expresso como uma função de y_i e μ_i, ou seja, $r_i = f_i(y_i;\mu_i)$. A forma da função f_i é escolhida de acordo com o objetivo da análise de resíduos, sendo usualmente definida de modo a obter-se homogeneidade da variância ou simetria na distribuição de r_i. A escolha de f_i deve ainda satisfazer propriedades de segunda ordem, tais como

$$E[r_i] = 0, \ var[r_i] = ctes. \ e \ cov[r_i;r_j] = 0 \qquad (3.10)$$

o que, em geral, sugere a forma da distribuição de r_i.

Plotagens de resíduos são diagnósticos muito importantes na

análise de casos. Quando o modelo não é adequado, algumas das propriedades dos resíduos falham, e estas falhas tornam-se aparentes na plotagem dos resíduos.

Entre os tipos de resíduos de uso mais comum em modelos lineares generalizados são: Pearson, Anscombe e *deviance*. O mais simples destes é o resíduo Pearson, definido como

$$r_P = \frac{y - \mu}{\sqrt{V(\mu)}}, \qquad (3.11)$$

onde V(.) é a função variância, já apresentada na tabela 3.2. Este tipo de resíduo apresenta como desvantagem sua distribuição bastante assimétrica para modelos não normais, podendo não satisfazer as propriedades (3.10) desejadas a todo tipo de resíduo.

ANSCOMBE (1953) apud CORDEIRO (1986) propôs um resíduo que vence as desvantagens do resíduo Pearson. O autor não compara os valores observados e calculados, mas uma transformação destes, aqui denotada por A(.), que objetiva normalizar a distribuição dos resíduos. Além disto, a variância de $A(y)-A(\mu)$ é estabilizada através de sua divisão pela raiz quadrada da função variância, onde A(.) é expressa por

$$A(.) = \int \frac{d\mu}{\left[V(\mu)\right]^{1/3}} \qquad (3.12)$$

Temos, então, usando (3.12), para a distribuição Normal a mesma expressão (3.11) e para a Gama

$$r_A = \frac{3.(y^{1/3} - \mu^{1/3})}{\mu^{1/3}} \quad . \qquad (3.13)$$

Já para a distribuição Binomial não se tem uma função explícita, tendo-se que integrar numericamente a função Beta incompleta. Este é da forma

$$r_A = \frac{t(y/m) - [t(p) + (p.q)^{-1/3}(2.p-1)/(6.m)]}{(p.q)^{1/6}/\sqrt{m}} \quad , \qquad (3.14)$$

onde $t(u) = \int_0^u s^{-1/3}(1-s)^{-1/3}\, ds.$

Outra classe de resíduos pode facilmente ser empregada, levando-se em conta que cada observação contribui com uma dada quantidade d_i para a *deviance* total do modelo sob investigação. As fórmulas de d_i para os modelos Normal e Gama podem ser facilmente obtidas da tabela 3.3, bastando-se para isto não considerar o sinal Σ. Pode-se então definir os resíduos *deviance* como

$$r_D = \pm \sqrt{d_i} \qquad (3.15)$$

tal que $\sum_i r_D^2 = D$. O sinal de (3.15) é o mesmo da diferença $(y_i - \mu_i)$. Este resíduo possui calculo muito simples, sendo aproximadamente equivalente ao resíduo Anscombe, já apresentado, o que pode ser verificado por expansão dos mesmos em séries de Taylor, como mostram McCULLAGH, NELDER (1989) e CORDEIRO (1986).

Outra vantagem de r_D é que se o modelo descreve bem os dados

temos, conforme (3.5), $d_i \sim \chi^2_1$, implicando em $\pm\sqrt{d_i} \sim N(0,\sigma^2)$, ou seja, este tipo de resíduo não necessita de uma função normalizadora.

III.1.5 Transformação Box-Cox da variável resposta e função de ligação potência

A suposição de normalidade da variável resposta pode ser verificada através de plotagens Q-Q (*Normal Plots*) e coeficiente de correlação Filliben. Quando se tem evidência de não-normalidade pode-se (i) assumir uma outra distribuição pertencente à família exponencial ou (ii) procurar uma transformação da variável resposta que leve simultaneamente a efeitos aditivos da componente sistemática do modelo e constância da variância na componente aleatória. Quanto a (ii) não se tem garantia que a aditividade e constância da variância sejam obtidas por uma mesma transformação, como afirma NELDER, WEDDERBURN (1972).

No contexto de MLGs a aditividade na parte sistemática do modelo é obtida pelo uso da função de ligação, e para a componente aleatória pode-se assumir outra distribuição, diferente da Normal, também pertencente à família exponencial, ou melhor, uma função de variância não constante. Embora a consideração de outra distribuição seja possível, nem sempre isto é o melhor, depende do problema e da facilidade de interpretação do modelo.

Logo, pode ser necessária a transformação da variável resposta y, sendo esta da seguinte forma

$$y^{(\lambda)} = \begin{cases} \dfrac{(y^{\lambda} - 1)}{\lambda}, & \lambda \neq 0 \\ \ln(y), & \lambda = 0 \end{cases} \qquad (3.16)$$

onde λ é o valor para o qual $y^{(\lambda)}$ tem distribuição Normal com média $\underline{b}'.\underline{x}_i$ e variância constante σ^2. Como a transformação é definida por um parâmetro,

λ, pode-se estimá-lo por máxima verossimilhança como indicado por AITKIN et alli (1989). Assim, assumindo para $\lambda \neq 0$ a aproximação $y^{(\lambda)} = y^{\lambda}$ e para cada valor de λ estimando os parâmetros $\underline{\beta}(\lambda)$ e $\sigma(\lambda)$ por máxima verossimilhança e substituindo na função log-verossimilhança para $\lambda \neq 0$

$$l(\underline{\beta},\sigma,\lambda) = -n.\ln(\sigma) + n.\ln|\lambda| + (\lambda-1).\Sigma \ln(y_i) - \Sigma (y_i^{\lambda} - \underline{b}'.\underline{x}_i)^2/(2.\sigma^2), \quad (3.17)$$

e para $\lambda = 0$

$$l(\underline{\beta},\sigma,0) = -n.\ln(\sigma) - \Sigma \ln(y_i) - \Sigma (\ln(y_i) - \underline{b}'.\underline{x}_i)^2/(2.\sigma^2) \quad (3.18)$$

resulta na função *profile log likelihood* maximizada em relação a $\underline{\beta}$ e σ com λ fixo, a menos de uma constante, para $\lambda \neq 0$

$$pl(\lambda) = -\frac{1}{2}.n.\ln(RSS(\lambda)) + n.\ln|\lambda| + (\lambda-1).\Sigma \ln(y_i), \quad (3.19)$$

e para $\lambda = 0$

$$pl(0) = -\frac{1}{2}.n.\ln(RSS(0)) - \Sigma \ln(y_i), \quad (3.20)$$

sendo $RSS(\lambda) = \Sigma \{y_i^{\lambda} - \underline{x}_i'.\underline{b}(\lambda)\}^2$. O valor de λ procurado é aquele correspondente ao máximo do gráfico de $-2.pl(\lambda)$ contra λ, com intervalo de confiança $100.(1-\alpha)\%$ aproximado consistindo dos valores de λ que estão a $\chi^2_{\alpha,1}$ unidades do referido máximo.

Como

$$\frac{1}{2} = P[f(y) \leq \underline{x}'.\underline{b}]$$

$$= P[y \leq f^{-1}(\underline{x}'.\underline{b})],$$

com $f(y) = y^{(\lambda)}$, segue que $f^{-1}(\underline{x}'.\underline{b})$ é a mediana de y, onde a distribuição de y é, em geral, assimétrica. Assim, na escala transformada λ, o modelo representa a variação na média da variável transformada $y^{(\lambda)}$, enquanto que na escala original representa a variação na mediana da variável y. Então, tem-se para $\lambda=0$

$$\text{Mediana}(y) = \exp(\underline{x}'.\underline{b}) \qquad (3.21)$$

$$E[y] = \exp(\underline{x}'.\underline{b} + \frac{1}{2}.\sigma^2)$$

$$\text{Var}[y] = \{\exp(\sigma^2)-1\}.\exp(2.\underline{x}'.\underline{b} + \sigma^2),$$

e para $\lambda\neq0$

$$\text{Mediana}(y) = \mu^{1/\lambda} \qquad (3.22)$$

$$E[y] \simeq \mu^{1/\lambda}.\{1+\sigma^2.(1-\lambda)/(2.\lambda^2.\mu^2)\}$$

$$\text{Var}[y] \simeq \mu^{2/\lambda}.\sigma^2/(\lambda^2.\mu^2).$$

Surge uma questão: como escolher entre a transformação da variável resposta e a função de ligação ? Para responder a esta pergunta, sejam os modelos sob investigação em termos de dois parâmetros, λ_L e λ_Y:

$$M1 : \lambda_L = 1; \quad \lambda_Y = \lambda_*^1$$

e

$$M2 : \lambda_L = \lambda_*^2;\ \lambda_Y = 1,$$

sendo M1 o modelo que faz uso da transformação da variável resposta cujo parâmetro é λ_Y e M2 o modelo com função de ligação potência e parâmetro λ_L, onde λ_*^1 e λ_*^2 são os valores estimados para λ_Y e λ_L, respectivamente. Logo, pode-se comparar o valor de $-2.pl(\lambda_*^1)$ para o modelo M1 com o valor $N.\ln(deviance)$ para M2, onde N é o número de observações. Este teste de razão de verossimilhança abilita-nos escolher entre a transformação da variável resposta e a função de ligação.

III.1.6 Considerações sobre o parâmetro de escala

O parâmetro de escala ϕ não é necessariamente o mesmo para todas observações, sendo possível admitir um parâmetro de escala para cada observação. Supõe-se aqui que este parâmetro é o mesmo para todas observações, em geral não conhecido a priori, sendo calculado, como já mencionado, a cada iteração do algoritmo de estimação.

O sistema GLIM fornece, quando não definido, um valor específico pelo usuário, como estimativa de ϕ

$$\frac{deviance}{N-p},$$

onde N é o número de observações e p o número de parâmetros do modelo sob investigação. Esta estimativa, embora não sendo de máxima verossimilhança, é consistente para o modelo Normal e coincide com a estimativa da variância, σ^2. CORDEIRO (1986) considera esta estimativa inconsistente para o modelo Gama quando os dados tenham uma grande dispersão, sugerindo como uma aproximação à estimativa de máxima verossimilhança o valor

$$\frac{2.deviance}{N}.\left\{1 + \left[1 + \frac{2.deviance}{3.N}\right]^{1/2}\right\}^{-1} .$$

No algoritmo implementado, as estimativas utilizadas para o parâmetro de escala foram a de máxima verossimilhança corrigida para a distribuição Normal $\frac{deviance}{N-p}$, e a de máxima verossimilhança para as distribuições Binomial (ϕ = 1) e Gama. Esta última não tem solução explícita, sendo obtida a partir da solução da seguinte equação

$$\ln(k) - \Psi(k) = \frac{deviance}{2.N}, \qquad (3.23)$$

onde $k = \phi^{-1}$ e $\Psi(k) = \frac{\Gamma'(k)}{\Gamma(k)}$ é a função digama, aqui estimada pelo algoritmo AS 103 de BERNARDO (1976). Para solução da equação (3.23) é utilizado o algoritmo da figura 3.1.

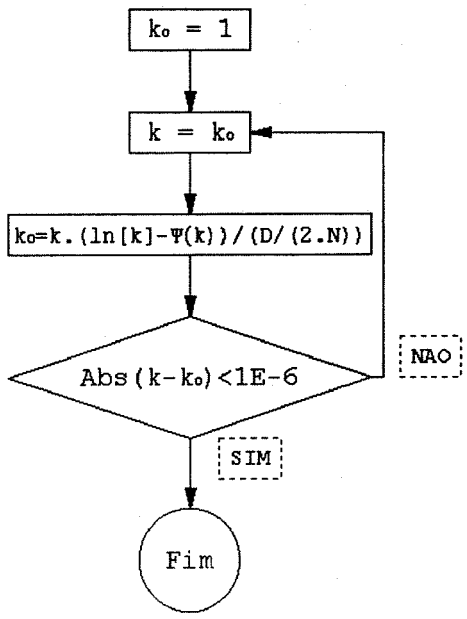


Figura 3.1 - Parâmetro de escala (k) do modelo Gama
D - *deviance* N - N° de observações
$\Psi(.)$ - Função digama

III.1.7 Procedimento de ajuste de máxima verossimilhança para MLGs

O algoritmo de estimação de um MLG, (3.1), é baseado no método de mínimos quadrados ponderados iterativos proposto por GREEN (1984). As estimativas de máxima verossimilhança dos parâmetros $\underline{\beta}$ de (3.1) podem ser obtidos pelo algoritmo escore de Fisher, embora resultado similar possa ser obtido usando-se Newton-Raphson.

Assim, utilizando o algoritmo escore, tem-se que sua k-ésima iteração pode ser assim descrita:

$$\underline{\beta}_{k+1} = \underline{\beta}_k + (\underline{X}^T \underline{W}_k.\phi \underline{X})^{-1} \underline{X}^T \underline{W}_k.\phi \underline{u}_k \qquad (3.24)$$

com

$$\underline{W}_k = [w_i] \quad \therefore \quad w_i^{-1} = \left[g'(\mu_i)\right]^2 . V(\mu_i)/m_i \qquad (3.25)$$

e

$$\underline{u}_k = (\underline{y} - \underline{\mu}_k).\left[g'(\underline{\mu}_k)\right]^T , \qquad (3.26)$$

sendo V(.) a função variância apresentada na tabela 3.2, m_i pesos definidos a priori e $\underline{X}$ a matriz de covariáveis. O parâmetro de escala ou dispersão ϕ é estimado a cada iteração conforme descrito no item III.1.6. A *deviance* D, critério de parada do algoritmo, é calculada a cada iteração k,

interrompendo o processo iterativo quando $|D^k - D^{k-1}|$ for menor do que a precisão desejada.

A equação (3.24) pode ser reescrita da seguinte forma

$$\underline{\beta}_{k+1} = (\underline{X}^T \underline{W}_k.\phi \underline{X})^{-1} \left[(\underline{X}^T \underline{W}_k.\phi \underline{X}) \underline{\beta}_k + \underline{X}^T \underline{W}_k.\phi \underline{u}_k \right] \qquad (3.27)$$

$$= (\underline{X}' \underline{W}_k.\phi \underline{X})^{-1} \underline{X} \underline{W}_k.\phi \underline{z}_k$$

com

$$\underline{z}_k = \underline{X}.\underline{\beta}_k + \underline{u}_k \qquad (3.28)$$

$$= \underline{\eta}_k + \underline{u}_k.$$

O método de inicialização para $\underline{\beta}$ pode ser obtido conforme a sugestão de NELDER, WEDDERBURN (1972). Assim, toma-se como uma primeira aproximação $\underline{\mu} = \underline{y}$, calculando a partir desta o preditor linear $\underline{\eta}$ por $\eta_i = g(\mu_i)$, a matriz diagonal de pesos iterativos $\underline{W}$ conforme (3.25) e fazendo-se $\underline{z} = \underline{\eta}$, sendo $\underline{z}$ conhecido como variável dependente ajustada.

A variável dependente ajustada $\underline{z}$, como definida em (3.28), pode ser utilizada para verificar a adequacidade da função de ligação adotada através de sua plotagem contra o preditor linear $\underline{\eta}$. Se o gráfico resultante for aproximadamente linear não se tem evidência contra a função de ligação adotada, g(.) (tabela 3.1).

III.2 TÉCNICAS DIAGNÓSTICAS

III.2.1 Introdução

O uso de técnicas diagnósticas faz-se necessário para

identificar observações de algum modo estranhas à massa global dos dados, e também para decidir se o modelo sob investigação é satisfatório ou obter informações de como aperfeiçoá-lo, através, por exemplo, da análise dos resíduos.

Fórmulas análogas dos resíduos apresentados no item III.1.4 serão utilizadas posteriormente na definição de diagnósticos, são os resíduos padronizados. Os resíduos Pearson padronizado ($\bar{r}_P$) e *deviance* padronizado ($\bar{r}_D$) são obtidos dividindo-se o lado direito de (3.11) e (3.15) por $\sqrt{\phi.(1-h_i)}$, respectivamente, sendo h_i o i-ésimo elemento na diagonal da matriz de projeção $\underline{H}$ dada por

$$\underline{H} = (\underline{W})^{1/2}.\underline{X}\ (\underline{X}^T\ \underline{W}\ \underline{X})^{-1}\ \underline{X}^T(\underline{W})^{1/2}. \qquad (3.29)$$

Plotagens destes resíduos contra $\underline{\mu}$ e contra cada covariável podem revelar padrões que indicam um modelo mal especificado, ou observações não comuns que tenham uma forte influência no valor de $\underline{\beta}$ e na adequacidade do modelo.

A estrutura geral dos procedimentos diagnósticos para MLGs é similar a aqueles para o Modelo Linear Clássico, embora ocorram algumas complicações. Estas ocorrem devido ao método de estimação ser iterativo para maioria dos MLGs, fazendo-se necessário escolher entre métodos baseados em aproximações de um passo e métodos exatos que se baseiam no ajuste do modelo em estudo N+1 vezes, o que é necessário para a avaliação exata do efeito da simples deleção de cada uma das N observações. Devido ao esforço computacional necessário aos métodos exatos, prefere-se aqui utilizar aproximações de um passo para definir resíduos e medir influência de simples casos em aspectos diferentes do modelo ajustado.

Quando um simples caso é deletado do conjunto de observações, não se pode, para a maioria dos MLGs, obter-se as estimativas de máxima verossimilhança dos N-1 casos como funções explícitas dos resultados do ajuste dos N casos. Uma solução a este problema é proposta por PREGIBON

(1981) que considera a seguinte aproximação

$$\hat{\underline{\beta}}_{(i)} \simeq \hat{\underline{\beta}} - \sqrt{w_i}.(1-h_i)^{-1/2}.\bar{r}^2_{p_i}.(\underline{X}^T.\underline{W}.\underline{X})^{-1}.\underline{x}_i \qquad (3.30)$$

onde $\hat{\underline{\beta}}_{(i)}$ e $\hat{\underline{\beta}}$ são as estimativas de MV de $\underline{\beta}$ para N-1 casos, sendo deletado o caso i, e N casos, respectivamente, $\bar{r}_{p_i}$ é o resíduo Pearson padronizado, w_i o peso iterativo, i-ésimo elemento da matriz diagonal $\underline{W}$ e h_i o i-ésimo elemento da diagonal principal de $\underline{H}$. Usando (3.30) para estimar $\hat{\underline{\beta}}_{(i)}$ e uma aproximação em séries de Taylor para *deviance* o autor deriva aproximações de um passo para mudanças na *deviance* quando um caso simples é deletado:

. A1 - decréscimo em $\sum_{j \neq i} d_j^2$ é aproximadamente $\phi.h_i.\bar{r}^2_{p_i}$

. A2 - acréscimo em d_j^2 é aproximadamente $\phi.h_i.(2-h_i).(1-h_i)^{-1}.\bar{r}^2_{p_i}$

. A3 - acréscimo em $D = \sum_j d_j^2$ é aproximadamente $\phi.h_i.(1-h_i)^{-1}.\bar{r}^2_{p_i}$

Destas três A1 é a que apresenta o menor erro percentual, enquanto que A3 o maior erro percentual.

III.2.2 Detecção de outlier

Os resultados do item III.2.1 podem ser utilizados para detectar se o i-ésimo caso é um *outlier* através de testes de hipóteses. A estatística razão de verossimilhança é a redução G_i em D/ϕ quando o caso i é deletado. Para se calcular o valor exato de G_i seria necessário N+1

ajustes do modelo, o que implica em esforço computacional considerável. Fazendo-se uso de A1 pode-se definir uma aproximação a G_i

$$r^2_{G^i} = d^2_i/\phi + h_i.\bar{r}^2_{P^i} \qquad (3.31)$$
$$= (1-h_i).\bar{r}^2_{D^i} + h_i.\bar{r}^2_{P^i},$$

onde o termo d^2_i/ϕ corresponde a exata contribuição do caso i e $h_i.\bar{r}^2_{P^i}$ a aproximação $A1/\phi$ para a contribuição dos N-1 casos restantes. Segundo WILLIAMS (1987), $r^2_{G^i}$ pode identificar a maioria dos casos de *outliers* e $\max_i r^2_{G^i}$ fornece uma estatística para testar a presença de um simples *outlier*.

A partir de (3.31) pode-se definir uma nova classe de resíduos

$$r_{G^i} = \pm \sqrt{(1-h_i).\bar{r}^2_{D^i} + h_i.\bar{r}^2_{P^i}}\ , \qquad (3.32)$$

com o sinal de (3.32) sendo o mesmo de $(y_i-\mu_i)$. Pela expressão (3.32) fica claro que o valor de r_{G^i} é intermediário aos valores de $\bar{r}_{D^i}$ e $\bar{r}_{P^i}$, sendo mais próximo a $\bar{r}_{D^i}$ já que o valor médio de h_i, p/N, é pequeno.

WILLIAMS (1987) mostra que $r^2_{G^i}$ só segue exatamente χ^2_1 se o modelo é o Normal linear, implicando que para outros MLGs r_{G^i} não são distribuidos assintoticamente conforme N(0,1). Apesar deste resultado, o autor considera que esta classe de resíduos é muito útil, principalmente quando utilizados em plotagens, tais como: r_{G^i} x i, r_{G^i} x h_i, r_{G^i} x $\hat{\eta}_i$ e plotagens Q (*half Normal*) aliadas a envelopes simulados como propostos por ATKINSON (1981). Embora r_{G^i} não siga N(0,1) para a maioria dos MLGs, plotagens dos resultados médios das plotagens Q simuladas apresentam um comportamento linear. O uso destas plotagens diagnósticas é abordado no item III.2.4.

III.2.3 Medidas de influência de simples casos

O i-ésimo elemento na diagonal da matriz de projeção $\underline{H}$ (3.29), h_i, ou *leverage*, como é denominado, fornece uma medida de afastamento da i-ésima observação das n-1 restantes. AITKIN et alli (1989) indicam o valor de 2.p/N como sendo uma importante influência, tornando observações com h_i > 2.p/N suspeitas. Assim, este patamar, 2.p/N, deve ser tomado apenas como uma referência, sugerindo uma análise cuidadosa das observações suspeitas para decidir a rejeição ou não das mesmas. Plotagens de h_i contra i ou μ_i são, em geral, utilizadas para visualizar pontos com alta *leverage*.

Outra medida de influência do caso i na estimativas de máxima verossimilhança $\hat{\underline{\beta}}$ é a estatística de Cook modificada. Esta medida de distância para um MLG é $2.p^{-1}.\left\{L(\hat{\underline{\beta}})-L(\hat{\underline{\beta}}_{(i)})\right\}$ que pode ser estimada pela aproximação

$$C_i = p^{-1}.h_i.(1-h_i)^{-1}.\bar{r}^2_{p^i} \ , \qquad (3.33)$$

utilizando a aproximação A3 do item III.2.1. Esta estimativa subestima o valor exato para casos de grande influência, contudo trabalha bem para casos de pequena influência. Embora isto ocorra, esta estimativa geralmente identifica a maioria dos casos influentes confiavelmente. Assim, uma vez identificados estes casos é possível avaliar os efeitos exatos de deleção de cada um deles.

Deve-se lembrar que um dos objetivos de ajuste de um MLG é o teste de hipóteses que considera um subconjunto dos parâmetros e trata os outros parâmetros como parâmetros incômodos. Se H_o representa o modelo mais simples e H_1 o mais complexo, H_o pode ser testada pela comparação de $\phi^{-1}.\left\{D(H_o) - D(H_1)\right\}$ com a distribuição assintótica χ^2, como em III.1.3. A mudança no valor desta estatística razão de verossimilhança mede a

influência do caso i neste teste, sendo estimada pela aproximação de um passo

$$r^2_{Gi}(H_1) - r^2_{Gi}(H_0). \qquad (3.34)$$

Assim, uma vez obtidos os resíduos r_{Gi} para H_0 e H_1, pode-se facilmente avaliar a expressão (3.34) e identificar os casos que mais influenciam a estatística razão de verossimilhança.

III.2.4 Plotagens diagnósticas

Como para cada quantidade diagnóstica tem-se N valores a serem analisados, surgem dificuldades de interpretação, sendo N o número de observações. Assim plotagens de resíduos e de outras quantidades diagnósticas, como as que foram apresentadas nos itens III.2.2 e III.2.3, podem ser muito úteis na identificação dos vários modos que os dados não concordam com o modelo, seja pela forma da função variância adotada, não linearidade do modelo se for o caso , não-normalidade dos resíduos, *outliers* e até a não adequacidade do modelo hipotetisado.

Entre estas plotagens pode-se citar: a plotagem de r_i contra μ_i que pode ser útil na identificação de observações que não pertençam à distribuição proposta para os dados, e a plotagem indexada[4] de r_i que pode indicar observações que apresentam-se dependentes ou exibindo alguma forma de correlação serial. O uso de plotagens indexadas é muito eficiente na identificação de valores anômalos de quantidades diagnósticas.

Além das já mencionadas, tem-se que destacar o uso de plotagens de quantidades diagnósticas versus pontos percentuais de alguma

---

4 Sendo v_i o i-ésimo valor da quantidade diagnóstica v, chama-se plotagem indexada de v o gráfico de v_i contra i.

distribuição de probabilidade de referência, em geral, a Normal $\Phi(.)$. Estes pontos podem ser aproximadamente determinados por

$$\Phi^{-1}\left[\frac{i-\alpha}{N-2.\alpha+1}\right], \qquad (3.35)$$

sendo em geral adotado $\alpha = 3/8$. Se a distribuição é a Normal, tem-se as plotagens *Normal* e *half-Normal*. Se a quantidade diagnóstica segue a distribuição *Normal* ou *half-Normal*, sua respectiva plotagem será uma linha reta. Estas plotagens para variável diagnóstica v_i são obtidas da seguinte forma:

. *half-Normal* :

(i) ordenação dos valores de $|v_i|$ para obter-se a seqüência ordenada $v_{(i)}$, com os $v_{(i)}$s sendo denominados estatísticas ordenadas;

(ii) obtenção do valor exato esperado da i-ésima estatística ordenada para a distribuição *Normal*, s_i, bastando para isto utilizar, alternativamente à aproximação (3.35), o algoritmo AS 177 de ROYSTON (1982) com a correção de KÖNIGER (1983);

(iii) finalmente, plota-se $v_{(i)}$ versus s_i.

. *Normal* :

Para obter-se a plotagem *Normal*, procede-se da mesma forma que para *half-Normal*, salvo por trabalhar-se com v_i ao invés de $|v_i|$.

Plotagens normais e os testes correspondentes podem não ser eficientes quando aplicados a resíduos, devido estes serem combinações

lineares de variáveis aleatórias, o que faz que eles sejam mais normais do que a distribuição do erro se esta é não Normal. Assim uma plotagem reta não significa necessariamente que a distribuição dos resíduos é Normal.

Além disto é difícil definir se tais plotagens são suficientemente retas, ou se as inevitáveis irregularidades são devido somente a flutuações aleatórias. ATKINSON (1981) sugere um método, descrito a seguir, para interpretar plotagens probabilísticas, baseado em uma forma de teste Monte Carlo no qual um *envelope* é construido por simulação para plotagens *Normal* e *half-Normal*.

Para gerar o envelope por simulação mantemos fixa a matriz de covariância dos resíduos $\underline{I} - \underline{H}$, onde $\underline{H}$ é a matriz de projeção (3.29) e $\underline{I}$ a matriz identidade. A partir da matriz $\underline{I} - \underline{H}$ gera-se para um determinado número de simulações, em geral 19, um conjunto de resíduos que pode ser transformado posteriormente em outras quantidades diagnósticas. Nada impede de usar-se um número maior de simulações, mas 19 simulações já proporciona uma chance de 1 em 20 para que o valor observado do maior v_i caia fora do *envelope*.

Logo, para obter-se *envelopes* da quantidade diagnóstica v_i gera-se 19 conjuntos de v_i, sendo o k-ésimo conjunto gerado da seguinte forma

$$\underline{v}^k = (\underline{I} - \underline{H}).\underline{\xi}^k \ , \qquad (3.36)$$

onde $\underline{\xi}^k \sim N(0,1)$. Uma vez obtido v_i^k ou $|v_i^k|$ conforme a plotagem seja *Normal* ou *half-Normal* respectivamente, ordena-se estes de modo conseguir a seqüência ordenada $v_{(i)}^k$, i = 1, ..., N. Procedendo-se da mesma maneira para as 19 simulações, os limites inferiores e superiores que formam o *envelope* são dados, respectivamente, por

$$v_{(i)}^I = \min_k \{ v_{(i)}^k \} \quad e \quad v_{(i)}^U = \max_k \{ v_{(i)}^k \}. \qquad (3.37)$$

Este método provê uma região onde a plotagem da quantidade diagnóstica $\underline{v}$ deve ser esperada, e não uma região de aceitação ou rejeição de observações. ATKINSON (1981,1982) afirma que para amostras de tamanhos moderados ($\sim$ 60) plotagens *half-Normal* são mais informativas para detectar observações influentes e *outliers*, enquanto que para grandes amostras plotagens *Normal* apresentam-se mais informativas.

Para algumas estatísticas faz-se necessário a mudança de escala antes da obtenção destas plotagens. Mais especificamente, para estatística Cook modificada é necessário multiplicar C_i por N-p e deste valor obter sua raiz quadrada (N - número de observações e p - número de parâmetros).

A plotagem *Normal*, mesmo sem *envelope*, permanece uma plotagem diagnóstica útil e mais simples, podendo ser empregada para detectar observações excessivamente influentes ou *outliers* e saídas sistemáticas na distribuição do erro.

III.3 MODELAGEM DA HETEROGENEIDADE DA VARIÂNCIA EM REGRESSÃO NORMAL

A aplicação do modelo linear clássico supõe, como já mencionado, a homogeneidade da variância. A falha desta suposição pode ser, algumas vezes, retificada por uma transformação Box-Cox da variável resposta como descrito em III.1.5, o que pode provê benefícios adicionais de aproximações à normalidade e à aditividade. No contexto de MLGs estas suposições são relaxadas pela possibilidade de escolha da distribuição e por permitir a aditividade dos efeitos sistemáticos na escala transformada, sendo a escolha da escala (função de ligação) independente da escolha da distribuição. Uma vez escolhidas a distribuição e a função de ligação, a variância fica determinada como uma função da média vezes o parâmetro de escala ou dispersão (item III.1.6).

Embora as abordagens acima possam ser adotadas, freqüentemente é

de interesse permitir explicitamente a heterogeneidade da variância (heterocedasticidade) na análise. A dificuldade é encontrar e ajustar um modelo satisfatório para variância.

AITKIN (1987) sugere um procedimento que fornece estimadores por máxima verossimilhança para os parâmetros dos dois modelos, média e dispersão. Este método possibilita considerar o parâmetro de escala dependente de covariáveis do modelo para média e/ou outras variáveis não incluidas no modelo para média. O autor propõe para o modelo de regressão Normal

$$\begin{aligned} & y_i \sim N(\mu_i,\sigma^2) \;;\; \xi_i \sim N(0,\sigma^2) \\ & \eta_i = \underline{\beta}'.\underline{x}_i \\ & \underline{\eta} = \underline{\mu} \end{aligned} \qquad (3.38)$$

o modelo para heterogeneidade da variância

$$\text{var}(\xi_i) = \sigma_i^2 = \exp(\underline{\gamma}'.\underline{z}_i), \qquad (3.39)$$

sendo $\underline{z}$ a matriz de covariáveis e $\underline{\gamma}$ os parâmetros do modelo da variância. Intrisicamente o que se faz é generalizar mais ainda a classe de MLGs de modo a considerar a relação entre médias e variâncias

$$\sigma_i^2 = \phi.w_i^{-1}.V(\mu) \qquad (3.40)$$

da seguinte forma

$$\sigma_i^2 = \phi_i.w_i^{-1}.V(\mu) \qquad (3.41)$$

e assumir que o parâmetro de dispersão ϕ_i, agora variável, pode ser modelado por

$$\begin{aligned} \phi_i &\sim \mathcal{G}(k, \mu_i^D/k) \\ \eta_i^D &= \underline{\gamma}' \cdot \underline{z}_i \\ \underline{\eta}^D &= \ln(\underline{\mu}^D) \end{aligned} \qquad (3.42)$$

onde $\mathcal{G}(k,\mu_i^D/k)$ é uma variável Gama com parâmetro de forma k e de escala μ_i^D/k, η_i^D e μ_i^D são, respectivamente, o preditor linear e o valor ajustado para o modelo de dispersão. Como para o modelo Normal a função variância, $V(\mu)$, e os pesos iterativos, w_i, assumem o valor unitário, tem-se $\sigma_i^2 = \phi_i$ e que as equações (3.39) e (3.41) são idênticas. O valor positivo de σ_i^2 é garantido pela função de ligação logarítmica do modelo de dispersão. O modelo para dispersão considera os componentes de *deviance* do modelo da média como sua variável dependente.

Tanto o modelo (3.38) e (3.42) são MLGs e devido ao fato que média e dispersão de um MLG serem ortogonais, pode-se estimar $\underline{\beta}$ e $\underline{\gamma}$ essencialmente ao mesmo tempo. O procedimento de AITKIN (1987) consiste em:

(i) igualar $m_i = 1$, $i = 1,\ldots,N$;

(ii) ajustar o modelo (3.38) com parâmetro de escala 1 pelo algoritmo fornecido em III.1.7;

(iii) utilizando como variável dependente do modelo de dispersão os componentes de *deviance* do modelo (3.38) ajustado

$$d_i = (y_i - \mu_i)^2, \qquad (3.43)$$

ajustar o modelo de dispersão (3.42) com parâmetro de escala 2 também pelo

algoritmo fornecido em III.1.7;

(iv) calcular a *deviance*

$$D = \sum_i [d_i/\mu_i^D + \ln(\mu_i^D) + \ln(2.\pi) \qquad (3.44)$$

(v) de (iii) calcular $m_i = 1/\mu_i^D$, $i = 1,\ldots,N$, onde μ_i^D é o valor ajustado para o modelo de dispersão;

(vi) enquanto a *deviance* D não convergir repete-se os passos de (ii) a (v).

A *deviance* $(-2.\text{Ln}(\mathbb{L}(\underline{b};y))$ calculada em (ii) no primeiro passo do procedimento, ou seja, quando $m_i = 1$, serve como um teste global de homogeneidade de variância, onde $\underline{b}$ é a estimativa de máxima verossimilhança dos parâmetros do modelo para média. A homogeneidade ou não da variância também pode ser verificada informalmente através do uso de plotagens de resíduos contra os valores ajustados e contra as covariáveis do modelo. O ajuste conjunto dos dois modelos, como já explicitado, reduz sensivelmente os desvios padrões dos parâmetros do modelo para média.

III.4 MODELOS PARA GERAÇÃO DE VAZÃO

III.4.1 Introdução

A concepção original do modelo Thomas-Fiering, no contexto de MLGs, pode ser formulado da seguinte forma

$$\begin{aligned} Y &= Q_t^i \sim N(\mu_t^i, \sigma^2) \qquad (3.45)\\ \eta_t^i &= \beta_t^o + \beta_t^1 . Q_{t-1}^i \\ \underline{\eta}_t &= \underline{\mu}_t \end{aligned}$$

onde Q_t^i é a i-ésima observação da vazão no mês t; t = 1,...,12. Esta concepção consiste de 12 equações de regressão (componentes sistemáticas) da vazão do mês t sobre a do mês anterior t-1 e de *variáveis aleatórias* ξ_t, provenientes da distribuição Normal com variância apropriada que em MLGs é o parâmetro de escala ϕ do item III.1.6.

O modelo (3.45) é freqüentemente inadequado para modelagem de vazões mensais, o que se torna evidente, como já mencionado, pela freqüência de vazões negativas na série gerada, bem como pela incapacidade do modelo em reproduzir o padrão de dispersão, que aumenta com o aumento da vazão do mês anterior como mostram as figuras 3.2 e 3.3.

É claro que pode-se generalizar o modelo (3.45) das seguintes formas:

(i) pela inclusão de termos de maior retardo na componente sistemática de (3.45);

(ii) pela inclusão de vazões de outros postos na componente sistemática de (3.45);

(iii) pela utilisação de outra distribuição, diferente da Normal, também pertencente à família exponencial, para componente aleatória do modelo;

(iv) e, finalmente, pelo uso de uma função de ligação outra que a identidade.

Estas alternativas serão apresentadas no item a seguir.

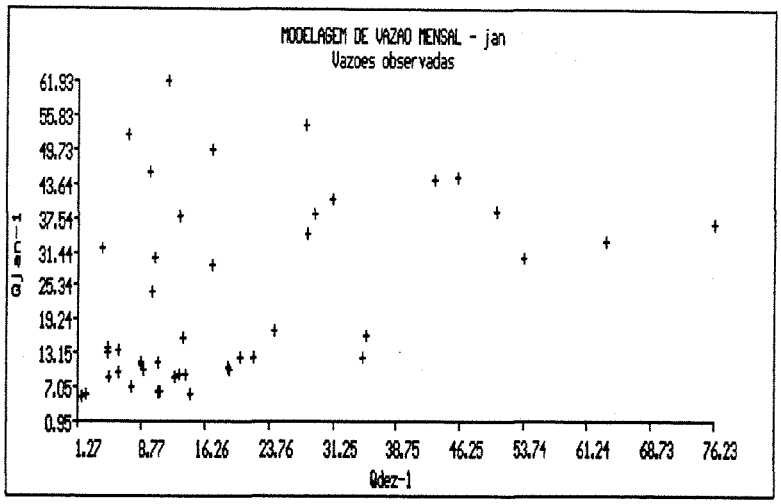


Figura 3.2 - Vazões observadas (janeiro x dezembro)
1a. Estação: 70300000

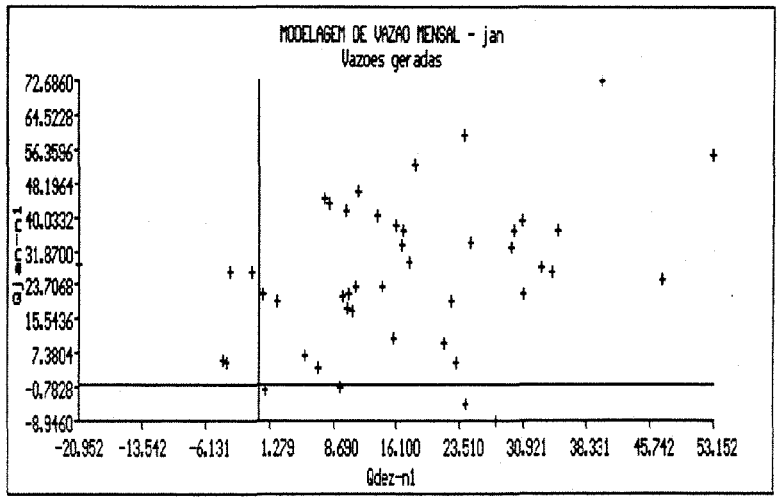


Figura 3.3 - Vazões geradas (janeiro x dezembro)
1a. Estação: 70300000

III.4.2 Modelagem univariada

Possivelmente um modelo apropriado para vazões mensais pode ser o MLG Gama com função de ligação logarítmica ou identidade. Logo, sejam

Q_t^i : i-ésima observação da vazão no mês t e

Q_{t-1}^i : i-ésima observação da vazão no mês t-1,

ambas seguindo a distribuição Gama com função densidade de probabilidade

$$f(Q_t^i,\mu_t^i) = \frac{(k_t/\mu_t^i)^{k_t} \cdot (Q_t^i)^{(k_t-1)} \cdot \exp(-k_t \cdot Q_t^i/\mu_t^i)}{\Gamma(k_t)}, \qquad (3.46)$$

onde k_t é o parâmetro de forma e μ_t^i/k_t o parâmetro de escala, sendo k_t diferente a cada mês. Assim, o modelo Gama com função de ligação logarítmica ou identidade, para cada mês t, é

$$\begin{aligned} Y &= Q_t^i \sim \mathcal{G}(k_t, \mu_t^i/k_t) \qquad (3.47)\\ \eta_t^i &= \beta_t^0 + \beta_t^1 . Q_{t-1}^i \\ \underline{\eta}_t &= \ln(\underline{\mu}_t) \quad \text{ou} \quad \underline{\eta}_t = \underline{\mu}_t, \end{aligned}$$

onde $\mathcal{G}(k_t, \mu_t^i/k_t)$ é uma variável Gama com parâmetro de forma k_t e de escala μ_t^i/k_t.

Uma vez estimados os $\underline{\beta}_t$, com t = 1,...,12, por máxima verossimilhança empregando o algoritmo descrito em III.1.7, pode-se obter séries sintéticas de vazões (Q^*) procedendo-se da seguinte forma:

(i) assumindo que o último valor da série histórica seja Q_{12}^{N}, N-ésima observação do mês de dezembro, e este inicie o processo de geração;

(ii) cálcula-se μ_1 através da equação

$$\mu_1 = h(\beta_1^0 + \beta_1^1 . Q_{12}^N),$$

onde h(.) é a inversa da função de ligação adotada;

(iii) obtem-se Q_1^* utilizando-se o parâmetro de forma k_1 de (3.46), inverso do parâmetro de escala ϕ do MLG em questão e estimado por (3.23), e o de escala μ_1/k_1 de (3.46) para gerar $\mathcal{G}(k_1, \mu_1/k_1)$. Para $k_1 \leq 5$ utiliza-se o algoritmo de FISHMAN (1973) apresentado na figura 3.4, resultando $Q_1^* = \mathcal{G}(k_1, \mu_1/k_1)$, enquanto que, para $k_1 > 5$ emprega-se a transformação Wilson-Hilferty, o que fornece:

$$Q_1^* = k_1 . \left[\frac{1}{3\sqrt{k_1}} . \left(X - \frac{1}{3\sqrt{k_1}} \right) + 1 \right]^3,$$

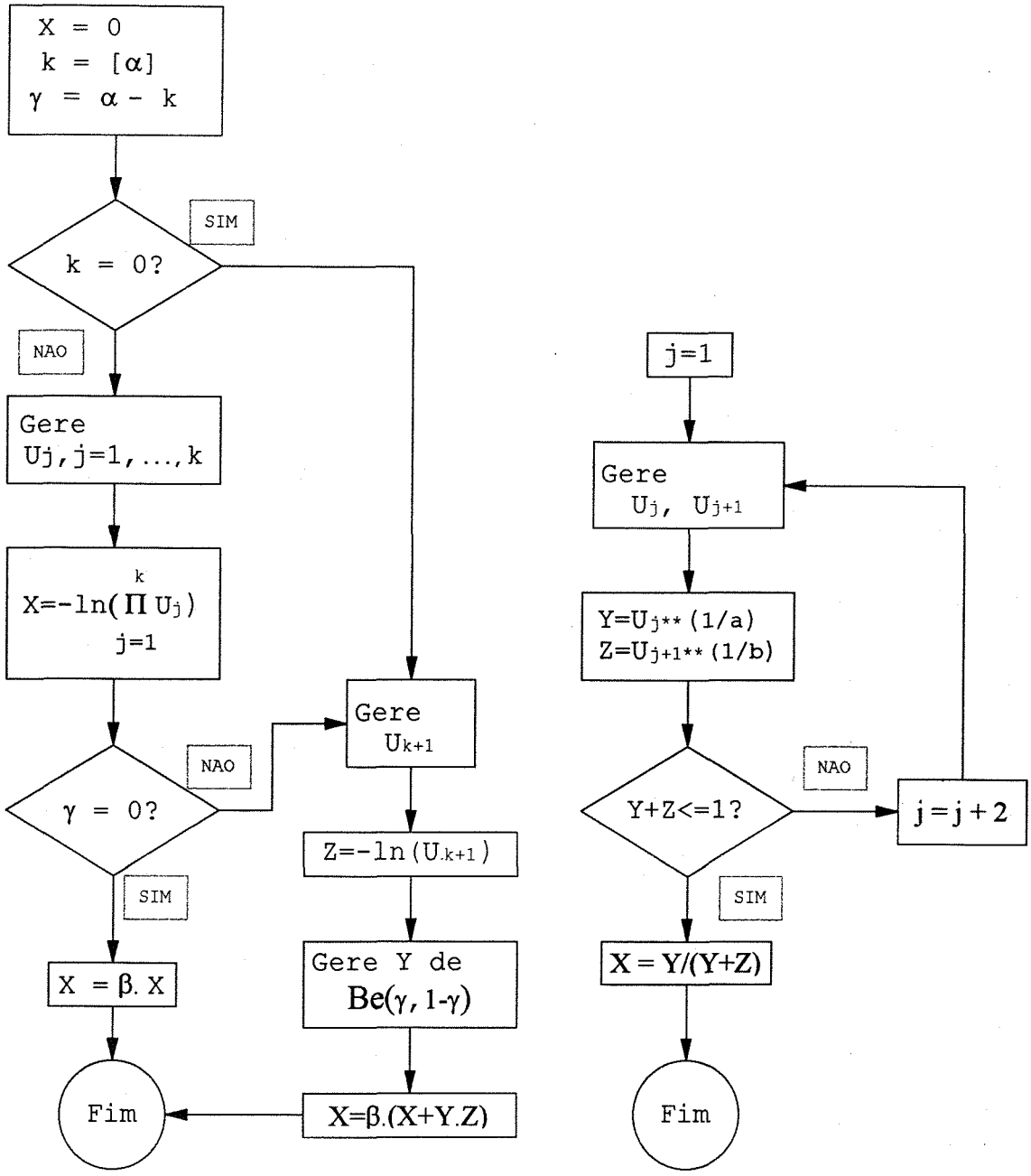
onde $X \sim N(0,1)$;

(iv) repete-se o processo utilizando Q_1^* para obter Q_2^* e assim por diante, até que se tenha obtido o tamanho desejado da série sintética de vazões.

Outros modelos com distribuição do erro relacionada à *Normal* pelas transformações BOX-COXs, a partir de agora identificados como modelos λ-*Normal*, podem ser adequados. Assim,

$(Q_t^i)^{(\lambda_t)}$: i-ésima observação da vazão transformada no mês t,

seguindo a distribuição Normal com função densidade de probabilidade



(a) Geração de $\mathcal{G}(\alpha,\beta)$ (b) Geração de $\mathcal{B}e(a,b)$

Figura 3.4 - Geração de variáveis Gama $\mathcal{G}(\alpha,\beta)$, FISHMAN (1973)

α - parâmetro de forma β - parâmetro de escala

$\mathcal{B}e(a,b)$ - variável Beta com parâmetros a e b.

$$f((Q_t^i)^{(\lambda_t)}, \mu_t^i, \sigma_t^2) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \cdot \exp\left[-\frac{1}{2\sigma_t^2}((Q_t^i)^{(\lambda_t)} - \mu_t^i)^2\right], \tag{3.48}$$

sendo o parâmetro da transformação, λ_t, estimado conforme indicado no item III.1.5. Assim, o modelo Normal com função de ligação identidade, para cada mês t, é

$$\begin{aligned} Y &= (Q_t^i)^{(\lambda_t)} \sim N(\mu_t^i, \sigma^2) \\ \eta_t^i &= \beta_t^0 + \beta_t^1 . Q_{t-1}^i \\ \underline{\eta}_t &= \underline{\mu}_t . \end{aligned} \tag{3.49}$$

Com uma pequena modificação, pode-se da mesma forma que antes, uma vez estimados os $\underline{\beta}_t$ (t = 1,...,12) por máxima verossimilhança empregando o algoritmo descrito em III.1.7, obter séries sintéticas de vazões (Q^*) procedendo-se da seguinte forma:

(i) assumindo que o último valor da série histórica seja Q_{12}^N, N-ésima observação do mês de dezembro, e este inicie o processo de geração;

(ii) cálcula-se μ_1 através da equação

$$\mu_1 = \beta_1^0 + \beta_1^1 . Q_{12}^N ,$$

(iii) passando μ_1 para escala original tem-se a variação na mediana de Q_1, enquanto que a estimativa para média e variância de Q_1, conforme valor de λ_1, seria

$$\begin{aligned} E[Q_1] &= \mu_1 \ , \ \lambda_1 = 1 \\ var[Q_1] &= \sigma_1^2 \end{aligned} \tag{3.50}$$

$$E[Q_1] = \exp(\mu_1 + \frac{1}{2}.\sigma_1^2) \ , \ \lambda_1=0 \qquad (3.51)$$

$$var[Q_1] = \{\exp(\sigma_1^2)-1\}.\exp(2.\mu_1+\sigma_1^2)$$

$$E[Q_1] \approx (\mu_1)^{1/\lambda_1}.\{1+\sigma_1^2.(1-\lambda_1)/(2.\lambda_1^2.(\mu_1)^2)\}, \ \lambda_1\neq0 \text{ e } \lambda_1\neq1 \qquad (3.52)$$

$$var[Q_1] \approx (\mu_1)^{2/\lambda_1}.\sigma_1^2/(\lambda_1^2.(\mu_1)^2)$$

como já mencionado em III.1.5. Gera-se Q_1^*, conforme o valor de λ_1, a partir de

$$Q_1^* = \mu_1 + \xi_1.\sigma_1 \ , \ \lambda_1 = 1 \qquad (3.53)$$

$$Q_1^* = \exp(\mu_1 + \xi_1.\sigma_1) \ , \ \lambda_1 = 0 \qquad (3.54)$$

$$Q_1^* = (\mu_1 + \xi_1.\sigma_1)^{1/\lambda_1} \ , \ \lambda_1 \neq 0 \text{ e } \lambda_1 \neq 1. \qquad (3.55)$$

com média e variância dadas, respectivamente, por (3.50), (3.51) e (3.52) e $\xi_1 \sim N(0,1)$ sendo gerado conforme o método Polar, *Marsaglia* citado por ATKINSON (1976). Este método é um versão aperfeiçoada do método Box-Müller pela substituição das funções trignométricas, consistindo em

(a) geração de V_1, V_2 uniformes entre $(-1,1)$[5]:

$$V_1 = -1 + 2.U_1 \quad e \quad V_2 = -1 + 2.U_2,$$

com U_1, U_2 uniformes entre (0,1).

(b) cálculo de $W = V_1^2 + V_2^2$ e, se $W > 1$, então voltar ao passo (a). Então W é distribuida uniformemente entre (0,1) independentemente de V_1 e V_2;

(c) A partir dos valores de V_1, V_2 e W obtem-se o par

$$X_1 = V_1 . \left[\frac{-2.\ln(W)}{W}\right]^{1/2} \quad e \quad X_2 = V_2 . \left[\frac{-2.\ln(W)}{W}\right]^{1/2}. \tag{3.56}$$

com X_1 e X_2 independentemente distribuidos segundo N(0,1). Deste resultado decorre que $\xi_1 = X_1$, sendo X_2 empregado para a próxima iteração do processo de geração, ou melhor, $\xi_2 = X_2$;

(iv) repete-se o processo utilizando Q_1^* para obter Q_2^* e assim por diante, até que se tenha obtido o tamanho desejado da série sintética de vazões.

Utilizando a metodologia indicada em III.1.5, é possível escolher entre a transformação Box-Cox da variável resposta ou o uso da função de

---

[5] Suponha que X siga distribuição uniforme no intervalo (a,b)

$$f_x(x) = \begin{cases} \dfrac{1}{(b-a)}, & a \le x \le b \\ 0, & \text{para outros valores de } x \end{cases}$$

gerando $U \sim \mathcal{U}(0,1)$, pode-se obter $X \sim (a,b)$ da seguinte forma

$$U = \frac{X - a}{b - a} \quad \Rightarrow \quad X = a + (b-a).U.$$

ligação potência ambos com o mesmo parâmetro λ_t.

Outro modelo possível para a modelagem de vazões mensais seria o modelo Lognormal, ou seja, supõe-se que $(\ln(Q_t), \ln(Q_{t-1}))$ seguem uma distribuição Normal bivariada. Pode-se comparar o modelo Lognormal com o modelo log-Gama definido por (3.47) e o modelo Lognormal definido por (3.49) utilizando um procedimento Monte Carlo gráfico como sugere ATKINSON (1982, 1987). Este procedimento consiste em:

(i) ajustar os modelos Lognormal e log-Gama aos dados, sendo a *deviance* do primeiro designada por L_o e do último por G_o. Uma vez ajustados os dois modelos aos dados, gerar 100 amostras, como sugerido por ATKINSON (1982), de tamanho N, tamanho da amostra observada, com cada um dos modelos;

(ii) para cada amostra gerada pelo modelo Lognormal y_L em (i) ajustar o modelo Lognormal obtendo-se os valores ajustados μ_L, calculando com y_L e μ_L a *deviance* para distribuição Normal e Gama, conforme as fórmulas apresentadas na tabela 3.3, identificadas respectivamente por L e G. No caso da *deviance* para o modelo Gama deve-se transformar, antes do cálculo da mesma, y_L e μ_L para a escala original de medida;

(iii) plotar L contra G, indicando no gráfico o ponto (L_o, G_o) para a seqüência histórica;

(iv) repetir o processo, utilizando no passo (ii) a amostra gerada pelo modelo log-Gama em vez da gerada pelo modelo Lognormal, fazendo interpretação análoga.

Quando, nas duas plotagens L contra G, o ponto (L_o, G_o) se encontra na região definida pelos pontos (L,N) das 100 amostras geradas os dois modelos não diferem significativamente. Caso isto não ocorra, se na plotagem das 100 amostras geradas com o modelo Lognormal G_o é grande quando comparado com L_o, fazendo que o *ponto observado* (L_o, G_o) caia distante da região definida pelos pontos (L,G) das 100 amostras, o modelo Gama é o mais apropriado. De maneira análoga procede-se a análise da plotagem das amostras geradas com o modelo Gama. Este teste para famílias separadas pode

ser utilizado para testar outras combinações de modelos.

Os modelos, até agora apresentados, consideram a vazão do mês t dependente apenas da vazão do mês t-1 como no modelo original de Thomas-Fiering. Conforme apresentado em III.4.1, pode-se incluir termos de maior retardo escolhidos de acordo com o correlograma de vazões mensais e vazões de outros postos escolhidas conforme correlogramas cruzados de vazões mensais e plotagens *scatter* como a da figura 3.5.

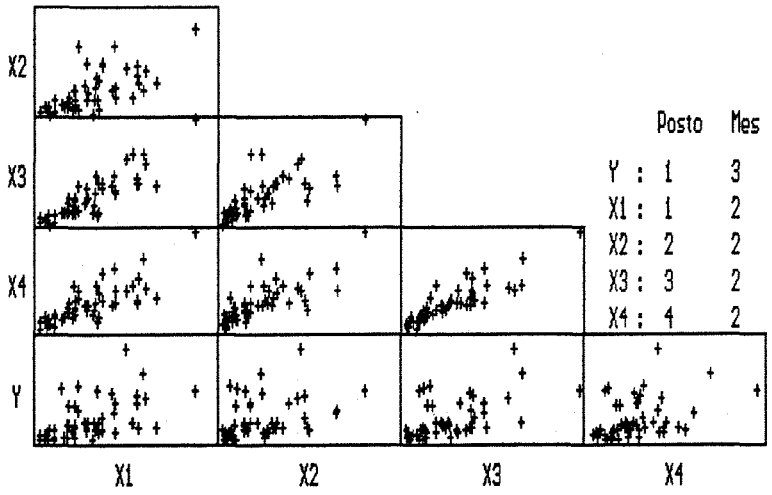


Figura 3.5 - Plotagem *Scatter* de vazões mensais,
Postos: (1) 70300000 (3) 70700000
(2) 70500000 (4) 71300000

III.4.3 Modelagem Multivariada

Às vezes torna-se necessário a geração de seqüências simultâneas de vazões mensais de vários postos, visando não somente preservar a estrutura correlacional para vazão mensal de cada posto, mas também a estrutura correlacional cruzada entre vazões mensais de cada par de postos.

Considere, assim, o caso de geração de vazões mensais em dois postos, cujo resultado pode facilmente ser extendido para mais de dois postos. Logo, sendo

$Q^i_{t(1)}$: i-ésima vazão no mês t do posto 1

$Q^i_{t(2)}$: i-ésima vazão no mês t do posto 2

e

$Q^i_{t-1(1)}$: i-ésima vazão no mês t do posto 1

$Q^i_{t-1(2)}$: i-ésima vazão no mês t do posto 2

considere a distribuição condicional, das vazões $Q^i_{t(1)}$, $Q^i_{t(2)}$, dadas as vazões $Q^i_{t-1(1)}$, $Q^i_{t-1(2)}$

$$f(Q^i_{t(1)},Q^i_{t(2)}|Q^i_{t-1(1)},Q^i_{t-1(2)}) = f(Q^i_{t(1)}|Q^i_{t(2)},Q^i_{t-1(1)},Q^i_{t-1(2)}).f(Q^i_{t(2)}|Q^i_{t-1(1)},Q^i_{t-1(2)}) \quad (3.57)$$

com f(.) sendo a função densidade de probabilidade das vazões mensais.

A equação (3.57) sugere que, para gerar as vazões $Q^i_{t(1)}$, $Q^i_{t(2)}$, dadas as vazões $Q^i_{t-1(1)}$, $Q^i_{t-1(2)}$, deve-se:

(i) gerar $Q^i_{t(2)}$ $\left(f(Q^i_{t(2)}|Q^i_{t-1(1)},Q^i_{t-1(2)})\right)$ com média $\mu^i_{t(2)}$, utilizando as vazões $Q^i_{t-1(1)}$, $Q^i_{t-1(2)}$ do mês anterior:

$$E[Q^i_{t(2)}] = \mu^i_{t(2)} = h\left(\beta^0_{t(2)} + \beta^1_{t(2)}.Q^i_{t-1(1)} + \beta^2_{t(2)}.Q^i_{t-1(2)}\right), \quad (3.58)$$

onde h(.) é a inversa da função de ligação adotada;

(ii) gerar $Q^i_{t(1)}$ $\left(f(Q^i_{t(1)}|Q^i_{t(2)},Q^i_{t-1(1)},Q^i_{t-1(2)})\right)$ com média $\mu^i_{t(1)}$, utilizando as vazões $Q^i_{t-1(1)}$, $Q^i_{t-1(2)}$ do mês anterior, postos 1 e 2 respectivamente, e a vazão $Q^i_{t(2)}$ do mês t, posto 2 gerada em (i):

$$E[Q^{i}_{t(1)}] = \mu^{i}_{t(1)} \qquad (3.59)$$
$$= h\left(\beta^{0}_{t(1)} + \beta^{1}_{t(1)}.Q^{i}_{t(2)} + \beta^{2}_{t(1)}.Q^{i}_{t-1(1)} + \beta^{3}_{t(1)}.Q^{i}_{t-1(2)}\right)$$

O método acima também pode ser generalizado mais ainda, pela inclusão de vazões de maior retardo do posto e de outros postos que venham a fazer parte do modelo multivariado.

III.4.4 Modelagem de vazões mensais em rios intermitentes

Um dos maiores problemas em aplicar modelos de geração a rios intermitentes é a modelagem de vazões nulas. Para estes rios o modelo Thomas-Fiering para geração de vazões mensais não gera o número adequado de vazões nulas.

Os modelos descritos pelas equações (3.47) e (3.49), ou ainda uma variação de (3.49) que utiliza a função de ligação potência em vez da transformação da variável resposta, já mencionada no item III.4.2, assumem a existência de uma dependência entre meses consecutivos, não levando em consideração a ocorrência de vazões nulas. Como é pouco provável que exista, para meses consecutivos em rios intermitentes, dependência entre vazão nula e não nula, torna-se necessário adaptar estes modelos para geração de seqüências sintéticas de vazão em rios que apresentem vazões nulas durante a estação seca.

O procedimento aqui empregado é uma modificação do método sugerido por Clarke (1973) e adaptado por Filho (1978). O fluxograma do procedimento utilizado é apresentado na figura 3.6.

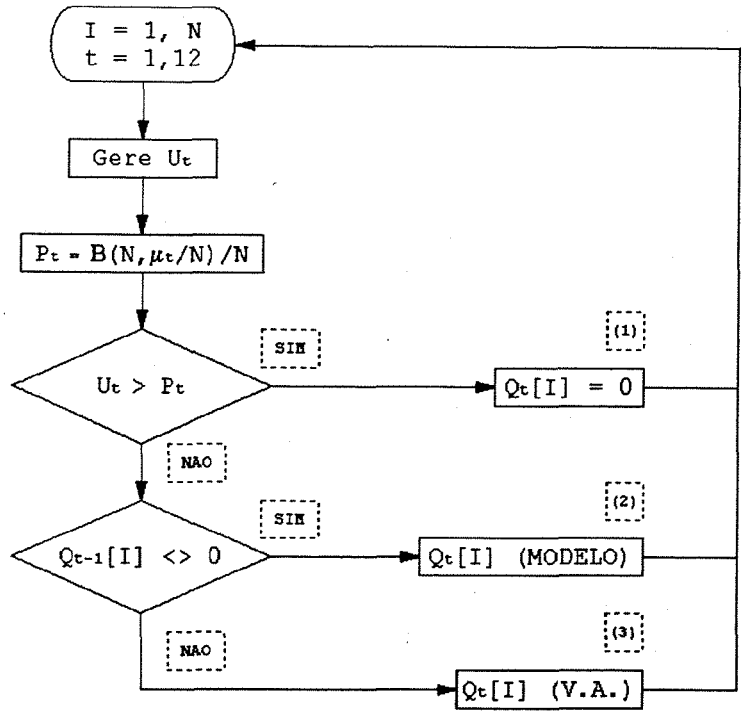


Figura 3.6 - Procedimento de CLARKE (1973) modificado

U_t - v.a. $\sim \mathcal{U}(0,1)$ P_t - $P[J(t) = 1]$

$B(N,\mu_t)$ - procedimento de geração de v.a. Binomiais (Figura 3.7)

A modificação, aqui proposta, consiste em utilizar MLGs para descrever adequadamente a ocorrência de vazão mensal visando considerar a variabilidade da freqüência de vazões nulas no mês e não adotar as freqüências observadas das mesmas. Seja, então,

$$J(t) = \begin{cases} 0 & \text{se no mês t ocorre vazão nula} \\ 1 & \text{se no mês t ocorre vazão não nula} \end{cases}$$

com t = 1,...,12, uma cadeia de Markov de dois estados. Supondo que uma cadeia de Markov de ordem 1 descreve adequadamente os dados, implicando em

$$P[J(t)=1|J(t-1),J(t-2),J(t-3),...] = P[J(t)=1|J(t-1)], \qquad (3.60)$$

com t = 1,...,12. Sendo as proporções de sucessos $p_{i1}(t)$, t = 1,...,12, supostas independentes nos conjuntos 1,...,12, segue que o modelo de estudo destas proporções, no contexto de MLGs, tem variável resposta $N_{i1}(t)$ Binomial, ou de outra forma, $N_{i1} \sim \mathcal{B}(N_i(t),\mu_{i1}(t))$, ligação *logit* (tabela 3.1) e estrutura linear definida por

$$\eta_{i1}(t) = a_o + \sum_{k=1}^{m} [a_k.\text{sen}(2\pi.k.t/12) + b_k.\cos(2\pi.k.t/12)] \qquad (3.61)$$

e valor estimado para $p_{i1}(t)$ dado por

$$\mu_{i1}(t) = \frac{\exp(\eta_{i1}(t))}{[1 + \exp(\eta_{i1}(t))]} \qquad (3.62)$$

onde

m : Número de harmônicos definido por análise de deviance;

$N_{ij}(t)$: Número de dias com $J(t) = j$, $J(t-1) = i$, $t = 1,\ldots,12$;

$N_i(t)$: Número de dias com $J(t-1) = i$, $t = 1,\ldots,12$, sendo $N_i(t) = N_{io}(t) + N_{i1}(t)$;

$p_{ij}(t)$: Probabilidades observadas de transição do estado i para o estado j, sendo calculadas por $p_{ij}(t) = \dfrac{N_{ij}(t)}{N_i(t)}$;

$\mu_{ij}(t)$: Probabilidades estimadas de transição do estado i para o estado j.

No fluxograma da figura 3.6 considera-se que a probabilidade de que no mês t a vazão não seja nula pode ser adequadamente descrita pela modelagem da proporção de vazões não nulas no mês t, não sendo necessário considerar os estados dos meses anteriores. Logo, tem-se

$$P_t = \mathcal{B}(N,\mu_t)/N, \tag{3.63}$$

onde μ_t é o valor estimado para as proporções de vazões não nulas no mês t e $\mathcal{B}(N,\mu_t)$ gerado conforme fluxograma da figura 3.7. Como N é moderado, o método da figura 3.7 é baseado na geração de variáveis Bernoulli e indicado por FISHMAN (1973).

Na verdade é imperativo explorar a natureza do processo markoviano para determinar a ordem necessária da cadeia markoviana para descrever adequadamente a ocorrência de vazão.

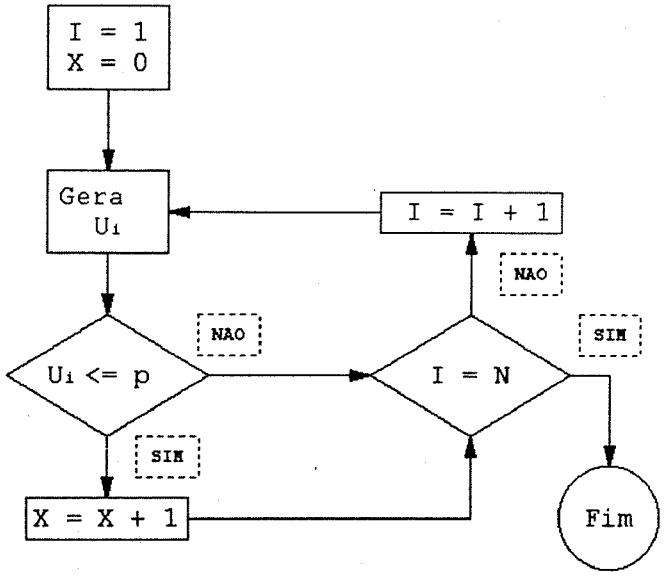


Figura 3.7 - Geração de v.a. Binomiais $\mathcal{B}(N,p)$

N - número de experimentos (anos)

p - proporções de sucessos

Caso esta exploração demonstre que a ordem necessária seja 1, o cálculo de P_t no referido fluxograma deve ser modificado para

$$P_t = \mathcal{B}(N_o(t),\mu_{o1}(t))/N_o(t) + \mathcal{B}(N_1(t),\mu_{11}(t))/N_1(t), \qquad (3.64)$$

devendo-se proceder de forma análoga para ordens superiores a 1.

No algoritmo notam-se três pontos de geração de vazão, o ponto (1) gera vazões nulas quando $U_t > P_t$, o (2) gera vazões conforme o modelo adotado (equações 3.47, 3.49 e uma modificação de 3.49) caso $U_t <= P_t$ e a vazão no mês anterior for diferente de zero, e, finalmente, o (3) gera vazões seguindo uma determinada distribuição de probabilidade (equações 3.46 ou 3.48) caso $U_t <= P_t$ e $Q_{t-1}[I] = 0$.

CAPÍTULO IV RESULTADOS

IV.1 - CASOS DE ESTUDO E PREENCHIMENTO DE FALHAS

Os modelos estudados no item III.4, Lognormal, Gama e Normal-Gama, foram aplicados a rios da região Nordeste e Sul do Brasil, sendo os rios da primeira intermitentes e fortemente marcados pela ocorrência de vazões nulas. As estações fluviométricas utilizadas, sua localização, bem como outras características das mesmas, estão apresentadas na tabela 4.1.

Tabela 4.1 - Estações fluviométricas

CÓDIGO	NOME	RIO - ESTADO	LAT.	LONG.	ÁREA DE DRENAGEM (km²)
35210000	FAZ. CAJAZEIRAS	ACARAÚ-CE	06°34'	39°31'	1550.0
36125000	SÍTIO POÇO DANTAS	BASTIÕES-CE	04°26'	40°34'	3700.0
70300000	FAZ. MINEIRA	LAVA TUDO-SC	28°07'	50°01'	1147.0
70500000	COXILHA RICA	PELOTINHAS-SC	28°09'	50°26'	638.7
70700000	PASSO SOCORRO	PELOTAS-RS	28°22'	50°48'	8365.0
71300000	RIO BONITO	CANOAS-SC	27°42'	49°50'	1971.8
71498000	PASSO MAROMBAS	MAROMBAS-SC	27°20'	50°44'	3722.0
72680000	PASSO COLOMBELLI	LIGEIRO-RS	27°34'	51°51'	3627.0
72980000	RIO URUGUAI	DO PEIXE-SC	27°27'	51°52'	5239.0

* FONTE: DNAEE (1987)

As duas bacias escolhidas na região Nordeste, ambas localizadas no Ceará figura 4.1, são as do rio Acaraú e do rio Bastiões, tendo um regime pluviométrico altamente marcado pela heterogeneidade da repartição temporal das chuvas, com sua concentração no primeiro semestre. A intermitência é uma forte característica tanto do rio Acaraú no trecho à

montante do açude Araras, estação Fazenda Cajazeiras, como do rio Bastiões, estação Sítio Poço das Antas. Em termos de distribuição espacial pluviométrica, a bacia do rio Acaraú é dominada pelas isoietas de 800 e 900mm e a bacia do Rio Bastiões pelas isoietas de 700 e 800 mm, estando a de 700 mm mais próxima. Na região Sul a área de estudo, fronteira entre os estados de Santa Catarina e Rio Grande do Sul, corresponde a uma parte da bacia do Rio Uruguai, compreendida entre 27'20 e 28'22 de latitude sul e entre 49'50 e 51'52 de longitude oeste, conforme figura 4.2. O clima na bacia do Rio Uruguai tem características de zona temperada, tendo como principais fatores genéticos: o anticiclone móvel polar da América do Sul e o anticilone do Atlântico Sul; e como fator estático de maior influência na bacia, sua orografia. As precipitações médias anuais são superiores a 1400mm.

As falhas da região Sul haviam sido preenchidas empregando-se regressão múltipla Normal. Apesar disto decidiu-se adotar o procedimento *MULTMISS* do pacote computacional GENSTAT 5, versão 1.3. Este procedimento, utilizando uma técnica de regressão iterativa, provê estimativas de unidades pertencentes a um conjunto multivariado de dados, exigindo apenas que as falhas sejam aleatórias e, como já mencionado anteriormente, é razoável considerar que o mesmo não faz qualquer suposição sobre a distribuição da amostra. Na linguagem do programa GENSTAT 5, a sintaxe do comando deste procedimento é muito simples e consiste em

```
> MULTMISS A
```

onde A é a matriz de dados com suas falhas indicadas por *.

A condição de aleatoriedade das falhas foi verificada através de procedimento Monte Carlo. Para isto foram geradas 20 amostras aleatórias (N=45,p=6) com vetor de médias $\underline{\mu}$ e matriz de covariância $\underline{\Sigma}$ estimados a partir dos dados. Uma vez obtidas estas, foram eliminados 5% de seus elementos de duas maneiras, uma aleatória e outra não. Depois de preenchidos os valores eliminados com o procedimento *MULTMISS*, foi ajustado um modelo para cada matriz, conseguindo-se assim, uma amostra de tamanho 20

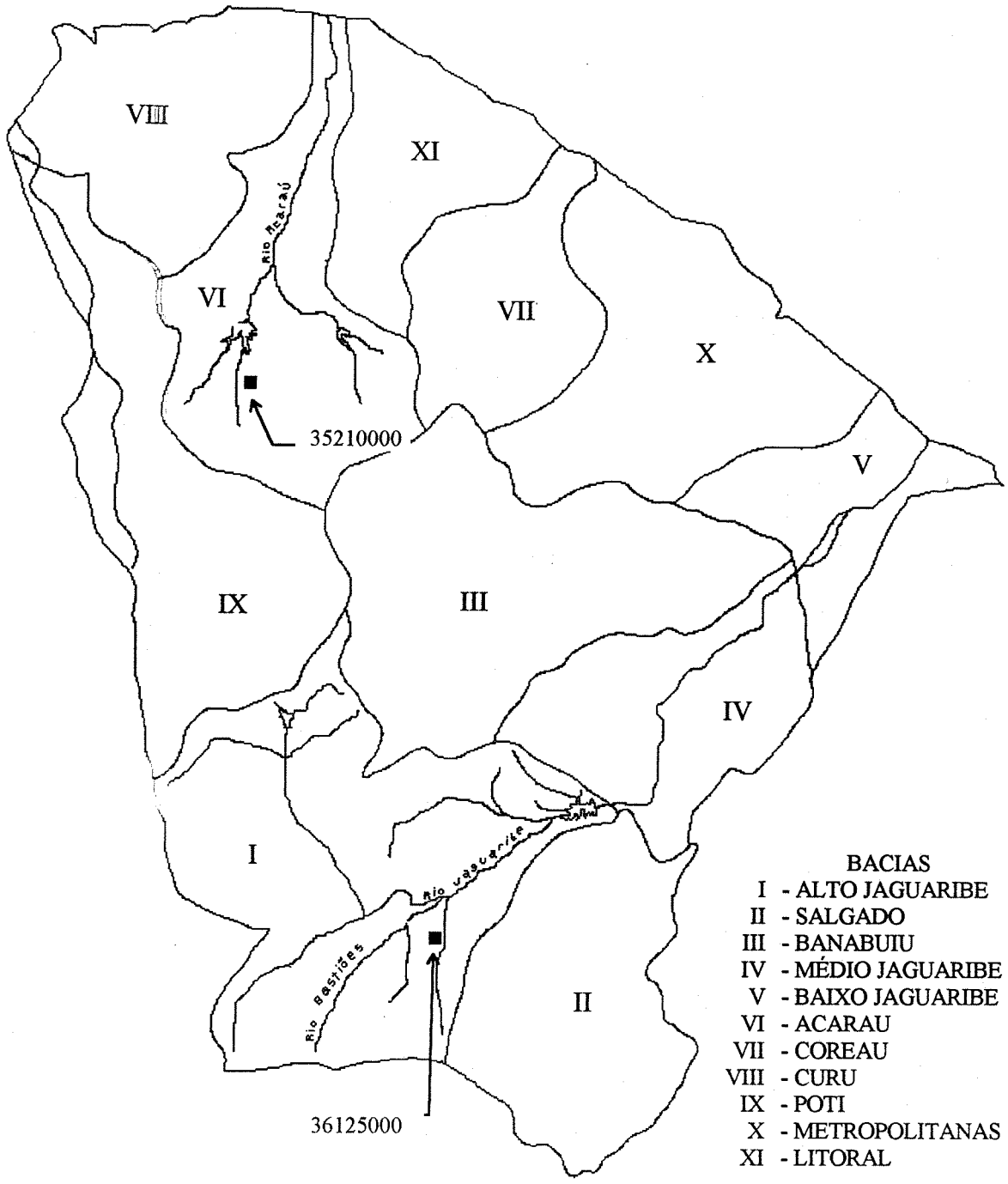


Figura 4.1 - Localização das Estações fluviométricas da Região Nordeste. Relação das estações utilizadas: 35210000 (Bacia do Rio Acaraú) e 36125000 (Bacia do Alto Jaguaribe).

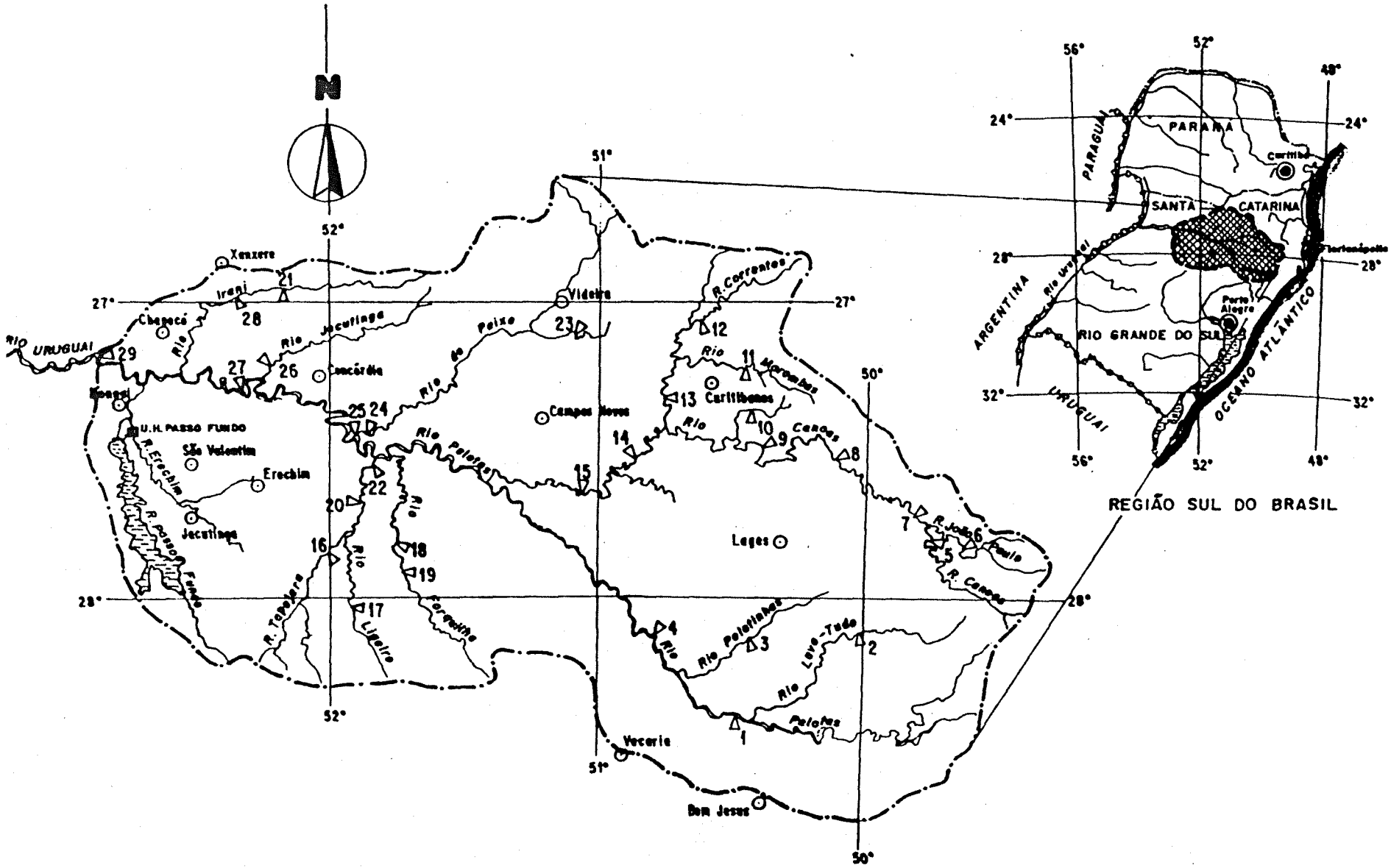


Figura 4.2 - Localização das estações fluviométricas da Região Sul. Relação das estações utilizadas: 2 - 70300000, 3 - 70500000, 4 - 70700000, 7 - 71300000, 13 - 71498000, 22 - 72680000 e 24 - 72980000.

para cada parâmetro do modelo ajustado e modo de escolha das falhas, aleatório ou não. A análise comparativa das duas distribuições de cada parâmetro do modelo mostra que não se tem evidência significativa contra o uso do procedimento *MULTMISS* quando as falhas não são aleatórias.

IV.2 - TÉCNICAS DIAGNÓSTICAS

O uso das técnicas diagnósticas apresentadas no item III.2 permitiram de maneira muito prática a identificação de observações excessivamente influentes e *outliers*, em especial, com o uso de plotagens diagnósticas.

O uso do resíduo r_G nas estatísticas $\max_i r^2_{G^i}$ e $r^2_{G^i}(H_1) - r^2_{G^i}(H_o)$, ambas definidas em III.2.3, fornece resultados razoáveis quando a estimativa do parâmetro de escala ϕ do MLG não varia consideravelmente após a deleção do caso i, na primeira estatística, e para os diferentes modelos (H_o e H_1), na segunda.

A deleção de observações quando o modelo é λ-*normal*, equação 3.49, pode afetar a estimativa do parâmetro da transformação Box-Cox. Assim, após a deleção de um conjunto de observações faz-se necessário estimar novamente o referido parâmetro, verificando se o mesmo muda significativamente. Uma vez estimado, procura-se identificar novamente a necessidade ou não de deletar novas observações. A figura 4.3 mostra dois casos em que a estimativa deste parâmetro mudou significativamente, para estação 70300000, meses de julho e outubro. Em (1) a estimativa de máxima verossimilhança $\hat{\lambda} = -0.02$ com intervalo de confiança 95% (-0.30;0.22) passa, após a deleção, a $\hat{\lambda} = 0.35$ com intervalo de confiança 95% (-0.06;0.78). Já em (2), tem-se antes da deleção $\hat{\lambda} = 0.02$ com intervalo de confiança 95% (-0.28;0.33) passando, após a deleção, a $\hat{\lambda} = 0.23$ com intervalo de confiança 95% (-0.17;0.66).

As figuras 4.4 e 4.5 apresentam algumas plotagens diagnósticas antes e após a deleção de um simples caso, mês de fevereiro do modelo bivariado Lognormal. Neste exemplo as plotagens dos resíduos não mostram qualquer característica

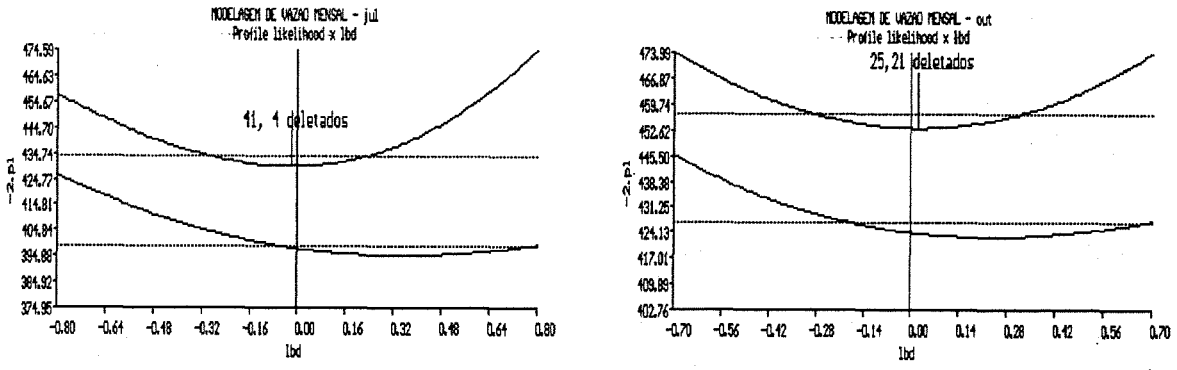


Figura 4.3 - Efeito das observações deletadas na estimativa de máxima verossimilhança do parâmetro λ da transformação Box-Cox. (1) Estação: 70300000. Mês de julho: casos 4, 41 deletados; (2) Estação: 70300000. Mês de outubro: casos 21, 25 deletados.

incomum, sem tendências ou padrões óbvios. Na figura 4.4, a plotagem indexada de C e de h sugerem que o caso 2 pode ter uma influência suficientemente grande para induzir alguma anomalia, uma vez que $\max_i C_i = C_2 = 0.72623$, enquanto os restantes não ultrapassam o valor de 0.18. Em (3) e (4) da mesma figura, o caso também é identificado como, de algum modo, estranho à massa restante dos dados, sendo para (3) esperado que pontos de alta *leverage* tenham um pequeno resíduo. A plotagem *half-normal* de r_G (5) não mostra nenhum resíduo excessivamente grande, o que provavelmente indicaria uma saida do modelo, enquanto que a plotagem *half-normal* de C mostra que o o valor de C_2 é excessivamente influente. Esta estatística tem como vantagem chamar atenção ao pontos de alta *leverage* onde variável resposta e explanatórias não concordam, não sendo possível dizer aonde reside o erro, na primeira ou segunda. Uma vez deletado o caso 2, não se identifica mais observações excessivamente influentes ou *outliers* como mostra a figura 4.5.

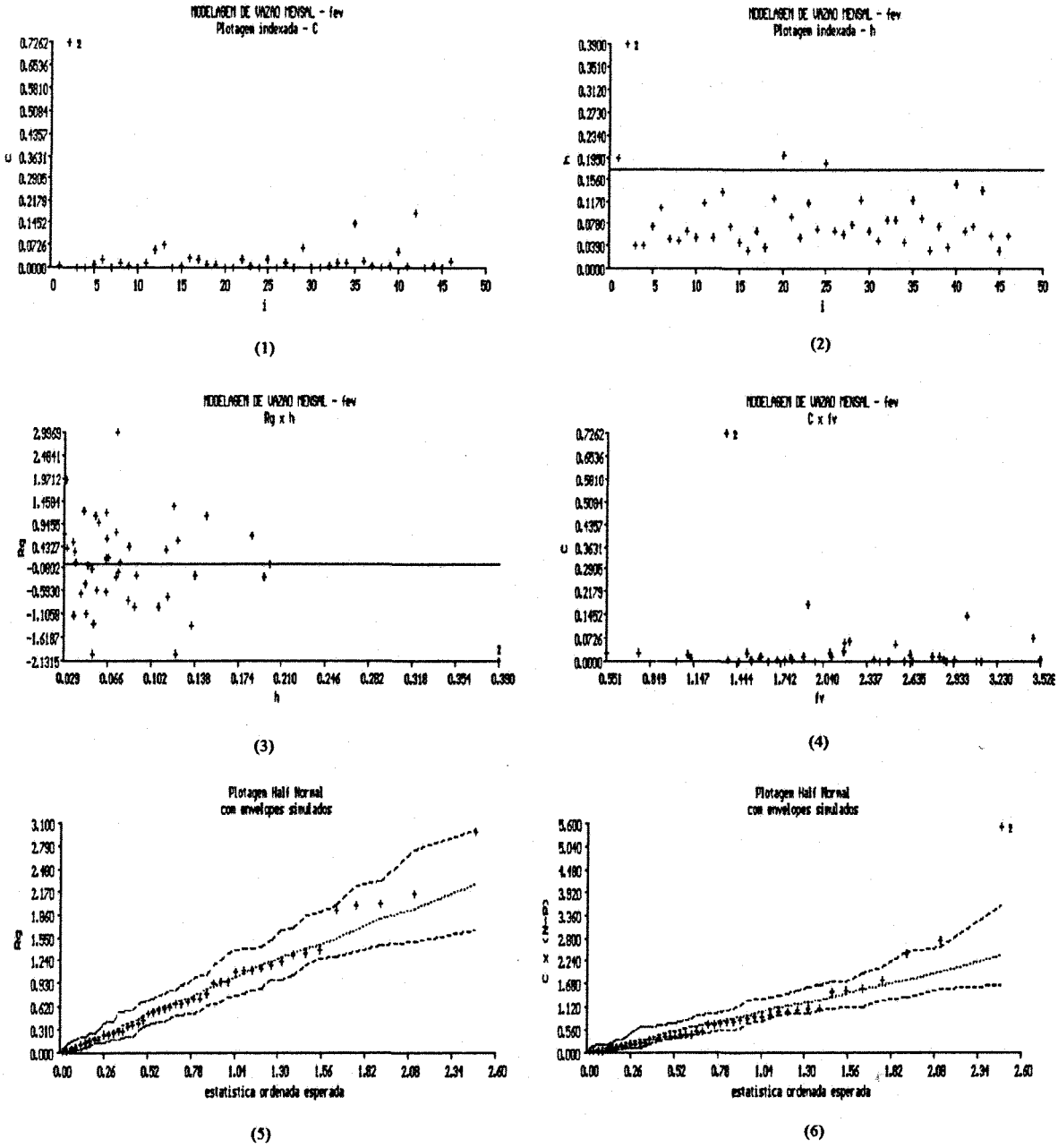


Figura 4.4 - Plotagens Diagnósticas. (1) C x i; (2) h x i; (3) Rg x h; (4) C x fv; (5) *half-normal* de R_G; (6) *half-normal* de C. Caso 2 não deletado. Modelo Lognormal bivariado. Estações: 70300000-70500000.

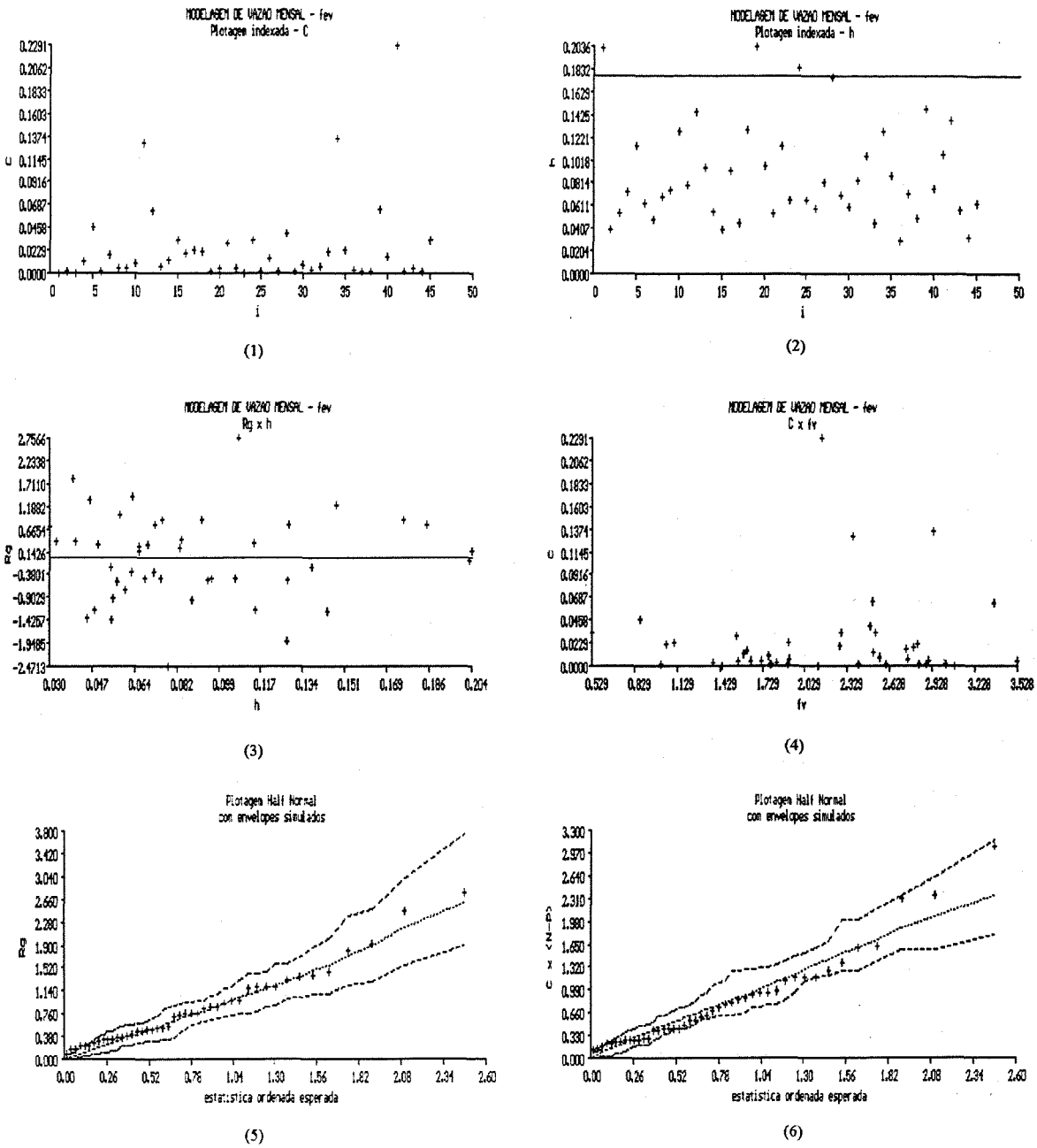


Figura 4.5 - Plotagens Diagnósticas. (1) C x i; (2) h x i; (3) Rg x h; (4) C x fv; (5) *half-normal* de R_G; (6) *half-normal* de C. Caso 2 deletado. Modelo Lognormal bivariado. Estações: 70300000-70500000.

IV.3 - MODELOS DE GERAÇÃO DE VAZÃO MENSAL

O desempenho dos modelos de geração foi avaliado pela comparação de estatísticas da série histórica com as médias das respectivas estatísticas de 50 séries geradas com extensão igual à observada. Como mencionado no capítulo III, o valor Q^N_{12} (o último valor da série histórica) foi utilizado como *semente* do processo de geração. Na tentativa de eliminar quaisquer efeitos que este valor, Q^N_{12}, tenha sobre as séries geradas, HAAN (1982) sugere que os primeiros 50 anos sejam descartados.

IV.3.1 - Modelo Lognormal

No caso univariado, o modelo conseguiu reproduzir adequadamente tanto as variações mensais das principais estatísticas (média, desvio padrão, coef. de variação, correlação com a vazão do mês anterior), assim como preservar os correlogramas de vazões mensais e anuais, apresentados na figura 4.6. O modelo não conseguiu preservar adequadamente as flutuações mensais da assimetria, superestimando-a a nível anual. As principais estatísticas a nível anual foram adequamente preservadas sendo apresentadas na tabela 4.2.

O esquema multivariado de geração apresentou melhores resultados a nível mensal e anual para a 'estação', variável resposta, da equação mais simples, (3.58), havendo uma tendência em superestimar as principais estatísticas para as demais 'estações', variáveis resposta, das equações restantes do esquema utilizado, que incluem um número maior de termos, como em (3.59). As figuras 4.7 e 4.8 apresentam as estatísticas mensais, respectivamente, para os modelos bi e trivariado, sendo as anuais apresentadas na tabela 4.2. Este esquema conseguiu preservar a estrutura correlacional cruzada entre vazões mensais para cada par de postos, apresentando a nível anual resultados razoáveis.

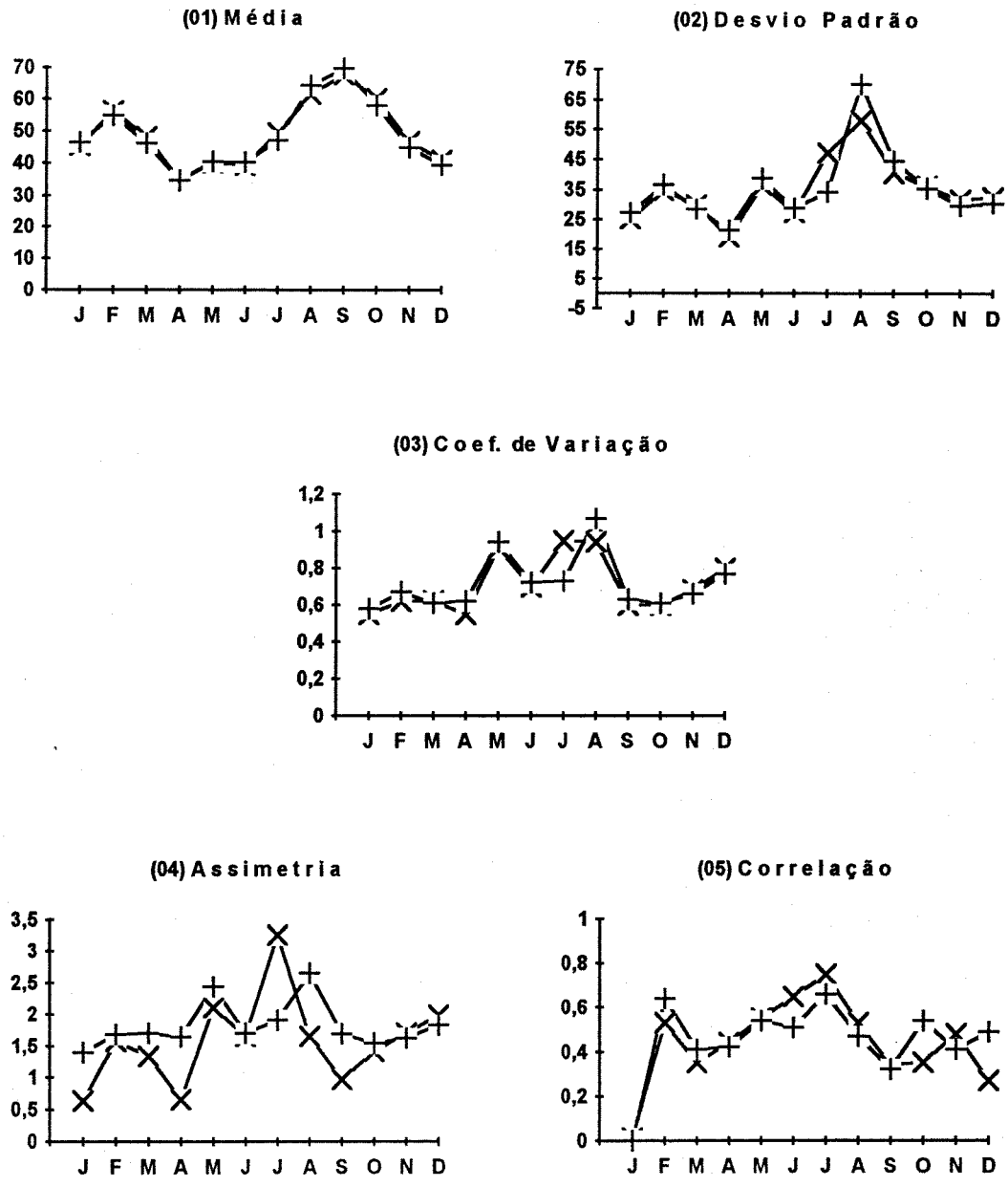


Figura 4.6 - Estatísticas históricas e geradas. Modelo Lognormal univariado. Estação: 71300000. x - observado + - gerado
(o_1) Média; (o_2) Desvio padrão; (o_3) Coeficiente de variação; (o_4) Assimetria, e (o_5) Correlação.

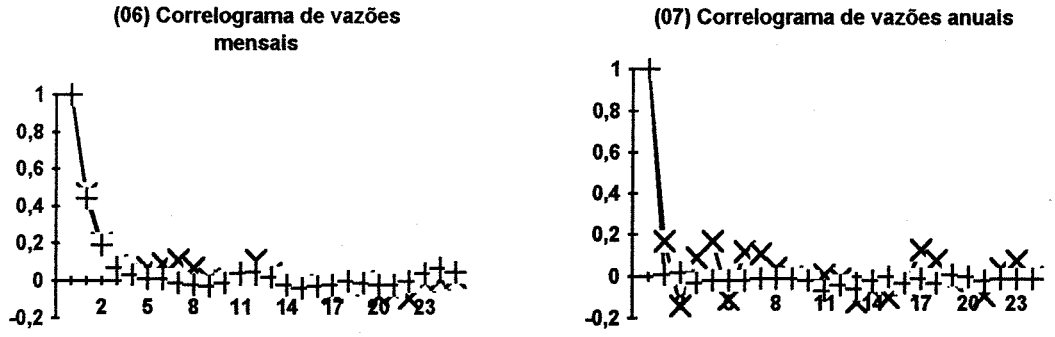


Figura 4.6 (Cont.) - Estatísticas históricas e geradas. Modelo Lognormal univariado. Estação: 71300000. x - observado + - gerado
(o_6) Correlograma de vazões mensais, e (o_7) Correlograma de vazões anuais.

Tabela 4.2 - Estatísticas anuais observadas e geradas.
Modelo Lognormal.

ESTAT.	UNIVARIADO	BIVARIADO		TRIVARIADO		
	71300000	70300000-70500000		70300000-70500000-71300000		
$\overline{X}_O$	49.34	27.52	17.03	27.52	17.05	49.34
$\overline{X}_G$	48.98	28.03	19.98	33.60	19.59	48.47
S_O	18.00	11.45	8.26	11.45	8.27	18.00
S_G	17.60	11.09	11.26	20.49	12.56	18.52
CV_O	0.36	0.42	0.49	0.42	0.49	0.36
CV_G	0.36	0.39	0.56	0.60	0.64	0.38
A_O	1.14	1.01	1.00	1.01	1.00	1.14
A_G	1.16	1.16	1.60	1.84	1.83	1.19
H_O	0.69	0.74	0.75	0.74	0.75	0.69
H_G	0.64	0.65	0.64	0.62	0.62	0.62

* SUBSCRITO: O-OBSERVADO; G-GERADO. $\overline{X}$-MÉDIA; S-DESVIO PADRÃO; CV-COEF. DE VARIAÇÃO; A-COEF. DE ASSIMETRIA; h-COEF. HURST.

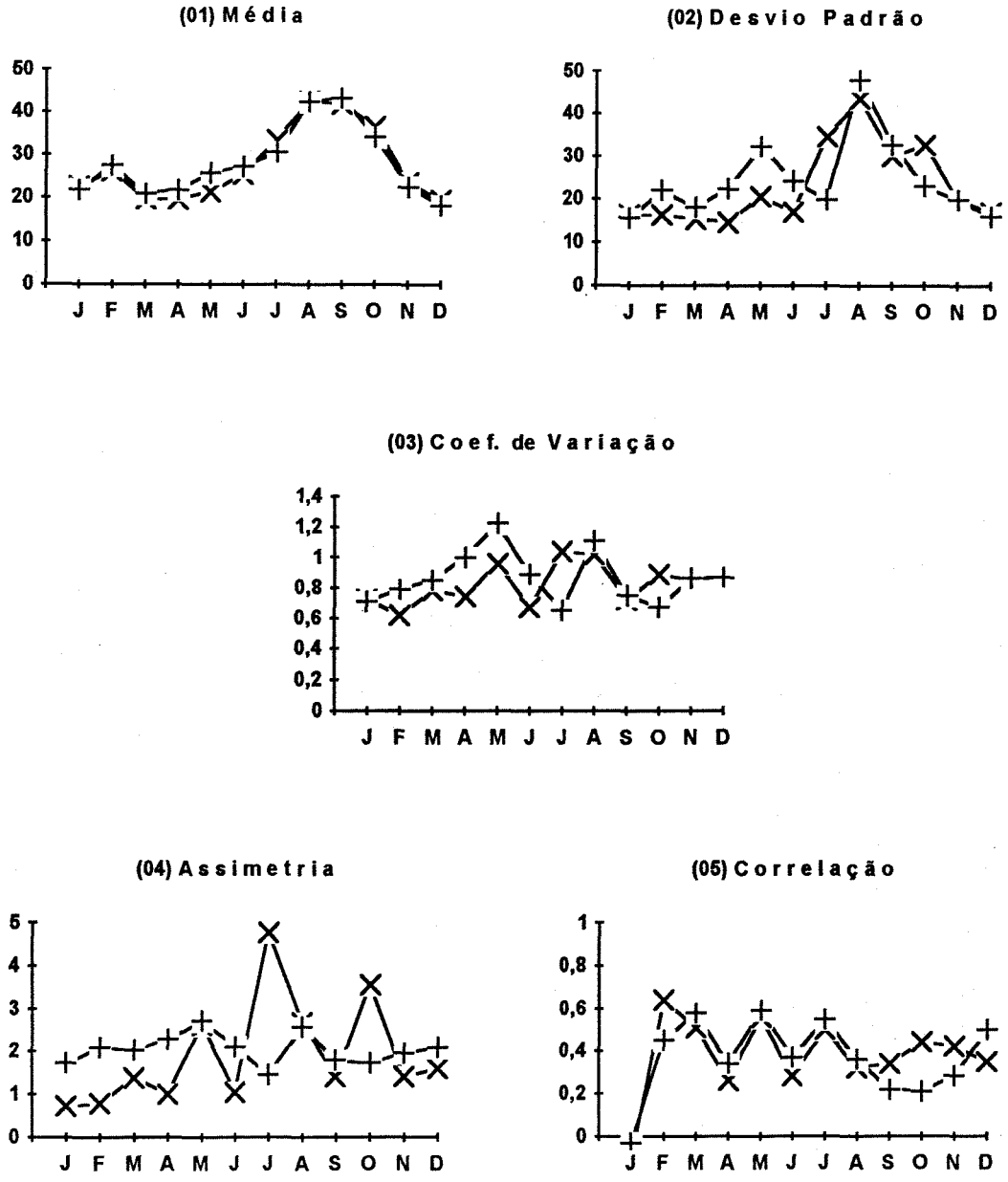


Figura 4.7 - Estatísticas históricas e geradas. Modelo Lognormal bivariado. Estações: 70300000-70500000. x - observado + - gerado
(o_1) Média-1; (o_2) Desvio padrão-1; (o_3) Coeficiente de variação-1; (o_4) Assimetria-1, e (o_5) Correlação-1.

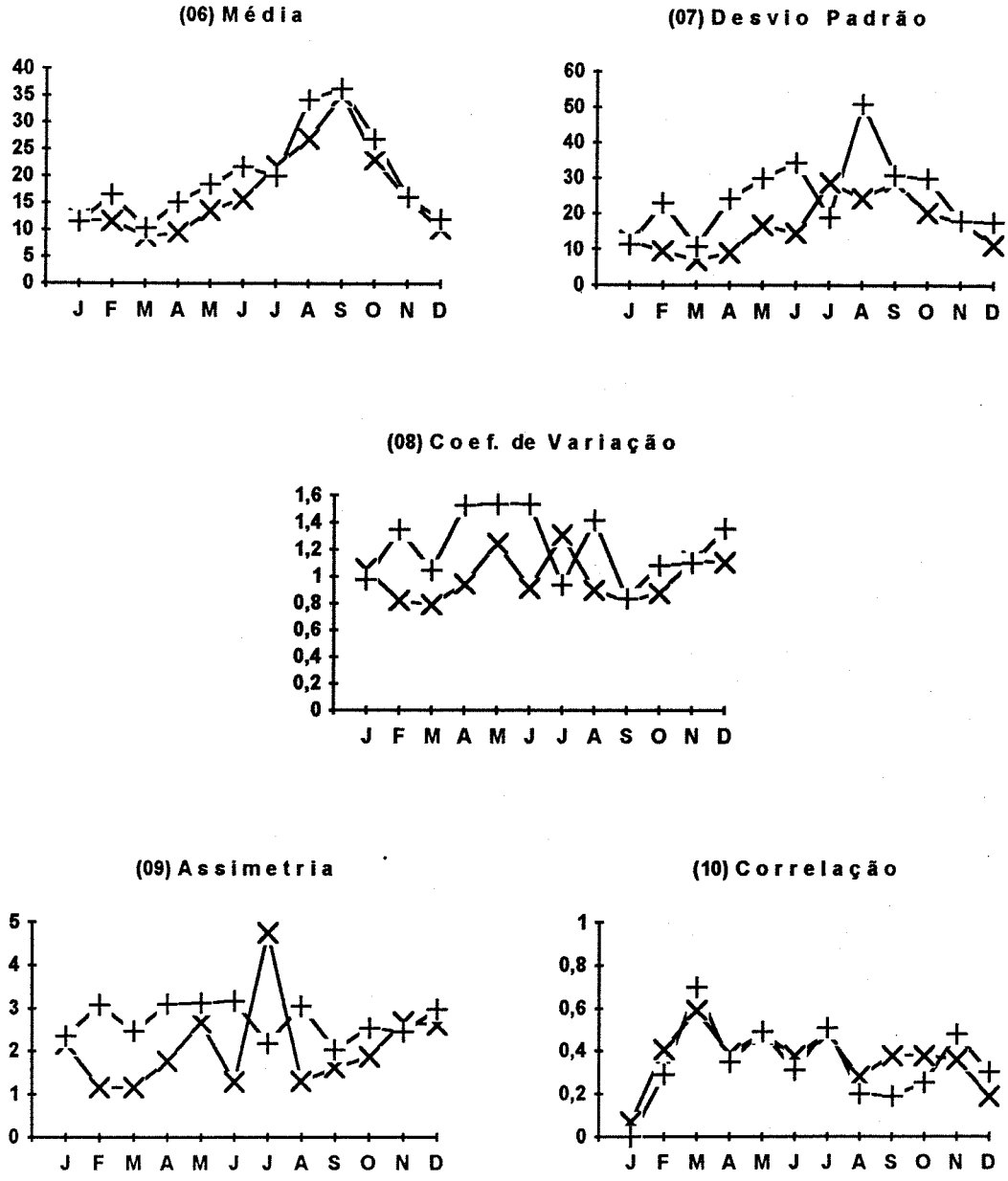


Figura 4.7 (Cont.) - Estatísticas históricas e geradas. Modelo Lognormal bivariado. Estações: 70300000-70500000. x - observado + - gerado
($_{06}$) Média-2; ($_{07}$) Desvio padrão-2; ($_{08}$) Coeficiente de variação-2; ($_{09}$) Assimetria-2, e ($_{10}$) Correlação-2.

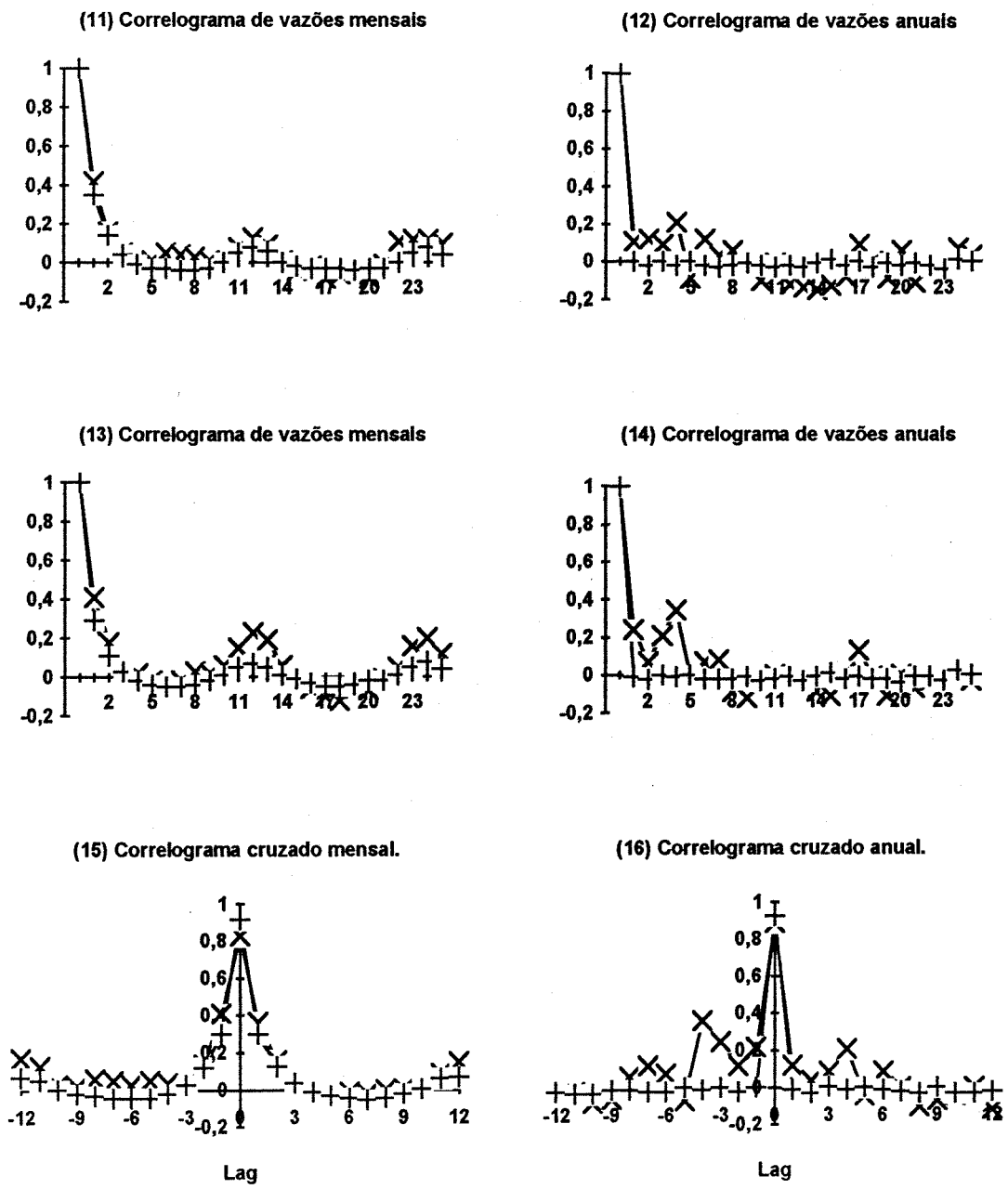


Figura 4.7 (Cont.) - Estatísticas históricas e geradas. Modelo Lognormal bivariado. Estações: 70300000-70500000. x - observado + - gerado
($_{11}$) Correlograma de vazões mensais-1; ($_{12}$) Correlograma de vazões anuais-1; ($_{13}$) Correlograma de vazões mensais-2; ($_{14}$) Correlograma de vazões anuais-2; ($_{15}$) Correlograma cruzado mensal 1-2, e ($_{16}$) Correlograma cruzado anual 1-2.

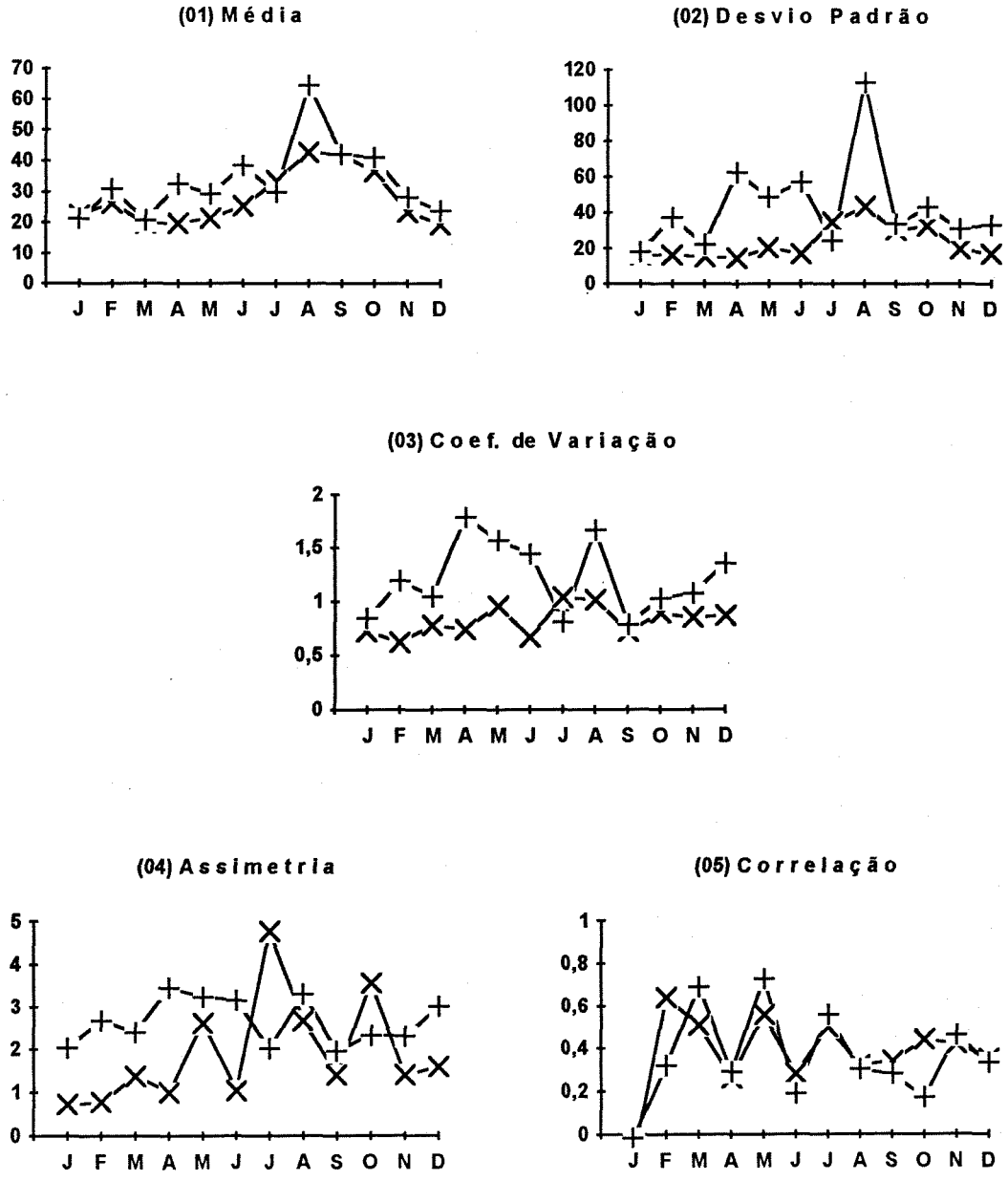


Figura 4.8 - Estatísticas históricas e geradas. Modelo Lognormal trivariado. Estações: 70300000-70500000-71300000. x - observado + - gerado (o_1) Média-1; (o_2) Desvio padrão-1; (o_3) Coeficiente de variação-1; (o_4) Assimetria-1, e (o_5) Correlação-1.

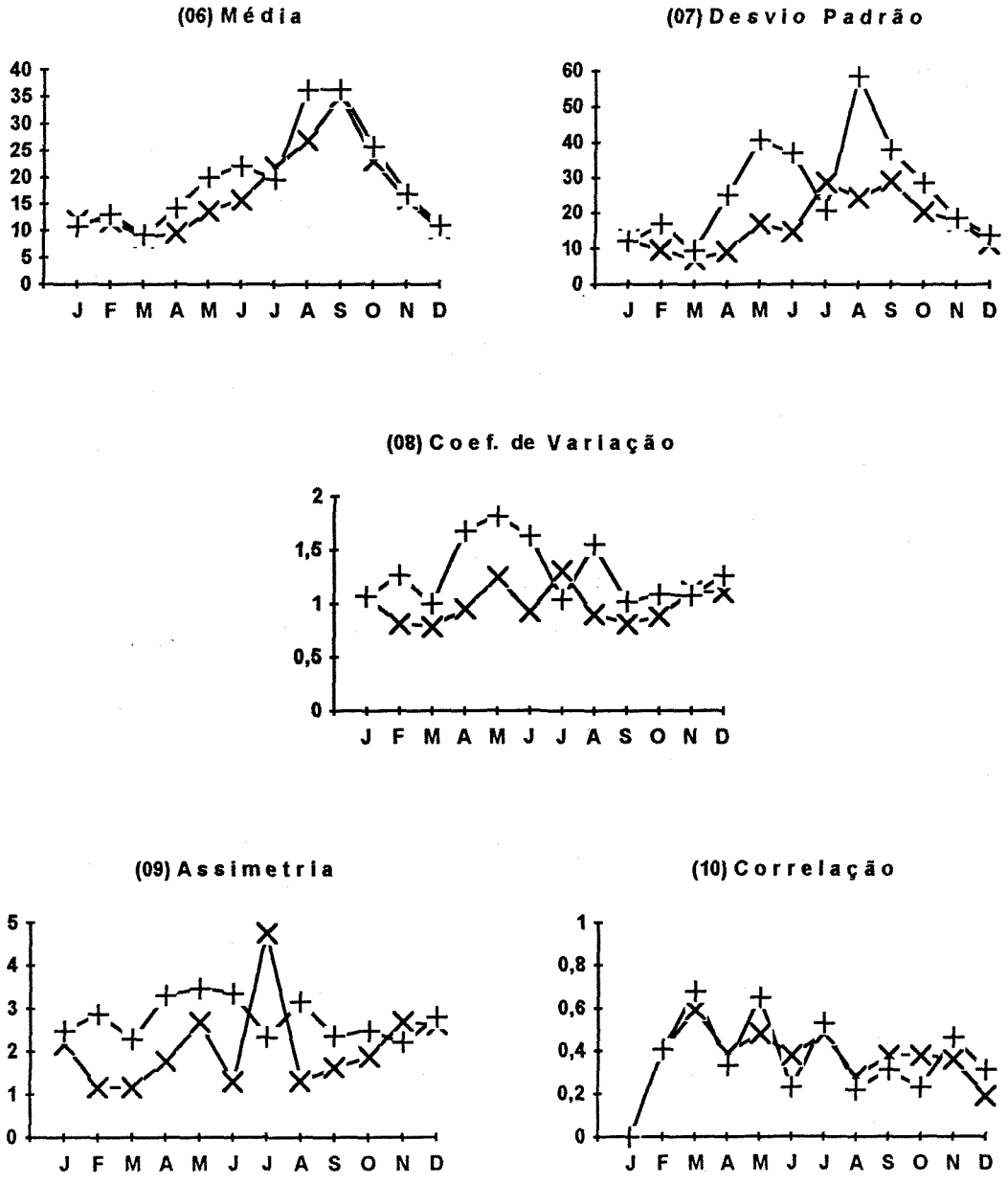


Figura 4.8 (Cont.) - Estatísticas históricas e geradas. Modelo Lognormal trivariado. Estações: 70300000-70500000-71300000. x - observado + - gerado (o_6) Média-2; (o_7) Desvio padrão-2; (o_8) Coeficiente de variação-2; (o_9) Assimetria-2, e ($1o$) Correlação-2.

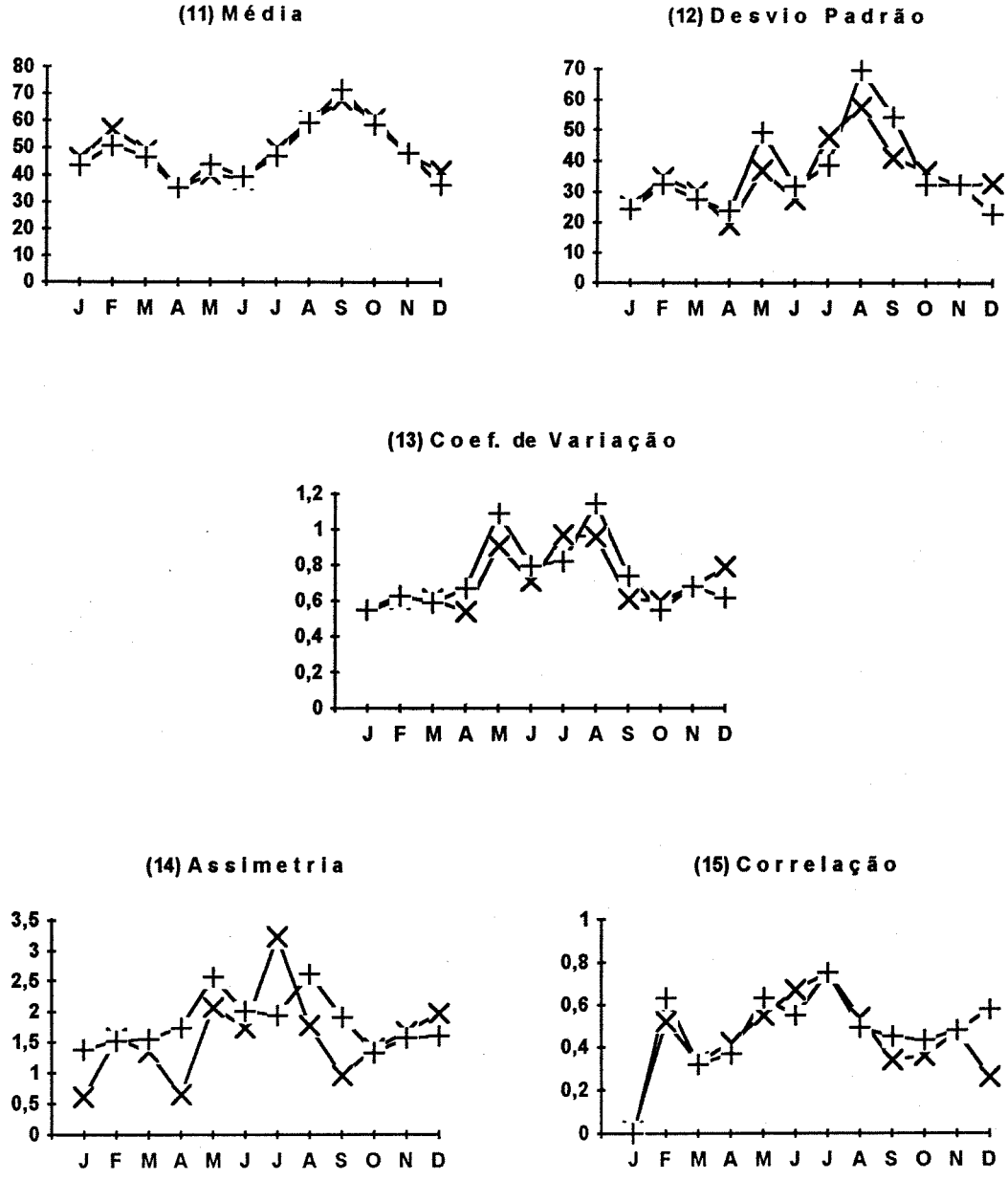


Figura 4.8 (Cont.) - Estatísticas históricas e geradas. Modelo Lognormal trivariado. Estações: 70300000-70500000-71300000. x - observado + - gerado ($_{11}$) Média-3; ($_{12}$) Desvio padrão-3; ($_{13}$) Coeficiente de variação-3; ($_{14}$) Assimetria-3, e ($_{15}$) Correlação-3.

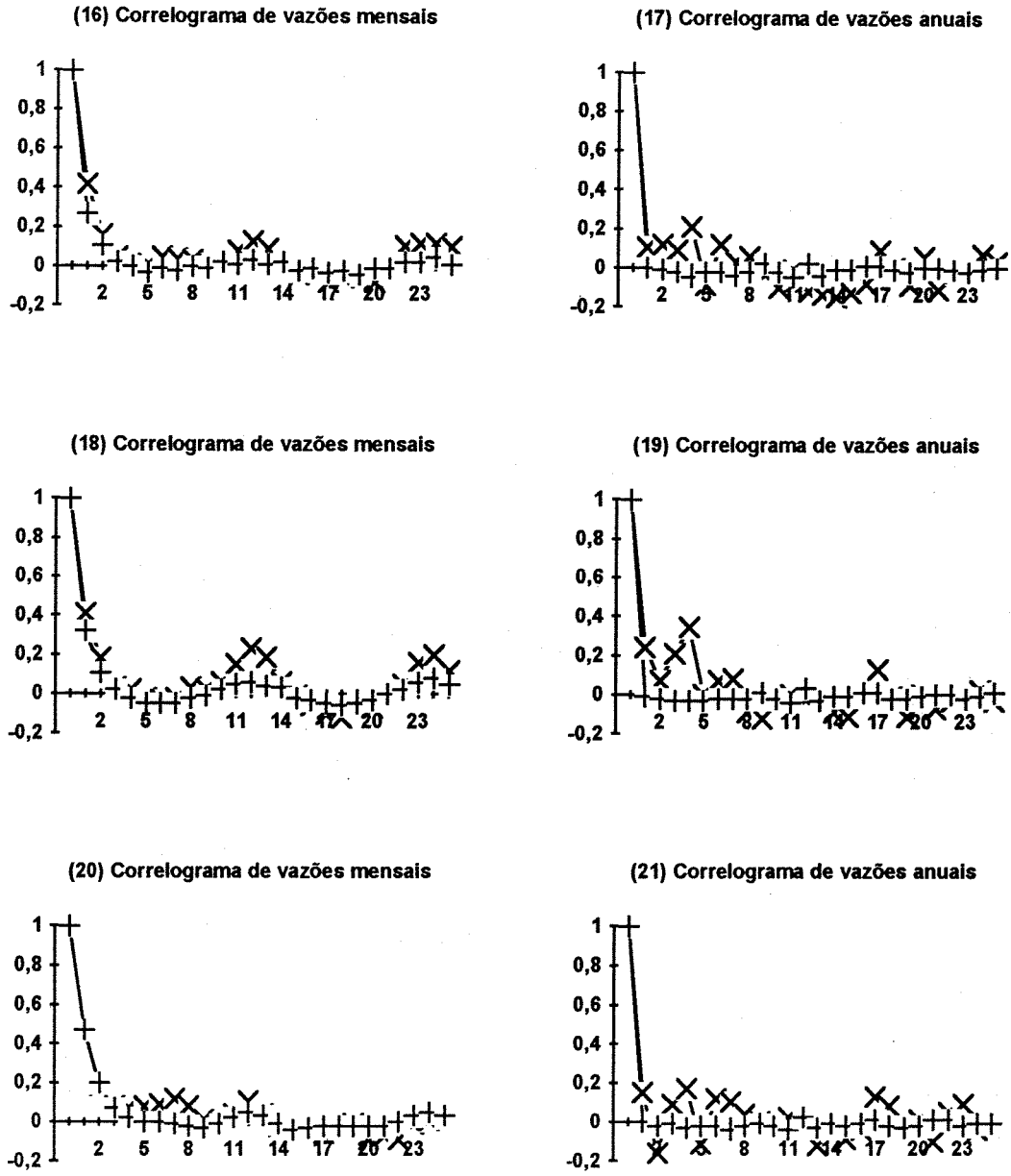
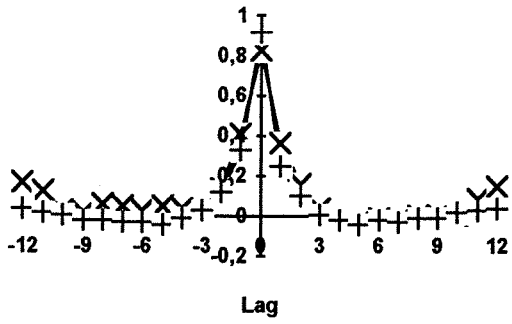
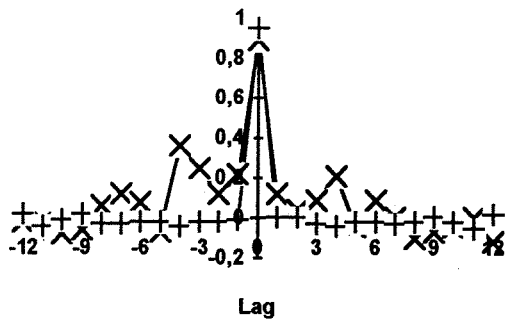


Figura 4.8 (Cont.) - Estatísticas históricas e geradas. Modelo Lognormal trivariado. Estações: 70300000-70500000-71300000. x - observado + - gerado ($_{16}$) Correlograma de vazões mensais-1; ($_{17}$) Correlograma de vazões anuais-1; ($_{18}$) Correlograma de vazões mensais-2; ($_{19}$) Correlograma de vazões anuais-2; ($_{20}$) Correlograma de vazões mensais-3, e ($_{21}$) Correlograma de vazões anuais-$_3$.

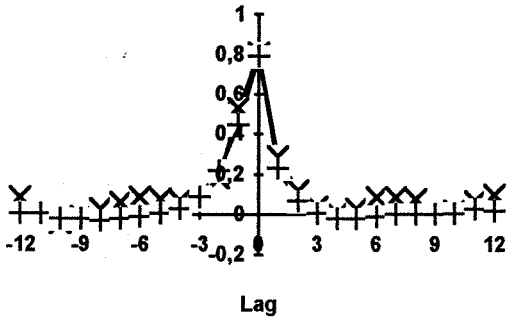
(22) Correlograma cruzado mensal.



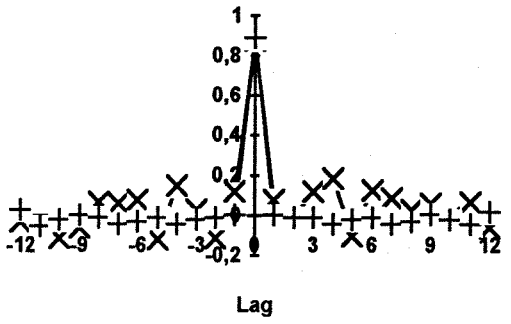
(23) Correlograma cruzado anual.



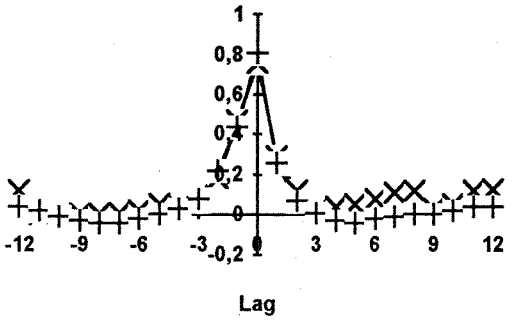
(24) Correlograma cruzado mensal.



(25) Correlograma cruzado anual.



(26) Correlograma cruzado mensal.



(27) Correlograma cruzado anual.

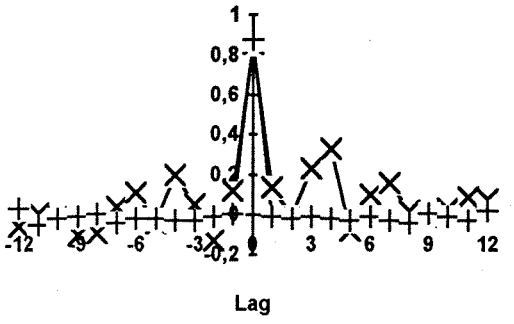


Figura 4.8 (Cont.) - Estatísticas históricas e geradas. Modelo Lognormal trivariado. Estações: 70300000-70500000-71300000. x - observado + - gerado ($_{22}$) Correlograma cruzado mensal 1-2; ($_{23}$) Correlograma cruzado anual 1-2, ($_{24}$) Correlograma cruzado mensal 1-3; ($_{25}$) Correlograma cruzado anual 1-3, ($_{26}$) Correlograma cruzado mensal 2-3,e ($_{27}$) Correlograma cruzado anual 2-3.

IV.3.2 - Modelo Gama

O modelo Gama-Identidade (3.47), caso univariado, bem como o Lognormal, conseguiu preservar adequadamente as flutuações mensais das principais estatísticas, como também os correlogramas de vazões mensais e anuais, apresentadas na figura 4.9. Do mesmo modo que o modelo Lognormal, as variações mensais da assimetria não são preservadas adequadamente, também superestimando-as a nível anual. As estatísticas anuais, na tabela 4.3, de maneira geral foram preservadas satisfatoriamente.

A modelagem multivariada com o modelo Gama comportou-se de maneira análoga ao caso univariado, conservando as mesmas estatísticas, e, além disto, como no modelo anterior, a correlação cruzada entre vazões mensais para cada par de postos foi preservada adequadamente, também apresentando a nível anual resultados razoáveis. As estatísticas mensais dos modelos bivariado e trivariado são apresentadas, respectivamente, nas figuras 4.10 e 4.11, enquanto que as estatísticas anuais constam na tabela 4.3.

O modelo Gama com a função de ligação logaritimica (3.47) apresenta problemas na geração com a ocorrência de *overflow*, da mesma forma que o modelo λ-*normal* (3.49), o que impossibilita o processo de geração.

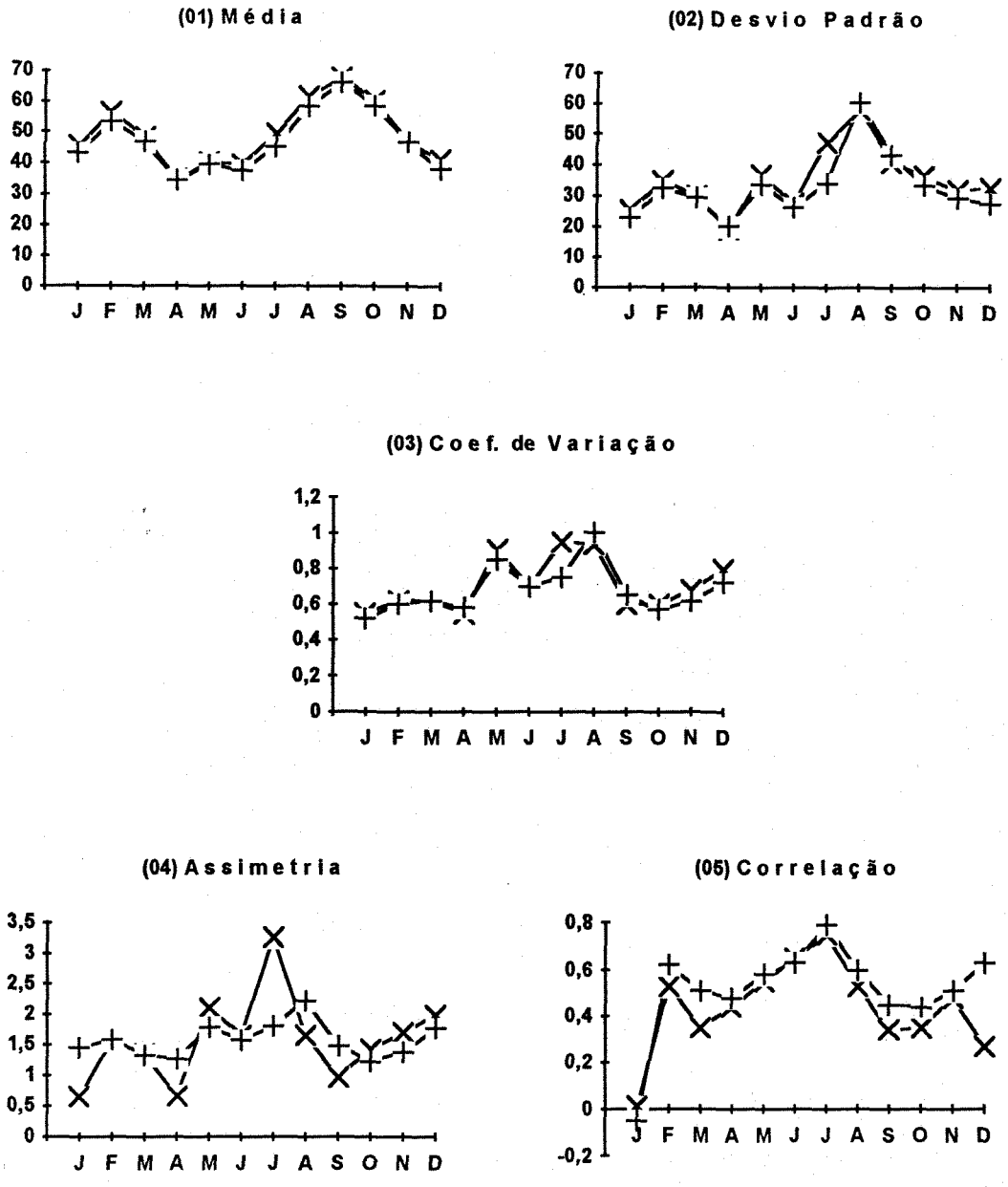
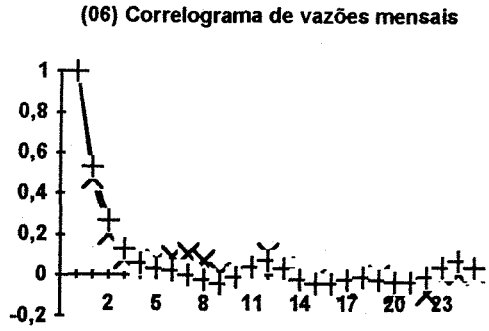


Figura 4.9 - Estatísticas históricas e geradas. Modelo Gama univariado. Estação: 71300000. x - observado + - gerado
(o_1) Média; (o_2) Desvio padrão; (o_3) Coeficiente de variação; (o_4) Assimetria, e (o_5) Correlação.



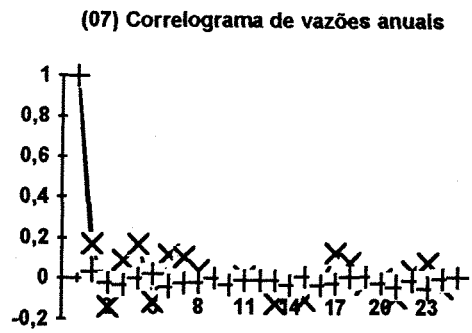


Figura 4.9 (Cont.) - Estatísticas históricas e geradas. Modelo Gama univariado. x - observado + - gerado
Estação: 71300000. (o_6) Correlograma de vazões mensais, e (o_7) Correlograma de vazões anuais.

Tabela 4.3 - Estatísticas anuais observadas e geradas.
Modelo Gama.

ESTAT.	UNIVARIADO	BIVARIADO		TRIVARIADO		
	71300000	70300000-71300000		70700000-71300000-72680000		
$\bar{X}_O$	49.34	27.52	49.34	194.96	49.34	88.45
$\bar{X}_G$	47.31	25.59	47.18	171.61	44.29	76.98
S_O	18.00	11.45	18.00	72.41	18.00	36.73
S_G	17.71	8.96	15.00	71.52	16.16	30.91
CV_O	0.36	0.42	0.36	0.37	0.36	0.42
CV_G	0.37	0.35	0.32	0.41	0.36	0.40
A_O	1.14	1.01	1.14	0.69	1.14	0.81
A_G	1.42	1.50	1.42	1.40	1.48	1.46
H_O	0.69	0.74	0.69	0.66	0.69	0.67
H_G	0.65	0.66	0.65	0.63	0.65	0.63

* SUBSCRITO: O-OBSERVADO; G-GERADO. $\bar{X}$-MÉDIA; S-DESVIO PADRÃO; CV-COEF. DE VARIAÇÃO; A-COEF. DE ASSIMETRIA; h-COEF. HURST.

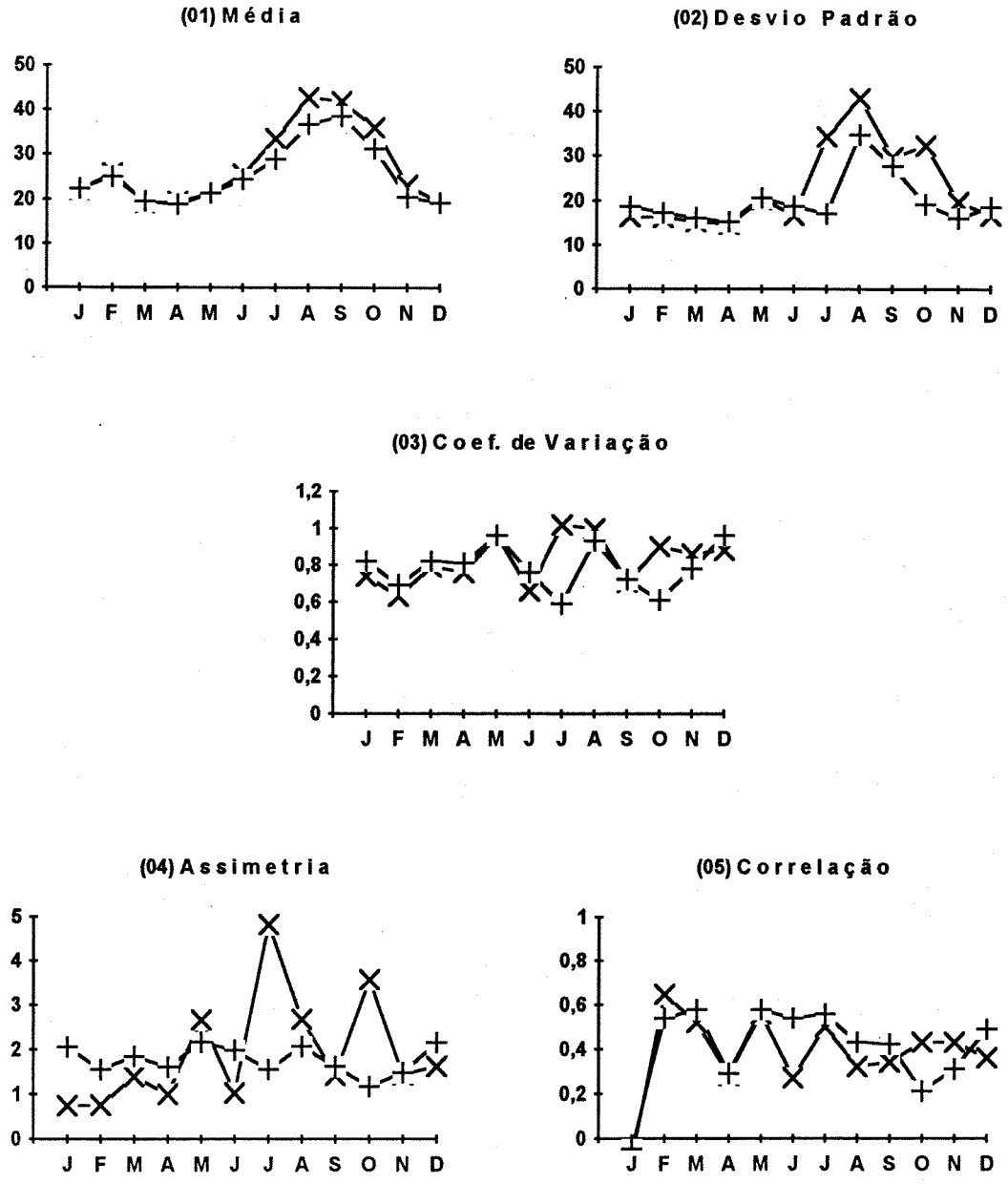


Figura 4.10 - Estatísticas históricas e geradas. Modelo Gama bivariado. Estações: 70300000-71300000. x - observado + - gerado
(o_1) Média-1; (o_2) Desvio padrão-1; (o_3) Coeficiente de variação-1; (o_4) Assimetria-1, e (o_5) Correlação-1.

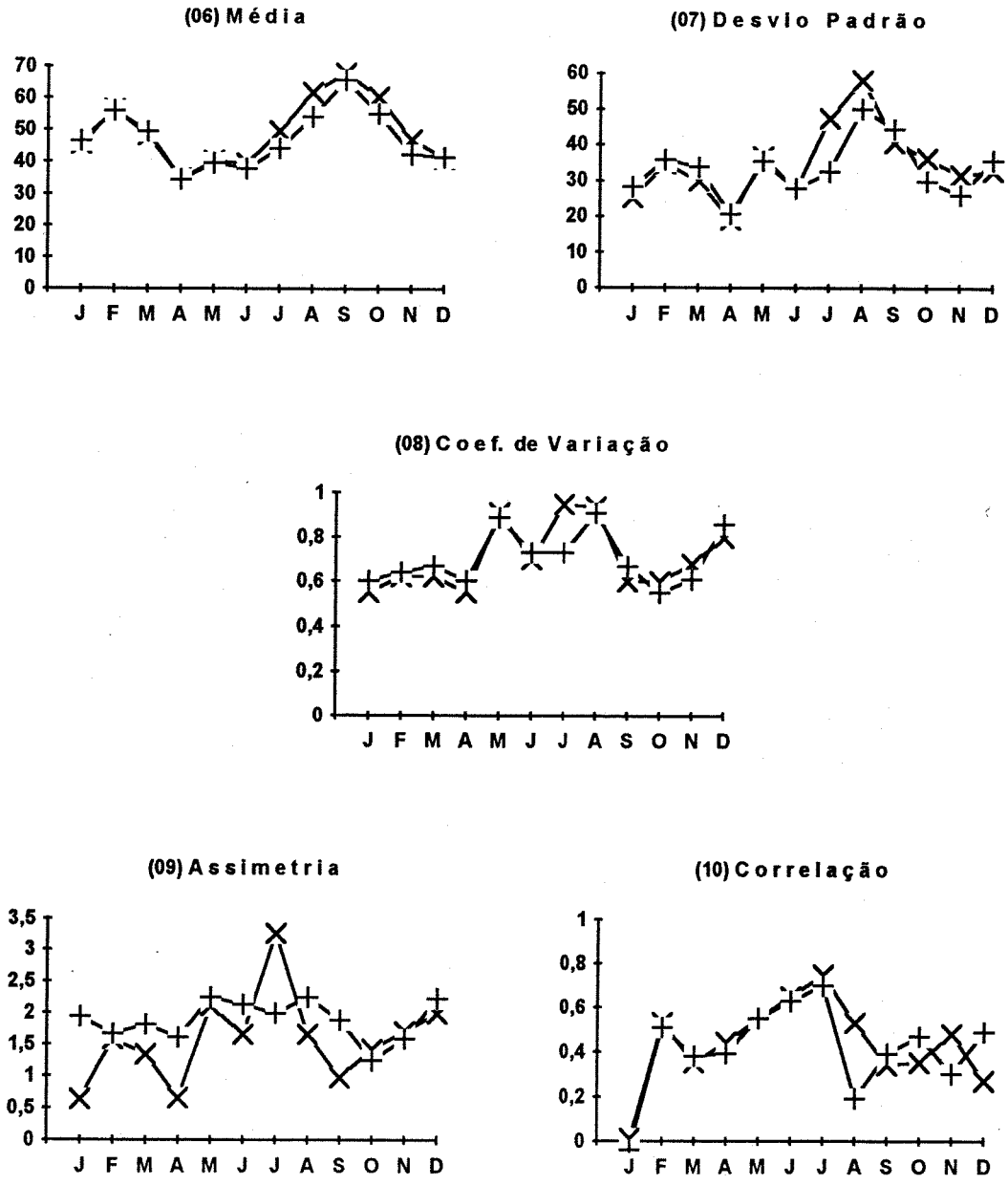
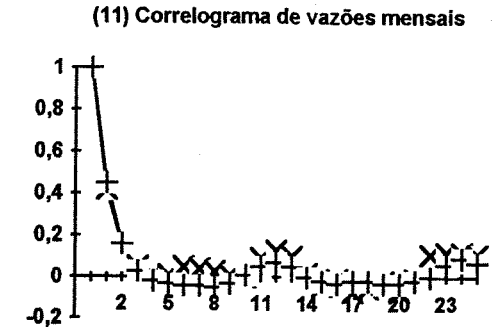
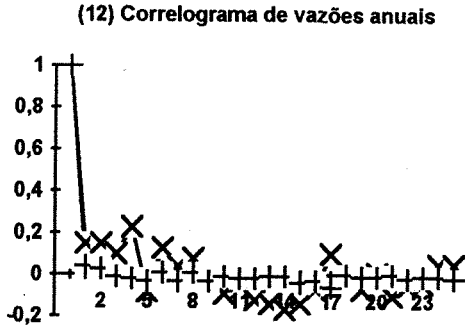
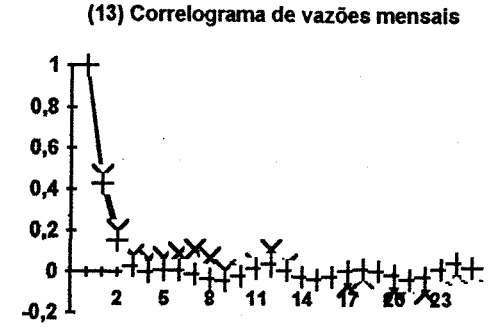
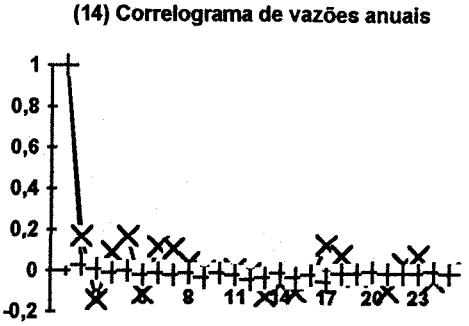


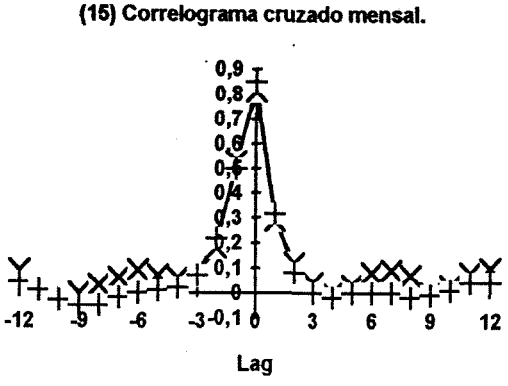
Figura 4.10 (Cont.) - Estatísticas históricas e geradas. Modelo Gama bivariado. Estações: 70300000-71300000. x - observado + - gerado
(o_6) Média-2; (o_7) Desvio padrão-2; (o_8) Coeficiente de variação-2, (o_9) Assimetria-2, e (o_{10}) Correlação-2.











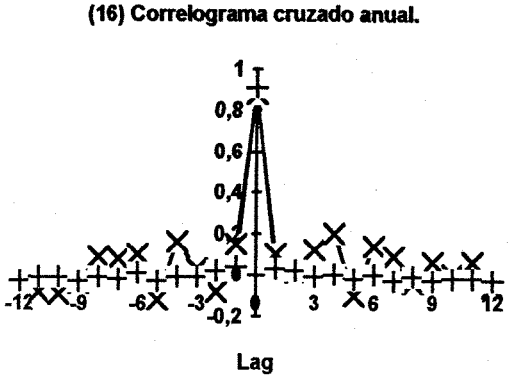


Figura 4.10 (Cont.) - Estatísticas históricas e geradas. Modelo Gama bivariado. Estações: 70300000-71300000. x - observado + - gerado
($_{11}$) Correlograma de vazões mensais-1; ($_{12}$) Correlograma de vazões anuais-1; ($_{13}$) Correlograma de vazões mensais-2; ($_{14}$) Correlograma de vazões anuais-2; ($_{15}$) Correlograma cruzado mensal 1-2, e ($_{16}$) Correlograma cruzado anual 1-2.

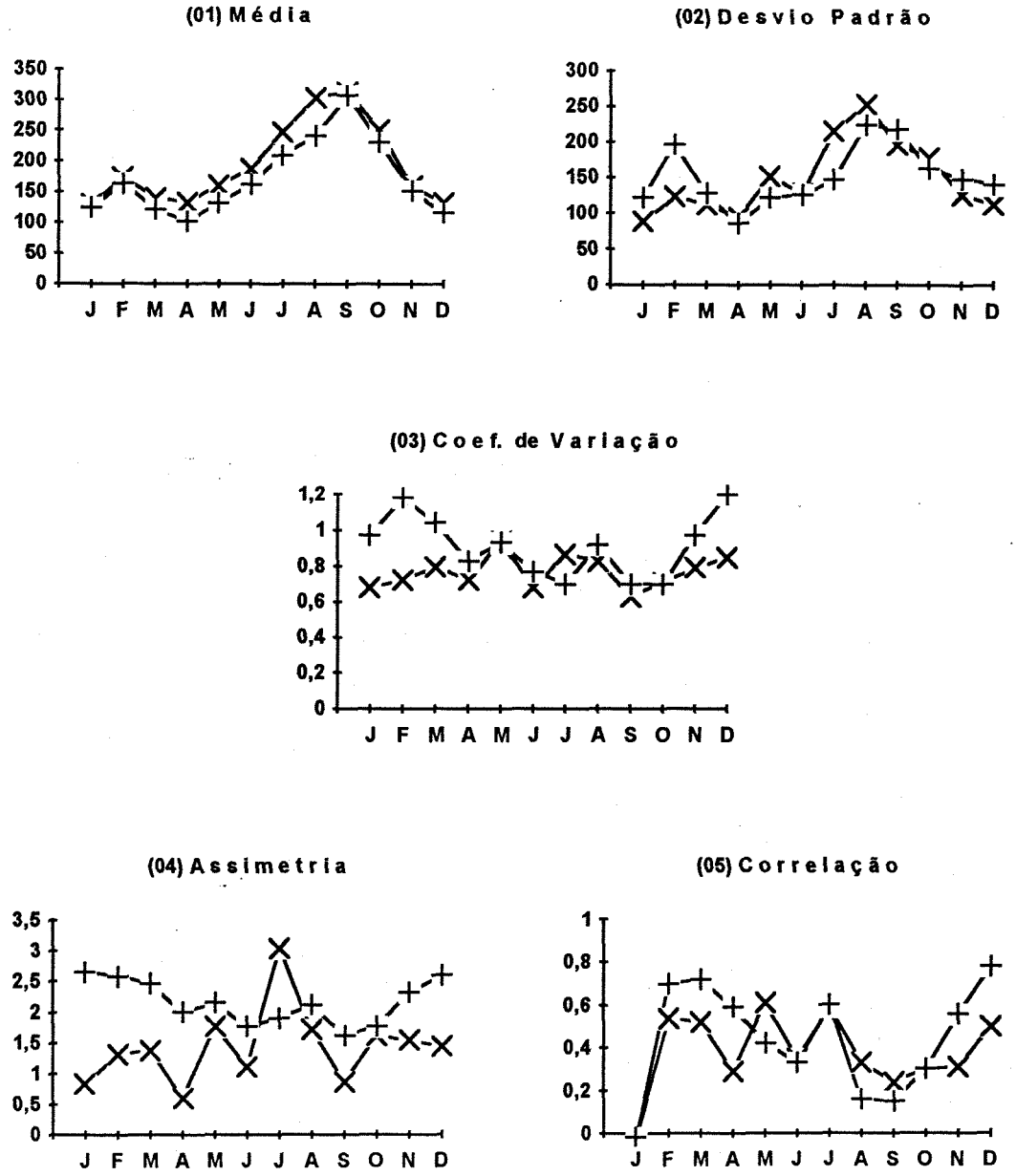


Figura 4.11 - Estatísticas históricas e geradas. Modelo Gama trivariado. Estações: 70700000-71300000-72680000. x - observado + - gerado (o_1) Média-1; (o_2) Desvio padrão-1; (o_3) Coeficiente de variação-1; (o_4) Assimetria-1, e (o_5) Correlação-1.

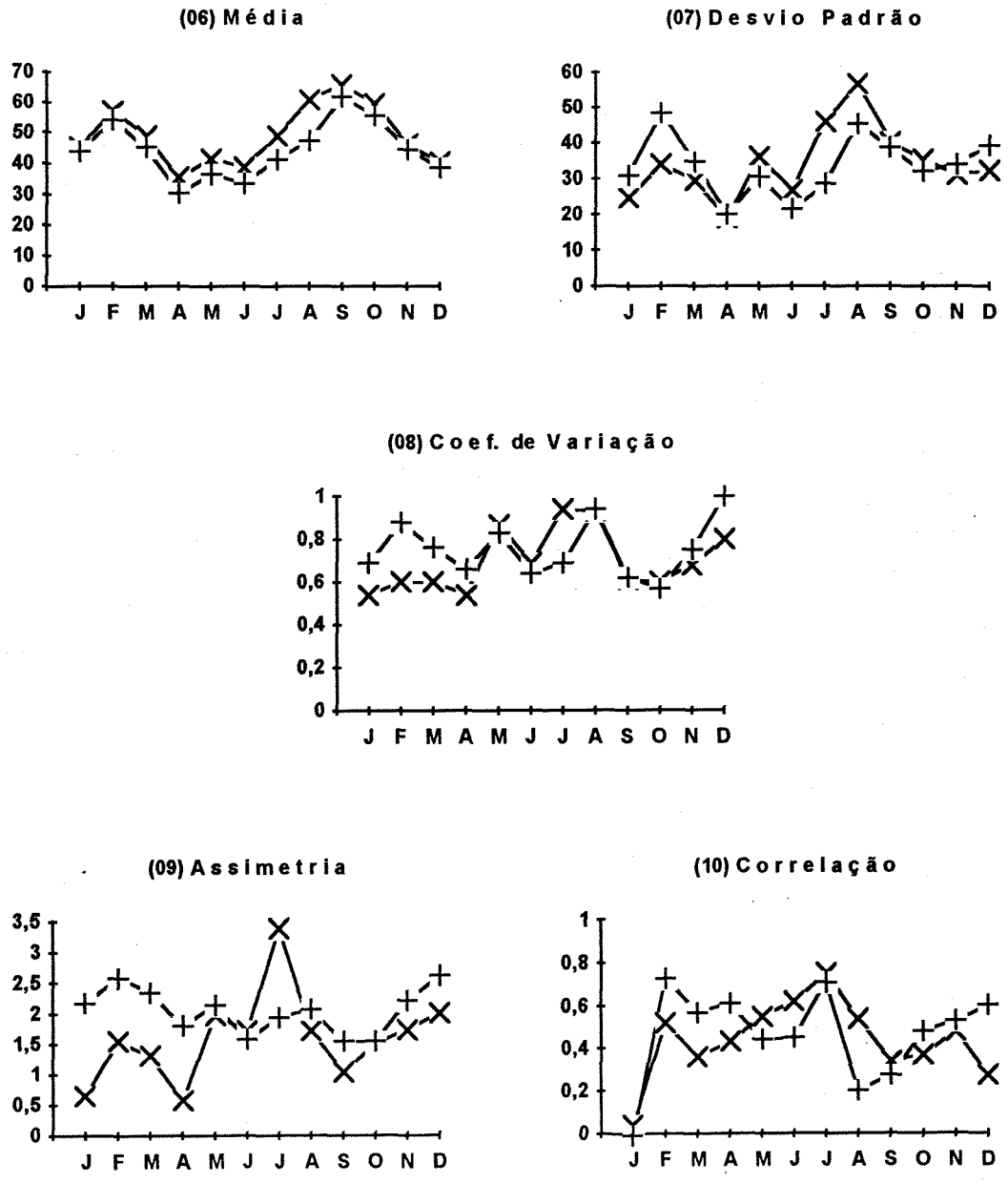


Figura 4.11 (Cont.) - Estatísticas históricas e geradas. Modelo Gama trivariado. Estações: 70700000-71300000-72680000. x - observado + - gerado (o_6) Média-2; (o_7) Desvio padrão-2; (o_8) Coeficiente de variação-2; (o_9) Assimetria-2, e (o_{10}) Correlação-2.

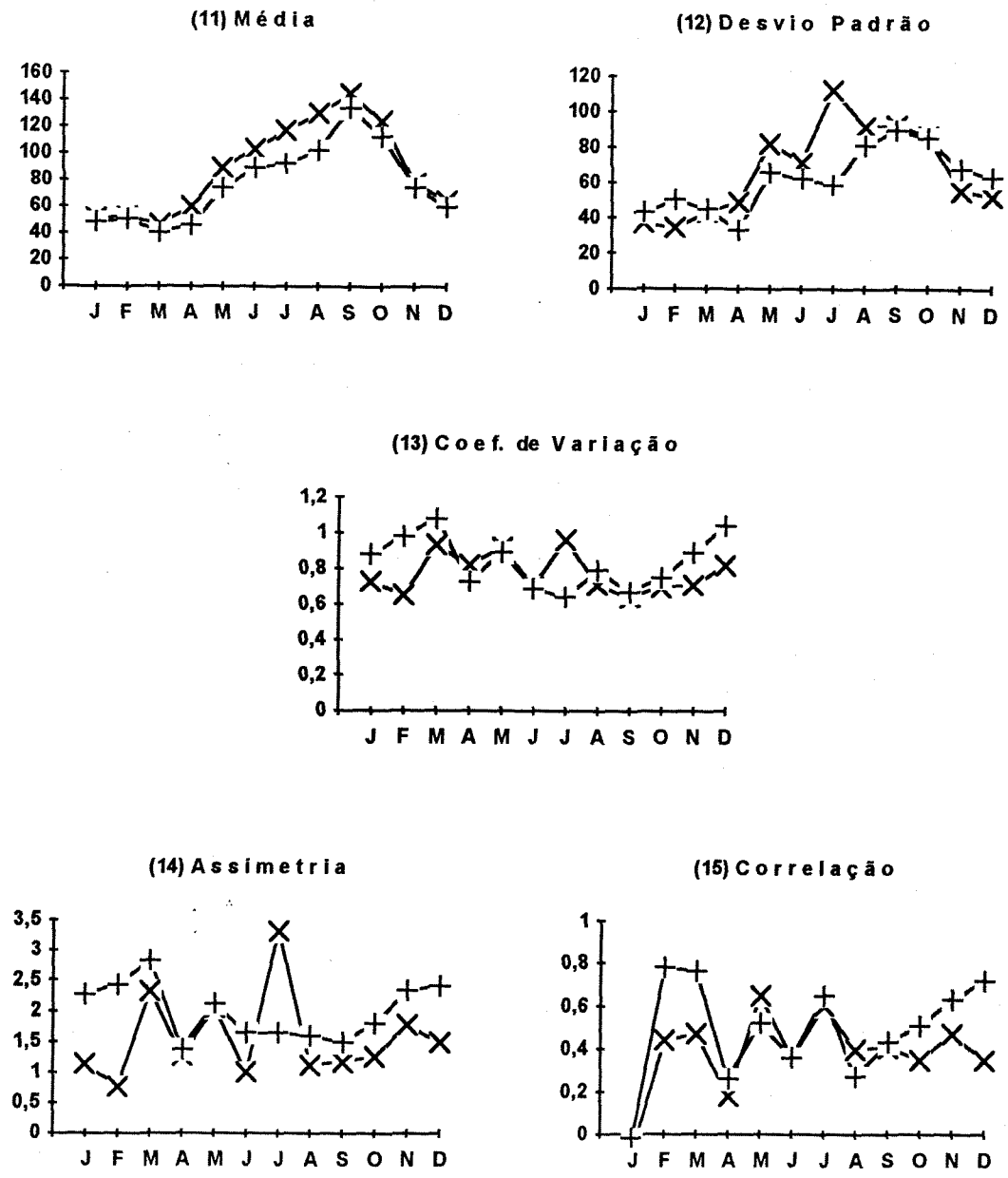


Figura 4.11 (Cont.) - Estatísticas históricas e geradas. Modelo Gama trivariado. Estações: 70300000-71300000-72680000. x - observado + - gerado ($_{11}$) Média-3; ($_{12}$) Desvio padrão-3; ($_{13}$) Coeficiente de variação-3; ($_{14}$) Assimetria-3, e ($_{15}$) Correlação-3.

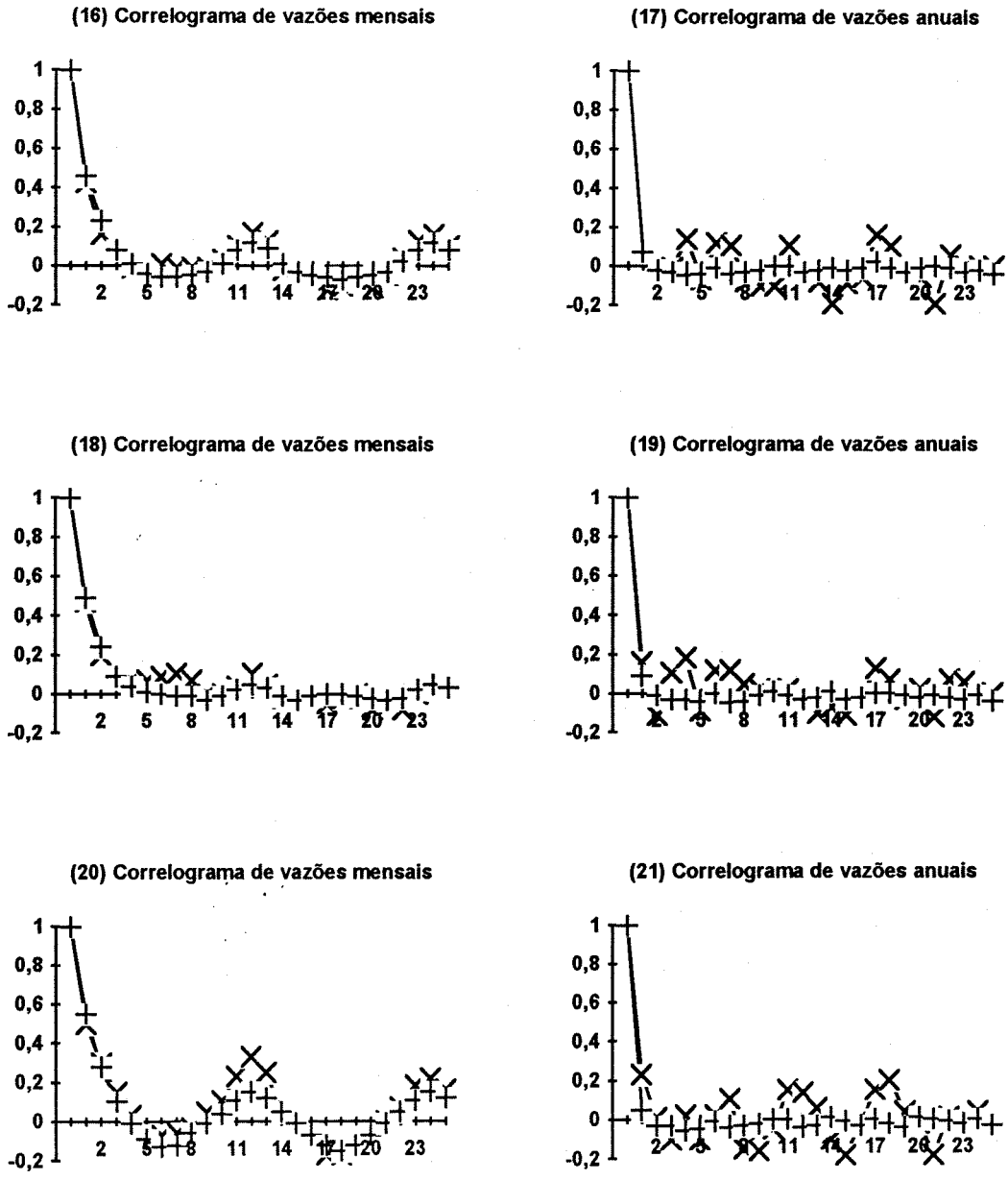
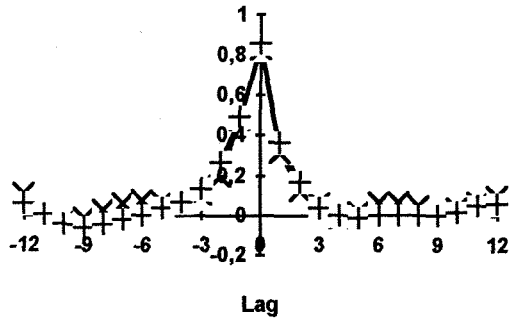
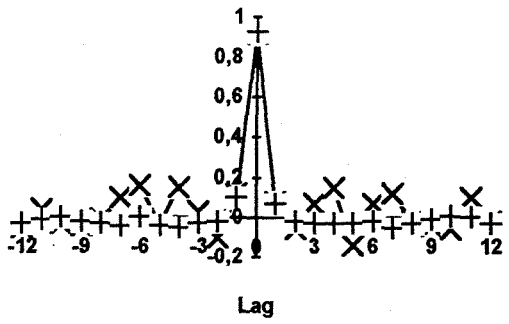


Figura 4.11 (Cont.) - Estatísticas históricas e geradas. Modelo Gama trivariado. Estações: 70300000-71300000-72680000. x - observado + - gerado ($_{16}$) Correlograma de vazões mensais-1; ($_{17}$) Correlograma de vazões anuais-1; ($_{18}$) Correlograma de vazões mensais-2; ($_{19}$) Correlograma de vazões anuais-2; ($_{20}$) Correlograma de vazões mensais-3, e ($_{21}$) Correlograma de vazões anuais-$_3$.

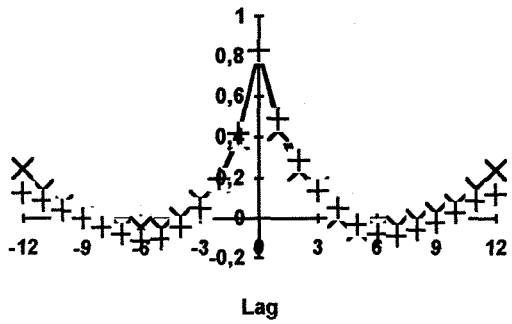
(22) Correlograma cruzado mensal.



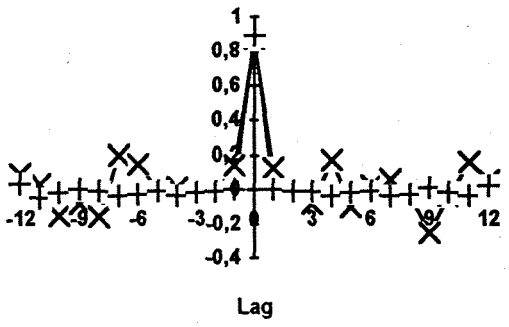
(23) Correlograma cruzado anual.



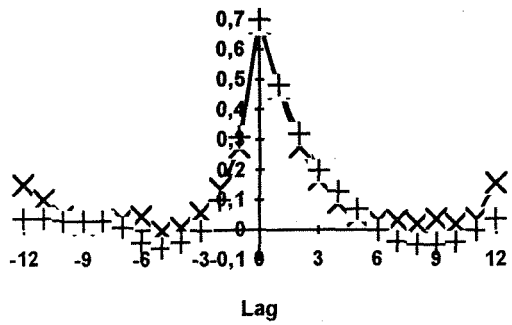
(24) Correlograma cruzado mensal.



(25) Correlograma cruzado anual.



(26) Correlograma cruzado mensal.



(27) Correlograma cruzado anual.

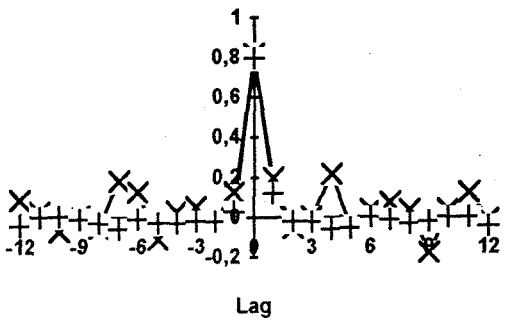


Figura 4.11 (Cont.) - Estatísticas históricas e geradas. Modelo Gama trivariado. Estações: 70300000-71300000-72680000. x - observado + - gerado ($_{22}$) Correlograma cruzado mensal 1-2; ($_{23}$) Correlograma cruzado anual 1-2, ($_{24}$) Correlograma cruzado mensal 1-3; ($_{25}$) Correlograma cruzado anual 1-3, ($_{26}$) Correlograma cruzado mensal 2-3,e ($_{27}$) Correlograma cruzado anual 2-3.

IV.3.3 - Modelo Média Normal - Variância Gama

O algoritmo apresentado em III.3, devido a AITKIN (1987), apresentou problemas para convergência principalmente quando se tratava da modelagem multivariada. Visando transpor este problema, uma transformação logarítmica das vazões foi realizada para o caso multivariado, garantindo o ajuste.

Para verificar a heterogeneidade da variância foi verificada a mudança na *deviance* do modelo da média; por exemplo, para o modelo univariado, aqui apresentado, com variável resposta a vazão de fevereiro: a *deviance* calculada no passo (ii) do algoritmo em III.3 (regressão normal não ponderada) foi de 520007.7, passando a 49.0 após atingida a convergência do algoritmo em III.3. Alternativamente, pode-se aplicar os testes apresentados no item III.1.3 ao modelo da dispersão. Sejam então, por exemplo, ainda continuando com o caso anterior, D_0 a *deviance* do modelo nulo da dispersão e D_1 a *deviance* do modelo que inclui a vazão do mês de janeiro:

$$D_0 = \phi.(138.012),\ D_1 = \phi.(123.820) \text{ e } \phi = 2,$$

$\Delta D = 28.384 > \chi^2_{1,95\%} = 3.84$ e $F = 5.502 > F_{1,48,95\%} = 4.04$[1], leva a concluir que não se pode ignorar os efeitos da vazão de janeiro sobre a dispersão da vazão de fevereiro. Ou melhor, deve-se rejeitar, ao nível de confiança de 95%, a hipótese de que a variância constante provê uma boa descrição da dispersão em favor da hipótese alternativa, que a inclusão da vazão de janeiro no modelo de dispersão fornece uma descrição melhor dos dados, a dispersão da vazão de fevereiro.

O modelo univariado, cujos resultados estão apresentados na figura 4.12 e tabela 4.4, conseguiu preservar, assim como os dois modelos

---

[1] Os valores tabelados $\chi^2_{1,95\%} = 3.84$ e $F_{1,48,95\%} = 4.04$ foram obtidos em HALD (1970).

anteriores, as principais estatísticas, e ainda acompanhou a tendência das flutuações mensais da assimetria, embora subestimando-a. A nível anual, a assimetria foi adequadamente preservada. Além da assimetria, as principais estatísticas anuais foram também preservadas.

No caso bivariado o modelo consegue apropriadamente acompanhar as flutuações mensais da média, coeficiente de variação, correlação mensal e o correlograma de vazões mensais, bem como o correlograma cruzado de vazões mensais. Por outro lado, o modelo não consegue descrever adequadamente as variações mensais no desvio padrão e assimetria, superestimando-os a nível anual. Na figura 4.13 são exibidos as estatísticas mensais do modelo bivariado e, na tabela 4.4, as estatísticas anuais.

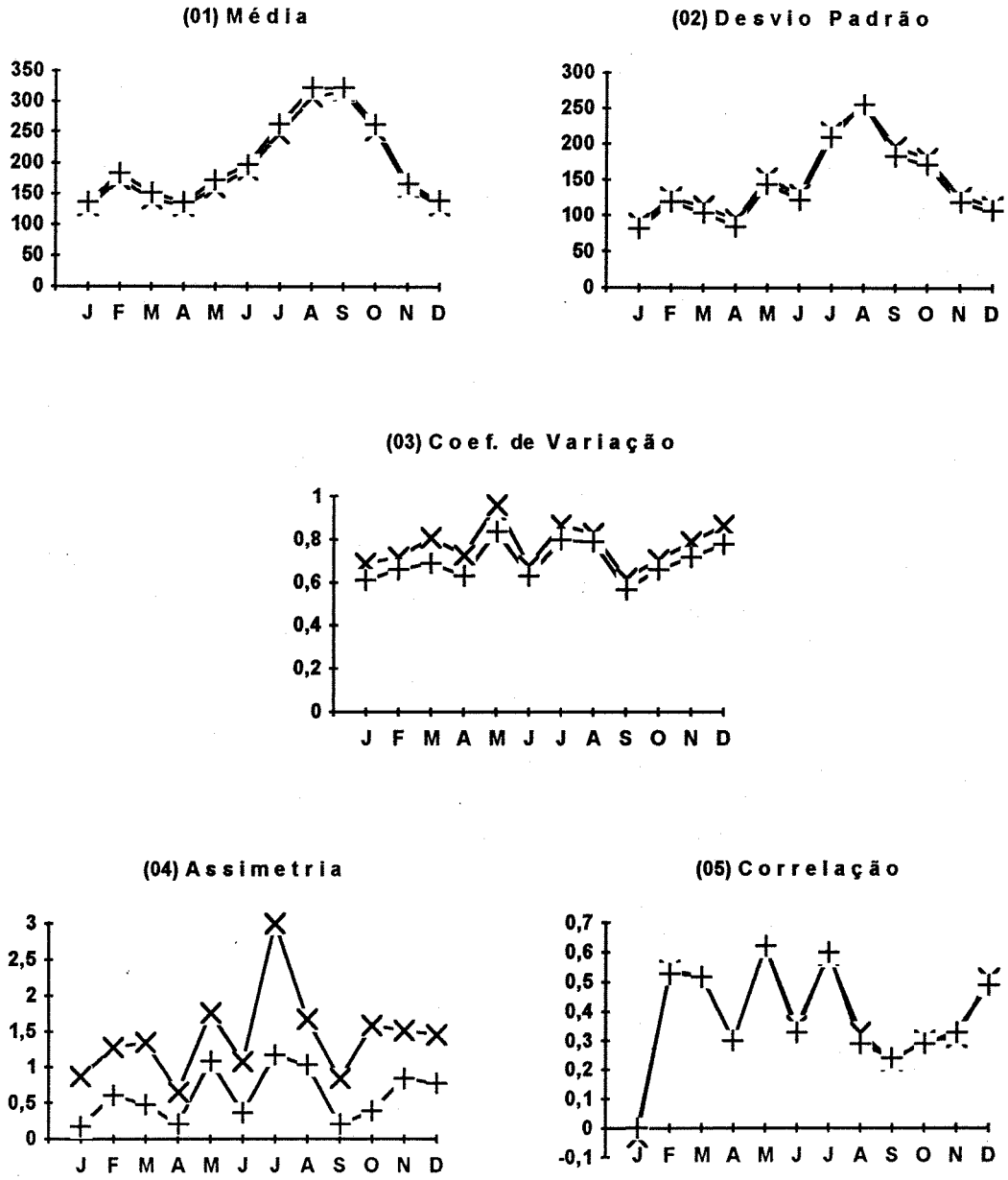


Figura 4.12 - Estatísticas históricas e geradas. Modelo Normal-Gama univariado. Estação: 70700000. x - observado + - gerado
(o_1) Média; (o_2) Desvio padrão; (o_3) Coeficiente de variação; (o_4) Assimetria, e (o_5) Correlação.

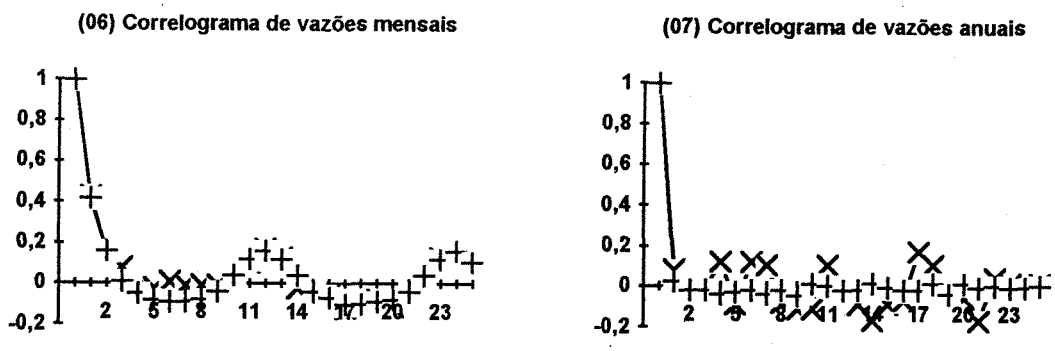


Figura 4.12 (Cont.) - Estatísticas históricas e geradas. Modelo Normal-Gama univariado. Estação: 70700000. x - observado + - gerado
(o_6) Correlograma de vazões mensais, e (o_7) Correlograma de vazões anuais.

Tabela 4.4 - Estatísticas anuais observadas e geradas.
Modelo Normal-Gama.

ESTAT.	UNIVARIADO	BIVARIADO	
	70700000	70500000-70700000	
$\overline{X}_O$	194.96	17.04	194.96
$\overline{X}_G$	204.15	17.42	223.55
S_O	72.41	11.09	72.41
S_G	63.64	10.76	122.08
CV_O	0.37	0.49	0.37
CV_G	0.31	0.55	0.53
A_O	0.69	1.00	0.69
A_G	0.71	1.53	1.54
H_O	0.66	0.75	0.66
H_G	0.63	0.62	0.63

* SUBSCRITO: O-OBSERVADO; G-GERADO. $\overline{X}$-MÉDIA; S-DESVIO PADRÃO; CV-COEF. DE VARIAÇÃO; A-COEF. DE ASSIMETRIA; h-COEF. HURST.

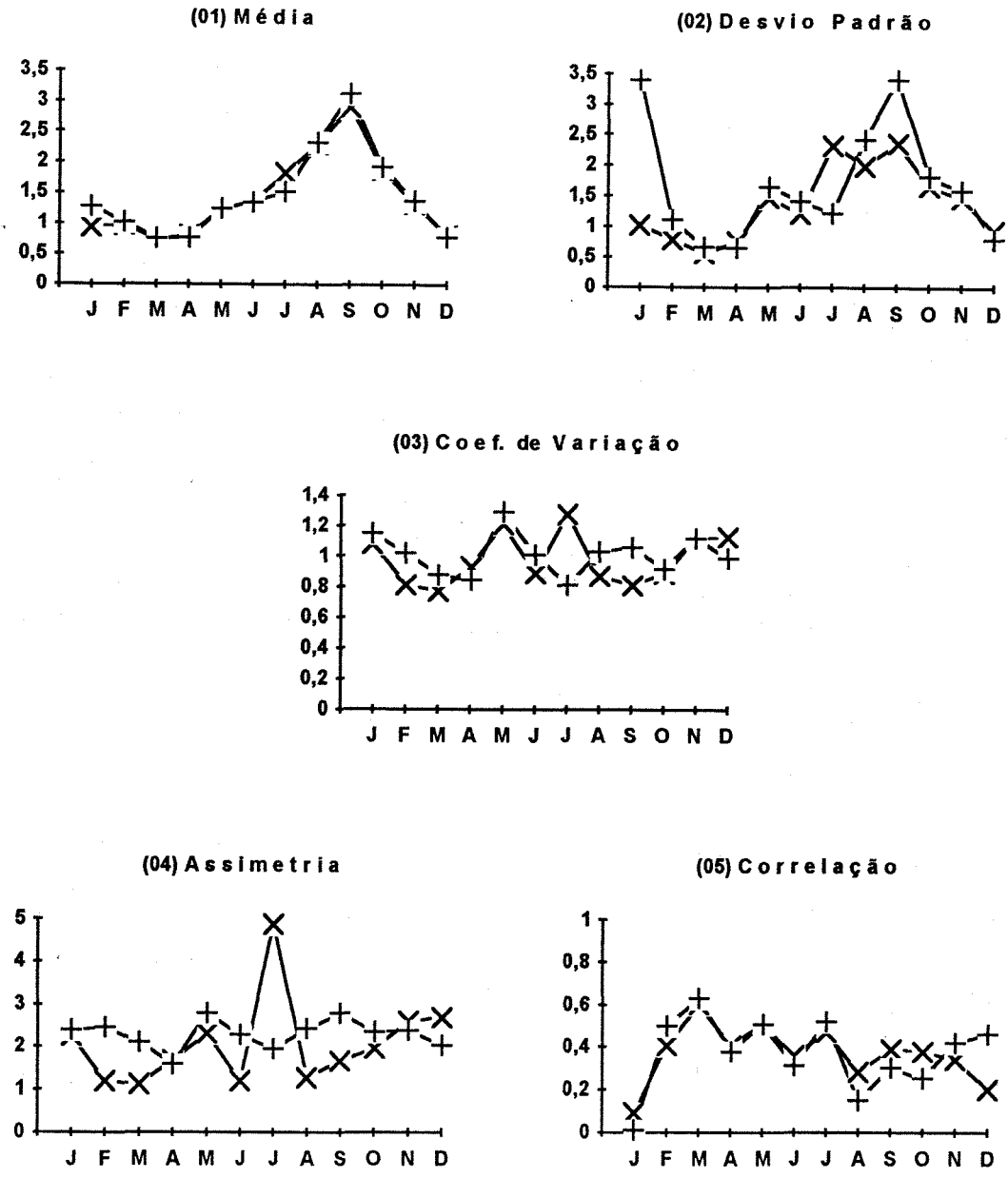


Figura 4.13 - Estatísticas históricas e geradas. Modelo Normal-Gama bivariado. Estações: 70500000-70700000. x - observado + - gerado (o_{01}) Média-1; (o_{02}) Desvio padrão-1; (o_{03}) Coeficiente de variação-1; (o_{04}) Assimetria-1, e (o_{05}) Correlação-1.

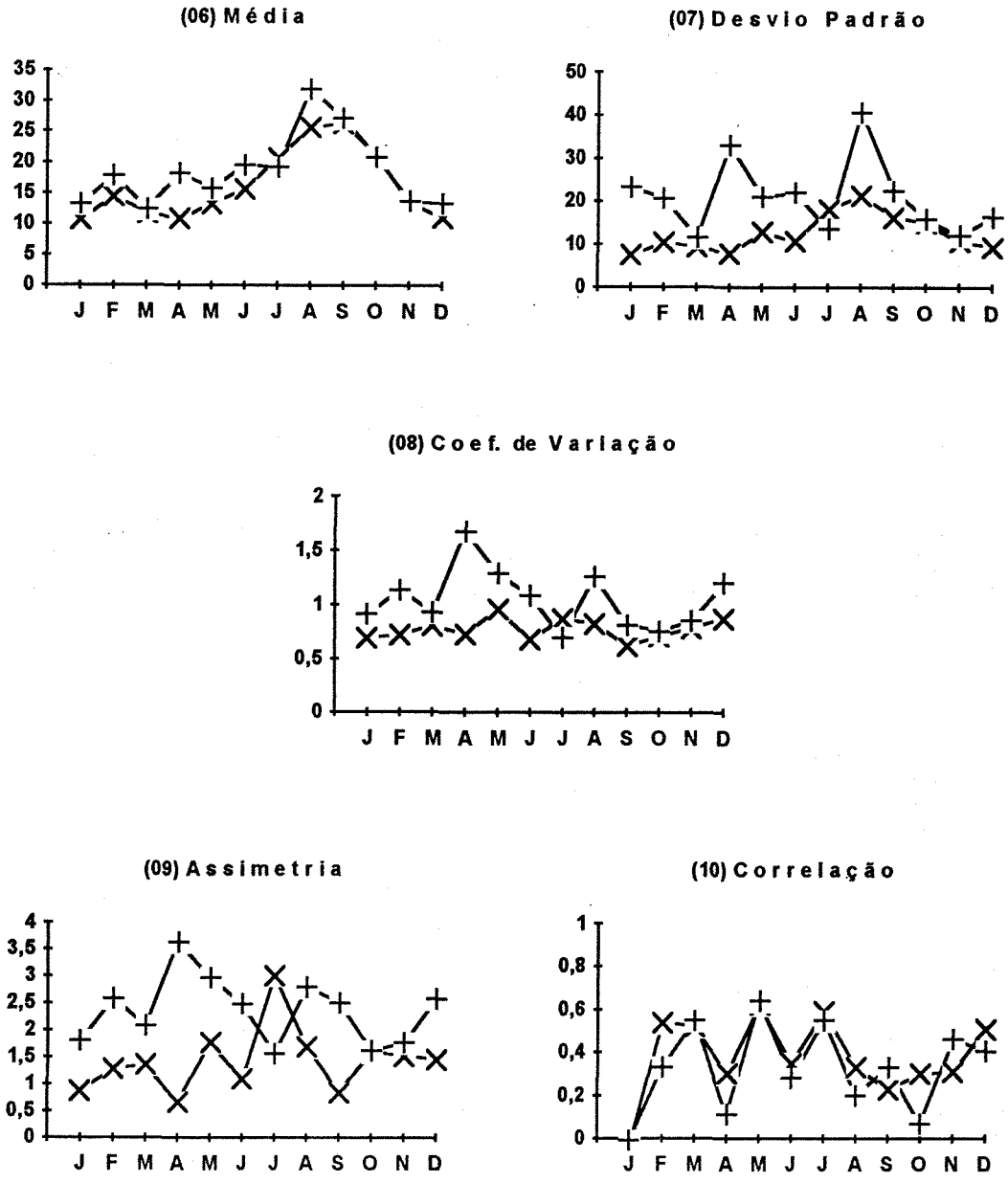


Figura 4.13 (Cont.) - Estatísticas históricas e geradas. Modelo Normal-Gama bivariado. Estações: 70500000-70700000. x - observado + - gerado
($_{06}$) Média-2; ($_{07}$) Desvio padrão-2; ($_{08}$) Coeficiente de variação-2; ($_{09}$) Assimetria-2, e ($_{10}$) Correlação-2.

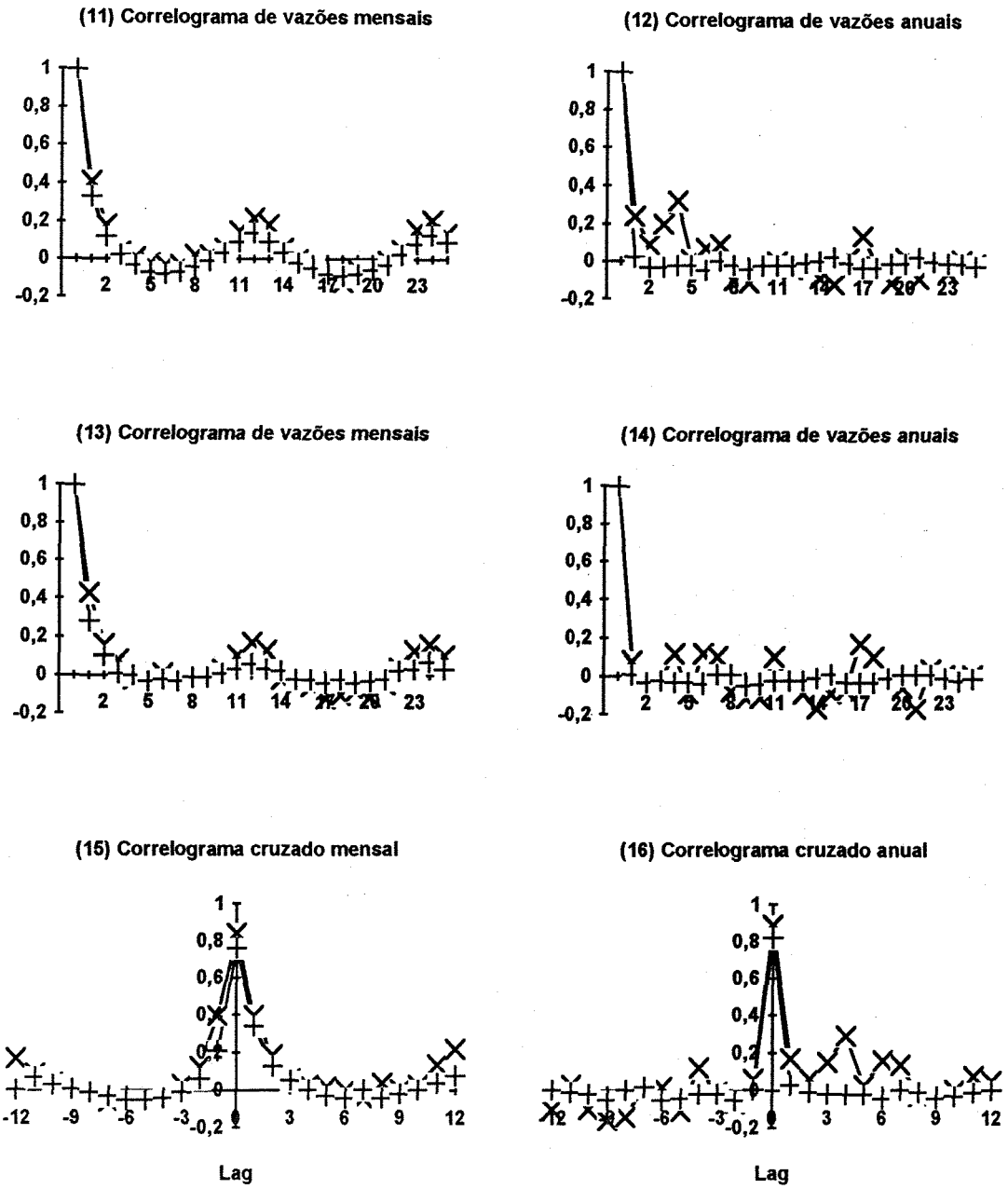


Figura 4.13 (Cont.) - Estatísticas históricas e geradas. Modelo Normal-Gama bivariado. Estações: 70500000-70700000. x - observado + - gerado
($_{11}$) Correlograma de vazões mensais-1; ($_{12}$) Correlograma de vazões anuais-1; ($_{13}$) Correlograma de vazões mensais-2; ($_{14}$) Correlograma de vazões anuais-2; ($_{15}$) Correlograma cruzado mensal 1-2, e ($_{16}$) Correlograma cruzado anual 1-2.

IV.3.4 - Modelagem de vazões mensais em rios intermitentes

A exploração da natureza do processo markoviano mostrou ser desnecessário levar em consideração a ocorrência ou não de vazão no mês anterior, ou melhor, os estados dos meses anteriores, sendo a equação 3.63 válida. A utilização do algoritmo apresentado na figura 3.6 para modelagem de vazões mensais seguindo a distribuição Lognormal consegue descrever adequadamente as flutuações mensais das principais estatísticas, sendo melhores os resultados para estação 36125000. A nível anual não foi possível obter uma descrição adequada das principais estatísticas. Os resultados a nível mensal para este caso são apresentados nas figuras 4.14 e 4.15 e a nível anual na tabela 4.5. Na tabela 4.6 são apresentadas as probabilidades de ocorrência observada, esperada e média gerada.

A distribuição Gama proporcionou em geral, relativamente à Lognormal, melhores resultados a nível anual, sendo os resultados a nível mensal comparáveis. Os resultados a nível mensal para este caso são apresentados nas figuras 4.16 e 4.17 e a nível anual na tabela 4.7. Na tabela 4.8 são apresentadas as probabilidades de ocorrência observada, esperada e média gerada.

Esta mescla de MLGs, ocorrência e quantidade, em uma tentativa de considerar a variabilidade não só das quantidades, mas também da ocorrência, promove o que com o esquema original não era possível: a ocorrência de vazões nulas em meses que na série observada não apresentaram vazões nulas.

Os parâmetros da distribuição, ponto (3) do algoritmo da figura 3.6, foram estimados pelo método da máxima verossimilhança.

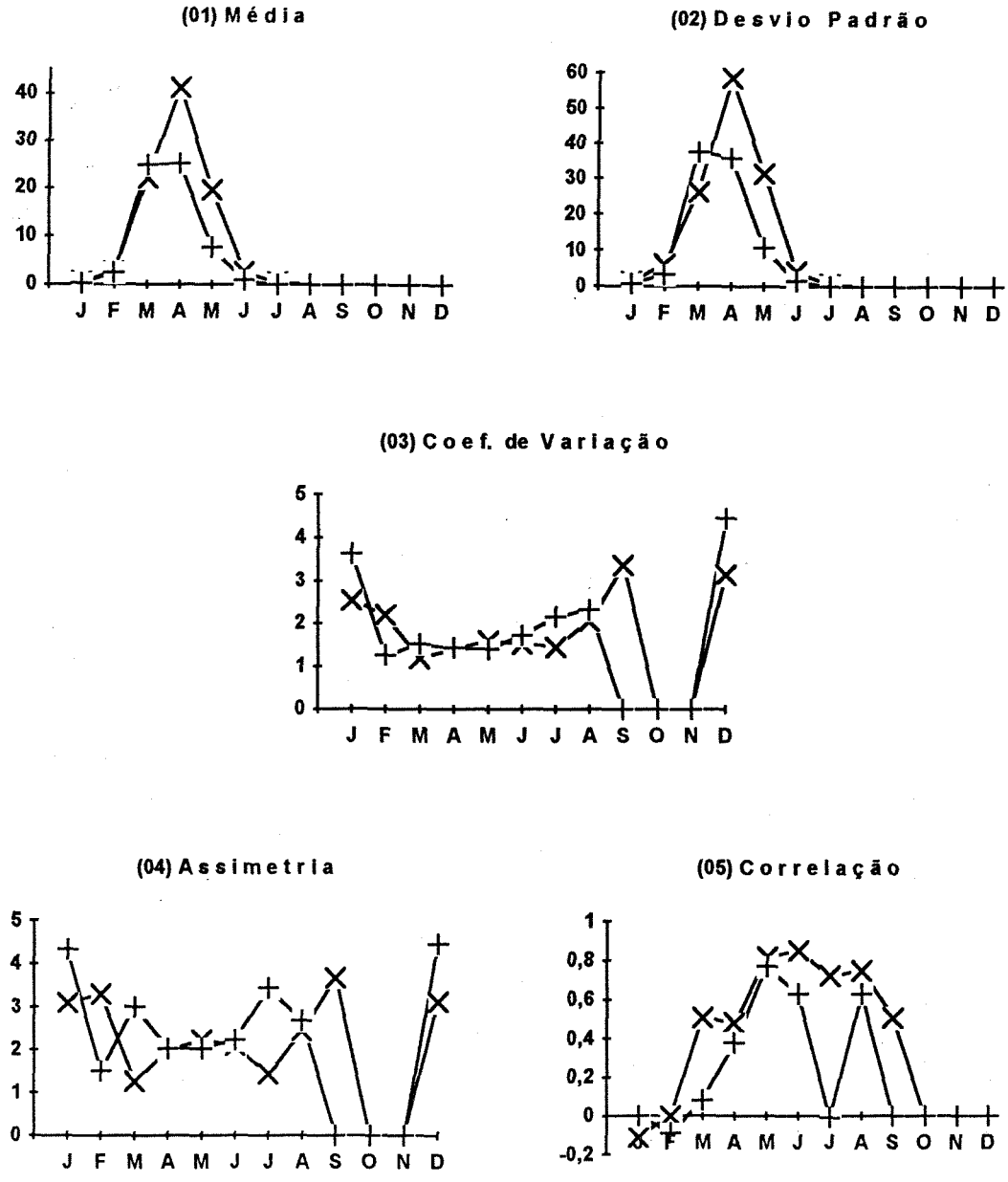


Figura 4.14 - Estatísticas históricas e geradas. Modelo Lognormal univariado. Estação: 35210000. x - observado + - gerado
(o_1) Média; (o_2) Desvio padrão; (o_3) Coeficiente de variação; (o_4) Assimetria, e (o_5) Correlação.

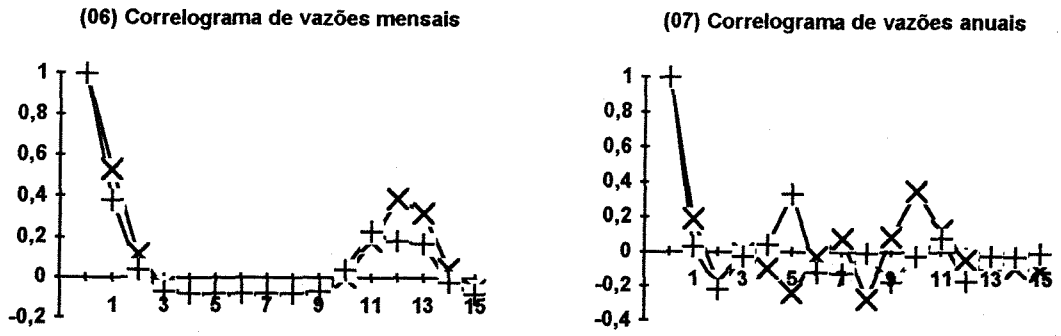


Figura 4.14 (Cont.) - Estatísticas históricas e geradas. Modelo Lognormal univariado. Estação: 35210000. x - observado + - gerado
(o_6) Correlograma de vazões mensais, e (o_7) Correlograma de vazões anuais.

Tabela 4.5 - Estatísticas anuais observadas e geradas. Modelo Lognormal.

	35210000	36125000
$\overline{X}_O$	7.45	3.89
$\overline{X}_G$	6.77	4.71
S_O	9.26	5.60
S_G	5.90	5.16
CV_O	1.24	1.44
CV_G	1.16	1.10
A_O	1.98	3.17
A_G	2.14	1.70
H_O	0.60	0.63
H_G	0.71	0.80

* SUBSCRITO: O-OBSERVADO; G-GERADO. $\overline{X}$-MÉDIA; S-DESVIO PADRÃO; CV-COEF. DE VARIAÇÃO; A-COEF. DE ASSIMETRIA; h-COEF. HURST.

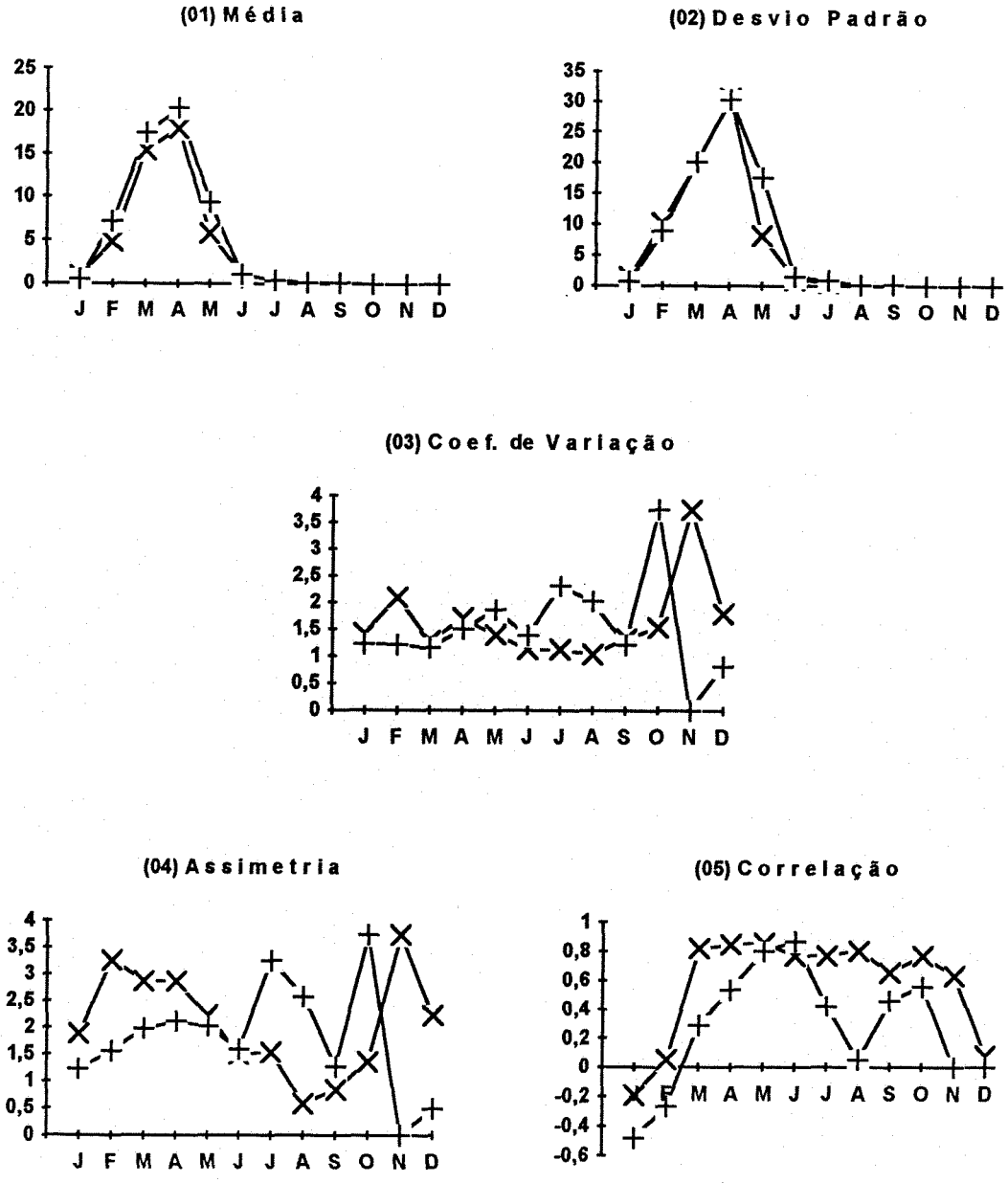


Figura 4.15 - Estatísticas históricas e geradas. Modelo Lognormal univariado. Estação: 36125000. x - observado + - gerado
(o_1) Média; (o_2) Desvio padrão; (o_3) Coeficiente de variação; (o_4) Assimetria, e (o_5) Correlação.

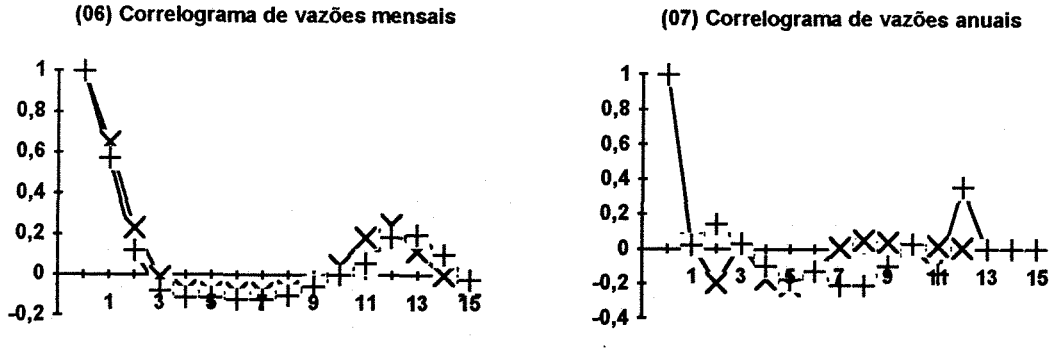


Figura 4.15 (Cont.) - Estatísticas históricas e geradas. Modelo Lognormal univariado. Estação: 36125000. x - observado + - gerado
(o_6) Correlograma de vazões mensais, e (o_7) Correlograma de vazões anuais.

Tabela 4.6 - Ocorrência de vazões mensais.
Modelo Lognormal.

MÊS	35210000			36125000		
	P	$\hat{P}$	P_G	P	$\hat{P}$	P_G
01	35.0	27.0	20.0	85.7	75.8	78.6
02	65.0	78.5	90.0	85.7	96.1	100.0
03	100.0	96.0	95.0	100.0	98.8	100.0
04	100.0	98.0	100.0	100.0	98.5	100.0
05	100.0	96.6	100.0	100.0	95.4	92.9
06	80.0	88.4	90.0	78.6	86.3	78.6
07	55.0	62.5	55.0	78.6	74.8	50.0
08	40.0	29.0	35.0	64.3	63.9	42.9
09	10.0	10.6	35.0	42.9	46.4	5701
10	0.0	4.3	0.0	35.7	26.3	7.1
11	0.0	3.1	0.0	7.1	19.1	0.0
12	10.0	5.9	5.0	35.7	33.0	78.6

* P - *Probabilidade observada de ocorrência*
$\hat{P}$ - *Probabilidade esperada (modelo binomial)*
P_G - *Probabilidade de ocorrência média gerada*

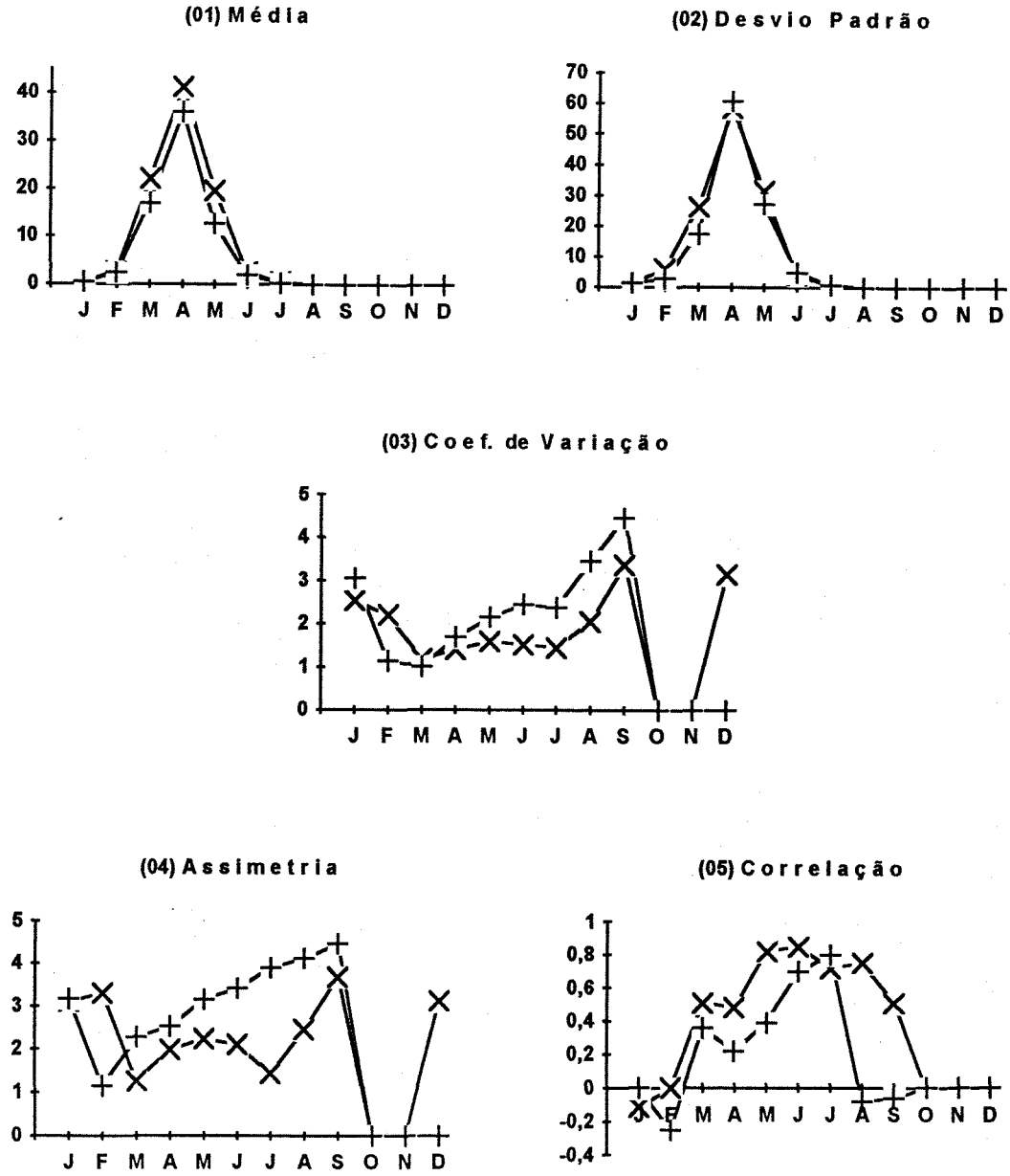


Figura 4.16 - Estatísticas históricas e geradas. Modelo Gama univariado. Estação: 35210000. x - observado + - gerado
(o_1) Média; (o_2) Desvio padrão; (o_3) Coeficiente de variação; (o_4) Assimetria, e (o_5) Correlação.

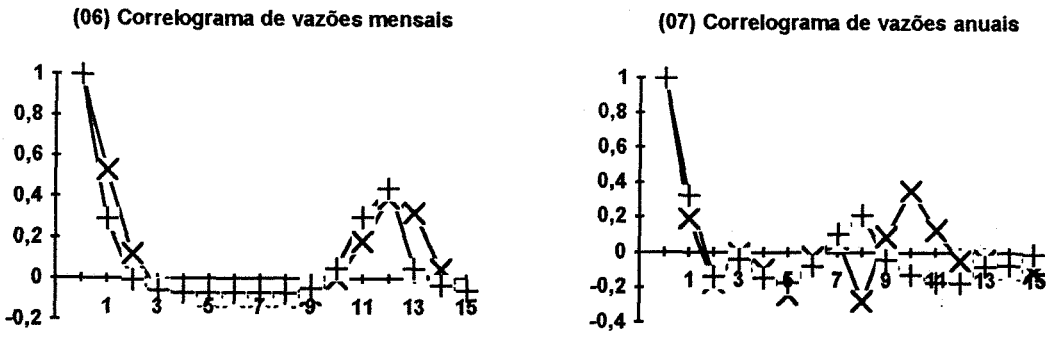


Figura 4.16 (Cont.) - Estatísticas históricas e geradas. Modelo Gama univariado. Estação: 35210000. x - observado + - gerado
($_{06}$) Correlograma de vazões mensais, e ($_{07}$) Correlograma de vazões anuais.

Tabela 4.7 - Estatísticas anuais observadas e geradas. Modelo Gama.

ESTAT.	35210000	36125000
$\overline{X}_o$	7.45	3.89
$\overline{X}_G$	7.64	3.87
S_o	9.26	5.60
S_G	6.92	4.48
CV_o	1.24	1.44
CV_G	0.91	1.06
A_o	1.98	3.17
A_G	1.64	1.82
H_o	0.60	0.63
H_G	0.69	0.64

* SUBSCRITO: O-OBSERVADO; G-GERADO. $\overline{X}$-MÉDIA; S-DESVIO PADRÃO; CV-COEF. DE VARIAÇÃO; A-COEF. DE ASSIMETRIA; h-COEF. HURST.

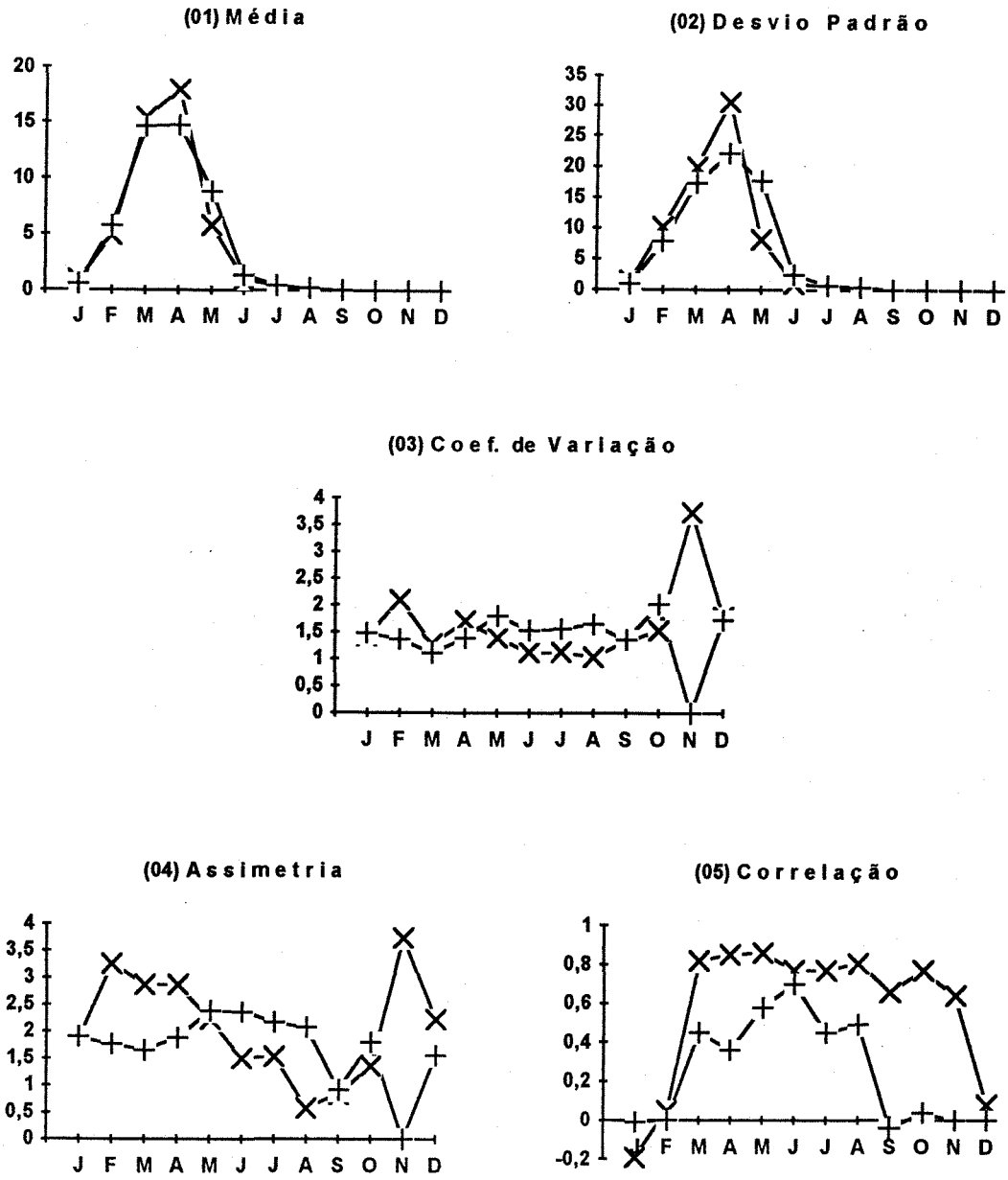


Figura 4.17 - Estatísticas históricas e geradas. Modelo Gama univariado. Estação: 36125000. x - observado + - gerado
(o_1) Média; (o_2) Desvio padrão; (o_3) Coeficiente de variação; (o_4) Assimetria, e (o_5) Correlação.

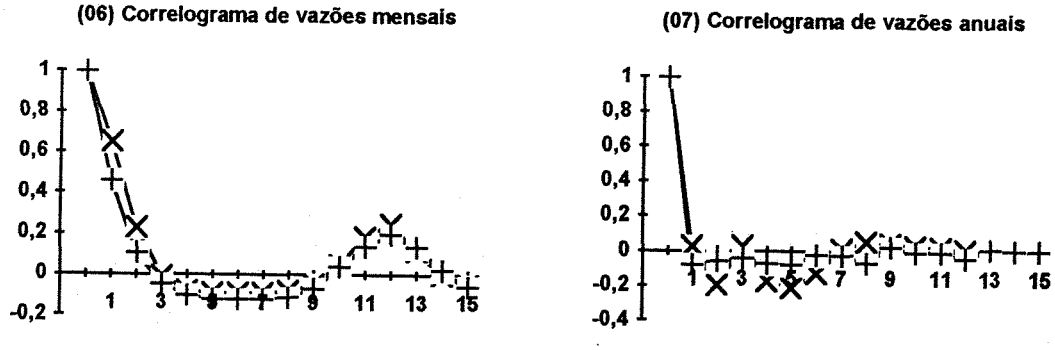


Figura 4.17 (Cont.) - Estatísticas históricas e geradas. Modelo Gama univariado. Estação: 36125000. x - observado + - gerado
(o6) Correlograma de vazões mensais; (o7) Correlograma de vazões anuais.

Tabela 4.8 - Ocorrência de vazões mensais.
Modelo Gama.

MÊS	35210000			36125000		
	P	$\hat{P}$	P_G	P	$\hat{P}$	P_G
01	35.0	27.0	20.0	85.7	75.8	75.7
02	65.0	78.5	70.0	85.7	96.1	95.3
03	100.0	96.0	100.0	100.0	98.8	98.0
04	100.0	98.0	95.0	100.0	98.5	97.7
05	100.0	96.6	95.0	100.0	95.4	95.3
06	80.0	88.4	70.0	78.6	86.3	85.6
07	55.0	62.5	75.0	78.6	74.8	76.6
08	40.0	29.0	25.0	64.3	63.9	62.7
09	10.0	10.6	5.0	42.9	46.4	44.1
10	0.0	4.3	0.0	35.7	26.3	23.6
11	0.0	3.1	0.0	7.1	19.1	0.0
12	10.0	5.9	0.0	35.7	33.0	35.7

* P - *Probabilidade observada de ocorrência*

$\hat{P}$ - *Probabilidade esperada (modelo binomial)*

P_G - *Probabilidade de ocorrência média gerada*

IV.3.5 - Escolha entre o modelo Lognormal e o Gama

O procedimento Monte Carlo gráfico, devido a ATKINSON (1982,1987), foi aplicado às estações do Rio Grande do Sul para verificar qual o modelo é mais apropriado para descrever os dados. O teste indicou que o modelo Lognormal como mais apropriado, embora para algumas estações no mês de fevereiro o melhor modelo tenha sido o Gama, como mostra a figura 4.18.

O método é eficaz na escolha do modelo, embora seja necessária a geração de 100 amostras para cada um dos modelos a cada mês, o que implica na geração e ajuste de 2400 amostras para cada estação. Assim, como o método implica em um grande esforço computacional, possivelmente é mais prático analisar as estatísticas médias das 50 amostras geradas, comparando-as com as estatísticas observadas.

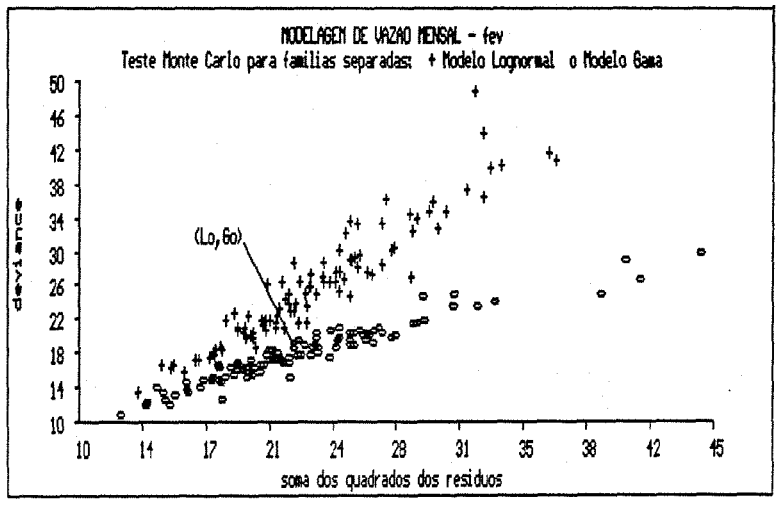


Figura 4.18 - Procedimento gráfico de Monte Carlo
Estação: 30700000. Mês de fevereiro.

CAPÍTULO V CONCLUSÕES E RECOMENDAÇÕES

O uso de MLGs representou uma grande ferramenta para simulação hidrológica, incluindo modelos que se apresentaram muito úteis na geração de vazões mensais com o benefício adicional de preservar satisfatoriamente as principais estatísticas anuais. Além disto, os MLGs mostraram-se também úteis na tentativa de reproduzir mais fielmente o processo estocástico natural pela consideração da variabilidade não só das vazões mensais, mas também de sua ocorrência.

Mais especificamente, no que diz respeito aos objetivos apresentados no capítulo I, pode-se concluir que:

(i) A formulação do modelo Thomas-Fiering como um MLG, seja este Lognormal ou Gama, consegue adequadamente acompanhar as flutuações das estatísticas mensais, sendo que o primeiro apresenta, em média, resultados melhores. Além das estatísticas mensais, as anuais também são adequadamente preservadas.

Uma vez que no caso univariado são preservadas apropriadamente as estatísticas anuais, não existe necessidade de recorrer a geração de vazão mensal a partir de modelos de geração de vazão anual, já que, em geral, um melhor ajuste anual faz-se às custas de prejuízos na descrição das estatísticas mensais.

Com os modelos log-Gama (3.47) e λ-*Normal* (3.49) não foi possível empregar o método de geração indicado em III.4.2 devido a explosão numérica, ou melhor, *overflow* na escala natural de medida. Provavelmente com outro método de geração seja possível obter resultados compatíveis com os já apresentados no capítulo anterior;

(ii) Com relação à utilização de técnicas diagnósticas, deve-se ter em mente que a deleção seqüencial de simples casos é realizada supondo que a mesma não indica um novo modelo, mesmo que o conjunto de observações deletadas tenham um alto *leverage*.

Os resíduos r_G, definidos em 3.32, foram muito informativos na identificação de *outliers* quando utilizados nas plotagens contra os valores ajustados e *half-Normal*, enquanto que a plotagem contra o preditor linear foi pouco informativa. Apesar da não normalidade deste resíduo, a sua plotagem *half-Normal* conjuntamente com os *envelopes* simulados, consegue, quase sempre, identificar as observações *outlying*.

A aproximação da estatística Cook, definida em (3.33), trabalha muito bem, embora seja subestimada para casos com alta *leverage*. As plotagens indexadas, *half-Normal*, contra os valores ajustados e contra as covariáveis do modelo são muito informativas, sendo que nas duas últimas as informações obtidas não diferiram marcadamente.

Quanto aos *envelopes* simulados utilizados ao lado de plotagens *half-Normal* e *Normal*, conforme capítulo III, são muito úteis na verificação da natureza das irregularidades destas plotagens. Assim como verificado por ATKINSON (1987), observou-se, como esperado, que o método leva a envelopes diferentes quando não utilizados com o mesmo gerador de números aleatórios, principalmente quando este é baseado no *clock* do sistema, podendo levar a conclusões diferentes. No caso de utilizar este tipo de gerador recomenda-se realizar testes de independência e uniformidade (teste qui-quadrado, Kolmogorov-Smirnov, serial, de *runs*, correlação, etc...) para avaliar a qualidade do gerador.

Esta etapa deve ser encarada como necessária para evitar que observações de algum modo estranhas influenciem significativamente as estimativas dos parâmetros, e conseqüentemente a geração;

(iii) Na modelagem multivariada foram obtidos bons resultados com a preservação das flutuações das principais estatísticas mensais, assim como é mantida adequadamente a correlação cruzada de vazões mensais e anuais, além de preservar, com um certo grau de aproximação, as principais estatísticas anuais. Além disto, a maior vantagem do método é sua simplicidade;

(iv) A modelagem de vazões mensais em rios intermitentes com a modificação do método de CLARKE (1973) adaptado por FILHO (1978), aqui proposta, conseguiu descrever adequadamente tanto a ocorrência de vazão mensal como as quantidades, considerando a variabilidade de ambas. A análise dos resultados nos leva a concluir que o modelo Gama é mais adequado que o modelo Lognormal para as duas estações estudadas.

O método é muito sensível aos parâmetros estimados para a distribuição de probabilidades adotada (ponto (3) do algoritmo da figura 3.6), devendo-se assegurar uma boa estimativa dos mesmos. De maneira geral, recomenda-se o uso do método da máxima verossimilhança para estimar estes parâmetros, embora nem sempre seja garantida uma boa estimativa, em especial, quando o número de vazões ocorridas em determinado mês é bem pequeno;

(v) A modelagem da variância de vazões mensais, juntamente com suas médias, propiciou excelentes resultados na modelagem univariada de vazões mensais, seguindo as flutuações das principais estatísticas mensais e preservando adequadamente as anuais. No caso bivariado as flutuações mensais das principais estatísticas são descritas satisfatoriamente, assim como as estatísticas anuais e a correlação cruzada entre vazões mensais e anuais.

Possivelmente outras combinações de modelos para média e dispersão sejam mais adequadas, devendo-se buscar outro algoritmo de ajuste, provavelmente na direção apontada por SMYTH (1989).

Quanto a característica relacionada com o armazenamento, coeficiente Hurst, o mesmo foi adequadamente preservado tanto para os rios perenes como para os intermitentes (modelo Gama); havendo no primeiro caso uma tendência a subestimá-lo e no último a superestimá-lo.

O procedimento gráfico de Monte Carlo, apesar de não ser objetivo central do trabalho, mostrou-se eficaz na escolha do modelo, embora o fazendo de maneira ineficiente por demandar grande esforço computacional, como já mencionado em IV.3.5.

CAPÍTULO VI REFERÊNCIAS BIBLIOGRÁFICAS

1 AITKIN, M. 1987. Modelling variance heterogeneity in normal regression using GLIM. Journal of the Royal Statistical Society. Série C: Applied Statistics, London, v. 36, n. 3, p. 332-339.

2 AITKIN, M. et al. 1989. Statistical modelling in GLIM. Oxford: Oxford University Press. 374 p.

3 ATKINSON, A. C. 1981. Two graphical displays for outlying and influential observations in regression. Biometrika, London, v. 68, n. 1, p. 13-20.

4 ATKINSON, A. C. 1982. Regression diagnostics, transformations and constructed variables (with discussion). Journal of the Royal Statistical Society. Série B: Methodological, London, v.44, n. 1, p. 1-36.

5 ATKINSON, A. C. 1987. Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis. Oxford: Oxford University, 282 p.

6 ATKINSON, A. C., PEARCE, M. C. 1976. The computer generation of beta, gamma and normal random variables. Journal of the Royal Statistical Society. Série A: General, London, v. 139, p. 431-460.

7 BEALE, E. M. L., LITTLE, R. J. A. 1975. Missing values in multivariate analysis. Journal of the Royal Statistical Society. Série B: Methodological, London, v. 37, p. 129-145.

8 BOBÉE, B., ASHKAR, F. 1991. The gamma family and derived distributions applied in hydrology. Littleton: Water Resources publications. 203 p.

9 BOX, G. E. P., COX, D. R. 1964. An analysis of transformations. Journal of the Royal Statistical Society. Série B: Methodological, London, v. 26, n. 3, p. 211-252.

10 BRAS, R. L., RODRIGUEZ-ITURBE, I. 1985. Random functions and hydrology. Massachusetts: Addison-Wesley publishing company. 559 p.

11 CHENG, R. C. H. 1977. The generation of gamma variables with non-integral shape parameter. Journal of the Royal Statistical Society. Série C: Applied Statistics, London, v. 26, n. 1, p. 71-75.

12 CHENG, R. C. H., FEAST, G. M. 1979. Some simple gamma variate generators. Journal of the Royal Statistical Society. Série C: Applied Statistics, London, v. 28, n. 3, p. 290-295.

13 CLARKE, R. T. 1973. Mathematical models in hydrology, Irrigation and Drainage Paper No. 19, Food and Agricultural Organization of UN, Rome.

14 COOK, R. D. 1977. Detection of influential observation in linear regression. Technometrics, Washington, v. 19, n. 1, p. 15-18, Feb.

15 COOK, R. D. 1979. Influential observations in linear regression. Journal of the American Statistical Association, Washington, v. 74, n. 365, p. 169-174, Mar.

16 COOK, R. D., WEISBERG, S. 1983. Diagnostics for heterosedasticity in regression. Biometrika, London, v. 70, n. 1, p. 1-10.

17 COOK, R. D., WEISBERG, S. 1986. Residuals and influence in regression. New York: Chapman and Hall, 230 p.

18 CORDEIRO, G.M. 1986. Modelos Lineares Generalizados. Campinas:Associação Brasileira de Estatística. 286 p.

19 COX, D. R., SNELL, E. J. 1968. A general definition of residuals (with discussion). Journal of the Royal Statistical Society. Série B: Methodological, London, v. 30, p. 248-275.

20 DEMPSTER, A.P., LAIRD, N.M., RUBIN, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society. Série B: Methodological, London, v. 39, p. 1-38.

21 DNAEE. 1987. Inventário de estações fluviométricas. Brasília: Dnaee.

22 DOBSON, A.J. 1990. An Introduction to Generalized Linear Models. London: Chapman and Hall, 174 p.

23 EFRON, B. 1982. Transformation theory: how normal is a family of distributions ? The Annals of Statistics, Hayward, v. 10, n. 2, p. 323-339.

24 FIERING, M. B. 1967. Streamflow Synthesis. Cambridge: Harvard University Press. 139 p.

25 FILHO, A. A. do C. 1978. Um modelo de simulação e operação de um sistema de irrigação com reservatórios em rios intermitentes. Tese M.Sc., COPPE/UFRJ, Brasil.

26 FISHMAN, G. S. 1973. Concepts and methods in discrete event digital simulation. New York: Wyley-Interscience. 380 p.

27 GENSTAT 5: procedure library manual release 1.3 [2] 1989. Oxford: NAG. 175 p.

28 GILROY, E. J. 1970. Reliability of a variance estimate obtained from a sample augmented by multivariate regression. Water Resources Research, Washington, v. 6, n. 6, p. 1595-1600.

29 GLIM 3.77. Reference guide. In GLIM 3.77 Reference manual. 1985. Oxford: NAG.

30 GOLDER, E. R., SETTLE, J. G. 1976. The Box-Müller method for generating pseudo-random normal deviates. Journal of the Royal Statistical Society. Série C: Applied Statistics, London, v. 25, n. 1, p. 12-20.

31 HAAN, C. T. 1982. Statistical methods in Hydrology. Ames: Iowa State University Press. 378 p.

32 HALD, A. 1970. Statistical tables and formulas. New York: John Wiley & Sons. 97 p.

33 HARMS, A. A., CAMPBELL, T. H. 1967. An extension to the Thomas-Fiering model for the sequential generation of streamflow. Water Resources Research, Washington, v. 3, n. 3, p. 653-661.

34 HAWKINS, D. M., BRADU, D., KASS, G. V. 1984. Location of several outliers in multiple-regression data using elemental sets. Technometrics, Washington, v. 26, n. 3, p. 197-208, Aug.

35 HINKLEY, D. V., RUNGER, G. 1984. The analysis of transformed data. Journal of the American Statistical Association, Washington, v. 79, n. 386, p. 302-320, Jun.

36 HURST, H. E. 1951. Long-term storage capacity of reservoirs. Transactions of American Society of Civil Engineers, New York, v. 116, p. 770-808.

37 JOHN, J. A., DRAPER, N. R. 1980. An alternative family of transformations. Journal of the Royal Statistical Society. Série C: Applied Statistics, London, v. 29, n. 2, p. 190-197, 1980.

38 JOHNSON, R. A., WICHERN, D. W. 1992. Applied multivariate statistical analysis. London: Prentice-Hall. 642 p.

39 JØRGENSEN, B. 1984. The delta algorithm and GLIM. International Statistical Review, Edinburgh, v. 52, n. 3, p. 283-300.

40 JØRGENSEN, B. 1989. Generalized linear models and extensions. Rio de Janeiro: IMPA. 29 p. (Informes de matemática, B-053).

41 KLEMEŠ, V. 1974. The Hurst phenomenon: A puzzle ? Water Resources Research, Washington, v. 10, n. 4, p. 675-688.

42 LANNA, A. E. 1971. Geração de escoamentos mensais através de modelo estocástico que utiliza precipitações. Tese M.Sc., IPH/UFRGS, Brasil.

43 LINSLEY, R. K., KOHLER, M. A., PAULHUS, J. L. H. 1988 Hydrology for Engineers. New York: McGraw Hill. 492 p.

44 MAK, T. K. 1992. Estimation of Parameters in Heteroscedastic Linear Models. Journal of the Royal Statistical Society. Série B: Methodological, London, v. 54, n. 2, p. 649-655.

45 MANDELBROT, B. B. 1972. Broken line process derived as an approximation to fractional noise. Water Resources Research, Washington, v. 8, n. 5, p. 1354-1356.

46 MANDELBROT, B. B., WALLIS, J. R. 1968. Noah, Joseph, and operational hydrology. Water Resources Research, Washington, v. 4, n. 5, p. 909-918.

47 MANDELBROT, B. B., WALLIS, J. R. 1969a. Computer experiments with fractional gaussian noises, 1, Averages and variances; 2, Rescaled range and spectra; 3, Mathematical appendix. Water Resources Research, Washington, v. 5, n. 1, p. 228-267.

48 MANDELBROT, B. B., WALLIS, J. R. 1969b. Some long-run properties of geophysical records. Water Resources Research, Washington, v. 5, n. 2, p. 321-340.

49 MANDELBROT, B. B., WALLIS, J. R. 1969c. The robustness of the rescaled range R/S in the measurement of non-cyclic long-run statistical dependence. Water Resources Research, Washington, v. 5, n. 5, p. 967-988.

50 MATALAS, N. C. 1967. Mathematical assessment of synthetic hydrology. Water Resources Research, Washington, v. 3, n. 4, p. 937-945.

51 MATALAS, N. C., HUZZEN, C. S. 1967. A property of the range of partial sums. in: Proceedings of the International Hydrology Symposium, Colorado, v. 1, p. 252-257.

52 MATALAS, N. C., JACOBS, B. 1964. A correlation procedure for augmenting hydrology data. U.S. Geological Survey Professional paper, 434-E:E1-E7.

53 MEJIA, J.M, ROUSSELLE, J. 1976. Disaggregation models in hydrology revisited. Water Resources Research, Washington, v. 12, n. 2, p. 185-186.

54 McCULLAGH, P., NELDER, J.A. 1989. Generalized linear models. London: Chapman and Hall. 261 p.

55 NELDER, J.A., LEE, Y. 1992. Likelihood, quasi-likelihood and pseudo-likelihood: some comparisons. Journal of the Royal Statistical Society. Série B: Methodological, London, v. 54, n. 1, p. 273-284.

56 NELDER, J.A., WEDDERBURN, R.W.M. 1972. Generalized linear models (with discussion). Journal of the Royal Statistical Society. Série A: General, London, v. 135, n. 3, p. 370-384.

57 O'CONNELL, P. E. 1971. A simple stochastic modeling of Hurst's law. in: Mathematical models in hydrology, Warsaw, v. 1, 169-187.

58 PEGRAM, G. G. S., JAMES, W. 1972. Multilag multivariate autoregressive model for the generation of operational hydrology. Water Resources Research, Washington, v. 8, n. 4, p. 1074-1076.

59 PIERCE, D. A., SCHAFER, D. W. 1986. Residual in generalized linear models. Journal of the American Statistical Association, Washington, v. 81, n. 396, p. 977-986, Dec.

60 PREGIBON, D. 1980. Goodness of link tests for generalized linear models. Journal of the Royal Statistical Society. Série C: Applied Statistics, London, v. 29, n. 1, p. 15-24.

61 PREGIBON, D. 1981. Logistic regression diagnostics. The Annals of Statistics, Hayward, v. 9, n. 4, p. 705-724, 1981.

62 RAUDKIVI, A. J. 1979. Hydrology. An advanced introduction to hydrological processes and modelling. Oxford: Pergamon Press. 479 p.

63 RODRIGUEZ-ITURBE, I., MEJIA, J. M., DAWDY, D. R. 1972. Streamflow simulation, 1, A new look at Markovian models, fractional Gaussian noise and crossing theory; 2, The broken line process as a potential model for hydrologic simulation. Water Resources Research, Washington, v. 8, n. 4, p. 921-941.

64 SALAS, J. D., BOES, D. C. 1974. Expected range and adjusted range of hydrologic sequences. Water Resources Research, Washington, v. 10, n. 3, p. 457-463.

65 SALAS, J. D., BOES, D. C., YEVJEVICH, V., PEGRAM, G. G., S. 1979. Hurst phenomenon as a pre-asymptotic behavior. Journal of Hydrology, Amsterdã, v. 44, n. 1, p. 1-15.

66 SARMENTO, F. J. 1989. Aplicabilidade de modelos de geração de vazão no semi-árido do Nordeste do Brasil. Tese M.Sc., UFC, Brasil.

67 SCALLAN, A., GILCHRIST, R., GREEN, M. 1984. Fitting parametric link functions in generalized linear models. Computational Statistics & Data Analysis, Amsterdan, v. 2, p. 37-49.

68 SMYTH, G. K. 1989. Generalized linear models with varying dispersion. Journal of the Royal Statistical Society. Série B: Methodological, London, v. 51, n. 1, p. 47-60.

69 SRIKANTHAN, R., McMAHON, T. A. 1980 Sthocastic generation of monthly flows for ephemeral streams. Journal of Hydrology, Amsterdã, v. 47, p. 19-40.

70 STEDINGER, J. R., VOGEL, R. M. 1984. Disaggregation procedures for generating serially correlated flow vectors. Water Resources Research, Washington, v. 20, n. 1, p. 47-56.

71 VALENCIA, D., SCHAAKE, J. C. 1973. Disaggregation processes in stochastic hydrology. Water Resources Research, Washington, v. 9, n. 3, p. 580-585.

72 WALLIS, J. R., MATALAS, N. C. 1970. Small properties of H & K - Estimators of the Hurst coeficient h. Water Resources Research, Washington, v. 6, n. 6, p. 1583-1594.

73 WALLIS, J. R., MATALAS, N. C. 1971a. In hydrology h is a household word. in: Mathematical models in hydrology, Warsaw, v. 1, 196-203.

74 WALLIS, J. R., MATALAS, N. C. 1971b. Correlogram analysis revisited. Water Resources Research, Washington, v. 7, n. 6, p. 1448-1459.

75 WEDDERBURN, R. W. M. 1974b. Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. Biometrika, London, v. 61, n. 3, p. 439-447.

76 WILLIAMS, D. A. 1987. Generalized linear model diagnostics using the deviance and single case deletions. Journal of the Royal Statistical Society. Série C: Applied Statis-tics, London, v. 36, n. 2, p. 181-191.