A Influência da Variabilidade dos Dados na Qualidade de Imputação de Dados Faltantes

Elisandra Lúcia Moro Stochero[1], Luciane Flores Jacobi[2], Alessandro Dal'Col Lúcio[3]

1. Introdução

É comum deparar-se com alguns fatores que interferem na qualidade dos resultados de uma análise estatística, como a perda de informações durante o processo de coleta dos dados no decorrer do estudo. Problemas que podem ser gerados com a perda de dados são destacados por Nunes, Klück e Fachel (2010) como ineficiência e viés nas estimativas.

Schafer e Graham (2002) levantam o questionamento sobre tais dificuldades criadas pelos dados faltantes na pesquisa científica, segundo eles isso se deve ao fato de que a maioria dos procedimentos de análise de dados não foram projetados para eles. A falta é geralmente um "contratempo" encontrado durante a pesquisa e não o foco principal, mas a maneira como tratado pode levantar dificuldades conceituais e desafios computacionais.

Como uma forma de contornar possíveis complicações geradas pela ausência de dados, de acordo com Dias e Albieri (1992), foram desenvolvidas técnicas de imputação de dados, onde são inseridas nas caselas vazias estimativas dos valores ausentes.

Técnicas de Imputação de Dados permitem preencher o banco de dados e assim ter uma matriz de dados completa, porém, há uma certa incerteza na validade dos resultados obtidos pelo fato de não serem os valores reais.

Inicialmente, de acordo com Nunes, Kluck e Fachel (2009), os métodos aplicados eram simples, a substituição era feita pela média ou mediana da variável, por interpolação ou por regressão linear.

Os principais avanços em pesquisas envolvendo dados ausentes surgiram na década de 70 com a Estimação de Máxima Verossimilhança (EM) e Imputação Múltipla (IM), (ENDERS, 2010). Ainda, segundo o autor, Rubin foi responsável por estabelecer um sistema de classificação para problemas de dados ausentes.

Os métodos de imputação podem ser divididos, de acordo com Engels e Diehr (2003), em: Imputação Única (IU), ou simples, que ocorre quando os dados ausentes são substituídos uma única vez e Imputação Múltipla (IM) que consiste em imputar vários valores para cada dado ausente obtendo-se para cada valor estimado um banco de dados completo que são avaliados aplicando-se diferentes métodos, após é determinada uma estimativa pontual de um parâmetro obtido através das imputações realizadas.

Little e Rubin (1987), apresentaram o método de Imputação Múltipla e desde então, segundo Schafer e Graham (2002), estudos relacionados a este tema vem crescendo, novos métodos foram desenvolvidos e a aplicação dos mesmos utilizadas em diversas áreas.

Embora o número de pesquisas neste contexto tenha sido crescente, principalmente na área da saúde, nas Ciências Agrárias onde os métodos estatísticos de análise e delineamento de experimentos são distintos não há muitos trabalhos considerando métodos de imputação.

[1]UFSM. email: *elismoro2016@gmail.com*.
[2]UFSM. email: *lucianefj8@gmail.com*.
[3]UFSM. email: *adlucio@ufsm.br*.

Uma vez que a perda também pode ocorrer na experimentação agrícola, seja por rasuras, falta de preenchimento ou até mesmo perda de unidades experimentais, geralmente em pesquisas de campo, juntamente com a preocupação de obter resultados precisos e de qualidade, é interessante e de grande importância aprofundar e disseminar o estudo e aplicação de métodos de imputação.

Alarcón e Dias (2009) e Bergamo, Dias e Krzanowski (2008), por exemplo, trazem pesquisas mais recentes envolvendo as ciências agrárias no contexto de dados ausentes. Os dois estudos são direcionados para experimentos com interação Genótipo x Ambiente. Bergamo, Dias e Krzanowski (2008) inclusive propõem um novo método que abre mão das suposições referentes à distribuição dos dados ausentes.

No entanto, nenhum dos estudos tem como foco principal a influência que as características dos dados observados podem causar sobre os resultados dos métodos de imputação, sendo que para determinar as estimativas dos dados faltantes são utilizados os valores presentes no banco de dados.

Portanto, diante do exposto, o tema dessa pesquisa é verificar se a variabilidade dos dados provenientes de um experimento no Delineamento em Blocos Casualizados (DBC) influencia a qualidade da imputação dos dados perdidos.

2. Materiais e Métodos

Nessa pesquisa foi aplicado o método de imputação Método de Imputação Múltipla Livre de Distribuição (IMLD) aos conjuntos de dados reais de um experimento balanceado desenvolvido por alunos do Programa de Pós-graduação em Agronomia do Centro de Ciências Rurais da Universidade Federal de Santa Maria.

Os bancos de dados contêm informações sobre dois experimentos realizados para verificar as pressuposições do modelo matemático e avaliar o efeito da aplicação de bioproduto de batata no lugar da adubação mineral na produtividade, qualidade de frutos e coloração de folhas de tomateiro. Um dos experimentos foi conduzido em túnel plástico e outro a campo, a partir do delineamento blocos ao acaso com três repetições e 12 tratamentos.

Para o presente estudo foi considerado somente os bancos de dados completos obtidos na terceira colheita conduzida em campo e na terceira colheita em túnel plástico. Foram formados 8 bancos de dados balanceados, os quais foram chamados D1, D2, D3, D4, D5, D6, D7 e D8, com dados reais.

Onde, D1, D2 D3 e D4 representam os bancos de dados obtidos no experimento em campo ao avaliar peso de frutos por planta (g), média do número de frutos por planta, comprimento de frutos (mm) e largura média de frutos (mm), respectivamente. D5, D6, D7 e D8 representam os bancos de dados obtidos no experimento em túnel ao avaliar peso de frutos por planta (g), média do número de frutos por planta, comprimento de frutos (mm) e largura média de frutos (mm), respectivamente.

Após organizar os bancos de dados com as colunas sendo os blocos e as linhas os tratamentos, foram determinadas aleatoriamente as posições das observações a serem excluídas em todos os bancos de dados balanceados. Assim, foram gerados três novos bancos de dados desbalanceados a partir de cada banco de dado completo inicial, com a exclusão de 5%, 15% e 30% das observações.

Seguindo, foi aplicado o método de imputação IMLD e, a partir do "novo" banco de dados, agora incompleto, foram determinadas estimativas para cada dado faltante e as mesmas

comparadas com o respectivo valor real que foi retirado no primeiro momento deste processo, seguindo os mesmos passos usados por Bergamo, Dias e Krzanowski (2008).

Com os bancos de dados iniciais completos, desbalanceados e completos com o método de imputação, foi dado início a análise da variabilidade dos dados e a precisão dos resultados obtidos.

Para determinar a variabilidade considerou-se o coeficiente de variação dos dados, a média de todas as observações de cada banco e de cada coluna (blocos), pois a média de cada coluna é o ponto de partida para desenvolvimento do método de imputação.

Foram determinadas as medidas de precisão utilizadas por Bergamo, Dias e Krzanowski (2008), V_E, VQM e T_{acc}.Onde, T_{acc} é uma medida geral de desempenho, V_E representa a variação agrupada entre imputações dentro de posições, portanto, quanto maior for seu valor, menor será a precisão do método de múltiplas imputações e VQM representa o viés médio quadrado entre os valores estimados e valores observados, então quanto menor o viés, maior é o número de imputações que são semelhantes aos valores originais e maior é a sua precisão.

Também, verificou-se o coeficiente de variação experimental, o qual relaciona o desvio padrão em termos da porcentagem média aritmética. Em um primeiro momento é apresentado o coeficiente de variação dos valores presentes no banco de dados e em um segundo, o coeficiente de variação experimental.

Por fim, foram comparados os resultados obtidos com o objetivo de verificar se houve diferença entre os resultados dos dados com maior e menor variabilidade, se em alguns dos bancos de dados a qualidade de imputação foi melhor que em outros.

Estes procedimentos foram realizados no software R Core Team (2017) e no RStudio (2009 - 2017), utilizando o script implementado por Gandolfi (2010).

3. Resultados

Ao gerar os bancos de dados incompletos e completos com 5%, 15% e 30% de dados em falta e organizá-los, foi possível obter as médias, desvio padrão e coeficiente de variação dos dados de cada experimento.

Estes resultados foram considerados importantes no processo de análise devido ao fato de que, para iniciar o processo de geração dos bancos de dados imputados nas caselas vazias, é inserida a média da respectiva coluna.

Após os primeiros passos, foi dado início a obtenção dos resultados das medidas de avaliação de precisão VE, VQM e Tacc.

Na Tabela 1, é possível verificar estas medidas e compará-las com o resultado do coeficiente de variação do conjunto de observações dos bancos de dados iniciais.

Dos oito bancos de dados reais completos (balanceados), que foram disponibilizados para esta pesquisa, quatro apresentaram um baixo coeficiente de variabilidade e quatro um maior coeficiente de variabilidade.

Os bancos de dados D1 e D5 apresentaram uma maior variação agrupada entre imputações, indicada pela medida de VE, pode-se então dizer que a precisão do método é pequena. Também se nota que quanto maior o número de dados retirados menor o valor de VE, resultando em uma melhor precisão.

Ao avaliar esta medida em D2, D3, D4, D6, D7 e D8, percebe-se que os valores são baixos, indicando que a variação entre as imputações é baixa, portanto a precisão é boa.

Tabela 1 - Coeficiente de variação do conjunto de observações do banco de dados inicial e medidas de precisão dos valores imputados.

BANCO DE DADOS	COEFICIENTE DE VARIAÇÃO	MEDIDAS DE PRECISÃO								
		VE			VQM			Tacc		
	BANCO DE DADO INICIAL	5%	15%	30%	5%	15%	30%	5%	15%	30%
D1	51%	18	12	7	519930	170525	243185	519948	170538	243192
D2	49%	0,00007	0,00004	0,00003	0,00001	2	6	0,00008	2	6
D3	11%	0,00354	0,00003	0,00004	354	23	33	354	23	33
D4	11%	0,01709	0,00056	0,00028	778	76	45	778	76	45
D5	49%	15	6	3	16802	713611	496391	16817	713616	496395
D6	43%	0,00055	0,00005	0,00002	50	6	3	50	6	3
D7	15%	0,00342	0,00236	0,00135	65	23	60	65	23	60
D8	16%	0,00040	0,00003	0,00010	1634	286	274	1634	286	274

Fonte: Elaborada pelo autor.

Comparando os valores de VE com os coeficientes de variação, os resultados mostram que em todos os bancos de dados que apresentaram baixo coeficiente de variação a precisão entre imputações foram boas. Porém, quando o coeficiente de variação é mais alto, alguns bancos de dados apresentaram boa precisão entre imputações e outros uma baixa precisão.

Com objetivo de verificar se causaria alguma influência nos resultados da análise do experimento, se determinou os coeficientes de variação experimental dos bancos de dados originais, desbalanceados e balanceados após inserir as estimativas dos valores retirados, o que pode ser visto na Tabela 2.

Tabela 2–Coeficiente de variação experimental dos bancos de dados iniciais balanceados, desbalanceados e completos a partir da imputação.

BANCO DE DADOS	COEFICIENTE DE VARIAÇÃO						
	BANCO DE DADOS						
	COMPLETO	INCOMPLETO			COM IMPUTAÇÃO		
		5%	15%	30%	5%	15%	30%
D1	49,86%	50,24%	52,17%	47,98%	49,60%	47,20%	42,65%
D2	49,99%	51,03%	52,03%	45,10%	49,99%	47,44%	39,00%
D3	10,78%	10,71%	11,61%	13,50%	10,70%	10,45%	10,55%
D4	11,62%	11,33%	11,85%	12,06%	11,20%	11,22%	11,90%
D5	48,08%	47,45%	50,00%	51,94%	47,69%	46.61%	40,46%
D6	39,63%	39,60%	42,02%	47,07%	43,76%	39,28%	36,34%
D7	15,29%	12,67%	12,78%	10,93%	14,19%	14,44%	12,82%
D8	15,35%	12,65%	12,88%	10,87%	12,42%	12,67%	10,55%

Fonte: Elaborada pelo autor.

A partir dos resultados apresentados na Tabela 2, temos que D1, D2, D5 e D6 apresentaram um maior coeficiente de variação, o que indica uma baixa precisão. Estes coeficientes sofreram um aumento nos dados desbalanceados e se mantiveram próximos dos valores referentes aos dados iniciais quando imputadas as estimativas para as caselas vazias.

Já em D3, D4, D7 e D8, os coeficientes experimentais encontrados são baixos e, portanto, a precisão é alta. Nos dados desbalanceados e completos com a imputação ao analisar D3 e D4 pode-se observar que os resultados se mantiveram bem próximos dos originais.

Porém, em D7 a precisão melhorou com os dados desbalanceados e os valores do banco de dados com imputação ficaram mais próximos daqueles encontrados nos dados iniciais.

4. Conclusão

Conclui-se que, para o método aplicado, os resultados desta pesquisa indicam que nem sempre são obtidas boas aproximações para os valores reais dos dados que foram retirados.

Ao investigar o que poderia estar ocasionando isso, percebe-se que os dados onde a variabilidade é consideravelmente baixa o valor observado e o valor imputado são mais próximos. Porém, quando considerados dados com maior variabilidade nem sempre foram tão satisfatórios os resultados quanto a precisão.Portanto, para os dados analisados, a variabilidade dos mesmos influenciou de maneira negativa nos resultados obtidos ao aplicar o método de imputação.

Neste trabalho, dispomos dos dados iniciais para poder comparar, mas na realidade estes valores são desconhecidos, sendo impossível saber se o dado que se encontra em falta na realidade era mais alto ou menor que os demais.

Desta forma, é de grande importância que seja considerada a variabilidade dos dados em questão. Percebe-se que o fato de que os dados possuam uma baixa variabilidade é o que mais assegura resultados mais próximos dos reais.

Referências Bibliográficas

ALARCÓN, S. A.; DIAS, C. T. S. Imputação de dados em experimentos com interação genótipo por ambiente: uma aplicação a dados de algodão. Revista Brasileira de Biometria. São Paulo, v.27, n.1, p. 125 – 138, 2009.

BERGAMO, G. C.; DIAS, C. T. S.; KRZANOWSKI, W. J. Distribuition-freeMultipleImputation in na Interactionmatrixthrough singular valuedecomposition. SciênciaAgricola. Piracicaba, v.65, n.4, p. 422 – 427, 2008.

DIAS, A. J. R.; ALBIERI, S. Uso de imputação em pesquisas domiciliares. VIII Encontro Nacional de Estudos Populacionais. Anais. Volume 1: Informação Demográfica, Fecundidade, Demográfica Histórica.ABEP, São Paulo, p. 11 – 26, 1992.

ENDERS, C. K. AppliedMissing Data Analysis. Series Editor's Note by Todd D. Little, The Guilford Press, New York, London, 2010.

ENGELS, J.M.; DIEHR, P. Imputationofmissing longitudinal data: a comparisonofmethods. Journalof ClinicalEpidemiology, 2003.

GANDOLFI, M. Imputação Múltipla via algoritmo MICE e método IMLD. Dissertação (mestrado), Universidade Estadual de Maringá, Maringá, 2010.

LITTLE, R. J. A.; RUBIN, D. StatisticalAnalysiswithMissing Data. JournalofEducationalStatistics, v.16, n.2, p. 150 – 155, 1987

NUNES, L. N.; KLÜCK, M. M.; FACHEL, J. M. G. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. Caderno de Saúde Pública, Rio de Janeiro, v.25, n.2, p. 268 – 278, 2009.

NUNES, L. N.; KLÜCK, M. M.; FACHEL, J. M. G. Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica. Revista brasileira de Epidemiologia, v.13, n.4, p. 596 – 606, 2010.

SCHAFER, J.L.; GRAHAM, J.W. Missing data: Ourviewofthestateofthe art. PsychologicalMethods, v.7, n.2, p. 147 – 177, 2002.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 2017.

RStudio Team. RStudio: Integrated Development for R. RStudio, Inc., Boston, MA . URL http://www.rstudio.com/. 2009 - 2017.