Revista Brasileira de Meteorologia, **v.29**, **n.4**, 527 - 536, 2014
http://dx.doi.org/10.1590/0102-778620130005

IMPUTAÇÃO DE DADOS CLIMÁTICOS UTILIZANDO A DECOMPOSIÇÃO POR VALORES SINGULARES: UMA COMPARAÇÃO EMPÍRICA

MARISOL GARCÍA-PEÑA; SERGIO ARCINIEGAS-ALARCÓN; DÉCIO BARBIN

Universidade de São Paulo, Escola Superior de Agricultura "Luiz de Queiroz" (USP/ESALQ), Piracicaba, SP, Brasil

luzmara@gmail.com; sergio.arciniegas@gmail.com; decio.barbin@usp.br

Recebido Julho de 2013 - Aceito Janeiro de 2014

RESUMO

Um problema comum em dados climáticos é a informação ausente. Recentemente, foram desenvolvidos quatro métodos de imputação que têm como base a decomposição por valores singulares de uma matriz (DVS). O objetivo deste artigo é avaliar os novos desenvolvimentos fazendo uma comparação por meio de um estudo de simulação baseado em duas matrizes completas de dados reais. Uma matriz corresponde à precipitação histórica de Piracicaba/SP – Brasil, enquanto a outra matriz corresponde às características meteorológicas multivariadas na mesma cidade desde o ano 1997 até 2012. No estudo foram feitas retiradas aleatórias de diferentes porcentagens com posterior imputação, comparando as metodologias através de três critérios: a raiz quadrada normalizada do erro quadrático médio, a estatística de similaridade de Procrustes e o coeficiente de correlação não paramétrico de Spearman. Concluiu-se que a DVS deve ser utilizada unicamente quando sejam analisadas matrizes multivariadas e, no caso de matrizes de precipitação, a imputação pela média mensal supera o desempenho de métodos baseados na DVS.
Palavras-chave: Imputação, DVS, observações ausentes.

ABSTRACT CLIMATE DATA IMPUTATION USING THE SINGULAR VALUE DECOMPOSITION: AN EMPIRICAL COMPARISON

A common problem in climate data is missing information. Recently, four methods have been developed which are based in the singular value decomposition of a matrix (SVD). The aim of this paper is to evaluate these new developments making a comparison by means of a simulation study based on two complete matrices of real data. One corresponds to the historical precipitation of Piracicaba / SP - Brazil and the other matrix corresponds to multivariate meteorological characteristics in the same city from year 1997 to 2012. In the study, values were deleted randomly at different percentages with subsequent imputation, comparing the methodologies by three criteria: the normalized root mean squared error, the similarity statistic of Procrustes and the Spearman correlation coefficient. It was concluded that the SVD should be used only when multivariate matrices are analyzed and when matrices of precipitation are used, the monthly mean overcome the performance of other methods based on the SVD.
Keywords: Imputation, SVD, missing values.

1. INTRODUÇÃO

Frequentemente, nos estudos de climatologia são necessárias observações completas (sem falta de informação) para realizar as análises apropriadas dos dados que são coletados durante um determinado período de tempo, em diferentes estações de uma ou várias regiões de interesse. É comum encontrar séries de dados climatológicos com dados omissos devido a varias razões, como falhas dos instrumentos de medição, condições climáticas extremas e erros na digitação. Uma maneira muito comum de analisar dados provenientes de estudos com informação faltante é imputar as observações ausentes e posteriormente, aplicar procedimentos clássicos sobre os dados completados (observados + imputados). Um método amplamente usado na literatura é utilizar a média como imputação.

Existem muitos métodos estatísticos para resolver o problema de dados omissos como Little e Rubin (2002), Schafer (1997), McLachlan e Krishnan (1997), Tanner e Wong (1987). Outras metodologias com desenvolvimentos recentes são os métodos de imputação simples e múltipla. Eles são uma alternativa às aproximações baseadas em verossimilhanças e são recomendados por alguns autores (Rubin, 1987; Allison 2001; Schafer, 1997; Schafer e Graham, 2002). Em meteorologia, os primeiros estudos usavam estimações de regressão linear simples ou simplesmente a média (Paulhus e Kohler, 1952). Outros autores compararam métodos de imputação como a análise discriminante múltipla, regressão linear múltipla, razão normal e os algoritmos de Esperança-Maximização ou EM (Young, 1992; Makhuvha et al., 1997; Xia et al. 1999a,b). Estudos recentes usaram aproximações para completar séries de temperatura (mínima e máxima) e precipitação total, como descrito em Eischeid et al. (2000). Entretanto, Teegavarapu e Chandramouli (2005), usaram o método modificado da distância entre as estações com dados faltantes e as estações de referência para estimação; também são usados conceitos de redes neurais, interpolação estocástica e kriging. Schneider (2001) trabalhou com o algoritmo EM regularizado e sugere seu uso quando o número de variáveis seja maior do que o tamanho da amostra. Cano e Andreu (2010), estudaram a imputação múltipla definindo uma estratégia de construção da base de dados (usando *lags* como variáveis de suporte e desta maneira criar uma aproximação para manejar o efeito da distância) para evitar ruído nas simulações.

Outros autores usaram métodos alternativos para a imputação de dados como o mapa de auto-ordenamento, perceptron multicamada, método do vizinho mais próximo (Junninen et al., 2004; Ramos-Calzado et al., 2008; Coulibaly e Evora, 2007; Lucio et al., 2007; Kalteh e Berndtsson, 2007; Kalteh e Hjorth, 2009). Smith e Aretxabaleta (2007) apresentaram um método alternativo à análise da função ortogonal empírica, baseado no algoritmo EM para estimar os parâmetros de um modelo de mistura Gaussiano, indicando, também, que este método permite uma interpretação clara da variabilidade associada com cada regime. Aly et al. (2009) compararam quatro métodos de interpolação estocástica e determinística para imputar dados faltantes em precipitação diária, por outro lado, Lo Presti et al. (2010), propuseram um método em duas etapas, a primeira, consiste em identificar as estações vizinhas e similares à estação com dados faltantes (coeficiente de similaridade) e a segunda, usar um método de regressão para realizar as estimações. Existem outros estudos no âmbito de imputação de dados como os descritos em Yozgatligil et al. (2013).

Os métodos de imputação utilizados neste estudo serão a Média, EM-DVS, imputação biplot, imputação por meio de uma aproximação de posto inferior ponderada (ou EMSJ) e imputação GabrielEigen. A comparação será feita através da raiz quadrada normalizada do erro quadrático médio, a estatística de similaridade de Procrustes e o coeficiente de correlação de Spearman. Neste trabalho, serão criados três cenários de dados omissos, 10%, 20% e 40% para os dados descritos na subseção 2.2. Depois, esses valores são estimados usando os métodos já mencionados e para os dados completados (observados + imputados) são calculados e comparados os valores dos critérios.

2. MATERIAL E MÉTODOS

2.1 A decomposição por valores singulares de uma matriz (DVS)

Esta ferramenta é a base que utiliza a maioria dos métodos de imputação considerados neste estudo, daí, a importância para apresentá-la inicialmente. Qualquer matriz **S** (n x p) pode ser decomposta por valor singular na forma $\mathbf{S}=\mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$, em que $\mathbf{U}^{\mathrm{T}}\mathbf{U}=\mathbf{V}^{\mathrm{T}}\mathbf{V}=\mathbf{V}\mathbf{V}^{\mathrm{T}}=\mathbf{I_p}$ e $\mathbf{D}=diag(d_1, \ldots, d_p)$ com $d_1 \geq d_2 \geq, \ldots, \geq d_p \geq 0$. As matrizes $\mathbf{S}^{\mathrm{T}}\mathbf{S}$ e $\mathbf{S}\mathbf{S}^{\mathrm{T}}$ têm os mesmos autovalores e os elementos d_i são a raiz quadrada destes autovalores; a i-ésima coluna $\mathbf{v}_{\mathrm{i}}=(v_{i1}, \ldots, v_{ip})^{\mathrm{T}}$ da matriz **V** (p x p) é o autovetor correspondente ao i-ésimo maior autovalor d_i^2 de $\mathbf{S}^{\mathrm{T}}\mathbf{S}$; enquanto a j-ésima coluna $\mathbf{u}_{\mathrm{j}}=(u_{1j}, \ldots, u_{nj})^{\mathrm{T}}$ da matriz **U** (n x p) é o autovetor correspondente ao j-ésimo maior autovalor d_j^2 de $\mathbf{S}\mathbf{S}^{\mathrm{T}}$. A decomposição tem sua representação elementar como na Equação 1 (Krzanowski, 1988)

$$s_{ij} = \sum_{h=1}^{p} u_{ih} d_h v_{jh} \qquad (1)$$

2.2 Métodos de imputação

***Média*:** É um método de imputação simples muito usado nas ciências sociais e consiste em imputar cada valor ausente pela média da correspondente variável ou coluna se os dados estão arranjados em uma matriz **X** de dimensão (n x p) com n linhas e p colunas. É um método rápido, mas dependendo do conjunto de dados e da porcentagem de dados faltantes pode apresentar desvantagens, tais como a compressão da distribuição das variáveis e a distorção das relações entre elas (Durrant, 2005).

***EM-DVS*:** Perry (2009) apresenta um método de imputação que mistura o algoritmo EM (Esperança-Maximização) com a decomposição em valores singulares de uma matriz (DVS). A metodologia é apresentada a seguir. Considere a matriz **A** de dimensão (n x p) com elementos A_{ij} ($i=1,\ldots,n; j=1,\ldots,p$) e com alguns deles faltantes.

EM-DVS passo 1: Seja $I = \{(i,j): A_{ij} \neq *\}$, ou seja, o conjunto I representa todos os valores observados.

EM-DVS passo 2: Para $1 \leq j \leq p$ seja μ_j a média dos valores não faltantes na coluna j de **A** ou 0 se todas as caselas na coluna j são faltantes.

EM-DVS passo 3: Defina $\mathbf{A}^{(0)}$ por

$$A_{ij}^{(0)} = \begin{cases} A_{ij} \text{ se } (i,j) \in I \\ \mu_j \text{ em caso contrário} \end{cases}$$

EM-DVS passo 4: Iniciar a contagem das iterações em zero, $N \leftarrow 0$.

EM-DVS passo 5: É feita a maximização calculando a DVS de $\mathbf{A}^{(N)}$, assim:

$$\mathbf{A}^{(N)} = \sum_{i=1}^{p} d_i^{(N)} \mathbf{u}_i^{(N)} \mathbf{v}_i^{(N)T}$$

e calculando $\mathbf{A}_k^{(N)}$, ou seja a DVS truncada com k termos, tal que:

$$\mathbf{A}_k^{(N)} = \sum_{i=1}^{k} d_i^{(N)} \mathbf{u}_i^{(N)} \mathbf{v}_i^{(N)T}$$

EM-DVS passo 6: É calculada a esperança definindo a matriz $\mathbf{A}^{(N+1)}$ de dimensão (n x p) como

$$A_{ij}^{(N+1)} = \begin{cases} A_{ij} \text{ se } (i,j) \in I \\ A_{k,ij}^{(N)} \text{ em caso contrário} \end{cases}$$

EM-DVS passo 7: Calcular $RSS(N) = \left\| \mathbf{A} - \mathbf{A}_k^{(N)} \right\|_{F,I}^2$, se $|RSS^{(N)} - RSS^{(N-1)}|$ é pequena, declare convergência e obtenha a matriz $\mathbf{A}_k^{(N)}$ que conterá as imputações dos valores ausentes. Caso contrário, aumente $N \leftarrow N+1$ e volte para **EM-DVS passo 5.**

***Imputação biplot*:** Recentemente, Yan (2013) descreveu um método de imputação utilizando a DVS e a qual é a técnica básica para análise biplot (Gabriel, 1971; 2002). O método é apresentado a seguir.

Biplot passo 1. Considere a matriz **X** de dimensão (n x p) com elementos x_{ij} ($i=1,...,n$; $j=1,...,p$), em que alguns deles são ausentes x_{ij}^{aus}. Inicialmente, os dados faltantes são imputados pela média dos valores observados na respectiva coluna, obtendo assim uma matriz **X** completada.

Biplot passo 2. As colunas da matriz **X** completada são padronizadas subtraindo de cada elemento m_j e dividindo o resultado por s_j (em que m_j e s_j representam a média e desvio padrão da j-ésima coluna). Os elementos padronizados serão notados por p_{ij} e modelados por meio de um biplot bidimensional, ou seja:

$$p_{ij} = \frac{(x_{ij} - m_j)}{s_j} = \sum_{k=1}^{2} \lambda_k \alpha_{ik} \gamma_{jk} + \varepsilon_{ij}$$

Os valores p_{ij} são decompostos em dois componentes principais (CP), com valores singulares λ_k, autovetores para as linhas α_{ik} e autovetores para as colunas γ_{jk} para cada um dos k-ésimos CP´s. ε_{ij} é o resíduo para a linha i na coluna j. A matriz com elementos padronizados p_{ij} será denotada por **P.**

Biplot passo 3: É calculada a DVS da matriz **P** e os valores p_{ij} são atualizados utilizando apenas dois CP´s da DVS obtendo uma nova matriz chamada $\mathbf{P^{(2)}}$ com elementos $p_{ij}^{(2)}$.

Biplot passo 4: Todos os elementos $p_{ij}^{(2)}$ em $\mathbf{P^{(2)}}$ devem ser retornados à sua escala original, assim $\hat{x}_{ij}^{(2)} = m_j + s_j p_{ij}^{(2)}$. Desta maneira é obtida uma nova matriz $\mathbf{X^{(2)}}$ de dimensão (n x p). Os elementos ausentes x_{ij}^{aus} na matriz **X** original são imputados pelo correspondente valor $\hat{x}_{ij}^{(2)}$ de $\mathbf{X^{(2)}}$.

Biplot passo 5: O processo é iterado (voltando ao **Biplot passo 2**) até alcançar estabilidade nas imputações. Por exemplo, as iterações devem ser realizadas até que $d/\ \bar{y} < 0{,}01$. Definindo

$$d = \left[\left(\frac{1}{na} \right) \sum_{i=1}^{na} \left(x_i - x_i^A \right)^2 \right]^{\frac{1}{2}} \text{ e } \bar{y} = \left[\left(\frac{1}{N} \right) \sum_{i=1}^{n} \sum_{j=1}^{p} y_{ij}^2 \right]^{\frac{1}{2}}$$

Em que d representa a diferença entre os valores preditos para todos os valores ausentes na iteração atual e na iteração anterior. Nessa estatística "*na*" é o número total de valores ausentes na matriz **X**, x_i é o valor predito para o i-ésimo dado faltante na iteração atual e x_i^A na iteração anterior. Entretanto, uma grande média pode ser calculada como $\bar{y}$, em que y_{ij} é o valor observado (não ausente) na i-ésima linha e na j-ésima coluna, sendo N o número total de valores observados.

Imputação por meio de uma aproximação de posto inferior ponderada (ou EMSJ): Srebro e Jaakkola (2003) apresentam um algoritmo EM simples e eficiente para calcular a aproximação de posto inferior ponderada de uma matriz. No caso de dados faltantes as ponderações podem ser 0 se for ausente e 1 se for observado. Este método recentemente foi considerado por Canas (2012) para propor os modelos WAMMI (*Weighted Additive Main effects and Multiplicative Interaction*) na análise da interação genótipo-ambiente com heterogeneidade de variâncias. A metodologia é apresentada a seguir.

Considere a matriz **Y** de dimensão (n x p) com elementos y_{ij} ($i=1,...,n$; $j=1,...,p$) sendo alguns deles faltantes. Construa uma matriz **W** (n x p) com valores $w_{ij} = 0$ se y_{ij} for faltante e $w_{ij}=1$ em caso contrário. Construa a matriz **1** (n x p) com uns em todas as posições. De maneira iterativa calcule

$$\mathbf{X}^{(t+1)} = DVS\left(\mathbf{W} \langle \bullet \rangle \mathbf{Y} + (\mathbf{1} - \mathbf{W}) \langle \bullet \rangle \mathbf{X}^{(t)} \right)$$

em que t faz referência ao número da iteração no processo, $\langle \bullet \rangle$ representa o produto de Hadamard. Para t=0, **X** deve ser iniciada como $\mathbf{X^{(0)}} = \mathbf{0}$. O processo é iterado enquanto a soma de quadrados entre duas iterações consecutivas $\mathbf{X}^{(t+1)}$ e $\mathbf{X}^{(t)}$, seja maior do que um valor especificado (por exemplo, 10^{-9}). A saída deste procedimento são as matrizes $\mathbf{U}_k$, $\mathbf{D}_k$ e $\mathbf{V}_k$, tal que $\hat{\mathbf{Y}} = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^T$, em que k é o posto da aproximação. Da matriz $\hat{\mathbf{Y}}$ são obtidas as imputações para os valores ausentes da matriz **Y** original.

***Imputação GabrielEigen*:** Arciniegas-Alarcón et al. (2010) propuseram um método de imputação que mistura a

regressão com aproximação de posto inferior utilizando a DVS. O método é apresentado a seguir. Suponha uma matriz **X** de dimensão (n x p) com elementos x_{ij} ($i=1,...,n; j=1,...,p$), em que alguns deles são ausentes.

GabrielEigen passo1: Os valores ausentes são imputados, inicialmente, pela média da respectiva coluna obtendo uma matriz **X** completada.

GabrielEigen passo2: São padronizadas as colunas da matriz **X**, subtraindo de cada elemento m_j e dividindo o resultado por s_j (em que m_j e s_j representam a média e desvio padrão da j-ésima coluna).

GabrielEigen passo 3: Sobre a matriz padronizada é recalculada a imputação de cada valor x_{ij} ausente usando-se

$$x_{ij}^{(m)} = \mathbf{x}_{\mathbf{i}\bullet}^{T}\mathbf{V}\mathbf{D}^{-1}\mathbf{U}^{T}\mathbf{x}_{\bullet\mathbf{1}}$$

em que os vetores $\mathbf{x}_{\mathbf{i}\bullet}^{T}$, $\mathbf{x}_{\bullet\mathbf{1}}$ e as matrizes **V**, **D** e **U**, são obtidos da partição**,**

$$\mathbf{X} = \begin{bmatrix} x_{ij} & \mathbf{x}_{\bullet\mathbf{1}}^{T} \\ \mathbf{x}_{\bullet\mathbf{1}} & \mathbf{X}_{11} \end{bmatrix}$$

com ,

$$\mathbf{X}_{\mathbf{1}} = \sum_{k=1}^{m} \mathbf{u}_{(k)} d_k \mathbf{v}_{(k)}^{T} = \mathbf{U}\mathbf{D}\mathbf{V}^{T}$$

sendo $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_m]$, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_m]$, $\mathbf{D} = diag(d_1, ..., d_m)$ e $m \leq \min\{n - 1, p - 1\}$. Para cada observação faltante os componentes da partição considerada serão diferentes e dita partição é obtida por meio de operações elementares nas linhas e colunas da matriz **X**.

GabrielEigen passo 4: O processo de imputação depende da escolha do valor para m e o critério será o seguinte: m tal que,

$$\frac{\sum_{k=1}^{m} d_k^2}{\sum_{k=1}^{\min\{n-1,\,p-1\}} d_k^2} \approx 0{,}75$$

GabrielEigen passo 5: Finalmente, os valores imputados devem ser retornados à sua escala original assim,

$$x_{ij} = m_j + s_j \hat{x}_{ij}^{(m)}$$

substituindo-os na matriz **X**. Então, o processo é iterado (voltando ao **GabrielEigen passo2**) até alcançar estabilidade nas imputações. Todo o processo deve ser feito sobre a matriz **X**, tal que $n > p$, caso contrário, a matriz deve ser transposta. .

2.3 Características dos dados

Para avaliar os métodos de imputação foram considerados dois conjuntos de dados completos, obtidos da série de dados climatológicos do campus Luiz de Queiroz na cidade de Piracicaba – SP, coletada pela estação meteorológica (localizada juntamente com a pluviométrica) do Departamento de Engenharia de Biossistemas da Universidade de São Paulo. Os dados podem ser acessados em http://www.leb.esalq.usp.br/posto.html.

O primeiro conjunto de dados corresponde à matriz histórica de totais mensais de precipitação em milímetros desde o ano 1917 até o ano 2012. Assim, a matriz "univariada" tem dimensão (96 × 12), em que as linhas representam os anos e as colunas os meses.

O segundo conjunto de dados corresponde à matriz "multivariada" de dados diários (coletados entre as 0h e as 24h) a partir do mês de janeiro de 1997 até dezembro de 2012. Em 1997 começou a coleta das seguintes 4 variáveis: radiação global (cal/cm), umidade relativa média (%), amplitude térmica (graus Celsius) e chuva (mm). Para obter um conjunto completo foram considerados somente os dias com a informação coletada em todas as variáveis, obtendo uma matriz de dimensão (5818 × 4), em que as linhas representam os dias e as colunas representam as variáveis.

Desde o ponto de vista prático, considerar uma matriz "multivariada" é muito importante, pois, a maioria dos problemas na natureza não são univariados e qualquer variável das consideradas pode ser afetada pela ocorrência de dados faltantes. Desde o ponto de vista estatístico, a presença conjunta de várias variáveis é justificada pela matriz de correlação entre elas (Tabela 1), sendo todas as correlações significativas (valor-p<0,001). Assim, por exemplo, a umidade relativa (UR) média tem uma correlação moderada positiva com a chuva (r = 0,40) e moderada negativa com a radiação global (r = -0,49). Também pode se observar uma alta correlação negativa com a amplitude térmica (r = -0,73).

2.4 Estudo de simulação

Cada matriz de dados (univariada e multivariada) foi submetida a retiradas aleatórias de diferentes porcentagens. Foram consideradas as porcentagens de 10%, 20%, e 40%. O processo foi repetido em cada conjunto de dados 1000 vezes para cada porcentagem de retirada, obtendo 3000 matrizes diferentes com valores omissos. No total, foram gerados 6000 conjuntos de dados incompletos e em cada um, os dados foram imputados com os 5 algoritmos já descritos através de um programa computacional implementado no R (R Core Team 2013).

O processo de retirada aleatória para uma matriz **X** (n x p) foi o seguinte. Números aleatórios entre 0 e 1 foram gerados no R com a função *runif*. Para um valor fixo de r ($0 < r < 1$) , se o ($pi + j$) -ésimo número aleatório foi menor do que r, então o elemento na posição (i +1, j) da matriz foi deletado (i = 0,1,..., n-1; j = 1,...,p). A proporção esperada de dados ausentes na matriz será r (Krzanowski, 1988). Essa técnica foi utilizada com r = 0,1, 0,2 e 0,4.

Para os métodos de imputação pela Média, EM-DVS e EMSJ foi utilizada a implementação computacional fornecida por Wong (2013). Na aplicação dos algoritmos EM-DVS e EMSJ, em cada uma das matrizes incompletas simuladas, devia ser escolhido de antemão o número de componentes *k* a ser utilizado na DVS e para isso foi utilizada a função *cv.SVDImpute,* que faz validação cruzada sobre a informação disponível em uma matriz da seguinte maneira: dos dados observados na correspondente matriz é eliminado, aleatoriamente, 30% deles e posteriormente, é feita a imputação por EM-DVS. Usando os dados completados, é, finalmente calculada a raiz do erro quadrático médio (RMSE) sobre a porção de dados que foi artificialmente removida. O valor de *k* com menor RMSE é o escolhido para fazer a imputação. Neste estudo, sobre cada matriz incompleta simulada foi repetido o processo de validação cruzada 100 vezes e o *k* selecionado foi aquele que mais frequentemente minimizasse a RMSE.

2.5 Critérios de comparação

No conjunto de dados "univariado" foram adotados três critérios para comparar os resultados obtidos nas simulações: a estatística M^2 de Procrustes (Krzanowski, 2000), a raiz normalizada do erro quadrático médio - NRMSE (Ching et al., 2010) e o coeficiente de correlação não paramétrico de Spearman (Arciniegas-Alarcón e Dias, 2009).

Assim, cada matriz de dados completada (observados+imputados) $\mathbf{Y}_{\text{imp}}$ foi comparada com a matriz original $\mathbf{X}_{\text{orig}}$ através de

$$M^2 = traço(\mathbf{X}_{orig}\mathbf{X}_{orig}^T + \mathbf{Y}_{imp}\mathbf{Y}_{imp}^T - 2\mathbf{X}_{orig}\mathbf{Q}\mathbf{Y}_{imp}^T)$$

em que $\mathbf{Q} = \mathbf{V}\mathbf{U}^T$ é a matriz de rotação e pode ser calculada com elementos da decomposição por valores singulares da matriz $\mathbf{X}_{orig}^T\mathbf{Y}_{imp} = \mathbf{U}\Sigma\mathbf{V}^{\mathbf{T}}$. Com a estatística M^2 se obtém uma medida da diferença entre duas configurações de pontos e o método de imputação que minimize essa diferença indicará os melhores resultados.

O segundo critério utilizado foi a NRMSE definida como:

$$NRMSE = \frac{\sqrt{média\left(\mathbf{a}_{imp} - \mathbf{a}_{orig}\right)^2}}{dp\left(\mathbf{a}_{orig}\right)}$$

em que $\mathbf{a}_{\text{imp}}$ e $\mathbf{a}_{\text{orig}}$ são vetores contendo os respectivos valores preditos e os valores verdadeiros das observações ausentes simuladas. $dp(\mathbf{a}_{\text{orig}})$ é o desvio padrão dos valores contidos no vetor $\mathbf{a}_{\text{orig}}$. Quanto menor seja a NRMSE melhor será o método de imputação.

O terceiro critério de comparação no conjunto de dados "univariado" foi o coeficiente de correlação de Spearman. Foi calculado este coeficiente de correlação não paramétrico entre cada valor ausente e seu correspondente dado verdadeiro. Assim, quanto maior for a correlação entre os valores imputados e os valores originais, melhor será o método de imputação. Usou-se essa medida não paramétrica para evitar problemas de distribuição nos dados, uma vez que o coeficiente de correlação de Pearson é fortemente dependente da distribuição normal das variáveis.

Para o conjunto de dados "multivariado" foram utilizados dois critérios de comparação: a estatística M^2 apresentada anteriormente e por causa das diferentes escalas, uma versão modificada da NRMSE segundo Audigier et al. (2013) chamada NRMSE2. Para isto, deve ser construída para cada matriz incompleta simulada uma matriz indicadora de observações ausentes **W** com elementos w_{ij} (como foi descrito no método de imputação EMSJ na seção 2.2)

$$NRMSE2 = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{p} w_{ij}\left(\frac{x_j - \hat{x}_j}{s_{\mathbf{X}_j}}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{p} w_j}}$$

Na NRMSE2 x_{ij} representa o valor original, $\hat{x}_{ij}$ representa o valor imputado e $s_{\mathbf{X}_j}$ representa o desvio padrão verdadeiro da variável *j* na qual foi simulada a observação ausente. Quanto menor for a NRMSE2, melhor será o método de imputação.

3. RESULTADOS E DISCUSSÃO

Matriz univariada de precipitação

Na Tabela 2 apresentam-se a média e a mediana da NRMSE. Observa-se que em todas as porcentagens o método menos recomendável é o EMSJ, pois maximizou o critério (com valores médios de 2,0299, 1,4459 e 2,4480) e o mais

Tabela 1 - Matriz de correlação.

Correlações	UR média	Chuva	Rad.Global	Ampl. Term.
UR media	1,00	0,40	-0,49	-0,73
Chuva	0,40	1,00	-0,27	-0,37
Rad.Global	-0,49	-0,27	1,00	0,49
Ampl. Term.	-0,73	-0,37	0,49	1,00

recomendável seria a imputação por Média mensal porque o minimizou com um valor aproximadamente de 0,68 em todas as porcentagens. Estabelecendo uma ordem do mais eficiente ao menos eficiente segundo a NRMSE, os métodos seriam Média, GabrielEigen ou EM-SVD, Biplot e finalmente o EMSJ.

Estudando com mais detalhe o método EMSJ e procurando uma explicação do baixo desempenho, encontrou-se que existiram múltiplos problemas de convergência, especialmente quando a porcentagem de retirada foi 40% (em geral, nem 20000 iterações foram suficientes para convergir). Os efeitos de tais problemas de convergência podem ser vistos adicionalmente, quando foi estudada a correlação entre os dados imputados com os correspondentes dados verdadeiros na Figura 1.

Quando se imputaram 10% e 20% os métodos apresentaram distribuições de correlações aproximadamente simétricas, sendo o Biplot o método com correlação mais baixa (mediana próxima de 0,6) e a Média com correlação mais alta (mediana próxima de 0,8). No entanto, nessas duas porcentagens, independente do método, as correlações sempre foram moderadas ou altas. Na porcentagem de 40 %, a imputação pela Média manteve seu bom desempenho e o EMSJ confirmou que não deve ser utilizado quando a quantidade de dados faltantes for grande, pois a variação de resultados pode ser alta. De acordo com os resultados da NRMSE e da correlação de Spearman decidiu-se não considerar mais neste conjunto de dados o EMSJ, por se tornar um método não competitivo e pouco prático por causa de sua convergência lenta.

Na Figura 2 se apresenta a distribuição da estatística M^2 de Procrustes utilizando diferentes porcentagens de imputação. Lembre-se que quanto menor seja dita estatística, melhor será o método de imputação. Assim, com 10% de imputação poderiam ser utilizados os métodos GabrielEigen ou Média, mas, conforme aumenta a porcentagem a Média torna-se novamente o método mais recomendado. Segundo a M^2, a ordem do método mais eficiente ao menos eficiente seria: Média, EM-SVD, GabrielEigen e por último o Biplot.

Todos os resultados anteriores foram verificados por meio de testes não paramétricos de Friedman na comparação de todos os métodos, com posteriores testes não paramétricos de Wilcoxon para comparações dois a dois (Sprent e Smeeton, 2001). Em todos os casos houve significância; não se apresentam, por questão de espaço e simplicidade para o leitor.

Matriz multivariada de dados climáticos

Na Tabela 3 apresentam-se a média e a mediana da NRMSE2. Observa-se que, para todas as porcentagens de retirada aleatória o método GabrielEigen foi o melhor porque sempre minimizou o critério. A diferença dos resultados encontrados na matriz univariada de precipitação é que o GabrielEigen apresentou melhor desempenho do que a imputação utilizando a Média da variável. Por exemplo, com 10% de imputação a média da NRMSE2 para GabrielEigen foi 0,8057, enquanto para imputação por Média foi de 0,9987. Pode-se observar que em todas as porcentagens o algoritmo EMSJ teve o mais baixo desempenho.

Pode ser estabelecida uma ordem para os métodos de acordo com sua porcentagem de imputação. Assim, com 10% e 20% de imputação, a ordem do algoritmo mais eficiente ao menos eficiente será: GabrielEigen, Média, Biplot, EM-DVS e EMSJ. Com 40% de imputação a ordem já descrita é alterada porque o Biplot supera a Média. Vale a pena comentar que problemas de convergência do EMSJ similares aos encontrados na matriz univariada se mantiveram na matriz multivariada.

Na Figura 3 se apresenta a distribuição da estatística M^2 de Procrustes utilizando diferentes porcentagens de imputação na matriz multivariada. Com este critério, o método EMSJ teve uma dispersão muito grande, razão pela qual não permitiu fazer a comparação simultânea com os outros quatro métodos. Por essa característica, é o método menos recomendado e excluído da análise quando foi considerada a M^2. O método que minimizou M^2 em todas as porcentagens foi o GabrielEigen, quer dizer, com este método se obteve a maior similaridade entre as matrizes completadas por imputação e a matriz original a partir da qual foi feito o estudo de simulação.

Tabela 2 - Média e mediana da NRMSE na matriz de precipitação

	Porcentagem de retirada aleatória					
	10%		20%		40%	
Método	**Média**	**Mediana**	**Média**	**Mediana**	**Média**	**Mediana**
Biplot	0,8171	0,8107	0,8453	0,8391	0,8919	0,8868
EMSJ	2,0299	2,0130	1,4459	1,4407	2,4480	1,0894
Média	0,6856	0,6828	0,6862	0,6831	0,6882	0,6875
EM-SVD	0,7409	0,7364	0,7507	0,7492	0,8031	0,7916
GabrielEigen	0,7156	0,7127	0,7328	0,7309	0,8096	0,8082

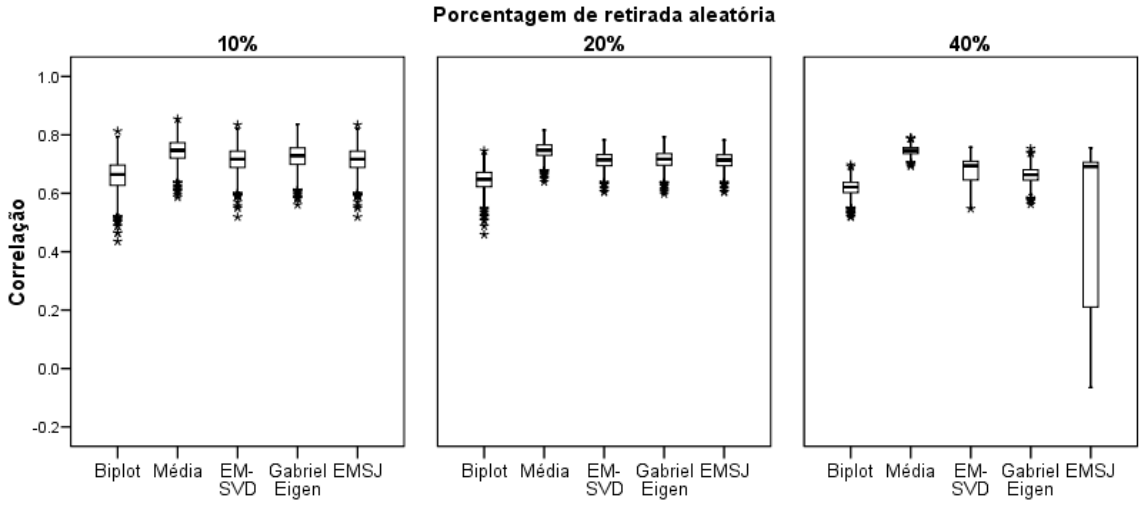

Figura 1 - Correlação com diferentes porcentagens de imputação na matriz de precipitação.

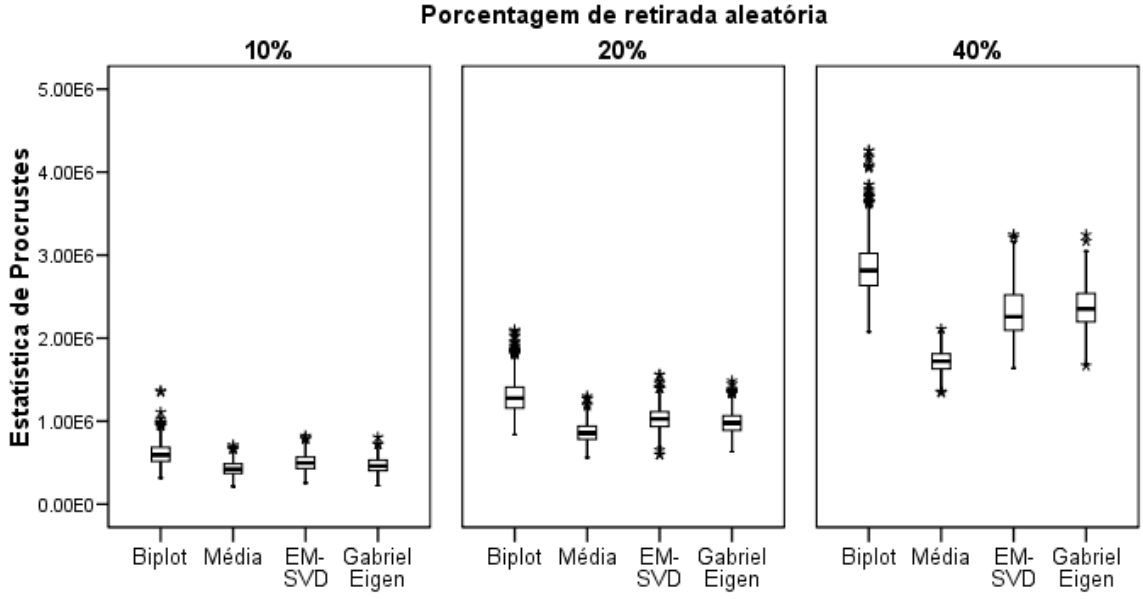

Figura 2 - M^2 com diferentes porcentagens de imputação na matriz de precipitação.

Por outro lado, o método com mais baixo desempenho foi o EM-DVS, porque em todas as porcentagens consideradas maximizou o critério. Os métodos, Média e Biplot, foram melhores que o EM-DVS, mas, foram superados pelo GabrielEigen.

4. CONCLUSÕES

Os resultados obtidos neste estudo a partir de duas matrizes reais de dados climáticos oferecem alguns guias para análises e pesquisas futuras sobre observações ausentes em dados climatológicos. Primeiro, se for considerada uma matriz univariada com dados faltantes, a imputação pela Média oferece melhores resultados do que outro algoritmo que envolva a DVS. Segundo, se a matriz de análise for multivariada e exista alguma razão para assumir correlação não nula entre as variáveis, o método GabrielEigen deveria ser considerado. Em geral, pode-se concluir também, que o algoritmo EMSJ deve ser utilizado unicamente com porcentagens baixas de observações ausentes, ou seja, menos de 20%.

5. AGRADECIMENTOS

O primeiro autor agradece ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, Brasil e à

Tabela 3 - Média e mediana da NRMSE2 na matriz multivariada.

	Porcentagem de retirada aleatória					
	10%		20%		40%	
Método	**Média**	**Mediana**	**Média**	**Mediana**	**Média**	**Mediana**
Média	0,9987	0,9964	0,9993	0,9972	1,0002	0,9997
Biplot	1,0483	1,0472	1,0149	1,0140	0,9729	0,9736
GabrielEigen	0,8057	0,8013	0,8232	0,8213	0,8703	0,8703
EM-SVD	1,2285	1,2280	1,2040	1,2040	1,1490	1,1488
EMSJ	2,1279	2,1229	2,5231	2,5220	7,2241	9,1977

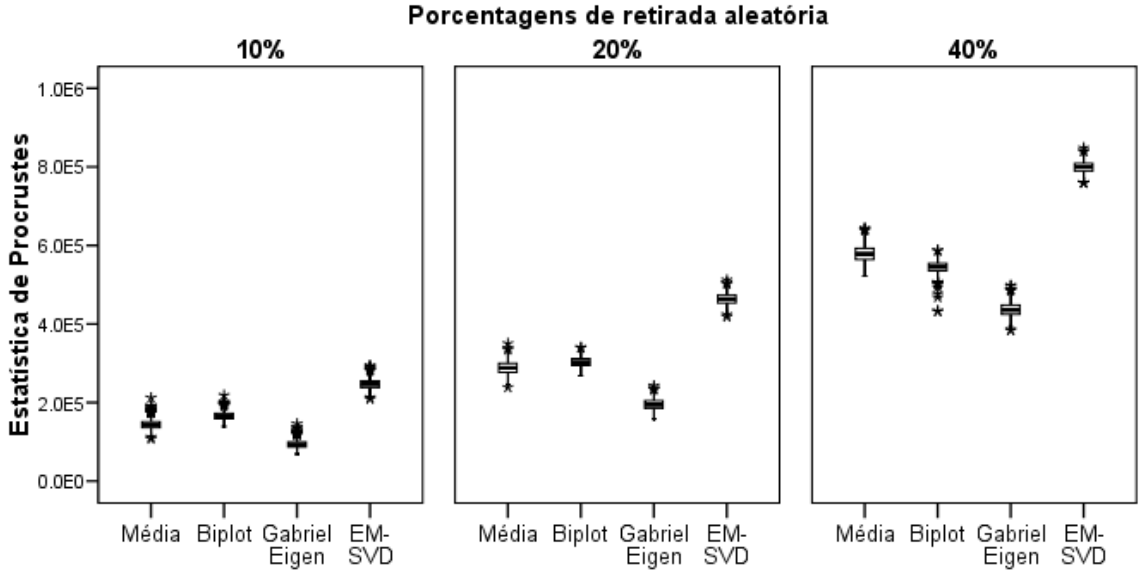

Figura 3 - M^2 com diferentes porcentagens de imputação na matriz multivariada**.**

Academia de Ciências para o Mundo em Desenvolvimento - TWAS, Itália (programa CNPq-TWAS) pelo apoio financeiro. O segundo autor agradece à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES, Brasil (programa PEC-PG) pelo apoio financeiro.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ALLISON, P.D. **Missing data**. Sage university papers series on quantitative applications in the social sciences. Sage, Thous and Oaks, pp 07–136, 2001.

ALY, A.; PATHAK, C.; TEEGAVARAPU, R. S. V.; AHLQUIST, J.; FUELBERG, H. Evaluation of improvised spatial interpolation methods for infilling missing precipitation records. **Proceedings World Environment Water Resources Congress**, doi:10.1061/41036(342)598, 2009.

ARCINIEGAS-ALARCÓN, S.; DIAS, C.T.S. Data imputation in trials with genotype by environment interaction: an application on cotton data. **Biometric Brazilian Journal**, v.27, p.125-138, 2009.

ARCINIEGAS-ALARCÓN, S.; GARCÍA-PEÑA, M.; DIAS, C.T.S.; KRZANOWSKI, W.J. An alternative methodology for imputing missing data in trials with genotype-by-environment interaction. **Biometrical Letters**, v.47, p.1-14, 2010.

AUDIGIER, V.; HUSSON, F.; JOSSE, J. A principal components methods to impute missing values for mixed data. **Arxiv**. Disponível em: http://arxiv.org/abs/1301.4797. Accesso em: 20 jul. 2013.

CANAS, P.J. **New strategies to detect and understand genotype-by-environment interactions and QTL-by-environment interactions**. 2012. 145f. Tese (Doutorado em Estatística e Gestão do Risco), Universidade Nova de Lisboa, Lisboa, 2012.

CANO, S.; ANDREU, J. Using multiple imputation to simulate time series: a proposal to solve the distance

effect. **WSEAS Transactions Computers**, v.9, n.7, p.768–777, 2010.

CHING, W.; LI, L.; TSING, N.; TAI, C.; NG, T. A weighted local least squares imputation method for missing value estimation in microarray gene expression data. **International Journal of Data Mining and Bioinformatics**, v.4, n.3, p.331-347, 2010.

COULIBALY, P.; EVORA, N.D. Comparison of neural network methods for in filling missing daily weather records. **Journal of Hydrology**, v.341, p.27–41, 2007.

DURRANT, G.B. Imputation methods for handling item-nonresponse in the social sciences: a methodological review, **NCRM (NCRM Working Paper Series, (NCRM-002)**, 2005.

EISCHEID, J.K.; PASTERIS, P.A.; DIAZ, H.F.; PLANTICO, M.S.; LOTT, N.J. Creating a serially complete, national daily time series of temperature and precipitation for the western United States. **Journal of Applied Meteorology**, v.39, n.9, p.1580–1591, 2000.

GABRIEL, K.R. The biplot graphic display of matrices with application to principal component analysis. **Biometika**, v.58, n.3, p.453-467, 1971.

GABRIEL, K.R. Le biplot - outil d´exploration de données multidimensionelles. **Journal de la Société Française de Statistique**, v.143, n.3-4, p.5–55, 2002.

JUNNINEN, H.; NISKA, H.; TUPPURAINEN, K.; RUUSKANEN, J.; KOLEHMAINEN, M. Methods for imputation of missing values in air quality data sets. **Atmospheric Environment**, v.38, n.18, p.2895–2907, 2004.

KALTEH, A.M.; BERNDTSSON, R. Interpolating monthly precipitation by self-organizing map (SOM) and multilayer perceptron (MLP). **Hydrologycal Sciences Journal**, v.52, n.2, p.305–317, 2007.

KALTEH, A.M.; HJORTH, P. Imputation of missing values in a precipitation-runoff process database. **Hydrology Research**, v.40, n.4, p.420–432, 2009.

KRZANOWSKI, W.J. Missing value imputation in multivariate data using the singular value decomposition of a matrix. **Biometrical Letters**, v.25, p.31-39, 1988.

KRZANOWSKI, W.J. **Principles of multivariate analysis: A user's perspective**. Oxford: University Press, 2000, 586 p.

LITTLE, R.J.A.; RUBIN, D.B. **Statistical analysis with missing data**, 2nd edn. Wiley, New York. 2002, 408 p.

LO PRESTI, R.; BARCA, E.; PASSARELLA, G.A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). **Environmental Monittoring and Assessment**, v.160, n.1–4, p.1–22, 2010.

LUCIO, P.S.; CONDE, F.C.; CAVALCANTI, I.F.A.; SERRANO, A.I.; RAMOS, A.M.; CARDOSO, A.O. Spatiotemporal monthly rainfall reconstruction via artificial neural network—case study: south of Brazil. **Advances in Geosciences**, v.10, p.67–76, 2007.

MAKHUVHA, T.; PEGRAM, G.; SPARKS, R.; ZUCCHINI, W. Patching rainfall data using regression methods. 2. Comparisons of accuracy, bias and efficiency. **Journal of Hydrology**, v.198, n.1-4, p.308-318, 1997.

McLACHLAN, G.; KRISHNAN, T. **The EM algorithm and extension**. Wiley, New York, 1997.

PAULHUS, J.L.H.; KOHLER, M.A. Interpolation of missing precipitation records. **Monthly Weather Review**, v.80, p.129-133, 1952.

PERRY, P. **Cross-validation for unsupervised learning**. 2009. 165f. Tese (Doutorado em Estatística), Stanford University, Stanford, 2009.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Disponível em: http://www.R-project.org/. Com acesso em: 20 jul. 2013.

RAMOS-CALZADO, P.; GÓMEZ-CAMACHO, J.; PÉREZ-BERNAL, F.; PITA-LÓPEZ, M.F. A novel approach to precipitation series completion in climatological datasets: application to Andalusia. **International Journal of Climatology**, v.28, n.11, p1525–1534, Rev A 45, 3403, 2008.

RUBIN, D.B. **Multiple imputation for nonresponse in surveys**. Wiley, New York, 1987, 258p.

SCHAFER, J.L. **Analysis of incomplete multivariate data**. Chapman and Hall/CRC, London, 1997, 444 p.

SCHAFER, J.L.; GRAHAM, J.W. Missing data: our view of the state of the art. **Psychological Methods**, v.7, n.2, p.147-177, 2002.

SCHNEIDER, T. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. **Journal of Climate**, v.14, p.853–871, 2001.

SMITH, K.W.; ARETXABALETA, A.L. Expectation–maximization analysis of spatial time series. **Nonlinear Processes in Geophysics**, v.14, n.1, p.73–77, 2007.

SPRENT, P.; SMEETON, N.C. **Applied nonparametric statistical methods**. 3th ed. Boca Raton: Chapman and Hall, 2001, 463 p.

SREBRO, N.; JAAKKOLA, T. Weighted low-rank approximations. **Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)**, Washington, D.C., 2003.

TANNER, M.A.; WONG, W.H. The calculation of posterior distributions by data augmentation. **Journal of the American Statistical Association**, v.82, n.398, p.528-540, 1987.

TEEGAVARAPU, R.S.V.; CHANDRAMOULI, V. Improved weighting methods, deterministic and stochastic data-driven

models for estimation of missing precipitation records. **Journal of Hydrology**, v.312, n.1–4, p.191–206, 2005.

WONG, J. **Imputation: imputation. R package version 2.0.1**. Disponível em: http://CRAN.R-project.org/package=imputation. Com acesso em: 20 jul. 2013.

XIA, Y.; FABIAN, P.; STOHL, A.; WINTERHALTER, M. Forest climatology: reconstruction of mean climatological data for Bavaria, Germany. **Agricultural and Forest Meteorology**, v.96, n.1-3, p.117-129, 1999a.

XIA, Y.; FABIAN, P.; STOHL, A.; WINTERHALTER, M. Forest climatology: estimation of missing values for Bavaria Germany. **Agricultural and Forest Meteorology**, v.96, n.1-3, p.131-144, 1999b.

YAN, W. Biplot analysis of incomplete two-way data. **Crop Science**, v.53, n.1, p.48-57, 2013.

YOUNG, K.C. A three-way model for interpolating for monthly precipitation values. **Monthly Weather Review**, v.120, p.2562-2569, 1992.

YOZGATLIGIL, C.; ASLAN, S.; IYIGUN, C.; BATMAZ, I. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. **Theoretical and Applied Climatology**, v.112, p.143-167, 2013.