PREENCHIMENTO DE FALHAS EM BANCOS DE DADOS METEOROLÓGICOS DIÁRIOS: UMA COMPARAÇÃO DE ABORDAGENS

RAMIRO RUIZ-CÁRDENAS[1] , ELIAS TEIXEIRA KRAINSKI[2]

[1] Engenheiro Agrônomo, D. Sc., bolsista de pós-doutorado CAPES/PRODOC, Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte – MG < ramiro@est.ufmg.br> [2] Estatístico, Professor Assistente, Departamento de Estatística, Universidade Federal do Paraná, Curitiba – PR < eliaskr@ufpr.br>

Apresentado no XVII Congresso Brasileiro de Agrometeorologia – 18 a 21 de Julho de 2011 – SESC Centro de Turismo de Guarapari, Guarapari - ES.

RESUMO: Neste trabalho são apresentadas e comparadas diferentes metodologias estatísticas para o preenchimento de falhas em bancos de dados meteorológicos diários, usando dados reais de temperatura máxima, temperatura mínima e precipitação. Abordagens baseadas no algoritmo EM e na metodologia de análise de componentes principais tiveram os menores erros de imputação para as três variáveis consideradas e tiveram um desempenho substancialmente melhor ao de um método de referência.

PALAVRAS-CHAVE: algoritmo EM, imputação múltipla, análise de componente principais

ABSTRACT: In this work we present and compare different statistical methodologies for gap filling in daily meteorological data sets using real data of maximum temperature, minimum temperature and precipitation. Two EM algorithm based approaches, as well as one based on principal component analysis had the lower imputation errors for all the three variables. They also had a much better performance than that found with a reference method.

KEYWORDS: EM algorithm, multiple imputation, principal component analysis.

INTRODUÇÃO : A maioria das séries temporais instrumentais são afetadas por uma certa proporção de dados faltantes. As razões dessa falta de dados são diversas, como por exemplo, interrupções ocasionais de estações automáticas, mau funcionamento dos instrumentos de medição, reorganização das redes de estações, etc. Uma forma de evitar lidar com esse tipo de dificuldade é excluir os períodos com dados faltantes das análises ou ignorar o problema, se a quantidade de falhas não for muito grande. Não entanto, esse tipo de abordagem desconsidera informação que pode ser relevante para a análise dos dados e pode induzir viés no resultado final. Várias técnicas têm sido desenvolvidas nas últimas décadas para tratar do problema de dados faltantes em séries temporais meteorológicas mensais ou anuais. Não entanto, o uso de estas técnicas em bancos de dados com uma resolução temporal maior (por exemplo, na escala diária), costuma ter erros de imputação grandes. Métodos para manipular dados faltantes na escala diária são, de fato, escassos na literatura, além de pouco conhecidos pela maioria de pesquisadores de áreas mais aplicadas. O objetivo do presente trabalho é apresentar e comparar diferentes metodologias estatísticas para o preenchimento de falhas em bancos de dados meteorológicos diários, usando dados reais de temperatura máxima, temperatura mínima e precipitação.

MATERIAL E MÉTODOS: A região de estudo escolhida foi o estado do Paraná (região sul do Brasil) e as variáveis meteorológicas de interesse foram temperatura mínima, temperatura máxima e precipitação diária. O acesso aos dados meteorológicos originais se deu através de solicitação formal nas respectivas agências (INMET, IAPAR, SIMEPAR e CPTEC). Um total de 113 estações, entre convencionais e automáticas, foram inicialmente identificadas no estado, todas com informação incompleta abrangendo o período de 01/01/1976 a 15/03/2011 (12858 registros por estação).

Foram avaliadas as seguintes abordagens estatísticas para o preenchimento de falhas diárias em bancos de dados:

- ***mtsdi*** (Multivariate time series data imputation): algoritmo para preenchimento de falhas em séries temporais normais multivariadas baseado no algoritmo EM, proposto por JUNGER ET AL. (2003). Além da estrutura de correlação entre estações levada em consideração na matriz de covariâncias dos dados, o método considera também a correlação temporal, através da modelagem independente das séries temporais em cada estação.

- ***ppca*** (Probabilistic Principal components analysis): Proposto inicialmente por TIPPING & BISHOP (1999), é uma reformulação em termos de um modelo probabilístico da análise de componentes principais convencional. A inferência é feita via máxima verossimilhança.

- ***mice*** (Multivariate Imputation by Chained Equations): é um algoritmo de imputação múltipla proposto por VAN BUUREN ET AL. (2006) em que o preenchimento dos dados faltantes é feito de forma iterativa considerando as densidades condicionais dos dados em cada estação.

- ***Amelia II*** (HONAKER, 2010): é outro algoritmo de imputação múltipla de dados multivariados baseado em técnicas de bootstrapping.

- ***regEM*** (Regularized EM algorithm): Também baseado no algoritmo EM, este método, proposto por SCHNEIDER (2001), realiza análises de regressão linear entre estações com dados faltantes e estações com dados disponíveis, em que os coeficientes de regressão são estimados via regressão penalizada (ridge regression).

- ***CIDW*** (Modified correlation coefficient with inverse distance weighting method): é uma modificação ao método da distância inversa, amplamente usado para preenchimento de falhas em bancos de dados meteorológicos, em que a ponderação é feita com base na correlação da estação "target" com os seus vizinhos mais próximos (TEEGAVARAPU ET AL., 2005). Por ser um dos métodos mais usados em dados meteorológicos este foi considerado como o método de referência para comparar com as cinco abordagens descritas anteriormente.

A avaliação do desempenho destas abordagens no preenchimento de dados faltantes diários foi realizada mediante um estudo de simulação que consistiu na criação de diferentes cenários, a partir da remoção de dados observados do banco de dados original, e na posterior medição do erro de imputação desses dados removidos obtido com cada um dos métodos utilizados. Para a criação desses cenários foram consideradas apenas estações com pelo menos 40% dos seus dados observados. Esse critério foi satisfeito para 41 das 113 estações iniciais. A partir desse conjunto de 41 estações, foram gerados cinco cenários de imputação para cada uma das três variáveis de interesse (temperatura máxima, temperatura mínima e precipitação), removendo em cada um deles 20 períodos de dados observados consecutivos, escolhidos aleatoriamente na parte observada das 41 estações. O comprimento de cada um desses 20 períodos de dados observados removidos foi de 30, 90, 180, 360 e 1100 dias consecutivos, para os cenários 1, 2, 3, 4 e 5, respectivamente.

A proporção total de dados faltantes nos cinco cenários, após a remoção dos dados observados, variou entre 17,5% e 22,1% para as variáveis temperatura máxima e mínima, e entre 16,5% e 20,5% para a variável precipitação. Em todos os casos o preenchimento de falhas foi feito para todos os dados faltantes no período de interesse (01/01/1976 a 15/03/2011) e não apenas para aqueles removidos artificialmente. As medidas de erro usadas para comparar os resultados obtidos foram a raiz do erro quadrático médio (REQM), o erro absoluto médio (EAM) e o índice de concordância de Willmott (d). As análises foram feitas usando a linguagem de programação *R* e seus pacotes "*mtsdi*", "*Amelia*", "*mice*" e "*pcaMethods*". No caso do algoritmo *regEM*, foi também usado o código Matlab fornecido pelos autores.

RESULTADOS E DISCUSSÃO: O desempenho das seis abordagens de imputação quando aplicadas à variável temperatura mínima diária, em três dos cinco cenários considerados, é ilustrada na Figura 1. Destaca-se o bom desempenho das cinco técnicas estatísticas, com erros absolutos médios em torno de 0,6 graus centigrados do valor real em todos os casos e índices de Willmott perto de um, indicando uma concordância muito boa entre dados observados (removidos) e preenchidos. Já o método da distância inversa modificado apresentou o pior resultado, com erros substancialmente maiores aos das outras metodologias. Uma tendência similar foi observada para a variável temperatura máxima diária.

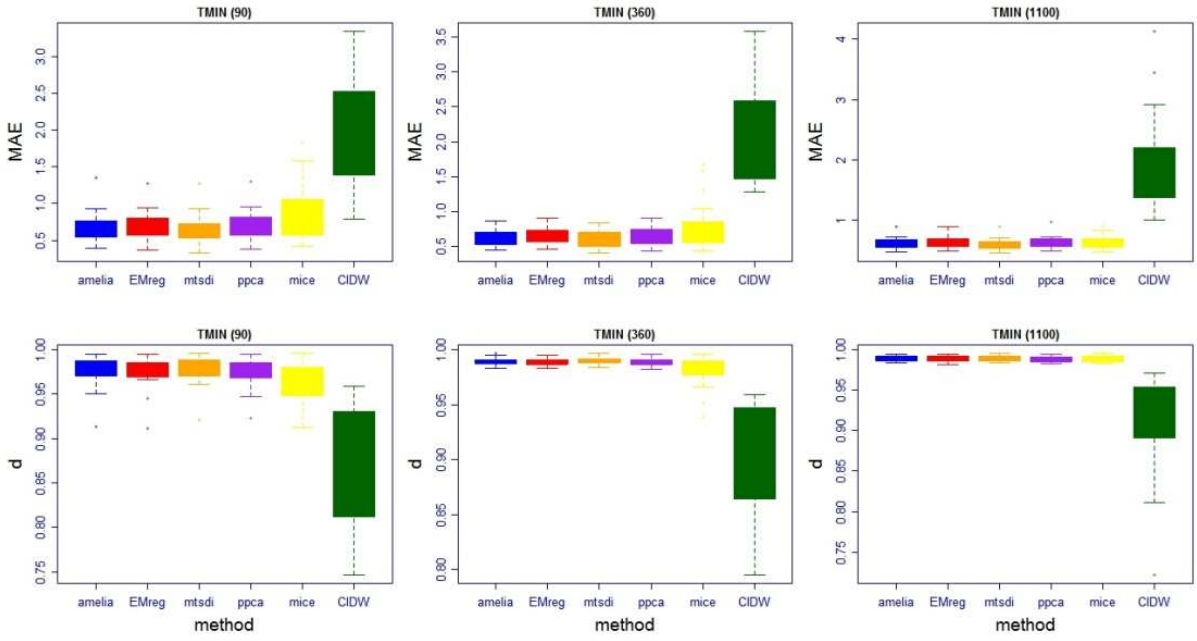

Figura 1. Erro absoluto médio (MAE) e índice de concordância de Willmott (d) para as seis abordagens de imputação, aplicadas à variável temperatura mínima diária nos cenários 2, 4 e 5 (períodos de 90, 360 e 1100 dias consecutivos removidos, respectivamente).

Os resultados para a variável precipitação são apresentados na Figura 2. De novo observam-se erros maiores para o método de referência (CIDW) em relação às cinco abordagens estatísticas, sendo essa diferença mais evidente em cenários em que os períodos de dados removidos eram maiores (360 e 1100 dias consecutivos). Dentre as abordagens estatísticas destacam-se as baseadas no algoritmo EM (*mtsdi* e *regEM*) e a baseada em análise de componentes principais (*ppca*), com erros absolutos médios em torno de 3mm, quando considerados todos os dados imputados. No entanto, a tendência em todos os casos foi de erros menores para os dias sem chuva e maiores para os dias com chuva.

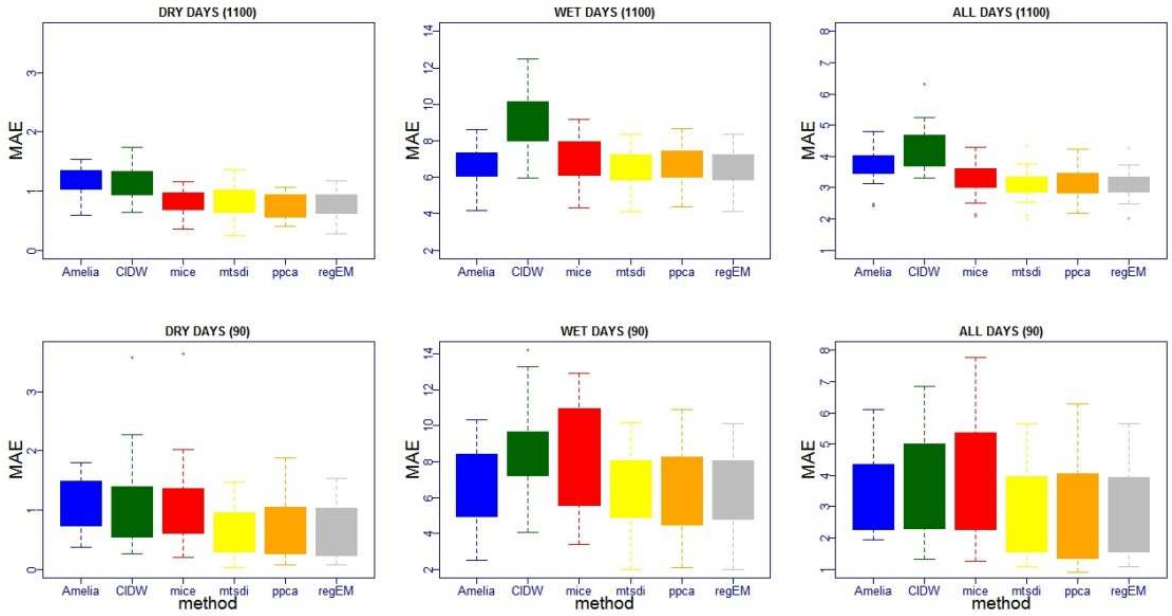

Figura 2. Erro absoluto médio (MAE) para as seis abordagens de imputação, aplicadas à variável precipitação diária, nos cenários 2 e 5 (períodos de 90 e 1100 dias consecutivos removidos), discriminando o erro nos dias sem chuva (secos), com chuva (úmidos) e no total.

O anterior é corroborado também nas Figuras 3 e 4, onde são apresentadas algumas séries de dados observados (removidos) e imputados usando os melhores métodos para as variáveis temperatura mínima e precipitação.

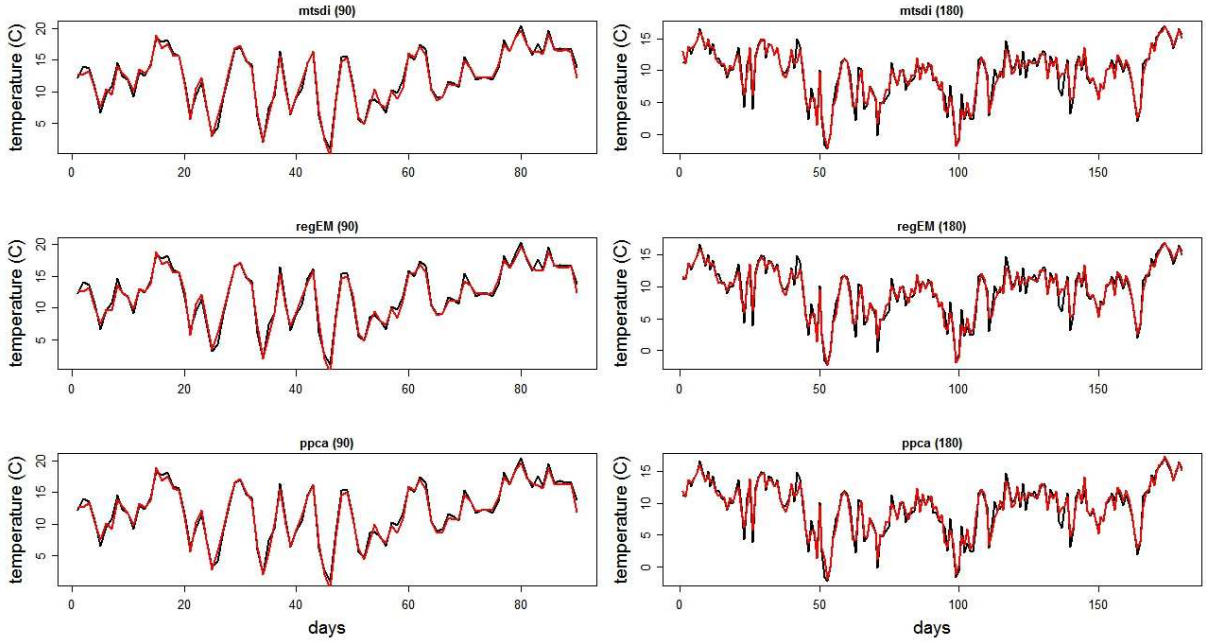

Figura 3. Séries de dados de temperatura mínima observados (em preto) e imputados (em vermelho) para as três abordagens de imputação com o melhor desempenho (mtsdi, regEM e ppca) nos cenários 2 e 3 (períodos de 90 e 180 dias consecutivos removidos).

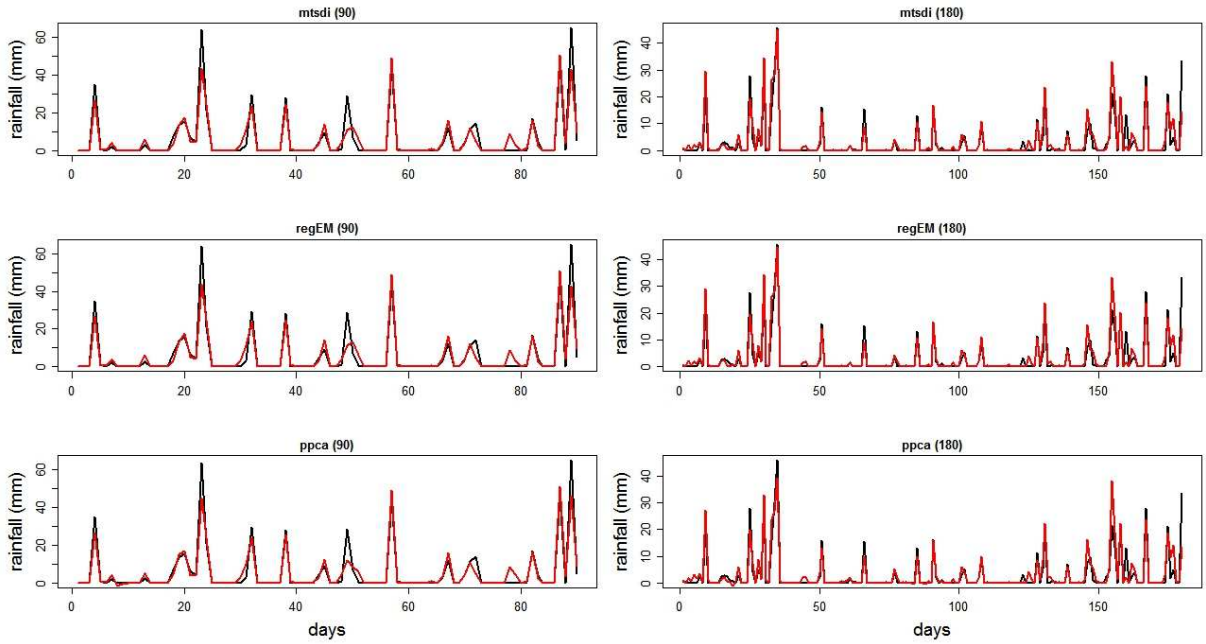

Figura 4. Séries de dados de precipitação observados (em preto) e imputados (em vermelho) para as três abordagens de imputação com o melhor desempenho (mtsdi, regEM e ppca) nos cenários 2 e 3 (períodos de 90 e 180 dias consecutivos removidos).

CONCLUSÕES: O uso adequado de técnicas estatísticas mostrou-se bastante eficiente no preenchimento de falhas em bancos de dados diários de temperatura e precipitação, mesmo no caso de longos períodos de dados faltantes.

REFERÊNCIAS BIBLIOGRÁFICAS

HONAKER, J., KING, G. **What to Do about Missing Values in Time-Series Cross-Section Data**. American Journal of Political Science, v. 54, p. 561-581, 2010.

JUNGER, W.L., PONCE DE LEON, A., SANTOS, N. **Missing data imputation in multivariate time series via EM algorithm**. Cadernos do IME, v. 15, p. 8-21, 2003.

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.

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, p. 191-206, 2005.

TIPPING, M.E., BISHOP, C.M. **Probabilistic principal component analysis**. Journal of the Royal Statistical Society, Series B, v. 61, p. 611-622, 1999.

VAN BUUREN, S., BRAND, J.P.L., GROOTHUIS-OUDSHOORN, C.G.M., RUBIN, D.B. **Fully conditional specification in multivariate imputation**. Journal of Statistical Computation and Simulation, v. 76, p. 1049-1064, 2006.