Ciência e Natura, Santa Maria, v. 36 Ed. Especial II, 2014, p.503–509
Revista do Centro de Ciências Naturais e Exatas - UFSM
ISSN impressa: 0100-8307 ISSN on-line: 2179-460X

CIÊNCIA e NATURA 35 anos

Preenchimento de falhas em dados de correlação de Anomalia da altura geopotencial (500 hPa)

Filling the gaps on *geopotential height* (*500 hPa) anomaly correlation* dataset

Rildo Gonçalves de Moura*[1], José Antônio Aravéquia[1], Alexandre Boleira Lopo[2]

[1]CPTEC/INPE - Cachoeira Paulista - São Paulo - rildo.moura@cptec.inpe.br
[2]Programa de Pós-Graduação em Ciências climáticas– PPGE/UFRN

Resumo

Esta pesquisa trata sobre a imputação múltipla de dados faltantes (missing data) da correlação de anomalia da altura geopotencial em 500 hPa. A imputação múltipla ocorreu via método da média preditiva ou Predictive Mean Matching (PMM). Os resultados permitiram inferir que a imputação dos dados teve uma qualidade adequada sendo que a série reconstruída de dados da correlação de anomalia da altura geopotencial em 500 hPa do modelo global possui em média 14.8% mais dados que a série original. O processo de imputação múltipla por PMM preencheu os dados ausentes com uma qualidade adequada, visto que há um forte coeficiente de correlação de Pearson (> 0.99), entre a série formada com dados originais e a série formada com dados recuperados, para todos os horários de previsão mostrados.

***Palavras-chave**: imputação múltipla, MICE, dados ausentes, Método da Média Preditiva*

Abstract

This research deals with the multiple imputation of missing data of Geopotential Height Anomaly Correlation at 500 hPa. The multiple imputation was reached by using the Predictive Mean Matching (PMM) method. Outcome showed that multiple imputation (PMM) resulted in an acceptable quality of data, once the reconstructed data serie of Geopotential Height Anomaly Correlation of the global model comprises, in average, 14.8% more data in comparison to the original data serie. Also, the multiple imputation process by PMM filled the missing data with adequate quality, since they showed a strong coefficient of Pearson correlation (> 0.99) between original data serie and reconstructed data serie, for every weather prediction period considered.

***Keywords**: Multiple Imputation, MICE, missing data, Predictive Mean Matching*

* rildo.moura@cptec.inpe.br
Recebido: 13/03/2014 Revisado: 19/05/2014 Aceito: 13/03/2014

1 Introdução

Uma das maiores dificuldades encontradas em investigações científicas está relacionada à escassez, a falha nas séries ou ainda a completa ausência dos dados (*missing data*) das variáveis em investigação. Um dos primeiros esforços no sentido de superar estes problemas está no preenchimento de falhas em dados faltantes que surgiram na década de 70, esses envolviam métodos simples, tais como, substituição dos dados faltantes pela média, pela mediana, por interpolação e por regressão linear (RUBIN, 1987; LITTLE & RUBIN, 2002). Atualmente, métodos estatísticos e meios computacionais mais avançados permitem um preenchimento de falhas mais eficaz.

O preenchimento de falhas ocorre através dos chamados métodos de imputação de dados. Nesses métodos os dados omissos são normalmente preenchidos com a média dos dados existentes, com valores previstos por uma regressão sobre os dados existentes, com valores constantes oriundos de uma fonte externa, pelo vizinho mais próximo ou outros métodos estatísticos. As modificações devem ser feitas para garantir diferenciação entre os dados reais e aqueles recuperados ou imputados (LITTLE & RUBIN, 2002).

Os métodos de imputação de dados vêm sendo bastante utilizado nos últimos anos na reconstrução de séries e consiste em preencher os dados faltantes com valores plausíveis e então aplicar os métodos tradicionais de análise de dados completos a fim de fazer inferências válidas e eficientes nas séries preenchidas (NUNES *et al.*,2009).

Um método de imputação pode ser classificado como simples ou única (IU), em que se realiza uma única imputação ou múltipla (IM) com várias imputações. O método de IM foi desenvolvido para solucionar a limitação do método de IU que não considera a incerteza associada à imputação (RUBIN, 1987, KENWARD; CARPENTER, 1997).

A IU tem sido bastante usada pela sua facilidade de aplicação, entretanto, existem desvantagens na utilização desse método, como a subestimação da variabilidade e a impossibilidade da utilização de outras variáveis do próprio conjunto de dados para melhorar o processo de imputação. Vale ressaltar a sua simplicidade e a importância de empregá-la exclusivamente para a pequena proporção de dados faltantes (NUNES *et al.*,2009; LITTLE & RUBIN, 2002).

O método de IM foi desenvolvido para solucionar a limitação do método de IU relacionado à incerteza da imputação. Na imputação múltipla os valores ausentes são substituídos por um conjunto de "k" valores plausíveis. Os "k" conjuntos de dados preenchidos produzem "k" diferentes conjuntos de estimativas de parâmetros e erros padrões, as estimativas são combinadas para fornecer uma única estimativa dos parâmetros de interesse, permitindo que a incerteza seja considerada. A técnica de IM foi proposta inicialmente para resolver o problema de não resposta em pesquisas (RUBIN, 1987, KENWARD; CARPENTER, 1997).

A partir de análises preliminares feitas nos dados de Correlação de Anomalia da Altura Geopotencial, optou-se por utilizar o método de IM, ou seja, o método que considera a incerteza associada à imputação. A técnica de imputação múltipla escolhida foi *Preditive Mean Matching* (PMM) ou Método da Média Preditiva (BUUREN & OUDSHOORN, 2011; LITTLE & RUBIN, 2002; RUBIN, 1987).

Este método foi aplicado para solucionar o problema de dados faltantes das séries originais de precipitação acumulada mensal no Distrito Federal gerando resultados promissores para a escala mensal, quando comparado com os dados originais (CONDE *et al.,* 2010) e em dados de Índice de Radiação Ultravioleta (IUV) da cidade de Natal-RN obtendo uma correlação alta (0,97) entre os grupos das médias formadas por dados com valores faltantes e o grupo das médias formadas por dados imputados (LOPO *et al.,* 2012).

PPM é considerado um método de baixa incerteza em razão de combinar elementos de regressão, vizinho mais próximo e imputação *hot deck* (DURRANT, 2005). O valor imputado através do PMM é calculado pelo modelo de regressão mais próximo do valor observado, e assim supera algumas outras técnicas de IM, entretanto Landerman *et al.*,(1997) explica que o desempenho do método varia consideravelmente com o poder preditivo do modelo de regressão de imputação e o percentual de casos com dados ausentes.

A finalidade deste artigo é aplicar o método de IM *Preditive Mean Matching* (PMM) aos dados ausentes da correlação de anomalia do geopotencial em 500 hPa e analisar a qualidade de seus resultados.

2. Material e métodos

O estudo foi dividido em 3 (três) etapas: (1) organização e consistência de dados; (2) imputação dos dados via PMM; e (3) análise da qualidade do método de imputação.

2.1. Dados

Foram utilizados dados mensais do índice de Correlação de Anomalia (CA), definido como a correlação linear entre as anomalias dos valores previstos e as anomalias das análises, ambas em relação à climatologia do modelo aplicado. Este índice foi proposto por Brankovic *et al.*(1990).

Dentro desse contexto, empregou-se especificamente o índice de CA fornecido pela altura geopotencial, no nível de 500 hPa, o qual é recomendado pela Organização Meteorológica Mundial – OMM, para verificação percentual do grau de acerto da previsão de um modelo

de PNT (Previsão Numérica de Tempo).

Neste trabalho, utilizaram-se 7 (sete) dias de previsão, seguindo a evolução das resoluções do modelo global do Centro de Previsões de Tempo e Estudos Climáticos (CPTEC) do Instituto Nacional de Pesquisas Espaciais (INPE) T062L28 [1996-2003], T126L28 [2004-2005] e T213L64 [2006-2012], apenas para região da América do Sul, compreendida entre [101.25W 11.25W – 60S 15N] no período total de janeiro de 1996 a dezembro de 2012.

2.2 Método de imputação múltipla preditive mean matching (pmm)

A imputação de dados pode ser utilizada desde que sejam respeitados rigorosamente os critérios de proporção de dados faltantes, relacionados a seguir: proporção menor ou igual a 5%: pode-se utilizar o método de IU ou ainda considerar apenas o banco de dados completo; proporção entre 5% e 15%: é possível utilizar o método de IU sendo aconselhável o método de IM; e proporção maior ou igual a 15%: indica-se o uso de IM (HARRELL Jr, 2001).

O método de IM PMM utilizado neste trabalho é considerado, como dito anteriormente, um método de baixa incerteza em razão de combinar elementos de regressão, vizinho mais próximo e imputação *hot deck* (técnicas paramétricas e não paramétricas). O PPM assim pode superar as dificuldades destas técnicas de imputação, dado o fato de que as técnicas paramétricas podem falhar quando o modelo não é adequado para os dados disponíveis e as técnicas não paramétricas exigem grande quantidade de observações (DI ZIO & GUARNERA, 2009).

Para Schafer (2011) e Li *et al* (1991) PMM é uma variante de regressão linear, que determina um valor imputado calculado pelo modelo de regressão mais próximo do valor observado. É considerado mais preciso que outros métodos de imputação por estas características conforme Horton & Lipsitz (2001).

PMM considera a seguinte formulação (equação 1.0) para cada i faltante em Y (SCHAFER, 2011; LI *et al*., 1991).

$$\hat{Y}^{obs} = \left\{ Y_i^{obs} = X_i' \beta^* : i \in obs(Y) \right\}$$

Sendo X uma variável sem dados faltantes, Y_j^{obs} o conjunto de valores observados; $Y_i^* = X_i' \beta^*$; e $\hat{Y}^{obs}$; e a observação encontrada correspondente ao valor mais próximo de Y_j^*.

Os métodos de imputação podem ser avaliados em termos de qualidade da imputação quanto a sua acurácia, concordância e dispersão. Existem indicadores de desempenho dos quais se destacam, o erro quadrático médio, o erro médio absoluto, o viés (bias), variância proporcional e coeficiente de correlação de Pearson (r), este último foi utilizado neste artigo, pois é o indicador mais adequado para avaliar o desempenho de métodos de IM, como o PMM (NUNES *et al*., 2009).

Ressalta-se que estudo de Horton & Lipsitz (2001) mostrou que os resultados da aplicação do PMM foram promissores para a escala mensal, quando comparados com os dados originais, embora o desempenho varie consideravelmente com o poder preditivo do modelo de regressão e o percentual de dados ausentes (LANDERMAN *et al*., 1997). Lembrando-se que os procedimentos para o preenchimento dos valores em falta utilizando o método da PMM seguem critérios e devem ser ajustadas localmente com viés reduzida (OKAMOTO, 2006).

A imputação de dados PMM foi realizada a partir de um software livre "R" desenvolvido para computação estatística (R Development Core Team, 2012) por meio do aplicativo (pacote) denominado MICE (*Multivariate Imputation by Chained Equatoins*). O MICE permite programar a sua própria função de imputação, ao mesmo tempo em que suporta uma variedade de métodos de imputação (HORTON & KLEINMAN, 2007). Este pacote pode ser usado em plataformas UNIX, Windows e MacOS.

3. RESULTADOS

Foram utilizados neste trabalho dados de correlação de anomalia da altura geopotencial em 500 hPa, referentes a informações de duas rodadas de modelo, uma inicializada as 00 e outra 12 horário GMT (Greenwich Mean Time), no período de 1996 a 2012, para cada um dos 7 (sete) dias de previsão, obtidos a partir do modelo global (CPTEC/ INPE), com exceção da previsão de 168 horas [sétimo dia de previsão], que teve seu início em setembro de 1999.

Sendo assim, caso os dados estivessem completos, deveriam existir 12.420 dados para os 6 (seis) primeiros dias de previsão e um número menor para o sétimo, da ordem 9.742 dados.

As análises mostraram diferentes quantidades de dados faltantes para cada um das previsões do modelo. Sendo que o valor médio desses dados, calculado entre as 7 (sete) previsões foi da ordem de 10.270 dados ou 14,8%. Estes valores são apresentados na Tabela 3.1.

Tabela 3.1 - Distribuição das médias de dados absolutos e percentuais faltantes da correlação de anomalia da altura geopotencial (500 hPa) por horário de previsão.

Previsão	Dados absolutos	% faltante
24 h	10768	13.3
48 h	10744	13.5
72 h	10733	13.6
96 h	10725	13.6
120 h	10655	14.2
144 h	10441	15.9
168 h	7825	19.7
Média	10270	14.8

A Figura 3.1 retrata a série temporal original da correlação de anomalia da altura geopotencial em 500 hPa, referente ao período de janeiro de 1996 a dezembro de 2012. No qual se observa a falha ocorrida, principalmente no ano de 2000 (área demarcada), decorrente de interrupção no processamento do modelo, além de outras pequenas falhas (pontuais) ocorridas ao longo da série de dados. Em seguida, é mostrada na Figura 3.2 a série de dados reconstruída, ou seja, incluindo os dados recuperados.

A Tabela 3.2 apresenta os resultados estatísticos tanto da série original quanto da imputada ou reconstruída, para o período de 17 anos de dados da correlação de anomalia da altura geopotencial em 500 hPa, referente a janeiro de 1996 e dezembro de 2012.

A região demarcada, na qual havia a maior falha de dados e também as falhas pontuais, foram preenchidas conforme indica a Figura 3.2.

Na Figura pode-se notar que os valores imputados pouco alteraram a variabilidade e a estatística da série original, uma vez que os valores obtidos com a imputação múltipla via PMM foram bastante semelhantes aos dados originais (Tabela 3.2).

Os índices de correlação de Pearson "r" entre os dados originais e os dados recuperados, para os 7 (sete): 24 [preto], 48 [vermelho], 72 [azul escuro], 96 [azul claro], 120 [amarelo], 144 [laranja] e 168 [verde], horários de previsão foram da ordem de 0,99.

A condição anterior pode ser verificada na Figura 3.3, a qual mostra o índice de correlação de Pearson referente ao período de janeiro de 1996 a dezembro de 2012.

Tabela 3.2 – Resultados estatísticos tanto da série original quanto da imputada ou reconstruída, por horários de previsão, para os 17 anos de dados.

PREV.	MÉDIA	MEDIANA	VARIÂNCIA	DESV, PAD
Dados originais				
24 h	98	98,4	6,3	**2,5**
48 h	95,1	96	15,5	3,9
72 h	90,3	92,1	46	6,8
96 h	83,5	86,5	119,1	10,9
120 h	75	78,8	231,4	15,2
144 h	66,4	70,2	331,6	18,2
168 h	59,6	62,7	384,5	19,6
Dados reconstruídos				
24 h	98	98,4	9,3	3
48 h	94,9	96	22,5	4,7
72 h	90,1	91,9	52	7,2
96 h	83,3	86,4	127,2	11,3
120 h	74,6	78,6	250,6	15,8
144 h	65,2	69,6	388,9	19,7
168 h	57,4	61,3	462,8	21,5

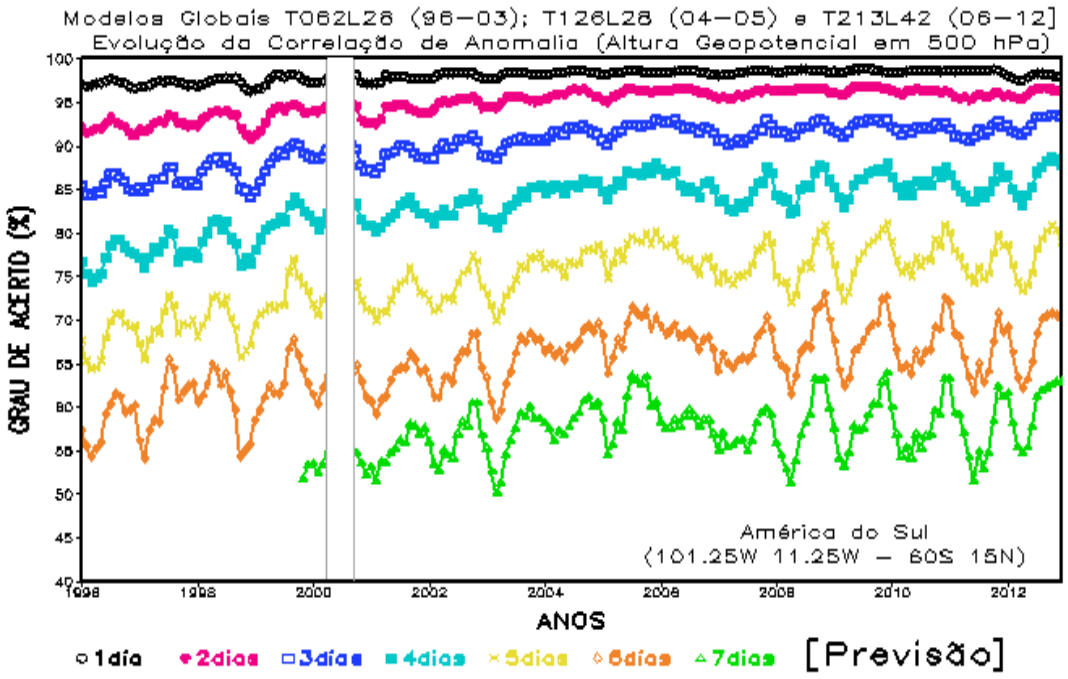

Figura 3.1 - Série temporal original da correlação de anomalia da altura geopotencial em 500 hPa, referente ao período de janeiro de 1996 a dezembro de 2012.

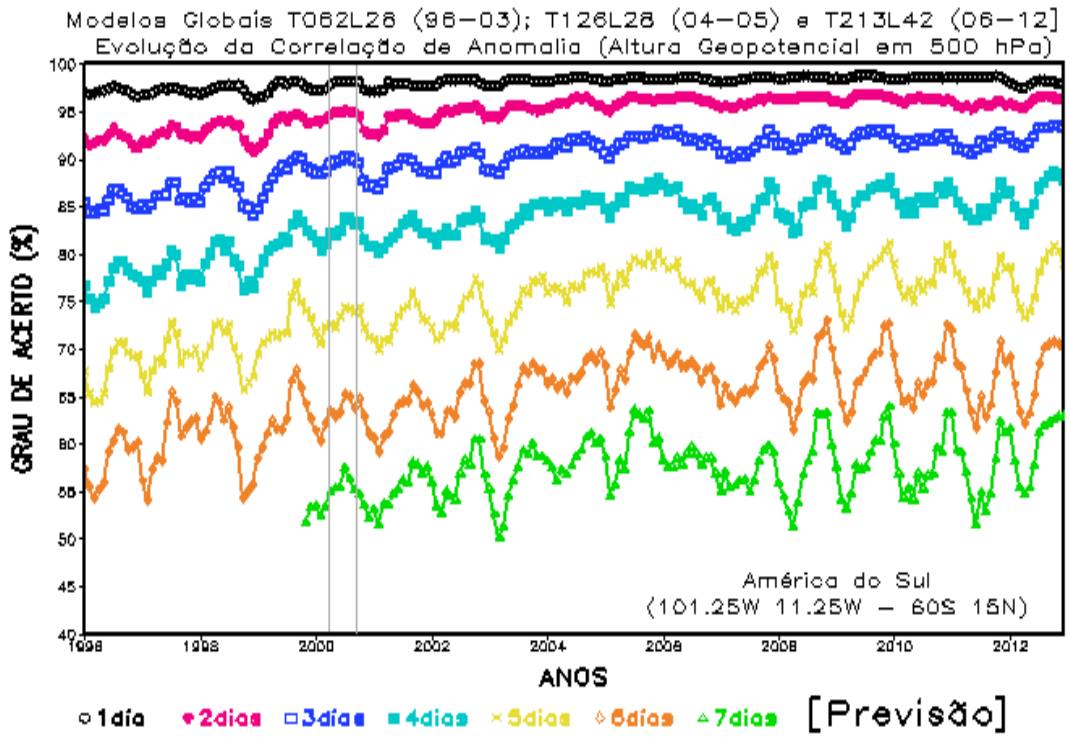

Figura 3.2 - Série temporal imputada [reconstruída] da correlação de anomalia do geopotencial em 500 hPa, referente ao período de janeiro de 1996 a dezembro de 2012.

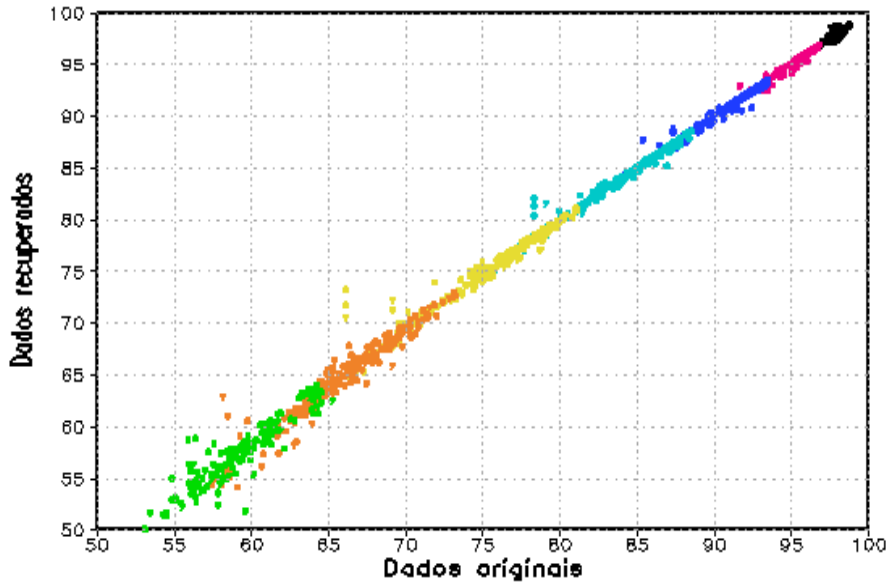

Figura 3.3 - Índice de correlação de Pearson, entre a correlação de anomalia da altura geopotencial em 500 hPa com dados originais e os dados imputados, referente ao período de janeiro de 1996 a dezembro de 2012, por horário de previsão: 24 [preto], 48 [vermelho], 72 [azul escuro], 96 [azul claro], 120 [amarelo], 144 [laranja] e 168 [verde].

4. Considerações finais

A série reconstruída de dados da correlação de anomalia da altura geopotencial em 500 hPa do modelo global possui em média 14,8% mais dados que a série original. O processo de imputação múltipla por PMM preencheu os dados ausentes com uma qualidade adequada, visto que há uma fortíssima correlação de 0.99, entre a série formada com dados originais e a série formada com dados recuperados, para todos os horários de previsão mostrados.

Os resultados apresentados neste trabalho mostram que a utilização do PMM para reconstrução de dados de séries de correlação de anomalia obteve resultados promissores para a escala usada, quando comparado com os dados originais, tornando-se assim uma ferramenta bastante útil para a meteorologia como um todo, inclusive para as variáveis usadas no clima, uma vez que para este tipo de estudo necessita-se de longas séries temporais de informações, as quais normalmente apresentam falhas.

Porém, é importante dizer que um estudo bem elaborado para preenchimento de falhas pode, entre outras coisas, aumentar consideravelmente a confiabilidade dos resultados obtidos. Além disso, as estratégias que lidam com dados faltantes visam principalmente aumentar o tamanho efetivo do conjunto de dados, tornando as análises mais poderosas.

Um aspecto interessante é mostrar que inúmeros trabalhos com foco neste tipo de metodologia são produzidos nas mais diversas áreas do conhecimento, o que indica que o estudo de séries sintéticas ou reconstruídas pelos métodos de imputação múltipla, como o PMM, pode ser uma ferramenta bastante útil para a investigação de variáveis climáticas.

Agradecimentos

Os autores agradecem ao Centro de Previsões de Tempo e Estudos Climáticos (CPTEC) do Instituto Nacional de Pesquisas Espaciais (INPE) pela gentileza em fornecer os dados.

Referências

BRANKOVIC, C. *et al*. Extended-range predictions with ECMWF models: Time-lagged ensemble forecasting. **Quarterly Journal of the Royal Meteorological Society**, v.116, n. 494 p.867-912, 1990

BUUREN .V, S., OUDSHOORN , K.G. `mice`: Multivariate Imputation by Chained Equations in `R`. **Journal of Statistical Software**, 45(3), 1-67, 2011.

CONDE, F.C. RAMOS A, M. SANTOS L, A. R. FERREIRA, D. B. Reconstrução de Séries de Precipitação Acumulada Mensal do Distrito Federal via PMM, XVI Congresso Brasileiro de Meteorologia, Belém, 2010.

DI ZIO, Marco; GUARNERA, Ugo. Semiparametric predictive mean matching. **AStA Advances in Statistical Analysis**, v. 93, n. 2, p. 175-186, 2009.

DURRANT, G.B. **Imputation Methods for Handling Item-Nonresponse in the Social Sciences**: A Methodological Review. ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute (S3RI), University of Southampto, 2005.

HARRELL Jr. F.E. Regression modeling strategies: with applications to linear models, logistic regression and survival analysis. New York: Springer-Verlag, 2001.

HORTON, Nicholas J.; KLEINMAN, Ken P. Much ado about nothing. **The American Statistician**, v. 61, n. 1, 2007.

HORTON, Nicholas J.; LIPSITZ, Stuart R. Multiple imputation in practice: comparison of software packages for regression models with missing variables. **The American Statistician**, v. 55, n. 3, p. 244-254, 2001.

KENWARD, Michael G.; CARPENTER, James. Multiple imputation: current perspectives. **Statistical Methods in Medical Research**, v. 16, n. 3, p. 199-218, 2007.

LANDERMAN, Lawrence R.; LAND, Kenneth C.; PIEPER, Carl F. An empirical evaluation of the predictive mean matching method for imputing missing values. **Sociological Methods & Research**, v. 26, n. 1, p. 3-33, 1997.

LITTLE, R.J.A. and RUBIN, D.B. **Statistical analysis with missing data**. 2nd ed. New York: Wiley, 2002.

LI, Kim-Hung; RAGHUNATHAN, Trivellore E.; RUBIN, Donald B. Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. **Journal of the American Statistical Association**, v. 86, n. 416, p. 1065-1073, 1991.

LOPO, A. B.; SPYRIDES, M. H. C.; LUCIO, P. S. Imputação Múltipla via Preditive Mean Matching (PMM) em dados do índice de radiação ultravioleta da cidade de Natal. XVII Congresso Brasileiro de

Meteorologia, Gramado-RS, 2012.

NUNES, L. N.; KLUCK, M. M.; FACHEL, J. M. G.. Multiple imputations for missing data: a simulation with epidemiological data. **Cad. Saúde Pública** [online]. 2009, vol.25, n.2, pp. 268-278. ISSN 0102-311X. http://dx.doi.org/10.1590/S0102-311X2009000200005

OKAMOTO, Masato. Bias-reduced multivariate imputation: use of the locally-adjusted Predictive Mean Matching method. 2006.

SCHAFER, Joseph L. **Analysis of incomplete multivariate data**. CRC press, 2010.

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

RUBIN, D.B. Multiple imputation for Nonresponse in Surveys. New York: Wiley, 1987.