X-026 - IMPUTAÇÃO DE DADOS FALTANTES EM SÉRIES TEMPORAIS DE CONCENTRAÇÕES DE MATERIAL PARTICULADO INALÁVEL

Wanderson de Paula Pinto[(1)]
Graduado em Matemática pela Faculdade da Região Serrana - FARESE. Mestre em Engenharia Ambiental pela Universidade Federal do Espírito Santo (UFES). Doutorando em Engenharia Ambiental no Programa de Pós-Graduação em Engenharia Ambiental da Universidade Federal do Espírito Santo (PPGEA/UFES). Professor do Curso de Engenharia Ambiental da FARESE e do Centro Universitário do Espírito Santo - UNESC.
Valdério Anselmo Reisen
Doutor em Estatística pela University of Manchester Institute of Science And Technology. Professor do Departamento de Estatística e do Programa de Pós-graduação em Engenharia Ambiental, da UFES.
Gemael Barbosa Lima
Engenheiro Ambiental pela FAESA - Faculdades Integradas Espírito-Santenses. Mestre em Engenharia Ambiental pela Universidade Federal do Espírito Santo (UFES). Professor do Curso de Engenharia Ambiental da FARESE e do Centro Universitário do Espírito Santo - UNESC.

Endereço[(1)]**:** Rua Jequitibá, 121 - Centro - Santa Maria de Jetibá – ES – CEP: 29645-000 - Brasil - Tel: (27) 3263-2010 - e-mail: **wandersondpp@gmail.com**.

RESUMO

Este trabalho teve por objetivo avaliar e comparar a performance de métodos univariados, média e mediana, algoritmo EM e algoritmo mice (Multivariate Imputation by Chained Equations) para imputação de dados faltantes em uma série temporal de concentrações médias diárias de Material Particulado Inalável (PM_{10}) monitorada no Bairro de Jardim Camburi, Vitória, E.S., Brasil, compreendida entre 01 de janeiro de 2003 e 31 de dezembro de 2004. Em particular, nota-se que ambos procedimentos fornecem bons resultados para porcentagem de 5% de dados faltantes. Para porcentagens maiores os melhores resultados foram obtidos através do algoritmo EM (expectation-maximisation).

PALAVRAS-CHAVE: PM_{10}, Dados faltantes, Séries temporais, Imputação de dados.

INTRODUÇÃO

Um problema frequente em séries temporais provenientes de monitoramentos da qualidade do ar é a presença de dados faltantes (*missing data*). Estes dados ocorrem, geralmente, pois os equipamentos de medição das concentrações de contaminantes na atmosfera podem apresentar defeitos que impossibilitem seu funcionamento por algum tempo, ocasionando perda de dados. A análise de dados, incluindo apenas as observações disponíveis sem um tratamento estatístico para os dados faltantes, pode produzir estimativa falsa da medida de efeito e subestimar sua precisão (JUNGER, 2008).

Entre as metodologias para o tratamento de dados faltantes em séries temporais estão os métodos de imputação usados em trabalhos que envolvem séries de concentração de material particulado inalável (PM_{10}), dentre estes, Junninen et al. (2004), que avaliaram dois contextos (univariada e multivariada) e Plaia e Bandi (2006) propuseram uma metodologia de imputação de dados faltantes denominada de Site-Depen-Dente (SDEM). Os procedimentos de imputação de dados faltantes consistem em preencher os valores em falta e analisar o conjunto de dados resultantes usando métodos convencionais. Alguns procedimentos de imputação são simples e implementados na maioria dos aplicativos estatísticos. A principal desvantagem dos métodos de imputação é que, em sua maioria, a imprecisão devida à imputação não é contemplada na análise, portanto, a variância dos estimadores é subestimada (JUNGER, 2008).

Desta forma, o objetivo deste trabalho foi avaliar e comparar a performace de métodos para imputação de dados faltantes em séries temporais de concentrações médias diárias de PM_{10} observadas na Região da Grande Vitória, ES, Brasil.

MATERIAIS E MÉTODOS

Esse trabalho foi realizado na Região da Grande Vitória (RGV), constituída pelos municípios de Vitória, Vila Velha, Cariacica, Serra e Viana, Espírito Santo. A região sofre com diversos tipos de problemas ambientais, dentre os quais está à deterioração da qualidade do ar, devido às emissões atmosféricas por indústrias e pela frota veicular.

A RGV possui uma Rede Automática de Monitoramento da Qualidade do Ar (RAMQAR) inaugurada em julho de 2000, de propriedade do Instituto Estadual de Meio Ambiente e Recursos Hídricos (IEMA). A referida rede é distribuída em oito estações localizadas nos municípios que compõem a RGV, da seguinte forma: o município Serra com duas estações localizadas nas regiões de Laranjeiras e Carapina; o município Vitória com três estações localizadas nas regiões de Jardim Camburi, Enseada do Suá e Centro de Vitória. O município de Vila Velha apresenta duas estações localizadas nas regiões do Ibes e Centro de Vila Velha e o município de Cariacica com uma estação em Cariacica. A localização espacial das estações de monitoramento da RAMQAR está ilustrada na Figura 1.

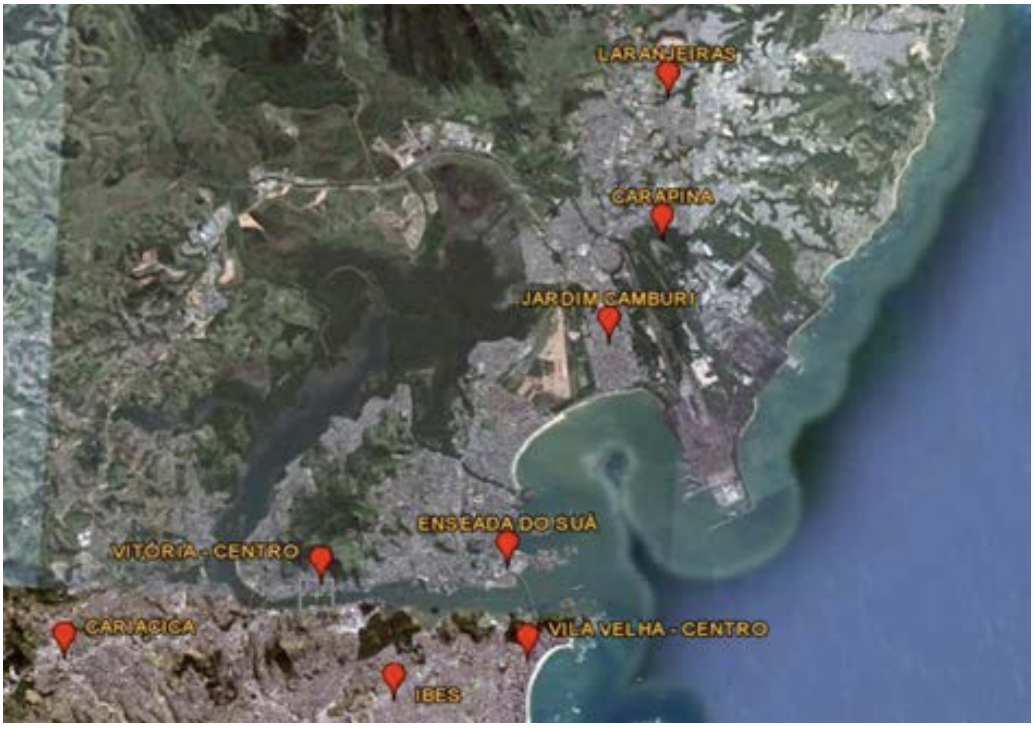

Figura 1: Localização espacial das estações de monitoramento da qualidade do ar da RGV. Fonte: Google Earth (2015).

A RAMQAR monitora os seguintes poluentes: Partículas Totais em Suspensão (PTS); Partículas Inaláveis (PM10); Ozônio (O3); Óxido de Nitrogênio (NOx); Monóxido de Carbono (CO) e Hidrocarbonetos (HC). E ainda, realiza-se o monitoramento dos seguintes parâmetros meteorológicos: Direção dos ventos (DV); Velocidade dos ventos (VV); Precipitação pluviométrica (PP); Umidade relativa do ar (UR); Temperatura (T); Pressão atmosférica (P) e Radiação solar (I).

Foram consideradas 731 observações compreendidas entre 01 de janeiro de 2003 e 31 de dezembro de 2004, de médias diárias das concentrações de material particulado (PM_{10}), medidas em $\mu g.m^{-3}$ na estação de Jardim Camburi, Vitória-ES. Todas as análises estatísticas e simulações foram realizadas no *software* R 2.15.1 (*software* livre) e seus pacotes "*mtsdi*", e "*mice*".

Imputação por constantes

Os métodos de imputação por constantes são os mais comuns dentre os métodos de imputação única (IU), o princípio deste método é imputar um valor para cada dado faltante da base de dados e, então, analisá-la como

se não houvesse dados faltantes (MCKNIGHT et al., 2007). Os métodos de imputação por constantes são os mais comuns dentre os métodos de IU. De forma geral, esses métodos substituem todos os valores faltantes de uma variável por um único valor, uma constante. A seguir, serão apresentados os dois métodos de IU utilizados neste trabalho, média e mediana.

Devido à sua facilidade de implementação a imputação da média se torna um método muito comum e bastante utilizado (MYRTVEIT et al., 2001). Nesta técnica, a média dos valores de um atributo que contém dados faltantes é usada para preencher os seus espaços com dados faltantes (FARHANGFAR et al., 2004). Mas, na maioria dos casos este método não é eficaz para o tratamento, pois os valores extremos ficam sub-representados, o que implica na perda de variabilidade, ou seja, a variância das variáveis com dados faltantes é subestimada (MCKNIGHT et al., 2007).

Outro método de IU é a mediana, a imputação da mediana é bem parecida com a imputação da média e apresenta as mesmas vantagens e desvantagens. Porém, segundo Veroneze (2011) este método é uma alternativa melhor para variáveis que não são normalmente distribuídas, pois a mediana representa melhor a tendência central de uma distribuição que possui grandes desvios da distribuição normal.

Imputação via algoritmo EM

O nome Expectatin Maximization (EM) vem de seus dois principais passos: Esperança e Maximização. A formalização deste método foi proposta por Dempster et al., (1977). Segundo Little e Rubin (1989) o algoritmo EM é um método geral para obter estimativas de máxima verossimilhança em bases de dados incompletos. Como estas estimativas podem ser difíceis de obter para bases de dados complexas, é necessário um procedimento para reduzir esta dificuldade, que é o objetivo do algoritmo EM (MACKNIGHT et al., 2007).

Neste trabalho utilizou o algoritmo EM proposto por Junger (2008), na qual vez uso da plataforma *msdti* (multivariante time-series data imputation) do R, que foi implementada pelo autor.

Imputação via algoritmo mice

O 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.

Para avaliar os métodos em termos de qualidade da imputação em 100 replicações de um padrão escolhido ao acaso foram utilizados os indicadores de performance propostos por Junnien et al., (2004). A saber: Coeficiente de correlação de Person (r), raiz do erro quadrático médio (REQM), erro absoluto médio (EAM), o viés (BIAS) e o índice de concordância de Willmot (d_2).

RESULTADOS E DISCUSSÃO

Para melhor compreensão da variável estudada a Tabela 1 apresenta algumas medidas descritivas da série. Verifica-se que a média diária de PM_{10} é de, aproximadamente, 27 $\mu g.m^{-3}$, sendo que a maior concentração foi de 67,8 $\mu g.m^{-3}$ e a menor foi 7 $\mu g.m^{-3}$.

Tabela 1: Estatísticas descritivas da série de concentração das médias diárias PM_{10}.

	Média	Desvio	Min	Max	Mediana
Série PM_{10}	27.593	7.816	7.083	67.830	27.333

Com a série temporal fez a análise dos métodos de imputação com as seguintes porcentagens de dados faltantes 5%, 20%, 30% e 40%. O trabalho de Greenlad e Rothman (1998) indica que, para uma pequena porcentagem de dados faltantes e um grande número de observações, a análise de dados completos produz bons resultados. Desta forma, a porcentagem de 5% de dados faltantes foi incluída como referência. A porcentagem de 40%, por outro lado, serve para avaliar os métodos de imputação sob condições extremas de informação perdida. A Tabela 2 apresenta os resultados para os indicações de performance para as porcentagens de dados faltantes estudadas.

Tabela 2: Performance dos métodos de imputação estudados para as quatro porcentagens de dados faltantes.

Porcentagem	Indicadores	Métodos de imputação			
		Média	Mediana	EM	Mice
5%	REQM	0.6412	0.6390	**0.6382**	0.6776
	EAM	0.4993	0.4975	**0.4905**	0.5246
	BIAS	-0.0182	-0.0164	**-0.0022**	0.0086
	r	0.9989	**0.9990**	0.9984	0.9983
	d_2	0.9812	0.9813	**0.9828**	0.9582
10%	REQM	1.2535	**1.2495**	1.2583	1.3077
	EAM	0.9966	0.9931	**0.9814**	1.0325
	BIAS	-0.0690	-0.0573	-0.0389	**-0.0353**
	r	0.9961	**0.9962**	0.9935	0.9944
	d_2	0.9602	0.9604	0.9631	**0.9162**
20%	REQM	2.3241	**2.3171**	2.3216	2.3617
	EAM	1.8718	1.8649	**1.8441**	1.8804
	BIAS	-0.1300	-0.0952	-0.0809	**-0.0469**
	r	0.9874	**0.9875**	0.9752	0.9834
	d_2	0.9139	0.9141	**0.9220**	0.8354
40%	REQM	4.0647	**4.0532**	4.0776	4.1359
	EAM	3.3128	3.3009	**3.2677**	3.3750
	BIAS	-0.2297	**-0.1421**	-0.2583	-0.1674
	r	**0.9566**	0.9569	0.8934	0.9411
	d_2	0.8028	0.8021	**0.8298**	0.6941

De acordo com os resultados (Tabela 2), observa-se um gradiente de crescimento nos indicadores em função do aumento da parcentagem de dados faltantes. Através da Tabela 2 nota-se que ambos os métodos forneceram bons resultados para a porcentagem de 5% de dados faltantes. Os métodos de imputação pela média e mediana consiste em substituir o valor faltante por uma constante, logo estes métodos apresentam uma Ineficiência quando aplicados aos dados com uma porcentagem maior que 5% de dados faltantes. O método de imputação pelo algoritmo EM, apresentou bom desempenho com valores baixos de RMSE e MAE e valores altos para o coeficiente de correlação e para o índice de concordância para porcentagens superiores a 5%. O mesmo observou-se para o algoritmo mice, porém, de forma geral, o desempenho deste método é inferior ao do algoritmo EM. Desta forma, pode-se concluir que o método baseado no algoritmo EM é o mais adequado, entre os testados neste trabalho, para imputação de dados faltantes em séries temporais de concentrações de PM_{10}.

O anterior é corroborado também nas Figuras 1 e 2, onde é apresentada a série temporal de dados observados de PM_{10} e imputados usando o algoritmo EM para as porcentagens de 5% (Figura 1) e 40% (Figura 2) de dados faltantes simulados.

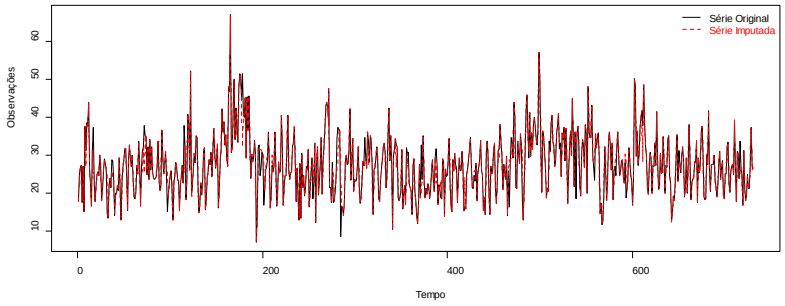

Figura 1: Série temporal de PM_{10} (em preto) e imputada (em vermelho) para a porcentagem de 5% de dados faltantes.

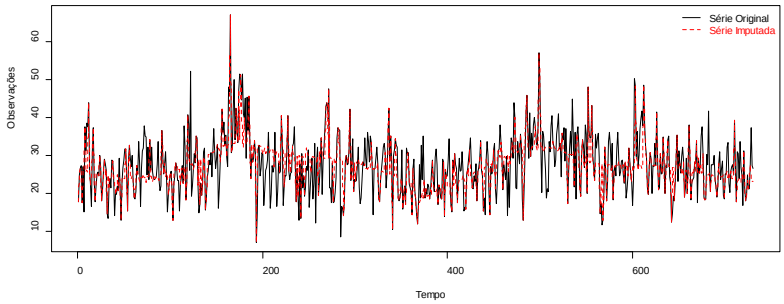

Figura 2: Série temporal de PM_{10} (em preto) e imputada (em vermelho) para a porcentagem de 40% de dados faltantes.

CONCLUSÃO

Devido à facilidade de implementação, a imputação pela média se torna um método muito comum e bastante utilizado. Mas, os resultados evidenciam que este método não é eficaz para o tratamento de dados faltantes para porcentagens superiores a 5%, pois os valores extremos ficam sub-representados, o que implica na perda de variabilidade, ou seja, a variância das variáveis com dados faltantes é subestimada.

Outro método testado foi à mediana, a imputação pela mediana é bem parecida com a imputação pela média e apresenta as mesmas vantagens e desvantagens. Porém, este método é uma alternativa melhor para variáveis que não são normalmente distribuídas, pois a mediana representa melhor a tendência central de uma distribuição que possui grandes desvios da distribuição normal. Os outros dois métodos de imputação testados no trabalho foram os algoritmos mice e EM, ambos implementados em pacotes do software R. De acordo com os resultados apresentados, o algoritmo EM foi o que apresentou os melhores resultados para imputação de dados faltantes em séries temporais de concentrações de PM_{10}.

REFERÊNCIAS BIBLIOGRÁFICAS

1. DEMPSTER A, Laird N, Rubin D. Maximum Likelihood from Incomplete Data via the Algorithm EM. Journal of the Royal Statistical Society, B., 39:1-38, 1977.
2. FARHANGFAR, A., KURGAN, L., PEDRYCZ, W. Experimental analysis of methods for imputation of missing values in databases. Proceedings of SPIE, Orlando, vol. 5421, pp.172-182, 2004.
3. GREENLAND S., FINKLE W. D. Modern epidemiology. 2 ed. Philadelphia, Lippincott-Raven, 1998.
4. JUNGER, W. L. Análise, imputação de dados e interfaces computacionais em estudos de séries temporais epidemiológicas. 178 f. Tese (Doutorado) – Programa de Pós-graduação em Saúde Coletiva, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2008.
5. JUNNINEN, H., NISKAA, H., TUPPURAINENC, K., RUUSKANENA, J., KOLEH-MAINEN, M. Methods for imputation of missing values in air quality data sets. Atmospheric Environment 38, 2895-2907, 2004.
6. LITTLE, R.J.A., RUBIN D.B. Statiscal analysis with missing data. New York, Wiley, 1989.
7. MCKNIGHT, P. C., MCKNIGHT, K. M., FIGUEREDO, A. J., Missing data: a gentle introduction. New York. The Guilford Press, 2007.
8. MYRTVEIT, I., STENSRUD, E., OLSSOM, U. H. Analyzing data sets with missing data: an empirical evaluation of imputation methods and likelihood-based methods. IEEE Transactions On Software Engineering, vol. 27, no. 11, pp. 999-1013, 2001.
9. PLAIA, A., BONDÌ AL. Single imputation method of missing values in environmental pollution data sets. Atmospheric Environment. 40: 7316-7330, 2006.
10. 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.

11. VERONEZE, R. Tratamento de dados faltantes empregando biclusterização com imputação múltipla. Dissertação de Mestrado, Campinas: Programa de Pós-Graduação em Engenharia Elétrica e de Computação: Universidade Estatual de Campinas, 2007.