CIÊNCIA e NATURA

UFSM

Ci. e Nat., Santa Maria v.42
Commemorative Edition: Statistic, e37, 2020
DOI: 10.5902/2179460X40446
ISSN 2179-460X

Statistics

Submissão 09/10/19 aprovação 26/11/19 publicação 30/11/20

Imputação de dados na análise de variância em experimentos no Delineamento Inteiramente Casualizado

Imputation of data in analysis of variance in experiments in the Completely Randomized Design

Elisandra Lucia Moro Stochero[I], Luciane Flores Jacobi[II], Alessandro Dal'Col Lúcio[III]

RESUMO

É comum a ocorrência de eventos não previstos no desenvolvimento de um experimento, principalmente em experimentos de campo, impossibilitando a mensuração das variáveis em determinadas parcelas experimentais. Pelo presente estudo, buscou-se verificar se há diferença no resultado do teste de Análise de Variância (ANOVA) para dados não balanceados no Delineamento Inteiramente Casualizado (DIC) quando realizada a inclusão de dados obtidos a partir da técnica de imputação de dados da Média Preditiva. Foram realizadas simulações de experimentos no DIC com 5 tratamentos e 10 repetições, gerando bancos de dados completos. De cada banco de dados, foi retirado 10% das parcelas e após aplicado o método de imputação, comparando os resultados da ANOVA em cada etapa. A imputação trouxe resultados aceitáveis, porém, não melhores do que os obtidos quando realizado o teste da ANOVA específico para dados desbalanceados.

Palavras-chave: Dados Faltantes; Métodos de Imputação; ANOVA; Análise de Dados Experimentais.

ABSTRACT

Unanticipated events often occur in the development of an experiment, especially in field experiments, often causing data loss. Through the present study, we sought to verify whether there is a difference in the result of the analysis of variance (ANOVA) test for unbalanced data in the Completely Randomized Design (DIC) when the inclusion of data obtained from the data imputation technique was performed of Predictive Average. Experiments were simulated in the DIC with 5 treatments and 10 repetitions, generating complete databases. From each database, 10% of the plots were removed and after the imputation method was applied, comparing the ANOVA results in each step. The imputation yielded acceptable results, but not better than those obtained when performing the specific ANOVA test for unbalanced data.

Keywords: Missing Data; Imputation Methods; ANOVA; Analysis of Experimental Data.

[I] Universidade Federal de Santa Maria, Santa Maria, Brasil. E-mail: elismoro2016@gmail.com.
[II] Universidade Federal de Santa Maria, Santa Maria, Brasil. E-mail: lucianefj8@gmail.com.
[III] Universidade Federal de Santa Maria, Santa Maria, Brasil. E-mail: adlucio@ufsm.br.

1 INTRODUÇÃO

A precisão dos resultados de um experimento, de acordo com Storck et al. (2010), está diretamente ligada ao processo de planejamento, condução e análise dos dados. Neste processo, os experimentos são conduzidos pelos seguintes princípios básicos, indispensáveis para que as conclusões obtidas se tornem válidas: princípio da repetição, casualização e controle local. Este sendo utilizado quando necessário, ou seja, nos casos onde as parcelas experimentais apresentam heterogeneidade entre si.

Também é importante definir o plano a ser utilizado, ou seja, a maneira como os tratamentos serão designados às unidades experimentais, o qual é conhecido como delineamento experimental. O mais simples destes é o Delineamento Inteiramente Casualizado (DIC), que utiliza apenas os princípios da repetição e casualização onde a alocação dos tratamentos às unidades experimentais é realizada inteiramente ao acaso, Banzatto e Kronka (2006). Sendo aconselhável utilizar esse delineamento quando as condições do ambiente são uniformes. O DIC apresenta algumas vantagens em relação a experimentos mais complexos e uma delas se refere ao número de repetições, que pode ser diferente entre os tratamentos. Nesta condição o experimento é classificado como desbalanceado.

Segundo Pimentel Gomes (1990), geralmente deve-se usar o mesmo número de repetições para todos os tratamentos em avaliação. No entanto, conforme o autor, alguns fatores podem levar à perda de parcelas, como a morte de animais ou plantas, entre outras causas. Também há a possibilidade de planejar experimentos desbalanceados utilizando o DIC, principalmente pela falta de parcelas homogêneas que permitam utilizar o mesmo número de repetições para todos os tratamentos. Assim a abordagem estatística para este tipo de situação não é a mesma adotada nos casos em que o experimento é balanceado.

A ANOVA é um procedimento estatístico utilizado para verificar a existência de diferença significativa entre as médias dos tratamentos. Para Padovani (2014), a lógica desta técnica consiste em desmembrar a variação total dos dados na variação devida aos tratamentos e na variação devida ao acaso (ou resíduo), para compará-las posteriormente.

O teste de hipóteses relativo a ANOVA tem como hipótese nula (H0) a não existência de efeito de tratamentos, ou seja, não existe diferença significativa entre as médias dos tratamentos e hipótese alternativa (H1) a existência de efeito de tratamentos, ou seja, pelo menos a média de um dos tratamentos difere significativamente das demais. A estatística de teste (F) é obtida através do Teste F (Fisher – Snedecor).

Para obter F é necessário determinar o valor referente ao quadrado médio dos tratamentos (QMTrat) e ao quadrado médio do resíduo (QMRes). A tabela geral da ANOVA é construída conforme a Tabela 1, e os valores de QMTrat e QMRes são determinados da seguinte forma:

$$SQ_{Tot} = \sum_{i=1}^{k}\sum_{j=1}^{r} y_{ij}^2 - C$$

Onde : $C = \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{r} y_{ij}\right)^2}{kr} = \frac{y_{..}^2}{kr}$

$$SQ_{Trat} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{r} y_{i.}^2}{r} - C$$

$$SQRes = SQTot - Sqtrat$$

Sendo y_{ij} a observação do tratamento i e repetição j, de um experimento com i = 1, ..., k tratamentos e j = 1, ..., r repetições.

Em que,

SQTot: soma de quadrados total; SQTrat: soma de quadrado dos tratamentos; SQRes: soma de quadrados do resíduo; GL: graus de liberdade.

Tabela 1– Tabela Geral da ANOVA de um DIC balanceado

Causa de Variação	GL	SQ	QM	F
Tratamentos	k-1	SQ_{Trat}	QM_{Trat}	QM_{Trat} / QM_{Res}
Resíduo	n-k	SQ_{Res}	QM_{Res}	
Total	n-1	SQ_{Tot}		

Em alguns estudos, conhecidos como não balanceados, o número de observações coletadas em cada tratamento não é o mesmo. O método de ANOVA também é válido para este caso, mas pequenas modificações devem ser feitas nas expressões das somas de quadrados. O que muda em relação ao caso balanceado é a soma de quadrados de tratamento e o valor de C. Assim,

n_i = número de observações do i-ésimo tratamento;

$N = \sum_{i=1}^{k} n_i$ = número total de observações

$$C = \frac{y_{..}^2}{N}$$

$$SQ_{Trat} = \sum_{i=1}^{k} \frac{y_{i.}^2}{n_i} - C$$

A soma de quadrado total e a soma de quadrado de resíduos seguem as equações do caso balanceado, assim como a tabela geral da ANOVA permanece igual à do DIC balanceado.

Para aplicar esta técnica é necessário, inicialmente, que sejam satisfeitas pressuposições básicas de normalidade e homogeneidade, as quais foram verificadas em todas as etapas através do teste de Shapiro Wilks e Bartlett, respectivamente.

Uma desvantagem é que um experimento, quando desbalanceado, pode levar a estimativas altas da variância residual. De acordo com Iemma (1995), quando os dados são provenientes de experimentos desbalanceados, a interpretação das hipóteses testadas também se torna mais trabalhosa, o que leva muitas vezes os pesquisadores a utilizarem de maneira inadequada os softwares estatísticos. Além disso, para Shafer e Graham (2002), com ou sem dados ausentes, o objetivo de um procedimento estatístico deve ser o de fazer inferências válidas e eficientes sobre uma população de interesse e não apenas estimar, prever ou recuperar observações ausentes.

Alguns métodos foram desenvolvidos com o objetivo de determinar estimativas para dados em falta. Inicialmente as estimativas de dados faltantes eram obtidas via procedimentos simples como: a média, mediana da variável, por interpolação ou regressão linear. Estes métodos são conhecidos como Métodos de Imputação Simples ou Única (IU), conforme Engels e Diehr (2003), por serem realizados apenas uma vez. Posteriormente foram desenvolvidos métodos baseados na Imputação Múltipla (IM), propostos inicialmente por D. B. Rubin na década de 1970 (ZHANG, 2003).

O Método da Média Preditiva (PMM – Predictive Mean Matching), descrito por Nunes et al. (2009), é um método de Imputação Múltipla e foi o utilizado por ser indicado quando nenhum critério determinado previamente é considerado para a

exclusão das observações e consequentemente o mecanismo gerador dos dados faltantes é o de perdas completamente ao acaso (MCAR – Missing Completely at Random). Nunes et al. (2009) descrevem o método PMM como um método de IM que parte do mesmo princípio do modelo ajustado sob o paradigma Bayesiano. Considera-se $Y = \alpha + \beta X$, com $Y \sim N(X\beta; I\sigma^2)$, sendo a variável resposta Y a variável a ser imputada. As estimativas dos parâmetros do modelo são determinadas por meio de uma distribuição a posteriori própria. São calculados os valores preditos para os y observados e y faltantes. O valor a ser imputado é o valor observado mais próximo do valor predito. A variabilidade entre imputações é gerada através dos passos usados para encontrar a estimativa de β e α.

Nunes et al. (2009) relatam que houve um aumento significativo dos estudos sobre imputação múltipla desde o início da década de 90, principalmente com o avanço computacional. A vantagem de sua aplicação se dá por este método considerar a variabilidade entre as imputações nos resultados. Nas ciências agrárias é possível encontrar trabalhos envolvendo métodos de imputação, embora o número seja menor em relação a outras áreas. Alarcon e Dias (2009) avaliaram a conveniência de definir o número de componentes multiplicativos dos modelos de efeitos principais aditivos com interação multiplicativa (AMMI) em experimentos de interações genótipo x ambiente de algodão com dados imputados ou desbalanceados, por exemplo. A eficiência da imputação múltipla foi estudada por Silva (2012) em experimentos multiambientais, e Oliveira (2012) avaliou seis métodos de imputação de médias faltantes em experimentos conjuntos incompletos de café conilon.

Geralmente as pesquisas buscam verificar qualidade, efeito ou realizar comparações entre métodos de imputação avaliando o modelo obtido, ou até mesmo propor novos métodos como Bergamo et al. (2008). Mesmo sendo encontrados estudos envolvendo a ANOVA, como a pesquisa desenvolvida por Ginkel e Kroonenberg (2014) que propõe uma reformulação do modelo ANOVA e do modelo de regressão, não se tem especificamente o objetivo de verificar o resultado do teste de ANOVA que de fato é aplicado em experimentos agrícolas.

A preocupação com a análise de experimentos, quando esses possuem caselas vazias já ocorre a muito tempo. Iemma (1995) alertava de que as hipóteses testadas em presença de caselas vazias deveriam ser estudadas com cautela por parte do pesquisador, pois pode ocorrer que o pesquisador julgue testar uma certa Hipótese H0(1) quando na verdade a estrutura de desbalanceamento dos dados pode induzir ao teste de uma outra hipótese H0(2), sem qualquer sentido prático para ele.

Para auxiliar os pesquisadores que buscam resultados com uma maior precisão e perderam parcelas em seus experimentos ou que não conseguiram unidades experimentais suficientes é que este trabalho traz como problema de pesquisa a seguinte questão: em casos de dados não balanceados, é preferível realizar a ANOVA para o caso desbalanceado ou realizar a imputação de dados faltantes e proceder a ANOVA para dados balanceados?

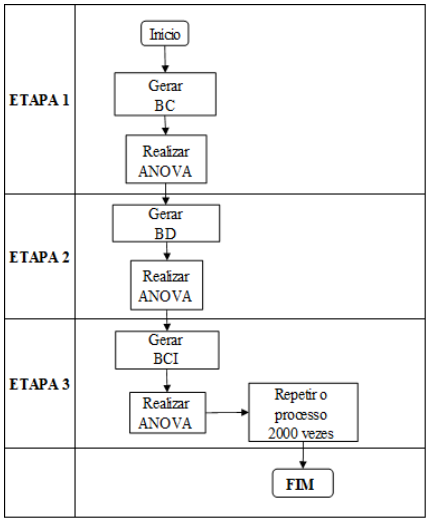
Portanto, o objetivo foi verificar se há diferença no resultado do teste da ANOVA para dados não balanceados se comparados ao resultado da ANOVA resultante de uma base com imputação de dados faltantes.

2 MATERIAL E MÉTODOS

O processo para o desenvolvimento da pesquisa foi dividido em três etapas (Figura 1). Os bancos de dados completos (BC) são simulações de um experimento no DIC com 5 tratamentos e 10 repetições. Realizou-se a ANOVA e foram consideradas 1000 séries com diferença significativa entre tratamentos e 1000 séries em que a diferença entre os tratamentos não foi significativa ao nível 5% de significância.

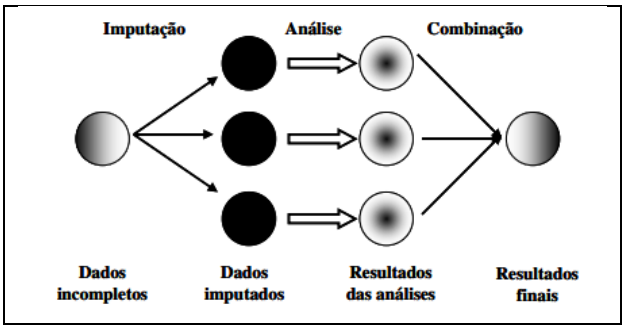
Os bancos de dados incompletos, ou seja, bancos de dados desbalanceados (BD), foram gerados a partir dos BC após a retirada de 10% dos dados, de forma aleatória, não sendo fixado previamente número de exclusões em cada tratamento. A aplicação do método de imputação foi realizada nos BD a fim de determinar as estimativas a serem incluídas nas caselas vazias, gerando bancos de dados completos com imputação (BCI).

Figura 1 – Etapas do processo



Para a imputação dos dados foi utilizada uma técnica de Imputação Múltipla que, conforme Nunes et al. (2009), foi proposta por Rubin para determinação da estimativa do valor a ser inserido no “espaço vazio”, e pode ser representado na forma gráfica como na Figura 2. A Imputação Múltipla constitui-se de três passos, no primeiro são obtidos m bancos de dados completos a partir da aplicação da técnica de imputação adequada, no segundo os m bancos de dados são analisados separadamente e por último é feita a combinação de maneira simples das respectivas estimativas de cada dado faltante, o valor será o valor final a ser imputado.

Figura 2 – Representação gráfica da Imputação Múltipla



As análises foram conduzidas a partir da implementaçãono software R Core Team (2017) e no Rstudio (2009 – 2017), pelo pacote MICE (Multivariate Imputation by Chained Equations). O número padrão de múltiplas imputações realizadas são m= 5, para dados numéricos o método de imputação padrão é o PMM.

3 RESULTADOS E DISCUSSÕES

A frequência dos resultados obtidos no teste da ANOVA em todas as etapas estão dispostos na Tabela 2, do total dos resultados, houve concordância entre Bancos de Dados Completos, Bancos de Dados Desbalanceados e Bancos de Dados Completos com Imputação em 88,05% das vezes. Nos quais, 49,45% rejeitou-se a hipótese nula e 38,6% não se rejeitou a hipótese nula. Houve concordância entre Banco de dados Completos e Bancos de Dados Desbalanceados em 6,9% dos resultados, e entre Bancos de Dados Completos e Bancos de Dados Completos com Imputação 0,8%.

Também, pode-se dizer que, das 2000 simulações, para os dados nos quais houve rejeição da hipótese nula no BC, houve concordância apenas com o BD em 0,05% dos resultados e com o BCI em apenas 0,3% das vezes. Para os dados nos quais não houve rejeição da hipótese nula no BC, houve concordância apenas com o BD em 6,85% dos resultados e com o BCI em 0,5% dos resultados.

No resultado final da ANOVA, a aplicação do método de imputação não se mostrou superior quando avaliado o número de resultados incorretos. Em contrapartida, houve 88,85% de acertos quando aplicado. A taxa de erro foi maior quando não se rejeitou a hipótese nula no BC e foi satisfatória quando a hipótese nula foi rejeitada no BC.

Tabela 2 - Frequência de resultados da ANOVA

Possíveis Resultados da ANOVA			Frequência	(%)
BC	BD	BCI		
Rejeita H_0	Rejeita H_0	Rejeita H_0	989	49,45
Rejeita H_0	Rejeita H_0	Aceita H_0	1	0,05
Rejeita H_0	Aceita H_0	Rejeita H_0	6	0,30
Rejeita H_0	Aceita H_0	Aceita H_0	4	0,20
Aceita H_0	Rejeita H_0	Rejeita H_0	81	4,05
Aceita H_0	Rejeita H_0	Aceita H_0	10	0,50
Aceita H_0	Aceita H_0	Rejeita H_0	137	6,85
Aceita H_0	Aceita H_0	Aceita H_0	772	38,6
TOTAL			2000	100

O erro Tipo I, que ocorre quando a hipótese nula é verdadeira e é rejeitada, foi identificado em 4,10% dos casos quando comparados BC e BD e em 10,9 % quando comparados BC e BCI, considerando correta a aceitação e rejeição da hipótese nula no BC. Logo, o erro tipo I ocorreu com uma maior frequência quando aplicado o método de imputação.

Nos resultados encontrados nesta pesquisa não houve uma diferença considerável quando aplicado método de imputação (PMM) em relação aos resultados obtidos com dados desbalanceados. Para Ginkel e Kroonenberg (2014), a realização de imputação (método PMM) não foi satisfatória quando verificados os graus de liberdade do erro, estes são sensíveis em relação ao número de imputações. Quando foram imputados 100 dados, o grau de liberdade do erro foi maior do que quando excluídas as informações relacionadas aos dados faltantes para obter um banco de dados completo.

Em um estudo comparativo do comportamento de diferentes métodos de imputação, Santos e Almeida (2014), utilizando para exemplificação duas bases de dados do repositório UCI, concluíram que todos os métodos obtiveram taxas de erro satisfatórias e equivalentes.

No presente estudo foram excluídos aleatoriamente 10% dos dados. Se o número de exclusões foi semelhante em cada tratamento talvez não tenha causado um grande viés, porém se o número foi maior em apenas um tratamento em relação aos demais, pode ter causado uma taxa de erro maior.

Para Nunes et al. (2009), o método PMM, o mesmo aplicado nesta pesquisa, foi eficiente. Para os dados que foram trabalhados, dados epidemiológicos, é necessário que o banco de dados seja completo para que seja viável a realização da análise estatística. Caso não seja aplicado um método de imputação a alternativa é a exclusão de todos os dados do indivíduo em que ao menos uma das observações esteja em falta. Esta exclusão dos dados acaba tornando a amostra pequena. Logo, trabalhar com dados imputados é a melhor opção, ao mesmo tempo que retorna estimativas com uma boa precisão, o tamanho da amostra não é afetado de forma negativa.

Ao contrário dos dados avaliados no estudo de Nunes et al. (2009), nesta pesquisa foram realizadas simulações do delineamento inteiramente casualizado para a obtenção dos bancos de dados iniciais completos, situação em que é possível trabalhar com um banco de dados mesmo que incompleto, apenas aplicando algumas alterações para a obtenção do quadro da ANOVA.

Desta forma, percebe-se que a indicação da aplicação de um método de imputação está inteiramente ligada as características dos dados a serem trabalhados, além de outros pontos essenciais a serem observados, como a causa da perda das observações e o padrão dos dados ausentes, por exemplo.

Portanto, ao se verificar especificamente o resultado do teste da ANOVA, ou seja, se a hipótese nula era aceita ou não, além de haver uma porcentagem pequena de discordância entre os resultados para dados não balanceados e os resultados dos dados onde foi realizada a imputação de dados faltantes, os resultados obtidos foram melhores para os BD em relação aos BCI. Embora a taxa de erro tenha sido menor nos Bancos Completos por Imputação do que nos Bancos Desbalanceados quando se rejeitou a hipótese nula no Banco Completo, no total a taxa de erro foi maior para os BCI.

Assim, aplicar o método de imputação não trouxe resultados satisfatórios. Em um banco de dados de um experimento no DIC, delineamento adotado para as simulações desta pesquisa, onde é possível realizar a análise estatística mesmo com o banco de dados desbalanceado, aplicar o método de imputação é mais trabalhoso e não garante uma probabilidade significativamente maior de não obter o resultado incorreto no teste da ANOVA.

No entanto, outros pontos que não foram avaliados podem ter sido beneficiados com a aplicação do método. Por exemplo, a precisão dos resultados não foi avaliada, a qual, segundo Cargnelutti Filho e Storck (2007), é importante para a validação dos resultados obtidos em um experimento. As estatísticas obtidas, conforme Lúcio (1997), através do cálculo do coeficiente de variação (CV), coeficiente de precisão (CP) e diferença mínima significativa (DMS) como um meio de descrever a qualidade de um

ensaio que, por sua vez, é um dos fatores que indicam a confiabilidade dos resultados obtidos.

Em estudos futuros, será avaliada a precisão experimental, a fim de verificar se a aplicação de um método de imputação leva a resultados mais precisos, com um menor viés.

4 CONCLUSÃO

De acordo com os resultados obtidos, os mesmos foram melhores para os Bancos de Dados Desbalanceados em relação aos Bancos de Dados Completos com Imputação. Embora a taxa de erro tenha sido menor nos Bancos de Dados Completos com Imputação do que nos Bancos de dados Desbalanceados quando se rejeitou a hipótese nula nos Bancos de Dados Completos, no total a taxa de erro foi maior para os Bancos de Dados Completos com Imputação. Neste caso, além de mais simples quando não aplicado o método de imputação, a taxa de erro é menor no resultado da ANOVA.

No entanto, é necessário ter cautela na generalização dos resultados deste trabalho, já que eles foram obtidos aplicando um único método, outros métodos e outras abordagens poderiam trazer resultados mais satisfatórios.

Portanto, conclui-se que não há diferença no resultado do teste da ANOVA para dados não balanceados se comparados ao resultado da ANOVA resultante de uma base com dados imputados pelo Método de Imputação da Média Preditiva, e que utilizar a imputação de dados para completar experimentos no Delineamento Inteiramente Casualizado não implicará em uma menor taxa de erro tipo I.

REFERÊNCIAS

ALARCÓN SA, DIAS CTS. Imputação de dados em experimentos com interação genótipo por ambiente: uma aplicação a dados de algodão. **Revista Brasileira de Biometria**. 2009;27(1):125-138.

BERGAMO GC, DIAS CTS, KRZANOWSKI WJ. Distribuition-free Multiple Imputation in Interaction matrix through singular value decomposition. **Sciencia Agricola**. Piracicaba. 2008;65(4):422-427.

CARGNELUTTI FILHO A, STORCK L. Estatísticas de avaliação da precisão experimental em ensaios de cultivares de milho. **Pesquisa Agropecuária Brasileira**. Brasília. 2007;42(1):17-24.

ENGELS JM, DIEHR P. Imputation of missing longitudinal data: a comparison of methods. **Journal of Clinical Epidemiology**. 2003; 56:968-976.

GINKEL JRV, KROONENBERG PM. Analysis of Variance of Multiply Imputed Data. **Multivariate Behav Res**. 2014; 49(1): 78–91.

IEMMA AF. Que hipóteses estatísticas testamos através do SAS em presença de caselas vazias. **Scientia Agricola**, Piracicaba. 1995;52(2):210-220.

LÚCIO AD, STORCK L. Parâmetros da precisão experimental das principais culturas anuais do Estado do Rio Grande do Sul. Santa Maria. **Ciência Rural**. 1997;27(3).

NUNES LN et al. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. **Caderno Saúde Pública**. Rio de Janeiro. 2009;25(2):268-278.

OLIVEIRA RLR. **Imputação de médias para análise de estabilidade e adaptabilidade em experimentos conjuntos incompletos: uma aplicação em café Conilon** [dissertation]. Viçosa: Universidade Federal de Viçosa (UFV); 2012. 61p.

PADOVANI CR. Delineamento de Experimentos. São Paulo: **Cultura Acadêmica: Universidade Estadual Paulista**, Pró-Reitoria de Graduação, 2014.

PIMENTEL GOMES F. **Curso de estatística experimental**. 13.ed. Piracicaba, São Paulo: Nobel 1990.

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.

SANTOS LI, ALMEIDA MFF. Tratamento de Dados Faltantes: Uma avaliação de métodos de imputação de dados a partir do desempenho de um classificador Bayesiano. **Anais do XIII Encontro Mineiro de Estatística**, Diamantina: UFVJM, 2014.

SCHAFER JL, GRAHAM JW. Missing data: Our view of the state of the art. **Psychological Methods**. 2002;7(2):147-177.

SILVA MJC. **Imputação múltipla: comparação e eficiência em experimentos multiambientais** [dissertation]. Piracicaba: Escola Superior de Agricultura Luiz Queiroz (USP); 2012. 125p.

STORCK L. et al. Avaliação da precisão experimental em ensaios de competição de cultivares de soja. Lavras. **Ciência Agropecuária**. 2010;34(3):572-578.

ZHANG P. Multiple Imputation: Theory and Method. **International Statistical Review**. 2003;71(3):581-592.