Revista Brasileira de Climatologia
ISSN: 1980-055x (Impressa) 2237-8642 (Eletrônica)

ANÁLISE DA APLICABILIDADE DE MÉTODOS ESTATÍSTICOS PARA PREENCHIMENTO DE FALHAS EM DADOS METEOROLÓGICOS

VENTURA, Thiago Meirelles - thiago@ic.ufmt.br
Universidade Federal de Mato Grosso

SANTANA, Luy Lucas Ribeiro - luylucas10@gmail.com
Universidade Federal de Mato Grosso

MARTINS, Claudia Aparecida - claudia@ic.ufmt.br
Universidade Federal de Mato Grosso

FIGUEIREDO, Josiel Maimone de - josiel@ic.ufmt.br
Universidade Federal de Mato Grosso

RESUMO: Os dados meteorológicos são de grande importância para os estudos científicos, pois auxiliam na tomada de decisões em diferentes áreas do conhecimento. As estações automáticas realizam o trabalho de captar esses dados, porém, problemas podem ocorrer nos instrumentos causando falhas nas séries de dados e inutilizando um período ou até mesmo toda a série. Visto que a análise desses dados é prejudicada com esse problema, as falhas devem ser tratadas para garantir uma maior qualidade na obtenção das informações. Este trabalho visa comparar métodos estatísticos de preenchimentos de falhas e verificar qual método possui melhores resultados no preenchimento de falhas em séries de dados meteorológicas. Falhas foram simuladas em séries de dados reais e o desempenho de quatro métodos foram comparados: média simples, média móvel, regressão linear simples e regressão linear múltipla. Para verificar os resultados obtidos, foram usados o erro médio absoluto e o coeficiente de correlação. Os resultados mostraram ótimo desempenho do método de regressão linear múltipla para as variáveis de temperatura, umidade e ponto de orvalho, enquanto que a média simples teve o melhor resultado para a variável de pressão atmosférica. Nenhum dos quatro métodos obteve bons resultados para a variável de radiação solar.
PALAVRAS-CHAVE: tratamento, processamento, séries temporais, regressão linear.

ANALYSIS METHODS OF APPLICATION FOR STATISTICAL DATA IN METEOROLOGY

ABSTRACT: The meteorological data are important for scientific studies, to assist in decision-making in different areas of knowledge. The automatic stations make it possible to obtain meteorological data. However, problems may occur in equipment causing failures in data series and making it useless a period or even the entire series. Analysis of these data is impaired by this problem, so gaps should be treated to ensure a higher quality in obtaining the information. This work has the objective of comparing statistical gap filling methods and check which method has better results in gap filling in meteorological datasets. Gaps were simulated in real datasets and the performance of four methods was compared: simple average, moving average, simple linear regression and multiple linear regression. To verify the results obtained were used the mean absolute error and the correlation coefficient. The results showed good performance of the multiple linear regression method for temperature, humidity and dew point, while the simple average had the best result for the atmospheric pressure variable. None of the four methods achieved good results for the solar radiation variable.
KEY-WORDS: pre-processing, data processing, time series, linear regression.

ISSN: 1980-055x (Impressa) 2237-8642 (Eletrônica)

1. INTRODUÇÃO

Os dados meteorológicos são de grande importância, pois auxiliam a tomada de decisões em diferentes áreas do conhecimento, seja na agricultura, engenharia, na própria meteorologia, transportes, na construção civil, entre outras (BAMBINI; FURTADO, 2011). A utilização desses dados pode ser observada em trabalhos como o de Oliveira (2009) que realizou uma análise de precipitação para a cultura de arroz, em Fuentes et al. (2013) que verificou se a velocidade do vento pode ser a causadora de grandes ondas e naufrágios (FUENTES et al, 2013), em Gianotti et al. (2013) verificando o nível de precipitação e sua influência na distribuição vegetal, em Barboza (2006) na tentativa de evitar uma tragédia envolvendo tempestades no litoral e em Magina & Souza (2007) utilizando esse tipo de dado em projeto de linha de transmissão de energia.

Para trabalhar com dados meteorológicos, primeiramente, é necessário obtê-los. Bambini & Furtado (2011) descreve o histórico do avanço dos equipamentos meteorológicos, comentando desde as especulações de Aristóteles até os tempos atuais com as automatizações e instrumentos meteorológicos. Porém, medições automáticas podem falhar, seja devido à falha do próprio instrumento ou na transmissão dos dados (TARDIVO; BERTI, 2014). As falhas também podem ocorrer por avaria, desligamento de equipamentos, manutenção, calibração, limitações físicas ou fenômenos climáticos (HUI, 2004). Qualquer um dos problemas citados pode ocorrer e, assim, diminuir a consistência e a qualidade dos dados.

Visto que esses dados são utilizados para tomar decisões sobre as análises realizadas, as falhas devem ser tratadas para garantir uma maior qualidade na obtenção das informações. Alguns métodos de preenchimento de falhas foram aplicados para correção de dados temporais. Em Hassan & Croke (2013), é utilizado o método estatístico Poisson-Gamma Distribution para preencher falhas em séries de precipitação, Tardivo & Berti (2014) descrevem um método baseado em regressão para preencher falhas em séries temporais de temperatura diária, Dengel et al. (2013) descrevem o uso de técnicas de inteligência artificial para preencher falhas em séries de CH_4, Wanderley et al. (2012) fazem uso da geoestatística para realizar o preenchimento de falhas em séries temporais pluviométricos.

Percebe-se nos trabalhos citados que diferentes métodos são aplicados para variáveis específicas. Dessa forma, há uma dificuldade para avaliar qual método se comporta melhor em determinadas variáveis meteorológicas. Assim, o objetivo deste trabalho foi de comparar e analisar quatro métodos diferentes de preenchimento de falhas em séries temporais de dados meteorológicos, avaliando o desempenho de cada um para diferentes variáveis climáticas, a fim de avaliar a sua aplicabilidade.

Este trabalho está dividido da seguinte forma: na Seção 2 são descritos os dados, suas unidades e também os métodos utilizados neste estudo. Na Seção 3 são descritos os resultados, criando uma discussão sobre o desempenho de cada método por meio dos erros encontrados e tempo de processamento. Por fim, o estudo é concluído na Seção 4.

ISSN: 1980-055x (Impressa) 2237-8642 (Eletrônica)

2. MATERIAIS E MÉTODOS

Os quatro métodos de preenchimento de falhas selecionados foram testados com dados meteorológicos reais. Simulações de falhas foram geradas para poder medir o desempenho dos métodos. Nesta seção serão descritos os métodos testados e os preparativos para realização dos experimentos.

2.1. Métodos avaliados

Os métodos utilizados para este estudo foram: média aritmética simples, média móvel, regressão linear simples e regressão linear múltipla. A média aritmética simples para preenchimento de falhas consiste em somar o valor anterior à falha e posterior à falha e dividi-los por 2. A Eq. (1) define a média aritmética simples para preenchimento de falhas. Caso algum dos valores, que serão utilizados para encontrar a média, sejam nulos, eles são ignorados e o anterior ou posterior são utilizados.

$$\overline{X_i} = \frac{x_{i-1} + x_{i+1}}{2} \quad (1)$$

No caso da média móvel, Gençay (1996) comenta que a média é calculada adicionando os n valores mais recentes e dividindo-os por n. A Eq (2) apresenta esse cálculo, que no caso de preenchimento de falhas pode ser utilizado os dados anteriores e posteriores à falha:

$$\overline{X_i} = \frac{\sum_{k=\, i-\frac{n}{2}}^{i-1} x_k + \sum_{j=i+1}^{i+\frac{n}{2}} x_j}{n} \quad (2)$$

Para este estudo optou-se por utilizar n=10. Em uma série de dados horária, um valor maior do que esse poderia prejudicar o desempenho do método, uma vez que estaria sendo utilizado dados muito distante da falha para calcular o dado a ser preenchido.

A análise de regressão é realizada de forma a determinar as correlações entre duas ou mais variáveis que mantenham relações de causa-efeito, realizando previsões utilizando a relação (UYANıK; GÜLER, 2013). No caso da regressão linear simples, os valores para se deduzir se baseiam em apenas uma variável. A Eq. (3) define a regressão linear simples:

$$y = \alpha + \beta x \quad (3)$$

Onde y é a variável dependente, α é o coeficiente linear, β o coeficiente angular e x a variável independente. Para calcular os coeficientes foram utilizadas as Eq. (4) e (5):

ISSN: 1980-055x (Impressa) 2237-8642 (Eletrônica)

$$\alpha = \frac{\sum x^2 \sum y - \sum(xy) \sum x}{n \sum x^2 - (\sum x)^2} \qquad (4)$$

$$\beta = \frac{n \sum(xy) - \sum x \sum y}{n \sum x^2 - (\sum x)^2} \qquad (5)$$

No caso da regressão linear múltipla, o valor a deduzir se baseia em n variáveis. A Eq. (6) define a regressão linear múltipla:

$$y = \alpha_0 + \sum_{i=n}^{n} \alpha_i x_i \qquad (6)$$

Onde α é um vetor de coeficientes calculados para utilização neste método.

2.2. Dados utilizados e simulações

Para este estudo foram utilizados dados das estações automáticas do INMET (www.inmet.gov.br). Foram obtidas três séries de dados de cidades diferentes: Manaus - AM, Rio de Janeiro – RJ e Porto Alegre – RS. Cada série de dados possui medidas de hora em hora, durante 90 dias. Os sensores escolhidos das séries de dados foram: temperatura (T), umidade relativa do ar (UR), ponto de orvalho (dew), pressão atmosférica (P) e radiação solar (Rg). A Tabela 1 mostra exemplos de dados de uma das séries utilizadas e a Tabela 2 mostra as unidades de cada variável climática utilizada.

Tabela 1 - Exemplo de dados utilizados neste trabalho

hora	T	UR	*dew*	P	Rg
21	25.1	89	23.1	1002.4	190.1
22	24.8	89	22.9	1003.0	43.34
23	24.6	88	22.5	1003.6	-3.25
0	25.9	87	23.6	1004.4	-2.18
1	25.7	89	23.8	1004.9	-2.98
2	25.7	88	23.6	1005.7	-2.82

Tabela 2 - Unidades das variáveis climáticas utilizadas

Variável climática	Unidade
Temperatura	ºC
Umidade relativa do ar	%
Ponto de orvalho	ºC
Pressão atmosférica	hPa
Radiação solar	KJ/m²

ISSN: 1980-055x (Impressa) 2237-8642 (Eletrônica)

Pelo fato dos dados não possuírem falhas em suas séries, um algoritmo foi criado para simular falhas nos sensores. Dessa forma foi possível comparar os dados estimados pelos métodos com os dados que foram realmente registrados pelos sensores.

Duas simulações foram feitas, variando a quantidade de falhas aleatórias inseridas nas séries de dados. A primeira simulação removeu 5% dos dados da variável climática analisada, com o propósito de avaliar os métodos quando há poucas falhas nas séries de dados. A segunda simulação removeu 30% da série, objetivando avaliar os métodos quando há várias falhas, inclusive algumas em sequência.

2.3. Avaliação estatística

A avaliação estatística de cada método de preenchimento de falhas testado foi baseada nos erros individuais e_i (i=1,2,...,n) de cada estimativa, mostrado na Eq. (7), onde P_i são os valores estimados e O_i são os valores reais (WILLMOTT; MATSUURA, 2005):

$$e_i = P_i - O_i \qquad (7)$$

Com o erro individual de cada estimativa, é calculado o desempenho do modelo usando o Erro Médio Absoluto (EMA), mostrado na Eq. (8). Segundo Willmott et al. (2009), essa é a melhor forma de avaliação para modelos ambientais, devido principalmente a presença de outliers e dados com desvio de normalidade.

$$EMA = \frac{\sum_{i=1}^{n} |e|}{n} \qquad (8)$$

Além disso, também foi calculado para os testes de preenchimento de falhas o coeficiente de correlação de Pearson (PEARSON, 1896), mostrado na Eq. (9).

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (9)$$

Além do EMA e do coeficiente de correlação, o tempo médio de processamento foi registrado para as devidas comparações.

3. RESULTADOS E DISCUSSÕES

Os resultados obtidos no preenchimento das falhas podem ser visualizados nas Tabelas 3 e 4. Foram realizados os cálculos do EMA, sendo que na Tabela 3 são exibidos os resultados para a simulação com 5% de falhas e na Tabela 4 os resultados para a simulação com 30% de falhas.

ISSN: 1980-055x (Impressa) 2237-8642 (Eletrônica)

Tabela 3 – Valores do Erro Médio Absoluto para 5% de falhas.

Métodos	T	UR	dew	P	Rg
Média Simples (MS)	0,49	2,59	0,37	**0,31**	**280,24**
Média Móvel (MM)	1,09	4,92	0,5	0,87	506,06
Regressão Linear Simples (RLS)	1,54	5,79	1,47	2,48	661,43
Regressão Linear Múltipla (RLM)	**0,29**	**1,23**	**0,24**	12,29	520,02

Tabela 4 – Valores do Erro Médio Absoluto para 30% de falhas.

Método	T	UR	dew	P	Rg
Média Simples (MS)	0,63	2,97	0,39	**0,41**	**381,31**
Média Móvel (MM)	1,33	5,81	0,55	1,01	663,1
Regressão Linear Simples (RLS)	1,5	5,73	1,51	2,35	611,57
Regressão Linear Múltipla (RLM)	**0,25**	**1,15**	**0,24**	11,92	535,13

Para a simulação com 5% de falhas, os melhores métodos foram a média simples, para as variáveis de pressão e radiação, e regressão linear múltipla, para as variáveis de temperatura, umidade e ponto de orvalho. Apenas para a variável de radiação houve um desempenho ruim (erro de 381,31 KJ/m^2 no melhor caso).

A Tabela 4 mostra os valores obtidos na simulação com 30% de falhas nas séries de dados que, apesar do aumento das falhas, houve pouca redução da precisão da estimativas dos valores ausentes. Para os métodos de média simples e média móvel, os testes de preenchimento de falhas em todas as variáveis tiveram uma menor precisão com o aumento das falhas, conforme o esperado. Entretanto, os métodos de regressão linear foram mais robustos com relação ao aumento das falhas. Assim como na simulação de 5% de falhas, na de 30% os métodos de regressão linear múltipla e média simples tiveram os melhores resultados.

Para uma melhor comparação entre os métodos, os coeficientes de correlação obtidos nos testes foram agrupados nos gráficos das Figuras 1 e 2, representando os resultados para as simulações de 5% e 30%, respectivamente.

ISSN: 1980-055x (Impressa) 2237-8642 (Eletrônica)

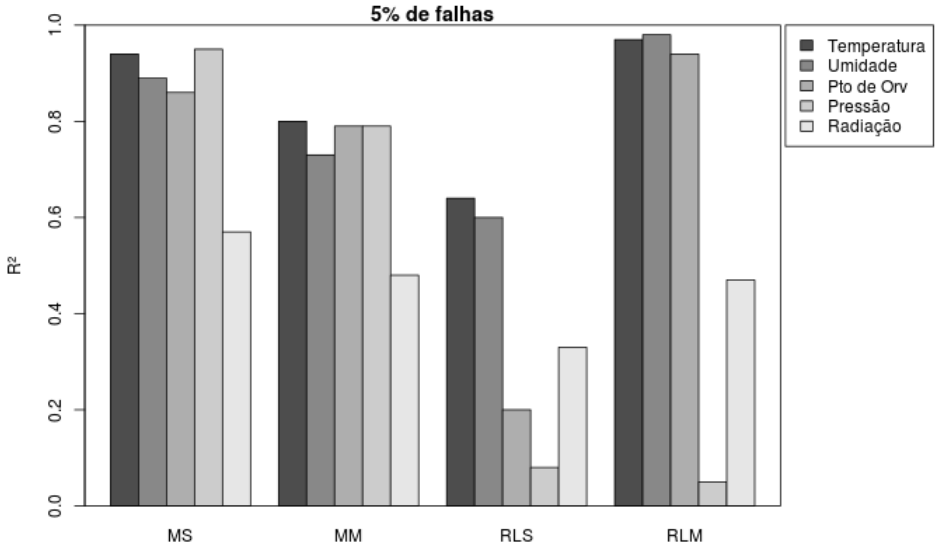

Figura 1 – Coeficiente de correlação das variáveis testadas para séries com 5% de falhas.

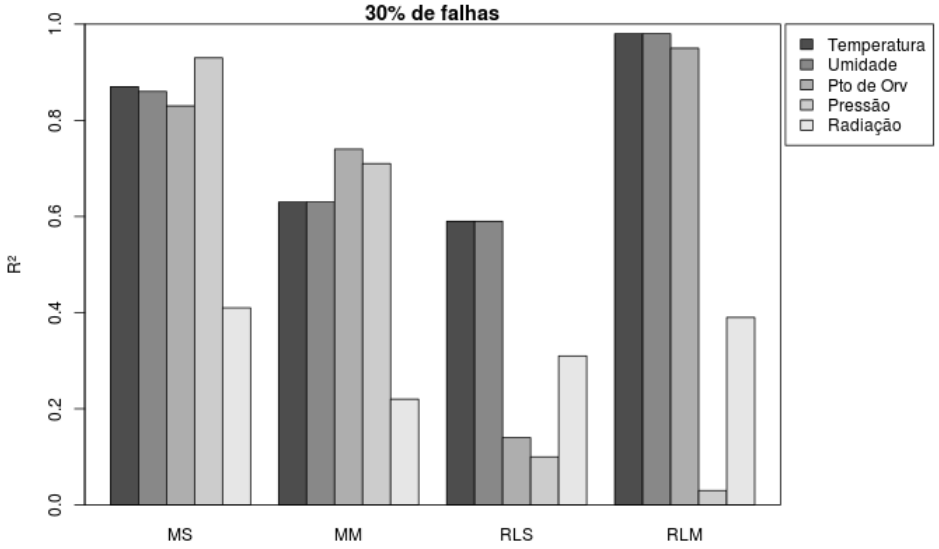

Figura 2 – Coeficiente de correlação das variáveis testadas para séries com 30% de falhas.

É possível perceber que para temperatura, umidade e ponto de orvalho, a MS e a RLM tiveram ótimos resultados (r^2 >= 0,86 nas duas simulações). O preenchimento na variável de pressão também obteve sucesso, com um r^2 de 0,93 (5% de falhas) e de 0,95 (30% de falhas) com a MS.

ISSN: 1980-055x (Impressa) 2237-8642 (Eletrônica)

Outra observação importante é que para a variável de pressão, a MS possui um valor de erro muito mais baixos que os outros métodos, ao mesmo tempo que a RLM, método que teve ótimos resultados para as outras variáveis, possui um erro consideravelmente alto para a pressão. Isso mostra que nenhum método tem a capacidade de modelar todas as variáveis climáticas, já que cada uma tem um comportamento diferente.

Como mostrado na análise do MAE, apenas a variável de radiação teve resultados ruins, r^2 de 0,22 no pior caso (MM em ambas as simulações) e r^2 de 0,57 no melhor caso (MS com 5% de falhas). Essa variável tem um comportamento muito dinâmico, sendo difícil a modelagem da mesma com uma série de dados horária.

O tempo de processamento de cada método também foi calculado. A Tabela 5 mostra a média do tempo necessário para a execução de cada teste para cada método.

Tabela 5 – Tempo médio de processamento para cada método

Método	Tempo
Média Simples (MS)	00:00.001
Média Móvel (MM)	00:00.002
Regressão Linear Simples (RLS)	00:00.001
Regressão Linear Múltipla (RLM)	00:00.004

Nenhum dos quatro métodos avaliados demanda um grande tempo de processamento. Portanto, a escolha do método a ser utilizado pode ser feita apenas avaliando a sua respectiva precisão.

4. CONCLUSÃO

Este trabalho comparou o desempenho de quatro métodos estatísticos para preenchimento de falhas em séries temporais: média simples, média móvel, regressão linear simples e regressão linear móvel. A comparação foi baseada nos resultados dos testes que envolveram diferentes variáveis climáticas, com diferentes proporções de falhas nas séries de dados.

Os resultados mostram que a regressão linear múltipla teve ótimo desempenho em três das cinco variáveis climáticas testadas (temperatura, umidade e ponto de orvalho), enquanto a média simples modelou com precisão a variável climática de pressão atmosférica. Entretanto, não pode ser dito que houve bom desempenho para o preenchimento de falhas para a variável climática de radiação, uma vez que o melhor resultado ficou abaixo de 0,6 para o coeficiente de correlação. Ou seja, o preenchimento de falhas pode ser realizado com precisão para as variáveis climáticas de temperatura, umidade, ponto de orvalho e pressão.

ISSN: 1980-055x (Impressa) 2237-8642 (Eletrônica)

É possível perceber também que nenhum dos quatro métodos avaliados foi ótimo para todas as variáveis. Logo, escolher o método de preenchimento de falhas de acordo com a variável climática que está sendo tratada pode resultar em uma maior precisão no procedimento.

Como trabalhos futuros, novos testes podem ser realizados para avaliar a aplicabilidade do preenchimento de falhas em outras variáveis climáticas. Outros métodos também podem ser testados, principalmente para verificar se algum tem a capacidade de preencher falhas na variável de radiação em séries de dados horárias.

5. Agradecimentos

Os autores agradecem o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo auxílio financeiro em bolsas de pesquisa.

6. Referências Bibliográficas

BAMBINI, M. D.; FURTADO, A. T. Redes observação e a evolução tecnológica contribuindo para o desenvolvimento de modelos matemáticos na Meteorologia no século XX In: *SEMINÁRIO NACIONAL DE HISTÓRIA DA CIÊNCIA E DA TECNOLOGIA, 12.; CONGRESSO LATINO-AMERICANO DE HISTÓRIA DA CIÊNCIA E DA TECNOLOGIA*, 7., 2010, Salvador. Anais... [S.l.]: Sociedade Brasileira da História da Ciência, 2010.

BARBOZA, C. H. *História da Meteorologia no Brasil*. Disponível em www.sbmet.org.br/userfiles/Historia_Meteorologia.pdf>. Acesso em: 26/06/2015.

FUENTES, E. V.; BITENCOURT, D. P.; FUENTES, M. V. Análise da velocidade do vento e altura de onda em incidentes de naufrágio na costa brasileira entre os estados do Sergipe e do Rio Grande Do Sul. *Revista Brasileira de Meteorologia*, V. 28, Nº 3, 257-266, 2013.

GENÇAY, R. Non-linear prediction of security returns with moving average rules. *Journal of Forecasting*, 1996. John Wiley and Sons, Ltd., v. 15, n. 3, p. 165–174,1996.

GIANOTTI, A. R. C.; SOUZA, M. J. H.; MACHADO, E. L. M.; PEREIRA, I. M.; VIEIRA, A. D.; MAGALHÃES, M. R. Análise microclimática em duas fitofisionomias do cerrado no alto vale do Jequitinhonha, Minas Gerais. *Revista Brasileira de Meteorologia*, V. 28, Nº 3, 246 - 256, 2013.

HASAN, M. M.; CROKE, B. F. W. Filling gaps in daily rainfall data: a statistical approach. In: *International Congress on Modelling and Simulation*, 20, 2013, Austrália.

ISSN: 1980-055x (Impressa) 2237-8642 (Eletrônica)

HUI, D. *Gap-filling missing data in eddy covariance measurements using multiple imputation (mi) for annual estimations*. Agricultural and Forest Meteorology, 2004. v. 121, n. 1–2, p. 93–111, 2004.

INMET. *Instituto Nacional de Meteorologia*. 2014. Http://www.inmet.gov.br. Acessado em Junho/2015.

MAGINA, F. C.; SOUZA, L. E. Rede automática de coleta de dados meteorológicos para utilização em projetos e operação de linhas de transmissão de energia elétrica In: *SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO*, 12., 2007, Florianópolis. Anais... Florianópolis: INPE/SELPER.

OLIVEIRA, A. G. *A importância dos dados das variáveis climáticas nas pesquisas em geografia: um estudo de caso empregando a precipitação pluviométrica*. Uberlandia – GO, V. 10, Nº 32, 9-21, 2009.

PEARSON, K. *Mathematical contributions to the theory of evolution*. III. Regression, heredity, and panmixia. Philosophical Transactions of the Royal Society of London. Series A, 1896. The Royal Society, v. 187, p. 253–318, 1896.

TARDIVO, G.; BERTI, A. The selection of predictors in a regression-based method for gap filling in daily temperature datasets. *Int. J. Climatol.,* 34: 1311–1317. JUN/2013.

UYANıK, G. K.; GÜLER, N. A study on multiple linear regression analysis. Procedia - *Social and Behavioral Sciences*, 2013. v. 106, n. 0, p. 234 – 240, 2013.

VENTURA, T. M. Criação de um Ambiente Computacional para Detecção de Outliers e preenchimento de Falhas em Dados Meteorológicos. Cuiabá, 2015, 96f*. Tese* (Doutorado em Física Ambiental) - Instituto de Física, Universidade Federal de Mato Grosso.

WILLMOTT, C.; MATSUURA, K. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, 2005. v. 30, p. 79–82, 2005.

WILLMOTT, C. J.; MATSUURA, K.; ROBESON, S. M. *Ambiguities inherent in sums-of-squares-based error statistics.* Atmospheric Environment, 2009. Elsevier Ltd, v. 43, n. 3, p. 749–752, 2009.

Texto submetido à RBClima em 01/02/2016