RBRH – Revista Brasileira de Recursos Hídricos Volume 18 n.4 –Out/Dez 2013,173-184

Amostragem de Séries Sintéticas Hidrológicas

Daniel H. Marco Detzel[1], Marcelo Rodrigues Bessa[1], Miriam Rita Moro Mine[2]
daniel@lactec.org.br; bessa@lactec.org.br; mrmine.dhs@ufpr.br

Recebido: 14/11/12 - revisado: 03/04/13 - aceito: 06/08/13

RESUMO

O uso de modelos estocásticos para a geração de séries temporais sintéticas tem grande aceitação em diversas áreas nas quais o uso de informações históricas é limitado, dentre elas o planejamento dos recursos hídricos. Dependendo de sua aplicação, entretanto, o uso de todos os cenários sintéticos pode se mostrar oneroso computacionalmente, fazendo com que seja necessária a simplificação dos sistemas utilizados. Sob esse contexto, propõe-se um estudo de amostragem de séries sintéticas, na intenção de reduzir o número de cenários gerados sem perder a representatividade obtida com o modelo estocástico. O método está calcado em duas etapas: (i) agrupamento das séries sintéticas através da determinação de distâncias de Mahalanobis entre elas e as séries históricas e (ii) aplicação de amostragem estratificada sobre o conjunto resultante. Como estudo de caso, selecionaram-se séries de vazões afluentes a 62 usinas hidrelétricas brasileiras, cujas séries sintéticas foram geradas a partir de um modelo autorregressivo contemporâneo multivariado CARMA(p,q). Os resultados confirmaram a plausibilidade do método de amostragem, permitindo a redução do número de séries sem alterar a distribuição empírica de probabilidades conseguida com o conjunto de séries sintéticas originalmente geradas. Análises extras relativas à estabilidade do método frente ao número de séries geradas estão presentes.

Palavras-chave: *Amostragem. Distância de Mahalanobis. Modelo contemporâneo. Séries sintéticas mensais.*

INTRODUÇÃO

Modelos estocásticos para geração de séries sintéticas de vazões tiveram uso crescente a partir do momento em que se percebeu que a série histórica sozinha era insuficiente para o planejamento apropriado de sistemas de recursos hídricos (JACKSON, 1975). Com emprego inicial no dimensionamento de reservatórios (VOGEL e STEDINGER, 1988 e referências citadas por esses autores), as aplicações se expandiram para operação de reservatórios considerando eventuais usos múltiplos (FREVERT *et al.*, 1989), abastecimento urbano (FRICK *et al.*, 1990), sistemas de irrigação (MORENO *et al.*, 2008), entre outros.

De particular interesse, estudos relacionados à operação ótima de múltiplos reservatórios têm evoluído rapidamente graças ao contínuo avanço computacional. Dessa maneira, ganharam espaço modelos de maior complexidade, capazes de contemplar variáveis multidimensionais, não lineares e estocásticas. Mesmo assim, Labadie (2004) aponta que algumas classes de modelos ainda oferecem grandes desafios, principalmente os que fazem uso de cenários sintéticos de afluências na otimização estocástica explícita de grandes sistemas (e.g. KELMAN *et al.*, 1990; PEREIRA; PINTO, 1985; SEIFI; HIPEL, 2001).

Como forma de aliviar o peso computacional de tais modelos, alguns autores desenvolveram algoritmos eficientes de solução, baseados principalmente em técnicas de decomposição (SAADOULI, 2010; PEREIRA e PINTO, 1985). Outros autores optaram por reduzir o número de cenários nas simulações, sem alterar a estrutura principal do modelo (FABER; STEDINGER, 2001).

O artigo aqui apresentado, em concordância com o contexto do segundo método supracitado, trabalha com a redução do número de cenários sintéticos a serem utilizados para modelos que exigem grande capacidade de processamento computacional. No estudo de Faber e Stedinger (2001), foram usadas combinações de cenários de previsão de afluências futuras semelhantes. No presente artigo, utilizou-se uma técnica de amostragem estatística para reduzir o número de séries sintéticas de vazões geradas por um modelo estocástico multivariado.

Para se chegar ao número reduzido de cenários, trabalhou-se em duas etapas: (i) ranquear as

[1] - Instituto de Tecnologia para o Desenvolvimento – LACTEC
[2] - Universidade Federal do Paraná – UFPR

matrizes de séries sintéticas de acordo com um critério de similaridade com a matriz de séries históricas e (ii) selecionar as séries de forma que a amostra final contenha elementos de diferentes probabilidades de ocorrência na população, garantindo assim representatividade de eventos hidrológicos diversos.

Os cenários sintéticos foram gerados para um grupo de 62 usinas hidrelétricas integrantes do Sistema Interligado Nacional através de um modelo multivariado autorregressivo com médias móveis contemporâneo – CARMA(p,q). Para o critério de similaridade utilizado no ranqueamento dos cenários, buscou-se fundamentação em estudos de agrupamentos de séries temporais, nos quais a determinação de distâncias entre elementos das séries é bastante comum. A fim de tornar possível o cálculo das distâncias sem prejudicar a estrutura de correlações espaciais entre as vazões das usinas, empregou-se a distância generalizada de Mahalanobis (FERREIRA, 2004). A amostragem, por sua vez, foi desenvolvida de acordo com os princípios da amostragem estratificada clássica, na qual uma população heterogênea é dividida em subpopulações homogêneas (COCHRAN, 1977). A amostragem é, então, feita individualmente a cada subpopulação.

O artigo está estruturado da seguinte forma: a seção seguinte se trata de uma breve revisão de literatura relativa ao agrupamento de séries temporais e amostragem estratificada. Na sequência, o método proposto no presente artigo é explicado em maiores detalhes. Depois, o dados utilizados são descritos mostrando características gerais das usinas e séries empregadas. Resultados são apresentados na seção seguinte, acompanhados de análises diversas. Por fim, a última seção conclui o estudo.

FUNDAMENTAÇÃO TEÓRICA

O agrupamento de séries temporais é um tema rico na literatura e que possui respaldo em muitas áreas do conhecimento. Liao (2005) oferece uma revisão detalhada sobre os diversos algoritmos existentes e a aplicabilidade de cada um, dado o tipo de série com a qual se está trabalhando. Em comum a um grande número de estudos, determinam-se critérios de similaridade de acordo com distâncias entre elementos das séries consideradas. Sejam dois vetores genéricos X e $Y \in \mathbb{R}^p$ e uma matriz qualquer Ψ. A distância entre os elementos x e y (pertencentes aos respectivos vetores X e Y) é dada pela equação (1) (FERREIRA, 2004):

$$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})'\boldsymbol{\Psi}(\boldsymbol{x}-\boldsymbol{y})} \quad (1)$$

A métrica Ψ pode assumir diferentes matrizes positivas definidas. Quando se emprega a matriz identidade (e.g. $\Psi = \mathrm{I}$), tem-se a distância Euclidiana clássica, métrica utilizada em diversos trabalhos (KOSMELJ e BATAGELJ, 1990; POLICKER e GEVA, 2000; KALPAKIS *et al.*, 2001; PICCOLO, 1990; CORDUAS e PICCOLO, 2008), mas que possui a limitação de não considerar a estrutura de correlações cruzadas entre as séries (CAIADO *et al.*, 2006).

Como métrica alternativa, pode-se utilizar a inversa da matriz de covariâncias ($\Psi = \boldsymbol{\Sigma}^{-1}$), obtendo-se a chamada distância generalizada de Mahalanobis. De Maesschalck *et al.* (2000) detalham as principais características dessa distância, comparando-a com a distância Euclidiana e apontando diferenças entre as duas. Jouan-Rimbaud *et al.* (1998) utilizam a distância de Mahalanobis como um de três critérios para identificar a representatividade de grupos de dados multidimensionais. Farber e Kadmon (2003) a usam na modelagem bioclimática de plantas, obtendo melhorias quando comparada a formulações anteriormente utilizadas. Picard (2012) desenvolveu um método de agrupamento baseado na distância de Mahalanobis para aplicação em casos nos quais se tem mais de uma população, cada uma com seu número de elementos. Corduas (2011) aplicou a distância de Mahalanobis entre os coeficientes harmônicos da regressão que definiu a sazonalidade de 89 séries de vazões de rios americanos, na intenção de classificá-las. Demais aplicações da métrica de Mahalanobis podem ser encontradas na detecção de outliers (ATKINSON e RIAN, 2007; FILMOSER e HRON, 2008; GIMÉNEZ *et al.*, 2012), aprendizagem de máquinas (MANOLOVA e GUÉRIN-DUGUÉ, 2008; CHANG, 2012; XIANG *et al.*, 2008), entre outras.

No campo da amostragem estratificada, dois pontos são recorrentes em qualquer aplicação: a definição dos limites superior e inferior de cada grupo (ou estrato) e o número de elementos a serem amostrados em cada estrato. A primeira questão foi tratada nos estudos de Dalenius e Hodges (1959), Ekman (1959), Gunning (2004), Kozak (2004) e Nicolini (2001), enquanto que a alocação das amostras foi tema dos trabalhos de Chaddha *et al.* (1971), Huddleston *et al.* (1970) e Bretthauer *et al.* (1999). Ainda, Keskintürk e Er (2007) aplicaram algoritmos genéticos para estudar ambas as questões simultaneamente.

No presente artigo, a métrica de Mahalanobis é utilizada na determinação das distâncias entre

as séries sintéticas de vazões e as séries históricas. O vetor de distâncias resultantes foi submetido a uma amostragem estratificada para tornar possível a redução do número de séries sintéticas, sem perder as diferentes características hidrológicas obtidas com o modelo estocástico. Na sequência são detalhados os métodos empregados para a classificação das séries e subsequente amostragem.

MÉTODOS DE ANÁLISE

Modelo para geração das séries mensais

A geração dos cenários sintéticos mensais de vazão se deu através de um modelo estocástico linear, não periódico, multivariado do tipo CARMA(p,q), ou autorregressivo com médias móveis contemporâneo de ordens p e q. Sejam os vetores $Z_t = (Z_{t1}, Z_{t2}, \dots, Z_{tk})'$ e $a_t = (a_{t1}, a_{t2}, \dots, a_{tk})'$, definidos para k séries temporais no tempo t. O modelo CARMA(p,q) é dado, genericamente, pela equação (2):

$$\varphi_i(\mathrm{B})Z_{ti} = \theta_i(\mathrm{B})a_{ti}, i = 1,2,\dots,k \quad (2)$$

onde φ_i é o i-ésimo operador AR de ordem p:

$$\varphi_i(B) = 1 - \varphi_{ii1}B - \varphi_{ii2}B^2 - \cdots - \varphi_{iip}B^p$$

Igualmente, θ_i é o i-ésimo operador MA de ordem q:

$$\theta_i(B) = 1 - \theta_{ii1}B - \theta_{ii2}B^2 - \cdots - \theta_{iip}B^q$$

B é o operador de defasagem do modelo. Para respeitar as condições de invertibilidade e estacionariedade da formulação (Box *et al.*, 1994), $\varphi_i(B) = 0$ e $\theta_i(B) = 0$ devem permanecer fora do círculo unitário. Por fim, assume-se o vetor de resíduos independentes e normalmente distribuídos $a_t \sim NID(0, \Delta)$, onde Δ é a matriz de variância-covariância.

Hipel e McLeod (1994, cap. 21) trazem uma descrição completa das minucias do modelo CARMA. A principal diferença desta formulação para um modelo ARMA multivariado tradicional reside no fato de que as matrizes de parâmetros do modelo contemporâneo são diagonais. Assim, ele respeita, além das estatísticas básicas, as autocorrelações individuais de cada série histórica. A correlação espacial é preservada sucintamente (*lag* zero) através do vetor de resíduos, modelado a partir da equação (3):

$$a_t = M\boldsymbol{\varepsilon}_i \quad (3)$$

onde $\boldsymbol{\varepsilon}_i$ é um vetor de variáveis aleatórias independentes e identicamente distribuídas $\boldsymbol{\varepsilon}_i \sim IID(0,1)$ e M é uma matriz de parâmetros. Aplicações do modelo contemporâneo podem ser conferidas nos estudos de Camacho *et al.* (1987), Haltiner e Salas (1988), Wang (2008) e Stedinger *et al.* (1985). Neste último, em particular, os autores comparam o modelo contemporâneo CARMA com uma tradicional formulação multivariada ARMA, ambos de primeira ordem, concluindo que a performance dos dois pouco variou.

Para aproximar as séries de vazões de uma distribuição normal, aplicou-se transformação logarítmica. Ademais, a sazonalidade das séries foi removida através de padronização individual por média e desvio padrão (HIPEL e McLEOD, 1994). Esse procedimento foi empregado em detrimento do uso de um modelo periódico, na intenção de simplificar todo o método.

A ordem dos modelos foi determinada para cada usina utilizando-se das funções de autocorrelação e autocorrelação parcial em conjunto com o critério de informação bayesiano (BIC – SCHWARTZ, 1978). Como regra geral, limitou-se a ordem máxima em dois (CARMA(2,2)), suficiente para o ajuste de modelos lineares a séries estacionárias (BOX *et al.*, 1994). Os parâmetros individuais p e q foram estimados a partir do método da máxima verossimilhança, enquanto que a matriz de parâmetros M foi calculada a partir da decomposição da matriz de correlações entre os resíduos de cada série (HALTINER; SALAS, 1988).

Método para amostragem das séries sintéticas

A utilização de técnicas de amostragem estatística encontra aplicação em estudos nos quais se necessita informações de uma população, mas não se dispõe de recursos suficientes para realizar inferências sobre todos os dados. No contexto do presente trabalho, os recursos são associados ao tempo computacional gasto por modelos que farão uso das séries sintéticas.

Para a diminuição do número dos cenários sintéticos gerados, poderia ser aplicado o método da amostragem aleatória simples (COCHRAN, 1977, cap. 2) sobre as séries obtidas, porém dois pontos o tornam indesejável: (i) o modelo de geração é multivariado, fazendo com que a amostragem individual prejudique a correlação espacial entre as usinas e (ii) entende-se que existem diversos eventos hidro-

lógicos, com diferentes probabilidades de ocorrência, estimulando o emprego de técnicas de amostragem não equiprovável.

Dessa maneira, optou-se, numa primeira etapa, por trabalhar com as similaridades entre as matrizes de séries sintéticas e a matriz de séries históricas a partir da determinação de distâncias de Mahalanobis. Sua formulação é mostrada na equação (4), obtida diretamente a partir da equação (1):

$$d(x,y) = \sqrt{(x-y)'\Sigma^{-1}(x-y)} \quad (4)$$

onde Σ^{-1} é a inversa da matriz de covariâncias. Para calcular as distâncias, as variáveis aleatórias x e y são substituídas pelas médias das séries históricas e sintéticas, $\bar{x}$ e $\bar{y}$ respectivamente, e a matriz Σ^{-1} assume a inversa da matriz de covariâncias conjunta amostral $\boldsymbol{S}_{xy}^{-1}$. Dessa forma, a equação final para determinação das distâncias de Mahalanobis é:

$$d(\bar{x},\bar{y}) = \sqrt{(\bar{x}-\bar{y})'S_{xy}^{-1}(\bar{x}-\bar{y})} \quad (5)$$

As distâncias foram determinadas individualmente entre cada série sintética e a série histórica. Sejam t o tamanho da série, k o número de séries históricas e N o número de séries sintéticas geradas, segue-se o procedimento:

- Monta-se a matriz de séries históricas $\boldsymbol{x}$ de tamanho (t x k):

$$x = \begin{vmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{t1} & \dots & x_{tk} \end{vmatrix}$$

- Organizam-se as matrizes de séries sintéticas y_N de tamanho (t x k):

$$y_{\mathbf{1}} = \begin{vmatrix} y_{11} & \cdots & y_{1k} \\ \vdots & \ddots & \vdots \\ y_{t1} & \dots & y_{tk} \end{vmatrix}_1 ; (\cdots); y_N = \begin{vmatrix} y_{11} & \cdots & y_{1k} \\ \vdots & \ddots & \vdots \\ y_{t1} & \dots & y_{tk} \end{vmatrix}_N$$

- Calculam-se os vetores de médias para cada matriz:

$$\bar{x} = \frac{\begin{vmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{t1} & \dots & x_{tk} \end{vmatrix}}{|\bar{x}_{.1} \quad \cdots \quad \bar{x}_{.k}|}; \; \bar{y}_N = \frac{\begin{vmatrix} y_{11} & \cdots & y_{1k} \\ \vdots & \ddots & \vdots \\ y_{t1} & \dots & y_{tk} \end{vmatrix}_N}{|\bar{y}_{.1} \quad \cdots \quad \bar{y}_{.k}|_N}$$

- Calcula-se a matriz de diferenças entre as médias de cada série, de tamanho (k x N):

$$(\bar{x}-\bar{y}) = \begin{vmatrix} (\bar{x}_{.1}-\bar{y}_{.1})_1 & \cdots & (\bar{x}_{.k}-\bar{y}_{.k})_1 \\ \vdots & \ddots & \vdots \\ (\bar{x}_{.1}-\bar{y}_{.1})_N & \dots & (\bar{x}_{.k}-\bar{y}_{.k})_N \end{vmatrix}$$

- Calcula-se a matriz de covariâncias conjunta entre as séries (McLachlan, 1999):

$$S_{xy} = \frac{(n_x-1)S_x + (n_y-1)S_y}{N}$$

onde S_x e S_y são as matrizes de covariâncias amostrais e n_x e n_y são os números de elementos de cada série;

- Substituem-se $(\bar{x}-\bar{y})$ e S_{xy} na equação (5) e determinam-se as distâncias.

A lógica envolvida no método é obter um ranque das séries sintéticas de acordo com sua semelhança com a série histórica. Quanto menor o valor da distância de Mahalanobis, mais próxima da série histórica é a série sintética.

A estratificação da população de distâncias é feita através do método de Dalenius e Hodges (1959), também conhecido por método MVS (*minimum variance stratification*). Mesmo se tratando de um método simples, é o mais utilizado nos estudos com amostragem estratificada, fato que comprova sua eficiência (KESKINTÜRK e ER, 2007). A ideia básica é obter a função cumulativa das raízes quadradas das frequências observadas. Os estratos são obtidos determinando-se intervalos iguais da série acumulada resultante.

A amostragem estratificada de uma população com N elementos é caracterizada por alguns parâmetros representativos. Sendo h o estrato, N_h, n_h e s_h^2 representam o número total de elementos, o número de elementos amostrados e a variância do estrato h, respectivamente. Definem-se o peso do estrato por $W_h = N_h/N$, a fração amostral do estrato por $f_h = n_h/N_h$ e a variância total da amostra pela equação (6):

$$s^2 = \sum_{h=1}^{L} W_h^2 \frac{s_h^2}{n_h}(1-f_h) \quad (6)$$

Para determinar o tamanho global da amostra (equivalente à soma do número de elementos em cada estrato), aplicou-se a equação (7), deduzida para calcular o tamanho da amostra na intenção de estimar a média de uma população finita (COCHRAN, 1977):

$$n = \frac{Z_\alpha^2 \sigma^2 N}{d^2(N-1) + Z_\alpha^2 \sigma^2} \quad (7)$$

onde Z_α é a abcissa normal padrão para nível de significância α, σ^2 é a variância populacional, N é o tamanho da população e d é o erro amostral esperado. A determinação do tamanho da amostra representa a única variável que foi calculada por usina, usando como referências as médias individuais de cada série sintética. Assim, σ^2 se refere à variância das médias das séries sintéticas em cada usina e N é o total de séries geradas. Adotou-se como nível de significância 95%, enquanto que para o erro amostral foi considerado 2% do valor médio populacional, equivalente à média das médias das séries sintéticas de cada usina.

Com relação ao número de estratos, não é usual aplicar métodos que definam precisamente a quantidade a ser utilizada. O que se encontra na literatura são ensaios com números variados (arbitrários) de estratos para um mesmo conjunto de dados, sempre buscando a minimização da variância resultante (KESKINTÜRK; ER, 2007; HUDDLESTON; CLAYPOOL, 1970; KOZAK, 2004). No entanto, é fácil notar que a variância somente será, de fato, mínima quando o número de estratos for igual ao número de elementos da amostra (COCHRAN, 1977). Por esse motivo, a escolha deste parâmetro fica condicionada à particularidade de cada estudo; neste trabalho, fixou-se o número de estratos em cinco.

DADOS UTILIZADOS

Para aplicar os métodos propostos foram coletadas séries de vazões mensais afluentes às usinas hidrelétricas que compõem o Sistema Interligado Nacional (SIN). As séries estão disponíveis no *site* do Operador Nacional do Sistema Elétrico (ONS) (http://www.ons.org.br/operacao/vazoes_naturais.aspx) e seu histórico é atualizado anualmente. Na coleta de dados realizada para este trabalho, o período histórico disponível era de jan./1931 a dez./2007 para todas as usinas. Vale lembrar que os registros disponibilizados pelo ONS se referem a vazões naturalizadas, ou seja, sem a influência do barramento das usinas ou das evaporações líquidas nos lagos dos reservatórios, além dos possíveis usos secundários dos mesmos (BRAGA *et al.*, 2009). Todas as séries são consistidas e não apresentam falhas.

Em dez./2007 estavam em operação 146 aproveitamentos hidrelétricos, dos quais foram selecionadas 62 usinas para geração de séries sintéticas e amostragem. O critério de seleção se baseou na escolha de usinas localizadas nas cabeceiras das bacias hidrográficas e usinas cuja potência instalada fosse superior a 1 GW.

A Tabela1 mostra a relação de usinas consideradas, juntamente com a potência, vazão média de longo termo (MLT) e desvios padrão das afluências a cada reservatório.

Nota-se que a coleção de usinas escolhidas forma uma gama de características diferentes encontradas dentro do SIN. Tanto em termos de potência, quanto em termos de hidrologia, a maior usina considerada é a de Itaipu, enquanto que a menor é a de Jaguari. Entre esses limites, observa-se uma grande variabilidade de casos, escolhidos propositalmente para validar o método proposto neste artigo.

Previamente à aplicação do modelo estocástico para geração das séries sintéticas, a condição de estacionariedade estatística das séries históricas foi verificada. Todas foram submetidas a testes de hipóteses e a procedimentos de correção, caso apresentassem indícios de não estacionariedade (ver DETZEL *et al.*, 2011).

Estudo de caso

O estudo de caso foi dividido em duas etapas: a primeira se focou puramente na determinação das distâncias de Mahalanobis e subsequente amostragem das séries sintéticas de forma a verificar se esse procedimento alteraria significativamente a distribuição de probabilidades empíricas originalmente conseguidas com o modelo estocástico. A segunda etapa se focou em uma análise de sensibilidade do método quando submetido à geração de diversos conjuntos de cenários sintéticos com tamanhos de população variados. A intenção foi verificar a consistência e estabilidade das estratificações obtidas através do método proposto.

RESULTADOS E ANÁLISES

O modelo estocástico CARMA foi utilizado inicialmente para a geração de uma população de N=2000 cenários sintéticos de afluências às 62 usinas consideradas. As séries sintéticas foram submetidas a extensivos testes de validação, baseados em estatísticas básicas (média, desvio padrão, assimetria, autocorrelações, mínimos e máximos) e verificações mais criteriosas (sequências de vazões abaixo da média e déficits acumulados), além da análise das correlações espaciais. Os resultados foram positivos para

Tabela 1 – Usinas consideradas no estudo de caso

Usina	Potência (MW)	MLT (m^3/s)	Desv. Pad. (m^3/s)	Usina	Potência (MW)	MLT (m^3/s)	Desv. Pad. (m^3/s)
Camargos	46,0	132	84	**Paraibuna**	85,0	68	33
Furnas	1.216,0	924	615	**Picada**	50,0	37	21
Estreito	1.104,0	1.058	704	**Sobragi**	60,0	74	41
Caconde	80,4	54	36	**Salto Grande**	102,0	146	106
Agua Vermelha	1.396,2	2.093	1.306	**Candonga**	140,0	154	87
Batalha	53,6	113	84	**Baguari**	140,0	574	358
Emborcação	1.192,0	489	367	**Irapé**	360,0	152	176
Nova Ponte	510,0	300	200	**São Domingos**	48,0	127	25
Corumbá IV	127,0	133	95	**Itapebi**	450,0	391	440
Itumbiara	2.082,0	1.564	1.084	**Retiro Baixo**	82,0	158	127
Cach. Dourada	658,0	1.641	1.131	**Três Marias**	396,0	689	603
São Simão	1.710,0	2.404	1.567	**Queimado**	105,0	56	36
Ilha Solteira	3.444,0	5.588	3.278	**Sobradinho**	1.050,3	2.687	1.976
Porto Primavera	1.540,0	7.857	4.122	**Itaparica**	1.479,6	2.768	2.055
Jurumirim	97,8	251	146	**Comp. P. Afonso**	1.419,2	2.786	2.064
Itaipu	14.000,0	11.615	5.258	**Xingó**	3.162,0	2.786	2.064
Santa Clara PR	120,2	113	88	**Pedra do Cavalo**	162,0	104	158
Gov. Bento Munhoz	1.676,0	722	537	**Boa Esperança**	237,3	465	246
Segredo	1.260,0	852	627	**Guil. Amorim**	140,0	74	51
Salto Santiago	1.420,0	1.135	848	**Jauru**	121,5	85	16
Salto Osorio	1.078,0	1.188	888	**Guaporé**	120,0	39	9
Salto Caxias	1.240,0	1.524	1.135	**Salto Pilão**	182,3	138	109
Barra Grande	698,4	300	246	**Rosal**	55,0	33	24
Campos Novos	880,0	347	282	**Serra da Mesa**	1.275,0	778	697
Machadinho	1.140,0	814	654	**Lajeado**	902,5	2.467	2.293
Itá	1.450,0	1.149	941	**Curuá Una**	30,0	188	120
Passo Fundo	220,0	63	52	**Tucuruí**	8.370,0	11.003	9.241
Quebra Queixo	121,5	93	76	**Manso**	210,9	178	124
Castro Alves	130,0	176	150	**Ponte Pedra**	176,1	83	19
São José	51,0	280	235	**Santa Clara MG.**	60,0	99	86
Jaguari	27,6	26	13	**Itiquira I**	95,2	77	34

todas as usinas (DETZEL *et al.*, 2012), validando sua utilização na amostragem proposta.

Como descrito anteriormente, na primeira etapa do estudo de caso as distâncias de Mahalanobis foram determinadas entre a série histórica e cada matriz de séries sintéticas, resultando em 2000 distâncias. A Figura 1 mostra o histograma obtido para o grupo de distâncias. Nota-se que a distribuição de frequências resultante se aproxima de uma distribuição normal com uma leve assimetria positiva. Evidentemente, quanto maior a distância obtida, mais diferente é a série sintética da série histórica.

Na determinação do tamanho da amostra, obtiveram-se resultados com alguma variabilidade. Analisando a equação (7), nota-se que este parâmetro depende diretamente dos desvios padrão de cada série; usinas cujas afluências tenham desvio padrão elevado necessitam de maior número de amostras e vice-versa. Assim, considerando a população de 2000 séries geradas, a menor amostra calculada foi para a usina de São Domingos, com apenas 18 séries. Por outro lado, a maior amostra foi observada para a usina de Pedra do Cavalo com 766 séries. Considerando-se a distribuição do número de amostras entre todas as usinas, optou-se por adotar o valor padrão de n=300 séries para todas as usinas, número este próximo à mediana da distribuição obtida. A amostragem estratificada sobre todo o conjunto de dados gerou os resultados mostrados na Tabela 2.

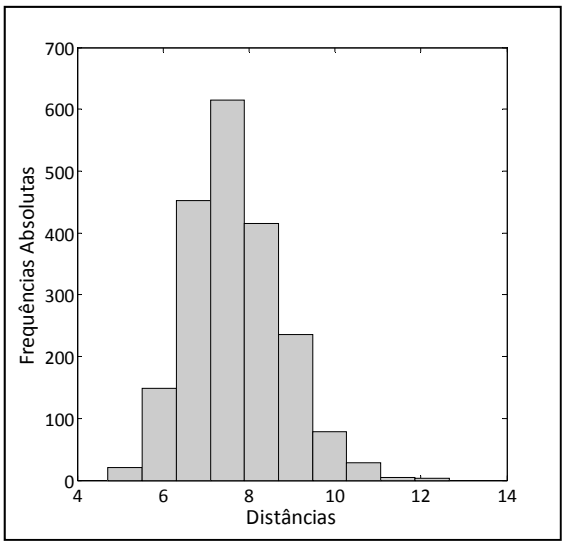

Figura 1 – Histograma para as 2000 distâncias

Tabela 2 – Parâmetros da amostragem estratificada

#	N_h	W_h	n_h	s_h^2
1	118	0,06	18	0,08
2	368	0,19	55	0,04
3	682	0,34	102	0,07
4	566	0,28	85	0,10
5	266	0,13	40	0,42
Total	**2000**	**1,00**	**300**	**$3{,}8\times10^{-4}$**

As amostras foram obtidas proporcionalmente ao peso amostral W_h, aplicando-o diretamente sobre o valor de n_h. Analisando as variâncias individuais, nota-se que o quinto estrato apresentou um valor discrepante em relação aos demais. Isso ocorreu por ser neste estrato que se encontram as séries cujas distâncias tiveram maior valor absoluto. Em outras palavras, o estrato cinco reuniu o conjunto de séries mais dissimilares em relação à série histórica. Ainda assim, a variância total da amostra estratificada resultou em um valor muito baixo, validando o método de estratificação utilizado.

Com relação à forma de seleção das séries dentro de cada estratificação, vale citar o estudo de Grafström (2010), no qual o autor avaliou oito métodos de diferentes complexidades para seleção de amostras a partir dos pesos amostrais obtidos. Utilizando a entropia como medida de aleatoriedade, a conclusão foi a de que todos os oito métodos são igualmente precisos. Baseado nesta evidência optou-se por um método de simples permutação aleatória (COCHRAN, 1977) das matrizes de séries sintéticas em cada estrato.

Uma vez compostas as amostras, elas foram submetidas às análises de distribuição empírica de probabilidades. A Figura 2 mostra a comparação entre as funções de distribuição de probabilidades empíricas acumuladas (FDA) da série histórica, séries sintéticas sem amostragem (N=2000) e séries sintéticas com amostragem (n=300) para quatro usinas: Itaipu, representando a hidrologia do Sudeste do país, Segredo, representando a hidrologia do Sul do país, Sobradinho, representando a hidrologia do Nordeste do país e Pedra do Cavalo, usina cujo tamanho teórico da amostra foi o maior calculado.

Percebe-se que a amostragem não afetou significativamente as funções empíricas em nenhum dos casos. As maiores diferenças foram observadas na porção superior, relativas ao intervalo entre 0,8 e 1,0, no qual as séries amostradas ligeiramente subestimaram as vazões de maior magnitude. O resultado positivo foi verificado inclusive para a usina de Pedra do Cavalo que, como mencionado anteriormente, necessitava de uma amostra composta por 766 séries. Mesmo utilizando menos da metade do número calculado, as amostras selecionadas foram capazes de manter a distribuição de probabilidades empíricas. Estes resultados foram também verificados para as demais usinas do estudo.

Na segunda etapa de verificações, foram gerados cinco conjuntos de cenários, com populações de 2.000, 3.000, 4.000, 5.000 e 10.000 elementos cada. Os histogramas das distâncias de Mahalanobis obtidos para cada população são mostrados na Figura 3. Analisando-os fica evidente que a forma geral da distribuição de frequências não varia de acordo com o aumento do tamanho das populações. Dessa forma, a classe de maior frequência em todos os casos se localiza no entorno do ponto de distância igual a sete.

A causa da não variabilidade do histograma final pode ser encontrada no próprio modelo estocástico de geração de séries sintéticas. Os parâmetros da formulação foram todos estimados utilizando unicamente as informações das séries históricas e, portanto, espera-se que o conjunto de séries sintéticas mantenha um comportamento parecido independentemente do número de cenários gerados. Do ponto de vista do método de agrupamento das séries sintéticas, o fato dos histogramas finais manterem sua forma é desejável, pois confirma a consistência da técnica para este caso.

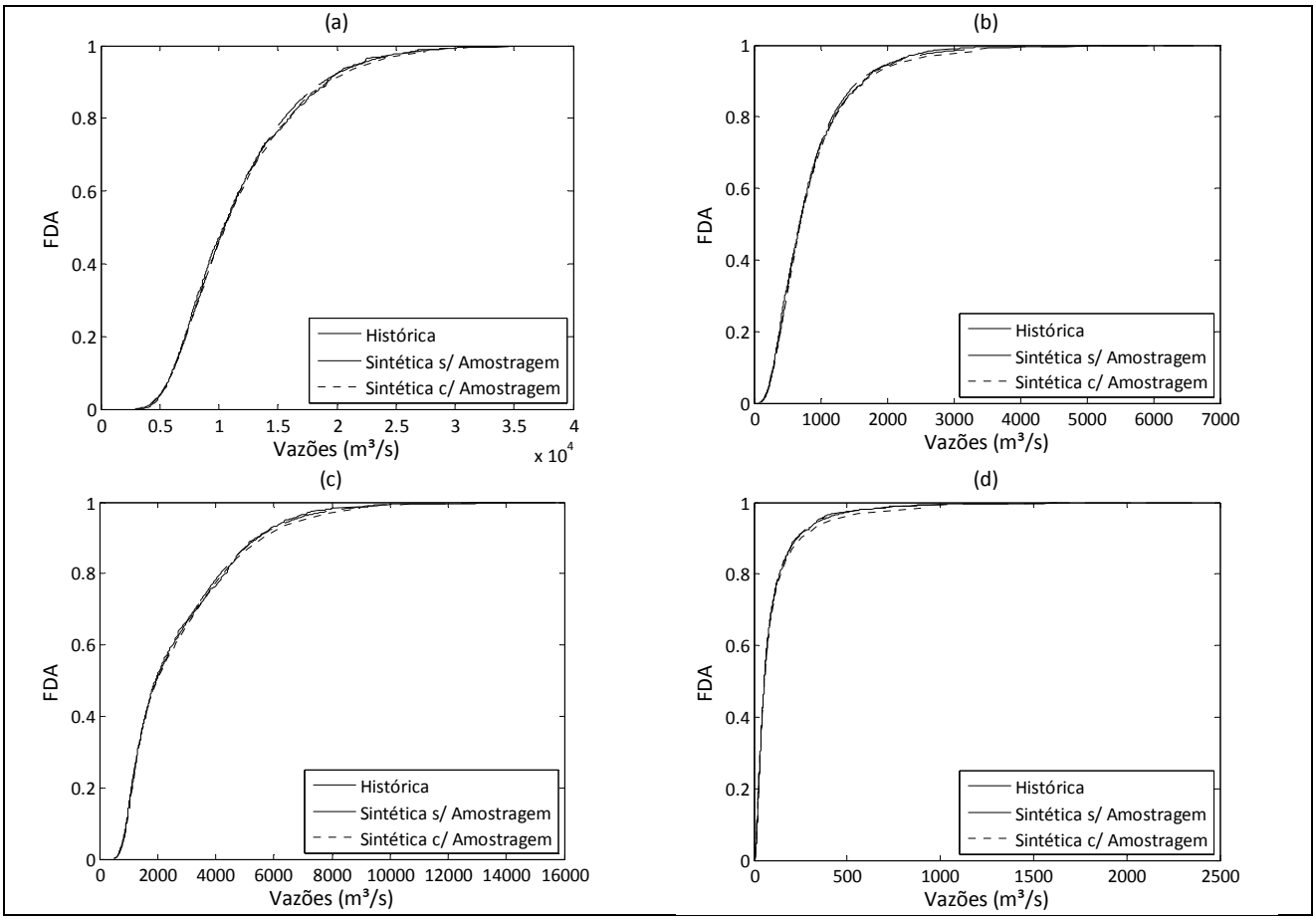

Figura 2 – Distribuições de probabilidades empíricas acumuladas. (a) Itaipu; (b) Segredo; (c) Sobradinho; (d) Pedra do Cavalo

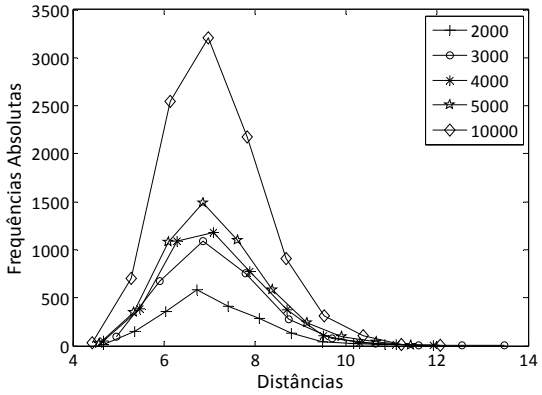

Figura 3 – Histograma de distâncias para diversos tamanhos de população

Para o processo de estratificação, notou-se que o aumento progressivo do tamanho da população não resultou em um aumento significativo dos tamanhos das amostras necessárias (255, 257, 310, 307 e 285 elementos, respectivos às populações de 2.000, 3.000, 4.000, 5.000 e 10.000 distâncias). Isso ocorreu devido a pouca variação do desvio padrão das populações, fato cuja justificativa é a mesma dada no parágrafo anterior. Assim, foi mantido o tamanho da amostra em 300 indivíduos, para os quais a estratificação resultou nos parâmetros apresentados na Tabela 3.

Como esperado, na estratificação da população com 2000 elementos os resultados foram semelhantes aos mostrados na primeira etapa de validação. Entretanto, para as demais populações testadas observa-se um fato interessante: o estrato três resultou em menos séries do que seus estratos vizinhos, ocorrência contrária à população de 2.000 elementos. Analisando novamente a Figura 3, percebe-se uma sutil diferença nos picos dos histogramas, condizente com a distribuição geral dos pesos dos estratos resultantes. O histograma da população

de 4.000 elementos, por exemplo, apresenta seu pico mais deslocado do que os demais, fato que refletiu em uma maior diferença entre os pesos do terceiro e quarto estratos para esta população na Tabela 3. Ainda assim, para todos os casos não foram observadas diferenças significativas em termos de variância.

Tabela 3 – Parâmetros da estratificação para diversos tamanhos de população

N	#	N_h	W_h	n_h	s_h^2
	1	172	0,09	27	0,07
	2	385	0,19	57	0,03
2.000	3	596	0,30	90	0,05
	4	503	0,25	75	0,07
	5	344	0,17	51	0,63
	1	214	0,07	21	0,11
	2	755	0,25	75	0,07
3.000	3	614	0,20	60	0,02
	4	977	0,33	99	0,12
	5	440	0,15	45	0,54
	1	337	0,09	27	0,10
	2	855	0,21	63	0,05
4.000	3	633	0,16	48	0,02
	4	1333	0,33	99	0,10
	5	842	0,21	63	0,56
	1	406	0,07	21	0,07
	2	1192	0,24	72	0,04
5.000	3	884	0,18	54	0,01
	4	1632	0,33	99	0,07
	5	886	0,18	54	0,41
	1	569	0,06	18	0,07
	2	2453	0,25	75	0,06
10.000	3	1836	0,18	54	0,02
	4	3536	0,35	105	0,10
	5	1606	0,16	48	0,52

O fato comentado no parágrafo anterior serve como evidência da relação intrínseca entre as séries de distâncias e a estratificação de cada população. Hipoteticamente, se as séries sintéticas geradas fossem todas extremamente semelhantes às séries históricas, o histograma resultante teria todas as frequências absolutas localizadas perto da distância igual à zero. Nesse caso, a estratificação sobre tal população não surtiria o efeito desejado, pois os primeiros estratos reuniriam grande parte da amostra. Para os ensaios mostrados neste artigo, observou-se uma grande variabilidade nas características das séries sintéticas quando comparadas às séries históricas. Dessa maneira pode-se considerar que o método de separação das séries e subsequente amostragem estratificada torna possível a seleção de amostras representativas de diferentes eventos hidrológicos, mesmo utilizando um número reduzido de séries sintéticas.

CONCLUSÃO

Apresentou-se neste trabalho um método para amostragem de séries hidrológicas sintéticas geradas através de um modelo estocástico, na intenção de reduzir o número total de cenários em estudos que exigem grande processamento computacional. O modelo utilizado para geração das séries sintéticas foi do tipo autorregressivo com médias móveis contemporâneo, aplicado às afluências de 62 usinas hidrelétricas que operam no SIN.

O processo de amostragem foi feito em duas etapas: (i) agrupamento das séries sintéticas através da determinação de distâncias entre elas e as séries históricas e (ii) aplicação de técnicas de amostragem estratificada clássica sobre o conjunto de distâncias resultante. É importante ressaltar que as séries sintéticas geradas são multivariadas e, portanto, possuem correlação espacial significativa. Por esse motivo, a métrica utilizada para a determinação das distâncias foi a de Mahalanobis, que contempla a matriz de covariâncias em sua formulação. Com essa técnica, foi possível resumir a população de matrizes de séries sintéticas em uma população de distâncias, simplificando a etapa de amostragem estratificada.

Os resultados apresentados foram extremamente positivos, mostrando ser possível a utilização de um número menor de séries sintéticas sem prejudicar a estrutura de correlações espaciais ou a distribuição de probabilidades empíricas obtidas com o modelo estocástico. Desta maneira, o método encontra aplicabilidade em estudos diversos como, por exemplo, trabalhos focados na otimização estocástica da operação de múltiplos reservatórios, reconhecidamente onerosos do ponto de vista computacional (KELMAN *et al.*, 1990; PEREIRA & PINTO, 1985; SRIFI & HIPEL, 2001; SAADOULI, 2010; FABER & STEDINGER, 2001).

Ainda assim, recomenda-se cautela na determinação das distâncias de Mahalanobis. Inconsistências nos dados utilizados podem levar à condição de singularidade da matriz de covariâncias, na qual não é possível a determinação de sua inversa e, consequentemente, das distâncias. Essa ocorrência é verificada em casos nos quais os dados apresentam elevado grau de colinearidade (De MAESSCHALCK *et al.*, 2000; JOUAN-RIMBAUD *et al.*, 1998). Para as

séries de vazões, por exemplo, a colinearidade pode ser consequência de métodos de regressão eventualmente aplicados entre os diferentes postos de medição, na intenção de corrigir falhas ou estender séries de menor duração. Ainda que esta condição não tenha sido observada no presente estudo, podem ser aplicadas técnicas de análise de componentes principais sobre os dados originais, visando a escolha de variáveis significativas para a determinação das distâncias.

AGRADECIMENTOS

Esta pesquisa/trabalho foi possível graças ao financiamento da ANEEL através do Projeto Estratégico de Pesquisa e Desenvolvimento – ANEEL PE-6491-0108/2009, "Otimização do Despacho Hidrotérmico", com o apoio das seguintes concessionárias: COPEL, DUKE, CGTF, CDSA, BAESA, ENERCAN, CPFL PAULISTA, CPFL, PIRATININGA, RGE, AES TIETÊ, AES URUGUAIANA, ELETROPAULO, CEMIG e CESP.

REFERÊNCIAS

ATKINSON, A. C.; RIANI, M. Exploratory tools for clustering multivariate data. *Computational Statistics & Data Analysis*, v. 52, p. 272-285, 2007.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. *Time Series Analysis*: Forecasting and Control. 3 ed. New Jersey: Prentice Hall, 1994.

BRETTHAUER, K. M.; ROSS, A.; SHETTY, B. Nonlinear integer programming for optimal allocation in stratified sampling. *European Journal of Operational Research,* v. 116, p. 667-680, 1999.

BRAGA, R. S.; ROCHA, V. F.; GONTIJO, E. A. Revisão das séries de vazões naturais nas principais bacias hidrográficas do sistema interligado nacional. In.: ANAIS DO XVI SIMPÓSIO BRASILEIRO DE RECURSOS HÍDRICOS. Campo Grande: ABRH, 2009.

CAIADO, J.; CRATO, N.; PEÑA, D. A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis*, v. 50, n. 10, p. 2668-2684, 2006.

CAMACHO, F.; McLEOD, A. I.; HIPEL, K. W. Multivariate contemporaneous ARMA model with hydrological applications. *Stochastic Hydrology and Hydraulics*, v. 1, p. 141-154, 1987.

CHADDHA, R. L.; HARDGRAVE, W. W.; HUDSON, D. J.; SEGAL, M.; SUURBALLE, J. W. Allocation of total sample size when only the stratum means are of interest. *Technometrics*, v. 13, p. 817-816, 1971.

CHANG, C-C. A boosting approach for supervised Mahalanobis distance metric learning. *Pattern Recognition*, v. 45, p. 844-862, 2012.

COCHRAN, W.G. *Sampling Techniques*. 3 ed. New York: John Wiley & Sons Inc., 1977.

CORDUAS, M. Clustering streamflow time series for regional classification. *Journal of Hydrology*, v. 407, n. 1-4, p. 73-80, 2011.

CORDUAS, M.; PICCOLO, D. Time series clustering and classification by the autoregressive metric. *Computational Statistics & Data Analysis*, v. 52, n. 4, p. 1860-1872, 2008.

DALENIUS, T.; HODGES, J. L. Minimum variance stratification. *Journal of the American Statistical Association*, v. 54, n. 285, p. 88-101, 1959.

DE MAESSCHALCK, R.; JOUAN-RIMBAUD, D.; MASSART, D. L. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, v. 50, n. 1, p. 1-18, 2000.

DETZEL, D. H. M.; BESSA, M. R.; VALLEJOS, C. A. V.; SANTOS, A. B.; THOMSEN, L. S.; MINE, M. R. M.; BLOOT, M. L.; ESTRÓCIO, J. P. Estacionariedade das Afluências às Usinas Hidrelétricas Brasileiras. *Revista Brasileira de Recursos Hídricos*, v. 16, n. 3, p. 95-111, 2011.

DETZEL, D. H. M.; MINE, M. R. M.; BESSA, M. R. Cenários Sintéticos de Vazões Para Grandes Sistemas Hídricos Através de Modelos Contemporâneos e Amostragem. Manuscrito submetido à Revista Brasileira de Recursos Hídricos, dez./2012.

EKMAN, G. An approximation useful in univariate stratification. *Annals of Mathematical Statistics*, v. 30, n. 1, p. 219-229, 1959.

FABER, B. A.; STEDINGER, J. R. Reservoir optimization using sampling SDP with ensamble streamflow prediction (ESP) forecasts. *Journal of Hydrology*, v.

249, p. 113-133, 2001.

FARBER, O.; KADMON, R. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. *Ecological Modelling*, v. 160, n. 1-2, p. 115-130, 2003.

FERREIRA, D. F. *Estatística Multivariada*. Lavras: Ed. UFLA, 2004, 662 p.

FILZMOSER, P.; HRON, K. Outlier detection for compositional data using robust methods. *Mathematical Geoscience*, v. 40, p. 233-248, 2008.

FREVERT, B. D. K.; COWAN, M. S., LANE, W. L. Use of stochastic hydrology in reservoir operation. *Journal of Irrigation and Drainage Engineer*ing, v. 115, n. 3, p. 334-343, 1989.

FRICK, D. M.; BODE, D.; SALAS, J. D. Effect of drought on urban water supplies: I. Drought analysis. Journal of Hydraulic. *Engineering*, v. 116, n. 6, p. 733-753, 1990.

GIMÉNEZ, E.; CRESPI, M.; GARRIDO, M. S.; GIL, A. J. Multivariate outlier detection based on robust computation of Mahalanobis distances. Application to positioning assisted by RTK GNSS Networks. *International Journal of Applied Earth Observation and Geoinformation*, v. 16, p. 94-100, 2012.

GRAFSTRÖM, A. Entropy of unequal probability sampling designs. *Statistical Methodology*, v. 7, n. 2, p. 84-97, 2010.

GUNNING, P.; HORGAN, J. M. A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology*, v. 30, n. 2, p. 159-166, 2004.

HALTINER, J. P., SALAS, J. D. Development and testing of a multivariate, seasonal ARMA(1,1) model. *Journal of Hydrology*, v. 104, p. 247-272, 1988.

HIPEL, K. W.; McLEOD, A. I. Time series modelling of water resources and environmental systems, 1994. Disponível em: http://www.stats.uwo.ca/ faculty/aim/1994Book/. Acesso em: 12/11/12.

HUDDLESTON, H. F., CLAYPOOL, P. L., HOCKING R. R. Optimal sample allocation to strata using convex programming. *Journal of the Royal Statistical Society*: Series C, v. 19, n. 3, p. 273-278, 1970.

JACKSON, B. B. The use of streamflow models in planning. *Water Resources Research*, v. 11, n. 1, p. 54-63, 1975.

JOUAN-RIMBAUD, D., MASSART, D. L., SABY, C. A., PUEL, C. Determination of the representativity between two multidimensional data sets by a comparison of their structure. *Chemometrics and Intelligent Laboratory Systems*, v. 40, n.2, p. 129-144, 1998.

KALPAKIS ,K., GADA, D., PUTTAGUNTA, V. Distance measures for effective clustering of ARIMA time-series, Proceedings of the 2001 IEEE International Conference on Data Mining, San Jose, CA, November 29–December 2, 2001, p. 273–280.

KELMAN, J., STEDINGER, J., COOPER, L., HSU, E., YUAN, S.-Q. Sampling stochastic dynamic programming applied to reservoir operation. *Water Resources Research*, v. 26, n. 3, p. 447–454, 1990.

KESKINTÜRK, T., ER, S. A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. *Computational Statistics & Data Analysis*, v. 52, n. 1, p. 53-67, 2007.

KOŠMELJ, K., BATAGELJ, V. Cross-sectional approach for clustering time varying data. *Journal of Classification*, v. 7, p. 99–109, 1990.

KOZAK, M. Optimal Stratification using random search method in agricultural surveys. *Statistics in Transition*, v. 6, n. 5, p. 797-806, 2004.

LABADIE, J. W. Optimal Operation of Multireservoir Systems : State-of-the-Art Review. *Journal of Water* Resources Planning and Management, v. 130, n. 2, p. 93-111, 2004.

LIAO, T. W. Clustering of time series data—a survey. Pattern Recognition, v. 38, n. 11, p. 1857-1874, 2005.

MANOLOVA, A.; GUÉRIN-DUGUÉ, A. Classification of dissimilarity data with a new flexible Mahalanobis-like metric. *Pattern Analysis and Applications*, v. 11, p. 337-351, 2008.

McLACHLAN, G. J. Mahalanobis Distance. Resonance, p. 20-26, jun. 1999.

MORENO, M. A.; PLANELLS, P.; ORTEGA, J. F.; TARJUELO, J. New Methodology to Evaluate Flow Rates in On-Demand Irrigation Networks. *Journal of Irrigation and Drainage Engineering*, v. 113, p. 298-306,

2008.

NICOLINI, G. *A method to define strata boundaries. Departmental Working Paper*, 2001–01. Department of Economics University of Milan, Italy, 2001.

PEREIRA, M. V. F.; PINTO, L. M. V. G. Stochastic Optimization of a Multireservior Hydroelectric System: A Decomposition Approach. *Water Resources Research*, v. 21, n. 6, p. 779-792, 1985.

PICARD, N. A Criterion Based on the Mahalanobis Distance for Cluster Analysis with Subsampling. *Jounral of Classification*, v. 49, p. 23-49, 2012.

PICCOLO, D. A distance measure for classifying ARMA models. *Journal of Time Series Analysis*, v. 11, n. 2, p. 153–163, 1990.

POLICKER, S.; GEVA, A.B. Nonstationary time series analysis by temporal clustering, IEEE Transactions on Systems, Man, and Cybernetics – Part B: *Cybernetics*, v. 30, n. 2, p. 339–343, 2000.

SAADOULI, N. Computationally efficient solution algorithm for a large scale stochastic dynamic program. *Procedia Computer Science*, v. 1, n. 1, p. 1397-1405, 2010.

SCHWARTZ, G. Estimating the dimension of a model. *Annals of Mathematical Statistics*, v. 6, n. 2, p. 461-464, 1978.

SEIFI, A., HIPEL, K. Interior-point method for reservoir operation with stochastic inflows. *Journal of Water Resources Planning and Management*, v. 127, n. 1, p. 48–57, 2001.

STEDINGER, J. R.; LETTENMAIER, D. P.; VOGEL, R. M. Multisite ARMA(1,1) and disaggregation models for annual streamflow generation. *Water Resources Research*, v. 21, n. 4, p. 497-509, 1985.

VOGEL, R. M.; STEDINGER, J. R. The value of stochastic streamflow models in overyear reservoir design applications. *Water Resources Research*, v. 24, n. 9, p. 1483-1490, 1988.

WANG, Q. J. A Bayesian method for multi-site stochastic data generation: dealing with non-concurrent and missing data, variable transformation and parameter uncertainty. *Environmental Modelling & Software*, v. 23, p. 412-421, 2008.

XIANG, S., NIE, F., ZHANG, C. Learning a Mahalanobis distance metric for data clustering and classification. *Pattern Recognition*, v. 41, n. 12, p. 3600-3612, 2008.

Sampling Synthetic Hydrologic Series

ABSTRACT

Application of stochastic models for generating synthetic time series is widely accepted in several areas where the use of historical information is limited, including water resources planning and management. Depending on its application, however, employing all synthetic scenarios could prove computationally burdensome, forcing simplifications of the systems used. In this context, a synthetic series sampling study is proposed to reduce the number of scenarios generated without losing the representation obtained with the stochastic model. The method is based on two steps: (i) synthetic series grouped by determining Mahalanobis distances between them and the original time series and (ii) stratified sampling application on the resulting set. As a case study, streamflow series for 62 hydroelectric plants in Brazil were selected, whose synthetic series were generated from a contemporary multivariate autoregressive model CARMA (p, q). The results confirm the sampling method plausibility, allowing reducing the number of scenarios without changing the empirical probability distribution achieved with the set of synthetic series originally generated. Extra analyses regarding the method stability towards the number of generated series are presented.
Keywords: *Sampling. Mahalanobis distance. Contemporary model. Monthly synthetic series.*