DESENVOLVIMENTO DE MÉTRICAS PARA DIAGNÓSTICO DE REGRESSÃO HIDROLÓGICA BAYESIANA COM MODELO GLS

Andrea M. Gruber[1], Dirceu S. Reis Jr.[2], Jery R. Stedinger[3]

RESUMO --- Reis et al. (2005) desenvolveram um procedimento Bayesiano para análise de modelos de regressão (B-GLS) para regionalização de variáveis hidrológicas, utilizando o modelo de Mínimos Quadrados Generalizados (GLS) [Stedinger e Tasker, 1985]. Este artigo apresenta os avanços que vêm sendo realizados para tornar a análise de regressão B-GLS em uma metodologia operacional de regionalização de variáveis hidrológicas. O artigo descreve o procedimento B-GLS e diversas estatísticas de diagnóstico para serem utilizadas em regionalização hidrológica. A regionalização do coeficiente de assimetria para bacia do Rio Illinois (EUA) ilustra a utilidade destas métricas de diagnóstico, que incluem Variância Média de Predição para avaliar a precisão da predição do modelo; pseudo-R^2, Valor de Plausibilidade Bayesiano, Razão entre Variâncias dos Erros (EVR), e Grau de Inadequação da Estimativa da Variância de Beta (MBV) para avaliar quão bem o modelo descreve os dados; e *Leverage* (alavanca), Influência e Influência-σ para identificar e avaliar os impactos dos erros dos dados na regressão. O procedimento B-GLS em conjunto com as estatísticas de diagnóstico apresentadas neste artigo estão sendo empregadas num estudo maior, em andamento, para obtenção de um estimador regional do coeficiente de assimetria de todo o sudeste americano, que inclui dados de mais de 800 estações fluviométricas.

ABSTRACT --- Reis et al. (2005) developed a Bayesian approach to analysis of a Generalized Least Squares (GLS) regression model for regional analyses of hydrologic data. Advances have been made over the last year to develop that Bayesian procedure into an operational Bayesian GLS regional hydrologic regression procedure. This paper describes a Bayesian Generalized Least Squares (B-GLS) framework together with diagnostic statistics introduced by Reis et al. (2004), Reis (2005) and Griffis and Stedinger (2006) that can be used to develop regional hydrologic relationships. An example using data from the Illinois River Basin illustrates useful diagnostic statistics including pseudo R^2, Bayesian plausibility, Error Variance Ratio (EVR), Misrepresentation of Beta Variance (MBV), Leverage, Influence and σ-Influence. The B-GLS framework and diagnostic statistics developed in this analysis are being applied to an ongoing study in the southeast United States which will produce a regional skew estimator.

Palavras-chave: Regionalização, análise Bayesiana, diagnóstico de regressão.

[1] Doutoranda, School of Civil and Environmental Engineering, Cornell University – amg66@cornell.edu
[2] Pesquisador, Fundação Cearense de Meteorologia e Recursos Hídricos – FUNCEME – dirceu.reis@gmail.com
[3] Professor, School of Civil and Environmental Engineering, Cornell University – jrs5@cornell.edu

1 - INTRODUÇÃO

Regionalização hidrológica baseada em modelos de regressão é uma das maneiras que os hidrólogos têm para estimar variáveis hidrológicas em locais onde não há disponibilidade de dados, ou para aumentar a precisão destes estimadores onde a série histórica não é longa o suficiente para fornecer estimativas com a precisão necessária (Benson e Matalas, 1967; Matalas e Gilroy, 1968; Thomas e Benson, 1970; Shane e Gaver, 1970; Moss e Karlinger, 1974; Vicens et al., 1975; IACWD, 1982; Kuczera, 1982; Stedinger, 1983; Jennings et al., 1994; Madsen e Rosbjerg, 1997; Fill e Stedinger, 1998; Martins e Stedinger, 2000; Walker e Krug, 2003; Shu e Burn, 2004; Reis et al., 2005; Reis, 2005).

Os modelos estatísticos empregados em estudos de regressão hidrológica vêm sendo aperfeiçoados ao longo do tempo. O popular procedimento dos mínimos quadrados (OLS) foi substituído, pois a premissas do modelo, independência dos dados e erros com mesma variância, muitas vezes não satisfaziam os dados hidrológicos. A premissa de mesma variância dos erros pôde ser removida com o uso do procedimento dos mínimos quadrados ponderados (WLS) [Tasker, 1980], cujo modelo estatístico leva em consideração que as variâncias amostrais dos estimadores locais da variável de interesse são diferentes, já que o tamanho da série histórica varia de estação para estação. Mais tarde, Stedinger e Tasker (1985) desenvolveram o procedimento mínimos quadrados generalizados (GLS), cujo modelo estatístico leva também em consideração a possível correlação espacial existente entre os estimadores da variável de interesse, tornando-se um modelo mais adequado para regionalização hidrológica. O procedimento GLS vem sendo empregado em várias regiões do mundo para regionalização de variáveis hidrológicas como vazões máximas, vazões mínimas, precipitação intensa, concentração de contaminantes e parâmetros de distribuições de probabilidades.

Mais recentemente, Reis et al. (2005) desenvolveram uma abordagem Bayesiana para análise do modelo de regressão GLS para regionalização de variáveis hidrológicas. Esta abordagem Bayesiana supre algumas deficiências do procedimento GLS clássico, como por exemplo, não fornecer uma estimativa da precisão da variância do erro do modelo, e ainda fornece uma descrição mais completa e realista dos possíveis valores da variância do erro do modelo, especialmente quando a variância dos erros amostrais são maiores do que a variância do erro do modelo. Esta é uma característica importante, em especial quando se pretende aumentar a precisão da estimativa da variável hidrológica de interesse através da combinação das estimativas regional e local. Uma estimativa equivocada da variância do erro do modelo regional pode subestimar/superestimar a precisão do estimador regional, resultando numa combinação que dará, equivocadamente, menos/mais peso ao estimador regional.

Este artigo contribui para o aperfeiçoamento da abordagem Bayesiana, introduzida por Reis et al. (2005), para regionalização de variáveis hidrológicas com base no modelo de Mínimos Quadrados Generalizados (B-GLS). Este método operacional de regressão hidrológica pode ser utilizado tanto para a estimativa regional de parâmetros de forma de distribuições de probabilidade, quanto para a estimativa de quantis de cheias.

Este artigo foca na implementação da abordagem B-GLS e na discussão de uma série de estatísticas utilizadas para diagnóstico de modelos de regressão B-GLS, apresentadas em Reis et al. (2004), Reis (2005), e Griffis e Stedinger (2006), de forma a construir uma estrutura metodológica completa para análise de regressão hidrológica. As novas estatísticas apresentadas neste artigo para diagnóstico de regressão incluem pseudo-R^2 ajustado (R^2_{GLS}), Valor de Plausibilidade Bayesiano (ψ) e Influência-σ. Estas novas estatísticas de diagnóstico, em conjunto com a variância média de predição (VMP), razão de variância do erro do modelo (EVR), grau de inadequação da estimativa da variância dos parâmetros β do modelo (MBV), pseudo Tabela ANOVA, *Leverage* e Influência, permitem que se faça um exame mais extenso e completo dos modelos de regressão baseados na abordagem B-GLS.

2 – ANÁLISE DE REGRESSÃO DE VARIÁVEIS HIDROLÓGICAS

A abordagem de regressão denominada B-GLS, desenvolvida por Reis et al. (2005), pode ser utilizada com registros fluviométricos para derivar relações empíricas entre características hidrológicas de uma dada estação, como por exemplo, o coeficiente de assimetria no espaço logarítmico, necessário para o ajuste da distribuição log-Pearson Tipo III, e características climáticas e fisiográficas da bacia contribuinte a esta mesma estação.

O modelo GLS utilizado na análise Bayesiana foi introduzido por Stedinger e Tasker (1985, 1986). Este modelo assume que o erro total de regressão é resultado da soma de dois termos, o erro do modelo propriamente dito $\delta \sim N(0, \delta^2)$, e os erros amostrais η, dado que o verdadeiro valor da variável de interesse y_i é desconhecido, e apenas uma estimativa está disponível. Em notação matricial,

$$\hat{y} = X\beta + \eta + \delta = X\beta + \varepsilon \qquad (1)$$

onde $\hat{y}$ contém as estimativas da variável de interesse, X é a matriz que contém 1 na primeira coluna e valores das *k* variáveis explanatórias (características da bacia de contribuição) nas colunas restantes, e $\boldsymbol{\eta}$ contém os erros amostrais. O erro total $\boldsymbol{\varepsilon}$ possui valor esperado zero e matriz de covariância igual a

$$E[\varepsilon\varepsilon^T] = \Lambda(\sigma_\delta^2) = \sigma_\delta^2 I + \Sigma \quad (2)$$

onde Σ é a matriz de covariância dos erros amostrais.

Os modelos OLS e WLS, mencionados anteriormente, são de fato casos especiais do modelo GLS descrito acima. Quando os elementos fora da diagonal da matriz Σ forem iguais a zero, o que significa que não há correlação entre os estimadores da variável de interesse, o modelo GLS se transforma no modelo WLS. Se, além disso, os elementos da diagonal de Σ também forem iguais a zero, o modelo GLS reduz-se ao modelo OLS.

A abordagem Bayesiana para análise de regressão hidrológica, desenvolvida por Reis et al. (2005), requer a especificação de uma distribuição a *priori* para os parâmetros β do modelo de regressão e para a variância do erro do modelo σ_δ^2.

Uma distribuição Normal multivariada, com valor esperado zero e matriz de covariância com valores altos na diagonal, é utilizada como distribuição a priori não-informativa para os parâmetros β do modelo de regressão. Valores altos de variância fazem com que esta distribuição a *priori* seja relativamente suave na faixa de interesse dos parâmetros β.

Uma distribuição exponencial com parâmetro λ é utilizada para modelar a informação a *priori* que se tem sobre a variância do erro do modelo σ_δ^2 [ver Reis et al. (2005) para uma discussão mais aprofundada acerca desta priori] o parâmetro λ é igual ao inverso do valor esperado da priori de σ_δ^2. No caso da regionalização do coeficiente de assimetria das vazões máximas anuais nos espaço logarítmico, o exemplo de caso deste artigo, segue-se o raciocínio de Reis et al. (2005) adotando-se λ = 6, mas admitindo que à medida que se for ganhando experiência com a regionalização desta variável hidrológica, um valor maior para λ pode vir a ser justificado.

Definidas as distribuições a priori para β e σ_δ^2, Reis et al. (2005) mostraram como calcular os momentos a posteriori dos parâmetros β do modelo regional e a função completa de distribuição de probabilidades a posteriori da variância do erro do modelo. Fazendo isso, eles mostraram que a abordagem B-GLS fornece uma descrição mais realista dos possíveis valores da variância do erro do modelo, especialmente nos casos onde as variâncias dos erros amostrais são maiores do que a variância do erro do modelo, fato muito comum quando se trata de regionalização de parâmetros de forma de distribuições teóricas de probabilidade, como o coeficiente de assimetria no caso da log-Pearson Tipo III, e o parâmetro κ da distribuição GEV. A correta estimativa da variância do erro do modelo é imprescindível para uma adequada combinação entre a estimativa regional e a estimativa local da variável de interesse.

3 – SELEÇÃO DO MODELO REGIONAL

A seleção de um modelo de regressão regional consiste em definir qual o conjunto de variáveis independentes que deve ser empregado. Para isto, diversas estatísticas descritivas foram desenvolvidas para avaliar quão bem o modelo de regressão em pauta descreve os dados que se tem em mãos.

O objetivo do processo de seleção do modelo de regressão é definir o conjunto de possíveis variáveis explanatórias que melhor se adequa aos dados, fornecendo a mais precisa predição da variável de interesse em questão, mantendo o modelo o mais simples possível, ou seja, empregando o menor número possível de variáveis explanatórias.

As estatísticas mais tradicionais para a seleção de modelos de regressão incluem R^2, razão de verossimilhança (*likelihood ratio*), estatística de Mallow (Cp), Critério de Informação de Akaike (CIA), e Critério de Informação Bayesiana (CIB) [Linhart and Zucchini, 1986; Gelman et al., 2004]. Várias destas estatísticas penalizam o aumento de complexidade do modelo, resultante da inclusão de uma variável explanatória, de modo que seja necessária a ocorrência de um aumento substancial na capacidade de predição do modelo de regressão para que a inclusão de uma nova variável possa ser justificada. Na seqüência, são introduzidas algumas estatísticas descritivas desenvolvidas exclusivamente para a avaliação dos parâmetros do modelo de regressão B-GLS.

3.1 – Variância média de predição (VMP)

A Variância Média de Predição é uma métrica natural para ser utilizada na avaliação de modelos de regressão, já que a maior motivação em derivar um modelo deste tipo é fazer predições acerca da variável hidrológica de interesse tanto em locais com dados (série histórica curta) quanto em locais sem dados. Como a VMP leva em consideração não apenas a variância do erro do modelo, mas também a variância amostral dos parâmetros, quanto maior o número de parâmetros do modelo, maior será a penalidade imposta por esta métrica.

Como observado em Reis et al. (2005) [seguindo Tasker e Stedinger, 1986], a métrica VMP_{nova} assume implicitamente que as estações empregadas na análise de regressão são representativas dos locais onde haverá interesse no futuro de se conhecer a variável hidrológica de interesse já que os valores das variáveis explanatórias destes locais são utilizados para estimar a variância média de predição para uma nova estação, ou novo local, VMP_{nova}, definida pela seguinte expressão,

$$VMP_{nova} = E[\sigma_\delta^2] + \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_i Var[\beta \mid \hat{\mathrm{y}}]\mathrm{x}_i^T \qquad (4)$$

Se o interesse for predizer o valor da variável de interesse numa estação que já está sendo empregada na análise de regressão, que é o caso, por exemplo, quando se deseja obter uma estimativa regional de uma variável hidrológica numa estação com série histórica curta, a métrica a ser utilizada deveria ser a variância média de predição para uma estação antiga, denominada VMP_{antiga} (Reis et al., 2004), estimada por:

$$VMP_{antiga} = E[\sigma_\delta^2] + \frac{1}{n}\sum_{i=1}^{n}\left\{ x_i Var[\beta \mid \hat{y}] x_i^T - 2E\left[\sigma_\delta^2 x_i (X^T \Lambda^{-1} X)^{-1} X^T \Lambda^{-1} e_i\right]\right\} \quad (5)$$

onde o vetor-coluna e_i contém 1 na i-ésima linha e zero nas demais.

3.2 – Valor de Plausibilidade Bayesiano (*Bayesian plausibility value*)

O Valor de Plausibilidade Bayesiano, ψ, desenvolvido por Reis (2005), descreve se zero é um valor plausível para cada parâmetro β do modelo de regressão, dada as distribuições a priori, descritas na seção 2, e os dados que se tem em mãos.

Na abordagem Bayesiana, a função densidade de probabilidades (fdp) a posteriori dos parâmetros β é obtida. Como discutido em Lindley (1965) e Zellner (1971), considerando-se que tanto os dados quanto a fdp a posteriori de β estão disponíveis, pode-se construir uma região de credibilidade para os parâmetros de regressão. Essa região de credibilidade pode ser vista como um resumo do que se conhece, a posteriori, sobre os parâmetros, e pode, portanto, servir de base para a realização de um teste de hipóteses para concluir se zero está incluído nas regiões de 90 ou 95% de credibilidade. Isto permite que se realize, dentro de uma abordagem Bayesiana, o equivalente ao teste de hipóteses da estatística clássica, utilizando para isso as fdps a posteriori de cada parâmetro β.

Portanto, o nível de plausibilidade do parâmetro β pode ser definido como a menor probabilidade ψ de que zero esteja na região de credibilidade de 100(1- ψ) do parâmetro em questão. Este valor seria o equivalente ao p-valor utilizado na estatística clássica para descrever a significância estatística de uma estimativa, e para avaliar se o parâmetro deve ou não ser incluído no modelo, ao invés de partir do pressuposto de que o valor verdadeiro do parâmetro é zero, que é o que se faz usualmente nestes casos.

Um valor de plausibilidade de mais de 5% (ou 10% se uma região de credibilidade de 90% é empregada na análise), sugere que o modelo poderia ser melhorado se o parâmetro em questão fosse igualado a zero. O valor de plausibilidade é calculado da seguinte maneira,

$$\psi = 2 \underset{\sigma_\delta^2}{E}\left\{ \Phi\left[-\upsilon \frac{b(\sigma_\delta^2)}{\sigma_b(\sigma_\delta^2)} \right] \right\} \quad (6)$$

em que Φ é a função acumulada de probabilidade da Normal padrão, e a média condicionada $b(\sigma_\delta^2)$ e o erro-padrão $\sigma_b(\sigma_\delta^2)$ para o parâmetro β_i são ambos dependentes da variância do erro do modelo σ_δ^2; $\upsilon = \text{sign}[\mu_\beta] = 1$ for $\mu_\beta \geq 0$ and -1 for $\mu_\beta < 0$, sendo μ_β a média a posteriori do parâmetro β.

Na literatura da estatística Bayesiana é comum encontrar a estatística p-Valor Bayesiano. Portanto, é importante mostrar a diferença entre esta estatística e o valor de plausibilidade definido acima. O p-Valor Bayesiano, discutido em Bayarri & Berger (2000) e em Robins et al. (2000), corresponde à probabilidade de que uma outra amostra aleatória X iria gerar um valor mais extremo de uma dada estatística teste do que o valor que foi observado na presente amostra, de modo que o p-Valor Bayesiano é uma estatística mais próxima do p-Valor da estatística clássica. Estes autores, e outros, tentaram desenvolver um p-Valor Bayesiano que refletisse apenas os dados, e não a distribuição a priori. Por outro lado, o Valor de Plausibilidade Bayesiano reflete a visão Bayesiana de que a distribuição a priori também representa uma informação válida e importante sobre os parâmetros, e que, portanto, deve ser utilizada na decisão de incluir ou não uma determinada variável explanatória no modelo de regressão.

3.3 – Pseudo-R^2_{GLS} e pseudo tabela ANOVA

Uma das maneiras de avaliar o quanto o modelo regional explica a variabilidade espacial observada nos dados é através de uma análise de variância, usualmente baseada na baseada na partição do somatório dos quadrados dos resíduos, e correspondentes graus de liberdade. Tradicionalmente em modelos de regressão OLS, o somatório dos quadrados dos desvios em relação à média (SST) é dividido em duas partes, a soma dos quadrados dos erros explicados pelo modelo (SSR) e a soma residual dos quadrados dos erros (SSE), onde SST = SSR + SSE. A métrica R^2 é então utilizada para representar o quanto da variabilidade observada nos dados é explicada pelo modelo,

$$R^2 = 1 - \text{SSE/SST} \tag{6}$$

Entretanto, essa métrica não é adequada para os modelos WLS e GLS porque as quantidades SSE e SST não fazem qualquer distinção entre variância amostral, $\boldsymbol{\Sigma}(\hat{\mathbf{y}})$, e variância do erro do modelo $\sigma_\delta^2\mathbf{I}$.

Na abordagem B-GLS, o que de fato interessa entender é a redução observada apenas na variância do erro do modelo porque o erro amostral, além de não poder ser explicado pelo modelo, representa um ruído que acaba por complicar a análise. Portanto, faz-se necessário o

desenvolvimento de uma nova métrica que seja capaz de separar o que é variância do erro do modelo e o que é variância do erro amostral. Esta métrica apresentada aqui é chamada de Pseudo-R^2_{GLS}, desenvolvida em Reis (2005). A proposta é que ela seja estimada através da seguinte expressão:

$$\text{Pseudo - } R^2_{GLS} = \frac{n\left[\hat{\sigma}^2_\delta(0) - \hat{\sigma}^2_\delta(k)\right]}{n\hat{\sigma}^2_\delta(0)} = 1 - \frac{\hat{\sigma}^2_\delta(k)}{\hat{\sigma}^2_\delta(0)} \tag{7}$$

onde $\hat{\sigma}^2_\delta(k)$ é a variância do erro do modelo estimada com *k* variáveis explanatórias, e $\hat{\sigma}^2_\delta(0)$ é a variância do erro do modelo quando nenhuma variável explanatória é empregada, ou seja, quando o modelo é igual à média regional. A Pseudo-R^2_{GLS} é uma extensão direta do tradicional $R^2_{ajustado}$, no sentido que emprega uma razão entre estimadores não-tendenciosos da variância do erro δ e da variância da variável hidrológica de interesse y. Se $\hat{\sigma}^2_\delta(k) = 0$, Pseudo-$R^2_{GLS} = 1$, exatamente como deveria ser, muito embora o modelo não seja perfeito, já que $\text{Var}[\eta_i + \delta_i]$ não é igual a zero porque $\text{Var}[\eta_i] > 0$.

A Tabela 1 apresenta uma pseudo tabela ANOVA para os modelos WLS e GLS. Esta tabela descreve o quanto da variabilidade das observações pode ser atribuída ao modelo regional, e quanto da variância residual pode ser atribuída ao erro do modelo e ao erro amostral, respectivamente. O problema é que não se pode resolver a questão de quais são os erros do modelo porque não se conhece os valores dos erros amostrais η_i para cada *i*. Entretanto, é possível descrever o somatório total dos quadrados dos erros amostrais através do seu valor médio, dado por $tr[\Sigma(\hat{\mathbf{y}})]$, onde tr[**A**] é o traço da matriz **A**. Dado que há *n* equações, uma para cada estação utilizada na análise de regressão, a variação total devida ao erro do modelo, δ, para um modelo com *k* variáveis explanatórias, possui uma média igual a $n\sigma^2_\delta(k)$. Essas estatísticas, $tr[\Sigma(\hat{\mathbf{y}})]$ e $n\sigma^2_\delta(k)$, fornecem uma descrição de duas das três fontes de variação.

Para um modelo que não possui qualquer variável explanatória, a não ser a própria média, a estimativa da variância do erro do modelo, $\sigma^2_\delta(0)$, descreve toda a variação em $\hat{y}_i = y_i + \eta_i$ que não é explicada pelos erros amostrais η_i. Portanto, em média, $\sigma^2_\delta(0)$ deveria ser igual à verdadeira variação em y devido à regressão e a variação devido aos erros do modelo δ. Desta forma, o valor esperado da soma TOTAL dos quadrados dos resíduos devido ao modelo, erro do modelo e erro amostral é descrito como sendo igual a $n\sigma^2_\delta(0) + tr[\Sigma(\hat{\mathbf{y}})]$. Portanto, atribui-se ao modelo um valor esparado da soma dos quadrados igual a $n[\sigma^2_\delta(0) - \sigma^2_\delta(k)]$. Esta tabela é chamada de pseudo-

ANOVA porque a contribuição das três fontes de erro é estimada, ao invés de ser determinada através dos erros residuais e das predições do modelo nas *n* estações, e pelo fato de se ignorar o impacto da correlação entre os erros amostrais.

Tabela 1 – Pseudo Tabela ANOVA

Fonte	Graus de liberdade	Soma dos quadrados
Modelo	k	$n[\sigma_\delta^2(0) - \sigma_\delta^2(k)]$
Erro do modelo	n-k-1	$n\sigma_\delta^2(k)$
Erro amostral	n	$tr[\sum(\hat{y})]$
Total	2 n -1	$n\sigma_\delta^2(0) + tr[\sum(\hat{y})]$
EVR		$\frac{1}{n} tr[\sum(\hat{y})] / \sigma_\delta^2(k)$
MBV		$\frac{1}{n} w^T \Lambda \sigma_\delta^2 w$ onde w é o vetor $1/\sqrt{\Lambda_{ii}}$

3.4 – Razão entre Variâncias dos Erros (*Error Variance Ratio-EVR*)

A Razão entre Variâncias dos Erros (EVR) é uma estatística utilizada para diagnóstico do modelo de regressão. Ela é empregada para determinar se um simples modelo de regressão OLS é suficiente, ou se modelos mais sofisticados, WLS ou GLS, são mais apropriados para os dados. EVR é uma razão entre a variância média do erro amostral e a variância do erro do modelo. Valores de EVR acima de 20% indicam que a variância amostral não é desprezível quando comparada com a variância do erro do modelo, o que sugere o uso de uma análise WLS ou GLS. O EVR é estimado da seguinte maneira:

$$EVR = \frac{\text{SS(erro amostral)}}{\text{SS(erro do modelo)}} = \frac{\text{tr}[\Sigma(\hat{y})]}{n\sigma_\delta^2(k)} \tag{8}$$

3.5 – Grau de Inadequação da Variância de β (*Misrepresentation of the Beta Variance* – MBV)

Apesar de EVR ser capaz de avaliar a adequação ou não da análise OLS frente às análises WLS e GLS, esta métrica não dá nenhum indicativo de que tipo de regressão deve ser utilizada, WLS ou GLS, caso a análise OLS não se mostrar adequada. A métrica MBV foi criada exatamente

para tentar definir se a análise WLS é suficiente, ou se a análise GLS é de fato a mais indicada para modelar os dados (Griffis e Stedinger, 2006; Griffis, 2006).

A MBV procurar avaliar qual é o erro acarretado na estimativa da precisão do parâmetro b_0 do modelo, que é o estimador de β_0, caso utilize-se a análise WLS no lugar da GLS. A métrica MBV baseia-se em b_0 porque se sabe que a covariância observada nos dados possui uma influência maior na estimativa da constante do modelo de regressão do que nos outros parâmetros do modelo (Stedinger e Tasker, 1985). Se o valor de MBV for muito superior à unidade, há uma indicação de que a análise GLS deve ser empregada. O MBV é calculado da seguinte forma,

$$MBV = \frac{Var[b_0^{WLS} \mid análise GLS]}{Var[b_0^{WLS} \mid análise WLS]} = \frac{w^T \Lambda w}{n} \ em\ que\ w_i = \frac{1}{\sqrt{\Lambda_{ii}}} \tag{9}$$

3.6 – *Leverage* (alavanca) e Influência

Leverage e Influência são duas estatísticas descritivas utilizadas para avaliar o ajuste de modelos de regressão aos dados, adequabilidade de modelos, e qualidade dos dados. *Leverage*, como adotado em Tasker e Stedinger (1989, eq. 23), considera se um dado valor de variável explanatória é anormal, indicando que seja bastante provável que este ponto tenha um grande impacto na estimativa dos coeficientes do modelo de regressão. Quando a estatística *Leverage* é aplicada a modelos de regressão WLS e GLS, tanto o valor da variável explanatória, quanto o peso estatístico dado a esta observação na análise de regressão, são levados em consideração. *Leverage* mede o impacto marginal dos resíduos na predição dos valores de y em cada estação. A expressão abaixo segue o cálculo de *Leverage* para modelos WLS e GLS proposta por Tasker e Stedinger (1989), porém adaptada para a análise Bayesiana,

$$leverage(\hat{y}_i, \mathbf{x}_i) = \frac{\partial(\mathbf{x}_i \mathbf{b})}{\partial \varepsilon_i} = \underset{\sigma_\delta^2}{E}\left[\mathbf{x}_i (\mathbf{X}^T \mathbf{\Lambda}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{\Lambda}^{-1}\right]_i \tag{10}$$

onde $\mathbf{x}_i\mathbf{b}$ é o estimador de y_i assciado a $\mathbf{x}_i$. O valor esperado dos valores da estatística *Leverage* é igual a (k+1)/n, onde k é a dimensão de β e n é o número de estações empregadas na análise de regressão. Portanto, valores de Leverage maiores do que 2(k+1)/n são considerados elevados (Tasker e Stedinger, 1989).

Diferentemente de *Leverage*, que procura indicar aqueles pontos que provavelmente afetam o ajuste da regressão, a estatística Influência descreve aqueles pontos que tiveram um impacto anormal na análise de regressão. Em muitos casos, observações que possuem grande Influência,

também possuem grande Leverage. Alta Influência requer a combinação de *Leverage* alta com grande erro residual.

A estatística de Influência apresentada abaixo, proposta por Tasker e Stedinger (1989), é baseada na métrica Distância de Cook (*Cook´s Distance*) (Cook e Weisberg, 1982; Clarke, 1994),

$$D_i = \frac{k_{ii}\hat{\varepsilon}_i^2}{(k+1)(\lambda_{ii} - k_{ii})^2} \tag{11}$$

em que k_{ii} e λ_{ii} são os valores dos elementos da diagonal de K = $\mathbf{X}(\mathbf{X}^T\mathbf{\Lambda}^{-1}\mathbf{X})^{-1}\mathbf{X}^T$ e de Λ, respectivamente.

3.7 – Influência -σ

A estimativa da variância do erro do modelo é extremamente importante, pois o seu valor determina o peso que se dá à estimativa regional da variável hidrológica de interesse em relação à estimativa local, ou seja, àquela estimativa baseada apenas nos dados da estação de interesse. Portanto, é interessante entender quais, se é que existe alguma, as estações que possuem um impacto anormal na estimativa da variância do erro do modelo regional.

A estatística Influência-σ descreve a influência que uma dada observação tem na estimativa da variância do erro do modelo, identificando, portanto, aquelas observações que individualmente podem aumentar potencialmente a estimativa de σ_δ^2. A estatística Influência, descrita acima, identifica aqueles pontos que influenciam de forma significativa os valores de predição do modelo, revelando, portanto, a instabilidade das predições do modelo de regressão naqueles faixas de variáveis explanatórias. Entretanto, a estatística Influência descreve apenas a influência que uma observação tem nas predições do modelo em cada estação empregada na análise. Esta estatística não faz nenhuma inferência a respeito da influência deste ponto na estimativa da variância do erro do modelo. A Influência-s considera se um erro residual ε_i teve de fato um impacto importante na estimativa da variância do erro do modelo. A estatística Influência-σ é calculada por:

$$Influência - \sigma = \frac{\sum_{j=1}^{n} \hat{\varepsilon}_i (\Lambda^{-1})_{ij} \hat{\varepsilon}_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} \hat{\varepsilon}_i (\Lambda^{-1})_{ij} \hat{\varepsilon}_j} = \frac{\hat{\varepsilon}_i (\Lambda^{-1}\hat{\varepsilon})_i}{\hat{\varepsilon}^T \Lambda^{-1}\hat{\varepsilon}} \tag{12}$$

A soma dos quadrados dos resíduos padronizados, $\hat{\varepsilon}^T\Lambda^{-1}\hat{\varepsilon}$, utilizada no cálculo da função verossimilhança dos dados, e na estimativa da variância do erro do modelo pelo método dos

momentos generalizados (Stedinger e Tasker, 1985), é dividida entre as diferentes estações. Por construção, o valor esperado da estatística Influência-s é igual a 1/n, onde n é o número de estações empregadas na análise de regressão. Valores de Influência-s acima de 2/n são considerados valores elevados.

4 – APLICAÇÃO: ESTIMATIVA REGIONAL DO COEFICINETE DE ASSIMETRIA

A estimativa do parâmetro de forma de uma distribuição teórica de probabilidades costuma ter uma incerteza maior quando comparada com a estimativa de outros parâmetros. Quanto menor o tamanho da série histórica, maior a incerteza, resultando em estimativas de quantis de cheias com baixa precisão. Uma maneira de reduzir a incerteza na estimativa do parâmetro de forma em locais de série histórica relativamente curta consiste em combinar a estimativa local desta variável com uma estimativa regional, obtida através de modelos de regressão.

Nos Estados Unidos, todas as agências estaduais ou federais devem seguir os procedimentos para estimativa de vazões de chieas que estão preconizados na publicação Bulletin 17B (IACWD, 1981). O Bulletin 17B recomenda o uso da distribuição Log-Pearson Tipo 3 para realização de análise de freqüência de cheias, e sugere o uso de uma estimativa regional do parâmetro de forma desta distribuição, que é o coeficiente de assimetria, para ser empregada em conjunto com a estimativa local na tentativa de aumentar a precisão dos quantis de cheia [Hardison, 1975; McCuen, 1979 e 2001; IACWD, 1981; Stedinger et al., 1993; Griffis e Stedinger, 2007].

Embora o exemplo apresentado aqui trate da regionalização do coeficiente de assimetria, o mesmo procedimento poderia ser aplicado à regionalização do parâmetro de forma κ da distribuição de Valores Extremos Generalizada.

Reis et al. (2005) desenvolveram modelos regionais do coeficiente de assimetria para a bacia do Rio Tibagi (17 estações) e para a bacia do Rio Muskingum (44 estações). No presente estudo, uma base de dados maior foi utilizada: a bacia do Rio Illinois com 62 estações, cujas séries possuem entre 14 e 90 anos de dados de vazão. Um grande esforço vem sendo empregado pelo USGS para analisar dados de mais de 800 estações fluviométricas do sudeste americano, empregando a análise B-GLS descrita em Reis et al. (2005) em conjunto com as estatísticas de diagnóstico apresentadas neste artigo, para obter um modelo regional de assimetria. Esta iniciativa no sudeste americano é um projeto em andamento, cujos resultados obtidos até o momento são rapidamente discutidos ao final do artigo, bem como nossas expectativas de trabalhos futuros.

O estudo na bacia do Rio Illinois avaliou sete variáveis explanatórias, mais uma constante. Duas variáveis binárias (Z1, Z2) forma empregadas na tentativa de explorar a possibilidade de ver a variabilidade observada nos coeficientes de assimetria serem explicadas por região hidrológica. Estas duas variáveis binárias representam três regiões da bacia do Rio Illinois: Little Wabash (1,0),

Rock (0,1) e Sangamon (0,0), conforme descrito em Tasker e Stedinger (1986). As outras cinco variáveis explanatórias são: 1) área de drenagem, em mi^2; 2) declividade do canal principal, em pés/mi; 3) área de lagos em termos de porcentagem em relação à área da bacia; 4) cobertura florestal em termos de porcentagem em relação à área da bacia; 5) índice de permeabilidade do solo, que varia entre 1 (baixa permeabilidade) e 6 (alta permeabilidade). Empregaram-se, de fato, os logaritmos das cinco variáveis indicadas acima. Todas estas variáveis foram centradas em relação à média. Este procedimento permitiu que todas as variáveis explanatórias, mais a constante, tivessem a mesma escala, permitindo um cálculo simples da média regional do coeficiente de assimetria para cada uma das três regiões hidrológicas.

A matriz de covariância amostral para a bacia do Rio Illinois foi desenvolvida utilizando os procedimentos descritos em Reis et al. (2005). O coeficiente de correlação espacial entre as vazões $\rho(d_{ij})$ foi modelado como uma função da distância entre pares de estação,

$$\rho(d_{ij})^{\kappa} = \theta^{\left(\frac{\kappa d_{ij}}{\alpha . d_{ij}+1}\right)} \tag{13}$$

onde d_{ij} é a distância entre estações, em kilômetros, $\theta = 0{,}988$, $\alpha = 0{,}002$, e $\kappa = 3$.

As tabelas 2 e 3, junto com a Figura 1, apresentam os resultados da estimativa do coeficiente de assimetria regional para a bacia do Rio Illinois com base na análise de regressão B-GLS. Todas as combinações possíveis das sete variáveis explanatórias forem utilizadas para 128 possíveis modelos regionais de assimetria. O modelo B-GLS1, com uma variável explanatória, o ln(declividade do canal), foi escolhido como o melhor modelo pois era o que apresentava a menor variância média de predição para uma nova estação e a menor variância do erro do modelo. À título de comparação, o modelo B-GLS0, que representa a média regional, já que não possui nenhuma variável explanatória, também foi incluído na Tabela 2.

Como mostrado na Tabela 2, a variância do erro do modelo para o modelo B-GLS1 é igual a 0,133, enquanto que para o modelo de média regional é igual a 0,151. A VMP do modelo B-GLS1 para uma nova estação é igual a 0,158, com um tamanho de série equivalente (TSE) de 49 anos. Isto significa dizer que o estimador regional obtido pelo modelo B-GLS1 é, em média na região, equivalente a um estimador local baseado em 49 anos de dados. O pseudo-R^2 é igual a 0,12, o que significa que o ln(declividade do canal) explica apenas 12% da variabilidade espacial dos verdadeiros valores do coeficiente de assimetria. No caso da análise OLS, o modelo com o mesmo conjunto de variáveis explanatórias obteve um R^2 igual a 0,09, o que subestima o verdadeiro valor do modelo regional.

Tabela 2: Regressão do coeficiente de assimetria para a bacia do Rio Illionis (62 estações). Plausibilidade Bayesiana (%) e erro-padrão em parênteses.

Modelos	Constante	Ln(decliv.)	VEM	VMA	VMP_{nova}	Pseudo-R^2	TSE (anos)
B-GLS0	-0,149	-	0,151 (0,056)	0,015	0,166	0	49
B-GLS1	-0,588	0,127 (3,3%)	0,133 (0,052)	0,025	0,158	0,118	49
B-GLS1 (s/ est. 28)	-0,533	0,105 (7,8%)	0,120 (0,051)	0,024	0,144	0,071	53
B-GLS1 (s/ est. 49)	-0,594	0,125 (2,5%)	0,122 (0,049)	0,023	0,145	0,143	53

A Tabela 3 apresenta a pseudo Tabela ANOVA, onde se pode observar que o erro amostral é mais do que duas vezes maior que o erro do modelo, no caso do modelo B-GLS1. A métrica EVR é igual a 2,3, o que indica claramente que as análises WLS ou GLS são mais indicadas do que a análise OLS. Além disso, como a métrica MBV = 3, pode-se concluir que a análise GLS é claramente mais adequada para modelar os dados da bacia do Rio Illinois do que a análise WLS. Neste caso, a análise WLS iria certamente superestimar a precisão da constante do modelo.

Tabela 3: Pseudo Tabela ANOVA para a bacia do Rio Illinois (modelo B-GLS1)

Fonte	Graus de liberdade		Soma dos quadrados		
	Caso 1	Casos 2 e 3	Caso 1 (todas as est.)	Caso 2 (s/ est. 28)	Caso 3 (s/ est. 49)
Modelo	k =1	k = 1	1,10	0,56	1,24
Erro do modelo	N – k -1 = 60	n – k – 1 = 59	8,24	7,30	7,41
Erroa amostral	n = 62	n = 61	19,04	18,81	18,29
Total	2n – 1 = 123	2n – 1 = 121	27,28	26,11	25,70
EVR			2,31	2,58	2,47
MBV			3,00	3,03	3,01
Pseudo-R^2			0,12	0,07	0,14

A Figura 1 apresenta os resultados de Influência, Leverage e Influência-σ das estações com maior grau de influência na análise de regressão do modelo B-GLS1. As estações estão ordenadas em ordem decrescente de influência. Fica claro que a estação 28 possui a mais alta Influência dentre todas as 62 estações, e um alto grau de Leverage (acima do limite que define uma estação como sendo de alto Leverage).

Além de possui a menor declividade do canal principal dentre todas as bacias empregadas no estudo de regressão, a estação 28 possui uma série histórica relativamente longa, 60 anos de dados, e um resíduo extremamente alto, - 0,88. Isto acaba por caracterizar a estação 28 como um *outlier* no estudo de regressão, resultando em valores altos de Influência, Leverage, e Influência-s. Apenas à título de teste, a análise de regressão foi refeita sem a presença da estação 28. Como mostrado na

Tabela 2, a estimativa da variância do erro do modelo diminui de 0,133, quando toda a base de dados é empregada, para 0,120, quando a estação 28 é removida. A Tabela 3 apresenta a pseudo Tabela ANOVA para o modelo B-GLS1 sem a estação 28.

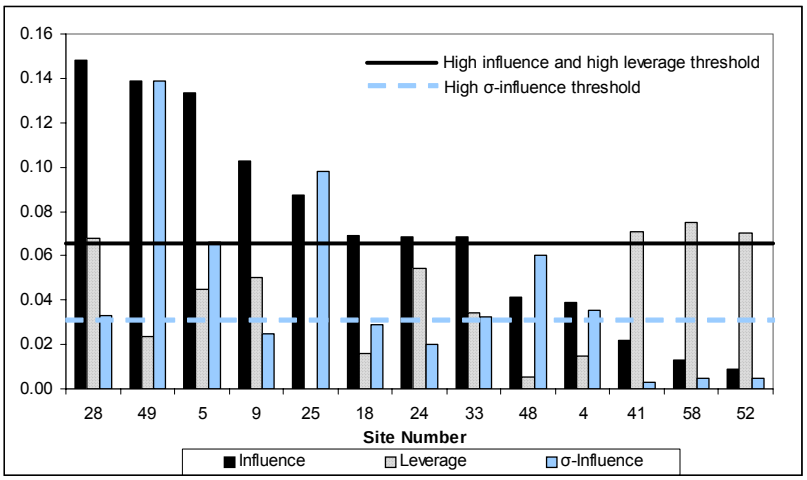

Figura 1: Diagnóstico de Regressão: Influência, Leverage e Influência-σ para a bacia do Rio Illinois (modelo B-GLS1).

Como mostrado na Figura 1, a estação 49 é a que possui maior Influência-s, apesar de apresentar baixos valores de Leverage e Influência. O altíssimo resíduo da estação 49, -1,53, é o terceiro maior dentre as 62 estações utilizadas no estudo. A estação 49 possui uma série histórica relativamente curta,apenas 16 anos, e um valor muito alto de assimetria amostral de -1,82. A predição da assimetria regional obtida pelo modelo B-GLS1 para a estação 49 é de -0,282. À título de teste, a estação 48 foi removida dos dados e a análise de regressão para o modelo B-GLS1 foi refeita. Conforme apresentado na Tabela 2, a estimativa da variância do erro do modelo foi reduzida de 0,133 para 0,122. Esta redução era esperada já que a estação com a terceira maior Influência-σ foi removida dos dados. Apesar da estação 49 ter um valor baixo de Influência, ela possui um impacto importante na estimativa da variância do erro do modelo, fato este caracterizado por sua alta Influência-σ. A estação 49 não foi identificada pela estatística de diagnóstico, Influência, mas foi reconhecida pela estatística aqui apresentada, Influência-σ, ilustrando a utilidade desta nova estatística de diagnóstico de regressão. A Tabela 3 apresenta a pseudo Tabela ANOVA para o modelo B-GLS1 quando a estação 49 é removida dos dados.

5 – CONCLUSÕES E TRABALHOS EM ANDAMENTO

Este artigo procurou apresentar os avanços que vêm sendo realizados para tornar a análise Bayesiana de modelos de regressão GLS, descrito em Reis et al. (2005), em uma metodologia operacional de regionalização de variáveis hidrológicas.

Diversas estatísticas para diagnóstico de modelos de regressão B-GLS foram desenvolvidas ou adaptadas para a abordagem Bayesiana, quais sejam, um pseudo-R^2 ajustado, uma pseudo Tabela ANOVA, Valor de Plausibilidade Bayesiano, EVR, MBV, *Leverage*, Influência e Influência-σ. Estas estatísticas permitem que se faça um exame mais extenso e completo dos modelos de regressão baseados na abordagem B-GLS.

O procedimento de regressão regional foi ilustrado através de um exemplo de regionalização do parâmetro de forma da distribuição Log-Pearson Tipo 3 para a bacia do Rio Illinois nos Estados Unidos, o que permitiu ilustrar a utilidade destas estatísticas de diagnóstico.

Esforços estão sendo realizados para aplicar o procedimento B-GLS a um estudo com mais de 800 estações fluviométricas, localizadas em mais de 10 estados do sudeste americano. Este estudo está aplicando o procedimento introduzido em Reis et al. (2005) em conjunto com as estatísticas de diagnóstico de regressão apresentadas neste artigo.

Um aspecto importante que este estudo do sudeste americano está indicando é a existência de vários *outliers* (vazões baixas), que acabam por resultar em estimativas anômalas do coeficiente de assimetria e, por conseguinte, uma variância muito grande dos estimadores regionais de assimetria. Isto é um sinal de que é necessário um tratamento prévio dos dados, aplicando-se a censura dos dados, que consiste em remover os *outliers* de vazões baixas, que acabam criando uma distorção na estimativa local da assimetria. Para isto, está sendo utilizado o procedimento de identificação do limite de corte dos dados, baseado no método do ajuste com base na probabilidade condicional (Conditional Probability Adjustment – CPA), preconizado no Bulletin 17B (IACWD, 1982), em conjunto com o Algoritmo dos Momentos Esperados (Expected Moments Algorithm – EMA), desenvolvido por Cohn (1997). Deste modo, será possível determinar um estimador regional adequado para o coeficiente de assimetria nesta região.

BIBLIOGRAFIA

BAYARRI, M.J., AND J.O. BERGER, P (2000), "*Values for Composite Null Models*", J. of the Am. Statistical Assoc. 95 (452), 1127-1142, December.

BENSON, M.A. AND N.C. MATALAS, (1967), "*Synthetic Hydrology Based on Regional Statistical Parameters*", Water Resources Research, 3(4), 931-935.

CLARKE, R. T. (1994), *Statistical Modeling in Hydrology*, John Wiley & Sons Inc.

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

COOK, R.D. AND WEISBEG, S. (1982), *Residuals and Influence in Regression*, Chapman and Hall, New York, NY, 230 pp.

FILL, H. D., AND J. R. STEDINGER (1998), "*Using Regional Regression within Index Flood Procedures and an Empirical Bayesian Estimator"*, Journal of Hydrology, 210, 128-145

GELMAN, A., CARLIN, J.B., STERN, H.S., AND RUBIN, D.B. (2004), *Bayesian Data Analysis*, Chapman & Hall/CRC, Boca Raton, FL.

GRIFFIS, V. W., AND J. R. STEDINGER (2006), *The Use of GLS Regression in Regional Hydrologic Analyses, manuscript*, Cornell University, July.

GRIFFIS, V. W. (2006), *Flood Frequency Analysis, Bulletin 17B and Regional Analysis*, Ph.D. Thesis, Cornell University, August, Ithaca, NY, EUA.

GRIFFIS, V. W., AND J. R. STEDINGER (2007), "*The LP3 distribution and its application in flood frequency analysis, 3. Sample Skew and Weighted Skew Estimators*", submitted J. of Hydrol. Engineering.

HARDISON, C. H. (1975), "*Generalized skew coefficients of annual floods in the United States and their application*", Water Resour Res., 11(6), 851-854.

INTERAGENCY ADVISORY COMMITTEE ON WATER DATA (1982), *Guidelines for Determining Flood Flow Frequency, Bulletin #17B, U.S. Department of the Interior*, U.S. Geological Survey, Office of Water Data Coordination, Reston Virginia.

JENNINGS, M.E., W.O. THOMAS, JR., AND H.C. RIGGS (1994), *Nationwide Summary of U.S. Geological Survey Regional Regression Estimates for Estimating Magnitude and Frequency of Floods for Ungaged Sites*, Water Resources Investigations Report 94-4002, U.S. Geological Survey: Reston, Virginia.

KUCZERA, G. (1982), "*Combining Site-Specific and Regional Information, An Empirical Bayes Approach*", Water Resources Research, 18(2), 306-314.

LINART, H. AND W. ZUCCHINI (1986), *Model Selection*, John Wiley and Sons, Inc., New York.

LINDLEY, D.V. (1965), *Introduction to Probability and Statistics from a Bayesian Viewpoint*, Part2. Inference. Cambridge: University Press.

MADSEN, H, AND D. ROSBJERG (1997), "*Generalized least squares and empirical Bayes estimation in regional partial duration series index-flood modeling*", Water Resources Research, 33(4), 771-782.

MARTINS, E.S. AND J.R. STEDINGER (2000), "*Generalized Maximum Likelihood GEV Quantile Estimators for Hydrologic Data*", Water Resources Research, 28(11), 3001-3010.

MATALAS, N.C. AND E.J. GILROY (1968), "*Some Comments on Regionalization in Hydrologic Studies*", Water Resoures Research, 4(6), 1361-1369.

MOSS, M.E. AND M.R. KERLINGER (1974), "*Surface Water Network Design by Regression Analysis Simulation*", Water Resources Research, 10(3), 427-433.

REIS, D. S., JR., J. R. STEDINGER, AND E. S. MARTINS (2004), "*Operational Bayesian GLS Regression for Regional Hydrologic Analyses*", In: Sehlke, G., D.F. Hayes, D.K. Stevens (Eds.), Proceedings of the World Water and Environmental Resources Congress: Critical Transitions in Water and Environmental Resources Management, June 27 - July 1, 2004, Salt Lake City, Utah, USA.

REIS, D. S., JR., J. R. STEDINGER, AND E. S. MARTINS (2005), "*Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation*", Water Resour. Res., 41, W10419, doi:10.1029/2004WR003445.

REIS, D.S.,Jr. (2005), *Flood Frequency Analysis Employing Bayesian Regional Regression and Imperfect historical Information*, Ph.D. Thesis, Cornell University, Ithaca – NY, EUA. January.

ROBINS, J. M. (2000), A. VAN DER VAART, AND V. VENTURA, "*Asymptotic Distribution of P Values in Composite Null Models"*, J. of the Am. Statistical Assoc. 95 (452), 1143-1156, December.

SHANE, R.M. AND GAVER, D.P. (1970), "*Statistical Decision Theory Techniques for the Revision of Mean Flood Flow Regression Estimates*", Water Resources Research, 6(6), 11649-1654.

SHU, C. AND D.H. BURN (2004), "*Homogeneous Pooling Group Delineation for Flood Frequency Analysis using a Fuzzy Expert System with Genetic Enhancement*", Journal of Hydrology, 291, 132-149.

STEDINGER, J.R. (1983), "*Design Events with Specified Flood Risk*", Water Resources Research, 19(2), 511-522.

STEDINGER, J.R., AND G.D. TASKER (1985), "*Regional Hydrologic Analysis, 1. Ordinary, Weighted and Generalized Least Squares Compared*", Water Resources Research, 21(9), 1421-1432.

TASKER, G.D. (1980). "*Hydrologic Regression with Weighted Least Squares*," Water Resources Research, 16(6), 11107-1113.

TASKER, G AND J.R. STEDINGER (1986a), "*Correction to "Regional hydrologic analysis, 1, Ordinary, weighted and generalized least squares compared*", Water Res. Research, 22(5), 844.

TASKER, G AND J.R. STEDINGER (1986b), "*Regional hydrologic analysis, 2: Model-error estimators, estimation of sigma and log-Pearson Type 3 distributions*", Water Resour. Res., 22(10), 1487-1499.

TASKER, G.D., AND J.R. STEDINGER (1989), "*An Operational GLS Model for Hydrologic regression"*, Journal of Hydrology, 111(1-4), 361-375.

THOMAS, D.M., AND M.A. BENSON (1970), "*Generalization of Streamflow Characteristics from Drainage-Basin Characteristics*", Water-Supply Paper 1975, U.S. Geological Survey, Washington, D.C.

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

WALKER, J.F., AND W.R. KRUG (2003), "*Flood-frequency characteristics of Wisconsin streams*", Water Resources Investigations Report 03-4250, U.S. Geological Survey, Reston Virginia.

ZELLNER, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, John Wiley and Sons, Inc., New York.