Preenchimento de Dados Limnimétricos Horários Via Modelos ARIMA

Daniel H. Marco Detzel[1], Ana Paula Oening[1], Angelo R. R. de Souza[1], Sérgio L. C. Cerminaro[2]
daniel@lactec.org.br; ana.oening@lactec.org.br; angelorrdesouza@gmail.com; sergio.cerminaro@cesp.com.br

Recebido: 06/06/13 - revisado: 26/06/13 - aceito: 27/08/13

RESUMO

O uso de dados hidrológicos é primordial em qualquer instância do planejamento e gerenciamento dos recursos hídricos de uma determinada localidade. Acompanhado da obtenção desses dados, é indispensável a execução de uma análise de consistência, a fim de detectar erros ou períodos com falhas em registros. O presente trabalho foca em dados limnimétricos horários, apresentando-se primeiramente um método simples para eliminação de dados errôneos baseado em técnicas de detecção de outliers. Na sequência, mostra-se o preenchimento de períodos com falhas nos dados a partir de modelos autorregressivos integrados de médias móveis (ARIMA) e conceitos de previsão de séries temporais. Faz-se uso de previsões forward-backward com subsequente ponderação das estimativas. Como estudo de caso, utilizaram-se limnigramas referentes a duas estações de monitoramento instaladas na bacia hidrográfica rio Paraná. Os resultados obtidos foram muito bons, com erros médios percentuais absolutos (MAPEs) baixos para as situações exibidas. Destaca-se a independência da metodologia apresentada para com o uso de dados externos, procedimento comum à tradicional análise de consistência.

Palavras-chave: *Análise de consistência. Dados limnimétricos. Modelos ARIMA. Escala horária*

INTRODUÇÃO

O uso de dados hidrológicos é primordial em qualquer instância do planejamento e gerenciamento dos recursos hídricos de uma determinada localidade. A partir deles, usualmente organizados em formato de séries temporais, é possível estimar variáveis relevantes para a elaboração de projetos de pequena e grande escala. Sua importância é reconhecida, inclusive, na Legislação do país, a qual aponta o Sistema de Informações de Recursos Hídricos como um instrumento de gestão da Política Nacional de Recursos Hídricos (Lei Federal 9433/97).

A medição de variáveis relacionadas com a água na natureza está inserida dentro das atividades da ciência hidrométrica. Em específico, registros limnimétricos se referem às medidas dos níveis de água, sejam eles de rios ou de reservatórios. Dentre suas diversas aplicações, destacam-se a obtenção de dados de vazões de rios através de curvas-chave e o monitoramento dos níveis de reservatórios. Tradicionalmente, as observações eram coletadas através de réguas limnimétricas instaladas às margens do corpo hídrico. Atualmente, muitos postos contam com medições automáticas através de limnígrafos, capazes de gravar informações contínuas e enviá-las via satélite.

Evidentemente a aquisição dos dados limnimétricos não está isenta de erros e, portanto, necessita ser submetida a análises de consistência. O uso dos limnígrafos praticamente elimina os erros grosseiros, principalmente por não depender de leituristas. Ainda assim, esses aparelhos não estão isentos de erros sistemáticos e fortuitos que podem provocar medições equivocadas ou interrupções nos registros. Encontra-se na literatura uma rotina clássica para analisar a consistência de dados fluviométricos em geral, que contempla desde visitas técnicas nas estações de monitoramento até avaliação conjunta com postos vizinhos (SANTOS *et al.*, 2001; Agência Nacional de Águas, 2011; World Meteorological Organization, 2008). Faz-se uso de curvas duplo-acumulativas para detecção de dados errôneos e regressões para suas correções. No caso de zeros nas séries, a análise é mais criteriosa e pode incluir dados de precipitação na bacia hidrográfica em questão. O uso de regressões entre postos vizinhos para o preenchimento dos vazios também é comum, principalmente se estes postos estão locados em um mesmo rio (SANTOS *et al.*, 2001).

[1] - Instituto de Tecnologia para o Desenvolvimento – LACTEC
[2] - Companhia Energética de São Paulo – CESP

No entanto em muitas situações essas técnicas de preenchimento de dados faltosos (ou errôneos) não são satisfatórias, fato que motivou alguns estudos mais aprofundados acerca do tema. Dos métodos propostos, observa-se uma divisão entre dois grupos: os que não possuem informações extras além da própria série em análise e os que contam com séries de postos vizinhos ou de outros dados hidrológicos pertinentes. No primeiro caso, frequentemente classificado como univariado, aplicam-se técnicas derivadas de previsões de séries temporais, tais como redes neurais artificiais (RNA) (KHALIL *et al.*, 2001; ELSHORBAGY *et al.*, 2002; KIM e AHN, 2008; STARRETT *et al.*, 2008) e modelos estocásticos lineares (BENNIS *et al.*, 1997; QUEVEDO *et al.*, 2010). Elshorbagy *et al.* (2000) utilizam-se dessas técnicas aplicadas a grupos com características hidrológicas semelhantes, identificados através de reconhecimento de padrões sobre a série mensal de vazões de um rio. Na comparação, os autores apontam melhor desempenho para as RNA. Em outras ocasiões, nas quais as variáveis não possuem tanta variabilidade, técnicas simples como interpolações lineares ou médias ponderadas são utilizadas (DINPASHOH *et al.*, 2011; YAWSON *et al.*, 2012).

Para o caso multivariado, o número de trabalhos na literatura é consideravelmente maior, sendo comum o uso de interpolações espaciais (HUGHES e SMAKHTIN, 1996; KUMAMBALA, 2010) e regressões múltiplas (SIMONOVIC, 1995; ABATZOGLOU *et al.*, 2009), além das já citadas RNA (ELSHORBAGY *et al.*, 2000; KIM e PACHEPSKY, 2010). Dastorani *et al.* (2010) também utilizam RNA para o preenchimento de séries mensais de vazão, comparando-as com uma técnica baseada em lógica neuro-fuzzy (ANFIS). Apesar de reforçar que RNA são boas técnicas para o preenchimento de dados faltantes, melhores estimativas foram fornecidas pela lógica ANFIS.

Encontram-se também estudos que fazem uso de outras variáveis hidrológicas, tais como precipitação e evaporação. Gyau-Boakye e Schultz (1994), por exemplo, as aplicam no preenchimento de dados diários de vazão, comparando o desempenho de oito técnicas. Dentre elas estão alguns modelo físicos, o que justifica o uso das informações extras.

Recentemente outras técnicas têm sido propostas com resultados promissores. Firat *et al.* (2012) aplicam o algoritmo EM (do inglês *Expectation-Maximization*) para o preenchimento de séries de temperatura em escala mensal. Trata-se de uma técnica iterativa na qual os valores esperados dos dados faltantes são estimados (fase E) e a função de log-verossimilhança formada pelas estimativas é maximizada (fase M). As estimativas iniciais são feitas com informações de postos vizinhos. Mwale *et al.* (2012), por sua vez, aplicam os conceitos de mapas auto organizáveis sobre séries hidrológicas diárias de níveis de rios, vazões e chuvas na intenção de identificar agrupamentos com características semelhantes, fornecendo subsídios para o preenchimento de períodos faltantes.

Todavia, os estudos citados nos parágrafos anteriores se limitam a escalas mensais e diárias. A literatura relacionada ao tema praticamente não contempla casos de análise de consistência e preenchimento de dados em escala horária. É nesse contexto que se apresenta este artigo, focado no preenchimento de falhas observadas em limnigramas, trabalhando-se com cotas médias hora a hora. Devido à impossibilidade de uso de informações adicionais e séries de postos vizinhos, utilizam-se modelos autorregressivos integrados de médias móveis (ARIMA) e conceitos comuns à previsão de séries temporais.

O restante do trabalho está organizado em três seções: a primeira apresenta a área de estudo e o método aplicado, com destaque à explicação de conceitos fundamentais da modelagem ARIMA. A segunda seção mostra os resultados e principais comentários pertinentes. Por fim, a última seção conclui o artigo.

MATERIAIS E MÉTODOS

Área de estudo e dados utilizados

A metodologia proposta neste trabalho foi aplicada a dados limnimétricos de rios localizados na bacia hidrográfica do rio Paraná, em sua porção norte. Referem-se a duas bobinas completas, respectivas aos postos de Fazenda Bálsamo (FB), no rio Aporé (afluente do rio Paranaíba) e Fazenda São Sebastião (FSS), no rio Sucuriú (afluente do rio Paraná) e cuja localização geográfica é mostrada no croqui da Figura 1.

O rio Aporé marca a divisa entre os estados de Goiás e Mato Grosso do Sul, enquanto que o rio Sucuriú se localiza inteiramente no estado do Mato Grosso do Sul. São considerados os dois principais afluentes da região do Alto Paraná, formando o Complexo Aporé-Sucuriú (PAGOTTO e SOUZA, 2006). Essa área pertence completamente ao Planalto da Bacia Sedimentar do Paraná, sendo marcada por altitudes médias entre 500 e 750 metros e relevo de chapadas, típicas do bioma cerrado. Os postos

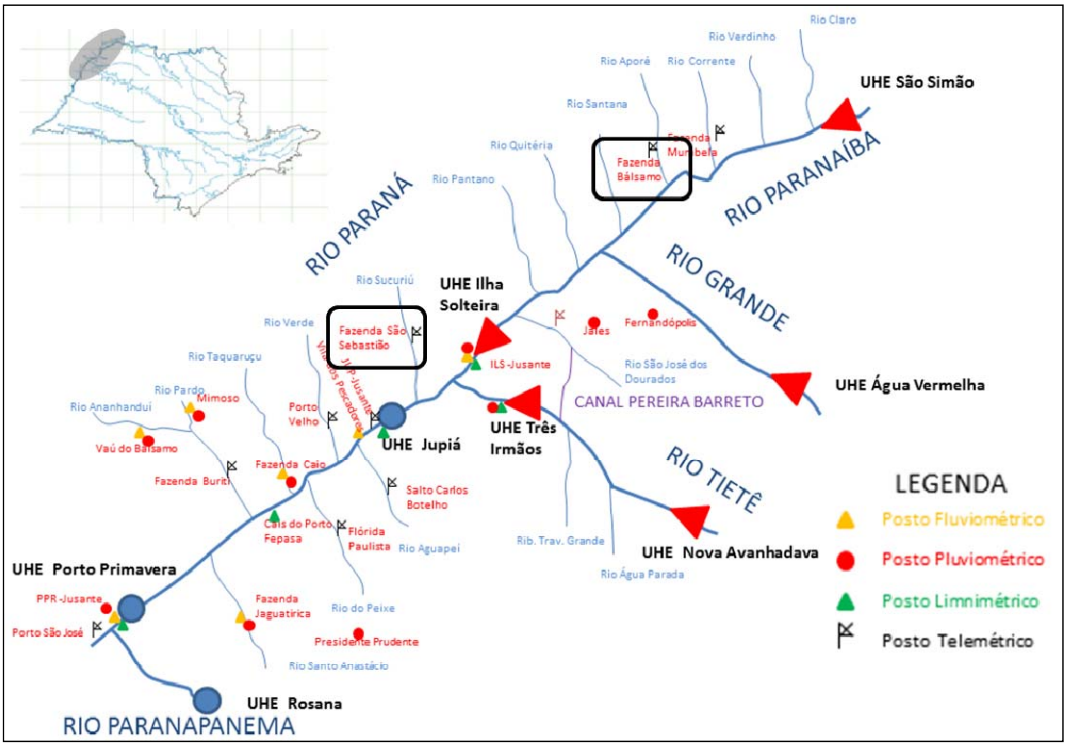

Figura 1 – Localização da área de estudo, com destaque aos postos limnimétricos utilizados

limnimétricos FB e FSS estão instalados nas proximidades da foz dos respectivos rios, onde encontram os reservatórios das usinas de Ilha Solteira e Jupiá, respectivamente.

Dado que as citadas usinas hidrelétricas são de grande relevância para o Sistema Elétrico Brasileiro, consideram-se essas estações de monitoramento de elevada importância. As bobinas limnimétricas foram cedidas pela Companhia Energética de São Paulo (CESP) para a realização de um projeto de Pesquisa e Desenvolvimento visando a digitalização e recuperação de dados históricos. As séries foram obtidas a partir de um algoritmo de reconhecimento de imagens elaborado no âmbito do projeto (leitores interessados no tema podem consultar Gonzalez e Woods (1992), que forneceu registros com escala de quatro minutos e precisão máxima de 0,1 milímetros para as cotas. Seguindo a exigência da resolução conjunta ANEEL/ANA nº 3 de agosto de 2010 e o padrão da própria CESP, transformaram-se os dados para a escala horária a partir da média dos registros contidos neste intervalo. A precisão numérica foi mantida a mesma.

Ao final, obtiveram-se séries horárias com 4350 registros para FSS, referentes ao período entre 16/06/2010 e 17/12/2010 e 2234 registros para FB, referentes ao período entre 14/06/2000 a 19/09/2000. Apesar das séries serem formadas por apenas seis (FSS) e três (FB) meses, ressalta-se que a escala considerada no trabalho é horária. Dessa forma, em termos estatísticos, consegue-se garantir amostras de tamanho representativo (superiores a 100 observações, Box *et al.*, 1994, p. 17) para os ajustes do modelo.

A fim de colocar em prática a metodologia proposta, escolheram-se períodos diversos das séries digitalizadas e induziram-se erros. Maiores detalhes são comentados nos próximos itens.

Detecção e erros nas séries

Antes do preenchimento das eventuais falhas contidas nos registros, aplicou-se um procedimento simples para auxílio na detecção automática de erros. Este passo é importante principalmente devido ao fato de que as séries são provenientes de digitalizações, que podem gerar algum tipo de equívoco.

O procedimento adotado se baseia na detecção de *outliers*, ou seja, valores inconsistentes com os demais registros da série. Aplicaram-se dois métodos simples, cuja eficiência foi comprovada em

Nascimento *et al.* (2012): Z-Score e Z-Score Modificado.

Os métodos se fundamentam na distribuição normal. O Z-Score é simplesmente o cálculo da variável normal padrão para todos os registros, como mostra a equação (1):

$$ZS_t = \frac{z_t - \bar{z}}{s_z} \quad (1)$$

onde z_t é a observação no instante t $(t = 1,2,\dots,n)$ e $\bar{z}$ e s_z são, respectivamente, os estimadores amostrais da média e desvio padrão. Dessa maneira, são calculados tantos ZS quanto for o tamanho da série de dados limnimétricos. Como regra, considera-se outlier a observação com ZS (em módulo) maior que 3 (NASCIMENTO *et al.*, 2012).

Contudo, se uma série contém muitos *outliers*, eles afetarão o cálculo dos parâmetros de média, desvio padrão e, consequentemente, a determinação do ZS. Por esse motivo utilizou-se também o método do Z-Score Modificado, que substitui a média pela mediana amostral e o desvio padrão pelo o desvio absoluto da mediana amostral (MAD) (IGLEWICZ e HOAGLIN, 1993). O MAD é calculado pela equação (2):

$$MAD = \text{mediana}(|z_t - \tilde{z}|) \quad (2)$$

onde $\tilde{z}$ é a mediana amostral da série. O Z-Score Modificado é, então, calculado pela fórmula (3):

$$ZSM_t = \frac{0{,}6745(z_t - \tilde{z})}{MAD} \quad (3)$$

Para este método, consideram-se *outliers* as observações cujo ZSM (em módulo) sejam superiores a 3,5 (IGLEWICZ e HOAGLIN, 1993). Por convenção, uma cota suspeita somente é retirada da série quando for identificado por ambos os métodos.

É importante frisar que o emprego das técnicas de detecção de *outliers* visa somente dar suporte à metodologia, visto que as séries são bastante longas. Se a intenção for executar uma detecção de *outliers* mais robusta, recomenda-se consulta a métodos que considerem a autocorrelação serial dos registros (e.g. BASU; MECKESHEIMER, 2007).

Preenchimento das falhas e os modelos ARIMA

Depois de passar pela detecção dos *outliers*, as séries resultantes são submetidas ao procedimento de preenchimento de falhas e correção dos erros. Elshorbagy *et al.* (2000) fornecem uma classificação das falhas comumente encontradas em dados hidrológicos que pode ser útil para o entendimento da metodologia adotada. Os três grupos são: (i) dados com falhas esparsas, ou isoladas; (ii) pequenas sequências de dados com falhas e (iii) sequências significativas de dados com falhas. No primeiro grupo se encontram aqueles registros incoerentes identificados pelos métodos descritos no item anterior, podendo ser resultado de algum erro na digitalização dos dados, por exemplo. O segundo e terceiro grupos são mais corriqueiros, podendo ser atrelados a períodos nos quais ocorreu alguma falha no limnígrafo, como secagem da tinta da pena, problemas com o relógio que controla o rolamento da bobina de papel, entre outros motivos. A diferença entre eles reside no fato de que no segundo os dados remanescentes são suficientes para caracterizar o perfil de comportamento da série. Já no terceiro, as falhas são tão extensas que as tentativas de preenchimento podem resultar em erros de grande magnitude.

A metodologia adotada para o preenchimento de falhas nos limnigramas digitalizados serve para os três grupos. Naturalmente sua precisão cai na medida em que se aumenta o período de falhas, sendo também dependente do comportamento da série em análise. A ideia é visualizar a questão do preenchimento sob a ótica da previsão de séries temporais. Como ressaltam Bennis *et al.* (1997), modelos que tradicionalmente são usados para prever dados futuros de uma série podem ser adaptados ao caso do preenchimento de valores em branco. Assim, a série final seria composta de uma parcela observada, seguida de uma parcela estimada e complementada por outra parcela observada, como mostra o esquema da Figura 2.

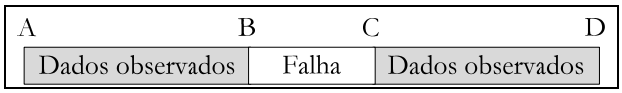

Figura 2– Esquema para preenchimento de falhas

O intervalo definido por BC marca o período de registros em branco a ser preenchido por técnicas de previsão. O maior desafio dessa metodologia é garantir que o último dado previsto (ponto C) coincida com o próximo registro observado, de forma a manter a coerência dos dados limnimétricos como um todo.

A forte autocorrelação inerente às cotas horárias fez com que a opção por modelos ARIMA,

também conhecidos por modelo Box e Jenkins (BOX *et al.*, 1994), fosse natural para o preenchimento dos períodos sem registros. Tais modelos resultam da combinação de três parcelas: o componente autorregressivo (AR), o filtro de integração (I) e o componente de médias móveis (MA). Uma série pode ser modelada pelos três filtros ou apenas um subconjunto deles. Existe uma grande variedade de modelos ARIMA, sendo que o modelo geral não sazonal é conhecido como ARIMA(p,d,q), no qual as letras p, d e q indicam a ordem da componente AR, o grau de diferenciação I e a ordem da componente MA, respectivamente.

Seja uma série de cotas qualquer representada pela variável x_t $(t = 1,2,\dots,n)$. Na modelagem de vazões em um rio tipicamente se adotam seus logaritmos, pois esta transformação aproxima a distribuição da série a uma normal (KELMAN, 1987). Essa convenção é também adotada para as cotas, de forma que $z_t = \log x_t$. Para início da descrição dos modelos ARIMA, introduz-se o operador de defasagem B, definido pela equação (4):

$$z_{t-k} = B^k z_t \tag{4}$$

Um modelo autorregressivo de médias móveis ARMA(p,q) é expresso pela equação (5):

$$\left(1 - \varphi_1 B - \cdots - \varphi_p B^p\right) z_t = (1 - \theta_1 B - \cdots - \theta_q B^q)\varepsilon_t \tag{5}$$

onde φ_p é o parâmetro autorregressivo de ordem p, θ_q é o parâmetro de médias móveis de ordem q e ε_t é uma série de resíduos (ou erros) aleatórios. De primordial importância para a correta estimação dos parâmetros do modelo de Box & Jenkins há de se considerar a série estatisticamente estacionária. Sabe-se, entretanto, que essa é uma condição raramente atendida em séries de fenômenos naturais. Por esse motivo, adiciona-se o filtro de integração I ao modelo, que consiste em tomar diferenças sucessivas da série original até obter uma série estacionária. A primeira diferença de z_t, por exemplo, é definida pela equação (6):

$$\mathrm{z}_t^d = z_t - z_{t-1} \qquad (\text{para } d = 1) \tag{6}$$

Comumente é suficiente tomar uma ou duas diferenças para que a série se torne estacionária (BOX *et al.*, 1994). O número d de diferenças necessárias é denominado ordem de integração do filtro I. A inclusão desse termo permite que sejam utilizados os modelos ARIMA(p,d,q) dados pela equação (7):

$$\left(1 - \varphi_1 B - \cdots - \varphi_p B^p\right)(1 - B - \cdots - B^d) z_t = \left(1 - \theta_1 B - \cdots - \theta_q B^q\right)\varepsilon_t \tag{7}$$

A metodologia Box e Jenkins utiliza uma estratégia de construção de modelos de forma iterativa em três passos, que consiste na identificação de um modelo candidato inicial, estimação de seus parâmetros e diagnóstico do modelo. Caso seja preciso, o modelo inicial é modificado e o processo é repetido até que o diagnóstico mostre que não seja necessária nova alteração.

O primeiro passo do procedimento iterativo tem grande importância para a modelagem das cotas. Em específico, este procedimento diz respeito a estudos acerca da classe dos modelos autorregressivos (com ou sem médias móveis) e suas respectivas ordens. A técnica mais tradicional de se identificar um modelo ARIMA é através da comparação gráfica entre as chamadas funções de autocorrelação (FAC) e de autocorrelação parcial (FACP). Esses gráficos são obtidos para as amostras e comparados a comportamentos teóricos esperados para cada modelo (SOUZA; CAMARGO, 2004).

Este exercício foi feito para as séries limnimétricas em estudo, na intenção de ajustar um modelo ARIMA apropriado para o preenchimento de falhas. Ao plotar a FAC das séries originais, notou-se que ela apresentou um decaimento muito lento, explicado pela forte correlação serial entre os dados horários. Dessa maneira a diferenciação das séries foi adotada. A questão resultante foi o número de diferenciações a ser empregado. A Figura 3 e a Figura 4 mostram, respectivamente, as FAC e FACP para as 1ª e 2ª diferenças no posto FSS. Utilizou-se 20 como defasagem (número de *lags*) máxima. As linhas tracejadas horizontais se referem ao intervalo dentro do qual as funções são estatisticamente iguais a zero.

Para a 1ª diferença, pode-se fazer um paralelo direto da FAC com o comportamento esperado de um modelo autorregressivo, pois apresenta um decaimento exponencial/senoidal em seus *lags*. No caso da FACP, contam-se três *lags* significativos, fazendo com que o modelo selecionado seja um ARIMA(3,1,0).

A 2ª diferença, por sua vez, apresenta uma FACP com decaimento exponencial aproximado, indicando um modelo de médias móveis. Ressalta-se que a análise do modelo MA é inversa à do modelo AR, ou seja, os componentes do modelo são indicados pela FACP e a ordem pela FAC. Seguindo esse roteiro, contam-se dois *lags* significativos na FAC, apontando para um modelo ARIMA(0,2,2).

Em resumo, têm-se dois modelos candidatos: ARIMA(3,1,0) e ARIMA(0,2,2). A escolha final foi feita relevando-se a complexidade em se estimar os parâmetros de cada um deles. Nesse quesito, modelos predominantemente autorregressivos são mais parcimoniosos, pois suas equações de estimação podem ser deduzidas analiticamente (BOX *et al.*, 1994). Além disso, esse modelo possui um apelo físico mais interessante para a série de cotas do que um modelo de médias móveis puro, pois as relaciona diretamente.

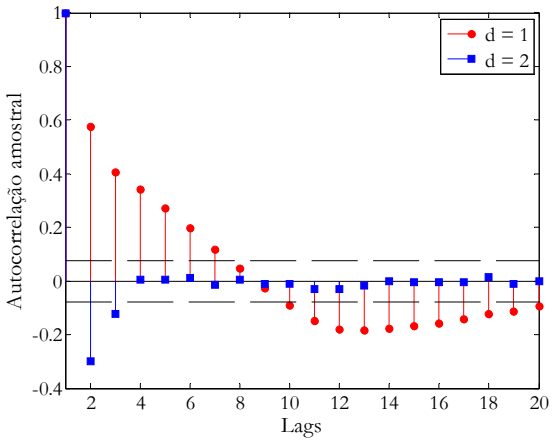

Figura 3 – FAC para posto FSS

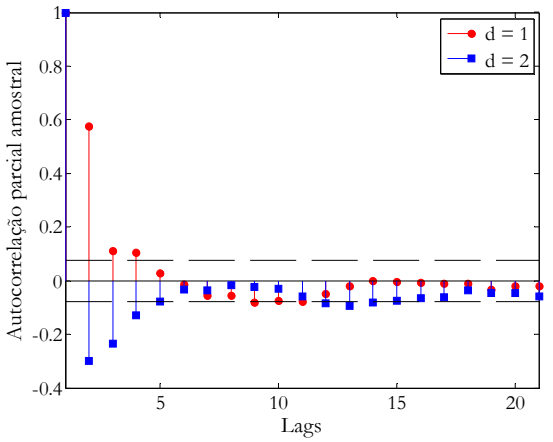

Figura 4 – FACP para posto FSS

O comportamento observado nas demais séries de cotas horárias digitalizadas seguiu o mesmo padrão mostrado. Porém, antes de fixar o modelo final, deve-se mencionar parte importante da teoria de Box & Jenkins: modelos ARIMA ajustados a séries estacionárias, não necessitam de ordens superiores a dois (BOX *et al.*, 1994). Lembra-se que o uso do filtro de integração I para as séries as deixa com essa característica. Portanto, sem perdas de generalidade, assume-se o modelo ARIMA(2,1,0) como padrão para o preenchimento dos eventuais dados faltosos ou errôneos nas séries. Este modelo é descrito formalmente pela equação (8):

$$(1 - \varphi_1 B - \varphi_2 B^2)(1 - B) z_t = \varepsilon_t \qquad (8)$$

Alternativamente, pode-se escrever o modelo sem o uso do operador B, chegando-se à equação de previsão utilizada no preenchimento:

$$z_{t+1} = z_t(1 + \varphi_1) - z_{t-1}(\varphi_1 - \varphi_2) - z_{t-2}\varphi_2 \qquad (9)$$

Atenta-se para a omissão do termo do resíduo da equação (9), pois se trata de uma formulação que não considera defasagens nos erros (e.g. MA = 0).

A identificação de modelos Box & Jenkins pode também ser feita de forma automática, aplicando-se os chamados critérios de informação. São equações que relevam a qualidade do ajuste promovido por cada modelo (a partir da função de verossimilhança) e uma penalidade atribuída ao número de parâmetros considerados. Os mais conhecidos são os critérios de informação de Akaike (AIC – AKAIKE, 1974) e de Bayes (BIC – SCHWARTZ, 1978). Entretanto, esses métodos têm a desvantagem de necessitar da estimação de todos os modelos candidatos, o que pode tornar o processo moroso.

A estimação dos parâmetros φ_1 e φ_2 do modelo foi feita empregando-se o método da máxima verossimilhança, objetivando-se a minimização da soma dos quadrados dos resíduos ε_t da equação (8). A formulação final foi, então, submetida à etapa de diagnóstico, na qual foram testadas a independência, homocedasticidade e a distribuição de probabilidades da série de resíduos (BOX *et al.*, 1994).

Ponderação das previsões

A questão pendente a ser resolvida é como garantir que o último dado previsto coincida com o próximo registro observado. Esse ponto foi levantado nos estudos de Elshorbagy *et al.* (2000) e Bennis *et al.* (1997), nos quais os autores sugeriram que uma interpolação dos valores previstos poderia ser uma saída. Utilizando essa indicação em conjunto com o emprego de previsões *forward-backward* (também considerada em BENNIS *et al.*, 1997), chegou-se a uma solução, explicada mais claramente com um exemplo: seja uma série de cotas horárias z_t na qual se detectou uma sequência de cinco registros

faltantes (n = 5). Essa descontinuidade observada na série quebra-a em dois segmentos, um antes e outro depois das falhas. Assim, dois modelos ARIMA(2,1,0) são ajustados, um para cada segmento de série. Para o segmento anterior à falha, o primeiro modelo é aplicado para prever as cinco próximas cotas; para o segmento posterior à falha, o segundo modelo é ajustado para prever as cinco cotas anteriores. Isso é possível através da manipulação dos índices temporais das variáveis na equação do modelo (equação (9)).

Depois de efetuado o procedimento descrito restará duas previsões para o mesmo período de falhas. Estas previsões são submetidas a uma ponderação para que a estimativa final das falhas seja determinada. A atribuição de pesos às previsões é feita de forma que se priorize o período anterior ou posterior à falha, o que for mais próximo da cota estimada. Tem-se, portanto, que a sequência final das cotas prevista é dada pela equação (10):

$$\hat{z}_t = \alpha_t \hat{z}_t^{\text{anterior}} + \beta_t \hat{z}_t^{\text{posterior}} \qquad (10)$$

onde $\hat{z}_t$ é a sequência final de cotas, $\hat{z}_t^{\text{anterior}}$ é a sequência de previsões utilizando o segmento anterior à falha (previsões *forward*), $\hat{z}_t^{\text{posterior}}$ é a sequência de previsões utilizando o segmento posterior à falha (previsões *backward*) e α_t e β_t são os pesos. Para o exemplo hipotético, $t = 1, \ldots, 5$ e os pesos são calculados para cada instante de tempo. Como condição, a soma do conjunto $\boldsymbol{\alpha}$, bem como do conjunto $\boldsymbol{\beta}$, para todos os instantes de tempo, deve ser um. A diferença é que $\boldsymbol{\alpha}$ inicia em um e decresce até zero quando $t = 5$, utilizando um decaimento de fator $1/(n-1)$. O peso $\boldsymbol{\beta}$ evolui de forma oposta, iniciando em zero e crescendo até um quanto $t = 5$, utilizando o mesmo fator $1/(n-1)$. Todo o procedimento é mais bem entendido através da Figura 5, que esquematiza essa metodologia.

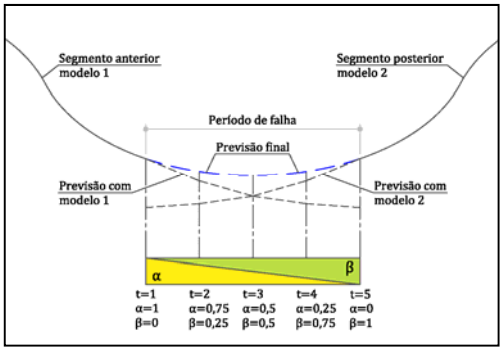

Figura 5 – Ponderação entre previsões para preenchimento do período de falhas

Evidentemente, em casos de *outliers* isolados detectados, o método também funciona, pois os pesos resultarão 0,5, equivalente à média das duas previsões.

Para a verificação do desempenho do modelo proposto, utilizou-se o erro médio percentual absoluto (MAPE, sigla inglesa para *mean absolut percentage error*), índice muito usado para avaliar a qualidade de previsão de séries temporais (MAKRIDAKIS *et al.*, 1998). Por ser medido em porcentagem, esse indicador pode ser empregado para comparar o desempenho em diferentes postos limnimétricos. O MAPE é calculado de acordo com a equação (11):

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{z_t - \hat{z}_t}{z_t}\right| \times 100 \qquad (11)$$

Onde z_t é a observação real retirada da série histórica, $\hat{z}_t$ é o valor previsto para o mesmo instante de tempo e n é o horizonte de previsão considerado. Nos experimentos numéricos realizados, falhas foram deliberadamente inseridas nas séries históricas de FSS e FB e submetidas ao procedimento de preenchimento. As estimativas foram, então, comparadas com os valores originais através do MAPE.

RESULTADOS E DISCUSSÕES

Antes de aplicar o modelo ARIMA(2,1,0) no preenchimento das falhas, suas propriedades teóricas foram verificadas através da aplicação de testes de hipótese para independência (PORTMANTEAU – LI e McLEOD, 1981), homocedasticidade (LEVENE – BROWN e FORSYTHE, 1974) e normalidade (LILLIEFORS – LILLIEFORS, 1967) da série de resíduos. A formulação foi ajustada a dez momentos da série de cotas horárias, cinco para cada posto linimétrico. Do total, os resíduos foram considerados independentes em 30% dos casos, homocedásticos em 100% dos casos e com distribuição aproximadamente normal em 30% dos casos, utilizando um nível de confiança de 5%. Apesar dos resultados pouco satisfatórios obtidos para o primeiro e último quesitos, lembra-se que essas são verificações balizadoras e que não invalidam o método. Uma melhora nesses índices poderia ser obtida caso fosse adotado o modelo ARIMA(3,1,0) inicialmente identificado. No entanto, a complexidade da formulação e o custo computacional para sua aplicação se elevariam, afetando a parcimônia do método.

Observou-se que o desempenho do modelo ARIMA dependeu de três fatores inter-relacionados: (i) a característica da série, (ii) o tamanho do período de falhas e (iii) a posição do período de falhas. O primeiro fator se refere ao comportamento geral da série. Quanto maior sua volatilidade, mais dificuldades o modelo terá para estimar valores faltantes. O segundo diz respeito ao horizonte a ser previsto; naturalmente, quanto maior esse valor, menos acurada será a estimativa. Finalmente, o terceiro fator diferencia falhas presentes em períodos considerados normais da série de picos ou recessões. A Figura 6 exibe a aplicação da metodologia para um caso considerado simples. Trata-se de um longo período de recessão observado na série de FB. Inseriram-se três falhas distintas, totalizando 133 registros vazios. Após aplicação do método, as cotas estimadas obtiveram um MAPE acumulado de apenas 0,36%.

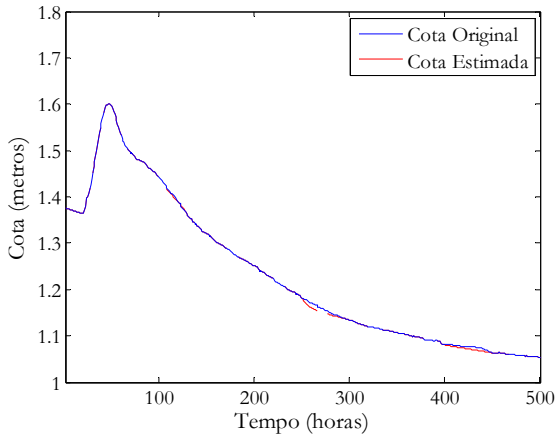

Figura 6 – Preenchimento de falhas para período de recessão da série FB (133 falhas; MAPE total: 0,36%)

A situação menos favorável ocorre quando a falha se dá justamente sobre valores extremos, sejam eles picos ou vales. Para demonstrar o comportamento do modelo nesses casos, selecionou-se uma sequencia de 200 registros de FSS relativos a um período da série relativamente volátil. Foi inserida uma falha contínua de 20 horas justamente sobre o pico existente. Após passar pelo modelo, o resultado obtido foi o mostrado pela Figura 7.

Nota-se que o modelo não foi capaz de reproduzir inteiramente o pico observado, ficando pouco abaixo dos valores reais. No entanto, o MAPE de 1,23% obtido para esse preenchimento pode ser considerado baixo quando comparado com índices padrão em estudos de previsão (MAPEs variando de 3% a 5%, dependendo da série analisada). Ademais, só o fato de o método ser capaz de reproduzir a curvatura do pico agrega avanços significativos em relação aos tradicionais métodos de interpolação. Esse resultado é ainda mais expressivo se for lembrado que a única informação utilizada para o preenchimento é a própria série, dispensando variáveis externas ou análises cruzadas entre postos de medição. Outros resultados são mostrados da Figura 8 à Figura 13, sendo que o posto, o tamanho das falhas e os MAPEs obtidos são identificados em suas respectivas legendas. Nesses resultados foi dada prioridade a falhas sobre picos e vales das séries, na intenção de avaliar o desempenho do modelo sob essas circunstâncias específicas.

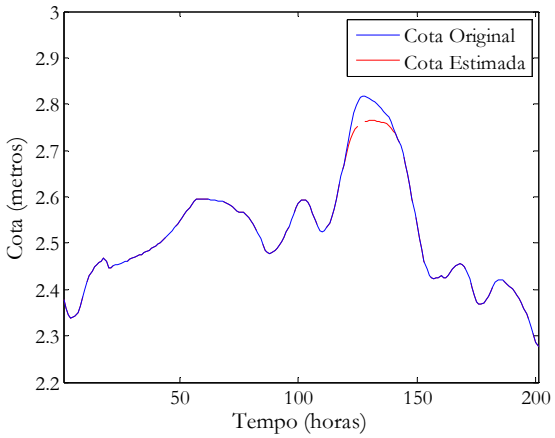

Figura 7 – Preenchimento de falhas para uma situação de pico de cheia na série FSS (20 falhas, MAPE: 1,23%)

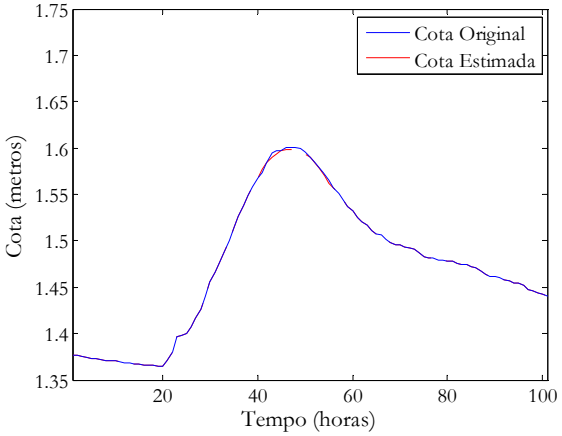

Figura 8 – Posto FB; 16 falhas; MAPE: 0,13%

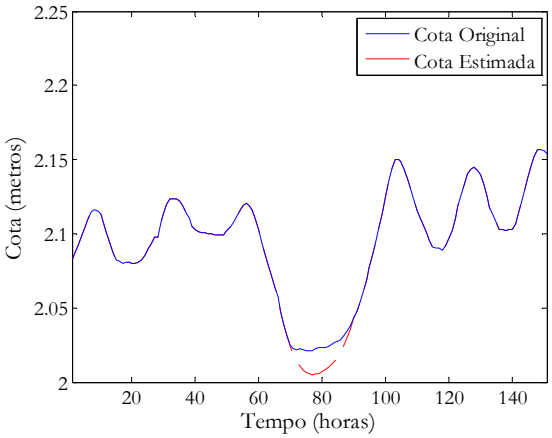

Figura 9 – Posto FSS; 20 falhas; MAPE: 0,52%

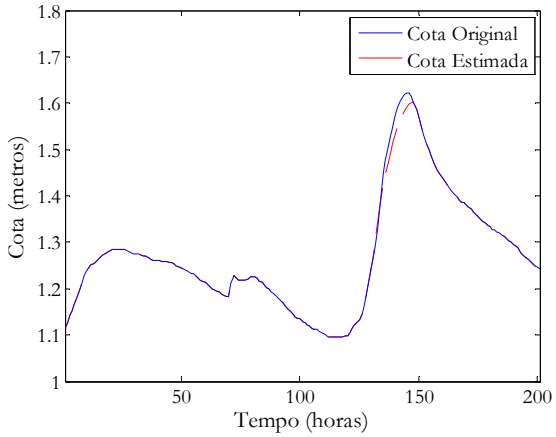

Figura 10 – Posto FB; 16 falhas; MAPE: 1,60%

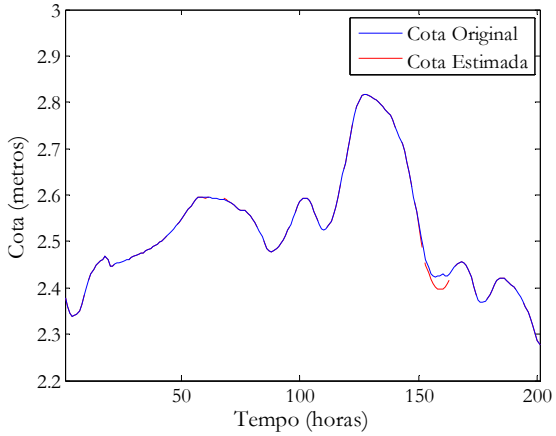

Figura 11 – Posto FSS; 11/16 falhas; MAPE total: 0,43%

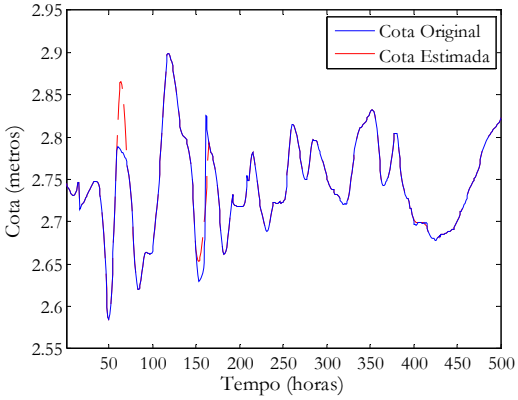

Figura 12 – Posto FSS; 11/16/16 falhas; MAPE total: 0,98%

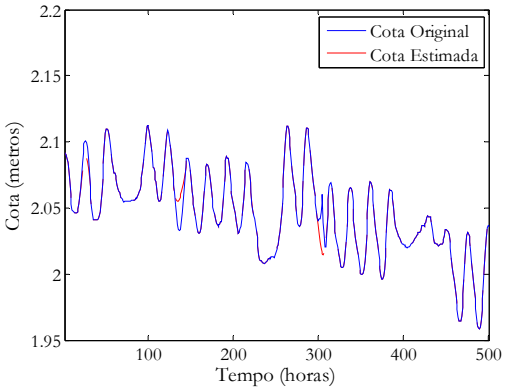

Figura 13 – Posto FSS; 11/6/16 falhas; MAPE total: 0,61%

A Tabela 1 apresenta estatísticas básicas de médias e desvios padrão para as séries originais e preenchidas. Percebe-se que as diferenças apresentadas foram mínimas, muitas delas diferindo apenas a quarta casa decimal (escala dos décimos de centímetros).

Nestas situações diversas, percebe-se que o modelo ARIMA obteve um bom desempenho, aparecendo como alternativa mais atrativa em relação às práticas comuns de análise de consistência em limnigramas. Tipicamente, em se deparando com algum período vazio, o operador é orientado a preencher as extremidades da falha com uma linha reta (SANTOS *et al.*, 2001). Se houver um pico neste período, o erro cometido será de grande magnitude. Por conter informações da série histórica em seus parâmetros, o modelo ARIMA melhora essas estimativas.

Tabela 1 – Estatísticas básicas comparativas das séries originais e preenchidas

Resultado	Média (m)		Desvio Padrão (m)	
	Original	Corrigida	Original	Corrigida
Figura 6	1,2321	1,2311	0,1568	0,1573
Figura 7	2,5222	2,5187	0,1240	0,1164
Figura 8	1,4734	1,4731	0,0737	0,0733
Figura 9	2,0950	2,0936	0,0362	0,0392
Figura 10	1,2737	1,2718	0,1292	0,1247
Figura 11	2,5222	2,5211	0,1240	0,1250
Figura 12	2,7422	2,7434	0,0563	0,0565
Figura 13	2,0466	2,0464	0,0322	0,0321

Ressalta-se, no entanto, que cuidado deve ser tomado principalmente na relação tamanho da falha *versus* posição do período de falha. Como dito anteriormente, se forem observadas falhas longas coincidentes a eventos hidrológicos significativos, o preenchimento pode não ser satisfatório.

Por outro lado, as estações limnimétricas de monitoramento consideradas nesse estudo, contam com a visita diária de um operador que realiza a inspeção do equipamento. Dessa maneira, salvo defeitos mais graves no limnígrafo, os períodos de falhas nos registros dificilmente chegam a 24 horas, o que torna a aplicação da metodologia proposta perfeitamente viável.

CONCLUSÃO

Apresentou-se neste trabalho um método para preenchimento de falhas em dados hidrológicos horários provenientes de limnigramas instalados em rios. Utilizaram-se conceitos de análise de séries temporais aplicadas à previsão, com foco na modelagem ARIMA de Box & Jenkins. Mostrou-se ser possível o ajuste de uma formulação simples e aplicação em séries cujas falhas apresentam tamanhos e características distintas.

A intenção deste trabalho foi oferecer uma alternativa às técnicas tradicionais de análise de consistência. A literatura oferece estudos com a mesma problemática, contudo aplicados a séries com escalas diárias e mensais. Diferentemente dessas escalas, entretanto, registros horários possuem como característica a falta de uma sazonalidade bem definida, a forte autocorrelação serial dos registros e a própria volatilidade das séries, tornando sua modelagem desafiadora. Contornando esses obstáculos, os modelos ARIMA produziram resultados satisfatórios.

Considera-se como grande vantagem da metodologia apresentada sua independência com relação a variáveis externas, sejam elas de postos limnimétricos vizinhos ou de outras variáveis hidrológicas (como chuvas, por exemplo). Relevando-se a reduzida rede de monitoramento hidrológica ativa no Brasil, a aplicação de técnicas como a mostrada aqui é bastante atrativa.

AGRADECIMENTOS

OMITIDO POR CONTER REFERÊNCIA AOS AUTORES E ENTIDADES.

REFERÊNCIAS

ABATZOGLOU, J. T.; REDMOND, K. T.; EDWARDS, L. M. Classification of Regional Climate Variability in the State of California. *Journal of Applied Meteorology and Climatology*, v. 48, n. 8, p. 1527–1541, 2009.

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, v. 19, n. 6, p. 716-723, 1974.

AGÊNCIA NACIONAL DE ÁGUAS. *Diretrizes e análises recomendadas para a consistência de dados fluviométricos*. Brasília: ANA, Superintendência de Gestão da Rede Hidrometeorológica, 2011. Disponível em: http://arquivos.ana.gov.br/infohidrologicas/cadastro/Diretrizes_Analises_Recomendadas_Consistencia_de_Dados_Fluviometricos.pdf. Acesso em: 08 mai. 2013.

BASU, S.; MECKESHEIMER, M. Automatic outlier detection for time series: an application to sensor data. *Knowledge and Information Systems*, v. 11, n. 2, p. 137-154, 2006.

BENNIS, S.; BERRADA, F.; KANG, N. Improving single-variable and multivariable techniques for estimating missing hydrological data. *Journal of Hydrology*, v. 191, p. 87-105, 1997.

BROWN, M. B.; FORSYTHE, A. B. Robust tests for the equality of variances. *Journal of the American Statistical Association*, v. 69, n. 346, p. 364-367, 1974.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. *Time Series Analysis: Forecasting and Control*. 3rd. Ed. New Jersey: Prentice Hall, 598 p., 1994.

DASTORANI, M. T.; MOGHADAMNIA, A.; PIRI, J.; RICO-RAMIREZ, M. Application of ANN and ANFIS models for reconstructing missing flow data. *Environmental monitoring and assessment*, v. 166, n. 1-4, p. 421–34, 2010.

DINPASHOH, Y.; JHAJHARIA, D.; FAKHERI-FARD, A.; SINGH, V. P.; KAHYA, E. Trends in reference crop evapotranspiration over Iran. *Journal of Hydrology*, v. 399, n. 3-4, p. 422–433, 2011.

ELSHORBAGY, A. A.; PANU, U. S.; SIMONOVIC, S. P. Group-based estimation of missing hydrological data: II. Application to streamflows. *Hydrological Sciences*, v. 45, n. 6, p. 867-880, 2000.

ELSHORBAGY, A.; SIMONOVIC, S. P.; PANU, U. S. Estimation of missing streamflow data using principles of chaos theory. *Journal of Hydrology*, v. 255, p. 123-133, 2002.

HUGHES, D. A.; SMAKHTIN, V. Daily flow time series patching or extension : a spatial interpolation approach based on flow duration curves. *Hydrological Sciences*, v. 41, June, p. 851-872, 1997.

FIRAT, M.; DIKBAS, F.; KOC, A. C.; GUNGOR, M. Analysis of temperature series: estimation of missing data and homogeneity test. *Meteorological Applications*, v. 19, n. 4, p. 397–406, 2012.

GONZALEZ, R. C.; WOODS R. E. *Digital Image Processing*. 3rd. Ed. Reading: Addison-Wesley, 1992, 730 p.

GYAU-BOAKYE, P.; SCHULTZ, G. A. Filling gaps in runoff time series in West Africa. *Hydrological Sciences*, v. 39, n. 6, p. 621-636, 1995.

IGLEWICZ, B.; HOAGLIN, D. Volume 16: How to Detect and Handle Outliers, In.: *The ASQC Basic References in Quality Control: Statistical Techniques*, MYKYTKA, E. F. (org.), Universidade da Califórnia, 1993, 87p.

KELMAN, J. Modelos estocásticos no gerenciamento de recursos hídricos. In:_____. *Modelos para Gerenciamento de Recursos Hídricos I.* São Paulo: Nobel/ABRH. 1987, cap. 4.

KHALIL, M.; PANU, U. S.; LENNOX, W. C. Groups and neural networks based streamflow data infilling procedures. *Journal of Hydrology*, v. 241, p. 153-176, 2001.

KIM, T.-W.; AHN, H. Spatial rainfall model using a pattern classifier for estimating missing daily rainfall data. *Stochastic Environmental Research and Risk Assessment*, v. 23, n. 3, p. 367–376, 2008.

KIM, J.-W.; PACHEPSKY, Y. A. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. *Journal of Hydrology*, v. 394, n. 3-4, p. 305–314, 2010.

KUMAMBALA, P. G. . Sustainability of Water Resources Development for Malawi with Particular Emphasis on North and Central Malawi, 2010. 412 f. Tese (Doutorado em Engenharia Civil) - Universidade de Glasgow, Reino Unido. Disponível em: http://theses. gla.ac.uk/1801/, Acesso em 15/07/2013.

LI, W. K.; McLEOD, A. I. Distribution of the residual autocorrelations in multivariate ARMA time series models. *Journal of the Royal Statistical Society*, series B, v. 43, n. 2, 231-239, 1981.

LILLIEFORS, H. W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*. v. 62, p. 399–402, 1967.

MAKRIDAKIS, S.; WHEELWRIGHT, S. C.; HYNDMAN, R. J. *Forecasting: methods and application*, 3rd Ed. New Jersey: John Wiley & Sons, 1998, 642 p.

MWALE, F. D.; ADELOYE, A. J.; RUSTUM, R. Infilling of missing rainfall and streamflow data in the

Shire River basin, Malawi – A self organizing map approach. *Physics and Chemistry of the Earth, Parts A/B/C*, v. 50-52, p. 34–43, 2012.

NASCIMENTO, R. M.; OENING, A. P.; MARCILIO, D. C.; AOKI, A. R.; ROCHA JR., E.; SCHIOCHET, J. Outliers' Detection and Filling Algorithms for Smart Metering Centers. In.: 2012 IEEE Power & Energy Society Transmission and Distribution, 2012, Orlando. *Proceedings*... Orlando: IEEE, 2012.

PAGOTTO, T. C. S.; SOUZA, P. R. de (org.). *Biodiversidade do Complexo Aporé-Sucuriú: subsídios à conservação e ao manejo do Cerrado*. Campo Grande: UFMS, 308 p., 2006. Disponível em: http://www.mma.gov.br/estruturas/chm/_arquivos/Complexo_Apore_Sucuriu.pdf. Acesso em: 16/07/2013.

QUEVEDO, J.; PUIG, V.; CEMBRANO, G.; BLANCH, J.; AGUILAR, J.; SAPORTA, D.; BENITO, G.; HEDO, M.; MOLINA, A. Control Engineering Practice Validation and reconstruction of flow meter data in the Barcelona water distribution network. *Control Engineering Practice*, v. 18, n. 6, p. 640-651, 2010.

SANTOS, I.; FILL, H. D. O. A.; SUGAI, M. R. von B.; BUBA, H.; KISHI, R. T.; MARONE, E.; LAUTERT, L. F. C. *Hidrometria Aplicada*. Curitiba: LACTEC, 2001, 372 p.

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

SIMONOVIC, S. P. Synthesizing missing streamflow records on several Manitoba streams using multiple nonlinear standardized correlation analysis. *Hydrological Sciences*, v. 40, n. 2, p. 183-203, 1995.

SOUZA, R. C.; CAMARGO, M. E. *Análise e previsão de séries temporais: os modelos ARIMA*. 2ª Ed. Rio de Janeiro: Ed. Regional, 2004, 187 p.

STARRETT, S. K.; STARRETT, S. K.; HEIERM, T.; SU, Y.; TUAN, D.; BANDURRAGA, M. Filling in missing peakflow data using artificial neural networks. ARPN Journal of Engineering and Applied Sciences, v. 5, n. 1, p. 49–55, 2010.

WORLD METEOROLOGICAL ORGANIZATION. *Guide to Hydrological Practices*. Vol. I e II, n. 168. Genebra: WMO, 2008, 598 p. Disponível em: http://www.whycos.org/hwrp/guide/index.php. Acesso em: 08 mai. 2013.

YAWSON, D. K.; KONGO, V. M.; KACHROO, R. K. Application of linear and nonlinear techniques in river flow forecasting in the Kilombero River basin, Tanzania. *Hydrological Sciences Journal*, v. 50, n. 5, p. 783–796, 2012.

[1]

Completing Hourly River Stage Data Via ARIMA Models

ABSTRACT

The use of hydrological data is vital in any instance of water resources planning and management in a particular area. Together with these data, it is essential to perform a consistency analysis in order to detect errors or flaws in the records. This paper focuses on river stage data, presenting, first, a simple method to detect erroneous data, based on outlier detection techniques. Next ARIMA models and concepts of time series forecasting are applied to complete gaps in the data. Forward-backward predictions with subsequent estimate weighting are used. As a case study, river stage data referring to two monitoring stations installed in the Paraná river basin are considered. The results were very satisfactory, with low MAPEs for the displayed situations. The independence of the presented methodology regarding external data is emphasized, a common procedure in traditional consistency analysis.

Keywords: *Consistency analysis. River stage data. ARIMA model. Hourly time scale.*