IV-135 – APLICAÇÃO DO MÉTODO DE ANÁLISE DE ESPECTRO SINGULAR PARA PREENCHER FALHAS DE DADOS EM SÉRIES TEMPORAIS DE PRECIPITAÇÃO ACUMULADA MENSAL

Cesar Augusto Medeiros Destro[(1)]
Engenheiro Sanitarista pela Universidade Federal do Mato Grosso (UFMT). Mestre em Recursos Hídricos pela Universidade Federal do Mato Grosso (UFMT). Estudante de Doutorado em Engenharia de Recursos Hídricos e Ambiental na Universidade Federal do Paraná (UFPR).
Acelmo de Jesus Brito
Bacharel em Matemática pela Universidade do Estado de Mato Grosso (UNEMAT). Mestre em Recursos Hídricos pela Universidade Federal do Mato Grosso (UFMT).
Pedro Cassiano Assumpção de Farias
Estudante do Curso de Engenharia Sanitária e Ambiental na Universidade Federal do Mato Grosso (UFMT).
Gilson Alberto Rosa Lima[(2)]
Bacharel em Física pela Universidade Federal de Santa Maria (UFSM). Mestre em Física pela Universidade Estadual de Campinas (UNICAMP). Doutor em Ciências pela Universidade de São Paulo (USP).

Endereço[(1)]**:** Universidade Federal do Paraná - Centro Politécnico, Bloco V, Caixa Postal 19011 - Jardim das Américas - Curitiba - PR - CEP: 81531-980 - Brasil - Tel: (41) 3361-3134 - e-mail: eng.destro@gmail.com
Endereço[(2)]**:** Universidade Federal do Mato Grosso - Departamento de Engenharia Sanitária e Ambiental, Bloco F, Caixa Postal 19011 - Boa Esperança - Cuiabá - MT - CEP: 78060-900 - Brasil - Tel: (65) 3615-8721 - e-mail: garlima@ufmt.br

RESUMO

Este estudo mostra o preenchimento falhas de dados em séries temporais de precipitação acumulada mensal utilizando o método de Análise de Espectro Singular (*Singular Spectrum Analysis* - *SSA*) e compara os resultados obtidos com o método de Regressão Linear Simples (*RLS*), um dos mais utilizados no Brasil e recomendado pela Agencia Nacional de Águas (ANA). O estudo foi realizado levando em conta séries temporais reais e sintéticas com características distintas (falhas regulares, aleatórias e períodos longos). Os resultados mostram que embora o valor do erro médio absoluto (*EMA*) e erro médio (*EM*) calculados para os dois métodos estejam na mesma ordem de grandeza, as magnitudes dos coeficientes de determinação e correlação indicam que os valores observados e estimados não podem ser explicados de forma satisfatória através da equação da reta de regressão. O preenchimento de falhas com o método *SSA* reproduziu o comportamento da série original nos intervalos de tempo onde não foram realizadas as medições demonstrando a robustez e eficiência do método *SSA* para estimar os dados faltantes.

PALAVRAS-CHAVE: Precipitação, Falhas de dados, Análise de Espectro Singular, Regressão Linear Simples.

INTRODUÇÃO

Dados meteorológicos constituem a base fundamental dos estudos ambientais e mudanças climáticas. A precipitação em forma de chuva é uma variável meteorológica fundamental para o entendimento de processos hidrológicos, caracterização do clima regional e todos os estágios da produção agrícola. Sendo a única forma na qual a água entra em uma bacia hidrográfica, pode ser utilizada no contexto da engenharia para quantificar a disponibilidade hídrica e dimensionar sistemas de irrigação, e de controle de inundações e erosão do solo. Como a precipitação oscila devido à dinâmica solar, atmosférica, oceânica e do solo (GHIL, 2002), os valores observados devem refletir a variação espacial e temporal (efeito sazonal) destes.

Os registros de precipitação são obtidos a partir de observações em campo, sob condições adversas, de maneira que as séries temporais podem conter falhas de dados. Os motivos que levam às falhas de dados estão associados a: i) quebra dos equipamentos de medição; ii) mudança do local de medição; iii) erros humanos e iv) condições do tempo, no caso do sensoriamento remoto. Harmanglucioglu et al. (1999) ressalta que além destes fatores, nos países em desenvolvimento, a falta de recurso financeiro também é um importante fator responsável

pelo surgimento das falhas de dados. No Brasil, a ANA disponibiliza através do sistema *Hidroweb*[1] séries temporais de diversas variáveis hidrológicas. Porém, a maior parte das séries temporais possui muitas falhas de dados.

As falhas de dados são prejudiciais, pois impõem um limite na qualidade da informação que pode ser extraída dos valores observados (QUIMPO e YANG, 1970; HARMANGLUCIOGLU et al., 1999; HAIR et al., 2005). Muitos modelos matemáticos para analisar séries temporais, partem do pressuposto de que as series são completas, isto é, as observações são realizadas com frequência constante, o que não ocorre devido às falhas de dados. Logo, preencher estas falhas é necessário para que toda a informação presente na série temporal possa ser extraída e a qualidade da informação maximizada. De outra maneira, a informação extraída pode conduzir a resultados e interpretações errôneas sobre o fenômeno estudado.

Existem vários métodos para contornar o problema das falhas de dados (LITTLE e RUBIN, 2002; BOX e PIERCE, 1970; BRILLINGER, 2001; BROCKWELL e DAVIS, 2002). Porém, estes métodos necessitam do conhecimento prévio de uma função matemática que descreva o comportamento dos valores observados antes do preenchimento das falhas. Além disso, para obter a função matemática é necessário estimar e/ou ajustar alguns parâmetros intrínsecos de cada modelo, introduzindo algum tipo de subjetividade no processo de estimação. Mais recentemente, a abordagem não paramétrica tem ganhado espaço na literatura como ferramenta para análise de séries temporais (BROOMHEAD e KING, 1986a; 1986b). Na abordagem não paramétrica, não é necessário conhecer previamente uma função matemática. Assume-se que a série temporal pode ser decomposta em um soma de componentes independentes e reconstruída por uma combinação ou superposição dos componentes.

O método de Análise de Espectro Singular (*SSA*), por ser auto-adaptativo e não-paramétrico, têm se mostrado eficiente para analisar, preencher falhas e reproduzir oscilações periódicas em séries temporais de diferentes áreas do conhecimento (SCHOELLHAMER, 2001; KONDRASHOV et al., 2005; KONDRASHOV e GHIL, 2006; GOLYANDINA et al., 2001; GOLYANDINA e OSIPOV, 2007; DESTRO, 2010; DESTRO et al., 2012), sendo adequada para a análise de séries cuja dinâmica é não linear. Neste contexto este estudo mostra como o método SSA pode ser utilizado para preencher falhas de dados de séries temporais de precipitação acumulada mensal.

METODOLOGIA

Os dados de precipitação acumulada mensal utilizados neste estudo foram obtidos do sistema *HidroWeb*. Foram pré-processados os dados das 234 estações pluviométricas localizadas no Estado de Mato Grosso. No pré-processamento, se realizou uma inspeção visual das séries e se selecionou àquelas passíveis de preenchimento (devido ao grande número de falhas em algumas séries, o preenchimento é inviável). O presente artigo mostra apenas o preenchimento e análises posteriores realizadas para a estação Santa Rosa (857000), porque esta possui um menor número de falhas de dados (5% do total de registros), sendo possível criar falhas artificiais na série temporal no intuito de verificar o comportamento dos métodos de preenchimento. Outro fator importante é a consistência dos dados. Nenhuma das 234 séries temporais consideradas neste estudo possui dados consistidos no nível dois do início ao final da série. Como mais da metade das séries contém muitas falhas de dados, torna-se difícil aplicar o método de dupla massa (SEARCY e HARDISON, 1960) para consisti-las.

Para preencher as falhas de dados com o método de *RLS*, foram obtidos dados de outras seis séries temporais de estações consideradas próximas à estação Santa Rosa. Porém, as estações Nova Monte Verde (957002), Trivelato (1057001), Paranaíta (956002) e Cotriguaçu (958001) foram descartadas, uma vez que as séries temporais dessas estações não cobrem o período de interesse do estudo ou possuem falhas de dados nos mesmos meses das falhas da série a ser preenchida. A estação Núcleo Ariel (1058002), dista 140*Km* da estação Santa Rosa, foi utilizada como referência para preencher as falhas de dados da estação Santa Rosa pois, em um raio de 150*Km*, foi a única estação com registros de precipitação no período de interesse do estudo (Figura 1).

[1] http://hidroweb.ana.gov.br/

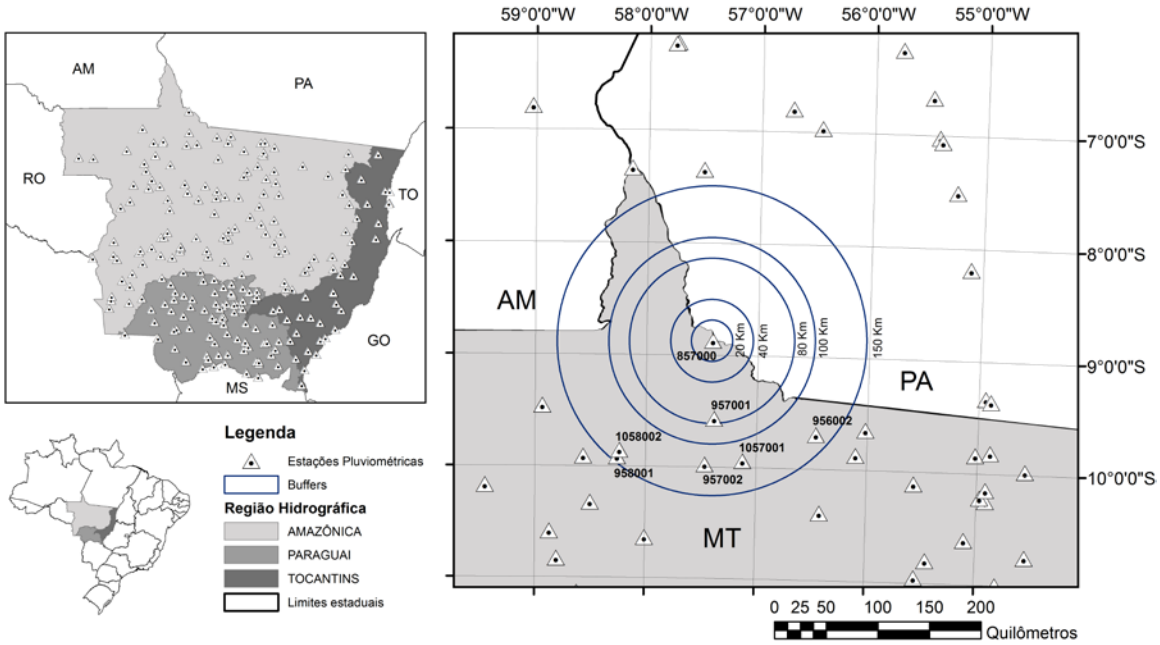

Figura 1: Localização das estações pluviométricas no Estado de Mato Grosso

Analise de Espectro Singular

O termo "espectro singular" provém da decomposição espectral (que conserva os autovalores) de uma matriz **X**, em um conjunto (espectro) de autovalores, λ. Estes autovalores, são os valores que tornam a matriz A - λ I singular. O algoritmo do método *SSA* contém quatro passos agrupados em dois estágios complementares entre si. O primeiro estágio decompõe a série original, enquanto o segundo estágio reconstrói a série. No caso de séries com falhas de dados, os valores faltantes são calculados durante a reconstrução da série. Detalhes sobre a modificação da *SSA* padrão para a análise de séries contendo falhas de dados pode ser obtidos em Golyandina e Osipov (2007).

Inserção da série no espaço vetorial - O primeiro passo da decomposição da série original corresponde à inserção da série em um espaço vetorial (*embedding*), isto é, transforma a série original (unidimensional e escalar) em uma série multidimensional cuja dimensão *L* é chamada de tamanho da janela. A série temporal multidimensional (sequência de vetores) forma a matriz trajetória (equação 1). Cada um dos vetores colunas é um sub-conjunto da série original e representa uma visão parcial da série. Sendo *K* o número de vetores defasados no tempo, $K = N - L + 1$.

$$\mathbf{X} = \left(x_{ij}\right)_{i,j=1}^{L,K} \begin{bmatrix} f_0 & f_1 & f_2 & \cdots & f_{K-1} \\ f_1 & f_2 & f_3 & \cdots & f_k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & f_{L+1} & \cdots & f_{N-1} \end{bmatrix} \tag{1}$$

Decomposição em valores singulares - O segundo passo decompõe a matriz trajetória em valores singulares (*Singular Value Decomposition*, (SVD)), em outras palavras, uma soma de matrizes bi-ortogonais de posto 1. Seja $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$ e $\lambda_1 \geq \ldots \geq \lambda_L \geq 0$ autovalores de **S,** e $\mathbf{U}_1,\ldots,\mathbf{U}_L$, os correspondentes autovetores, forma-se o sistema ortonormal. Representando os componentes principais da matriz trajetória como, $V_i = \mathbf{X}^{\mathbf{T}}\mathbf{U}_i / \sqrt{\boldsymbol{\lambda}_i}$, ($i$ = 1,2,3.....d), a decomposição da matriz **X** pode ser representada por,

$$\mathbf{X} = \left[\mathbf{X}_1 + \mathbf{X}_2 + \ldots.. + \mathbf{X}_d\right] \tag{2}$$

onde d denota o número de autovalores diferentes de zero da matriz **S** e $\mathbf{X}_i = \sqrt{\boldsymbol{\lambda}_i \mathbf{U}_i}\mathbf{V}_i^t$ as matrizes elementares.

Agrupamento - O primeiro passo da reconstrução é o agrupamento (*grouping*) das matrizes elementares $\mathbf{X}_i$ em vários grupos e a soma delas dentro de cada grupo, ou seja, particiona o conjunto de índices da equação (2) em subconjuntos disjuntos $I_1,\ldots, I_m$, isto é, se não tiverem nenhum elemento em comum. Fazendo $I = \{i_1, \ldots, i_p\}$ então a matriz resultante $\mathbf{X}_I$ correspondente ao grupo I é definida como $\mathbf{X}_i = \mathbf{X}_{iL} + \ldots + \mathbf{X}_{Ip}$. Estas matrizes são calculadas para $I = I_1 + \ldots + I_m$ e a expansão leva à decomposição representados por,

$$\mathbf{X} = \left[\mathbf{X}_{I1} + \mathbf{X}_{I2} + \ldots.. + \mathbf{X}_{\mathrm{Im}}\right], \tag{3}$$

o resultado desse passo é a representação da matriz trajetória como uma soma de matrizes resultantes ($\mathbf{X}_{I1} \ldots \ldots \mathbf{X}_{\mathrm{Im}}$). A escolha dos conjuntos $I_1, \ldots, I_m$ é baseada na cálculo da correlação ponderada entre pares de subséries, da seguinte forma,

$$\rho_{i,j}^{(w)} = \sum_{q=1}^{N} w_q \mathbf{Y}_q^i \mathbf{Y}_q^j \Big/ \left(\sqrt{\left\langle \mathbf{Y}_t^i, \mathbf{Y}_t^i \right\rangle_w}\right)\left(\sqrt{\left\langle \mathbf{Y}_t^j, \mathbf{Y}_t^j \right\rangle_w}\right), \tag{4}$$

onde os pesos w_c são definidos por; $w_c = mín\{c, L, N\text{-}c\}$, e assume-se que $L \leq N/2$. Se o valor absoluto da correlação ponderada for muito pequeno, diz-se que as duas séries são quase ortogonais (independentes e separáveis). A operação realizada neste último passo resulta em uma aproximação para os componentes da série original, ou seja, transforma cada matriz da decomposição agrupada, equação (3), em uma nova série de tamanho N, que pode ser considerada como uma aproximação da série original.

Diagonalização - O segundo passo da reconstrução transforma cada matriz resultante em uma nova série temporal, as quais representam os componentes (tendência, harmônicos e ruídos) presentes na série inicial. Isto é feito através de uma operação chamada diagonalização, que é uma operação linear aplicada na matriz trajetória e que a transforma na série inicial. Chamando de $\mathbf{Y}$ a matriz $L \times K$ com elementos y_{ij}, $1 \leq i \leq L$, $1 \leq j \leq K$. Definindo $L^* = mín\ (L, K)$, $K^* = máx\ (L, K)$ e $N = L + K - 1$. Fazendo $y^*_{ij} = y_{ij}$ se $L < K$ e $y^*_{ij} = y_{ji}$ caso contrário. A matriz $\mathbf{Y}$ é a matriz trajetória de uma série ($h_0, \ldots, h_{N\text{-}1}$) em outras palavras, se $\mathbf{Y}$ é uma matriz de Hankel, então $g_i = h_i$ para todo i. A diagonalização aplicada a matriz resultante $\mathbf{X}_{\mathrm{Ik}}$ produz a série $\tilde{F}^{(K)} = \tilde{F}_0^{(K)}, \ldots, \tilde{F}_{N-1}^{(K)}$ e, portanto, a série inicial $f_0, \ldots, f_{N-1}$ é decomposta em uma soma de m series.

$$f_n = \sum_{k=1}^{m} \tilde{f}_n^{(k)} \tag{5}$$

Tipos de falhas de dados e escolha das séries de testes

Após o preenchimento da série temporal da estação Santa Rosa pelos dois métodos, os valores estimados pelo método SSA foram a ela incorporados. Após esse procedimento, a série temporal completa foi modificada, criando artificialmente falhas de dados. Foram criadas cinco novas séries temporais (*Série 1*, *Série 2*, *Série 3*, *Série 4* e *Série 5*) no intuito de verificar o desempenho dos dois métodos no preenchimento de séries temporais com falhas de dados disposta de formas distintas, isto é, até que ponto eles são capazes de estimar valores aceitáveis para substituir os ausentes.

Estimativa dos erros e validação do preenchimento

A escolha de uma métrica é o critério utilizado para julgar se um método descreve de forma satisfatória um conjunto de medições, isto é, quão próximo um valor estimado está do valor observado. Embora o erro médio quadrático (EMQ) tenha sido tradicionalmente usado como métrica e citado na literatura de estudos climatológicos e ambientais, Willmott e Matsuura (2005) mostraram que existe um inconveniente na sua utilização. O valor calculado do erro não é representativo apenas dos erros médios, também varia com a magnitude da distribuição dos erros, com a raiz quadrada do número de erros e com magnitude do erro médio absoluto. O presente trabalho utiliza como métrica o erro médio absoluto (*EMA*) pois este proporciona um melhor indicador da qualidade do ajuste e sobre o desempenho global médio. O *EMA* é calculado pela seguinte equação,

$$EMA = \left[\frac{1}{N} \sum_{i=1}^{N} \left| x_i^{obs} - x_i^{est} \right| \right], \tag{6}$$

onde N é o número de observações, x_i^{obs} é o valor observado, x_i^{est} o valor estimado pelo método de preenchimento. O erro médio (*EM*) definido como o valor médio dos resíduos entre os valores observados e estimados, também foi utilizado neste estudo. O *EM* é calculado pela seguinte equação,

$$EM = \left[\frac{1}{N} \sum_{i-1}^{N} x_i^{obs} - x_i^{est} \right] \tag{7}$$

O *EM* avalia, em média, se o modelo tende a superestimar ou subestimar a magnitude dos valores faltantes. O grau de relação entre os valores estimados e observados foi verificado com a análise de correlação linear e a significância da relação foi medida pelo *p*-valor calculado para um nível de significância $\alpha = 0{,}05$.

Regressão Linear Simples

A análise de regressão é um dos métodos matemáticos mais utilizados para analisar relação entre variáveis. O relacionamento é estabelecido por uma equação matemática que associa a variável explicada com as variáveis explicativas. A regressão linear simples definida pela equação (8) estabelece a relação linear entre duas variáveis,

$$Y = \alpha + \beta X + \varepsilon \tag{8}$$

onde X é a variável explicativa medida sem erro (não aleatoria), Y é a variável explicada (aleatória), ε é a variável aleatória residual na qual se procuram incluir todas as influências no comportamento da variável Y que não podem ser explicadas pelo comportamento linear da variável X. α e β são parâmetros desconhecidos do modelo (a estimar). Mais detalhes sobre o método e inferência estatística dos parâmetros (R^2, F, t, e p-valor) da equação de regressão podem ser obtidos em Draper (1981) e em Wittink (1988).

RESULTADOS E DISCUSSÕES

Regressão Linear Simples

A estação Santa Rosa possui 334 observações (N=334) correspondentes ao período de agosto de 1982 a setembro de 2011. A série temporal possui 16 falhas, das quais 9 nos meses de janeiro a setembro do ano hidrológico 91/92, 5 nos meses de outubro a fevereiro do ano hidrológico 03/04 e falhas isoladas em dezembro de 1997 e fevereiro de 2009. Para preencher falhas de dados com o método de *RL* (simples ou múltipla), exige-se séries temporais completas como referência. Entre as estações consideradas próximas, esta condição somente foi satisfeita pela série temporal da estação Núcleo Ariel. Portanto, a série de precipitação acumulada mensal da estação Santa Rosa foi preenchida com o método de *RLS* utilizando os valores de precipitação observados na estação Núcleo Ariel. A equação (9) descreve a relação linear entre as séries das estações Santa Rosa e Núcleo Ariel.

$$P_{sr}^{est}(mm) = 0{,}786 P_{na}^{obs}(mm) + 51{,}70(mm) \tag{9}$$

Onde P_{sr}^{est} é a precipitação acumulada mensal estimada para a estação Santa Rosa, e P_{na}^{obs} é a precipitação acumulada mensal observada na estação Núcleo Ariel. A equação (9) informa que, em média, quando a precipitação é zero na estação Núcleo Ariel, a precipitação na estação Santa Rosa é aproximadamente 52*mm*, e a taxa de variação da precipitação na estação Santa Rosa é de aproximadamente 0,79*mm* quando a precipitação varia uma unidade na estação Núcleo Ariel. Esta informação é razoável, já que as estações estão a 140*Km* de distância uma da outra (ver Figura 1). A Figura 2 mostra o gráfico de dispersão dos valores da precipitação e os resultados da inferência estatística utilizada para avaliar a significância da relação entre os valores observados de precipitação nas duas estações e dos parâmetros R^2, α e β da equação de regressão.

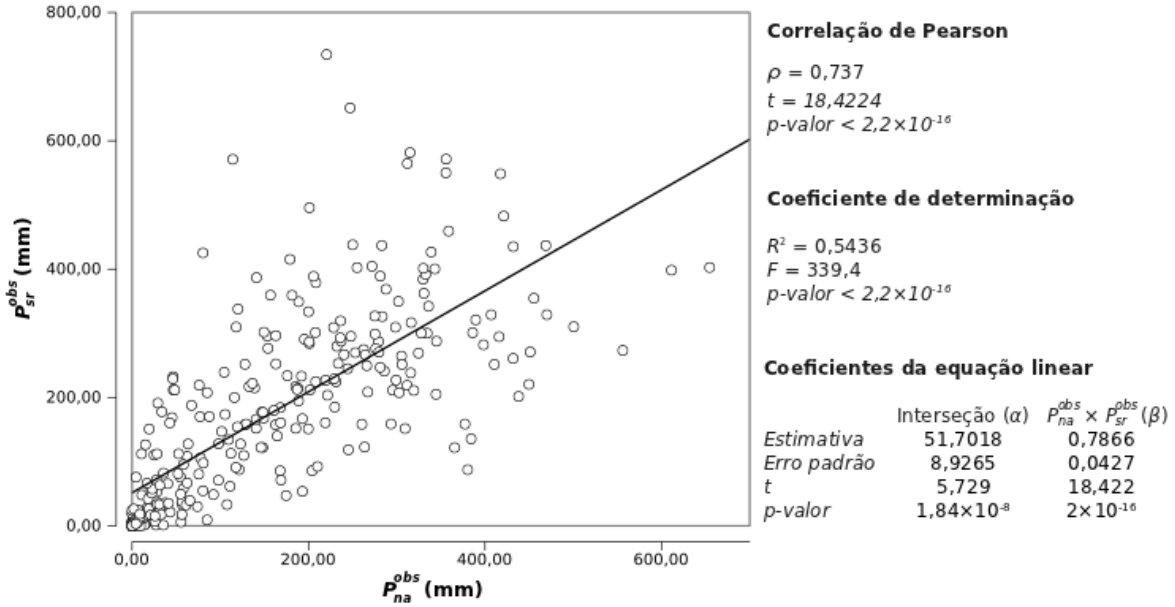

Figura 2: Gráfico de dispersão e parâmetros estatísticos (Santa Rosa X Núcleo Ariel)

O coeficiente de correlação de Pearson (ρ_{ij}) mede o grau da relação entre os valores da precipitação observados nas estações Núcleo do Ariel (*j*) e Santa Rosa (*i*). O valor $(\rho = 0,737)$ indica que existe uma relação linear substancial entre os valores observados da precipitação nas duas estações. O coeficiente de determinação R^2 mede a variação que pode ser explicada pela equação de regressão linear (9) e a variação total da precipitação na estação Núcleo do Ariel. O coeficiente de determinação da reta de regressão, $(R^2 = 0,544)$ indica que apenas 54% dos valores observados na estação Santa Rosa podem ser explicados pelos valores observados na estação Núcleo Ariel. Em outras palavras, 48% dos valores observados não podem ser explicados pela equação de regressão (9). A significância estatística dos valores estimados para ρ_{ij}, R^2, α e β foi verificada com testes de hipóteses usando os testes *t* e *F* e o *p*-valor. Os resultados implicam na rejeição da hipótese H_0, ou seja, pode-se afirmar com 95% de confiança que: $(\rho = 0,737)$, $(R^2 = 0,544)$, $(\alpha = 51,70)$, e $(\beta = 0,786)$ são estatisticamente significativos e a equação (9) é a melhor equação para descrever a relação linear entre as duas séries temporais. Após os dados faltantes terem sidos estimados com a equação linear (9), foram calculados os *EMAs* e *EMs*. O *EM* indicou que o método de *RLS* tende, em média, a subestimar a magnitude dos valores faltantes em 19,56*mm*. *EMA* indicou que o erro médio cometido ao estimar os valores de precipitação pela equação de regressão linear é de 84,59*mm*. Os erros foram considerados aceitáveis já que a precipitação mínima e máxima foram de 0*mm* e 734,5*mm*, respectivamente.

Análise de Espectro Singular

A Figura 3(a) mostra a série temporal da precipitação acumulada mensal original. Nota-se um comportamento irregular nos períodos de maior precipitação, e ocorrência de valores de precipitação iguais a zero, geralmente nos meses de junho, julho e/ou agosto caracterizando o período de estiagem na região. O tamanho da janela, *L*=48, foi escolhido conforme sugerido por Hassani (2007) e Golyandina e Osipov (2007). A decomposição da serie original resultou em um conjunto de 115 vetores defasados completos (*K*=115). A separação entre as componentes de tendência, harmônicos e ruídos foi realizada a partir do resultado da equação (4). O autovetor |1> é um componente de tendência da série e mostra que há uma tendência de decréscimo da precipitação ao longo do tempo (Figura 3b), enquanto os autovetores |2> e |3> são componentes harmônicos com periodicidade de 12 meses (Figura 3c e 3d).

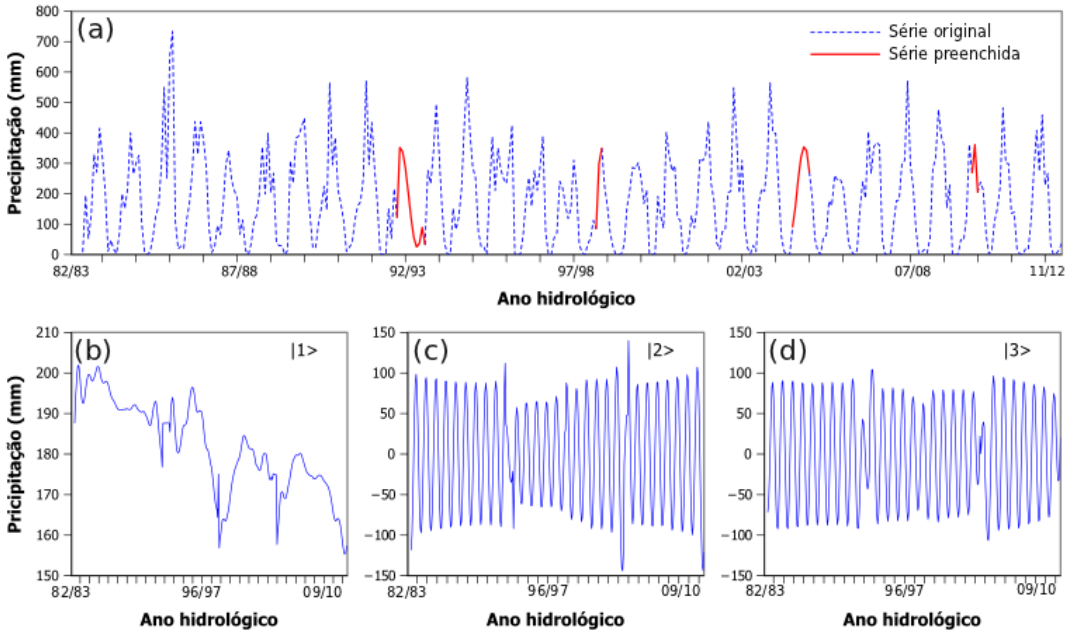

Figura 3: (a) Série temporal de precipitação da estação Santa Rosa (b) tendência (c) e (d) harmônicos

A periodicidade anual é atribuída ao ciclo cheia/estiagem e revela as variações sazonais de comportamento não linear. Estes autovetores, utilizados na reconstrução da série, reproduziram o comportamento da série original nos intervalos de tempo onde não foram realizadas as medições (Figura 3 (a)). O *EM* indicou que o método *SSA* tende, em média, a superestimar a magnitude dos valores faltantes em 2,95*mm*, e o *EMR* indicou que o erro médio cometido ao estimar os valores de precipitação com o método *SSA* é de 59,55*mm*.

Apesar dos *EMR* e *EM* estimados ao preencher as falhas de dados com o método de *RL* terem magnitudes maiores que os respectivos erros ao preencher as falhas com método *SSA*, os erros estimados para os dois métodos estão na mesma ordem de grandeza. Analisando somente os erros pode-se concluir que os dois métodos produziram resultados similares e que poderiam ser utilizados com o mesmo grau de confiança. Porém, análise de regressão deve levar em conta todos os parâmetros estatísticos que informam sobre o grau da relação entre os registros das duas séries utilizadas na regressão e a significância estatística dos parâmetros calculados.

Tipos de falhas de dados e comparação entre os métodos RLS e SSA

Os resultados apresentados a seguir, correspondem à segunda etapa de análise, na qual foram criadas diferentes tipos de falhas de dados na série temporal da estação Santa Rosa. A Tabela 1 mostra as equações de regressão utilizadas para preencher as falhas de dados e os resultados dos testes estatísticos utilizada para avaliar a significância da regressão.

Tabela 1: **Equações de regressão das séries sintéticas**

Séries	***N***	***Equação***	R^2	***F***	***p*-valor**
$P_{na}^{obs} \times P_{série1}^{obs}$	268	$P_{sr}^{est}(mm) = 0,799 P_{na}^{obs}(mm) + 53,85(mm)$	0,5643	346,8	$2,2 \times 10^{-16}$
$P_{na}^{obs} \times P_{série2}^{obs}$	272	$P_{sr}^{est}(mm) = 0,756 P_{na}^{obs}(mm) + 55,52(mm)$	0,5420	321,7	$2,2 \times 10^{-16}$
$P_{na}^{obs} \times P_{série3}^{obs}$	251	$P_{sr}^{est}(mm) = 0,786 P_{na}^{obs}(mm) + 47,20(mm)$	0,5723	335,5	$2,2 \times 10^{-16}$
$P_{na}^{obs} \times P_{série4}^{obs}$	276	$P_{sr}^{est}(mm) = 0,795 P_{na}^{obs}(mm) + 48,02(mm)$	0,5518	339,6	$2,2 \times 10^{-16}$
$P_{na}^{obs} \times P_{série5}^{obs}$	226	$P_{sr}^{est}(mm) = 0,821 P_{na}^{obs}(mm) + 36,34(mm)$	0,5665	295,1	$2,2 \times 10^{-16}$

Como pode-se observar na Tabela 1, a variação no número de falhas de dados na série, praticamente não alterou o valor do coeficiente de determinação da reta, R^2. A Figura 4 mostra as séries sintéticas com as falhas de

dados, e o respectivo preenchimento pelo método *SSA*. Para preencher as falhas foram utilizados os três primeiros autovetores, |1>, |2> e |3>, extraídos na etapa da decomposição da série.

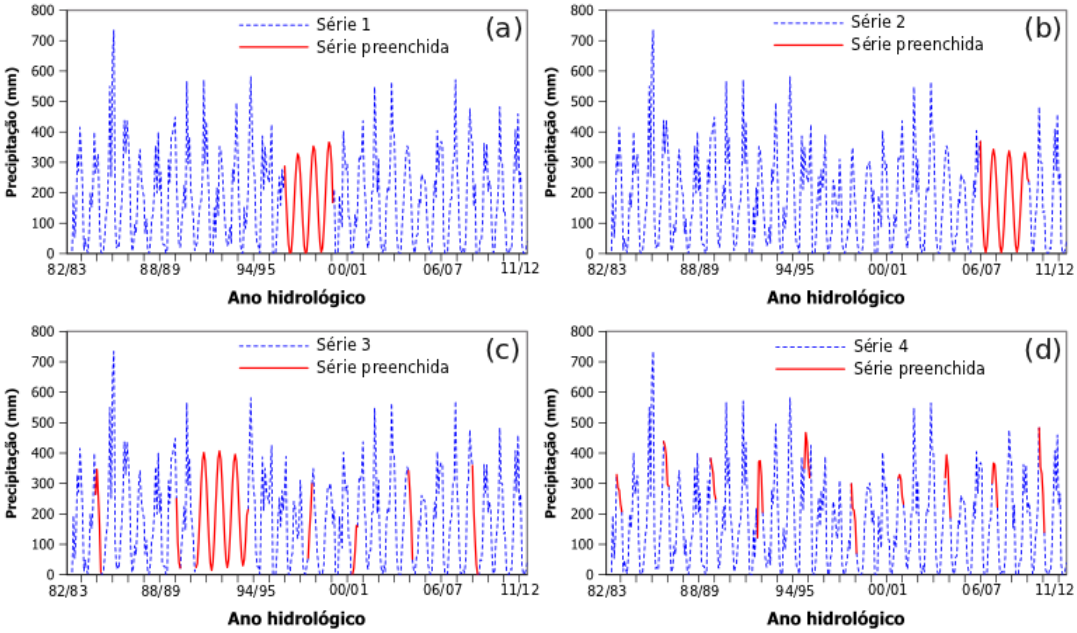

Figura 4: Séries sintéticas da estação Santa Rosa preenchidas com o método *SSA*

A Tabela 2 mostra que a magnitude dos erros tende a aumentar com o aumento do número de falhas de dados. O método *SSA* tendeu a apresentar erros menores que o método *RLS*, com exceção da *Série 5*. No entanto os erros estão na mesma ordem de grande.

Tabela 2: Erros do preenchimento das falhas das séries sintéticas

Método	Parâmetro	Série temporal precipitação acumulada mensal				
		Série 1	*Série 2*	*Série 3*	*Série 4*	*Série 5*
SSA	*EMA*	60,92	57,63	62,03	76,38	268,61
	EM	-34,24	5,81	-24,19	4,28	268,61
RL	*EMA*	78,73	74,87	68,91	88,62	107,45
	EM	-39,65	4,83	24,00	21,34	36,73

A Figura 5 mostra a série sintética 5 e a série preenchida com os três primeiros autovetores. Esta série foi enfatizada, pois nota-se claramente que o preenchimento não mais reproduz o comportamento da série. Há uma inversão dos valores estimados pelo método *SSA*. A inversão dos valores estimados deve-se a perda de informação provocada pelas falhas de dados consecutivas que foram criadas. Em termos do modelo *SSA* estas falhas provocaram a perda dos componentes harmônicos com periodicidade de 12 meses, tornando, portanto, mais difícil a separação do sinal e do ruído, o que influenciou na etapa de reconstrução da série.

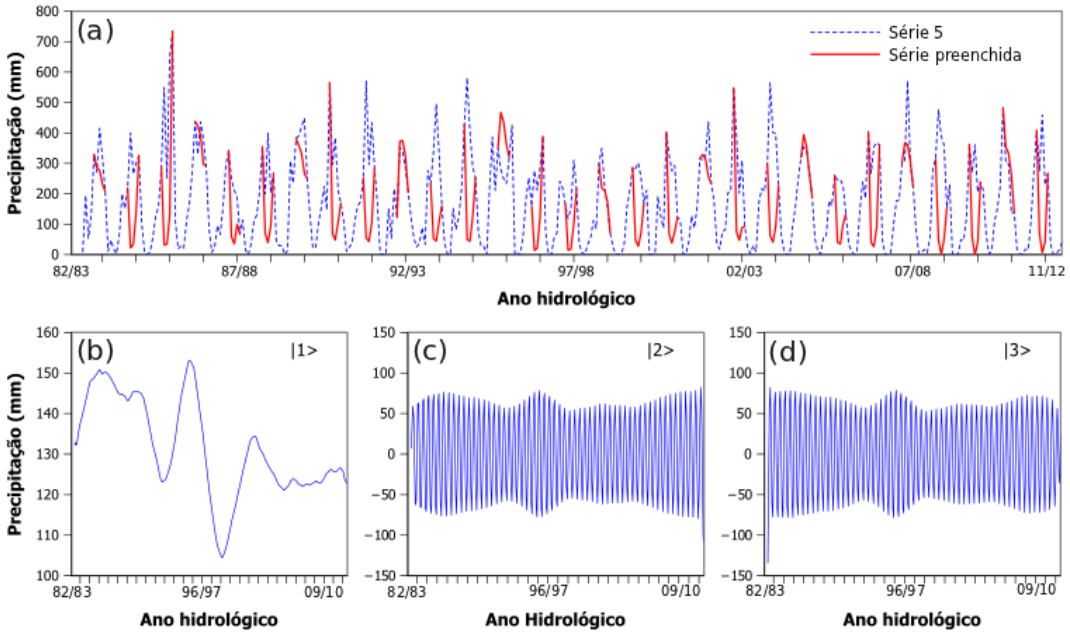

Figura 5: Preenchimento da série temporal da estação Santa Rosa (*SSA*)

Novamente, a separação entre as componentes de tendência, harmônicos e ruídos foi realizada a partir do resultado da equação (4). O autovetor |1> foi interpretado como a componente de tendência da série, porém há uma considerável perda de informação, como pode ser observado na Figura 5(b). Os autovetores |2> e |3> são componentes harmônicos de periodicidade de 6 meses - Figura 5c e 5d. Ressalta-se que não foi possível obter autovetores de periodicidade de 12 meses, que são característicos da região em estudo. O déficit de informação devido a ausência desta frequência influenciou na qualidade da série reconstruída, o que está de acordo com (MUSIAL, 2011). Assim, as falhas de dados prolongadas e regulares provocaram um déficit de informação, com a perda da freqüência de 12 meses. Neste caso, não é possível recuperar a informação perdidas.

CONCLUSÕES

A maior parte dos estudos brasileiros sobre preenchimento de falhas de dados de variáveis ambientais utiliza métodos de regressão linear e/ou média móvel, ponderação regional e krigagem. Porém, para que esses métodos sejam utilizados de forma eficiente, necessitam de séries temporais longas (geralmente mais de 10 anos) e completas em pelo menos três estações próximas. Não obstante, as estações devem estar em uma região climatológica semelhante àquela da estação que se deseja preencher. Tais condições são um fator limitante na grande maioria dos estudos que utilizam métodos clássicos. Além disso, quando o método de regressão é utilizado, geralmente o R^2 é o único parâmetro estatístico levado em consideração na análise da relação entre as variáveis. Este estudo mostra que outros parâmetros estatísticos relevantes devem ser inclusos na análise. Outro fator importante a considerar é que as observações são cíclicas e, consequentemente, não podem ser descritas por uma função de distribuição normal.

Os resultados do presente estudo mostram que o método *SSA* pode ser utilizado para preencher falhas de dados de séries incompletas. Ressalta-se algumas das principais vantagens do método *SSA* em relação ao método de *RLS*: a) a primeira deve-se a sua natureza auto-adaptativa que dispensa a necessidade de séries temporais de estações próximas; b) o preenchimento é realizado durante o processo de reconstrução da série usando exclusivamente os valores observados da série; c) o método *SSA*, é capaz de extrair informações sobre a tendência e variações periódicas existente na série; d) capacidade para reproduzir o comportamento não linear dos valores observados. A série com falhas regulares apresentou perda de informação e o preenchimento pelo método *SSA* não foi satisfatório.

Os resultados são promissores e demonstram a potencialidade do método *SSA* para analisar séries temporais completas (Destro, 2012) e preencher falhas de dados em séries incompletas de variáveis ambientais (Destro,

2010). O método *SSA*, como demonstrado neste estudo, pode ser uma ferramenta matemática útil na recuperação de informações perdidas, maximizando a qualidade da informação produzida.

REFERÊNCIAS BIBLIOGRÁFICAS

1. BOX, G.E.P.; PIERCE, D. A. (1970) Distribution of residual correlations in autoregressive-integrated moving average time series models. Journal of the American Statistical Association, v. 65, p. 1509-1526.
2. BRILLINGER, D. R. (2001) Time Series: Data analysis and Theory. Classics in Applied Mathematics. SIAM.
3. BROCKWELL, P.J.; DAVIS, R.A. (2002) Introduction to Time Series and Forecasting. 2nd ed. New York: Springer Texts in Statistics.
4. BROOMHEAD, D. S.; KING, G. P. (1986a) Extracting qualitative dynamics from experimental data. Physica D. 20, p. 217-236.
5. BROOMHEAD, D. S.; KING, G. P. (1986b) On the qualitative analysis of experimental dynamical systems. In: SARKAR, S. (Ed.) Nonlinear Phenomena and Chaos. Adam Hilger, Bristol. p. 113-144.
6. DESTRO, C. A. M.; ZEILHOFER, P.; LIMA, G. A. R. (2012) Análise de Séries Temporais de Vazão Média Mensal do Rio Cuiabá Através do Método de Análise de Espectro Singular. Revista Brasileira de Recursos Hídricos. v. 17, n. 2 – Abr/Jun 2012. p. 111-120.
7. DESTRO, C. A. M. (2010) Séries Temporais de Variáveis Ambientais: Análise, Preenchimento e Previsão Utilizando o Método de Análise de Espectro Singular (SSA). Dissertação de mestrado (Programa de Pós-Graduação em Recursos Hídricos), UFMT, Cuiabá.
8. DRAPER, N.R.; SMITH, H. (1981) Applied Regression Analysis, 2 ed, Wiley, New York, 709p.
9. GHIL, M.; ALLEN, M. R.; DETTINGER, M. D.; IDE, K.; KONDRASHOV, D.; MANN, M. E.; ROBERTSON, A. W.; SAUNDERS, A.; TYAN Y.; VARADI, F.; YIOU, P. (2002) Advanced spectral methods for climatic time series. Review of Geophysics, 40 (1), p. 1–41.
10. GOLYANDINA, N.; NEKRUTKIN, V.; ZHIGLJAVSKY, A. (2001) Analysis of Time Series Structure: SSA and Related Techniques. Chapman & Hall/CRC, London. 305p.
11. GOLYANDINA, N.; OSIPOV, E. (2007). The "Caterpillar"-SSA method for analysis of time series with missing values. Journal of Statistical Planning and Inference. 137, p. 2642–2653.
12. HAIR, J. F.; TATHAM, R. L.; ANDERSON, R. E.; BLACK, W. (2005) Análise Multivariada de Dados. 5ª edição, Bookman, 600p.
13. HARMANCIOGLU, N. B.; FISTIKOGLU, O.; OZKUL, O. D.; SINGH, V. P.; ALPASLAN, M. N. (1999) Water quality Network Design. Kluwer Academic Publishers, Water Science and Technology Library, v. 33. Dordrecht.
14. HASSANI, H. (2007) Singular Spectrum Analysis: Methodology and Comparison. Journal of Data Science. 5, p. 239-257.
15. LITTLE, R. J. A.; RUBIN, D. B. (2002) Statistical Analysis with Missing Data. (2nd ed.) New York, Wiley Publisher.
16. KONDRASHOV, D.; FELIKS, Y.; GHIL, M. (2005) Oscillatory modes of extended Nile River records (A.D. 622-1922). Geophysical Research Letters. 32 (10), L10702.
17. KONDRASHOV, D.; GHIL, M. (2006) Spatio-temporal filling of missing points in geophysical data sets. Nonlinear Processes in Geophysics. v.13, p. 151-159.
18. MUSIAL, J. P.; VERSTRAETE M. M.; GOBRONAND N. (2011) Technical Note: Comparing the effectiveness of recent algorithms to fill and smooth incomplete and noisy time series. Atmospheric Chemistry and Physics, 11, p. 7905–7923,
19. QUIMPO, R. G.; YANG, J-Y (1970) Sampling Consideration in Stream Discharge and Temperature Measurements, Water Resources Research, v.6 (6), p. 1711-1774.
20. SCHOELLHAMER, D. H. (2001) Singular Spectrum Analysis for time series with missing data. Geophysical Research Letters. 28 (16), p. 3187-3190.
21. SEARCY, J. K.; HARDISON, C. H. (1960) Double Mass Curves, U.S. Geological Survey Water-Suplay Paper 1541-B, USGS, Washington, 66p.
22. WITTINK, D. R. (1988) The Application of Regression Analysis, Allyn and Bacon, 324p.
23. WILLMOTT, C. J.; MATSUURA, K. (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, v. 30, p. 79–82.