Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Modelo de efeitos principais aditivos e interação multiplicativa generalizado (GAMMI) para imputações de dados em experimentos multiambientais

Pedro Marinho Amoedo

Tese apresentada para obtenção do título de Doutor em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2021

Pedro Marinho Amoedo
Bacharel em Estatística

Modelo de efeitos principais aditivos e interação multiplicativa generalizado (GAMMI) para imputações de dados em experimentos multiambientais

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Profa. Dra. **SÔNIA MARIA DE STEFANO PIEDADE**

Tese apresentada para obtenção do título de Doutor em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2021

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Amoedo, Pedro Marinho

Modelo de efeitos principais aditivos e interação multiplicativa generalizado (GAMMI) para imputações de dados em experimentos multiambientais / Pedro Marinho Amoedo. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2021 .

45 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1.Experimentos multiambientais 2.EM-AMMI generalizado 3.Imputação de dados 4.Imputação multipla GAMMI . I. Título.

DEDICATÓRIA

Dedico a meus pais;
Rossi Paes de Andrade Amoêdo e Iolanda Marinho Amoêdo

AGRADECIMENTOS

Agradeço à Deus pelo caminhar por este mundo maravilhoso, junto à família, amigos, colegas e tantos outros que encontramos na estrada.

À minha grande mãe Maria por sempre se fazer presente em minha vida, mesmo quando eu não a via, sei que estava lá.

Ao Profesor Carlos Tadeu dos Santos Dias, pela amizada, apoio e por me conduzir para o desenvolvimento desta tese.

À professora Sônia Maria de Stefano Piedade, pela sua orientação e pessoa maravilhosa que é.

Aos meus pais, Rossi e Iolanda, por todos os ensinamentos e por tudo a mim dedicado, muitas vezes, deixaram de realizar seus desejos em detrimento dos meus.

À minha esposa Fabiane Santos e meus filhos, Perseu e Rafael, o verdadeiro grande baú do pirata.

À minha família como um todo, que de forma muito especial me incentivou com muito amor e carinho, dando apoio e suporte a cada dia.

A todos os Professores e funcionários do Programa de Pos-graduação em Estadística e Experimentação Agronômica do Departamento de Ciências Exatas da ESALQ/USP, pela ajuda, ensinamento e apoio.

Aos amigos e colegas de curso que sempre estiverem dando apoio, estes vivenciaram comigo problemas, estudos e acima de tudo me ajudaram com suas partilhas. Em especial, Djair Durand, Marcello Neiva e Ana Garcez.

A todos que estiveram presentes em minha vida, que contribuiram de alguma forma para a minha formação. O meu muito obrigado!

EPÍGRAFE

Não vejo mais o teu brilho,
mas não vi o que teus olhos viram,
senti saudades, falei com Deus por ti,
aos poucos fostes desistindo,
sinto que fostes esquecendo,
da bondade, do amor, o amor! aos poucos foi morrendo,
ja no brilho da escuridão, não vejo mais a tua face,
não vi o que teus olhos viram,
mas meu brilho!
meu brilho, a bondade o amor, não vão acabar.

Fabi A. Santos

SUMÁRIO

RESUMO

Modelo de efeitos principais aditivos e interação multiplicativa generalizado (GAMMI) para imputações de dados em experimentos multiambientais

As análises mediante abordagem de modelo de efeitos principais aditivos e interação multiplicativa exigem que as matrizes de dados, provenientes de experimentos multiambientais, sejam completas, o que em geral não ocorre. Uma excelente alternativa para contornar o problema das ausências para posterior análise são os métodos de imputação de dados. Esta tese tem por objetivo desenvolver duas estratégias de imputação de dados para experimentos multiambientais por meio de modelos GAMMI. Um algoritmo de imputação simples e outro de imputação múltipla. O algoritmo de imputação simples foi desenvolvido a partir da combinação do método de imputação EM-AMMI com o modelo de efeitos principais aditivos e interação multiplicativa generalizado (GAMMI). O segundo algoritmo, de imputação múltipla, é uma extensão do algoritmo EM-AMMI generalizado (EM-GAMMI), desenvolvido a partir da proposta de imputação com uso dos resíduos simples do modelo de regressão linear. Para determinar o desempenho dos algoritmos de imputação, foram realizadas simulações de retiradas aleatória de valores em diferentes porcentagens em que foram tomados dois conjuntos de dados reais completos como base. Os dois conjuntos possuem dimensões, um 4×5 e o outro de dimensão 19×6. Então, a qualidade das imputações foi avaliada por meio das distribuições da estatística raiz normalizada do erro quadrático médio (NRMSE) e da estatística geral de acurácia (Tacc), obtidas a partir dos valores imputados em cada um dos níveis de retiradas. Concluiu-se que os novos métodos constituem ferramentas eficientes como técnicas de imputação de dados, útil para contornar o problema das ausências em experimento multiambientais.

Palavras-chave: Experimentos multiambientais, EM-AMMI generalizado, Dados ausentes, Imputação multipla GAMMI

ABSTRACT

Generalized additive main effect and multiplicative interaction (GAMMI) model for data imputations in multi-environmental experiments

Analyzes using a model of main additive effects and multiplicative interaction require that the data matrices, coming from multi-environmental experiments, be complete, which in general does not occur. An excellent alternative to circumvent the problem of absences for further analysis are the methods of data imputation. This thesis aims to develop two data imputation strategies for multi-environmental experiments using GAMMI models. A simple imputation algorithm and a multiple imputation algorithm. The simple imputation algorithm was developed from the combination of the EM-AMMI imputation method with the main additive effects and generalized multiplicative interaction (GAMMI). The second algorithm, of multiple imputation, is an extension of the generalized EM-AMMI algorithm (EM-GAMMI), developed from the imputation proposal using simple residues of the linear regression model. To determine the performance of the imputation algorithms, simulations were carried out with random remove values with different percentages in which two real data sets were taken as a reference.The two sets had dimensions, one 4 x 5 and the other 19 x 6. Then, the quality of the imputations were assessed through the distribution of the normalized root statistic of the mean square error (NRMSE) and the statistic overall accuracy statistic (Tacc), obtained from the imputed values in each of the levels of simulated withdrawals. It was concluded that the new methods are efficient tools such as data imputation techniques, being useful to circumvent the problems of missing samples in a multi-environmental experiment.

Keywords: Generalized EM-AMMI, Missing data, Multi-environmental experiments, Multiple imputation GAMMI

1 INTRODUÇÃO

Estudos experimentais, por mais bem planejados que sejam, estão sujeitos às condições humanas de acompanhamento e casualidades do meio em que se encontram, tendendo em determinada etapa, vir à apresentar alguma falha, seja por falta de controle, desatenção do controlador, medições erradas pelo responsável da coleta, transferências duvidosas para arquivos de dados, fatores climáticos, pragas e entre outros (Schafer e Graham, 2002; Bergamo, 2007; Arciniegas-Alarcón, 2015). Algumas destas imposições podem pruduzir perdas de dados e consequentemente, causando no experimento o seu desbalanceamento.

A ocorrêcia do desbalanceamento em experimentos levam pesquisadores a recorrerem a procedimentos teóricos específicos, pois a maioria dos procedimentos estatísticos de análise exigem a completude dos dados, não sendo possível serem analisados diretamente por metodologias clássicas. A situação ideal de análise sobre o experimento seria repeti-lo e obter novos valores para as observações ausentes, entretanto, esta solução é inviável em muitos casos, pois ocorre quase sempre a escassez de material e de tempo (Arciniegas-Alarcón, 2015). Dentre as alternativas, mais usuais, para contornar o problema das ausências de dados para posterior análise, estão aquelas apresentadas por Dodge (1985); Little e Rubin (2002). Dodge (1985) apresentou procedimentos teóricos para fazer as análises baseadas exclusivamente nos dados presentes, (os valores observados), enquanto Little e Rubin (2002) apresentaram métodos que predizem os valores ausentes por estimativas plausíveis, métodos de imputação de dados.

Os métodos de imputação de dados são excelentes opções para contornar o problema das ausências, viabilizando posteriores análises. Os primeiros métodos envolviam procedimentos relativamente simples, como a imputação pela média, atribuída a Wilks (1932) mas, foi a partir da técnica de imputação múltipla de Rubin (1987) que os métodos de imputação tornaram-se mais conhecidos e usuais. Métodos específicos para experimentos multiambientais como junção de outras técnicas foram desenvolvidos e, uma das abordagens com bastante êxito e bem aceita são os métodos de imputação baseados na decomposição em valores singulares (DVS). Gauch e Zobel (1990) apresentam o algoritmo EM-AMMI, introduzindo no algoritmo EM (Esperança-Maximização) o modelo de efeitos principais aditivos e interação multiplicativa (AMMI), Bergamo et al. (2008) apresentaram o método de imputação múltipla livre de distribuição (IMLD), um método livre de suposição sobre a distribuição dos dados ou restrição quanto ao padrão e mecanismo de ausência, Yan (2013) apresenta o método Biplot, Perry (2009) apresenta o EM+DVS, a junção do algoritmo EM (Esperança-Maximização) com a decomposição em valores singulares (DVS) de uma matriz e entre outros.

Para análise de dados balanceados em experimentos com interações genótipo por ambiente, é de ampla aceitação no meio científico que o modelo de efeitos principais aditivos e interação multiplicativa (AMMI) é um dos melhores procedimentos, explora melhor as informações contidas nos dados do que a ANOVA tradicional (Duarte e Vencovsky, 1999), a ocorrência da interação genótipo por ambiente exclui o uso de modelos básicos de interações em que se tem somente efeitos principais aditivos de genótipo e ambiente (Mandel, 1971). No entanto, para seu uso, há a exigência que a matriz de dados seja completa e que a variável de resposta seja normal, independente e identicamente distribuída (Hadi et al., 2010)

O modelo de efeitos principais aditivos e interação multiplicativa generalizado (GAMMI) é uma das melhores alternativas quando da impossibilidade ao uso do AMMI, amplia o leque de modelagem para os experimentos com interação genótipo por ambiente, possibilitando analisar dados com resposta Poisson, Binomial, Normal, entre outras. O procedimento permite realizar análises sem recorrer as populares transformações de dados para atender às pressuposições da análise. As estimativas dos parâmetros são obtidas por processo iterativo alternando regressões generalizadas de linhas e colunas. A determinação do número de eixos ou termos multiplicativos a serem retidos pelo modelo podem ser obtidos por generalização de teste da metodologia AMMI, teste da razão de verossimilhança, teste F ou teste de Gollob (Van Eeuwijk, 1995).

Levando em conta o exposto acima, o objetivo primário aqui é apresentar dois métodos de imputação de dados para experimentos multiambientais, um de imputação simples e outro de imputação múltipla, ambos tendo por base o algoritmo EM-AMMI e, com suporte na teoria dos modelos lineares generalizados. Deste modo, busca-se contribuir para o meio científico com novas ferramentas para predição das ausências.

Assim, o trabalho justifica-se pela adoção de novas ferramentas de imputação de dados mediante uso dos modelos GAMMIs, visando alternativas à análise do modelo de efeitos principais aditivos e interação multiplicativa (AMMI), seja por limitações do enfoque tradicional frente à violações das suposições do modelo aditivo ou por limitações junto ao cálculo da DVS, pois situações em que o experimento é desbalanceado, não é possível calcular a DVS da matriz, inviabiliza posteriores análises.

A tese está composta de dois capítulos e escrita de modo a se abordar cada capítulo independentemente. Os assuntos apresentados são descritos a seguir. No capítulo 2 é apresentado o estudo sobre um método de imputação simples para experimentos com interação genótipo por ambiente, um novo processo de imputação, uma junção do algoritmo EM com o modelo de efeitos principais aditivos e interação multiplicativa generalizado (GAMMI). No capítulo 3 se propõe uma nova metodologia de imputação múltipla, desenvolvida a partir do algoritmo EM-GAMMI, para imputação de ausências em tabelas incompletas de dupla entrada.

Referências

Arciniegas-Alarcón, S. (2015). *Imputação de dados em experimentos multiambientais: novos algoritmos utilizando a decomposição por valores singulares*. PhD thesis, Universidade de São Paulo.

Bergamo, G. C. (2007). *Imputaçao múltipla livre de distribuiçao utilizando a decomposiçao por valor singular em matriz de interaçao*. PhD thesis, Universidade de São Paulo.

Bergamo, G. C., Dias, C. T. d. S., e Krzanowski, W. J. (2008). Distribution-free multiple imputation in an interaction matrix through singular value decomposition. *Scientia Agricola*, 65(4):422–427.

Dodge, Y. (1985). Analysis of experiments with missing data. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1985*.

Duarte, J. e Vencovsky, R. (1999). Interação genótipos x ambientes: uma introdução à análise ammi. *Ribeirão Preto: Sociedade Brasileira de Genética*.

Gauch, H. e Zobel, R. W. (1990). Imputing missing yield trial data. *Theoretical and Applied Genetics*, 79(6):753–761.

Hadi, A. F., Mattjik, A., e Sumertajaya, I. (2010). Generalized ammi models for assessing the endurance of soybean to leaf pest. *Jurnal Ilmu Dasar*, 11(2):151–159.

Little, R. J. e Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: John Wiley. 381p .

Mandel, J. (1971). A new analysis of variance model for non-additive data. *Technometrics*, 13(1):1–18.

Perry, P. O. (2009). *Cross-validation for unsupervised learning*. PhD thesis, Stanford University, 153p.

Rubin, D. B. (1987). *Multiple imputation for survey nonresponse*. New York: John Wiley & Sons. 320 p.

Schafer, J. L. e Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, 7(2):147–177.

Van Eeuwijk, F. A. (1995). Multiplicative interaction in generalized linear models. *Biometrics*, pages 1017–1032.

Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *The Annals of Mathematical Statistics*, 3(3):163–195.

Yan, W. (2013). Biplot analysis of incomplete two-way data. *Crop Science*, 53(1):48–57.

2 IMPUTAÇÃO DE VALORES AUSENTES POR MEIO DE MODELO DE EFEITOS PRINCIPAIS ADITIVOS E INTERAÇÃO MULTIPLICATIVA GENERALIZADO

Resumo

O objetivo deste trabalho foi propor um método de imputação utilizando a mistura do modelo de efeitos principais aditivos e interação multiplicativa generalizado (GAMMI) e o algoritmo EM, para experimentos de interação genótipo por ambiente. Desta forma, modificações no algoritmo de imputação EM-AMMI foram feitas, de modo a comportar o ajuste do modelo multiplicativo em termos de modelo linear generalizado, afim de se realizar imputações. Estudos de simulações foram realizados com base em dois conjuntos de dados reais completos, ambos obtidos de experimentos de interação genótipo por ambiente, e a eficácia do método foi testada por meio das simulações feitas com retiradas aleatórias de 10%, 20%, 30% e até 40% dos dados. A qualidade das imputações obtidas a partir da junção do modelo de efeitos principais aditivos e interação multiplicativa generalizado com o algoritmo EM, em que se considerou o mecanismo de ausência aleatória - *MAR (Missing at Random)*, foi avaliada utilizando a raiz normalizada do erro quadrático médio (NRMSE) entre os dados originais de genótipo por ambiente e os dados imputados correspondentes. Também foram realizados testes de Kruskal-Wallis e testes de Wilcoxon afim de identificar possíveis diferenças entre os procedimentos de imputações. Os melhores resultados obtidos para os algoritmos de imputações, em termos da NRMSE, foram alcançados pelo algoritmo proposto EM-GAMMI. No entanto, não se constatou evidência estatística de diferença em relação ao algoritmo EM-AMMI tomado como padrão de comparação.
Palavras-chave: Imputação de dados; Interação genótipo por ambiente; Modelo GAMMI; Valores ausentes.

2.1 Introdução

É comum, nos diversos tipos de experimentos científicos, por mais bem planejados que sejam, a ocorrência de dados ausentes *(missing data)*, seja pelas condições de acompanhamento, erros humanos ou pelas casualidades naturais do meio em que se encontram, e que quase sempre, geram problemas e dificuldades de análises. A escolha de métodos estatísticos eficientes para tratar dados com ausências é bastante delicada, pois a maioria das técnicas de análise exigem a completude dos dados. Métodos inadequados para tratar ausências podem causar mais problemas do que resolvê-los, distorcem estimativas, erros padrões e inferências (Little e Rubin, 2002; Yan, 2013). Uma excelente alternativa, para contornar o problema das ausências e posterior análise dos dados, são os denominados métodos de imputação (Rubin, 1987; Schafer, 1999; Zhang, 2003). Os metodos de imputação concentram-se em preencher as ausências mediante estimativas plausíveis dos valores ausentes, as imputações, evitando recorrer ao desenvolvimento de um quadro teórico específico para análise (Robins et al., 1994; Little e Rubin, 2002)

Os primeiros métodos de imputação envolviam técnicas relativamente simples, estimavam (substituiam) os dados ausentes pela média, pela mediana, por interpolação ou por regressão linear (Dias e Albieri, 2016), mas a partir da técnica de imputação múltipla (IM) de Rubin (1987) os procedimentos de imputações tornaram-se mais conhecidos e recorrentes. Novas propostas, em diferentes áreas, foram aprimoradas com suporte da IM, e no contexto multiambiental, métodos específicos foram desenvolvidos, melhorados a partir de extensões de métodos existentes ou de suas misturas (Gauch e Zobel, 1990; Bergamo et al., 2008; Arciniegas-Alarcón et al., 2014).

Um trabalho bem aceito para estes experimentos foi o desenvolvido por Gauch e Zobel (1990), os autores implementaram a mistura do algoritmo EM-AMMI, introduzindo no algoritmo EM (Esperança-Maximização) o modelo de efeitos principais aditivos e interação multiplicativa (AMMI) para realizar imputações em um ensaio de soja. Neste método, conforme mostram os estudos de Piepho (1995);

Arciniegas-Alarcón e Dias (2009); Paderewski e Rodrigues (2014), os melhores resultados para imputação são alcançados pelo algoritmo com a inclusão de poucos termos multiplicativos no modelo AMMI.

Um outro método, de igual ou maior destaque, é o algoritmo EM+DVS, apresentado por Perry (2009). O EM+DVS usa mistura de dois procedimentos para realizar imputação. O método combina o algoritmo EM (Esperança-Maximização) com a decomposição em valores singulares (DVS) de uma matriz. Na etapa inicial, os valores ausentes são substituídos por valores arbitrários, obtendo uma matriz completada, em seguida, é calculada interativamente a DVS da matriz completada. No final do processo, quando as iterações alcançarem estabilidade, uma matriz contendo as imputações para as observações ausentes é obtida. Segundo Perry (2009), globalmente o cálculo da melhor aproximação de posto k é inviável computacionalmente, mas o algoritmo se concentra em soluções mais locais, k viável. Arciniegas-Alarcón (2015), em um estudo comparativo entre métodos que usam a DVS, foi animoso em concluir que o algoritmo EM+DVS apresentou melhores resultados que outros métodos de imputações analisados.

Bergamo et al. (2008) apresentaram o método de imputação múltipla livre de distribuição (IMLD), um método sem qualquer restrição quanto ao padrão e mecanismo de ausência de dados e livre de suposição sobre a distribuição ou estrutura dos dados. O método estima os valores a serem imputados por meio de uma modificação no procedimento de imputação simples desenvolvido por Krzanowski (1988). Ainda, dentre outros métodos que usam a DVS, destaque para a imputação Biplot e o GabrielEigen. O método de imputação Biplot, foi descrito por Yan (2013) para realizar a análise biplot, mas foram García-Peña et al. (2014) que o denominaram de imputação Biplot. Na imputação Biplot, as ausências iniciais são substituidas por valores arbitrários, obtendo uma matriz completada, em seguida, é calculado a DVS utilizando dois componentes apenas. No GabrielEigen, proposto por Arciniegas-Alarcón et al. (2010), as ausências iniciais são preenchidas por valores arbitrários, em seguida, as imputações são refinadas por meio de um esquema iterativo que utiliza regressão linear das colunas (ou linhas).

Em geral, os métodos de imputação foram classificados em imputação simples ou única e imputação múltipla (IM). Na imputação simples ou única, os dados ausentes são imputados uma única vez, e então, os dados completados são analisados como se não tivessem dados ausentes, mas pelo fato da imputação ocorrer uma única vez, não se tem como quantificar as incertezas associadas aos resultados, o que pode ser uma limitação da imputação simples (Enders, 2010). A imputação é dita múltipla, quando para cada valor ausente são imputados *m valores, gerando m bancos de dados com valores imputados, podendo serem analisados por meio de procedimentos convencionais. Em geral, a IM consiste de três etapas: imputação dos valores ausentes, análise dos m bancos de dados geradas e combinação dos resultados gerados nas m análises (Zhang, 2003). Na IM, as incertezas dos dados imputados são incorporadas aos resultados e as inferências realizadas (Bergamo et al., 2008)*

Conforme Duarte e Vencovsky (1999); Hadi e Sa'diyah (2016), o modelo AMMI, comumente usado em estudos de interação genótipo por ambiente, é uma das melhores opções de análise para experimentos multiambientais, além do que, com uso do recurso visual para análise do efeito da interação por meio do gráfico AMMI-biplot, por sua praticidade, o modelo tornou-se a técnica mais recorrente de análise. O modelo AMMI combina componentes aditivos para os efeitos principais de genótipos e ambientes e componentes multiplicativos para o efeito da interação (DIAS, 2005), mas para o ajuste do modelo, não pode haver ausências de dados e a variável resposta é admitida ser do tipo contínua. Por outro lado, a boa qualidade do modelo ajustado depende de suposições sobre a variável resíduo, que deve ser normalmente e independente distribuída e com variância constante (Hadi e Sa'diyah, 2016)

Na impossibilidade ao uso do AMMI, uma das melhores alternativas é o modelo de efeitos principais aditivos e interação multiplicativa genelalizado (GAMMI) (Hadi e Sa'diyah, 2016). O modelo GAMMI de Van Eeuwijk (1995) foi especificado a partir da incorporação da teoria dos modelos lineares generalizados (MLG) de Nelder e Wedderburn (1972) à dos modelos AMMI, tornando possível postular um modelo de efeitos principais aditivos e interação multiplicativa generalizado. Em geral, seu uso é de

aplicação para analisar a estabilidade e a adaptabilidade em estudos com interação genótipo por ambiente em que as variáveis de resposta possam ser contagem, proporção, contínua entre outras, ou quando as suposições sobre a distribuição da variável resíduo forem violadas (Hadi e Sa'diyah, 2016). Apesar do modelo GAMMI ser uma excelente opção de análise, tem-se o problema das ausências, haja visto que os métodos de imputações que contornam o problema das ausências para uso do AMMI, não são estendidos para o GAMMI.

Levando em conta o exposto acima, este trabalho tem como objetivo propor um método de imputação para dados multiambientais, utilizando a mistura do modelo de efeitos principais aditivos e interação multiplicativa generalizado (GAMMI) e o algoritmo EM, tendo como base o método EM-AMMI. Assim, buscou-se contribuir para o meio científico com uma nova ferramenta para predição das ausências em dados multiambientais, contornando o problema do desbalanceamento para posterior análise do experimento.

2.2 Materiais e Métodos

Seram descritos os procedimentos EM-AMMI, EM-GAMMI, o modelo AMMI, o modelo GAMMI e exemplos em termos de dois conjuntos de g genótipos que foram testados experimentalmente em a ambientes. A resposta média das combinações de genótipos e ambientes é representada pela matriz $\mathbf{X}_{(g\times a)}$ *de elementos* $[x_{ij}]$.

2.2.1 Algoritmo EM-AMMI

O Algoritmo de imputação EM-AMMI é apresentado por Gauch e Zobel (1990) para contornar o problema das ausências de informações em experimento com interação genótipo por ambiente. Os autores combinaram o algoritmo EM com o modelo de efeitos principais aditivos e interação multiplicativa (AMMI) em um único algoritmo, de modo a realizar imputação de dados, a junção foi denominada de EM-AMMI. As etapas pertinentes para realizar o processo de imputação mediante uso do algoritmo EM-AMMI em uma matriz $\mathbf{X}$ *de dimensão* $(g \times a)$ *com elementos* $[x_{ij}]$ *(i=1,...,g; j=1,...,a), em que alguns desses elementos estão ausentes, os* $[x_{ij}^{aus}]$*, é como descrito nos passos a seguir.*

Passo 1 - Os elementos ausentes $[x_{ij}^{aus}]$ *em* $\mathbf{X}$ *são inicialmente estimados pela média geral dos valores observados + média da linha i (efeito principal de linha) + média da coluna j (efeito principal de coluna), obtendo-se uma matriz completada pelas médias* $\mathbf{X}_c$*. Também é possível o preenchimento inicial por um valor arbitrário.*

Passo 2 - Os parâmetros do modelo AMMI com k termos multiplicativos são estimados. Para tal, considera-se as entradas colunas de $\mathbf{X}_c$ *como efeito do fator ambiente e as entradas linhas o efeito do fator genótipo. Deve-se ter clareza que é ajustado inicialmente o modelo de ANOVA de dupla entrada e em seguida, é realizada a decomposição em valores singulares (DVS) da matriz de resíduo.*

Passo 3 - As médias ajustadas são calculadas com base no modelo AMMI com k termos multiplicativos. Dependendo do número de termos utilizado, o método de imputação pode ser nomeado EM-AMMI0, EM-AMMI1,..., EM-EMMIk (Gauch e Zobel, 1990).

Passo 4 - Os valores ausentes x_{ij}^{aus} *em* $\mathbf{X}$ *são preenchidos (imputados) com as estimativas AMMI apropriadas (médias ajustadas).*

Passo 5 - Se a alteração máxima entre duas estimativas de valores ausentes em etapas de iteração sucessivas (distância de Chebyshev) for maior que a precisão assumida , as etapas de 2 a 5 serão repetidas. Caso contrário, o algoritmo para.

2.2.2 Modelo AMMI

O modelo de ANOVA de dupla entrada para experimentos com interação genótipo por ambiente é como descrito abaixo:

$$Y_{ij} = \mu + g_i + a_j + (ga)_{ij} + \epsilon_{ij} \quad i = 1, \ldots, g; \quad j = 1, \ldots, a \tag{2.1}$$

em que μ é a média geral, g_i representa o efeito principal de genótipo, a_j efeito principal de ambiente, $(ga)_{ij}$ o efeito da interação genótipo por ambiente e ϵ_{ij} o erro.

O modelo AMMI é a melhor alternativa para análise de experimento em que g genótipos tenham sido testados experimentalmente em a ambientes O modelo postula componentes aditivos para os efeitos principais de genótipos (g_i) e ambientes (a_j) e componentes multiplicativos para o efeito da interação $(ga)_{ij}$ (GAUCH, 1988). Deste modo, a resposta média do genótipo i em um ambiente j, obtida de r repetições do conjunto balanceado de dados, é modelada a partir de (2.1) por:

$$Y_{ij} = \mu + g_i + a_j + \sum_{k=1}^{K} \lambda_k \alpha_{ik} \gamma_{jk} + \rho_{ij} + \epsilon_{ij} \quad i = 1, 2, \ldots, g, \quad j = 1, 2, \ldots, a \tag{2.2}$$

em (2.2), Y_{ij} é a resposta média do i-ésimo genótipo no j-ésimo ambiente, o termo $(ga)_{ij}$ em 2.1 pode ser visto como o efeito da interação entre genótipos e ambientes e modelado por $\sum_{k=1}^{K} \lambda_k \alpha_{ik} \gamma_{jk} + \rho_{ij}$. Aqui, λ_k^2 é o k-ésimo autovalor, $\boldsymbol{\alpha}_{k(1\times a)}$ e $\boldsymbol{\gamma}^t_{k(g\times 1)}$ são os respectivos vetores singulares (vetor linha e vetor coluna) associados a λ_k (valor próprio) e $\rho_{ij} = \sum_{k=K+1}^{s} \lambda_k \alpha_{ik} \gamma_{jk}$ um resíduo adicional . O índice k pode variar 1,. . . ,s em que s=min(g-1,a-1).

2.2.3 Algoritmo de imputação simples EM-GAMMI

Para se fazer imputações de dados via procedimento EM-GAMMI, em experimentos com interação genótipo por ambiente, foi realizada no algoritmo EM-AMMI a seguinte modificação: o modelo AMMI foi substituído pelo modelo de efeitos principais aditivos e interação multiplicativa generalizado (GAMMI). Tal artifício foi possível graças ao algoritmo de Van Eeuwijk (1995), que possibilitou fazer uso da metodologia dos modelos lineares generalizados (MLG), de Nelder e Wedderburn (1972), como base para estimar o modelo AMMI generalizado. As etapas do algoritmo de imputação, proposta neste trabalho, e doravante denominado método de imputação EM-GAMMI, são como descritas nos passos a seguir:

Passo 1 - Os elementos ausentes $[x_{ij}^{aus}]$ em $\mathbf{X}$ são inicialmente estimados pela expressão: -média geral dos valores observados + média da linha i (efeito principal de linha) + média da coluna j (efeito principal de coluna), obtendo-se uma matriz completada $\mathbf{X}_c$.

Passo 2 - Assumindo um MLG particular, com função de ligação específica, estima-se os parâmetros do modelo GAMMI. Considera-se as entradas colunas de $\mathbf{X}_c$ como efeito do fator ambiente e as entradas linhas o efeito do fator genótipo para o ajuste.

Passo 3 - As médias ajustadas são calculadas com base no modelo GAMMI com k termos multiplicativos. Dependendo do número de termos utilizados, o método de imputação pode ser nomeado EM-GAMMI-0, EM-GAMMI-1, EM-GAMMI-2,. . ., EM-GAMMI-K.

Passo 4 - Os valores ausentes (x_{ij}^{aus}) em $\mathbf{X}$ são preenchidos (imputados) pelas estimativas EM-GAMMI apropriadas. Como a relação entre a E(Y) e o preditor linear η não se dá de forma direta, são ligados pela função ligação, os valores preditos são retornados à escala dos dados mediante $g^{-1}(\eta)$

Passo 5 - Se a alteração máxima entre duas estimativas de valores ausentes em etapas de iteração sucessivas (distância de Chebyshev) for maior que a precisão assumida, as etapas de 2 a 5 serão repetidas. Caso contrário, o algoritmo para.

2.2.3.1 Modelo linerar generalizado - MLG

A metodologia dos modelos GAMMIs têm por base as mesmas suposições básicas que as dos modelos lineares generalizados (MLG) de Nelder e Wedderburn (1972). Para Myers e Montgomery (1997), os modelos lineares generalizados são usados quando os erros não seguem uma distribuição normal e a suposição de homogeneidade for violada. Conforme Agresti (2003); Paulino e da Motta Singer (2006) e dentre outros, um MLG envolve na sua composição uma variável resposta univariada, uma ou mais variáveis explanatórias e uma função adequada de ligação. Além de Nelder e Wedderburn (1972), na literatura, existem diversos livros clássicos que tratam de MLG (Collett, 2002; McCullagh e Nelder, 1989) e em língua portuguesa pode-se encontrar Cordeiro e Demétrio (2008); Paula (2004). Assim, diz-se que um modelo linear generalizado é especificado por três componentes como descrito a seguir:

1. *Componente aleatório: consiste de um conjunto de variáveis aleatórias independentes $Y_1, Y_2, ..., Y_n$ obtidas de uma mesma distribuição e que pertence à família exponencial de distribuições, com médias $\mu_1, \mu_2, ..., \mu_n$, ou seja, $E(Y_i) = \mu_i$, $i = 1, ..., n$;*

2. *Componente sistemático: o componente sistemático especifica a estrutura linear das variáveis explicativas/explanatórias (quantitativas e/ou qualitativas), as quais entram no modelo na forma de uma soma linear de seus efeitos, dando origem ao preditor linear, $\eta_i = \sum_j^p x_{ij}\beta_j = \mathbf{x_i^T}\boldsymbol{\beta}$. Em que $\mathbf{X}$ é a matriz cujas linhas $\mathbf{x}_i^T$ são os valores das variáveis explicativas para a estrutura paramétrica da distribuição de todos os $\{y_i\}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ é o vetor de parâmetros e $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)^T$ é o preditor linear;*

3. *Função de ligação: uma função que relaciona o componente aleatório ao componente sistemático, isto é, $g(\mu_i) = \eta_i$, em que $g(.)$ é uma função monótona e diferenciável.*

A metodologia dos modelos lineares generalizados fornece uma estrutura teórica geral para muitos modelos estatísticos, permite ajustes de modelos de regressão para uma variável resposta que assume diferentes tipos de distribuições e que por vez pertença à família exponencial de distribuições. Em geral, as estimativas dos parâmetros β_j são obtidas por procedimentos numéricos, os chamados processos iterativos, como o método de Newton-Raphson ou o método escore. Conforme Cordeiro e Paula (1989) a expressão em (2.3) descreve a relação entre os componentes do modelo linear generalizado.

$$g(\mu_i) = \eta_i = \sum_j^p x_{ij}\beta_j \tag{2.3}$$

2.2.4 Modelo de efeitos principais aditivos e interação multiplicativa generalizado - GAMMI

Van Eeuwijk (1995) propôs um modelo de efeitos principais aditivos de interação multiplicativa generalizado (GAMMI) para dados de genótipos por ambientes, dispostos em tabela de dupla entrada. No algoritmo proposto, o autor incorporou a teoria dos modelos lineares generalizados à dos AMMI, possibilitando ao modelo assumir outras distribuições além da distribuição normal, comportando dados de contagem, proporção entre outros (Hadi et al., 2010). Para o caso em que distribuição da variável resposta é a normal e a função de ligação é a identidade, o modelo GAMMI pode ser definido como o modelo AMMI. Assim, o modelo AMMI generalizado (GAMMI) para a resposta média μ_{ij} em termos de preditor linear, é como descrito em 2.4,

$$\eta_{ij} = \mu + \alpha_i + \beta_j + \sum_{k=1}^{K} \sqrt{\lambda}_k \gamma_{ik} \delta_{jk} \tag{2.4}$$

em que μ é a média geral, α_i e β_j representam efeitos das linhas e colunas, γ_{ik} e δ_{jk} são valores para o k-ésimo componente multiplicativo dos termos de interação, $\sqrt{\lambda}_k$ é o valor singular do k-ésimo componente e K é o número de termos multiplicativos que é menor ou igual ao posto da matriz.

Note que $g(.)$ é a função que liga a média $E(Y_{ij}) = \mu_{ij}$ ao preditor linear η_{ij}. No caso da função de ligação ser a identidade, o modelo em (2.4) é o próprio modelo AMMI. A expressão em (2.5) mostra a obtenção de (μ_{ij}).

$$g(\mu_{ij}) = \eta_{ij} \Rightarrow \mu_{ij} = g^{-1}(\eta_{ij}) = E(Y_{ij}) \tag{2.5}$$

Em (2.4), fixando os valores β_j e δ_{jk} o modelo se reduz a um MLG por linha, enquanto se fixando os valores α_i e γ_{ik} o modelo é reduzido a um MLG por coluna. Esta característica serve como base para o procedimento de estimação dos parâmetros do modelo. O algoritmo para o ajuste do modelo GAMMI é bastante complexo, usa regressão alternada entre as regressões de linha e de coluna realizadas, e cada regressão de linha e de coluna inclui um MLG, e que por vez o processo é iterativo (Hadi et al., 2010). Portanto, o algoritmo envolve três convergências, regressão de linha, regressão de coluna e regressão alternada. As etapas para estimação dos parâmetros do modelo GAMMI e seu respectivo esboço são descritos em Van Eeuwijk (1995); Hadi et al. (2010); Acorsi et al. (2016); Turner e Firth (2018) e em língua portuguesa Acorsi (2010).

2.2.5 Os dados usados na pesquisa

Para avaliar o procedimento de imputação, proposta deste trabalho, foram considerados dois conjuntos de dados reais, completos e provenientes de experimentos com interação genótipo por ambiente. O primeiro conjunto de dados é um delineamento em blocos ao acaso, estudo da resistência de soja à praga foliar, publicado por Hadi et al. (2010). No experimento, foram utilizados quatro genótipos de soja resultantes de híbridos (Wilis, IAC-100, IAC-80 e W-80) e, avaliados aos 14 dias após o plantio, a contagem de praga foliar encontradas por planta. Na contagem, cinco tipos de pragas foliares foram classificadas nas variedades genótipos de soja. A Tabela 2.1 apresenta a média da população dos cinco tipos de praga foliar em quatro genótipos de soja. A escolha deste conjunto de dados foi particular, pois as respostas médias das repetições são expressas em escala intervalar e analisadas pela metodologia GAMMI (Hadi et al., 2010). Deste modo, foi possível fazer uso dos modelos AMMI e GAMMI para posterior uso dos algoritmos de imputações. Este conjunto foi nomeado de dados de praga foliar para fins de referência.

O segundo conjunto de dados utilizado foi obtido de parte de um delineamento em blocos aleatorizados cedido pelos pesquisadores Spitti et al. (2019). No estudo, foram usados 19 genótipos de feijoeiro observados em seis ambientes. Os genótipos foram avaliados quanto a cor do tegumento dos grãos em função do valor de luminosidade(L) em relação ao método de crescimento em condições de prateleira. A variável resposta é a tolerância (resistência) do genótipo a perda de pigmentos, ou seja, mudança gradativa da coloração dos grãos aos 60 dias. A Tabela 3.2 mostra os valores médios dos genótipos por ambiente obtidos das seis regiões consideradas no estudo. Este conjunto foi nomeado de dados de Acácia para fins de referência.

Tabela 2.1 – Média da população de cinco tipos de praga foliar em quatro genótipos de soja

Genótipos	Tipos de praga foliar				
	Bemissia	Emproosca	Agronyza	Lamprosema	Longitarsaus
IAC-100	0,50	1,75	2,25	0,50	1,75
IAC-80	3,00	2,75	1,00	1,75	3,25
W-80	3,50	4,00	1,25	2,00	2,00
Wilis	4,00	3,00	1,00	1,75	4,00

Fonte: Hadi et al. (2010)

Tabela 2.2 – Média de genótipos de feijoeiros avaliados quanto a cor do tegumento de grãos em função do valor de luminosidade(L)

Genótipos	Regiões					
	R1	R2	R3	R4	R5	R6
BRS Pérola	0,5041	0,4727	0,5036	0,4497	0,4840	0,4957
CHC 01-175-1	0,4987	0,4648	0,5105	0,4610	0,4747	0,5013
CNFC 11-948	0,5068	0,4703	0,5023	0,4618	0,5048	0,5110
CNFC 11-954	0,5013	0,4585	0,4867	0,4708	0,4992	0,4961
Gen 4-1F-19P	0,5263	0,5000	0,4909	0,4892	0,5241	0,5245
Gen 12-2F-67	0,5178	0,4681	0,5021	0,4790	0,5098	0,5184
Gen 20-4F-129	0,5122	0,4847	0,4844	0,4494	0,4987	0,5343
Gen 45-2F-293P	0,5244	0,4922	0,5083	0,4792	0,5326	0,5493
Gen 78-1A-59	0,5078	0,4907	0,4950	0,4717	0,5168	0,5291
Gen 86-12A-122	0,5055	0,4776	0,4907	0,4501	0,4878	0,5215
Gen 90-4A-160	0,5106	0,4692	0,4993	0,4588	0,5002	0,5228
Gen 104-1A-291	0,5314	0,4901	0,5109	0,4677	0,5197	0,5304
Gen 106-4A-317	0,5107	0,4882	0,4999	0,4497	0,5016	0,5346
Gen 106-6A-319	0,5195	0,4794	0,5014	0,4856	0,5143	0,5226
Gen 107-14A-336	0,5256	0,4777	0,5145	0,4563	0,5348	0,5552
Gen 125-10A-510	0,5123	0,4670	0,5103	0,4756	0,4987	0,5183
IAC Milênio	0,5219	0,4803	0,5063	0,4873	0,5017	0,5111
IAC Sintonia	0,5028	0,4682	0,4821	0,4588	0,4899	0,5276
LP 11-363	0,5394	0,4810	0,5201	0,4703	0,5018	0,5144

Fonte: Spitti et al. (2019)

2.2.6 Simulações geradas a partir dos dados reais

Para os dois conjunto de dados experimentais, usados no trabalho, foram realizadas simulações de retiradas aleatórias nas porcentagens de 10%, 20% *e* 30% *para o conjunto de praga foliar e de* 10%, 20%, 30% *e 40% para os dadaos de Acácia, pois o número de dados ausentes em interação genótipo ambiente é menor que 40% (Yan, 2013). Este processo foi repetido 100 vezes para cada porcentagem de retirada em cada um dos dois conjuntos de valores, obtendo 700 matrizes diferentes com valores ausentes simulados. Em seguida, para cada uma das 700 matrizes com valores ausentes simulados, foram feitas as imputações com os algoritmos EM-GAMMI-0, EM-GAMMI-1 e EM-GAMMI-2, obtendo 2100 matrizes completadas (valores observados + valores imputados).*

As etapas, simulações e predições, foram realizadas por meio de rotinas computacionais desenvolvidas e implementadas para o programa R Core Team (2020). Destaca-se no algoritmo desenvolvido, uso da função gnm para ajuste do modelo GAMMI com até dois termos multiplicativos. Para os dados de praga foliar, foram usados os modelos GAMMI Poisson e GAMMI Gaussiano com respectivas funções de ligações logarítmica e identidade. A escolha do modelo GAMMI Poisson foi decorrente do estudo de Hadi et al. (2010). Para os dados de Acácia, utilizou-se o modelo GAMMI Binomial com função de ligação logística, pois os dados representam uma proporção. Também foram geradas imputações pelo algoritmo

EM-AMMI, com uso da função EM-AMMI. As simulações de retiradas aleatórias, supondo o padrão de ausências aleatória - MAR (Missing at Random), deram-se mediante uso da função SimIm do pacote multivariado ImputeR. Assim, os dados de praga foliar e Acácia foram imputados neste trabalho.

2.2.7 Modelos ajustados

O modelo GAMMI é um dos melhores modelos para análise de experimentos com interação genótipo por ambiente, em que ocorrem violações das suposições do modelo de ANOVA, ou quando a resposta é uma contagem, uma proporção, entre outras. Por esta razão, para cada uma das matrizes com valores ausentes, obtidas por simulações a partir do conjunto original, foram geradas imputações pelo método EM-GAMMI com até k termos multiplicativos (k=0,1,2), em que se fez uso de modelos de efeitos principais aditivos e interação multiplicativa generalizado (GAMMI). Para o conjunto de praga foliar, foi assumido para o algoritmo EM-GAMMI os modelos Poisson e Gaussiano com funções de ligações logarítmica e identidade respectivamente e para os dados de Acácia, o modelo Binomial com função de ligação logística.

2.2.8 NRMSE como critério de seleção

*Foi utilizado para comparar os valores verdadeiros com os resultados obtidos pelo método de imputação o critério de menor valor da raiz do erro quadrático médio padronizado - NRMSE (Ching et al., 2010). Pelo critério da NRMSE, o algoritmo é comparado por meio das médias ajustadas, ou seja, os valores imputados são comparados com os correspondentes valores observados no conjunto dos dados originais por meio da expressão em (*2.6*), e é considerado como melhor método ou método com melhor desempenho, aquele que apresentar o menor valor da estatística NRMSE. Também foi explorado o uso de gráficos para, em forma visual, comparar e concluir sobre o melhor procedimento em análise.*

$$NRMSE = \frac{\sqrt{média(\mathbf{x}_{imp} - \mathbf{x}_{obs})^2}}{\mathbf{s}(\mathbf{x}_{obs})} \tag{2.6}$$

em que $\mathbf{x}_{imp}$ *e* $\mathbf{x}_{obs}$ *são vetores contendo os respectivos valores médios ajustados (imputados) e os valores verdadeiros das observações ausentes simuladas e* $\mathbf{s}(\mathbf{x}_{obs})$ *é o desvio padrão dos valores contidos no vetor* $\mathbf{x}_{obs}$*. Quanto menor for o valor da estatística NRMSE, melhor será o método de imputação.*

2.3 Resultados e Discussões

2.3.1 Dados de praga foliar

Nas Tabelas 2.3, 2.4 e 2.5 são apresentadas as estatísticas: valor mínimo, 1º quartil, mediana, média, 3º quartil e valor máximo da NRMSE para os procedimentos de imputações EM-GAMMI e EM-AMMI com k termos (k = 0, 1 e 2) multiplicativos. A análise conjunta destas estatísticas descreve o comportamento da distribuição da NRMSE e, consequentemente, a escolha do método com melhor desempenho. Quanto menor for o valor da NRMSE, melhor será o desempenho do método. Também, foram utilizadas para "comparar"as distribuições da NRMSE, o teste não-paramétrico de Kruskal-Wallis com posterior teste de Wilcoxon, se o primeiro for significativo. O teste de Kruskal-Wallis foi empregado para comparar k > 2 grupos e o teste e Wilcoxon compara os k grupos, um contra o outro, na tentativa de encontrar qual ou quais grupos foram responsáveis por produzir diferenças significativas pelo teste Kruskal-Wallis.

Para retiradas aleatórias de 10% (Tabela 2.3), observou-se que o procedimento EM-GAMMI-0, sem nenhum termo multiplicativo, apresentou melhor desempenho em termos médio e mediano da estatística quando comparado com os resultados dos procedimentos EM-GAMMI-1 e EM-GAMMI-2. Já,

para o procedimento EM-AMMI, verificou-se para o EM-AMMI-0 o menor valor da NRMSE em termos de mediana, indicando ser este o método com melhor desempenho quando comparado com o EM-AMMI-1 e EM-AMMI-2. Fazendo um comparativo entre os valores da estatística NRMSE para os procedimentos EM-GAMMI-0 e EM-AMMI-0, verificou-se não haver evidência estatística de diferença pelo teste de Wilcoxon, valor-p = 0,3965. Todavia, vale destacar pequena vantagem numérica em termos de mediana (menor mediana) em favor do procedimento EM-GAMMI-0.

A Tabela 2.4 apresenta resultados da NRMSE padronizada para a porcentagem de 20% de ausências. Nesta, observa-se para o EM-GAMMI-0 valor da NRMSE mediana de 0,960, menor que a dos procedimentos EM-GAMMI-1 e EM-GAMMI-2. Para o método EM-AMMI, observou-se menor valor da NRMSE mediana de 1,029 para EM-AMMI-0 quando comparada com as dos procedimentos EM-AMMI-1 e EM-AMMI-2. Já, quando comparadas as NRMSE dos métodos EM-GAMMI-0 e EM-AMMI-0, não foi possível constatar evidência estatística de diferença pelo teste de Wilcoxon (valor-p = 0,7031), mas vale destacar pequena vantagem em favor do procedimento EM-GAMMI-0 em termo da NRMSE mediana.

Para retiradas de 30% (Tabela 2.5), foi observado que o procedimento EM-GAMMI-2 apresentou menor valor mediano da NRMSE (1,036) em comparação com os EM-GAMMI-0 e EM-GAMMI-1. Neste caso, a melhor escolha foi pelo procedimento com dois termos multiplicativos, o EM-GAMMI-2. Em relação ao método EM-AMMI, observou-se menor valor da NRMSE mediana para o EM-AMMI-2 (1,032). Quando comparados os procedimentos EM-GAMMI-2 e o EM-AMMI-2 em termos das suas respectivas NRMSE, não foi possível constatar evidência estatística de diferença, valor-p = 0,8815, apesar do EM-AMMI-2 ser numericamente favorável pelo critério da NRMSE (1,032).

Foi observado que aumentos dos níveis de ausências, ocasionaram aumentos nos valores das estatísticas NRMSE (Tabelas 2.3, 2.4 e 2.5). Ainda, pelo teste de kruskl-Wallis, verificou-se haver diferenças estatísticas significativas, valor-p = 0,01338, quando comparadas as estatísticas NRMSE dos procedimentos EM-GAMMI-0 com ausências de 10%, o EM-GAMMI-0 com ausências de 20% e o EM-GAMMI-2 com ausências de 30%. Já, pelo teste de Wilcoxon, à NRMSE, para ausências de 10% difere da de 20% (valor-p =0,0234) e da de 30% (valor-p = 0,0063), e as de 20% e 30% não diferem entre si (valor-p = 0,5625). Vale observar que o conjunto de dados é de tamanho 20, o que o torna bastante sensível a retiradas a partir de um determinado nível. Por exemplo, para o nível de ausências de 30%, dois termos multiplicativos foram necessários ser incluídos, de modo a melhorar o modelo e consequentemente o método de imputação.

Tabela 2.3 – Descrição da NRMSE em termos de estatísticas básicas para ausências de 10%, em que foram usados os modelos GAMMI e AMMI no processo de imputação, dados de praga foliar

	Porcentagem 10%					
Estatistica	GAMMI-0	GAMMI-1	GAMMI-2	AMMI-0	AMMI-1	AMMI-2
mínimo	0,183	0,243	0,081	0,219	0,173	0,145
1º quartil	0,678	0,751	0,581	0,633	0,952	0,620
mediana	0,930	1,395	1,044	0,965	1,462	1,076
média	0,983	3,569	1,190	1,086	2,168	1,049
3º quartil	0,183	1,752	1,539	1,590	2,401	1,352
máximo	1,343	38,05	5,128	2,179	7,881	2,443

Tabela 2.4 – Descrição da NRMSE em termos de estatísticas básicas para ausências de 20%, em que foram usados os modelos GAMMI e AMMI no processo de imputação, dados de praga foliar

	Porcentagem 20%					
Estatistica	GAMMI-0	GAMMI-1	GAMMI-2	AMMI-0	AMMI-1	AMMI-2
mínimo	0,381	0,496	0,196	0,459	0,710	0,310
1º quartil	0,672	1,047	0,668	0,729	1,177	0,708
mediana	0,960	1,480	0,975	1,029	1,892	1,047
média	1,230	4,653	1,215	1,377	2,368	1,193
3º quartil	1,362	3,836	1,340	1,510	3,093	1,337
máximo	7,770	26,37	6,977	7,701	9,146	6,959

Tabela 2.5 – Descrição da NRMSE em termos de estatísticas básicas para ausências de 30%, em que foram usados os modelos GAMMI e AMMI no processo de imputação, dados de praga foliar

	Porcentagem 30%					
Estatistica	GAMMI-0	GAMMI-1	GAMMI-2	AMMI-0	AMMI-1	AMMI-2
mínimo	0,533	0,727	0,305	0,488	0,536	0,386
1º quartil	0,837	1,173	0,853	0,926	1,361	0,859
mediana	1,056	1,501	1,036	1,192	2,016	1,032
média	1,201	3,751	1,121	1,303	2,149	1,122
3º quartil	1,321	3,427	1,288	1,514	2,588	1,257
máximo	4,334	22,677	3,357	4,164	6,889	2,919

As Figuras 2.1, 2.2 e 2.3 apresentam o comportamento da NRMSE para os métodos com melhores desempenhos selecionados entre os EM-GAMMI-0, EM-GAMMI-1 e EM-GAMMI-2, bem como para os procedimentos EM-AMMI-0, EM-AMMI-1 e EM-AMMI-2, nos níveis de ausências de 10%, 20% e 30%. Para facilidade e clareza da representação gráfica, foram tomados resultados de 10 matrizes nas respectivas posições (m81, m82, . . . , m90), selecionadas entre 100 simuladas em cada nível de ausência e com a mesma semente. Para ausências de 10%, Figura 2.1, observa-se vantagem da técnica EM-GAMMI em relação a técnica EM-AMMI, exceção para as NRMSE das matrizes m86, m89 e m90. Todavia, pelo teste de Wilcoxon, não foi possível constatar evidências de diferenças estatísticas entre as NRMSE dos dois métodos comparados, valor-p=0,3254.

Para o nível de ausência de 20%, Figura 2.2, o algoritmo com mais baixo desempenho, pelo critério de menor valor da estatística, é o EM-AMMI, exceto para a matriz m90. As maiores diferenças numéricas ocorreram para as matrizes m87 com posições de retiradas em linha × coluna ((1,1), (2,5), (4,3) e (4,4)) e m88, com posições de retiradas ((1,1), (1,2), (2,2) e (2,5)). Entretanto, pelo teste de Wilcoxon, não foi constatado evidência estatística de diferença, valor-p = 0,315. Para as imputações,

geradas a partir das retiradas de 30%, Figura 2.3, o algoritmo com melhor desempenho é o EM-GAMMI, mas pelo teste de Wilcoxon não foi possível verificar evidência estatística de diferença, valor-p= 0,7959.

Em todos os três cenários de retiradas, observou-se vantagem a favor do método de imputação EM-GAMMI em termos da NRMSE em comparação com o EM-AMMI, apesar de estatisticamente não ser confirmada tal evidência. Vale observar que o método EM-AMMI é de ampla aceitação no meio cientifico como técnica de imputação, o que motivou a usá-lo como padrão de comparação. Conforme Arciniegas-Alarcón e Dias (2009), a imputação com modelos AMMI0, AMMI1 e AMMI2, em alguns casos, pode proporcionar melhores resultados do que a imputação com IMLD, Arciniegas-Alarcón (2015), em um estudo comparativo entre os algoritmos EM+DVS, EM-AMMI0, EM-AMMI1, GabrielEigen e IMLD, concluiu à favor dos métodos EM+DVS e EM-AMMI0.

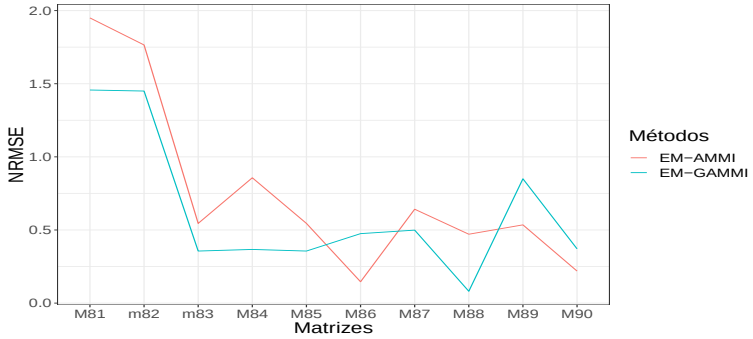

Figura 2.1 – Gráfico da estatística NRMSE com ausências de 10% para os procedimento EM-AMMI e EM-GAMMI, dados de praga foliar

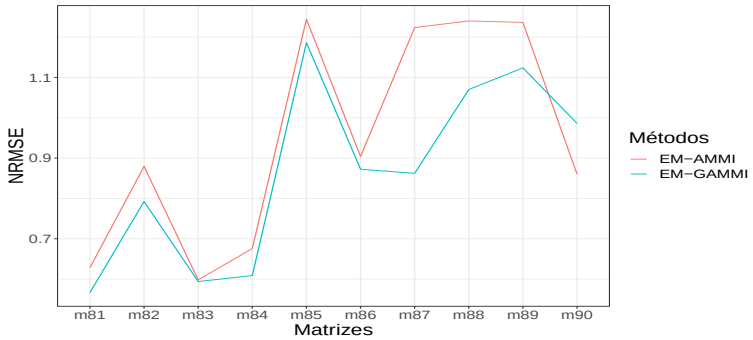

Figura 2.2 – Gráfico da estatística NRMSE com ausências de 20% para os procedimento EM-AMMI e EM-GAMMI, dados de praga foliar

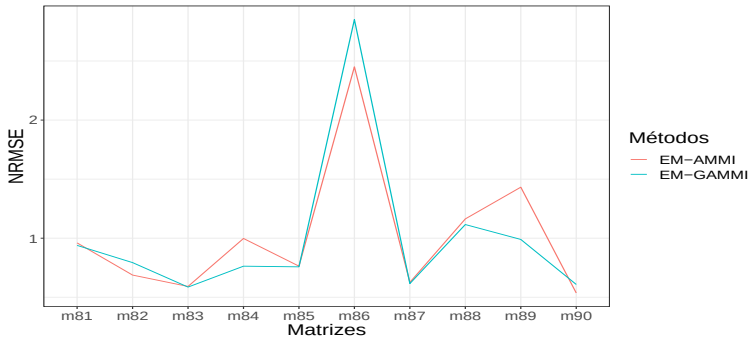

Figura 2.3 – Gráfico da estatística NRMSE com ausências de 30% para os procedimento EM-AMMI e EM-GAMMI, dados de praga foliar

2.3.2 Aplicação com dados de praga foliar

A Tabela 2.6 mostra os valores originais (VO) na posição p_{ij} da amostra 87 (matriz m87) para retiradas aleatórias de 10%, 20% e 30%, valores imputados pelas duas técnicas e a estatística NRMSE. Para retiradas de 10% e 30% respectivamente as imputações em que foram considerados os modelos GAMMI-0 e AMMI-0, GAMMI-2 e AMMI-2, os valores imputados correspondentes estão bem próximos. Pelo critério da NRMSE, o método de imputação EM-GAMMI foi o que apresentou menor valor em todos os três níveis de retiradas quando comparado ao EM-AMMI.

Tabela 2.6 – Valores originais (VO), NRMSE e valores imputados pelo melhor método selecionado entre os k-termos do procedimento EM-GAMMI e do EM-AMMI, conjunto de praga foliar

10%			20%			30%		
VO (p_{ij})	GAMMI-0	AMMI-0	VO (p_{ij})	GAMMI-1	AMMI-1	VO (p_{ij})	GAMMI-2	AMMI-2
1,75 (1,2)	1,76	2,19	0,50 (1,1)	1,98	2,64	4,00 (4,1)	2,82	2,81
3,50 (3,1)	2,79	2,71	1,00 (4,3)	1,82	2,22	1,75 (1,2)	2,12	2,12
			1,75 (4,4)	2,09	2,47	2,75 (2,2)	2,93	2,97
			3,25 (2,5)	2,11	2,21	4,00 (3,2)	3,23	3,21
						1,75 (1,5)	2,12	2,12
						3,25 (2,5)	2,95	2,95
NMRSE	0,499	0,642		0,862	1,224		0,616	0,626

2.3.3 Dados de Acácia

As Figuras 2.4, 2.5, 2.6 e 2.7 apresentam as distribuições da estatística NRMSE para retiradas de 10%, 20%, 30% e 40% do conjunto de dados de Acácia. Nestas, observou-se que as distribuições geradas a partir dos procedimentos EM-GAMMI-0, EM-GAMMI-1 e EM-GAMMI-2 mostraram pequena assimetria à direita para as porcentagens de 30% e 40%. Essa assimetria para retiradas maiores de porcentagens sinaliza pequena tendência no método, a NRMSE tende a apresentar valores maiores em seus resultados, diferenças entre os dados reais e imputados correspondentes. Ainda, apareceram alguns poucos valores fora do limite de amplitude padrão, o que não compromete a qualidade do método, pois vale observar, que quando a amostra é grande, é esperado que ocorram alguns poucos valores fora da amplitude gráfica, mesmo para modelos que apresentam bons resultados.

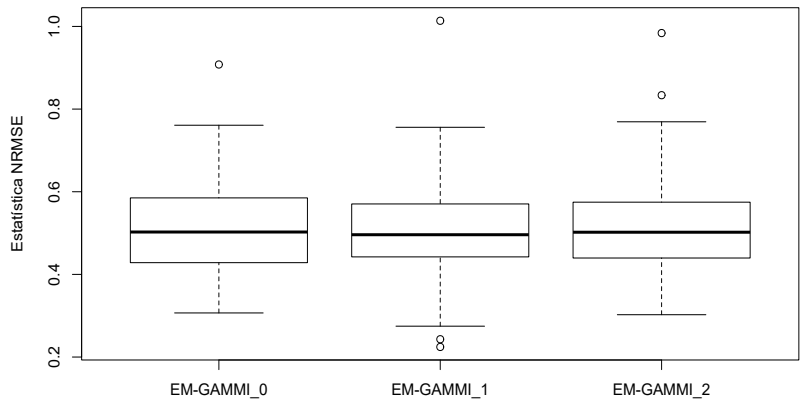

Figura 2.4 – Gráfico de caixa da distribuição da estatística NRMSE para ausências de 10%, dados de Acácia

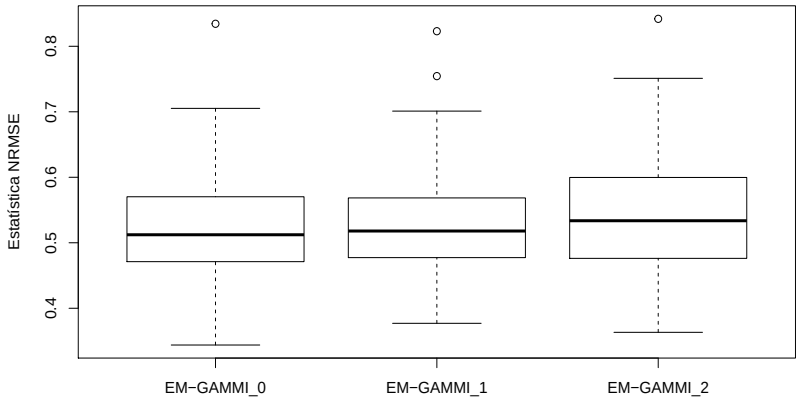

Figura 2.5 – Gráfico de caixa da distribuição da estatística NRMSE para ausências de 20%, dados de Acácia

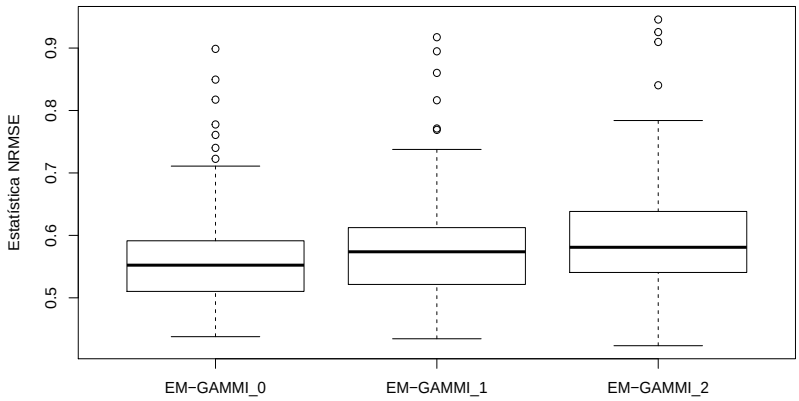

Figura 2.6 – Gráfico de caixa da distribuição da estatística NRMSE para ausências de 30%, dados de Acácia

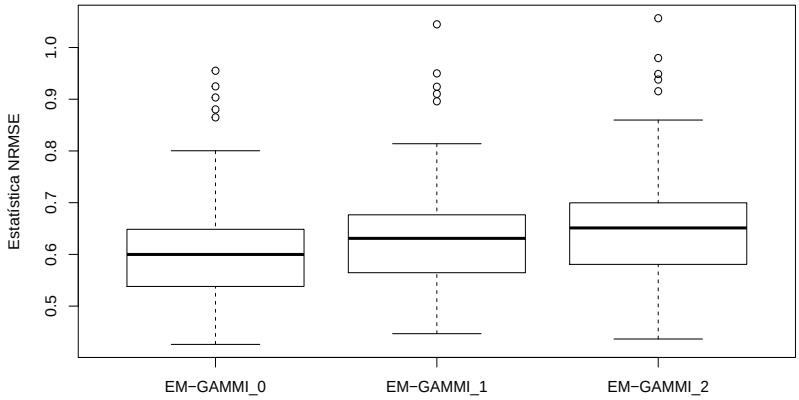

Figura 2.7 – Gráfico de caixa da distribuição da estatística NRMSE para ausências de 40%, dados de Acácia

A Tabela 2.7 apresenta a mediana e a média para a distribuição da estatística NRMSE e o teste de Kruskal-Wallis para comparar os procedimentos EM-GAMMI-0, EM-GAMMI-1 e EM-GAMMI-2, em cada um dos níveis de retiradas. Com ausências de 10%, os procedimentos EM-GAMMI-0, EM-GAMMI-1 e EM-GAMMI-2 têm mediana e média aproximadamente iguais, e em termos estatísticos, não se evidenciou diferença entre os procedimentos pelo teste Kruskal-Wallis, valor-p = 0,9021. Entretanto, o procedimento EM-GAMMI-1 foi o que apresentou melhor desempenho em termos de valores absoluto da NRMSE (mediana=0,4960 e média=0,5057). Para ausências de 20%, não se constatou evidência estatística de diferença entre os métodos pelo teste de Kruskal-Wallis, valor-p = 0,3181. Todavia, vale observar menor valor (mediana=0,5122 e média=0,5232) da distribuição da NRMSE ocorrendo para EM-GAMMI-0. Já, com ausências de 30%, pelo teste de Kruskal-Wallis, tem-se evidências estatísticas que pelo menos um dos métodos diferem, valor-p=0,0094. Para retiradas de 40%, resultado similar as de 30%, valor-p=0,0022.

A Tabela 2.8 apresenta a estatística de Wilcoxon para os níveis de ausências de 30% e 40% que apresentaram diferenças significativas pelo teste de Kruskal-Wallis, mostrado na Tabela 2.7. O teste Wilcoxon constatou para ausências de 30%, haver diferença significativa entre os procedimentos EM-GAMMI-0 e EM-GAMMI-2, valor-p=0,0028, as demais comparações não apresentaram diferenças estatísticas, sendo assim, pelo critério da NRMSE, o procedimento EM-GAMMI-0 foi o que apresentou melhor desempenho em termos de valores absolutos, mediana=0,5523 e média=0,5699. Para retiradas de 40%, o procedimento EM-GAMMI-0 difere estatisticamente dos demais, e pelo critério utilizado, este é o procedimento recomendável em termos de mediana=0,5999 e média=0,6111 da distribuição da NRMSE.

Tabela 2.7 – Mediana e média da distribuição da NRMSE e testes de Kruskal-Wallis, dados de Acácia

Porcentagem	NRMSE	GAMMI-0	GAMMI-1	GAMMI-2	Kruskal-Wallis	valor-p
10 %	Mediana	0,5026	0,4960	0,5021	0,2061	0,9021
	Média	0,5128	0,5057	0,5140		
20 %	Mediana	0,5122	0,5179	0,5336	2,2908	0,3181
	Média	0,5232	0,5287	0,5404		
30 %	Mediana	0,5523	0,5737	0,5809	9,3174	0,0094
	Média	0,5699	0,5863	0,6022		
40 %	Mediana	0,5999	0,6310	0,6512	12,232	0,0022
	Média	0,6111	0,6357	0,6550		

Tabela 2.8 – Teste de Wilcoxon para a distribuição da NRMSE, níveis de 30% e 40% que apresentaram significância pelo teste de Kruskal-Wallis, dados de Acácia

	30%		40%	
Comparação	Wilcoxon	valor-p	Wilcoxon	valor-p
EM-GAMMI-0 - EM-GAMMI-1	4320	0,0966	4169	0,04231
EM-GAMMI-0 - EM-GAMMI-2	3768	0,0026	3597	0,0006
Em-GAMMI-1 - EM-GAMMI-2	4402	0,1440	4359	0,1173

A Tabela 2.9 mostra o resultado do teste de Kruskal-Wallis para os quatros melhores procedimentos recomendados de acordo com o critério da NRMSE, apresentados nas Tabelas 2.7 e 2.8, e o teste de Wilcoxon. Nesta, observou-se diferenças significativas para pelo menos um dos procedimentos pelo teste de Kruskal-Wallis, valor-p = 0,0001. Todavia, pelo teste de Wilcoxon, só não foi possível constatar evidência estatística de diferença entre os procedimentos EM-GAMMI-1, com ausências 10% e o EM-GAMMI-0, com ausências de 20%, valor-p = 0,1111, demais comparações apresentam evidências estatísticas de diferenças entre si. Foi possível constatar, que a partir de um certo nível de retirada, o

procedimento tende a predizer valores imputados mais afastado dos valores observados correspondentes. Yan (2013) apresenta um limite para os níveis de retiradas de dados, argumenta que as ausências devem ser menores que 40%, afim de que o procedimento de imputação seja capaz de prever com êxito valores ausentes e recuperar padrões de dados, caso contrário, o método fornece resultados inconsistentes.

Tabela 2.9 – Teste de Kruskal-Wallis e teste Wilcoxon sobre a distribuição da NRMSE dos quatro melhores modelos selecionados, dados de Acácia

Comparação	Kruskal-Wallis	valor-p
	70,401	0,0001
Porcentagens e comparações	Wilcoxon	valor-p
10% (EM-GAMMI-1) - 20% (EM-GAMMI-0)	4348	0,1111
10% (EM-GAMMI-1) - 30% (EM-GAMMI-0)	3064	0,0001
10% (EM-GAMMI-1) - 40% (EM-GAMMI-0)	2253	0,0001
20% (EM-GAMMI-0) - 30% (EM-GAMMI-0)	3349	0,0001
20% (EM-GAMMI-0) - 40% (EM-GAMMI-0)	2299	0,0001
30% (EM-GAMMI-0) - 40% (EM-GAMMI-0)	3571	0,0005

2.3.4 Aplicação com os dados de Acácia

Foi observado na Tabela 2.10, os valores originais (V.O) na posição p_{ij} e os respectivos imputados (imp) pelo método EM-GAMMI de uma amostra, matriz de posição 9 entre 100 simuladas para cada um dos níveis de ausências (10%, 20%, 30% e 40%). Visualmente, percebe-se uma boa qualidade do método quando se observa a proximidade entre o verdadeiro valor e o imputado correspondente, os valores imputados são bem próximos. Por outro lado, aqui vale duas observações que de certo modo podem ou estão favorecendo à boa qualidade do método: 1) A matriz de dados apresenta dimensão relativamente grande, quando comparado com a matriz de dados de praga foliar e 2) os dados não apresentam variabilidade grande, pouca dispersão.

Para o conjunto de dados de praga foliar, em uma semente, diferente da utilizada, observou-se problema de convergência. Esta ocorrência foi observada em uma única matriz dentre 100 simuladas para o nível de retirada de 10%. Apesar desta deficiência, o uso do algoritmo EM-GAMMI é uma excelente opção para contornar o problema das ausências, por exemplo, casos em que o modelo AMMI não é apropriado, e de acordo com Hadi et al. (2010), o modelo GAMMI deve ser preferido, o que faz do método EM-GAMMI a melhor alternativa ao método EM-AMMI.

No geral, quando analisados os procedimentos EM-GAMMI e o EM-AMMI, em cada um dos níveis de ausências, constatou-se para o conjunto de dados de praga foliar, vantagem do método EM-GAMMI em termos médio e mediano da estatística NRMSE, todavia não sendo confirmado pelo teste de Wilcoxon. As Figuras 2.1, 2.2 e 2.3 mostram o comportamento em termos gráficos da distribuição da NRMSE para dez matrizes em que os dois métodos foram usados, com destaque a favor do método EM-GAMMI, mas pelo teste de Wilcoxon, não se constatou diferença estatística. Para os dados de Acácia, os resultados da NRMSE foram bem menores que os de praga foliar, indicando que o método reproduz resultados próximos aos verdadeiros, o que pode ser um ótimo indicativo da eficácia do método.

Outros estudos sobre experimentos incompletos com interação genótipo ambiente e com uso de modelos GAMMIs são necessárias e podem envolver, por exemplo, novos critérios de avaliações dos algoritmos, novos modelos, mecanismos de ausência de dados diferentes do completamente aleatório (MCAR) e definidos por (Little e Rubin, 2002) ou uso de outros algoritmos eficientes como padrão para comparação. Enquanto novas pesquisas não forneçam resultados contraditório, o método EM-GAMMI aparece como uma excelente alternativa de imputação de dados.

Tabela 2.10 – Valores originais (V.O) e valores imputados (imp) das retiradas aleatórias de 10%, 20%, 30% e 40% na posição ($p_{i,j}$), dados de Acácia

10%		20%		30%		40%	
V.O	(p_{ij}) imp	V.O	(p_{ij}) imp	V.O	(p_{ij}) imp	V.O	(p_{ij}) imp
0.5122	(7,1) 0,5056	0.5314	(12,1) 0,5204	0,5013	(4,1) 0,4981	0,5441	(1,1) 0,5005
0.5219	(17,1) 0,5135	0.4648	(2,2) 0,4675	0,5263	(5,1) 0,5211	0,5648	(3,1) 0,5199
0,4585	(4,2) 0,4672	0,4681	(6,2) 0,4829	0,5122	(7,1) 0,5020	0,5013	(4,1) 0,4981
0,4922	(8,2) 0,4977	0,4847	(7,2) 0,4797	0,5078	(9,1) 0,5204	0,5263	(5,1) 0,5237
0,5145	(15,3) 0,4909	0,4922	(8,2) 0,4962	0,5195	(4,1) 0,5195	0,5178	(6,1) 0,5112
0,5103	(16,3) 0,5019	0,4907	(9,2) 0,4840	0,5219	(7,1) 0,5118	0,5122	(7,1) 0,5057
0,5063	(17,3) 0,5026	0,4777	(15,2) 0,4948	0,4648	(2,2) 0,4669	0,5106	(11,1) 0,5097
0,4497	(13,4) 0,4677	0,4950	(9,3) 0,5059	0,4703	(3,2) 0,4749	0,5314	(12,1) 0,5216
0,4873	(17,4) 0,4686	0,4907	(10,3) 0,4919	0,5000	(5,2) 0,4950	0,4648	(2,2) 0,4699
0,4957	(11,6) 0,5026	0,5109	(12,3) 0,5076	0,4922	(8,2) 0,4977	0,4703	(3,2) 0,4849
0.5346	(13,6) 0,5234	0,4999	(13,3) 0,5002	0,4776	(10,2) 0,4690	0,5000	(5,2) 0,4888
		0,4497	(1,4) 0,4570	0,4794	(14,2) 0,4787	0,4847	(7,2) 0,4747
		0,4618	(3,4) 0,4640	0,4670	(16,2) 0,4827	0,4692	(11,2) 0,4757
		0,4708	(4,4) 0,4517	0,5105	(2,3) 0,4886	0,4901	(12,2) 0,4877
		0,4494	(7,4) 0,4685	0,5021	(6,3) 0,5054	0,4777	(15,2) 0,4977
		0,4756	(16,4) 0,4663	0,4907	(10,3) 0,4941	0,5023	(3,3) 0,5097
		0,4992	(4,5) 0,4878	0,5145	(5,3) 0,5080	0,5083	(8,3) 0,5049
		0,5016	(13,5) 0,5034	0,5063	(7,3) 0,4922	0,4950	(9,3) 0,4956
		0,5143	(14,5) 0,5091	0,4892	(5,4) 0,4851	0,4907	(10,3) 0,4940
		0,5017	(17,5) 0,5088	0,4790	(6,4) 0,4678	0,4993	(11,3) 0,4987
		0,4899	(18,5) 0,4891	0,4717	(9,4) 0,4771	0,5109	(12,3) 0,5113
		0,5013	(2,6) 0,5092	0,4677	(12,4) 0,4843	0,5014	(14,3) 0,5056
		0,5276	(18,6) 0,5058	0,4747	(2,5) 0,4915	0,4821	(18,3) 0,4917
				0,4992	(4,5) 0,4880	0,4618	(3,4) 0,4722
				0,5098	(6,5) 0,5018	0,4790	(6,4) 0,4636
				0,5326	(8,5) 0,5224	0,4497	(13,4) 0,4676
				0,5143	(14,5) 0,5121	0,4756	(16,4) 0,4577
				0,5348	(15,5) 0,5103	0,4873	(17,4) 0,4762
				0,4987	(16,5) 0,5073	0,4588	(18,4) 0,4544
				0,5018	(19,5) 0,5086	0,4703	(19,4) 0,4833
				0,5343	(7,6) 0,5118	0,4747	(2,5) 0,4930
				0,5111	(17,6) 0,5094	0,4992	(4,5) 0,4831
				0,5276	(18,6) 0,5053	0,5098	(6,5) 0,5015
				0,5144	(19,6) 0,5304	0,5326	(8,5) 0,5113
						0,5168	(9,5) 0,5030
						0,4878	(10,5) 0,5005
						0,5143	(14,5) 0,5088
						0,4987	(16,5) 0,5028
						0,5017	(17,5) 0,5053
						0,4957	(1,6) 0,5023
						0,5110	(3,6) 0,5266
						0,4961	(4,6) 0,4912
						0,5493	(8,6) 0,5251
						0,5291	(9,6) 0,5170
						0,5346	(13,6) 0,5241
						0,5276	(18,6) 0,5100

2.4 Conclusão

No presente trabalho, foi analisada uma nova metodologia estatística de imputação de dados em experimentos multiambientais, tendo como marco teórico o algoritmo EM-AMMI de Gauch e Zobel (1990). Os valores da NRMSE para os métodos EM-AMMI e EM-GAMMI permitiram constatar vantagem a favor do procedimento EM-GAMMI. Para o conjunto de praga foliar, alguns poucos valores imputados apresentaram-se um pouco mais afastado dos verdadeiros correspondentes em todos os procedimentos de imputações analisados. Para o conjunto de dados de Acácia, os resultados gerados pelo procedimento EM-GAMMI demonstraram eficiência da técnica de imputação. Portanto, neste trabalho de tese foi possível concluir que o procedimento de imputação EM-GAMMI, dado seu bom desempenho, é um método preferível como técnica de imputação de dados ao método EM-AMMI, útil para contornar o problema do desbalanceamento em experimentos estatísticos voltados aos estudos da interação genótipo por ambiente.

Referências

Acorsi, C., Guedes, T., Coan, M., Pinto, R., Scapim, C., Pacheco, C., Guimarães, P. d. O., e Casela, C. (2016). Applying the generalized additive main effects and multiplicative interaction model to analysis of maize genotypes resistant to grey leaf spot. Embrapa Tabuleiros Costeiros-Artigo em periódico indexado (ALICE).

Acorsi, C. R. L. (2010). Análise da estabilidade e adaptabilidade de genótipos de milho na resistência a doenças por meio dos modelos Gammi. PhD thesis, Universidade Estadual de Maringá.

Agresti, A. (2003). Categorical data analysis. 2nd ed. Hoboken: John Wiley & Sons. 710 p.

Arciniegas-Alarcón, S. (2015). Imputação de dados em experimentos multiambientais: novos algoritmos utilizando a decomposição por valores singulares. PhD thesis, Universidade de São Paulo.

Arciniegas-Alarcón, S. e Dias, C. T. d. S. (2009). Data imputation in trials with genotype by environment interaction: an application on cotton data. Biometric Brazilian Journal, 27:125–138.

Arciniegas-Alarcón, S., Dias, C. T. d. S., e García-Peña, M. (2014). Imputação múltipla livre de distribuição em tabelas incompletas de dupla entrada. Pesquisa Agropecuária Brasileira, 49(9):683–691.

Arciniegas-Alarcón, S., García-Peña, M., Dias, C. T. d. S., e Krzanowski, W. J. (2010). An alternative methodology for imputing missing data in trials with genotype-by-environment interaction. Biometrical Letters, 47(1):1–14.

Bergamo, G. C., Dias, C. T. d. S., e Krzanowski, W. J. (2008). Distribution-free multiple imputation in an interaction matrix through singular value decomposition. Scientia Agricola, 65(4):422–427.

Ching, W., Li, L., Tsing, N., Tai, C., Ng, T., Wong, A., e Cheng, K. (2010). A weighted local least squares imputation method for missing value estimation in microarray gene expression data. International journal of data mining and bioinformatics, 4(3):331–347.

Collett, D. (2002). Modelling binary data. 2 ed. Boca Raton: Chapman & Hall; CRC Press, 387p.

Cordeiro, G. M. e Demétrio, C. G. (2008). Modelos lineares generalizados e extensões. Piracicaba: USP.

Cordeiro, G. M. e Paula, G. A. (1989). Modelos de regressão para análise de dados univariados. IMPA.

Dias, A. J. R. e Albieri, S. (2016). Uso de imputação em pesquisas domiciliares. Anais, pages 11–26.

DIAS, C. d. S. (2005). Métodos para escolha de componentes em modelo de efeito principal aditivo e interação multiplicativa (AMMI). PhD thesis, Tese (Livre Docência)–Escola Superior de Agricultura Luiz de Queiroz, Piracicaba.

Duarte, J. e Vencovsky, R. (1999). Interação genótipos x ambientes: uma introdução à análise ammi. Ribeirão Preto: Sociedade Brasileira de Genética.

Enders, C. K. (2010). Applied missing data analysis. Guilford press, 382p.

García-Peña, M., Arciniegas-Alarcón, S., e Barbin, D. (2014). Imputação de dados climáticos utilizando a decomposição por valores singulares: uma comparação empírica. Revista Brasileira de Meteorologia, 29(4):527–536.

GAUCH, H. (1988). Model secretion and validation for yield trials with interaction. Biometrics, 44(3):705–715.

Gauch, H. e Zobel, R. W. (1990). Imputing missing yield trial data. Theoretical and Applied Genetics, 79(6):753–761.

Hadi, A. F., Mattjik, A., e Sumertajaya, I. (2010). Generalized ammi models for assessing the endurance of soybean to leaf pest. Jurnal Ilmu Dasar, 11(2):151–159.

Hadi, A. F. e Sa'diyah, H. (2016). On the development of statistical modeling in plant breeding: An approach of row-column interaction models (rcim) for generalized ammi models with deviance analysis. Agriculture and Agricultural Science Procedia, 9:134–145.

Krzanowski, W. (1988). Missing value imputation in multivariate data using the singular value decomposition of a matrix. Biometrical letters, 25(1-2):31–39.

Little, R. J. e Rubin, D. B. (2002). Statistical analysis with missing data. New York: John Wiley. 381p .

McCullagh, P. e Nelder, J. A. (1989). Generalized Linear Models. 2nd Edition. Chapman and Hall.

Myers, R. H. e Montgomery, D. C. (1997). A tutorial on generalized linear models. Journal of Quality Technology, 29(3):274–291.

Nelder, J. A. e Wedderburn, R. W. (1972). Generalized linear models. Journal of the Royal Statistical Society: Series A (General), 135(3):370–384.

Paderewski, J. e Rodrigues, P. C. (2014). The usefulness of em-ammi to study the influence of missing data pattern and application to polish post-registration winter wheat data. Australian Journal of Crop Science, 8(4):640–645.

Paula, G. A. (2004). Modelos de regressão: com apoio computacional. IME-USP São Paulo.

Paulino, C. D. e da Motta Singer, J. (2006). Análise de dados categorizados. Editora Blucher.

Perry, P. O. (2009). Cross-validation for unsupervised learning. PhD thesis, Stanford University, 153p.

Piepho, H.-P. (1995). Methods for estimating missing genotype-location combinations in multilocation trials-an empirical comparison. Informatik Biometrie und Epidemiologie in Medizin und Biologie, 26(4):335–349.

R Core Team (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Robins, J. M., Rotnitzky, A., e Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. Journal of the American statistical Association, 89(427):846–866.

Rubin, D. B. (1987). Multiple imputation for survey nonresponse. New York: John Wiley & Sons. 320 p.

Schafer, J. L. (1999). Multiple imputation: a primer. Statistical methods in medical research, 8(1):3–15.

Spitti, A. M. D. S., Carbonell, S. A. M., Dias, C. T. d. S., Sabino, L. G., Carvalho, C. R. L., e Chiorato, A. F. (2019). Genótipos de feijoeiro carioca para tolerância ao escurecimento de grão pelos métodos natural e acelerado. Ciência e Agrotecnologia, 43.

Turner, H. e Firth, D. (2018). Generalized nonlinear models in r: An overview of the gnm package2015. R package version, page 61.

Van Eeuwijk, F. A. (1995). Multiplicative interaction in generalized linear models. Biometrics, pages 1017–1032.

Yan, W. (2013). Biplot analysis of incomplete two-way data. Crop Science, 53(1):48–57.

Zhang, P. (2003). Multiple imputation: theory and method. International Statistical Review, 71(3):581–592.

3 IMPUTAÇÃO MÚLTIPLA IMGAMMI EM EXPERIMNTOS MULTIAMBIENTAIS DESBALANCEADOS

Resumo

Dados ausentes são comuns em experimetos multiambientais por mais bem planejados que sejam, por isso, o uso de métodos de análises apropriados é essencial para reduzir o impacto gerado pela perda de informações. A imputação de dados é uma das técnicas comumente usada para contornar o problema das ausências, estima os dados ausentes por valores plausíveis e posteriormente as análises são realizadas sobre os dados completados. O presente trabalho tem por objetivo propor um novo método de imputação múltipla, para dados provenientes de experimentos multiambientais, resultante da proposta dos resíduos simples do modelo de regressão linear. Deste modo, modificações no algoritmo de imputação simples EM-AMMI foram realizadas, de forma a comportar o modelo de efeitos principais aditivos e interação multiplicativa generalizado GAMMI. A qualidade do método de imputação múltipla foi avaliada por meio das distribuições da estatística geral de acurácia Tacc e da raiz normalizada do erro quadrático médio (NRMSE). Para tal, simulações de retiradas aleatória de valores nos níveis de 10%, 20%, 30% e até 40% foram geradas a partir de dois conjuntos de dados reais e as imputações correspondentes obtidas. Os resultados da medida geral acurácia e da NRMSE, pelos seus baixos valores obtidos em relação ao método proposto, servem de evidências da melhor qualidade do algoritmo de imputação múltipla IMGAMMI proposto.
Palavras-chave: Algoritmo IMGAMMI; Experimentos multiambientais; Imputação múltipla; Modelo GAMMI .

3.1 Introdução

Os experimentos multiambientais são planejados para serem balanceados, entretanto é comum a ocorrência de valores ausentes, seja por falta de controle, erros humanos ou por imposições naturais do meio em que os experimentos se encontram, como excesso de chuvas, ataque de pragas, invasão de animais e outros (Bergamo, 2007; Yan, 2013; Rodrigues et al., 2011). As ausências produzem nos experimentos seu desbalanceamento e como consequência não podem ser analisadas diretamente por métodos de análises tradicionais eficientes. Um exemplo bem comum é aquele em que cultivares são estudadas em diferentes ambientes e a variável resposta é a média das repetições de cada combinação dos níveis dos fatores, experimentos como este, os modelos de efeitos principais aditivos e interação multiplicativa são a melhor abordagem de análise, no entanto, na presença de dados ausentes é inadmissível sua aplicabilidade (Gauch e Zobel, 1990; Yan, 2013).

Várias são as estratégias empregadas para resolver o problema das ausências de dados, que comumente ocorrem nos diversos tipos de experimentos multiambientais, como exemplo, tem-se a eliminação do conjunto de dados as entradas linhas ou colunas que apresentam valores ausentes, obtendo um subconjunto balanceado; preenchimento dos valores ausentes por meio das médias ambientais (entradas colunas) ou preenchimento por meio de estimativas obtidas por algum método, como os modelos lineares ou modelos mistos multiplicativo. Cada um desses procedimentos pode ser utilizado, mas nenhum deles é simples e ou totalmente eficaz. O primeiro produz ainda mais perdas, visto que na obtenção do subconjunto completo tende a eliminar outros valores, diminuindo drasticamente a amostra, o que pode resultar em desvios de padrões; o segundo pode não ser adequado, pois podem ocorrer muitos valores ausentes e o terceiro é muito complicado, envolve diversas etapas (Yan, 2013).

Alguns Trabalhos bem aceitos para preencher as ausências em experimentos mutiambientais são os métodos de imputações que fazem uso da decomposição em valores singulares (DVS) de uma matriz, como o algoritmo EM-AMMI apresentado por Gauch e Zobel (1990), em que os autores introduziram o

modelo de efeitos principais aditivos e interação multiplicativa (AMMI) no algoritmo EM (Esperança-Maximização) para realizar imputações. Neste algoritmo, os melhores resultados são alcançados com a inclusão de poucos termos multiplicativos no modelo AMMI (Piepho, 1995; Arciniegas-Alarcón e Dias, 2009; Paderewski e Rodrigues, 2014). Ainda, o algoritmo EM+DVS apresentado por Perry (2009), o método de imputação múltipla livre de distribuição (IMLD) de Bergamo et al. (2008), um método sem qualquer restrição quanto ao padrão e mecanismo de ausência de dados e livre de suposição sobre a distribuição ou estrutura dos dados, o método de imputação Biplot descrito por Yan (2013) e entre outros de igual destaque que usam a DVS.

A imputação é o preenchimento dos dados ausentes por valores plausíveis para posterior análise. Ela é simples, quando os dados ausentes são imputados uma única vez, mas pelo fato de ocorrer uma única vez, não se tem como quantificar as incertezas associadas as imputações, o que pode ser uma limitação da imputação simples (Enders, 2010; Bergamo, 2007), ou múltipla (Rubin, 1978, 1987). Na imputação múltipla são imputados *m valores para cada valor ausente, gerando m conjuntos de dados com valores imputados. Em geral, a imputação múltipla (IM) consiste de três etapas: imputação dos valores ausentes, análise dos m conjuntos de dados gerados e combinação dos resultados gerados nas m análises (Zhang, 2003). Na IM, as incertezas das imputações são incorporadas aos resultados, fazendo com que a IM seja mais atrativa que a imputação simples (Bergamo et al., 2008).*

Apesar da existência à décadas de métodos que tratam de dados ausentes, o assunto ainda é bastante delicado, levando muitos pesquisadores a não utilizarem métodos adequados em suas análises por falta de conhecimento na maioria das situações, usam abordagens simples de eliminação ou substituição (Peugh e Enders, 2004). Resultados semelhantes aos de Peugh e Enders (2004) foram apresentados por Rousseau et al. (2012), que em mais de um terço dos trabalhos revisados não foi constatado indicação nenhuma sobre dados ausentes, ainda, metade dos trabalhos em que os valores ausentes foram reportados, o método adotado não estava entendível e entre os que estavam corretamente, a maioria realizava a simples eliminação das observações. Para o autor, pesquisadores fazem uso desses métodos por serem padrão dos pacotes estatísticos.

Procedimentos adequados de imputação são bem mais vantajosos que a simples eliminação das unidades ausentes, porque manter toda a amostra pode ajudar a evitar aumento de erros resultante da diminuição do tamanho amostral, dados completados podem ser analisados por métodos clássicos eficientes e que estão disponíveis em programas estatísticos usuais. Finalmente, quando os dados são analisados por vários indivíduos, imputar uma vez, antes de todas as análises, garante a unicidade do conjunto, possibilitando comparações dos resultados. Por outro lado, a imputação pode não ser bem implementada, alguns métodos podem apresentar deficiências, o que pode ser uma desvantagem (Schafer e Graham, 2002).

Assim, apresentado alguns aspectos literários sobre imputação de dados em experimentos multi-ambientais, tem-se que o objetivo desse capítulo é propor um algoritmo de imputação múltipla a partir de uma extensão do método EM-AMMI, uma combinação do modelo de efeitos principais aditivos e interação multiplicativa generalizado com o algoritmo EM.

3.2 Materiais e métodos

3.2.1 Algoritmo de imputação EM-GAMMI

O processo para gerar imputação simples por meio do algorítimo EM-GAMMI, a partir de uma Matriz $\mathbf{X}$ *de dimensão* $(g \times a)$ *com elementos* $[x_{ij}]$ $(i=1,...,g;\ j=1,...,a)$, *em que alguns desses elementos estão ausentes, os* $[x_{ij}^{aus}]$, *é como descrito nos passos a seguir.*

Passo 1 - Os elementos ausentes $[x_{ij}^{aus}]$ *de* $\mathbf{X}$ *são inicialmente estimados pela média geral dos valores observados mais média da linha i (efeito principal de linha) mais média da coluna j (efeito*

principal de coluna) , obtendo-se uma matriz $\mathbf{X}$ *completa. Também é possível o preenchimento inicial por um valor arbitrário.*

Passo 2 - Um MLG particular com função de ligação especifica é definido, então os parâmetros do modelo GAMMI são estimados. Considera-se as entradas colunas de $\mathbf{X}$ *completa como efeito do fator ambiente e as entradas linhas como efeito do fator genótipo para o ajuste.*

Passo 3 - As médias ajustadas são calculadas com base no modelo GAMMI com k termos multiplicativos. Dependendo do número de termos multiplicativos utilizado, o método de imputação pode ser nomeado EM-GAMMI-0, EM-GAMMI-1, EM-GAMMI-2,..., EM-GAMMI-K.

Passo 4 - Os valores ausentes (x_{ij}^{aus}) *em* $\mathbf{X}$ *são preenchidos (imputados) pelas estimativas EM-GAMMI apropriadas. Como a relação entre a E(Y) e o preditor linear* η *não ocorre de forma direta, são unidos pela função de ligação, os valores preditos são retornados a escala dos dados mediante* $g^{-1}(\eta_i)$.

Passo 5 - Se a alteração máxima (distância Chebyshev) entre dois vetores de valores ausentes, estimativas em etapas de iteração sucessivas for maior que a precisão assumida, as etapas de 2 a 5 serão repetidas. Caso contrário, o algoritmo para.

3.2.2 Imputação múltipla (IMGAMMI) com uso do resíduo simples da regressão linear

Esta proposta de imputação múltipla segue a fundamentação teórica dos trabalhos de Arciniegas-Alarcón et al. (2014); Srivastava e Dolatabadi (2009). Arciniegas-Alarcón et al. (2014) realizaram imputação múltipla a partir do método de "imputação biplot", com uso do resíduo simples do modelo de regressão linear, $\boldsymbol{Y} = \boldsymbol{Q\beta} + \boldsymbol{E}$, *em que* $\mathbf{Y}$ $(n \times 1)$ *é o vetor que representa a variável resposta;* $\boldsymbol{Q}$ $(n \times p)$ *é a matriz de delineamento;* $\boldsymbol{\beta}$ $(p \times 1)$ *é o vetor de parâmetros da regressão e* $\boldsymbol{E}$ $(n \times 1)$ *é o vetor dos erros aleatórios. Segundo os autores, as ausências somente podem ocorrer no vetor* $\mathbf{Y}$, *as variáveis explicativas a compor o modelo devem ser completas. Assim, e seguindo a notação de Arciniegas-Alarcón et al. (2014), o modelo de regressão linear pode ser reescrito como* $(\boldsymbol{Y}_0/\boldsymbol{Y}_A) = (\boldsymbol{Q}_0/\boldsymbol{Q}_A)\boldsymbol{\beta} + \boldsymbol{E}$, *em que* $\boldsymbol{Y}_0$ $(n_1 \times 1)$ *corresponde ao subvetor dos* n_1 *dados observados,* $\boldsymbol{Y}_A$ $(n_0 \times 1)$ *ao subvetor que contém* n_0 *valores ausentes,* $\boldsymbol{Q}_0$ $(n_1 \times p)$ *a submatriz dos* n_1 *dados observados e* $\boldsymbol{Q}_A$ $(n_0 \times p)$ *a submatriz dos* n_0 *dados ausentes, de modo que* $n_0 + n_1 = n$. *Então, a obtenção da IM se dá mediante ajuste do seguinte modelo:* $\widehat{\mathbf{Y}}_{\mathbf{At}} = \mathbf{Q_A}(\mathbf{Q_0^T Q_0})^{-\mathbf{1}}\mathbf{Q_0^T Y_0} + \mathbf{E_t}$, *em que* $\mathbf{t} = \mathbf{1}, \ldots, \mathbf{M}$, *são as* $\mathbf{M}$ *imputações para cada dado ausente; e* $\mathbf{E_t}$ *é a* $\mathbf{t}$*-ésima amostra aleatória com reposição de tamanho* n_0 *obtida do vetor de resíduos* $\mathbf{e} = \left(\frac{\mathbf{n_1}}{\mathbf{n_1 - p}}\right)^{\mathbf{0,5}}(\mathbf{Y_0} - \mathbf{Q_0 b_1})$, *em que* $\mathbf{b_1} = (\mathbf{Q_0^T Q_0})^{-\mathbf{1}}\mathbf{Q_0^T Y_0}$ *é a estimativa de mínimos quadrados de* $\boldsymbol{\beta}$, *baseada nos dados observados.*

A proposta de IM com uso do algoritmo de imputação EM-GAMMI e seguindo o entusiasmo de utilizar o resíduo simples é como segue: o método EM-GAMMI fornece no final do seu processo, uma matriz $\mathbf{X^c}$ *completada, cujos elementos são os valores imputados para os respectivos valores ausentes e também as estimativas para os valores observados. O passo seguinte consiste na obtenção da matriz de resíduos simples a partir dos dados observados, por meio da diferença entre a matriz original e a matriz que contém as estimativas para os valores observados, isto é,* $\widehat{\boldsymbol{\epsilon}} = \mathbf{X} - \mathbf{X}^c$. *Como o resíduo é obtido somente para os valores observados, a matriz* $\widehat{\boldsymbol{\epsilon}}$ *de dimensão* $(n \times p)$ *é incompleta, pois somente podem ser obtidos resíduos para os (np-na) dados observados. Posteriormente, e a partir da matriz* $\widehat{\boldsymbol{\epsilon}}$ *de resíduos, são construídas t matrizes diferentes* Ω_t *de dimensão* $(n \times p)$ *e com* $t = 1, \ldots, M$ *da forma seguinte: cada elemento à compor* Ω_t *é selecionado aleatoriamente com reposição da matriz* $\widehat{\boldsymbol{\epsilon}}$. *O processo da seleção aleatória com reposição é repetido sobre* $\widehat{\boldsymbol{\epsilon}}$, M *vezes, produzindo M matrizes, isto é,* $\Omega_1, \Omega_2, \ldots, \Omega_M$. *Uma vez obtida* Ω_t, *o passo seguinte é realizar a imputação múltipla, isto é feito ao se substituir os elementos ausentes* $[x_{ij}^{aus}]$ *da matriz* $\mathbf{X}$, *pelos valores correspondentes de cada uma das t matrizes que são construídas por* $\mathbf{X}^c + \Omega_t$, *assim, o processo da IM fornece* $\mathbf{X}^c + \Omega_1$, $\mathbf{X}^c + \Omega_2$, $\ldots, \mathbf{X}^c + \Omega_M$ *matrizes completadas. Após as imputações terem sido obtidas, as t matrizes completadas (observados e imputados) são combinadas*

pela média das t matrizes completadas, dando origem uma única matriz, então, os elementos ausentes em $\mathbf{X}$ *original são imputados com as correspondentes médias obtidas.*

Neste trabalho, utilizou-se t=5, número de imputações múltiplas, pois de acordo com Rubin (1996), $t = 5$ *imputações são suficientes para fazer inferências válidas. Para Van Buuren (2018), t=5 permite boa qualidade ao método, sendo improvável que conclusões importantes sejam alteradas substancialmente se o limite t for maior que 5. Assim, seguindo a proposta Srivastava e Dolatabadi (2009), foi possível obter o processo de imputação múltipla por meio de um modelo multiplicativo generalizado, o qual foi denominado de imputação múltipla GAMMI (IMGAMMI).*

3.2.3 Descrição dos dados usados na pesquisa

Para avaliar o procedimento de IM, foram considerados dois conjuntos de dados reais, completos e provenientes de experimentos com interação genótipo por ambiente. O primeiro conjunto de dados é um delineamento em blocos ao acaso, estudo da resistência de soja à praga foliar, publicado por Hadi et al. (2010). No experimento, foram utilizados quatro genótipos de soja resultantes de híbridos (Wilis, IAC-100, IAC-80 e W-80) e, avaliados aos 14 dias após o plantio, a contagem de pragas foliar encontradas por planta. Na contagem, cinco tipos de pragas foliares foram classificadas nas variedades (genótipos de soja). A Tabela 3.1 apresenta a média da população dos cinco tipos de pragas foliares em quatro genótipos de soja. A escolha deste conjunto de dados foi particular, pois as respostas médias das repetições são expressas em escala intervalar e analisadas pela metodologia GAMMI (Hadi et al., 2010). Deste modo, foi possível fazer uso dos modelos AMMI e GAMMI para posterior uso dos algoritmos de imputações. Este conjunto foi denominado de dados de praga foliar para fins de referência.

O segundo conjunto de dados utilizado, é parte de um estudo, obtido de um delineamento em blocos aleatorizados cedido pelos pesquisadores Spitti et al. (2019). No estudo, foram usados 19 genótipos de feijoeiro observados em seis ambientes. Os genótipos foram avaliados quanto a cor do tegumento dos grãos em função do valor de luminosidade (L) em relação ao método de crescimento em condições de prateleira. A variável resposta é a tolerância (resistência) do genótipo a perda de pigmentos, ou seja, mudança gradativa da coloração dos grãos aos 60 dias. A Tabela 3.2 mostra os valores médios dos genótipos por ambiente obtidos das seis regiões consideradas no estudo. Este conjunto foi denominado de dados de Acácia para efeito de referência.

Tabela 3.1 – Média da população de cinco tipos de pragas foliares em quatro genótipos de soja

Genótipos	Tipos de praga foliar				
	Bemissia	Emproosca	Agronyza	Lamprosema	Longitarsaus
IAC-100	0,50	1,75	2,25	0,50	1,75
IAC-80	3,00	2,75	1,00	1,75	3,25
W-80	3,50	4,00	1,25	2,00	2,00
Wilis	4,00	3,00	1,00	1,75	4,00

Fonte: Hadi et al. (2010)

3.2.4 Procedimento de simulação com base nos dados reais

Para os dois conjunto de dados experimentais, usados no trabalho, foram realizadas simulações de retiradas aleatórias nas porcentagens de 10%, 20% e 30% para os dados de praga foliar e de 10%, 20%, 30% e 40% para os dados de Acácia, pois conforme Yan (2013), o número de dados ausentes em experimentos de interação genótipo por ambiente é menor que 40%. Este processo foi repetido 100 vezes

Tabela 3.2 – Média de genótipos de feijoeiros avaliados quanto a cor do tegumento de grãos em função do valor de luminosidade(L)

Genótipos	Regiões					
	R1	R2	R3	R4	R5	R6
BRS Pérola	0,5041	0,4727	0,5036	0,4497	0,4840	0,4957
CHC 01-175-1	0,4987	0,4648	0,5105	0,4610	0,4747	0,5013
CNFC 11-948	0,5068	0,4703	0,5023	0,4618	0,5048	0,5110
CNFC 11-954	0,5013	0,4585	0,4867	0,4708	0,4992	0,4961
Gen 4-1F-19P	0,5263	0,5000	0,4909	0,4892	0,5241	0,5245
Gen 12-2F-67	0,5178	0,4681	0,5021	0,4790	0,5098	0,5184
Gen 20-4F-129	0,5122	0,4847	0,4844	0,4494	0,4987	0,5343
Gen 45-2F-293P	0,5244	0,4922	0,5083	0,4792	0,5326	0,5493
Gen 78-1A-59	0,5078	0,4907	0,4950	0,4717	0,5168	0,5291
Gen 86-12A-122	0,5055	0,4776	0,4907	0,4501	0,4878	0,5215
Gen 90-4A-160	0,5106	0,4692	0,4993	0,4588	0,5002	0,5228
Gen 104-1A-291	0,5314	0,4901	0,5109	0,4677	0,5197	0,5304
Gen 106-4A-317	0,5107	0,4882	0,4999	0,4497	0,5016	0,5346
Gen 106-6A-319	0,5195	0,4794	0,5014	0,4856	0,5143	0,5226
Gen 107-14A-336	0,5256	0,4777	0,5145	0,4563	0,5348	0,5552
Gen 125-10A-510	0,5123	0,4670	0,5103	0,4756	0,4987	0,5183
IAC Milênio	0,5219	0,4803	0,5063	0,4873	0,5017	0,5111
IAC Sintonia	0,5028	0,4682	0,4821	0,4588	0,4899	0,5276
LP 11-363	0,5394	0,4810	0,5201	0,4703	0,5018	0,5144

Fonte: Spitti et al. (2019)

para cada porcentagem de retirada em cada um dos dois conjuntos de valores, obtendo 300 matrizes diferentes para o conjunto de dados de praga foliar e 400 matrizes diferentes para o conjunto de dados de Acácia, totalizando 700 matrizes com valores ausentes simulados. Em seguida, para cada uma das 700 matrizes com valores ausentes simulados, foram feitas as imputações.

As etapas, simulações e predições, foram realizadas por meio de rotinas computacionais desenvolvidas e implementadas para o programa R (R Core Team, 2020). Destaca-se no algoritmo desenvolvido, uso da função gnm para ajuste do modelo GAMMI com até dois termos multiplicativos. Para os dados de praga foliar, foram usados os modelos GAMMI Poisson e GAMMI Gaussiano com respectivas funções de ligações logarítmica e identidade. A escolha pelo modelo GAMMI Poisson foi decorrente do estudo de Hadi et al. (2010). Para os dados de Acácia, utilizou-se o modelo GAMMI Binomial com função de ligação logística, pois os dados representam uma proporção. As imputações foram obtidas pelos algoritmos EM-GAMMI, IMGAMMI, EM-AMMI (com uso da função EM-AMMI) e EM+DVS (em que se utilizou a função impute.svd). As simulações de retiradas aleatórias de valores ou geração dos dados ausentes, supondo o mecanismo de ausência aleatória - MAR (Missing at Random), deram-se mediante uso da função SimIm do pacote multivariado ImputeR.

O modelo GAMMI é um dos melhores modelos para análise de experimentos com interação genótipo por ambiente, casos em que se constata violações das suposições do modelo de ANOVA, ou quando a resposta é uma contagem, uma proporção, entre outras. Por esta razão, para cada uma das matrizes com valores ausentes, obtidas por simulações a partir do conjunto original, foram geradas imputações a partir da junção do algoritmo EM com o modelo GAMMI com até $\mathbf{k}$ *termos multiplicativos (*$\mathbf{k}$*=0,1,2), com uso do resíduo simples do modelo de regressção linear. Para o conjunto de praga foliar, foi assumido para os algoritmos IMGAMMI e IM-GAMMI ou (IM-AMMI) os modelos Poisson e Gaussiano com funções de ligações logarítmica e identidade respectivamente. Já, para o conjunto de dados Acácia, o modelo Binomial com função de ligação logística foi utilizado pelo IMGAMMI, como também, pelo EM-GAMMI.*

3.2.5 Critérios usados para avaliação do método

Como critérios de avaliação, foram usadas as estatísticas: raiz do erro quadrático médio padronizado - NRMSE, variância entre imputações - V_E, viés quadrático médio - VQM e a medida total (ou geral) de acurácia (Tacc).

3.2.5.1 NRMSE

*Pelo critério da NRMSE (Ching et al., 2010), o algoritmo é comparado por meio das médias ajustadas, ou seja, os valores imputados são comparados com os correspondentes valores observados no conjunto dos dados originais por meio da equação (*3.1*). É considerado o método com melhor desempenho, aquele que apresentar menor valor da estatística NRMSE.*

$$NRMSE = \frac{\sqrt{média(\mathbf{x}_{imp} - \mathbf{x}_{obs})^2}}{\mathbf{s}(\mathbf{x}_{obs})} \tag{3.1}$$

em que $\mathbf{x}_{imp}$ e $\mathbf{x}_{obs}$ são vetores contendo os respectivos valores médios imputados e os valores verdadeiros das observações ausentes simuladas e $\mathbf{s}(\mathbf{x}_{obs})$ é o desvio padrão dos valores contidos no vetor $\mathbf{x}_{obs}$. Quanto menor for o valor da estatística NRMSE, melhor será o desempenho do método de imputação.

3.2.5.2 Estatística geral de acurácia - Tacc

Conforme Bergamo (2007), a medida geral da acurácia Tacc é uma medida de exatidão ou acurácia, usada para avaliar um particular procedimento de IM e, segundo o autor, pode ser decomposta em dois componentes Tacc=$V_E + VQM$. O primeiro, V_E representa a variância entre imputações, em geral, valores pequenos da V_E indicam boa precisão do método. Já, o VQM representa o viés quadrático médio entre a média das imputações ($\overline{Y}$) e o valor original retirado no estudo de simulação (VO). O método de imputação múltipla apresentará bom desempenhose se os valores de VQM forem pequenos . As estatísticas V_E e VQM são apresentadas em 3.2.

$$V_E = \frac{1}{na}\sum_{l=1}^{na}\left[\frac{\sum_{m=1}^{M}(\widehat{y}_{ij(m)} - \overline{Y}_l)^2}{M-1}\right] \quad e \quad VQM = \frac{1}{na}\sum_{l=1}^{na} M\frac{(\overline{Y}_l - VO_l)^2}{M-1} \tag{3.2}$$

em que, para cada posição (i, j) de retiradas aleatórias na matriz de dados são realizadas $\mathbf{M}$ imputações; VO_l valor original retirado aleatoriamente; o índice l representa a posição do valor retirado correspondentes a posição (i, j) com $l = 1, \ldots, na$; na é o número total de valores retirados; $\hat{y}_{ij}$ é o valor imputado para o respectivo valor VO_l e $\overline{Y}_l$ é a média das imputações para a posição l.

De forma geral, as estatísticas NRMSE e Tacc, usadas para comparar e avaliar, dão uma ótima visão sobre o desempenho do método em análise. Assim, neste trabalho, considerou-se como um bom método de imputação de dados aquele que apresentou conjuntamente, dentre os comparados, valor médio/mediano para a distribuição da NRMSE pequeno e também, valores pequenos das distribuições V_E, VQM e Tacc, visto que as imputações foram obtidas com base em 100 matrizes simuladas para diferentes níveis de retiradas aleatórias de valores.

3.3 Resultados e discussão

3.3.1 Dados de praga foliar

A Tabela 3.3 apresenta as médias e medianas da NRMSE para o conjunto de dados de praga foliar, mostrando os métodos de imputação múltipla proposto (IMGAMMI), o método EM+DVS e o

método EM-AMMI para cada nível de porcentagem de retiradas. Nesta, pelo critério da NRMSE utilizado, o método com melhor desempenho foi o IMGAMMI, independentemente do nível de retirada. Assim, para o nível de ausência de 10% de imputação, o procedimento com melhor desempenho foi o IMGAMMI0 (mediana=0,199); para o nível de 20%, foi o IMGAMMI0 que apresentou melhor desempenho (mediana=0,2076) e para o nível de 30%, foi o IMGAMMI0 (mediana=0,1956). Observa-se ainda, que o procedimento EM+DVS obteve melhores resultados em termos das NRMSE que o método clássico EM-AMMI com até dois termos multiplicativos para todos os níveis de retiradas

Tabela 3.3 – Média e mediana da distribuição da NRMSE, em que foram feitas retirada aleatória (10%, 20% e 30%) , do conjunto de dados de praga foliar

	10%		20%		30%	
Metodo	Media	Mediana	Media	Mediana	Media	Mediana
EM+DVS	0,925	0,971	1,1297	0,9518	1,0955	1,0006
EM-AMMI0	1,086	0,965	1,3770	1,0290	1,3025	1,1922
EM-AMMI1	2,168	1,462	2,3680	1,8920	2,1495	2,0158
EM-AMMI2	1,049	1,076	1,1930	1,0470	1,1218	1,0323
IMGAMMI0	0,216	0,199	0,2524	0,2076	0,2188	0,1956
IMGAMMI1	0,254	0,230	0,2454	0,2126	0,2230	0,2024
IMGAMMI2	0,227	0,221	0,2759	0,2341	0,2195	0,2046

A outra estatística usada para jugar o método foi a Tacc, e discutida mais adiante. A Tabela 3.4 mostra em termos de média e mediana, os valores da V_E e VQM. Nesta, verificou-se que os procedimentos de imputação múltipla IM-AMMI0, IM-AMMI1 e IM-AMMI2 forneceram a maior variância entre imputações (V_E), independentemente da percentagem de ausências para imputações, enquanto o algoritmo com menor variância entre imputações, foi o IMGAMMI0, para os níveis de retiradas de 10% e 20%, seguido do IMGAMMI2 para o nível de 30% de retiradas. No entanto, como complemento da análise V_E e de modo a tomar a melhor decisão sobre qual seria o procedimento com maior eficiência preditiva, é necessário que se analise o viés quadrático médio (VQM) como também a medida geral de acurácia Tacc.

Com relação ao VQM, os métodos com o menor viés de acordo com as percentagens de imputações adotadas foram: para ausências de 10% o IMGAMMI0, para ausências de 20% o IMGAMMI1 e para ausências de 30% o IMGAMMI2 (Tabela 3.4). Em todos os casos, os procedimentos de imputações com maiores valores de VQM foram IM-AMMI0, IM-AMMI1 e o IM-AMMI2. Por outro lado, o algoritmo IMGAMMI0, IMGAMMI1 e IMGAMMI2, por terem apresentados menores valores do VQM, permitiram que se atingisse a maior similaridade entre as imputações e seus respectivos valores originais, resultando em melhor precisão do método. Além disso, à medida que a percentagem de imputação fora aumentada, era esperado aumento nos valores do VQM para os procedimentos de imputações, o que não se constatou. Observou-se sim, pequeno aumento mediano do VQM para o procedimento IM-AMMI1 em relação ao IM-GAMMI0, o mesmo foi verificado para os procedimentos IMGAMMI1 em relação ao IMGAMMI0. Tal acontecimento, quando ocorre, pode ser justificado pela diminuição de valores geradas pelos níveis de retirada (diminuição da amostra), pois segundo Arciniegas-Alarcón et al. (2014), o erro da imputação tende a aumentar, visto que informações disponíveis na matriz de dados foram diminuídas pelos aumentos de percentagem de retiradas.

Para decidir qual o melhor método de imputação, a estatística geral de acurácia Tacc deve ser considerada. A estatística considera tanto a variância entre imputações quanto o viés quadrático médio (Tabela 3.4). Na Figura 3.1, tem-se a distribuição da Tacc em termos mediano para o procedimento IM-GAMMI0 (IMGA), IMGAMMI1 (IMGA1) e IMGAMMI2 (IMGA2) nos três níveis de retiradas. Nesta, o procedimento de imputação IMGAMMI0 apresentou menor valor da mediana para imputações com 10%

e 20% de retiradas aleatórias, seguido do procedimento IMGAMMI1, para imputações com 30% de retiradas. Assim, para 10% de retiradas aleatórias, as medianas da Tacc foram: 0,4131 para IMGAMMI0; 0,4758 para IMGAMMI1 e 0,4270 para IMGAMMI2, contra 0,4942 para IM-AMMI0; 0,4977 para IM-AMMI1 e 0,5305 para IM-AMMI2. Para percentagem de 20%, o procedimento IMGAMMI0 apresentou melhor desempenho, mediana = 0,3847. Para ausências de 30%, o procedimento IMGAMMI1 foi que apresentou melhor desempenho, mediana=0,3473. Deste modo, vale destacar o bom desempenho do método IMGAMMI, seja em comparação aos resultados dos métodos apresentados como também pelos baixos valores obtidos pelas estatísticas NRMSE e medida geral de acurácia (Tacc).

Tabela 3.4 – Média e mediana da variância combinada entre imputações (V_E) e do viés quadrático médio (VQM), correspondente aos níveis de retirada aleatória de dados (10%, 20% e 30%) do conjunto de dados de praga foliar

	10%		20%		30%	
Método	Média	Mediana	Média	Mediana	Média	Mediana
			V_E			
IM-AMMI0	0,4618	0,4064	0,3986	0,3628	0,3412	0,3425
IM-AMMI1	0,4231	0,4130	0,3835	0,3530	0,3341	0,3159
IM-AMMI2	0,4282	0,4031	0,4169	0,4249	0,3398	0,3140
IMGAMMI0	0,3557	0,3362	0,3256	0,3046	0,2927	0,2867
IMGAMMI1	0,4083	0,3784	0,3238	0,3098	0,2836	0,2710
IMGAMMI2	0,3777	0,3486	0,3310	0,3296	0,2968	0,2961
			VQM			
IM-AMMI0	0,1103	0,0646	0,1024	0,0807	0,0838	0,0687
IM-AMMI1	0,1064	0,0590	0,1022	0,0764	0,0771	0,0597
IM-AMMI2	0,1237	0,0850	0,0901	0,0796	0,0862	0,0740
IMGAMMI0	0,0769	0,0495	0,0814	0,0676	0,0730	0,0565
IMGAMMI1	0,1047	0,0662	0,0739	0,0631	0,0748	0,0591
IMGAMMI2	0,0776	0,0608	0,0947	0,0719	0,0697	0,0582

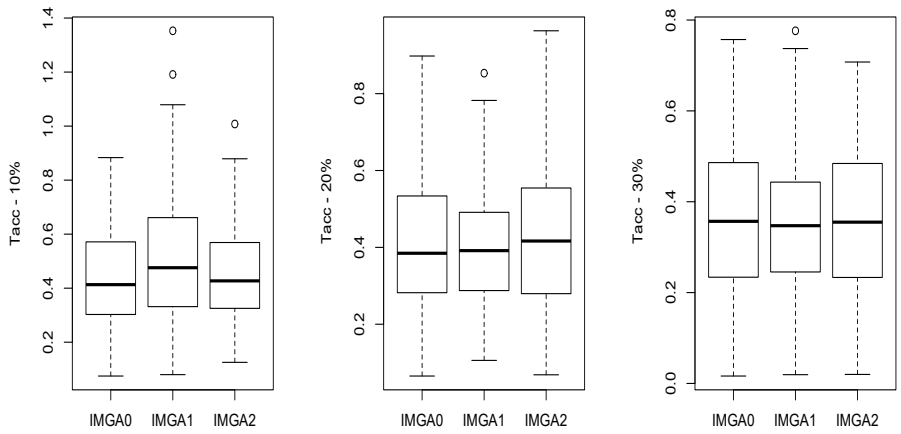

Figura 3.1 – Distribuição da medida de acurácia geral (Tacc), com uso dos métodos IMGAMMI0 (IMGA0), IMGAMMI1 (IMGA1) e IMGAMMI2 (IMGA2), para o conjunto de dados praga foliar nos níveis de 10%, 20% e 30% de retiradas

3.3.2 Segundo conjunto - dados Acácia

Para o conjunto de dados de Acácia, avaliado, o método de imputação IMGAMMI forneceu sempre o menor valor da estatística NRMSE, em termos médios e medianos, quando comparado com

método EM-GAMMI em todos os níveis de retiradas (Tabela 3.5). Os baixos valores obtidos da NRMSE sugerem melhores predições por parte do procedimento de imputação múltipla IMGAMMI, isto é, os valores imputados se aproximam aos observados correspondentes. Para ausências de 10%, os procedimentos IMGAMMI0, IMGAMMI1 e IMGAMMI2 apresentaram desempenho semelhantes, todavia, por facilidades computacionais e economia de parâmetros o IMGAMMI0 deve ser o escolhido. Para ausências de 20%, o IMGAMMI1 apresentou melhor desempenho (média=0,1725), com 30% de ausências, melhor desempenho do IMGAMMI2 e com 40% de ausências, o método com melhor desempenho foi o IMGAMMI2. Ainda, pode-se constatar para os procedimentos EM-GAMMI-0, EM-GAMMI-1 e EM-GAMMI-2 aumento da NRMSE à medida que se aumenta os níveis de retiradas, o mesmo pode ser observado também para os procedimentos IMGAMMI0, IMGAMMI01 e IMGAMMI02, em níveis bem mais moderados.

Tabela 3.5 – Média e mediana da distribuição da NRMSE, em que foram feitas retirada aleatória de 10%, 20% , 30% e 40%, do conjunto de dados de Acácia

	10%		20%		30%		40%	
Método	Média	Mediana	Média	Mediana	Média	Mediana	Média	Mediana
EM-GAMMI-0	0,5128	0,5026	0,5232	0,5122	0,5699	0,5523	0,6111	0,5999
EM-GAMMI-1	0,5057	0,4960	0,5287	0,5179	0,5863	0,5737	0,6357	0,6310
EM-GAMMI-2	0,5140	0,5021	0,5404	0,5336	0,6022	0,5809	0,6550	0,6512
IMGAMMI0	0,1664	0,1612	0,1741	0,1724	0,1785	0,1744	0,1738	0,1725
IMGAMMI1	0,1683	0,1702	0,1725	0,1726	0,1757	0,1758	0,1737	0,1709
IMGAMMI2	0,1671	0,1528	0,1743	0,1707	0,1738	0,1733	0,1714	0,1690

A Tabela 3.6, apresenta a variância entre imputações (V_E) e o viés quadrático médio (VQM), em termos de média e mediana das 100 matrizes submetidas aos procedimentos de imputações múltiplas IMGAMMI0, IMGAMMI1 e IMGAMMI2. Nesta, os métodos apresentaram variância V_E pequena com valores próximos, em todos os níveis de percentagem de imputação, pois com até cinco casas decimais os resultados médios e medianos da V_E foram de aproximadamente 0,00008. Com relação ao VQM, os procedimentos de imputações IMGAMMI0, IMGAMMI1 e IMGAMMI2, nos quatros níveis de ausências, apresentaram tendências muito pequenas, ou seja, valores de VQM muito próximo de zero, oscilando entre 0,0000194 à 0,0000207 em termos médios, o que indica excelente precisão por parte dos métodos. Vale destacar pequeno aumento no VQM, o que é esperado, pois à medida que se aumentou a percentagem de retiradas para imputações o tamanho da amostra diminuiu.

Tabela 3.6 – Média e mediana da variância combinada entre imputações (V_E) e do viés quadrático médio (VQM), correspondente aos níveis de retirada aleatória de dados (10%, 20%, 30% e 40%) do conjunto de dados Acácia

	10%		20%		30%		40%	
Método	Média	Mediana	Média	Mediana	Média	Mediana	Média	Mediana
				V_E				
IMGAMMI0	0,00008202	0,0000819	0,0000811	0,0000799	0,0000819	0,0000819	0,0000816	0,0000810
IMGAMMI1	0,00008282	0,0000830	0,0000817	0,0000796	0,0000793	0,0000786	0,0000822	0,0000831
IMGAMMI2	0,00008204	0,0000782	0,0000813	0,0000806	0,0000799	0,0000785	0,0000824	0,0000810
				VQM				
IMGAMMI0	0,0000194	0,0000171	0,0000204	0,0000197	0,0000210	0,0000198	0,0000207	0,0000204
IMGAMMI1	0,0000199	0,0000189	0,0000201	0,0000206	0,0000204	0,0000198	0,0000206	0,0000193
IMGAMMI2	0,0000195	0,0000183	0,0000205	0,0000199	0,0000199	0,0000197	0,0000201	0,0000188

A análise da V_E e do VQM deve ser complementada pela medida geral de acurácia Tacc, vista à decidir sobre qual seria o melhor método de imputação. Essa estatística leva em conta tanto a variância entre imputações quanto o viés quadrático médio. Na Figura 3.2, foram apresentadas as distribuições da estatística Tacc para os procedimentos IMGAMMI0 (IMGA0), IMGAMMI1(IMGA1) e IMGAMMI2 (IMGA2). Nesta, observou-se que os métodos apresentam distribuições aproximadamente simétricas em

torno da mediana para os níveis de imputações realizados. O método com menor valor para o parâmetro de centralidade mediana, com ausência de 10%, foi o IMGAMMI2, com ausências de 20% foi IMGAMMI2, com ausência de 30% foi o IMGAMMI1 e para ausências de 40% foi o IMGAMMI2. Por outro lado, vale observar que todos os procedimentos apresentaram valores pequenos da Tacc, próximos de zero, em todos os níveis de imputações realizadas (Tabela 3.7). Assim, o método IMGAMMI0 pode ser preferido, se considerado a ideia de economia de parâmetros e facilidades computacionais, pois dispensa a inclusão de termos multiplicativos.

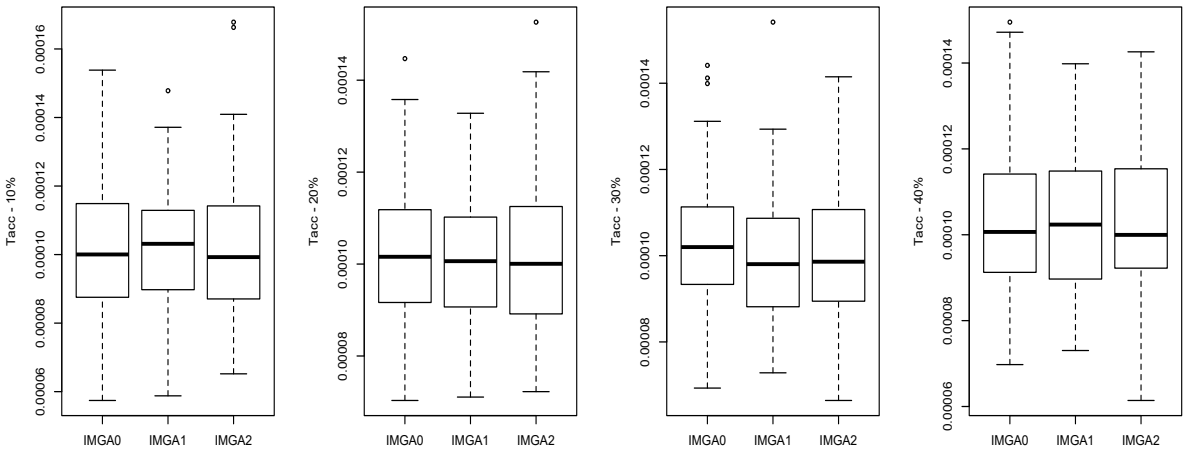

Figura 3.2 – Distribuição da medida de acurácia geral (Tacc), com uso dos métodos IMGAMMI0 (IMGA0), IMGAMMI1 (IMGA1) e IMGAMMI2 (IMGA2), para o conjunto de dados Acácia nos níveis de 10%, 20%, 30% e 40% de retiradas

Tabela 3.7 – Mediana da distribuição da Tacc para os níveis de retirada aleatória (10%, 20% , 30% e 40%), do conjunto de dados de Acácia

	Estatìstica Tacc			
Método	Mediana (10%)	Mediana (20%)	Mediana (30%)	Mediana (40%)
IMGAMMI0	0,0001000	0,0001016	0,0001020	0,000101
IMGAMMI1	0,0001031	0,0001006	0,0000980	0,000102
IMGAMMI2	0,0000993	0,0001001	0,0000986	0,000100

Por último, quando se analisou os resultados dos procedimentos para os dois conjuntos de dados multiambientais, foi possível constatar que o procedimento de imputação múltipla IMGAMMI apresentou melhores resultados que os procedimentos EM+DVS, EM-AMMI em termos de valores da NRMSE para o conjunto de praga foliar, bem como, apresentou resultados melhores para o conjunto de dados de Acácia, tanto em termos da NRMSE, que comparou os procedimentos EM-GAMMI0, EM-GAMMI1 e EM-GAMMI2 com os procedimentos IMGAMMI0, IMGAMMI1 e IMGAMMI2, quanto em termos da medida geral de acurácia Tacc, pelos seus baixos valores medianos obtidos. Conforme Rubin (1987), a IM é mais vantajosa que a imputação simples, permite aumento na eficiência das estimativas, permite fazer inferências válidas, refletindo a variabilidade adicional devido aos valores ausentes e permite comparar a sensibilidade das inferências obtidas por diferentes técnicas de imputação, simplesmente usando métodos de dados completos.

3.3.3 Aplicação - dados de praga foliar

Uma vez realizada a etapa de imputação, o passo seguinte é realizar as análises sobre o experimento, e geralmente, um modelo conveniente é usado para tal propósito. As Tabelas 3.8, 3.9 e 3.10 apresentam resultados da análise do desvio (Analysis of the Deviance - ANODEV), em que se utilizou modelos GAMMIs, para os níveis de retiradas de 10%, 20% e 30% respectivamente. Os valores verdadeiros retirados para o nível de 10% foram 1,25 e 2,00 com correspondentes 1,40 e 2,00 imputados, para o nível de 20%, os valores retirados foram 2,75; 3,00; 2,25 e 1,25 com imputados correspondentes 2,80; 3,00; 1,91 e 1,27 e para o nível de 30%, os valores retirados pelo processo de simulação foram 4,00; 3,00; 1,25; 1,00; 1,75 e 2,00 com imputados correspondentes 4,0; 3,0; 1,29; 0,65; 2,0; 2,0. Como esperado, para processos bem equilibrados, além do método reproduzir valores imputados próximos aos observados correspondentes, não se verificou alterações substanciais nas inferências para os três níveis de retiradas (Tabelas 3.8, 3.9 e 3.10) em comparação com os resultados apresentados por Hadi et al. (2010), significativos com até dois termos multiplicativos no modelo.

Tabela 3.8 – Análise do deviance para o conjunto de dados de praga foliar, após imputações pelo método IMGAMMI0, com retiradas aleatórias de 10%

F.V	G.L	Deviance	Deviance médio	Fc	valor de p
Ambiente	4	4,1067	1,0267	76,05	0,0132
Genotipos	3	2,8562	0,9521	70,52	0,0142
GAMMI1	6	3,6184	0,6031	44,67	0,0222
GAMMI2	4	0,9680	0,242	17,93	0,0542
Erro	2	0,0270	0,0135		
Total	19	11,5763	0,6093		

Tabela 3.9 – Análise do deviance para o conjunto de dados de praga foliar, após imputações pelo método IMGAMMI0, com retiradas aleatórias de 20%

F.V	G.L	Deviance	Deviance médio	Fc	valor de p
Ambiente	4	4,5929	1,1482	83,81	0,0120
Genotipos	3	3,2505	1,0835	79,088	0,0127
GAMMI1	6	2,9942	0,4990	36,426	0,0272
GAMMI2	4	0,9159	0,2289	16,714	0,0579
Erro	2	0,0274	0,0137		
Total	19	11,7809	0,620		

Tabela 3.10 – Análise do deviance para o conjunto de dados de praga foliar, após imputações pelo método IMGAMMI1, com retiradas aleatórias de 30%

F.V	G.L	Deviance	Deviance médio	Fc	valor de p
Ambiente	4	4,3151	1,0787	101,29	0,0099
Genotipos	3	2,8065	0,9355	87,84	0,0115
GAMMI1	6	4,0746	0,6791	63,77	0,0157
GAMMI2	4	1,0183	0,2545	23,90	0,0411
Erro	2	0,0213	0,0107		
Total	19	12.2358	0,6439		

Os resultados obtidos proporcionam alguns guias para pesquisas futuras relacionadas a dados ausentes. Por exemplo, usar de novos modelos GAMMIs para imputação em dados multiambientais com superdispersão, outros ou novos métodos de imputações podem ser tomados para fazer comparações com o método IMGAMMI, novos conjuntos de dados podem ser tomados para análises. Os métodos EM-AMMI-0, EM-AMMI-1, EM-AMMI-2, EM+DVS, apresentados na literatura, mostraram desempenho inferior ao IMGAMMI apresentado aqui. Visto que em estudos anteriores, como o de Arciniegas-Alarcón e Dias (2009), em que o método EM-AMMI1 apresentou melhor desempenho que a IMLD, já em Arciniegas-Alarcón et al. (2014), em um estudo comparativo, conclui sobre o bom desempenho dos métodos EM+DVS e EM-AMMI quando comparados a outros métodos de imputação. Tais indicativos e somados com os resultados apresentados aqui, são fortes constatações da boa qualidade do método IMGAMMI.

3.4 Conclusão

No presente trabalho, foi analisado uma abordagem estatística de imputação múltipla de dados para experimentos multiambientais e avaliada por meio das distribuições da estatística NRMSE e da medida geral de acurácia Tacc. O procedimento IMGAMMI apresentou melhores resultados como método de imputação, mostrou-se superior aos métodos EM-GAMMI, EM-AMMI e EM+DVS, em ambos conjuntos de dados usados no estudo. Portanto, neste trabalho de tese, foi possível concluir a favor da eficiência do procedimento de imputação múltipla IMGAMMI, o método mais eficiênte para realizar imputações, tanto em termos da NRMSE como também em termos da estatística geral de acurácia Tacc.

Referências

Arciniegas-Alarcón, S. e Dias, C. T. d. S. (2009). Data imputation in trials with genotype by environment interaction: an application on cotton data. Biometric Brazilian Journal, 27:125–138.

Arciniegas-Alarcón, S., Dias, C. T. d. S., e García-Peña, M. (2014). Imputação múltipla livre de distribuição em tabelas incompletas de dupla entrada. Pesquisa Agropecuária Brasileira, 49(9):683–691.

Bergamo, G. C. (2007). Imputaçao múltipla livre de distribuiçao utilizando a decomposiçao por valor singular em matriz de interaçao. PhD thesis, Universidade de São Paulo.

Bergamo, G. C., Dias, C. T. d. S., e Krzanowski, W. J. (2008). Distribution-free multiple imputation in an interaction matrix through singular value decomposition. Scientia Agricola, 65(4):422–427.

Ching, W., Li, L., Tsing, N., Tai, C., Ng, T., Wong, A., e Cheng, K. (2010). A weighted local least squares imputation method for missing value estimation in microarray gene expression data. International journal of data mining and bioinformatics, 4(3):331–347.

Enders, C. K. (2010). Applied missing data analysis. Guilford press, 382p.

Gauch, H. e Zobel, R. W. (1990). Imputing missing yield trial data. Theoretical and Applied Genetics, 79(6):753–761.

Hadi, A. F., Mattjik, A., e Sumertajaya, I. (2010). Generalized ammi models for assessing the endurance of soybean to leaf pest. Jurnal Ilmu Dasar, 11(2):151–159.

Paderewski, J. e Rodrigues, P. C. (2014). The usefulness of em-ammi to study the influence of missing data pattern and application to polish post-registration winter wheat data. Australian Journal of Crop Science, 8(4):640–645.

Perry, P. O. (2009). Cross-validation for unsupervised learning. PhD thesis, Stanford University, 153p.

Peugh, J. L. e Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. Review of educational research, 74(4):525–556.

Piepho, H.-P. (1995). Methods for estimating missing genotype-location combinations in multilocation trials-an empirical comparison. Informatik Biometrie und Epidemiologie in Medizin und Biologie, 26(4):335–349.

R Core Team (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rodrigues, P. C., Pereira, D. G. S., e Mexia, J. T. (2011). A comparison between joint regression analysis and the additive main and multiplicative interaction model: the robustness with increasing amounts of missing data. Scientia Agricola, 68(6):679–686.

Rousseau, M., Simon, M., Bertrand, R., e Hachey, K. (2012). Reporting missing data: a study of selected articles published from 2003–2007. Quality & Quantity, 46(5):1393–1406.

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In Proceedings of the survey research methods section of the American Statistical Association, volume 1, pages 20–34. American Statistical Association.

Rubin, D. B. (1987). Multiple imputation for survey nonresponse. New York: John Wiley & Sons. 320 p.

Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American statistical Association, 91(434):473–489.

Schafer, J. L. e Graham, J. W. (2002). Missing data: our view of the state of the art. Psychological methods, 7(2):147–177.

Spitti, A. M. D. S., Carbonell, S. A. M., Dias, C. T. d. S., Sabino, L. G., Carvalho, C. R. L., e Chiorato, A. F. (2019). Genótipos de feijoeiro carioca para tolerância ao escurecimento de grão pelos métodos natural e acelerado. Ciência e Agrotecnologia, 43.

Srivastava, M. S. e Dolatabadi, M. (2009). Multiple imputation and other resampling schemes for imputing missing observations. Journal of Multivariate Analysis, 100(9):1919–1937.

Van Buuren, S. (2018). Flexible imputation of missing data. Chapman and Hall/CRC, 416p.

Yan, W. (2013). Biplot analysis of incomplete two-way data. Crop Science, 53(1):48–57.

Zhang, P. (2003). Multiple imputation: theory and method. International Statistical Review, 71(3):581–592.