MAYUMI FURUYA DE ASSIS

MODELOS LINEARES GENERALIZADOS MISTOS E APLICAÇÕES DE REDES NEURAIS NO ESTUDO DA DIVERSIDADE GENÉTICA EM VARIEDADES DE *Coffea arabica*

Dissertação apresentada à Universidade Federal de Viçosa, campus Rio Paranaíba como parte das exigências do Programa de Pós-Graduação em Agronomia, Produção Vegetal para obtenção do título de *Magister Scientiae.*

Orientador: Pedro Ivo Vieira Good God
Coorientador: Everaldo Antônio Lopes

RIO PARANAÍBA - MINAS GERAIS

2020

Ficha catalográfica elaborada pela Biblioteca da Universidade Federal de Viçosa - Campus Rio Paranaíba

T
A848m

Assis, Mayumi Furuya de, 1996-
Modelos lineares generalizados mistos e aplicações de 2020 redes neurais no estudo da diversidade genética em variedades de *Coffea arabica* / Mayumi Furuya de Assis. – Rio Paranaiba, MG, 2020.
65 : il. (algumas color.) ; 29 cm.

Inclui apêndices.
Orientador: Pedro Ivo Vieira Good God.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Melhoramento genético do café. 2. Métodos estatísticos. 3. Mapas auto-organizáveis de Kohonen. 4. Modelos mistos. I. Universidade Federal de Viçosa. Ciências Agrárias. Mestrado em Agronomia (Produção Vegetal). II. Título.

633.73

MAYUMI FURUYA DE ASSIS

MODELOS LINEARES GENERALIZADOS MISTOS E APLICAÇÕES DE REDES NEURAIS NO ESTUDO DA DIVERSIDADE GENÉTICA EM VARIEDADES DE *Coffea arabica*

Dissertação apresentada à Universidade Federal de Viçosa, campus Rio Paranaíba como parte das exigências do Programa de Pós-Graduação em Agronomia, Produção Vegetal para obtenção do título de *Magister Scientiae.*

APROVADA: 7 de agosto de 2020.

Assentimento:

Mayumi Furuya de Assis
Autor

Pedro Ivo Vieira Good God
Orientador

A minha mãe, meu irmão e

a todos meus amigos,

que fizeram parte dessa tragetória e me deram apoio para continuar.

AGRADECIMENTOS

Gostaria de agradecer primeiramente à minha mãe, Konoe Furuya e meu irmão Nobuyuki Furuya pelo grande apoio emocional ao longo de toda minha trajetória acadêmica. Sem os valores, o amor, a alegria, educação e disciplina que veio da minha família, o caminho seria sem dúvidas muito mais difícil.

Ao meu orientador, Prof. Pedro Ivo Vieira Good God, que desde os primeiros estágios sempre esteve disposto a compartilhar seu conhecimento para me orientar e possibilitar o meu progresso profissional. Gostaria de expressar minha imensa gratidão pelo apoio para não desistir nos momentos mais complicados, pela paciência de ensinar desde a pipetar no laboratório até as análises estatísticas mais complexas, por confiar na minha competência em pesquisa e ensino. Também sou grata por sempre correr atrás de abrir portas para novas oportunidades para mim.

Gostaria de agradecer ao meu coorientador Professor Everaldo Antônio Lopes e todos os membros da banca avaliadora, Professores Moysés Nascimento, Vinícius Ribeiro Faria, Liliane Evangelista Visôtto e Luciano Bueno dos Reis por aceitarem compartilhar seus conhecimentos para o enriquecimento deste trabalho.

O presente trabalho foi realizado com o apoio do Consórcio Pesquisa Café (Edital 20/2018 Programa Café - Projeto Código: 10.18.20.037.00.00/170). Agradeço também ao Programa de Bolsas e Auxílio do Consórcio Pesquisa Café pela concessão de bolsa n° 1984, alocado(a) na solução de inovação/plano de ação 10.18.20.037.00.04.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

RESUMO

ASSIS, Mayumi Furuya de, M.Sc., Universidade Federal de Viçosa, agosto de 2020. **Modelos lineares generalizados mistos e aplicações de redes neurais para estudo da diversidade genética em variedades de *coffea arabica.*** Orientador: Pedro Ivo Vieira Good God. Coorientador: Everaldo Antônio Lopes.

A predição de parâmetros genéticos e o estudo da diversidade em variedades de café são etapas importantes para a seleção eficiente de genótipos que apresentem características superiores e garantir a manutenção da variabilidade genética. Muitas variáveis importantes para o melhoramento do café não possuem distribuições normais, dificultando a obtenção de predições acuradas e informações a respeito da estrutura genética das populações. Os modelos lineares generalizados mistos (MLGM) foram propostos para modelar variáveis que tenham distribuições diferentes das normais, por meio da inserção de uma função de ligação ao modelo. Essas características tornam os MLGM potenciais ferramentas para predição de componentes de variância, parâmetros genéticos e obtenção de valores de BLUP. O estudo da diversidade genética do café é comumente feito por métodos multivariados convencionais. Entretanto, alguns estudos têm utilizado um tipo de redes neural, chamada *Self-organizing maps* (SOM). Portanto, os objetivos deste estudo foram comparar modelos lineares mistos (MLM) com os MLGM na estimação de parâmetros genéticos, rankeamento dos genótipos através dos valores de BLUP e investigar a aplicabilidade das redes neurais SOM para o estudo da diversidade genética em variedades de *Coffea arabica*. Os MLGM detectaram variabilidade genética em um maior número de variáveis avaliadas. As variâncias genéticas e residuais estimadas pelos MLGM foram menores, entretanto, as análises de resíduos indicaram que os MLGM obtiveram desvios menores entre os valores ajustados e esperados pelos modelos em relação aos MLM. As interpretações dos parâmetros genéticos estimados via MLGM é mais complexa devido às diferentes escalas do modelo. Portanto, os MLGM foram mais eficazes para os ajustes dos dados, mas são necessárias investigações a respeito da influência as diferentes escalas atribuídas aos componentes do modelo na estimação dos parâmetros genéticos. As análises de redes neurais ofereceram diferentes informações a respeito das variáveis avaliadas, permitiram a

identificação de genótipos superiores para as variáveis analisadas e identificação de grupos genéticos similares e divergentes. Portanto, o método SOM é viável, informativo e refinado para o estudo da diversidade genética em populações de melhoramento.

Palavras-chave: Melhoramento genético do café. Métodos estatísticos. Mapas auto-organizáveis de Kohonen. Modelos mistos.

ABSTRACT

ASSIS, Mayumi Furuya de, M.Sc., Universidade Federal de Viçosa, August, 2020. **Generalized linear mixed models and applications of neural networks for the study of genetic diversity in *Coffea arabica* varieties.** Adviser: Pedro Ivo Vieira Good God. Co-adviser: Everaldo Antônio Lopes.

The prediction of genetic parameters and study of coffee varieties diversity are very important steps for the selection of important characteristics, superior genotypes and guarantee genetic variability maintenance. There is a interest for robust statistical tools that bring more acurated results for genetic breeding. Many important variables for the coffee breeding have non normal distribution, making it difficult obtain accurate predictions and good informations about varieties genetic structure. Generalized linear mixed models (GLMM) have been proposed to analyse variables that has non normal distributions adding a link function to the model. These characteristics turns GLMM a potential tool to predict components of variances, genetic parameters and BLUP values. The study of the genetic diversity of coffee commonly uses classical approaches. However, some studies have been using a type of neural network, called Self organizing maps (SOM). Therefore, the aims of this study were to compare linear mixed models (LMM) with GLMM in genetic parameters estimations and genotype ranking and to investigate the applicability of neural network SOM for the study of genetic diversity in *Coffea arabica* varieties. MLGM detected genetic variability in a greater number of evaluated variables. The genetic and residual variances estimated by the MLGM were smaller, however, the residual analyzes indicated that the MLGM obtained smaller deviations between the adjusted and expected values by the models in relation to the MLM. The interpretations of the genetic parameters estimated via MLGM are more complex due to the different scales of the model. However, the interpretations of the genetic parameters estimated using GLMM are more complex due to the different scales of the model. Therefore, GLMM was more efficient for data adjustments, but investigations and respect for the influence of the model's components on the prediction of genetic parameters are needed. The neural networks analysis offered different

information about the evaluated variables and allowed the identification of superior genotypes for the variables and identification of similar and divergent genetic groups. Therefore, the SOM method is viable, informative, and refined for the study of genetic diversity in breeding populations.

Keywords: Coffee breeding. Statistical methods. Kohonen’s self-organizing maps. Mixed models.

SUMÁRIO

INTRODUÇÃO GERAL 10
REFERÊNCIAS 12
CAPÍTULO 1 14
RESUMO 15
ABSTRACT 16
1. INTRODUÇÃO 17
2. MATERIAIS E MÉTODOS 18
2.1 Material genético e dados fenotípicos 18
2.2 Modelo linear misto 19
2.3 Modelo linear generalizado misto 20
2.4 Estimação de parâmetros e correlação de Spearman 21
2.5 Análise de resíduos 21
3. RESULTADOS 22
3.1 Distribuição de dados e ajustes de modelos 22
3.2 Média fenotípica, componentes de variância e parâmetros genéticos 22
3.3 Correlações de Spearman 25
3.4 Resíduos de Pearson x valores ajustados 25
4. DISCUSSÃO 29
5. CONCLUSÕES 33
6. REFERÊNCIAS 34
APÊNDICES 37
APÊNDICE A 37
APÊNDICE B 39
CAPÍTULO 2 44
RESUMO 45
ABSTRACT 46
1. INTRODUÇÃO 47
2. MATERIAIS E MÉTODOS 48
2.1 Material genético e dados fenotípicos 48
2.1 Modelos lineares mistos 50
2.2 *Self-organazing maps* (SOM) 51
2.3 Análise de componentes principais e agrupamento UPGMA 52
3. RESULTADOS 52
3.1 Ajustes de modelos lineares mistos e correlações de Pearson 52
3.2 *Self-organazing maps* (SOM) 55
3.3 Análise de componentes principais e agrupamento UPGMA 58
4. DISCUSSÃO 59
5. CONCLUSÕES 61
6. REFERÊNCIAS 62
CONCLUSÕES GERAIS 65

INTRODUÇÃO GERAL

As principais ferramentas estatísticas utilizadas em estudos de parâmetros genéticos do café são a análise de variância (ANOVA) e modelos lineares mistos (MLM) (RESENDE et al, 2001; BOTELHO et al., 2010; GUEDES et al., 2013; GILES et al., 2018; SOUSA et al., 2019; TASSONE et al., 2019). Entretanto, as acurácias destes modelos estão relacionadas a algumas pressuposições a respeito dos dados, como, por exemplo, a distribuição normal e a homogeneidade e independência dos resíduos (JANSEN, 1993; SMITH; CULLIS; THOMPSON, 2005; PIEPHO; ECKL, 2014). Aplicar estes métodos para analisar variáveis para o melhoramento do café, que comumente não respeitam tais pressuposições, pode resultar em predições pouco acuradas dos componentes de variância e, consequentemente, parâmetros genéticos do modelo.

Os modelos lineares generalizados mistos (MLGM) possuem particularidades, como a predição de efeitos aleatórios e fixos de um modelo a partir de variáveis cujas distribuições pertencem à família exponencial e a aplicação de uma função de ligação que relacione os parâmetros estimados aos valores esperados para uma variável (JANSEN, 1993). Devido às características flexíveis deste método, atualmente os MLGM tem sido frequentemente sugeridos como uma alternativa para estimar parâmetros genéticos em diferentes espécies animais e vegetais (BOLKER et al., 2009; WILSON et al., 2013; CAPPA; VARONA, 2013; MAKOUANZI et al., 2014; BALSALOBRE et al., 2016; WENG et al., 2016; MELO et al., 2020). Apesar das vantagens observadas para os MLGM, estes ainda não foram investigados em estudo de componentes de variância, parâmetros genéticos e valores genotípicos para o melhoramento genético do café.

Para os programas de melhoramento genético, também é importante que a variabilidade genética seja explorada e mantida para assegurar a seleção de cultivares produtivas e resistentes a diferentes fatores ambientais. As análises de componentes principais e de agrupamentos são frequentemente utilizados em estudos de diversidade genética no café (GUEDES et al., 2013; TEIXEIRA et al., 2013; RODRIGUES et al., 2016; MACHADO et al., 2017; GILES et al., 2019), porém, estes métodos podem apresentar baixa resolução devido a incapacidade para lidar com o volume e complexidade de dados, que são crescentes na evolução dos estudos de melhoramento genético (BARBOSA et al., 2011; IBRAHIM et al., 2016; SPANOGHE et al., 2020).

A investigação da viabilidade das redes neurais do tipo *Self-organizing maps* (SOM) (KOHONEN, 1982) para estudos de diversidade genética de diferentes espécies têm apresentado resultados satisfatórios (BARBOSA et al., 2011; IBRAHIM et al., 2016; SPANOGHE et al., 2020; SANTOS et al, 2020). A principal função dos SOM é mapear amostras em uma representação uni ou bidimensional de acordo com suas similaridades com poucas distorções em relação às suas posições em um espaço multidimensional mais complexo. Devido às suas características, os SOM são capazes de detectar e classificar padrões em diferentes dimensões dos conjuntos de dados e determinar agrupamentos naturais de acordo com as similaridades entre as amostras (KOHONEN, 2014).

Devido a necessidade de investigar novas ferramentas estatísticas para assegurar o potencial de ganho e selecionar genótipos superiores em programas de melhoramento, os objetivos deste trabalho foram: 1) Comparar os MLM e MLGM na estimação de componentes de variância e parâmetros genéticos para variáveis com diferentes distribuições de probabilidade, em variedades de *C. arabica*. 2) Explorar a aplicabilidade das redes neurais SOM para o estudo da diversidade genética em variedades de *C. arabica*.

REFERÊNCIAS

BALSALOBRE, T. W. A. et al. Mixed modeling of yield components and brown rust resistance in sugarcane families. **Agronomy Journal**, v. 108, n. 5, p. 1824–1837, 2016.

BARBOSA, C. D. et al. Artificial neural network analysis of genetic diversity in Carica papaya L. **Crop Breeding and Applied Biotechnology**, v. 11, n. 3, p. 224–231, 2011.

BOLKER, B. M. et al. Generalized linear mixed models: a practical guide for ecology and evolution. **Trends in Ecology and Evolution**, v. 24, n. 3, p. 127–135, 2009.

BOTELHO, C. E. et al. Adaptability and phenotype stability of Arabica coffee cultivars in Minas Gerais, Brazil. **Pesquisa Agropecuaria Brasileira**, v. 45, n. 12, p. 1404–1411, 2010.

CAPPA, E. P.; VARONA, L. An assessor-specific Bayesian multi-threshold mixed model for analyzing ordered categorical traits in tree breeding. **Tree Genetics and Genomes**, v. 9, n. 6, p. 1423–1434, 2013.

GILES, J. A. et al. Divergence and genetic parameters between coffea sp. genotypes based in foliar morpho-anatomical traits. **Scientia Horticulturae**, v. 245, n. May 2018, p. 231–236, 2019.

GUEDES, J. M. et al. Divergência genética entre cafeeiros do germoplasma Maragogipe. **Bragantia**, v. 72, n. 2, p. 127–132, 2013.

IBRAHIM, O. M. et al. Evaluating the Performance of 16 Egyptian Wheat Varieties Using Self-Organizing Map (SOM) and Cluster Analysis. **Journal of Applied Sciences**, v. 16, n. 2, p. 47–53, 2016.

JANSEN, J. Generalized linear mixed models and their application in plant breeding research. **Eindhoven: Technische Universiteit Eindhoven**, p. 143, 1993.

KOHONEN, T. Selg organized formation of topologically correct feature maps. **Biological Cybernetics**, v. 69, p. 59–69, 1982.

KOHONEN, T. **MATLAB Implementations and Applications of the Self-Organizing Map**. Unigraphia Oy: Helsinki, Finland, 2014.

MACHADO, C. M. S. et al. Genetic diversity among 16 genotypes of Coffea arabica in the Brazilian cerrado. **Genetics and Molecular Research**, v. 16, n. 3, p. 1–13, 2017.

MAKOUANZI, G. et al. Assessing the additive and dominance genetic effects of vegetative propagation ability in Eucalyptus—influence of modeling on genetic gain. **Tree Genetics and Genomes**, v. 10, n. 5, p. 1243–1256, 2014.

MELO, B.; SOUSA, L. B. Biologia da reprodução de Coffea arabica L . e Coffea canephora Pierre. **Revista verde de agroecologia e desenvolvimento sustentavelerde de agroecologia e desenvolvimento sustentavel**, v. 5 (3), p. 05–11, 2011.

MELO, R. C. DE et al. Statistical model assumptions achieved by linear models: classics and generalized mixed. **Revista Ciência Agronômica**, v. 51, n. 1, p. 1–9, 2020.

PIEPHO, H. P.; ECKL, T. Analysis of series of variety trials with perennial crops. **Grass and Forage Science**, v. 69, n. 3, p. 431–440, 2014.

RESENDE, M. et al. Estimativas de parâmetros genéticos e predição de valores genotípicos no melhoramento do cafeeiro pelo procedimento REML/BLUP. **Bragantia**. v. 60, n. 3, p. 185–193, 2001.

RODRIGUES, W. P. et al. Assessment of genetic divergence among coffee genotypes by Ward-MLM procedure in association with mixed models. **Genetics and Molecular Research**, v. 15, n. 2, 2016.

SANTOS, I. G. et al. Self-organizing maps in the study of genetic diversity among irrigated rice genotypes. **Acta Scientiarum - Agronomy**, v. 41, n. 1, p. 1–9, 2019.

SMITH, A. B.; CULLIS, B. R.; THOMPSON, R. The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. **Journal of Agricultural Science**, v. 143, n. 6, p. 449–462, 2005.

SOUSA, T. V. et al. Early selection enabled by the implementation of genomic selection in coffea arabica breeding. **Frontiers in Plant Science**, v. 9, n. January, p. 1–12, 2019.

SPANOGHE, M. C. et al. Genetic patterns recognition in crop species using self-organizing map: the example of the highly heterozygous autotetraploid potato (Solanum tuberosum L.). **Genetic Resources and Crop Evolution**, v. 67, n. 4, p. 947–966, 2020.

TASSONE, G. A. T. et al. Simultaneous selection in coffee progenies of mundot novo by selection indices. **Coffee Science**, v. 14, n. 1, p. 83–92, 2019.

TEIXEIRA, A. L. et al. Principal component analysis on morphological traits in juvenile stage arabica coffee. **Coffee Science**, v. 8, n. 2, p. 205–211, 2013.

TOUNEKTI, T. et al. Genetic Diversity Analysis of Coffee (<i>Coffee arabica</i> L.) Germplasm Accessions Growing in the Southwestern Saudi Arabia Using Quantitative Traits. **Natural Resources**, v. 08, n. 05, p. 321–336, 2017.

VOLSI, B. et al. The dynamics of coffee production in Brazil. **PLoS ONE**, v. 14, n. 7, p. 1–15, 2019
.
WENG, Y. et al. Genetic Parameters for Bole Straightness and Branch Angle in Jack Pine Estimated Using Linear and Generalized Linear Mixed Models. **Forest Science**, v. 63, n. 1, p. 111–117, 2016.

WILSON, B. J. et al. Estimated breeding values for canine hip dysplasia radiographic traits in a cohort of Australian German Shepherd dogs. **PloS one**, v. 8, n. 10, 2013.

CAPÍTULO 1

MODELOS LINEARES GENERALIZADOS MISTOS NA ESTIMAÇÃO DE PARÂMETROS GENÉTICOS EM VARIEDADES DE *Coffea arabica*

RESUMO

DE ASSIS, Mayumi Furuya, M.Sc., Universidade Federal de Viçosa, agosto de 2020. **Modelos lineares generalizados mistos na estimação de parâmetros genéticos em variedades de *Coffea arabica***. Orientador: Pedro Ivo Vieira Good God. Coorientador: Everaldo Antônio Lopes.

Os métodos estatísticos frequentemente utilizados em estudos de parâmetros genéticos no café são a ANOVA e modelos lineares mistos (MLM). Entretanto, esses métodos assumem pressuposições que comumente não ocorrem em variáveis biométricas, como a homogeneidade residual e distribuição normal dos dados. Os modelos lineares generalizados mistos (MLGM) são capazes de modelar variáveis cujas distribuições pertencem à família exponencial e aplicar uma função de ligação capaz de relacionar o preditor linear do modelo ao valor esperado para uma variável aleatória. Devido à suas particularidades, a aplicabilidade dos MLGM tem sido investigada para estimar parâmetros genéticos em diferentes espécies. O objetivo deste estudo foi comparar os componentes de variância, parâmetros genéticos e ordenação dos genótipos através dos valores de BLUP pelos MLM e MLGM para 16 características, com diferentes distribuições de probabilidade, avaliadas em 62 variedades experimentais de *C. arabica*. Foram testadas seis famílias de MLGM com as distribuições normal e gama, os modelos mais adequados foram selecionados pelo critério de AIC. Foram observadas diferenças entre os parâmetros estimados via MLM e MLGM, as variâncias residuais e genéticas estimadas via MLGM foram predominantemente menores e os valores das herdabilidades calculadas pelos dois modelos foram variáveis. Mesmo que as variâncias genéticas estimadas via MLGM tenham sido muito menores, o modelo detectou variabilidade genética para mais variáveis que os MLM. Não foi encontrada diferenças entre os modelos para ranqueamento dos genótipos pelos valores genéticos preditos. Os MLGM apresentaram melhor ajuste dos dados. Entretanto, a interpretação destes parâmetros sofre interferência das escalas envolvidas no modelo, que dependem da distribuição atribuída aos dados e das funções de ligação. Portanto, ainda são necessárias mais investigações a fim de facilitar as interpretações das predições obtidas através dos MLGM.

Palavras-chave: Melhoramento do café, modelos mistos, métodos estatísticos.

ABSTRACT

DE ASSIS, Mayumi Furuya, M.Sc., Universidade Federal de Viçosa, august, 2020. **Genetic parameter estimation in *Coffea arabica* varieties with generalized linear mixed models.** Adviser: Pedro Ivo Vieira Good God. Co-adviser: Everaldo Antônio Lopes.

Some natural characteristics of coffee culture turns the phenotyping and application of genetic parameters challenging for breeding programs. The statistical methods frequently used to estimate the genetic parameters in coffee are ANOVA and linear mixed models (LMM). However, these methods assume some assumptions that do not occurs in biometric data, such as residuals homogeneity and normal data distribution. Generalized linear mixed models (GLMM) are capable of modeling variables, whose distributions belong to exponential family and apply a link function capable of relating the linear predictor with the expected value for a random variable. Considering its particularities, MLGM have been used to estimate genetic parameters for different species. The aim of this study was to compare the variance components, genetic parameters and genotype classification according with BLUP values obtained by LMM and GLMM for 16 traits, with different probability distributions, evaluated in 62 experimental varieties of *C. arabica*. Six GLMM families were tested with the normal and gama distributions, the better fitted models were selected by the AIC criteria and the model's diagnostics were analyzed by Pearson's residual plots. It was found differences between the parameters estimated by models. The GLMM estimated residual and genetic variances predominantly smaller, and the values of heritability calculated were very diversified. Even though, the genetic variations estimated by GLMM were much smaller, this model detected genetic variability for more traits than LMM. No difference was found between the genotype ranking by genetic values predicted. Pearson's residual plots and the low residual variances indicates that GLMM better fit the data, mainly with gama distribution, can be a viable alternative for non-normal biometric data analysis and with heterogeneous variances in breeding programs. However, the interpretation of these parameters is affected by changes in GLMM scales, which depends on the distribution attributed to the data and the link functions. Therefore, even if GLMM has shown to be more efficient in data explanation, further investigation is still needed to improve interpretation of predicted components.

Keywords: coffe breeding, mixed models, statistical methods.

1. INTRODUÇÃO

O melhoramento genético na cultura do café apresenta certas particularidades. O *Coffea arabica* é uma espécie de desenvolvimento perene, ciclo reprodutivo longo, apresenta oscilação anual de produção, taxa de sobrevivência instável de indivíduos e sobreposição de gerações em ciclos de seleção. Estes fatos tornam o melhoramento do café desafiador por influenciarem nos métodos de seleção e pela necessidade de avalições repetidas ao longo do tempo. Além disso, é muito comum que sejam gerados dados desbalanceados, que influenciam diretamente na estimação dos parâmetros genéticos e na predição de valores genotípicos (RESENDE et al, 2001; CARVALHO, 2008).

Em *C. canephora* e *C. arabica*, a análise de variância (ANOVA) é o método mais utilizado em estudos genéticos. No entanto, as estimativas obtidas pelos métodos dos momentos podem ser de baixa acurácia e os componentes de variância estimados podem assumir valores negativos, fatos que interferem na estimação e interpretação dos parâmetros genéticos. Além disso, devido aos dados longitudinal obtidos pelos experimentos nas culturas de café não raramente são obtidas variâncias residuais heterogêneas e correlacionadas, não atendendo à pressuposição de homogeneidade residual exigida para ANOVA (RESENDE et al, 2001; SMITH; CULLIS; THOMPSON, 2005; PIEPHO; ECKL, 2014). O conjunto desses fatores demandam modelos estatísticos mais robustos e informativos, e que aumentem a eficiência de estimação componentes de variância e predição dos valores genotípicos.

Os modelos lineares mistos (MLM) são utilizados como uma alternativa para lidar com alguns destes problemas. MLM permitem o ajuste de efeitos fixos e aleatórios do modelo, resultando na estimação dos componentes de variância e do erro experimental de forma mais eficiente (HENDERSON, 1975; SMITH; CULLIS; THOMPSON, 2005). Em estudos de melhoramento de *C. arabica*, os MLM têm sido frequentemente utilizados na predição de valores genéticos pela melhor predição linear não viciada (BLUP) e estimação dos componentes de variância pela máxima verossimilhança restrita (REML) (RESENDE et al, 2001; RODRIGUES et al., 2016; SILVA et al., 2018; SOUSA et al., 2019; TASSONE et al., 2019). de café.

Mesmo que os MLM sejam amplamente utilizados para estudos genéticos no café, em seu ajuste deve-se assumir distribuição normal dos dados. Entretanto, muitas situações envolvem dados não normais, contínuos assimétricos e discretos, por exemplo o vigor das

plantas, número de ramos plagiotrópicos primários e secundários e número de internódios. Uma abordagem clássica a fim de atender as condições de homogeneidade e normalidade é a transformação logarítmica dos dados (MCCULLAHG; NELDER, 1989). Entretanto, a transformação não garante as adequações necessárias para os testes paramétricos, pode limitar a extrapolação das inferências para outros populações e levar a conclusões equivocadas (BOLKER et al., 2009).

Os modelos lineares generalizados (MLG) foram propostos para lidar com esses obstáculos (NELDER; WEDDERBURN, 1972; MCCULLAHG; NELDER, 1989).Os MLG permitem a modelagem de variáveis, cujas distribuições pertencem à família exponencial, e a utilização de uma função de ligação que relaciona os parâmetros do modelo (preditores lineares) ao valor esperado para uma variável aleatória (MCCULLAHG; NELDER, 1989; STROUP; KACHMAN, 1994; VILLEMEREUIL et al., 2016).Os modelos lineares generalizados mistos (MLGM) são uma extensão dos MLG que permitem a adição de efeitos aleatórios independentes ao preditor linear. Nos MLGM os fatores aleatórios e fixos são ajustados da mesma forma que em MLM, ou seja, os parâmetros de interesse são estimados por máxima verossimilhança (STROUP; KACHMAN, 1994; BOLKER et al., 2009). Devido à essas características, esta ferramenta tem sido crescentemente utilizada para estimar parâmetros genéticos em diferentes modelos biológicos (BOLKER et al., 2009; CAPPA; VARONA, 2013; WILSON et al., 2013; MAKOUANZI et al., 2014; BALSALOBRE et al., 2016; MELO et al., 2020; WENG et al., 2016).

No melhoramento genético do *C. arabica* muitas mensurações biométricas importantes são utilizadas para estimar parâmetros genéticos das populações. Entretanto, ainda não foi investigada a aplicação dos MLGM para estimar parâmetros genéticos em *C. arabica*. Neste contexto, o objetivo deste trabalho é comparar os MLM e MLGM na estimação de componentes de variância, parâmetros genéticos e classificação dos valores de BLUP de 62 variedades de *C. arábica* em 16 variáveis, com diferentes distribuições de probabilidade, de interesse para o melhoramento do café.

2. MATERIAIS E MÉTODOS

2.1 Material genético e dados fenotípicos

Foram utilizadas 62 variedades de *Coffea arabica* (Apêndice A). O experimento foi delineado em blocos ao acaso com três repetições, com dez plantas por parcela. O espaçamento utilizado foi de 3,80 x 0,70 m em uma área total de 7.798,23 m^2, localizado no munícipio de Rio Paranaíba, estado de Minas Gerais, Brasil (19°13'02" S, 46°13'59" W).

Foram avaliadas 16 variáveis nos anos de 2017, 2018 e 2019 (Tabela 2). A produção (PROD) foi mensurada em 10 plantas por parcela, os frutos foram coletados manualmente e com o auxílio de um medidor volumétrico quantificou-se a produção em L/ parcela. As características vigor (VIG), altura (ALT), diâmetro da copa (COPA) e diâmetro do caule (CAULE) foram avaliadas em três plantas por parcela. Para o VIG foram atribuídas notas de de 1 a 5, em função do desenvolvimento vegetativo das plantas, sendo 1 atribuídos às plantas pouco vigorosas e 5 aquelas com ótimo desenvolvimento. A ALT e COPA foram medidas com uma trena e um paquímetro digital foi utilizado para medir o diâmetro do caule 10 cm acima do solo. A área foliar (AF) foi mensurada em duas plantas por parcela no 3° ou 4° par de folhas em um ramo selecionado no terço médio da planta, de acordo com o método descrito por Kemp (1960). O comprimento de ramos plagiotrópicos 1° (CR1), n° de ramos plagiotrópicos 1° (NR1), n° de internódios 1° (NI1) e n° de ramos plagiotrópicos secundários (NR2) foram avaliados em três plantas centrais da parcela para remoção do efeito de bordadura. O CR1 foi medido com uma trena em um ramo do terço médio da planta a partir do ponto de inserção no ramo ortotrópico principal até a extremidade. O NR1 e NR2 foram mensurados por contagem direta no terço médio das plantas e o NI1 foi avaliado por contagem direta nos ramos selecionados para o avaliar o CR1. Para as avaliações das peneiras foram separados um volume de 5 L de café por parcela, as amostram foram secadas e revolvidas em terreiro suspenso. Foram avaliadas as quantidades em gramas retidas nas peneiras 19 (P19), 15 (P15) e 11 (P11), fundo de peneira (FUNDO) em g/ peneira, o peso de 100 grãos (P100) e o rendimento (REND) em proporção ao volume inicial pré-secagem.

2.2 Modelo linear misto

Os efeitos aleatórios e fixos, bem como seus componentes de variância para todas as variáveis avaliadas foram estimados de acordo com o modelo linear misto (MLM):

$$y = X\beta + Zg + e$$

em que: y é o vetor dos valores observados de tamanho n, β é o vetor de efeito fixo (blocos) associado à matriz de incidência X, g é o vetor de efeito aleatório (genótipo) associado às matrizes Z, $e \sim N(0, V = I_n\sigma^2_e)$ é o vetor residual e I é a matriz identidade associada ao erro.

Tabela 2. Variáveis, identificação (ID) e anos avaliados nos 62 genótipos de *C. arábica*.

Ano	Variáveis	ID
2017/ 2018	Produção (L/ parcela)	PROD
2018	Área Foliar (cm^2)	AF
	N° de ramos plagiotrópicos secundários	NR2
2018/ 2019	Vigor (1 – 5)	VIG
	Altura da planta (m)	ALT
	Diâmetro da copa (m)	COP
	Diâmetro do caule (cm)	CAU
	Comprimento de ramos plagiotrópicos primários (cm)	CR1
	N° de ramos plagiotrópicos primário	NR1
	N° de internódios primário	NI1
2019	Peneira 11 (g/peneira)	P11
	Peneira 15 (g/peneira)	P15
	Peneira 19 (g/peneira)	P19
	Fundo de peneira (g/peneira)	FUN
	Peso de 100 grãos	P100
	Rendimento (% de 5L)	REN

2.3 Modelo linear generalizado misto

Os modelos MLGM são representados da seguinte forma (PIEPHO, 2019):

$$\eta = X\beta + Zu$$

em que: η é o preditor linear do modelo, que contém os vetores fixos e aleatórios, assim como em MLM. O vetor de observação y possui uma expectativa relacionada a uma distribuição da família exponencial.

$$E(y|\eta) = \mu = g^{-1}(\eta)$$

em que, $g(.)$ é a função de ligação que relaciona o preditor linear ao valor esperado dos dados.

Inicialmente os dados foram submetidos às análises gráficas quantil-quantil (Q-Q plot) e densidade de Kernel para interpretação das distribuições dos dados. Foram selecionadas cinco famílias de MLGM (Tabela 3) para os ajustes. $g(.)$ identidade e inversa foram escolhidas por

serem padrões (canônicas) para distribuição gaussiana e gama respectivamente (MCCULLAHG; NELDER, 1989; CRAWLEY, 2007). $g(.)$ logarítmica foi assumida por ser mais adequado assumir uma relação exponencial entre η e a μ dos dados do que transformar os dados (MCCULLAHG; NELDER, 1989; BOLKER et al., 2009).

Tabela 3. Distribuições da família exponencial, funções de ligação, notações e identificação (ID) das famílias de MLGM selecionados para os testes de ajustes.

Distribuição	**Função de ligação**	**Notação**	**Abreviação**	**ID**
Gaussiana (μ, σ^2)	Identidade	$g^{-1}(\eta) = \mu_i$	~N(link=identidade)	M1
	Log	$g^{-1}(\eta) = \ln(\mu_i)$	~N(link=log)	M2
	Inversa	$g^{-1}(\eta) = \mu^{-1}$	~N(link=inversa)	M3
Gama (μ, ν)	Identidade	$g^{-1}(\eta) = \mu_i$	~Ga(link=identidade)	M4
	Log	$g^{-1}(\eta) = \ln(\mu_i)$	~Ga(link=log)	M5
	Inversa	$g^{-1}(\eta) = \mu^{-1}$	~Ga(link=inversa)	M6

O MLGM de melhor ajuste para cada característica, foi selecionado pelo critério de informação de Akaike (AIC) (AKAIKE, 1974). A estimativa do critério é feita de acordo com a equação:

$$AIC = -2\ log\ L + 2k$$

em que: $logL$ é o logaritmo da função de verossimilhança e k é o número de parâmetros estimados do modelo.

2.4 Estimação de parâmetros e correlação de Spearman

Os coeficientes de variação residual (CV_r) e variação genética (CV_g) foram utilizados para análise dos componentes de variância do modelo devido as diferenças de escala dos dados (RESENDE; DUARTE, 2007). A herdabilidade (H^2) no sentido amplo foi calculada e os valores do efeito aleatório (genotípico) do modelo foram estimados via BLUP. Também foi calculada a correlação de Spearman entre os valores de BLUP gerados via MLM e MLGM para analisar as diferenças de classificação dos valores genotípicos entre os modelos.

2.5 Análise de resíduos

Para todas as características em que os MLM e MLGM foram ajustados, foram plotados gráficos de resíduo de Pearson x valores ajustados. O resíduo de Pearson é definido por:

$$R_i^P = \frac{yi - \hat{u}i}{\sqrt{var(Yi)}}$$

em que: R_i^P corresponde à contribuição de cada observação para o cálculo da estatística de Pearson generalizada. Para modelos bem ajustados, espera-se baixa discrepância entre o valor observado (y_i) e valor ajustado ($\hat{u}_i$) e pouca ou nenhuma variância, ou seja, resíduos próximos do valor de 0 (TURKMAN, 2000). Todas as análises foram realizadas com o software R (R Development Core Team, 2020). Os pacotes utilizados para os ajustes dos modelos foram lme4, MASS e car (RIPLEY et al., 2002; BATES et al., 2015; FOX; WEISBERG, 2019; VERNABLES)

3. RESULTADOS

3.1 Distribuição de dados e ajustes de modelos

A análise exploratória baseada em gráficos Q-Q (Apêndice B) e densidade Kernel (Figura 1) indicam que o conjunto de variáveis analisados se aproximam da distribuição normal e gama. Exceto para a variável VIG, cuja escala é categórica, todas as demais variáveis apresentam dados contínuos e positivos. Além disso, a maioria das características apresentaram distribuição assimétrica, indicando que são variáveis potenciais para o ajuste de modelos gama.

Os MLM detectaram variabilidade genética em menos variáveis em relação aos MLGM (Tabela 4). Para todas características avaliadas via MLGM o efeito de genótipos foi detectado. Para a maioria dos cenários avaliados (anos e variáveis) menores valores de AIC foram obtidos para os MLGM, indicando melhores ajustes dos dados. Na seleção de funções de ligação para MLGM, quando a família gaussiana identidade (M1) apresentou o melhor ajuste (menor AIC), assumimos a relação linear simples entre a médias e o preditor linear, tornando desnecessário ajustar MLGM para tais variáveis (p.e, AF 2019 e NI1 2019). Entre os MLGM ajustados, a família de distribuições gama foi predominante. Houve variação da função de ligação, inclusive entre os anos de avaliação para a mesma característica. Os modelos ajustados com maior frequência foram M4 [~Ga(link=identidade)] e M6 [~Ga(link=inversa)].

3.2 Média fenotípica, componentes de variância e parâmetros genéticos

As médias fenotípicas das variáveis de crescimento e desenvolvimento foram maiores em 2019, o que é naturalmente esperado pelo desenvolvimento das plantas (Tabela 5). Com

exceção do COP, NR1 e NI1, que apresentaram redução na média fenotípica de 2018 para 2019. A produção média de 2017 foi 4,25 vezes maior que em 2018, evidenciando o comportamento produtivo bienal do café. Em 2019, a peneira que reteve maior proporção de grãos foi a de grãos chato médio (P15).

Tabela 4. Valores p para o teste de razão de verossimilhança para efeito de genótipo via MLM e MLGM, valores de AIC para os MLM e MLGM ajustados para todas as variáveis avaliadas e família de MLGM ajustada para as variáveis em cada ano avaliado.

		Valor p		AIC	
Variável	**Ano**	**MLM**	**MLGM**	**MLM**	**MLGM (Modelo)**
VIG	2018	0,2162	0,0032**	385	373 (M3)
	2019	0,0850	0,0023**	492	486 (M2)
ALT	2018	6,37e-12***	< 2,20e-16***	-199	-260 (M4)
	2019	1,77e-06***	5,25e-16***	-21	-64 (M5)
COP	2018	0,0079**	3,30e-09***	-180	-208 (M2)
	2019	0,4417	5,09e-05***	-107	-123 (M2)
CAU	2018	8,06e-03**	1,40e-08***	1069	1048 (M5)
	2019	0,3219	5,60e-05***	1259	1255 (M4)
AF	2018	0,5745	-	1575	-
CR1	2018	0,4423	9,87e-06***	1398	1375 (M4)
	2019	0,7687	1,16e-04***	1344	1327 (M6)
NR1	2018	0,6985	0,0046**	1395	1365 (M6)
	2019	0,0514	1,19e-06***	835	809 (M4)
NI1	2018	-	0,0156*	-	1652 (M6)
	2019	0,8962	-	1056	-
NR2	2018	0,6034	0,0045**	1071	1039 (M6)
PROD	2017	0,2194	-	1646	-
	2018	5,06e-06***	1,15e-09***	1318	1214 (M5)
P19	2019	6,11e-04***	7,23e-10***	1631	1583 (M4)
P15		0,1008	-	1881	-
P11		0,2511	-	1909	-
FUN		0,0021**	1,20e-13***	1295	1169 (M5)
P100		0,0011**	1,32e-11***	334	298 (M6)
REN		0,0903	1,05e-06***	-676	-697 (M2)

Códigos de significância: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 de valores *p* do teste razão de razão de verossimilhança para efeito de genótipo. Células preenchidas com '–' indicam o não ajuste do modelo aos dados.

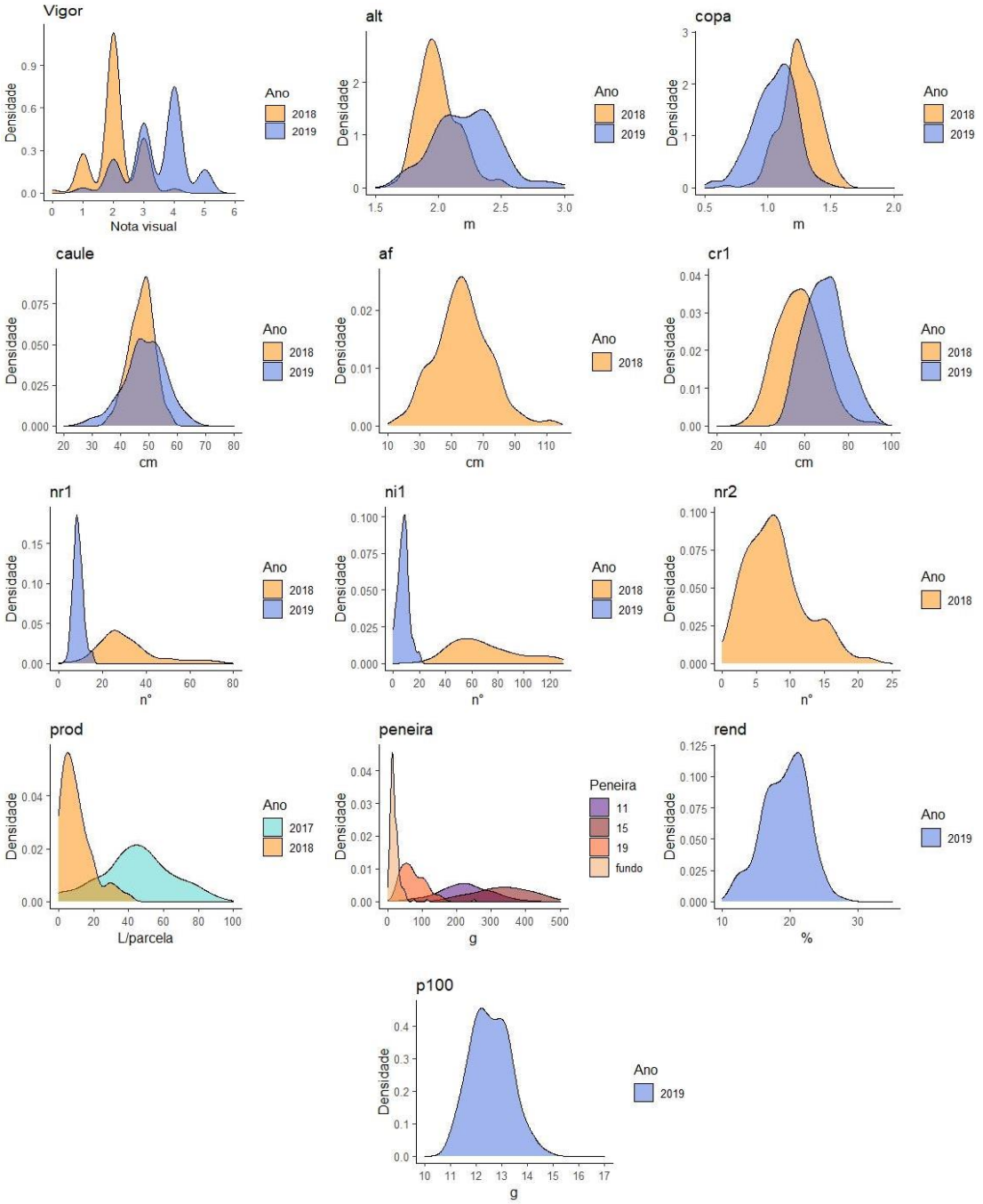


Figura 1. Gráficos de densidade de Kernel para as variáveis vigor (VIG), altura (ALT), diâmetro da copa (COP), diâmetro do caule (CAU), comprimento de ramos 1° (CR1), n° de ramos 1° (NR1) e n° de internódios 1° (NI1) avaliados em 2018 e 2019, n° de ramos 2° (NR2) e área foliar (AF) em 2018, produção (PROD) de 2017 e 2018, peneira 19, 15 e 11 (P19, P15 e P11), fundo de peneira (FUN), rendimento (REN) e peso de 100 grãos (P100) avaliados em 2019.

Os coeficientes de variação residual (CV_r) obtidos pelos MLGM foram predominantemente menores. Os CV_r estimados via MLM apresentaram uma variação entre 5,11 e 66,49 % e os valores obtidos pelos MLGM variaram entre 0,17 a 29,33% (Tabela 5).

Não foram observadas grandes diferenças nos CV_r estimados para as mesmas variáveis nos dois anos de avaliação.

Os coeficientes de variação genéticas (CV_g) estimados pelos MLGM também foram menores em relação aos MLM (Tabela 5). Sendo que os CV_g estimados pelos MLM variaram entre 1,93 e 49,06% e os valores obtidos pelos MLGM foram entre 0 e 42,94%. A produção de 2018 obteve o maior CV_g (49,06%) estimado via MLM (Tabela 5). Também não foram observadas diferenças significativas para os valores de CV_g entre os anos de avaliação, exceto para produção com aumento de 35,35% de 2017 para 2018. As características P19 e P100 foram as únicas que apresentaram estabilidade do CV_g entre os dois modelos.

As herdabilidades (H^2) estimadas pelos MLM foram relativamente baixas (≤0,4) para a maioria das características avaliadas, exceto para ALT e PROD em 2018 e P19 e FUN, que apresentaram $H^2 \geq 0,5$ (Tabela 5). A maioria das H^2 obtidas pelos MLGM foram mais baixas em relação aos MLM (≤ 0,3). Houveram exceções para as variáveis CAU em 2019, CR1, P19, REN e P100 que apresentaram valores de H^2 próximos a 0,9. Entre os anos de avaliação não houve grandes diferenças, exceto para ALT.

3.3 Correlações de Spearman

Foram observados valores altos de correlações para todas as variáveis (≥ 0,97) (Tabela 5), indicando que os MLM e MLGM não diferem em relação ao rankeamento dos genótipos pelos valores genotípicos. Correlações negativas foram observadas para os casos em que os MLGM com função de ligação inversa foram ajustados às variáveis.

3.4 Resíduos de Pearson x valores ajustados

Os gráficos de resíduos para os MLM e MLGM (Figuras 2 e 3) demonstraram, para quase todas as variáveis, pontos distantes do valor de 0 do eixo Y para os MLM e tendência (positivas ou negativas) em alguns casos. Para os MLGM, os resíduos foram bem menores, podendo ser observado pela diferença de escala do eixo Y entre os gráficos. Em alguns casos, como em NR1 em 2018, os MLGM apresentaram resíduos mais aleatorizados em relação aos MLM. Já para o CR1 em 2019 foi possível observar a homogeneização completa dos resíduos pelos MLGM. Não houve grandes diferenças entre os modelos em relação aos ajustes de *outliers*, exceto para COP em 2019.

Tabela 5. Valores de média experimental, coeficientes de variação residual (CV_r), coeficiente de variação genético (CV_g), herdabilidade (H^2) e correlação de *Spearman* entre os BLUP preditos via MLM e MLGM para todas as variáveis.

			$CV_{r\,(\%)}$		$CV_{g\,(\%)}$		H^2		
Variável	**Ano**	**Média**	**MLM**	**MLGM**	**MLM**	**MLGM**	**MLM**	**MLGM**	**Spearman**
VIG	2018	2,07	30,96	29,33	9,92	1,55	0,2354	0,0083	-0,97
	2019	3,46	24,61	23,54	9,55	2,79	0,3111	0,0405	0,99
ALT	2018	2,00	5,54	2,80	5,28	1,15	0,7314	0,3344	-1,00
	2019	2,22	8,49	8,50	6,58	1,27	0,6428	0,0629	-1,00
COP	2018	1,26	10,52	10,16	5,34	2,74	0,4359	0,1787	1,00
	2019	1,05	16,45	15,43	4,09	3,72	0,1568	0,1486	1,00
CAU	2018	47,32	8,35	0,17	4,22	0,21	0,4335	0,5088	0,99
	2019	48,45	14,49	0,29	4,13	6,84	0,1955	0,9994	0,99
AF	2018	57,07	28,92	0,48	6,03	0,00	0,1154	0,0002	-1,00
CR1	2018	57,74	17,17	0,28	4,23	8,27	0,1543	0,9996	1,00
	2019	69,46	12,80	0,17	1,93	5,77	0,0638	0,9997	1,00
NR1	2018	31,09	33,40	0,93	5,80	0,01	0,0829	0,0005	-1,00
	2019	8,94	23,67	2,52	9,94	0,15	0,3460	0,0104	-1
NI1	2018	71,33	-	0,41	-	0,00	-	0,0001	
	2019	7,77	52,89	-	5,24	-	0,0286	-	-
NR2	2018	7,83	54,66	5,82	10,99	0,28	0,1081	0,0070	-1,00
PROD	2017	44,70	42,28	-	13,72	-	0,2401	-	
	2018	10,51	66,49	5,81	49,06	0,32	0,6202	0,0093	0,99
P19	2019	73,1	38,56	0,52	24,31	24,71	0,5440	0,9998	-0,99
P15		331,47	23,84	-	11,70	-	0,4196	-	-
P11		222,17	32,00	-	10,42	-	0,2413	-	-
FUN		19,63	55,55	2,32	36,81	2,03	0,5687	0,6984	0,99
REN		19,23	15,20	14,43	6,08	42,94	0,3272	0,9637	1,00
P100		12,56	5,11	0,39	3,34	3,65	0,5622	0,9962	1,00

CV_r(%): coeficiente de variação residual, CV_g(%): coeficiente de variação genotípico, H^2: herdabilidade no sentido amplo para as características avaliadas nos anos de 2017, 2018 e 2019 via MLM e MLGM.

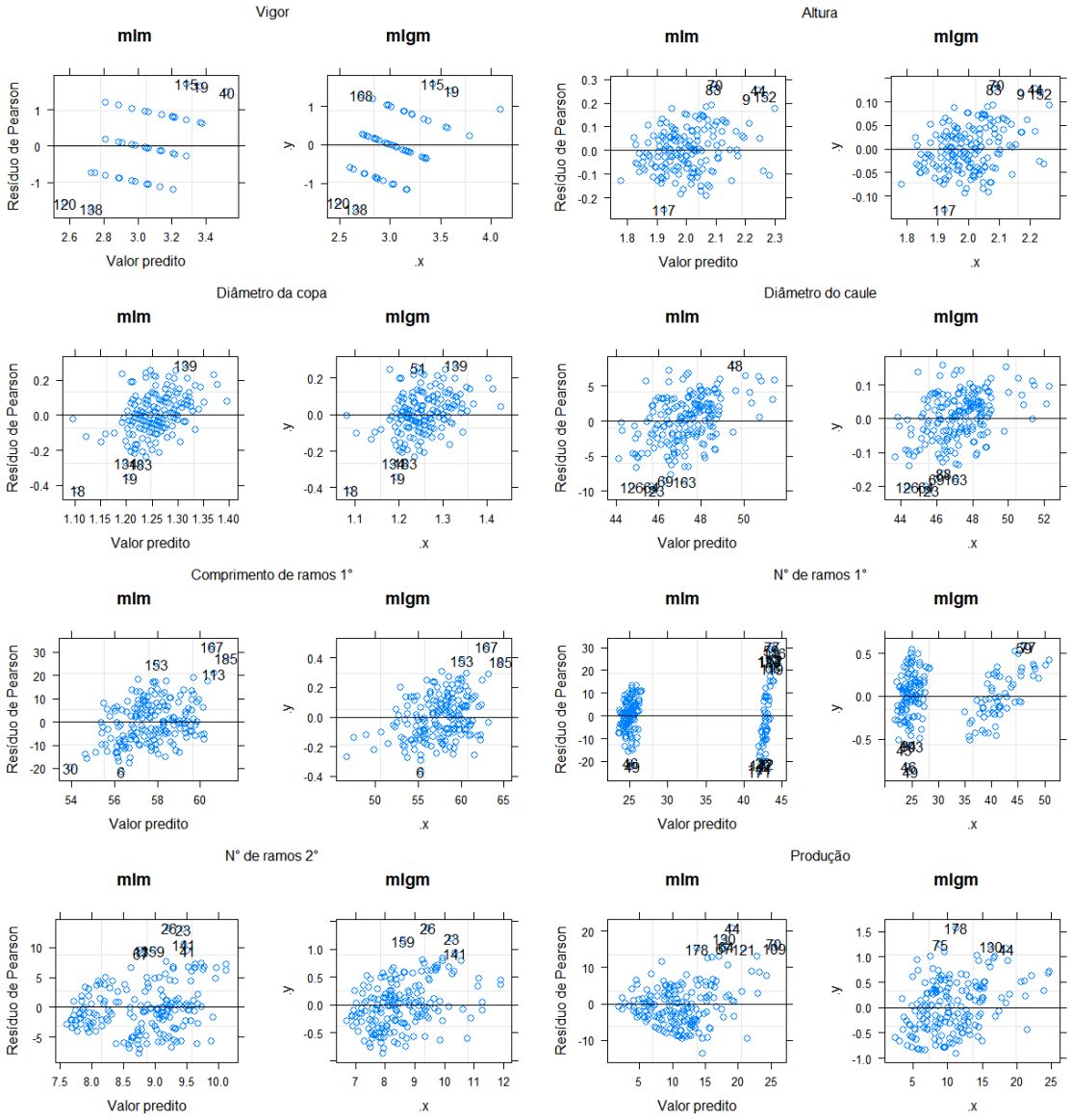


Figura 2. Valores preditos x Resíduos de Pearson para as variáveis vigor (nota visual de 1 à 5), altura (m), diâmetro da copa (m), diâmetro do caule (cm), comprimento de ramos 1° (cm), n° de ramos 1° e n° de ramos 2° ajustadas via MLM e MLGM no ano de 2018.

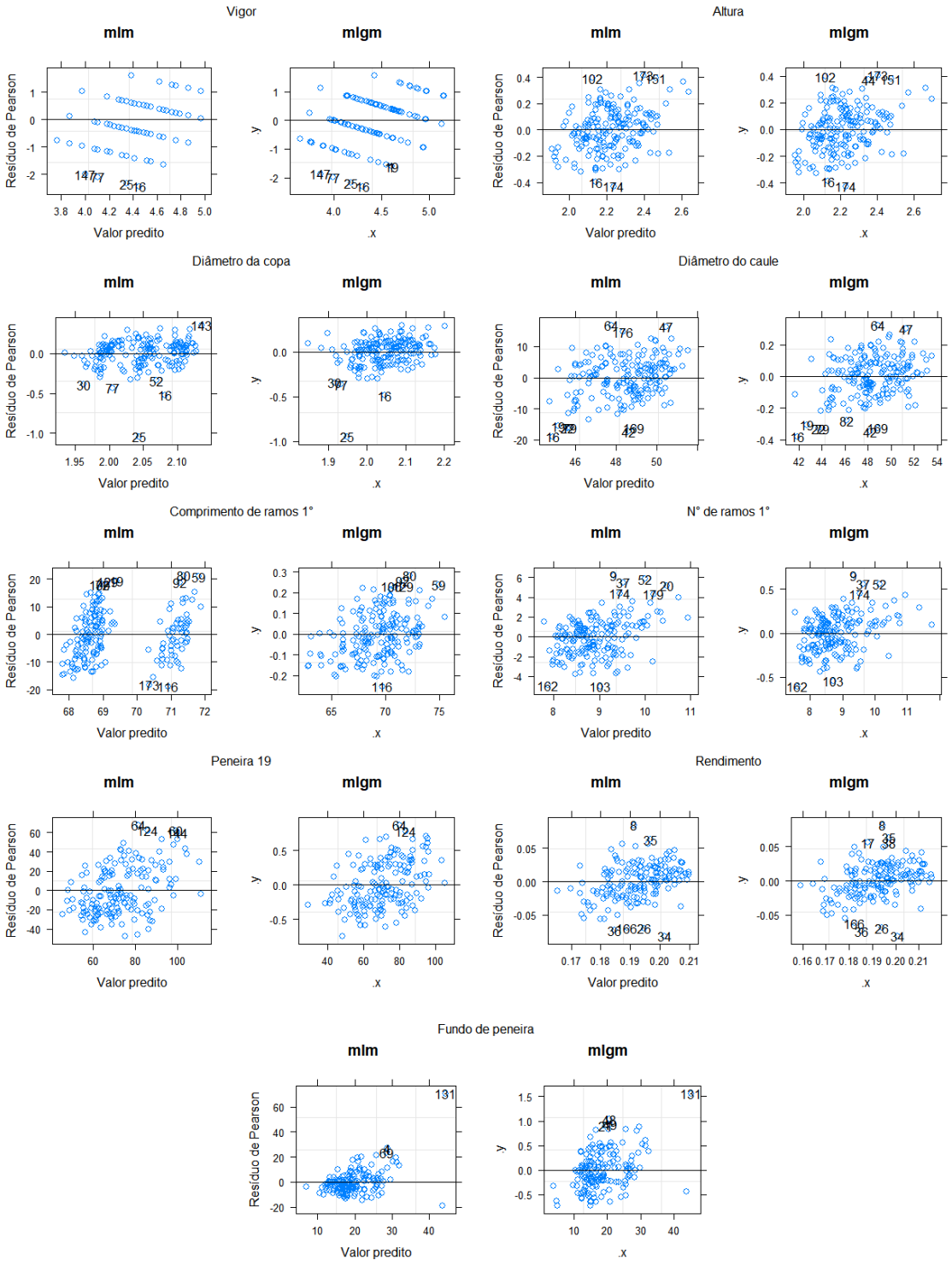


Figura 3. Valores preditos x Resíduos de Pearson para as variáveis vigor (nota visual de 1 à 5), altura (m), diâmetro da copa (m), diâmetro do caule (cm), comprimento de ramos 1° (cm), n° de ramos 1°, peneira 19 (g), rendimento (% de 5L), e fundo de peneira (g) ajustadas via MLM e MLGM no ano de 2019.

4. DISCUSSÃO

Características biológicas do café fazem com que a seleção de genótipos superiores se torne um desafio para programas de melhoramento. Recentemente, os MLGM têm sido utilizado para estudar parâmetros genéticos de algumas espécies animais e vegetais (BOLKER et al., 2009; CAPPA; VARONA, 2013; WILSON et al., 2013; MAKOUANZI et al., 2014; BALSALOBRE et al., 2016; WENG et al., 2016; MELO et al., 2020). Entretanto, até o momento, não há registros do uso de MLGM em estudos genéticos de características de interesse no processo de seleção de *C. arábica*. É importante que estudos comparativos entre MLM e MLGM sejam realizados para compreender os efeitos dessas abordagens para estimativas de parâmetros genéticos importantes (LOPES; HUBER; WHITE, 2000; WENG et al., 2016). Espera-se que os MLGM forneçam estimativas mais verossímeis das variâncias envolvidas no modelo, uma vez que permitem pressuposições mais flexíveis sobre a natureza dos dados.

Desde a proposta dos MLGM, foram sugeridos diferentes métodos para a escolha do modelo mais informativos para os dados (COX, 1961; ATKINSON, 1970; ROYSTON; THOMPSON, 1995). A estratégia escolhida neste trabalho foi o critério de informação de Akaike (AIC), o qual já foi investigado e tem sido preferido na escolha de MLGM candidatos (LINDSEY; JONES, 1998; DICK, 2004; BOLKER et al., 2009). Neste contexto, os gráficos de resíduos de Pearson aqui apresentados, demonstraram que o critério AIC foi eficiente para a escolha dos modelos para explicar os dados, já que os resíduos foram bem menores para os MLGM.

Para algumas variáveis (PROD 2017, AF 2018, NI1 e P11 2019) ao confrontar os MLM e MLGM, o melhor ajuste foi o linear misto, isso pode ser explicado pela própria natureza dos dados, contínua, positiva e predominantemente com valores distantes de 0 (Figura 1). A última característica, em especial, é importante para atender à suposição de simetria e variância constante dos dados (SMITH; CULLIS; THOMPSON, 2005). Em casos diferentes, como em dados contínuos assimétricos, a distribuição gama é uma alternativa para o ajuste de MLGM (MCCULLAHG; NELDER, 1989; TURKMAN, 2000; DICK, 2004; CRAWLEY, 2007). Este fato foi comprovado pelos nossos resultados, em que a maioria das variáveis possuem distribuições assimétricas e foram ajustadas em modelos com distribuição gama.

Uma vez determinada as distribuições que melhor se encaixam aos dados, as funções de ligações são especificadas para relacionar o valor esperado das observações (μ_i) com o preditor linear do modelo (η). A maioria das variáveis analisadas foram ajustadas ao M6 [~Ga(link=inversa)], isso significa que, os valores esperados, estão relacionados às observações de uma forma não linear, ou seja, pela função de ligação inversa $\{g(\mu_i) = \frac{i}{\mu_i}\}$, canônica para distribuição gama (MCCULLAHG; NELDER, 1989). O segundo modelo mais frequente em nossas análises foi o M4 [~Ga(link=identidade)], portanto, para essas variáveis, há uma relação linear simples entre o preditor linear e os valores esperados para as observações (MCCULLAHG; NELDER, 1989; CRAWLEY, 2007). A mesma interpretação é dada às variáveis ajustadas pelo M1 [~N(link=identidade)], no entanto, para observações com distribuição normal.

As funções de ligação ajustadas para as mesmas características avaliadas em dois anos diferentes variaram entre os anos, como ocorreu com VIG, ALT, CAU, CR1 e NR1. Nestes casos, mesmo que a variável tenha a mesma distribuição, pode haver variação da função que relacione a média com o preditor linear do modelo ao longo das avaliações anuais. Neste sentido, os MLGM permitem a flexibilização de diferentes escalas: i) escala latente aos dados; ii) escala esperada dos dados (ajustadas pelas funções de ligação) e iii) escala observada na distribuição dos dados (VILLEMEREUIL et al., 2016). Essa particularidade, portanto, torna os MLGM robustos para melhor explicar as variações dos dados em análise.

Os parâmetros dos MLM e MLGM são estimados via máxima verossimilhança (ML) ou máxima verossimilhança restrita (REML) (STROUP; KACHMAN, 1994). Entretanto, devido à maior complexidade dos MLGM, os métodos quadratura de Gauss-Hermite (PINHEIRO; BATES, 1995) e aproximação de Laplace (WOLFINGER, 1993) são comumente utilizados para maximizar a verossimilhança dos modelos (MCCULLAHG; NELDER, 1989; TURKMAN, 2000; PIEPHO, 2019). O método de Laplace foi utilizado neste estudo, que é a abordagem padrão do pacote glmer do software R®. Foi demonstrado que a aproximação de Laplace é tão boa ou melhor quanto à quadratura de Gauss-Hermite para modelos com função de ligação canônica e efeitos aleatórios multivariados normais (RAUDENBUSH; YANG; YOSEF, 2000). Neste sentido, são necessários estudos que investiguem a eficiência destes métodos para outras famílias de MLGM, por exemplo modelos gama, que foram predominantes em nossos resultados.

A estimação dos componentes de variância via MLGM é um processo complexo, que recentemente vêm sendo discutido com maior atenção, especificamente para modelos de distribuição Poisson (MAKOUANZI et al., 2014; VILLEMEREUIL et al., 2016; PIEPHO, 2019). Foi observado que as características que tiveram CV_r próximos para os MLM e MLGM (VIG, COPA e REND) foram ajustadas às famílias LGM com distribuição normal, independente da função de ligação, ou seja, a escala do preditor linear não interferiu na escala do CV_r para essas variáveis. Essas observações podem ser explicadas pelo fato da variância residual se dar na mesma escala da distribuição assumida para as variáveis (PIEPHO, 2019) Para o restante das características, o CV_r reduziu consideravelmente dos MLM para MLGM, indicando que ao assumir a distribuição gama como mais realista para os dados, houve uma redução do ruído residual e, portanto, maior qualidade de ajuste do modelo. Além disso, foi relatado que os parâmetros das distribuições assumidas para as variáveis, neste caso normal (μ,σ) e gama (μ,ν), podem interferir na variação estimada para os componentes do preditor linear do modelo (VILLEMEREUIL et al., 2016).

Também foi observada uma diminuição dos coeficientes de variação genética (CV_g) estimados via MLGM. Resultados semelhantes foram registrados para modelo Binomial (logit) em *Eucaliptus* sp. (MAKOUANZI et al., 2014). Entretanto, para algumas variáveis houve aumento do CV_g estimado. É importante considerar que MLGM podem assumir variações maiores que as variações latentes e esperadas no momento em que uma distribuição é atribuída aos dados (VILLEMEREUIL et al., 2016). Este aspecto não foi contemplado em nossos resultados, uma vez que a maioria das características com maior CV_g foram ajustadas em modelos com funções de ligação identidade, ou seja, não há mudanças nas escalas entre os dados observados e esperados para o modelo. Para as outras características ajustadas com as funções de ligação inversa e log (ALT, CR1, FUN E P100 em 2019, CAU, NR1, NI1 e PROD em 2018) não houve aumento da variação. Portanto, em nossos resultados, não foi possível observar a interferência das escalas das funções log e inversa na estimação das variâncias genéticas para as características avaliadas. Também foi observado que mesmo com baixos CV_g estimados via MLGM, houve maior detecção dos efeitos de genótipos com a utilização destes modelos. Neste sentido, é necessário refletir que ao adotar os MLM como ferramenta, os melhoristas podem assumir maiores influências genéticas na variação das características avaliadas, porém, podem não detectar variabilidade genética em seus materiais.

Estimar a herdabilidade (H^2) para uma característica é fundamental por quantificar a proporção transmissível do fenótipo e determinar a resposta a seleção (PIEPHO; MÖHRING, 2007; MARTÍNEZ-GARCÍA et al., 2017). As H^2 estimadas via MLM e MLGM indicaram a P19, FUN e P100 como as características com maior variação fenotípica atribuídas a causas genéticas. As H^2 calculadas via MLM foram predominantemente maiores que em MLGM, fato que pode ser explicado pela maior variação genética estimada pelos MLM para essas características. Além disso, as diferenças dos valores de H^2 são esperadas porque as estimativas não são feitas na mesma escala, dificultando a interpretação dos resultados (NAKAGAWA; SCHIELZETH, 2010; VILLEMEREUIL et al., 2016).

Para algumas variáveis (CAU, CR1, P19, FUN, REN e P100) houve um salto da H^2 ($\geq$0,9) estimada via MLGM. Este padrão também já foi observado em estudos com modelos Binomiais (LOPES; HUBER; WHITE, 2000), Poisson para *Eucalyptus* sp. (MAKOUANZI et al., 2014) e *Pinus banksiana* (WENG et al., 2016), porém, a acurácia seletiva variou entre MLM e MLGM para estes estudos. Portanto, é discutível a vantagem do uso de MLGM para estimar a H^2, uma vez que, exceto para as características citadas anteriormente, os valores estimados para as outras variáveis foram baixos (0 a 0,3). Em teoria, os baixos valores de H^2 observados tanto em MLM quanto em MLGM podem ser explicados pela alta variação experimental, refletindo baixa estabilidade durante a coleta, sensibilidade das plantas de café às mudanças ambientais e baixa variância genética detectada pelos modelos (MAKOUANZI et al., 2014).

Diferente dos outros resultados, os valores de BLUP foram extremamente correlacionados, mesmo que de forma não linear, para todas as características. A correlação alta entre valores genotípicos estimados via MLM e MLGM também foram registrados para estudos de variáveis categóricas no melhoramento florestal (CAPPA; VARONA, 2013) e em *Pinus banksiana* (WENG et al., 2016). O principal objetivo de computar valores de BLUP é predizer valores genotípicos que irão direcionar a seleção de genótipos superiores para determinados caracteres de interesse (PIEPHO; MÖHRING, 2007). Nossos resultados demonstraram, portanto, que não houve diferenças entre a aplicação de MLM e MLGM no ranqueamento dos genótipos para as características aqui avaliadas.

Também foi observada uma correlação negativa entre os valores de BLUP para as variáveis ajustadas a MLGM com função de ligação inversa. Esse fenômeno é explicado por uma das propriedades matemáticas do BLUP (WENG et al., 2016). Os componentes aleatórios

de modelos mistos, neste caso os efeitos genotípicos, se tornam parte da média, e então são calculados (PIEPHO; MÖHRING, 2007). No caso dos MLGM, a média é relacionada aos efeitos do modelo através da função de ligação, explicando a correlação negativa para MLGM com função inversa. Entretanto, não foi observada diferenças entre os valores de BLUP para as características aqui ajustadas à MLGM com função de ligação logarítmica. Ainda não existem interpretações desse tipo de relação entre valores de BLUP e as escalas envolvidas nos MLGM.

5. CONCLUSÕES

Os MLGM foram aplicáveis a quase todas as características avaliadas para os genótipos de *C. arabica*. Embora a interpretação dos componentes de variância estimados pelos MLGM, sobretudo modelos gama, ainda não seja consolidada, as famílias de MLGM com a distribuição gama parecem ser uma alternativa viável para modelar dados contínuos assimétricos e com variância heterogênea. Os parâmetros estimados via MLM e MLGM variam muito entre si, exceto os valores de BLUP. A variância residual estimada pelos MLGM é muito menor, o que pode indicar um ajuste mais adequado aos dados. Em geral, os MLGM retornam menores variações genéticas, mas são mais sensíveis para detectar o efeito de genótipo na variação dos dados fenotípicos. O comportamento das H^2 estimadas via MLM e MLGM é muito variável e de difícil interpretação, uma vez que as escalas envolvidas nos MLGM interferem no cálculo. Mais estudos a respeito destes aspectos precisam ser feitos para melhorar o entendimento das aplicações destes modelos.

Mesmo que os parâmetros genéticos estimados sejam muito diferentes, os modelos não interferem no ranqueamento dos genótipos, portanto, estudos a fim de selecionar genótipos superiores para as características de interesse no café, podem se restringir a utilização dos MLM, uma vez que possuem um *workflow* computacional e interpretação mais simples e mais rápidas. Os MLGM possuem vantagens matemáticas em relação aos MLM por considerarem distribuições diferentes da normal e funções de ligações para modelagem dos dados, e apesar da complexidade computacional e teórica dos MLGM, essa ferramenta pode ser utilizada para a complementação dos estudos genéticos envolvidos no processo de melhoramento do café.

6. REFERÊNCIAS

AKAIKE, H. A New Look at the Statistical Model Identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974.

ATKINSON, A. C. A Method For Discriminating Between Models. **Royal statistical society**, v. 32, n. 3, p. 323–353, 1970.

BALSALOBRE, T. W. A. et al. Mixed modeling of yield components and brown rust resistance in sugarcane families. **Agronomy Journal**, v. 108, n. 5, p. 1824–1837, 2016.

BOLKER, B. M. et al. Generalized linear mixed models: a practical guide for ecology and evolution. **Trends in Ecology and Evolution**, v. 24, n. 3, p. 127–135, 2009.

BOTELHO, C. E. et al. Adaptability and phenotype stability of Arabica coffee cultivars in Minas Gerais, Brazil. **Pesquisa Agropecuaria Brasileira**, v. 45, n. 12, p. 1404–1411, 2010.

CAPPA, E. P.; VARONA, L. An assessor-specific Bayesian multi-threshold mixed model for analyzing ordered categorical traits in tree breeding. **Tree Genetics and Genomes**, v. 9, n. 6, p. 1423–1434, 2013.

CARVALHO, C. H. S. **Cultivares de café: origem, características e recomendações**. Brasília, 2008.

COX, D. R. Separate Families of Hypotheses. **Proceedings of the fourth Berkeley symposium on mathematical statistics and probability.**, v. 1, n. 2, p. 96, 1961.

CRAWLEY, M. J. The R Book. 1° ed. London: John Wiley & Sons, 2007.

DICK, E. J. Beyond "lognormal versus gama": Discrimination among error distributions for generalized linear models. **Fisheries Research**, v. 70, n. 2- 3 SPEC. ISS., p. 351–366, 2004.

FERREIRA, R. T. et al. Seleção recorrente intrapopulacional em maracujazeiro-azedo via modelos mistos. **Revista Brasileira de Fruticultura**, v. 38, n. 1, p. 158–166, 2016.

GILES, J. A. et al. Divergence and genetic parameters between coffea sp. genotypes based in foliar morpho-anatomical traits. **Scientia Horticulturae**, v. 245, n. May 2018, p. 231–236, 2019.

GILES, J. A. et al. Genetic diversity of promising 'conilon' coffee clones based on morpho-agronomic variables. **Anais da Academia Brasileira de Ciencias**, v. 90, n. 2, p. 2437–2446, 2018.

GUEDES, J. M. et al. Divergência genética entre cafeeiros do germoplasma Maragogipe. **Bragantia**, v. 72, n. 2, p. 127–132, 2013.

HENDERSON, C. R. Best Linear Unbiased Estimation and Prediction under a Selection Model Author (s): C . R . Henderson Published by : International Biometric Society Stable URL : https://www.jstor.org/stable/2529430 International Biometric Society is collaborating with J. **Biometrics**, v. 31, n. 2, p. 423–447, 1975.

JANSEN, J. Generalized linear mixed models and their application in plant breeding research. **Eindhoven: Technische Universiteit Eindhoven**, p. 143, 1993.

LINDSEY, J. K.; JONES, B. Choosing among generalized linear models applied to medical data. **Statistics in Medicine**, v. 17, n. 1, p. 59–68, 1998.

LOPES, U. V.; HUBER, D. A.; WHITE, T. L. Comparison of methods for prediction of genetic gain from mass selection on binary threshold traits. **Silvae Genetica**, v. 49, n. 1, p. 50–56, 2000.

MAKOUANZI, G. et al. Assessing the additive and dominance genetic effects of vegetative propagation ability in Eucalyptus—influence of modeling on genetic gain. **Tree Genetics and Genomes**, v. 10, n. 5, p. 1243–1256, 2014.

MARTÍNEZ-GARCÍA, P. J. et al. Predicting breeding values and genetic components using generalized linear mixed models for categorical and continuous traits in walnut (Juglans regia). **Tree Genetics and Genomes**, v. 13, n. 5, 2017.

MCCULLAHG, P.; NELDER, J. A. Generalized linear models. 2° ed. London: Chapman and Hall, 1989.

MELO, R. C. DE et al. Statistical model assumptions achieved by linear models: classics and generalized mixed. **Revista Ciência Agronômica**, v. 51, n. 1, p. 1–9, 2020.

NAKAGAWA, S.; SCHIELZETH, H. Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. **Biological Reviews**, v. 85, n. 4, p. 935–956, 2010.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society**, v. 135, p. 370–384, 1972.

PIEPHO, H. P. A coefficient of determination (R2) for generalized linear mixed models. **Biometrical Journal**, v. 61, n. 4, p. 860–872, 2019.

PIEPHO, H. P.; ECKL, T. Analysis of series of variety trials with perennial crops. **Grass and Forage Science**, v. 69, n. 3, p. 431–440, 2014.

PIEPHO, H. P.; MÖHRING, J. Computing heritability and selection response from unbalanced plant breeding trials. **Genetics**, v. 177, n. 3, p. 1881–1888, 2007.

PINHEIRO, J. C.; BATES, D. M. Approximations to the log-likelihood function in the nonlinear mixed-effects model. **Journal of Computational and Graphical Statistics**, v. 4, n. 1, p. 12–35, 1995.

RAUDENBUSH, S. W.; YANG, M. L.; YOSEF, M. Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Aproximation. **Journal of Computational and Graphical Statistics**. v.9, n. 1, p. 141-157, 2000.

RESENDE, M. et al. Estimativas de parâmetros genéticos e predição de valores genotípicos no melhoramento do cafeeiro pelo procedimento REML/BLUP. **Bragantia**. v. 60, n. 3, p. 185–193, 2001.

RESENDE, M.; DUARTE, J. Precisão E Controle De Qualidade Em Experimentos De Avaliação De Cultivares. **Pesquisa Agropecuária Tropical (Agricultural Research in the Tropics)**, v. 37, n. 3, p. 182–194, 2007.

RODRIGUES, W. P. et al. Assessment of genetic divergence among coffee genotypes by Ward-MLM procedure in association with mixed models. **Genetics and Molecular Research**, v. 15,

n. 2, 2016.

ROYSTON, P.; THOMPSON, S. G. Comparing Non-Nested Regression Models. v. 51, n. 1, p. 114–127, 1995.

R DEVELOPMENT CORE TEAM. R. a language and environment for statistical computing. R foundation for Statistical Computin, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

SMITH, A. B.; CULLIS, B. R.; THOMPSON, R. The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. **Journal of Agricultural Science**, v. 143, n. 6, p. 449–462, 2005.

SILVA, D. O. et al. Genetic progress with selection of *Coffea canephora* clones of superior processed coffe yield. **Ciência Rural**. v48. n. 3, 2018.

SOUSA, T. V. et al. Early selection enabled by the implementation of genomic selection in coffea arabica breeding. **Frontiers in Plant Science**, v. 9, n. January, p. 1–12, 2019.

STROUP, W. W.; KACHMAN, S. D. Generalized Linear Mixed Models - an Overview. **Conference on Applied Statistics in Agriculture**, 1994.

TASSONE, G. A. T. et al. Simultaneous selection in coffee progenies of mundot novo by selection indices. **Coffee Science**, v. 14, n. 1, p. 83–92, 2019.

TURKMAN, M. A. A. Modelos Lineares Generalizados: da teoria à prática. **VIII Congresso Anual da Sociedade Portuguesa de Estatística**, p. 153, 2000.

VILLEMEREUIL, P. et al. General methods for evolutionary quantitative genetic inference from generalized mixed models. **Genetics**, v. 204, n. 3, p. 1281–1294, 2016.

WENG, Y. et al. Genetic Parameters for Bole Straightness and Branch Angle in Jack Pine Estimated Using Linear and Generalized Linear Mixed Models. **Forest Science**, v. 63, n. 1, p. 111–117, 2016.

WILSON, B. J. et al. Estimated breeding values for canine hip dysplasia radiographic traits in a cohort of Australian German Shepherd dogs. **PloS one**, v. 8, n. 10, 2013.

WOLFINGER, R. Laplace's approximation for nonlinear mixed models. **Biometrika**, v. 80, n. 4, p. 791–795, 1993.

APÊNDICES

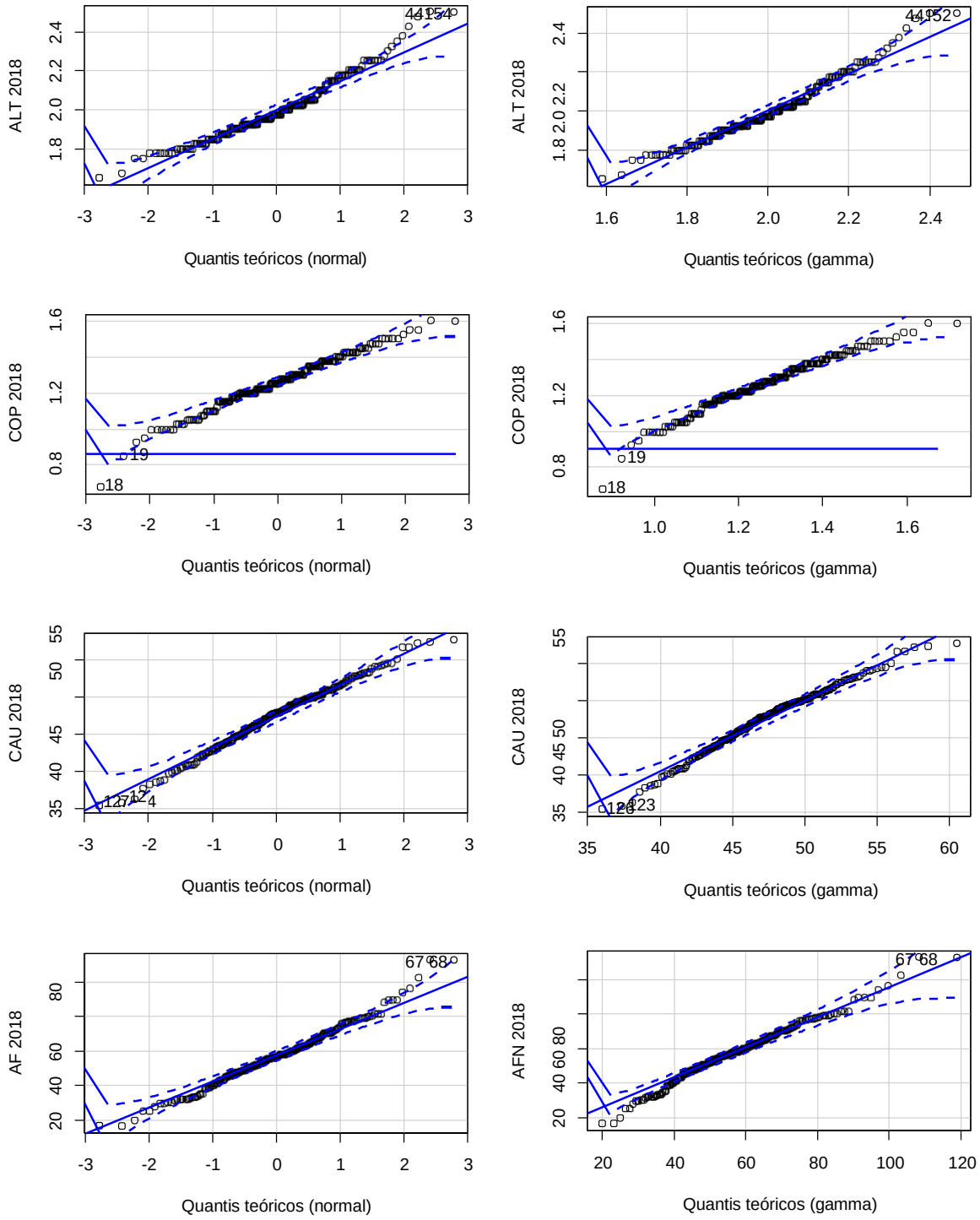
Apêndice A - Identificação das variedades, código original e origem (SETOTAW et al. 2013) referente aos 92 genótipos avaliados no estudo.

Genótipo	Variedade	Código original	Origem
1	Arara	44	Obatã x Icatu amarelo 2944
2	Sábia 398	26	Acaiá (Mundo Novo) x Catimor UFV 386
3	Icatu 3696	71	*C. canephora* x *C. arabica*
4	Siriema 14/8	10	*C. racemosa* x *C. arabica* cv. Blue Mountain
5	Catucaí Amarelo 3 SM	04	Icatu amarelo x Catuaí amarelo
6	Siriema 5/14	35	*C. racemos*a x *C. arabica* cv. Blue Mountain
7	Catucaí Vermelho 36/6	69	Icatu vermelho x Catuaí vermelho
8	Catucaí Amarelo	15	Icatu amarelo x Catuaí amarelo
9	Maracatia Vermelho	47	Acaiá x Catuaí Vermelho IAC 81
10	Maracatia Amarelo	47	Acaiá x Catuaí Vermelho IAC 81
11	Acauã 37	17	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
12	Catucaí Vermelho 20/15	18	Icatu vermelho x Catuaí vermelho
13	Acauã 65	27	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
14	Acauã 68/11	63	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
15	Icatu Amarelo 925	39	*C. canephora* x *C. arabica*
16	Icatu Vermelho 925	39	*C. canephora* x *C. arabica*
17	Catucaí Amarelo 2 SL	1	Icatu amarelo x Catuaí amarelo
18	Catuaí Vermelho 19/8	57	Caturra amarelo (IAC 467-11) x Mundo Novo (IAC 374-19)
19	Catucaí	11	Desconhecida
20	Catucaí x Catimor 357-77 (5/33)	16	Catuaí x Catimor 357-77 (5/33) FSA
21	Catucaí Amarelo 24/137	02	Icatu x Catuaí
22	Acauã Laranja	38	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
23	Acauã Amarelo	38	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
24	Acauã Vermelho	38	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
25	Acauã	55	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
26	Catucaí Amarelo 24/137	07	Icatu x Catuaí
27	Catucaí Vermelho 20/15	05	Icatu x Catuaí
28	Acauã	49	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
29	Palma III	56	Catuaí Vermelho IAC 81 x Catimor (UFV 353)
30	Catucaí Amarelo	43	Icatu x Catuaí
31	Acauã 68-2	61	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
32	Bem Te Vi	77	Catimor 391 x Catuaí amarelo 74
33	Icatu 925	73	*C. canephora* x *C. arabica*
34	Siriema Vermelho	58	*C. racemos*a x *C. arabica* cv. Blue Mountain
35	Siriema Amarelo	58	*C. racemos*a x *C. arabica* cv. Blue Mountain
36	Catucaí 785/15	09	Icatu Vermelho 785 x Catuaí vermelho
37	Catucaí Amarelo 3/5	81	Icatu x Catuaí
38	Catucaí Amarelo 3/5 LVA	81	Icatu x Catuaí
39	Catucaí Amarelo	01	Icatu x Catuaí
40	Catucaí Vermelho 20/15	21	Icatu x Catuaí
41	Acauã 3%	52	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
42	Acauã 5%	51	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
43	Catucaí Vermelho	86	Icatu x Catuaí
44	Acauã Amarelo (Tardio)	48	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
45	Acauã Amarelo (Médio)	48	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
46	Araponga	60	Catuaí amarelo IAC 86 x HT UFV 446-08
47	Acauã Amarelo	50	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
48	Icatu 925	83	*C. canephora* x *C. arabica*
49	Acauã	46	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
50	Catuaí 66	84	Caturra amarelo (IAC 476-11) x Mundo Novo (IAC 374-19)
51	Acauã Amarelo	54	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
52	Acauã Vermelho	54	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)

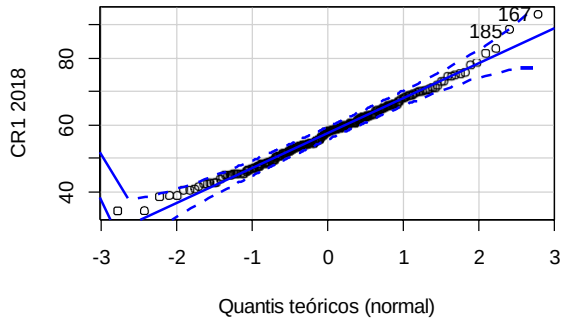
Continuação apêndice A -

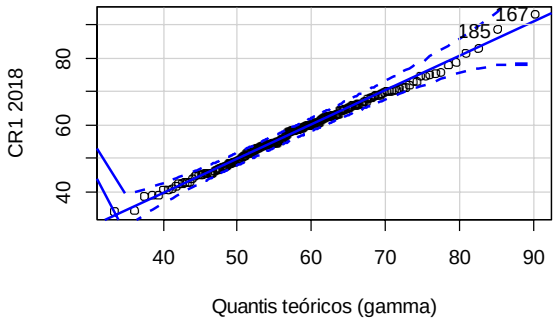
Genótipo	Variedade	Código original	Origem
53	IPR 103	N22	Catuaí x Icatu
54	Catucaí Vermelho 19/8	40	Icatu x Catuaí
55	3 SM	08	Desconhecida
56	Catucaí Amarelo 20/15	30	Icatu x Catuaí
57	Catucaí Vermelho 36/69	24	Icatu x Catuaí
58	Acauã Amarelo FSA	22	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
59	Catucaí Roxinho	82	Icatu x Catuaí
60	IPR 98	N19	Villa Sarchi CIFC 971/10 x HT CIFC 832/2
61	Acauã 5/20	28	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
62	Bourbon Vermelho	14	Desconhecida

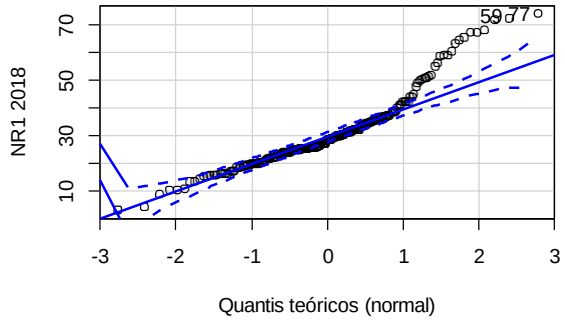
Apêndice B - Gráficos quantil-quantil para as distribuições normal e gama para todas as variáveis avaliadas nos anos de 2017, 2018 e 2019.

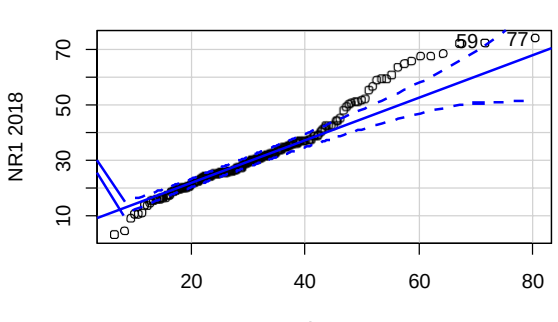


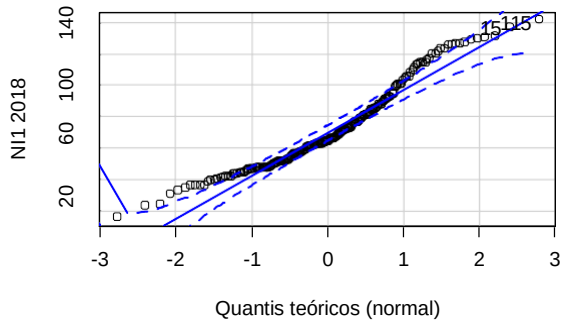
Continuação apêndice B -

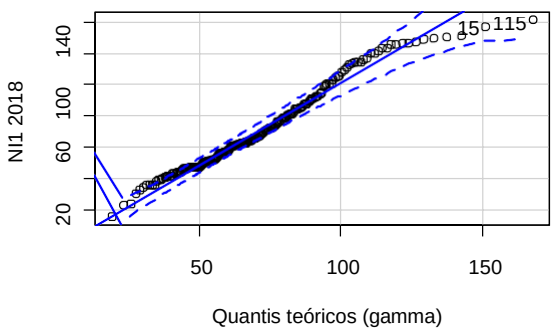


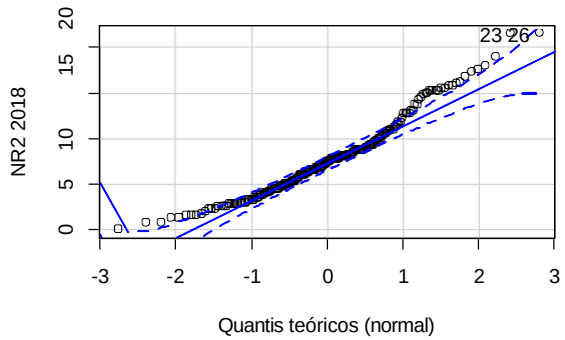


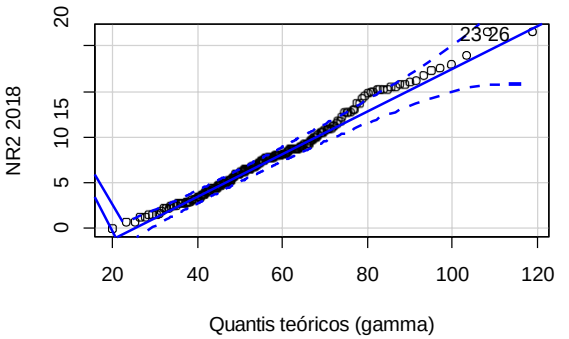


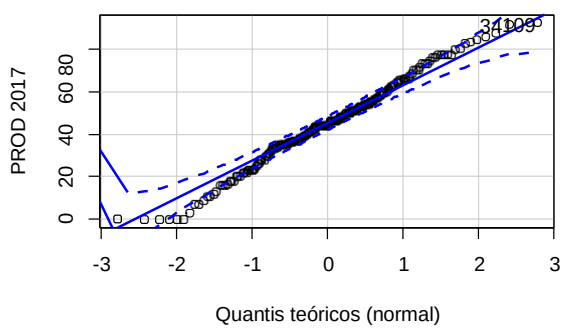


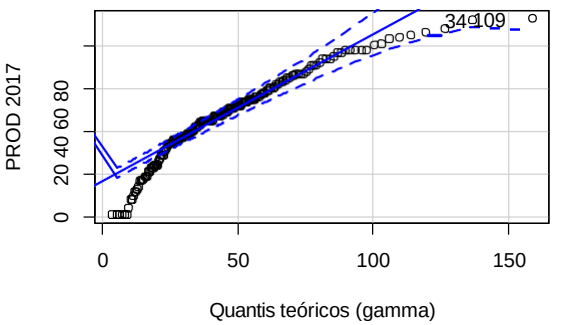




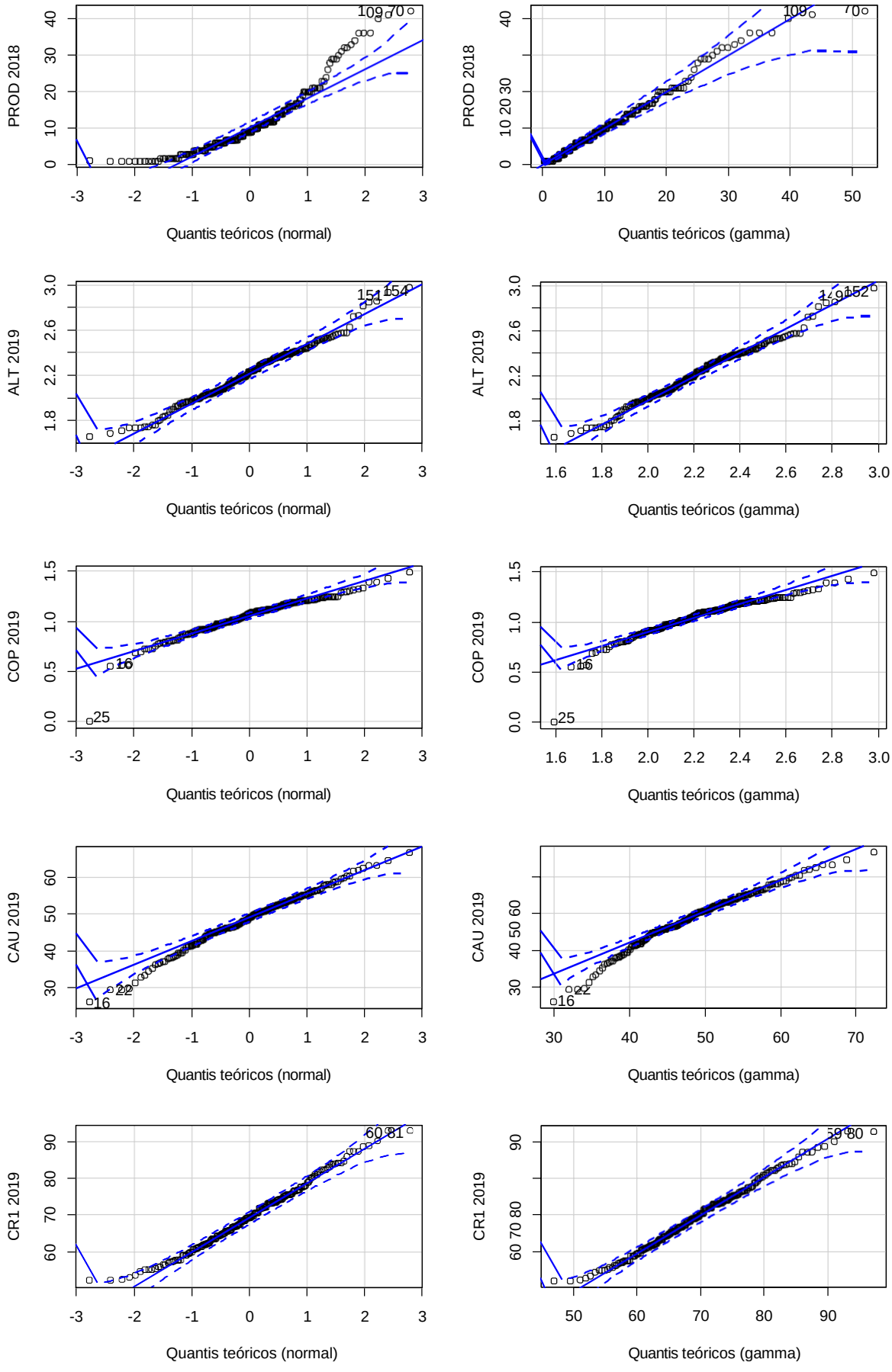




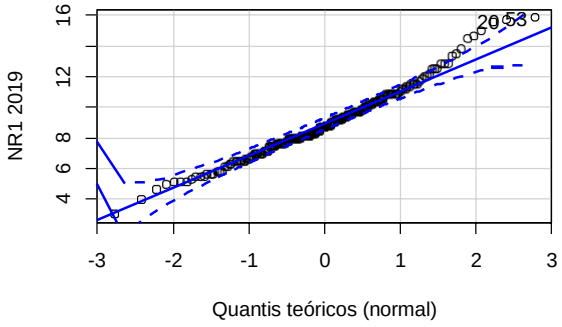


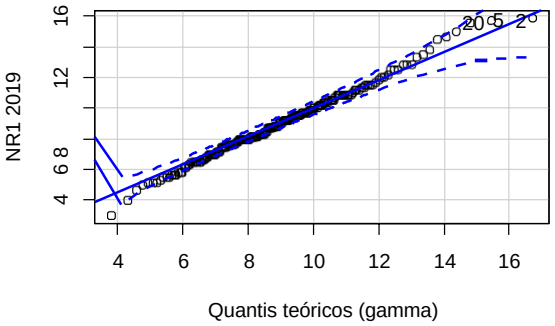


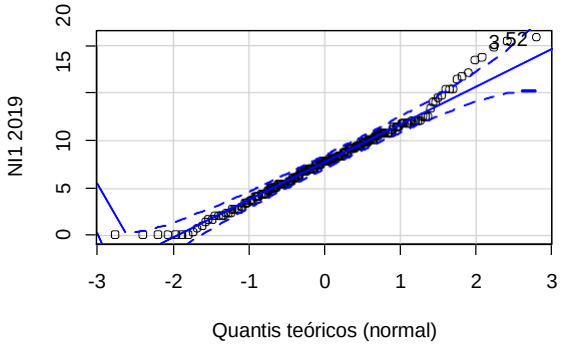
Continuação apêndice B -

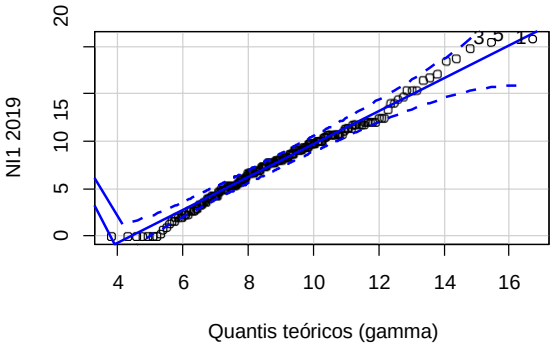


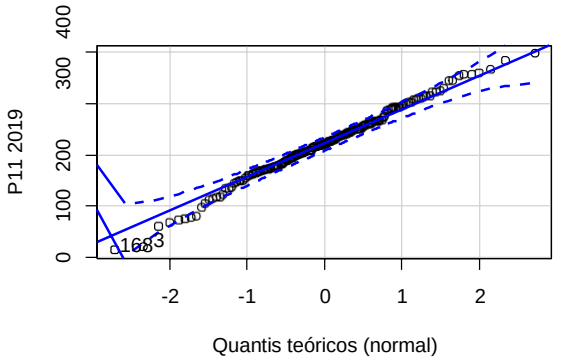
Continuação apêndice B -

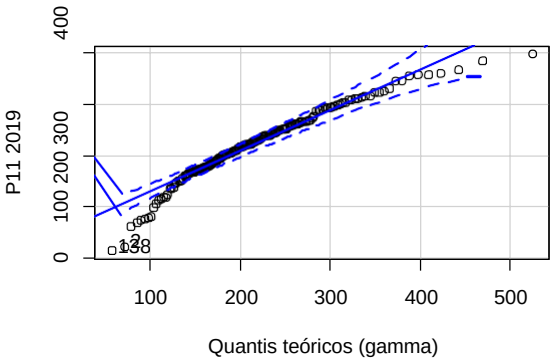


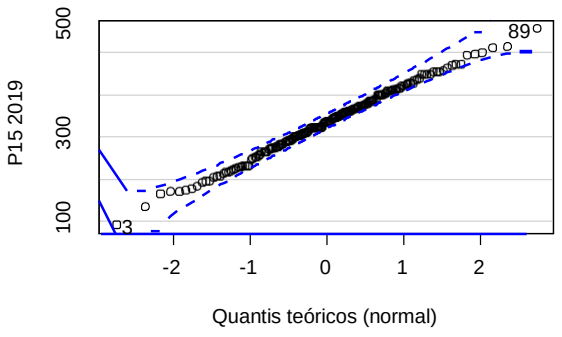


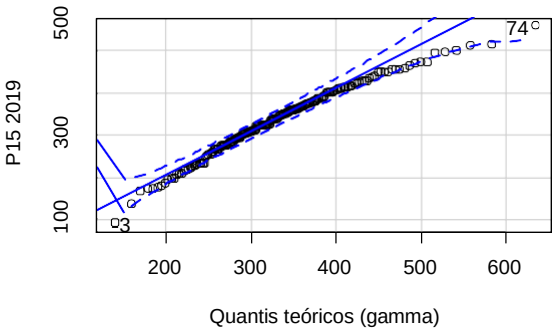


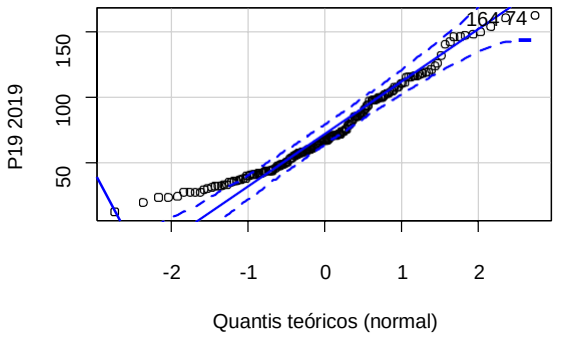


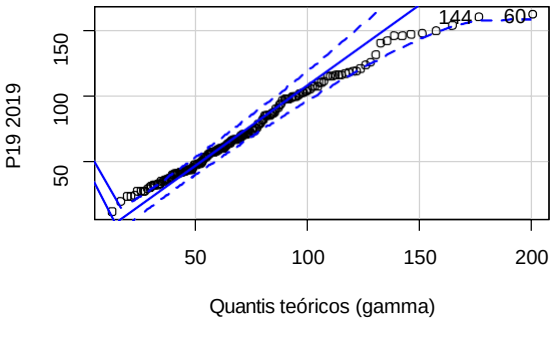




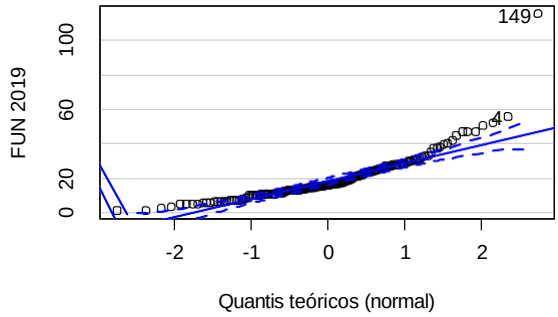


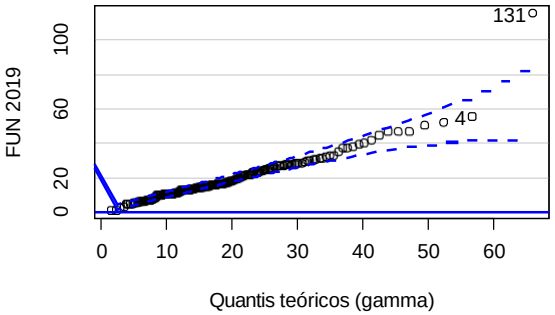


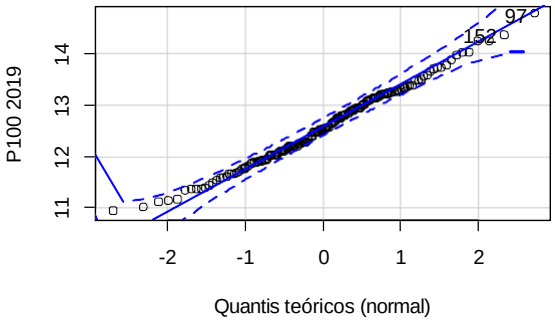


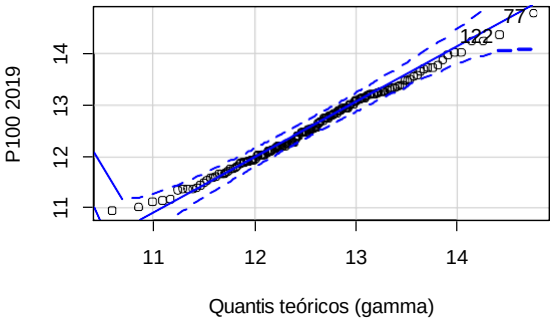


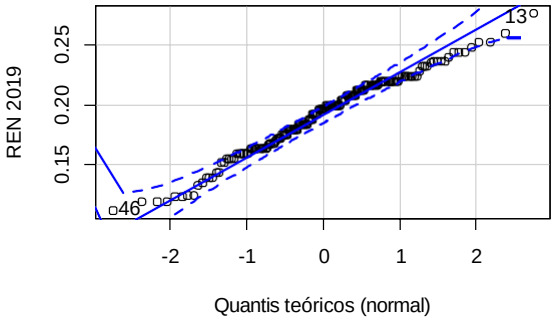
Continuação apêndice B -

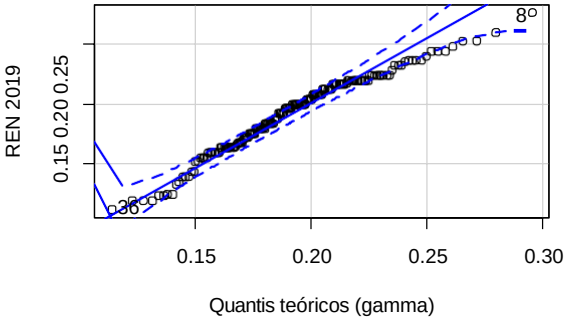












CAPÍTULO 2

MAPAS AUTO-ORGANIZÁVEIS DE KOHONEN PARA ESTUDO DE DIVERSIDADE GENÉTICA EM VARIEDADES DE *C. arabica*

RESUMO

DE ASSIS, Mayumi Furuya, M.Sc., Universidade Federal de Viçosa, agosto de 2020. **Mapas auto-organizáveis de Kohonen para estudo de diversidade genética em variedades de *Coffea. arabica***. Orientador: Pedro Ivo Vieira Good God. Coorientador: Everaldo Antônio Lopes.

Os programas de melhoramento genético do café, especificamente o *C. arabica*, enfrentam o desafio de assegurar a diversidade genética dos bancos de genótipos para garantir a obtenção de variedades de alta qualidade. Os métodos multivariados clássicos, como as análises de CP e agrupamentos hierárquicos são os mais comuns nos estudos de diversidade genética do café. Entretanto, esses métodos podem apresentar resultados pouco informativos, devido a capacidade reduzida de processar um volume de dados complexos sem distorções e de explicar as variações contidas nos dados de forma satisfatória. Esses aspectos exigem a utilização de técnicas mais avançadas para os estudos de diversidade. Os mapas auto-organizáveis (SOM) de Kohonen é um tipo de rede neural capaz de representar em baixas dimensões as distâncias entre as amostras de um espaço amostral multidimensional com o mínimo de distorção, a partir de um conjunto de dados diverso e complexo. O objetivo deste estudo foi avaliar a aplicabilidade das redes neurais SOM integradas aos métodos multivariados clássicos na exploração da diversidade genética e o comportamento de características biométricas de 62 variedades experimentais de *C. arabica*. O material genético em estudo possui variabilidade genética, detectado tanto pelos métodos convencionais quanto pelos SOM. Os agrupamentos dos genótipos de acordo som suas similaridades pelo método UPGMA e pelos SOM foram divergentes. Além disso, as redes neurais possibilitaram diferentes interpretações a respeito das características biométricas avaliadas, como correlações entre as características, identificação de genótipos superiores para diferentes padrões fenotípicos, quantificar a dissimilaridade entre os grupos de genótipos formados e ainda observar a variabilidade genética do material relacionando-os a seus parentais. Os resultados obtidos confirmam a aplicabilidade dos SOM em programas de melhoramento para auxiliar na exploração da variabilidade genética dos materiais, garantindo a obtenção de genótipos superiores e a manutenção da diversidade das variedades de *Coffea arabica*.

Palavras-chave: Métodos multivariados, mapas auto-organizáveis, melhoramento genético do café, variabilidade genética.

ABSTRACT

DE ASSIS, Mayumi Furuya, M.Sc., Universidade Federal de Viçosa, august, 2020. **Kohonen self-organizing maps for studying genetic diversity in *Coffea arabica* varieties** Adviser: Pedro Ivo Vieira Good God. Co-adviser: Everaldo Antônio Lopes.

Coffee breeding programs, specially of *Coffea arabica*, face the challenge of ensuring the genetic diversity of genotype to guarantee the obtaining of high-quality varieties. Classical multivariate methods, such as PC analysis and hierarchical clusters, are the most common in studies of the genetic diversity of coffee. However, these methods may present poor information, due to the reduced ability to process complex data without distortions and to explain the variations contained in the data in a satisfactory way. These aspects require the use of more advanced techniques for diversity studies. Self-organizing maps (SOM) proposed by Kohonen, is a type of neural network capable of representing in small dimensions the distances between samples in a multidimensional sample space with minimal distortion, based on a diverse and complex data set. The aim of this study was to evaluate the applicability of SOM neural networks integrated with classical multivariate methods in the exploration of genetic diversity and the behavior of biometric characteristics of 62 experimental varieties of *Coffea arabica*. The genetic material presented genetic variability, detected by both conventional and neural network methods. The groupings of the genotypes according to their similarities by the UPGMA method and by the SOM were divergent. In addition, the neural networks allowed different interpretations regarding the biometric characteristics evaluated, such as correlations between the characteristics, identification of superior genotypes for different phenotypic patterns, quantifying the dissimilarity between the groups of genotypes formed and also observing the genetic variability of the material relating those to their parents. The results obtained confirm the applicability of SOM in breeding programs to integrate the genetic variability studies, guaranteeing the achievement of superior genotypes and maintaining the diversity of *Coffea arabica* varieties.

Keywords: Multivariate methods, self-organizing maps, coffee genetic breeding, genetic variability.

1. INTRODUÇÃO

No melhoramento genético de *Coffea arabica* considera-se que há pouca diversidade genética, devido à baixa variabilidade em linhagens ancestrais e à escolha frequente da variedade Bourbon Vermelho como parental (SETOTAW et al., 2013). Estes aspectos podem resultar em uma menor eficiência na seleção de cultivares produtivas, adaptadas e resistentes a estresses bióticos e abióticos (ZHOU et al., 2002; SETOTAW et al., 2013). Para maximizar o potencial de ganho e a escolha de germoplasma elite, é fundamental conhecer a estrutura e diversidade populacional. Isso resulta na seleção de genótipos divergentes para cruzamentos, garantindo a manutenção de uma base populacional menos vulnerável geneticamente e aumentando a probabilidade de obtenção de genótipos superiores (BARBOSA et al., 2011).

No estudo da diversidade genética de populações, é muto comum o uso de técnicas multivariadas, como componentes principais (CP) e técnicas de agrupamentos hierárquicos (GUEDES et al., 2013; TEIXEIRA et al., 2013; RODRIGUES et al., 2016; MACHADO et al., 2017; GILES et al., 2019). Entretanto, mesmo que estes métodos sejam massivamente utilizados e tenham fornecido importantes diretrizes em programas de melhoramento, os mesmos apresentam algumas limitações. Por exemplo, foi demonstrado que análises de CP podem apresentar baixa resolução para explicar quantidade suficiente de variação contidas em conjuntos de dados (PARK et al., 2004). Também foi registrado que os métodos de agrupamentos hierárquicos apresentam baixa acurácia em determinar relacionamentos entre os genótipos, devido capacidade reduzida da representação de dados multidimensionais e com relações não lineares entre si (IBRAHIM et al., 2016; SPANOGHE et al., 2020). Além disso, estes métodos de classificação e ordenação tendem a resultar em distorções frente a dados desbalanceados ou contendo *outliers* (BARBOSA et al., 2011). Portanto, essas desvantagens podem resultar em conclusões equivocadas e limitarem as interpretações possíveis levando a agrupamentos pouco reais ou inespecíficos dos genótipos de uma população de melhoramento. Estes aspectos exigem a utilização de ferramentas estatísticas ou computacionais mais robustas, refinadas e que sejam capazes de fornecer um maior volume de informações para os melhoristas.

O *self-organazing map* (SOM) propostos por Kohonen é um tipo de rede neural utilizado como método de análise de dados capaz de lidar com um volume de informações multidimensionais complexos e com relação não lineares entre si (KOHONEN, 1982, 2014). A

principal função do SOM é o processamento e organização de dados de um espaço multidimensional para um espaço bidimensional, preservando a topologia dos dados originais de entrada, ou seja, mantendo a mesma distância entre os indivíduos em um espaço com mais dimensões, refletindo a similaridade entre os mesmos (WEHRENS, 2007; KOHONEN, 2014). As características dos SOM possibilitaram sua rápida adoção em diferentes campos tecnológicos (KOHONEN, 2014) e sobretudo para o melhoramento de espécies vegetais e animais (ZHAO et al., 2005; ROUX et al., 2007; BARBOSA et al., 2011; IBRAHIM et al., 2016; DOS SANTOS et al., 2019; SPANOGHE et al., 2020).

Devido aos princípios dos SOM, esse método é capaz de capturar diferentes padrões nas estruturas dos dados, classificar dados com poucas distorções e ainda extrair de forma eficiente informações, uma vez que não são necessários conhecimentos prévios da organização dos dados, seguindo o princípio do aprendizado não supervisionado de máquinas (RICHARDSON; RISI EN; SHILLINGTON, 2003; KOHONEN, 2014; SPANOGHE et al., 2020). Atualmente, os SOM têm sido explorados para estudos de diversidade genética de diferentes espécies vegetais e animais, como em arroz (*Oryza sativa* L.), batata (*Solanum tuberosum* L.), mamão (*Carica papaya* L.), traças (*Plutella xylostella*), variedades de trigo egípcias e esturjão-chinês (*Acipenser sinensis* Gray) (ZHAO et al., 2005; ROUX et al., 2007; BARBOSA et al., 2011; IBRAHIM et al., 2016; DOS SANTOS et al., 2019; SPANOGHE et al., 2020).

Devido à necessidade de acessar a variabilidade genética em programas de melhoramento do café e desenvolver estratégias de seleção, o objetivo deste estudo é avaliar a aplicabilidade das redes neurais SOM propostas por Kohonen para explorar a diversidade genética e o comportamento de variáveis biométricas de 62 variedades de *Coffea arabica*.

2. MATERIAIS E MÉTODOS

2.1 Material genético e dados fenotípicos

Para este estudo, foram utilizadas 62 variedades experimentais de *C. arabica* (Tabela 1) do campo experimental da Universidade Federal de Viçosa, município de Rio Paranaíba, Minas Gerais, Brasil. O experimento foi delineado em blocos ao acaso com três repetições, contendo dez plantas por parcela. Foram mensuradas 14 variáveis, envolvendo variáveis de crescimento e desenvolvimento (VIG, ALT, COP, CAU, CR1, NR1 e NR2) e produtivas (PROD, REN, P19, P15, P11, FUN e P100) (Tabela 2).

Tabela 1. Identificação das variedades, códigos originais e origens (SETOTAW et al, 2013) dos 62 genótipos de *Coffea arabica* avaliados.

Genótipo	Variedade	Código original	Origem
1	Arara	44	Obatã x Icatu amarelo 2944
2	Sábia 398	26	Acaiá (Mundo Novo) x Catimor UFV 386
3	Icatu 3696	71	*C. canephora* x *C. arabica*
4	Siriema 14/8	10	*C. racemosa* x *C. arabica* cv. Blue Mountain
5	Catucaí Amarelo 3 SM	04	Icatu amarelo x Catuaí amarelo
6	Siriema 5/14	35	*C. racemos*a x *C. arabica* cv. Blue Mountain
7	Catucaí Vermelho 36/6	69	Icatu vermelho x Catuaí vermelho
8	Catucaí Amarelo	15	Icatu amarelo x Catuaí amarelo
9	Maracatia Vermelho	47	Acaiá x Catuaí Vermelho IAC 81
10	Maracatia Amarelo	47	Acaiá x Catuaí Vermelho IAC 81
11	Acauã 37	17	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
12	Catucaí Vermelho 20/15	18	Icatu vermelho x Catuaí vermelho
13	Acauã 65	27	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
14	Acauã 68/11	63	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
15	Icatu Amarelo 925	39	*C. canephora* x *C. arabica*
16	Icatu Vermelho 925	39	*C. canephora* x *C. arabica*
17	Catucaí Amarelo 2 SL	1	Icatu amarelo x Catuaí amarelo
18	Catuaí Vermelho 19/8	57	Caturra amarelo (IAC 467-11) x Mundo Novo (IAC 374-19)
19	Catucaí	11	Desconhecida
20	Catucaí x Catimor 357-77 (5/33)	16	Catuaí x Catimor 357-77 (5/33) FSA
21	Catucaí Amarelo 24/137	02	Icatu x Catuaí
22	Acauã Laranja	38	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
23	Acauã Amarelo	38	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
24	Acauã Vermelho	38	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
25	Acauã	55	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
26	Catucaí Amarelo 24/137	07	Icatu x Catuaí
27	Catucaí Vermelho 20/15	05	Icatu x Catuaí
28	Acauã	49	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
29	Palma III	56	Catuaí Vermelho IAC 81 x Catimor (UFV 353)
30	Catucaí Amarelo	43	Icatu x Catuaí
31	Acauã 68-2	61	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
32	Bem Te Vi	77	Catimor 391 x Catuaí amarelo 74
33	Icatu 925	73	*C. canephora* x *C. arabica*
34	Siriema Vermelho	58	*C. racemos*a x *C. arabica* cv. Blue Mountain
35	Siriema Amarelo	58	*C. racemos*a x *C. arabica* cv. Blue Mountain
36	Catucaí 785/15	09	Icatu Vermelho 785 x Catuaí vermelho
37	Catucaí Amarelo 3/5	81	Icatu x Catuaí
38	Catucaí Amarelo 3/5 LVA	81	Icatu x Catuaí

Continuação Tabela 1.

Genótipo	Variedade	Código original	Origem
39	Catucaí Amarelo	01	Icatu x Catuaí
40	Catucaí Vermelho 20/15	21	Icatu x Catuaí
41	Acauã 3%	52	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
42	Acauã 5%	51	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
43	Catucaí Vermelho	86	Icatu x Catuaí
44	Acauã Amarelo (Tardio)	48	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
45	Acauã Amarelo (Médio)	48	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
46	Araponga	60	Catuaí amarelo IAC 86 x HT UFV 446-08
47	Acauã Amarelo	50	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
48	Icatu 925	83	C. canephora x C. arabica
49	Acauã	46	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
50	Catuaí 66	84	Caturra amarelo (IAC 476-11) x Mundo Novo (IAC 374-19)
51	Acauã Amarelo	54	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
52	Acauã Vermelho	54	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
53	IPR 103	N22	Catuaí x Icatu
54	Catucaí Vermelho 19/8	40	Icatu x Catuaí
55	3 SM	08	Desconhecida
56	Catucaí Amarelo 20/15	30	Icatu x Catuaí
57	Catucaí Vermelho 36/69	24	Icatu x Catuaí
58	Acauã Amarelo FSA	22	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
59	Catucaí Roxinho	82	Icatu x Catuaí
60	IPR 98	N19	Villa Sarchi CIFC 971/10 x HT CIFC 832/2
61	Acauã 5/20	28	Mundo Novo (IAC 388-17) x Sarchimor (IAC 1668)
62	Bourbon Vermelho	14	Desconhecida

2.1 Modelos lineares mistos

Os efeitos aleatórios (genótipos), bem como os componentes de variância para as 14 variáveis avaliadas foram estimados de acordo com o modelo linear misto (MLM):

$$y = X\beta + Zg + e$$

em que: y é o vetor dos valores observados de tamanho n, β é o vetor de efeito fixo (blocos) associado à matriz de incidência X, g é o vetor de efeito aleatório (genótipo) associado às matrizes Z, $e \sim N(0, V = I_n\sigma_e^2)$ é o vetor residual e I é a matriz identidade associada ao erro. Os valores genotípicos foram estimados via BLUP (*Best Linear Unbiased Prediction*) e utilizados como entradas nas redes neurais SOM e para a análise de CP. As correlações de

Pearson entre os valores de BLUP para todas as variáveis foram calculadas a fim de avaliar a correlação genética entre as mesmas.

Tabela 2. Metodologia de coleta para todas as variáveis avaliadas nas 62 variedades de *C. arabica*.

Características	ID	Metodologia
Produção de 2017 e 2018 (L/parcela)	PROD17 PROD18	Os frutos foram coletados manualmente e a produção em L/ parcela foi quantificada com auxílio de um medidor volumétrico.
Vigor vegetativo (1-5)	VIG	Foram determinadas notas de 1 a 5 relativas ao vigor vegetativo das plantas, sendo 1 atribuídas às plantas pouco vigorosas e 5 àquelas muito vigorosas.
Altura da planta (m)	ALT	Medição direta com auxílio de uma trena.
Diâmetro da copa (m)	COP	
Diâmetro do caule (cm)	CAU	Medição direta 10 cm acima do solo, com auxílio de um paquímetro digital
Comprimento de ramos primários (cm)	CR1	Medição direta em um ramo do terço médio da planta a partir do ponto de inserção no ramo principal até a extremidade.
Número de ramos primários	NR1	Contagem direta no terço médio das plantas.
Número de ramos secundários	NR2	
Rendimento pós secagem (%/5 L)	REN	Os grãos foram inicialmente coletados em um volume de 5 L e pesados, após a secagem e revolvimento dos grãos em terreiro suspenso a massa em gramas (g) foi novamente mensurada e feito o cálculo de proporção de rendimento em relação ao peso inicial.
Peneira19 (g)	P19	Os grãos foram inicialmente coletados em um volume de 5 L, após a secagem e revolvimento dos grãos em terreiro suspenso, as quantidades em gramas (g) retidas em cada peneira foram mensuradas com auxílio de uma balança.
Peneira15 (g)	P15	
Peneira 11 (g)	P11	
Fundo de peneira (g)	FUN	
Peso de 100 grãos (g)	P100	Os grãos foram coletados, posteriormente secos e revolvidos em terreiro suspenso, 100 grãos foram amostrados aleatoriamente e pesados.

2.2 *Self-organazing maps* (SOM)

Os valores de BLUP preditos foram utilizados como dados de entrada para o treinamento do SOM. As metodologias para definir o formato adequado para as unidades do mapa e o número de neurônios, referidos aqui como unidades, ainda não são estabelecidas na literatura (SPANOGHE et al., 2020), portanto, a estrutura hexagonal foi estabelecida para as unidades, uma vez que, permite melhor visualização da estrutura geral dos dados e não favorece direções horizontais ou verticais no mapa (KOHONEN, 2014). O grid, ou tamanho do mapa foi selecionado de acordo com a melhor configuração do número de genótipos mapeados por unidades e a distância entre as unidades mais próximas. De acordo com o volume de dados e número de variáveis analisadas, os tamanhos de mapa testados foram 2 x 2, 3 x 3, 4 x 4 e 5 x 5. O mapa com a distribuição mais homogênea do número de genótipos mapeados por unidades e sem distâncias discrepantes (ou seja, muito próximos ou muito distantes entre unidades) foi o de 4 x 4, e portanto, selecionado para o estudo da diversidade genética do material em estudo. Foi determinado um número de 1000 interações para o processo de aprendizagem. O pacote

estatístico utilizado foi o Kohonen (WEHRENS; KRUISSELBRINK, 2019) pelo software R (R Development Core Team, 2020).

Três representações gráficos do SOM foram geradas, a primeira do tipo *codes*, é uma representação das variáveis originais normalizadas utilizadas, o tamanho dos vetores para cada variável indica a magnitude dos valores para as amostras mapeadas em cada unidade (WEHRENS, 2007). Desta forma é possível visualizar ao longo do mapa quais são as características em comum para os genótipos mapeados em cada unidade. O segundo tipo de representação chamado *mapping* mostra quais foram os genótipos mapeados em cada unidade. A última representação foi a *dist.neighbours*, que expressa o grau de conectividade neural entre as unidades vizinhas (WEHRENS, 2007). Este tipo de mapa é importante porque apresenta *clusters* naturais para os dados, refletindo a distância genética entre os genótipos agrupados (SPANOGUE et al., 2020). A interpretação de um mapa SOM gerado para um conjunto de dados a partir das representações *codes*, *mapping* e *dist. neighbours* oferece informações que permitem inferir sobre a diversidade genética do material em estudo.

2.3 Análise de componentes principais e agrupamento UPGMA

As análises de componentes principais foram realizadas para quantificar a influência das variáveis mensuradas na explicação da variação total dos dados. Foi calculada a matriz de distância euclidiana e o método UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*) foi selecionado para o estudo da dissimilaridade do material genético. O critério de Mojena (1977) foi adotado para determinar o número ideal de grupos para o dendrograma e o coeficiente de correlação cofenética foi calculado para mensurar a consistência e qualidade do dendrograma.

3. RESULTADOS

3.1 Ajustes de modelos lineares mistos e correlações de Pearson

Houve detecção de variância genética para variáveis PROD18, VIG, ALT, COP, P19, P15, P11, FUN e P100 (Tabela 3). O coeficiente de variação ambiental (CV_e) apresentou grande amplitude, sendo consideravelmente alto entre as variáveis PROD18 e PROD17, indicando baixa precisão experimental e/ou instabilidade das variedades de café na produção diante as variações ambientais.

Tabela 3. Valores p para o teste de razão de verossimilhança pelos modelos lineares mistos, médias fenotípicas, coeficiente de variação ambiental (CV_e) e herdabilidade (H^2) para as variáveis produção 2017 e 2018 (PROD17 e PROD18), vigor (VIG), altura (ALT), diâmetro da copa (COP), diâmetro do caule (CAU), comprimento de ramos primários (CR1), número de ramos primários (NR1), número de ramos secundários (NR2), rendimento (REN), peneiras 19, 15, 11 (P19, P15 e P11), fundo de peneira (FUN) e peso de 100 grãos (P100).

Variáveis	valor P	Média	CVe (%)	H^2
PROD17 (L/parcela)	0,21^{ns}	44,7	42,44	0,23
PROD18 (L/parcela)	5,06e-06***	10,5	66,52	0,62
VIG (1-5)	0,04*	2,74	20,63	0,34
ALT (m)	1,21e-09***	2,11	6,09	0,72
COP (m)	3,96e-02*	1,15	9,22	0,36
CAU (cm)	0,36^{ns}	47,88	8,23	0,17
CR1 (cm)	0,58^{ns}	62,74	9,99	0,11
NR1	0,70^{ns}	16,31	23,64	0,08
NR2	0,96^{ns}	6,63	35,73	0,01
REN (%/ 5L)	0,09^{ns}	19,00	15,53	0,32
P19 (g)	6,11e-04***	73,10	38,56	0,54
P15 (g)	0,03*	333,37	23,08	0,39
P11 (g)	0,01*	222,17	28,62	0,49
FUN (g)	2,18e-03**	19,63	55,58	0,55
P100 (g)	1,11e-03**	12,56	5,12	0,56

*, **, ***: significativo a 0.05, 0.01 e 0.001 de probabilidade; ^{ns}: não significativo; CVe (%): coeficiente de variação ambiental.

As herdabilidades (H^2) calculadas variaram entre 0,01 e 0,72 (Tabela 3). Os menores valores observados (< 0,1) foram para NR2 (0,01) e NR1 (0,08) e as variáveis com maiores H^2 (> 0,5) foram ALT (0,72), PROD18 (0,62), P100 (0,56), FUN (0,55) e P19 (0,54).

As correlações genéticas obtidas entre as características avaliadas foram predominantemente baixas (Figura 1). As maiores correlações positivas foram observadas entre REN e P19 (0,52), CAU e ALT (0,45) e PROD18 e VIG (0,42). E a maior correlação negativa foi observada entre FUN e P11 (-0,53).

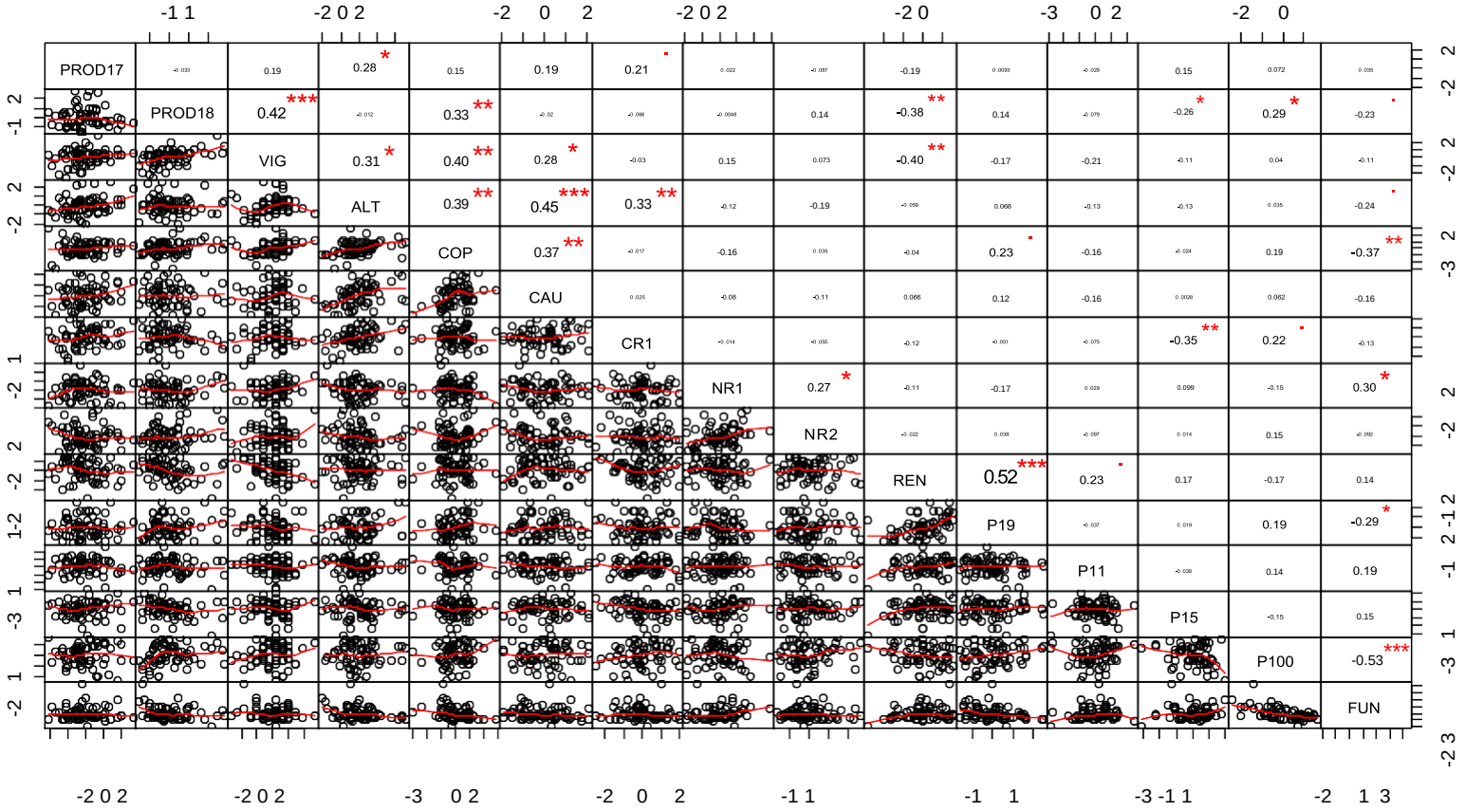


Figura 1. Diagonal inferior com gráficos de correlações genéticas e linhas de tendências (em vermelho) entre todas as características. Diagonal superior indicando os valores das correlações e os níveis de significância dos valores p. Diagonal principal contendo a identificação das variáveis produção 2017 e 2018 (PROD17 e PROD18), vigor (VIG), altura (ALT), diâmetro da copa (COP), diâmetro do caule (CAU), comprimento de ramos primários (CR1), número de ramos primários (NR1), número de ramos secundários (NR2), rendimento (REN), peneiras 19, 15 e 11 (P19, P15 e P11), peso de 100 grãos (P100) e fundo de peneira (FUN), os símbolos '*', '**' e '***' indicam 0,05; 0,01 e 0,001, respectivamente

3.2 *Self-organazing maps* (SOM)

O mapa do tipo *codes* (Figura 2a) apresentou diferentes padrões para as magnitudes dos valores genotípicos para os genótipos mapeados em cada unidade, evidenciando que existem diferenças genéticas entre os 62 indivíduos mapeados nas 16 unidades. Observou-se que, a depender das unidades, algumas variáveis de interesse ficaram em evidência (p.e, PROD17, PROD18, P19 e REN) indicando que os genótipos agrupados nessas unidades são superiores para estas características, por possuírem valores genotípico maiores em relação aos outros genótipos avaliados. Também foi possível observar no mapa *codes* a oscilação de produção bienal, que é característica das culturas de café. Genótipos com maiores valores de PROD17 (unidades 9 e 10) não são os mesmos com os valores altos de PROD18 (unidades 3 e 13) (Figura 2a).

Algumas correlações para as variáveis de produção e classificações dos grãos (P19, P15 e P11) foram detectadas (Figura 2a). Valores altos de PROD17 foram associados com maiores valores de VIG, ALT, COP e CAU (unidades 9 e 10) e os genótipos com maior produção em 2018 também possuem valores altos de VIG, NR1, NR2, ALT, COP, CAU e P19 (unidades 3 e 13). Em todas as unidades com valores genotípicos altos para REN (unidades 4, 8, 13, 14 e 15) houve associações com valores altos de P15.

O mapa do tipo *mapping* (Figura 2b) permitiu relacionar os genótipos agrupados nas unidades aos padrões genéticos representados através do mapa *codes*. Os genótipos (11, 16, 19, 37, 53, 28, 48, 56 e 57) mapeados nas unidades 9 e 10 foram superiores para a PROD17 e aqueles agrupados nas unidades 3 e 13, superiores para PROD18. E os indivíduos mapeados nas unidades 4, 9, 13 e 15 foram os genótipos com maior quantidade de variáveis com valores genotípicos altos. As cores diferentes para os números dos genótipos representados no mapa indicam os seis clusters determinados pelo método de Mojena para o agrupamento UPGMA (Figura 3). Portanto, os grupos formados pela rede neural SOM não foram semelhantes ao método UPGMA (Figura 4).

O mapa das distâncias das unidades em relação aos vizinhos mais próximos (Figura 2c) indicou nove agrupamentos naturais para os genótipos. Neste tipo de mapa (*dist. neighbour*), a interpretação se assemelha a mapas topológicos, onde unidades vizinhas com cores semelhantes estão em níveis mais próximos do que em relação a outras unidades de cores diferentes. A maior

distância foi observada para a unidade um, como um pico topológico. Este resultado indica que os genótipos mapeados nesta unidade são divergentes aos genótipos das unidades vizinhas (grupos 2 e 4). Ou seja, os genótipos 1, 3, 6 e 61 são geneticamente similares entre si e relativamente divergentes dos genótipos mapeados nas unidades vizinhas.

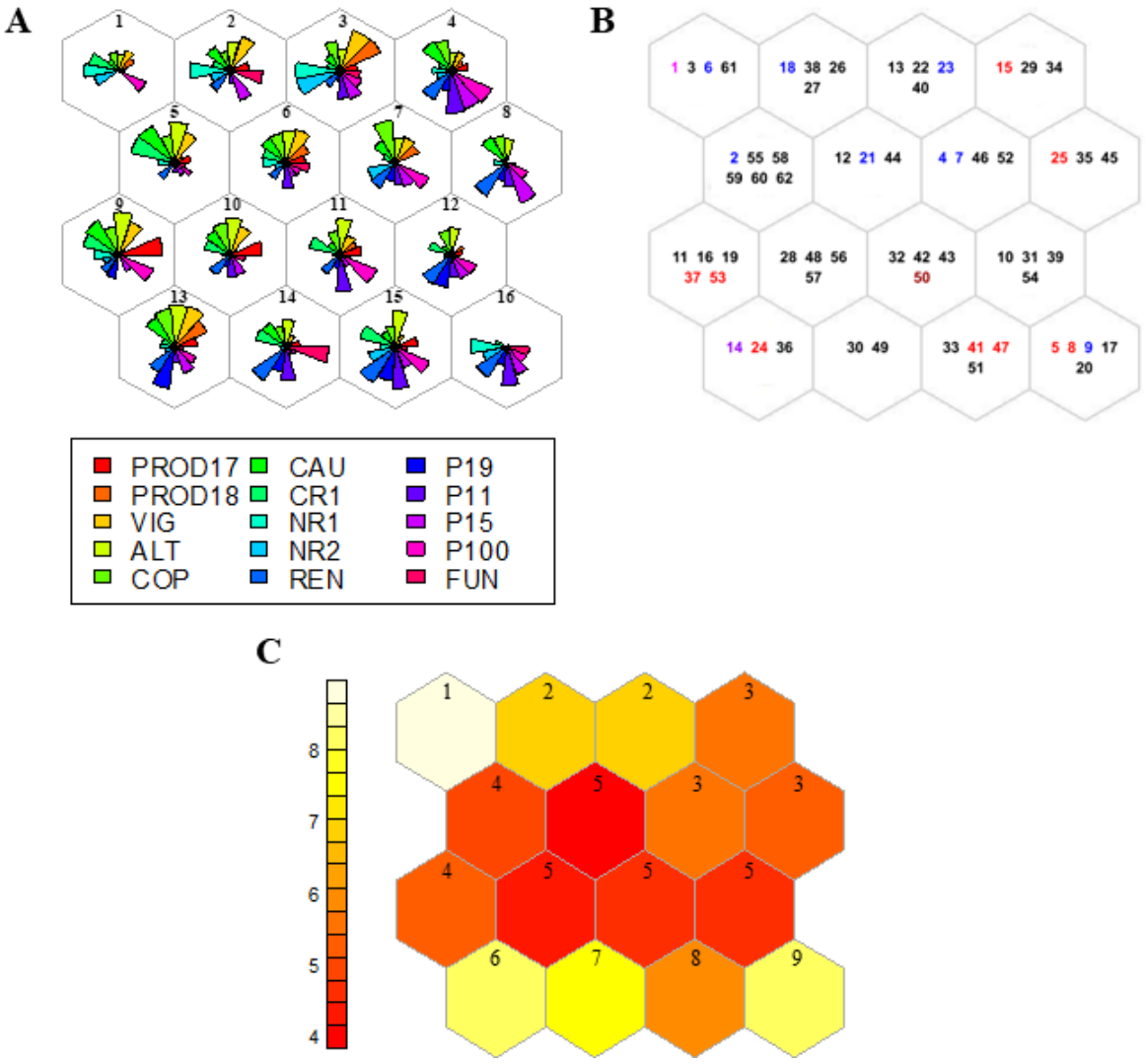


Figura 2. *Self-oganizin maps* (SOM) gerados A: representação *codes:* vetores indicando a magnitude dos valores para as variáveis (leque ao centro dos hexágonos) comum para os indivíduos mapeados em cada unidade, os números no topo dos hexágonos indicam o número da unidade. B: representação *mapping:* identificação dos genótipos mapeados em cada unidade, os números com as cores rosa, azul, vermelho, vermelho escuro, roxo e preto indicam os 6 grupos determinados para o agrupamento UPGMA. C: representação *dist. neighbour*: distâncias das unidades em relação aos vizinhos mais próximos.

A interpretação de todas informações geradas obtidas através da análise de redes neural (Tabela 4) permitiu a identificação de grupos de genótipos superiores para grupos diferentes de

variáveis. Os genótipos com maiores valores para produção, tanto em 2017 quanto para 2018 foram os genótipos 11,13, 14, 16, 19, 22, 23, 24, 36 e 37, 40 e 53, os quais, são descendentes das variedades Mundo novo (IAC388-17), Sarchimor (IAC 1668), Icatu, Catuaí, *C. canephora*, *C. arabica*, Icatu vermelho 785 e Catuaí vermelho. Também foi possível identificar os fatores que permitiam o agrupamento dos genótipos de acordo com as distâncias entre as unidades. Por exemplo, os genótipos mapeados dentro do grupo quatro possuem valores similares para VIG, ALT, CAU e CR1. Essas informações facilitam a interpretação dos agrupamentos formados para as variedades em estudo.

Tabela 4. Número de genótipos agrupados por unidades dos SOM, identificação dos genótipos mapeados, variáveis com maiores valores para cada unidade, identificação dos grupos formados pela distância entre as unidades vizinhas mais próximas e origens dos genótipos mapeados.

Unidade	**N° de genótipos**	**Genótipos**	**Variáveis em destaque**	**Grupos (*dist.neighbour*)**	**Grupos genealógicos**
1	4	1, 3, 6, 61	CR1, NR1 e P100	1	Mundo novo (IAC388-17) x Sarchimor (IAC 1668); *C. canephora* x *C. arabica; C. racemosa* x *C. arabica* cv. Blue Mountain; Obatã x Icatu amarelo.
2	4	18, 38, 26, 27	VIG, NR1, P15 e FUN	2	Icatu x Catuaí; Caturra amarelo (IAC466-11) x Mundo novo (IAC375-19).
3	4	13, 22, 23, 40	PROD18, VIG, NR1 e NR2	2	Mundo novo (IAC388-17) x Sarchimor (IAC 1668); Icatu x Catuaí.
4	3	15, 29, 34	P11, P15 e P100	3	*C. canephora* x *C. arabica*; *C. racemosa* x *C. arabica* cv. Blue Mountain; Catuaí vermelho (IAC81) x Catimor (UFV386).
5	6	2, 55, 58, 59, 60, 62	VIG, ALT, CAU e CR1	4	Mundo novo (IAC388-17) x Sarchimor (IAC1668); Icatu x Catuaí; Acaiá (Mundo novo) x Catimor (UFV386).
6	3	12, 21, 44	VIG, ALT, COP, CAU e P11	5	Mundo novo (IAC388-17) x Sarchimor (IAC1668); Icatu x Catuaí .
7	4	4, 7, 46, 52	COP, REN e P100	3	Mundo novo (IAC388-17) x Sarchimor (IAC 1668); *C. canephora* x *C. arabica*; C. racemosa x C. arabica cv. Blue Mountain; Catuaí amarelo (IAC86) x Hibrido Timor (UFV).
8	3	25, 35, 45	REN, e P11	3	Mundo novo (IAC388-17) x Sarchimor (IAC 1668); *C. racemosa* x *C. arabica* cv. Blue Mountain.
9	5	11, 16, 19, 37, 53	PROD17, VIG, ALT, CAU, CR1 e P100	4	Mundo novo (IAC388-17) x Sarchimor (IAC 1668); Icatu x Catuaí; *C. canephora* x *C. arabica*.
10	4	28, 48, 56, 57	PROD17, ALT e COP	5	Mundo novo (IAC388-17) x Sarchimor (IAC 1668); Icatu x Catuaí; *C. canephora* x *C. arabica*.
11	4	32, 42, 43, 50	ALT, CR1, P11 e P100	5	Mundo novo (IAC388-17) x Sarchimor (IAC 1668); Icatu x Catuaí; Caturra amarelo (IAC466-11) x Mundo novo (IAC374-19); Catimor 391 x Catuaí amarelo 74.
12	4	10, 31, 39, 54	REN e P19	5	Mundo novo (IAC388-17) x Sarchimor (IAC 1668); Icatu x Catuaí; Acaiá x Catuaí vermelho (IAC81).

Continuação Tabela 4

Unidade	N° de genótipos	Genótipos	Variáveis em destaque	Grupos (*dist.neighbour*)	Grupos genealógicos
13	3	14, 24, 36	PROD18, VIG, ALT, COP, CAU e REN	6	Mundo novo (IAC388-17) x Sarchimor (IAC 1668); Icatu vermelho 785 x Catuaí vermelho.
14	2	30, 49	REN, P11 e FUN	7	Mundo novo (IAC388-17) x Sarchimor (IAC 1668); Icatu x Catuaí.
15	4	33, 41, 47, 51	ALT, CR1, REN, P19, P11 e P100	8	Mundo novo (IAC388-17) x Sarchimor (IAC 1668); *C. canephora* x *C. arabica*.
16	5	5, 8, 9, 17, 20	NR1, REN e P11	9	Icatu amarelo x Catuaí amarelo; Acaiá x Catuaí vermelho (IAC81); Catuaí x Catimor 357.

3.3 Análise de componentes principais e agrupamento UPGMA

O gráfico das componentes principais apresentou um baixo percentual de explicação para a variação dos dados (Figura 3). Sendo que 19,6% foi retido pela CP1 e 14,1% pela CP2, totalizando em 33,7%. Com base nesse resultado, não é possível avaliar por meio da análise de componentes principais quais são as variáveis mais importantes para o estudo da variabilidade genética nas 62 variedades de *C. arabica* em estudo.

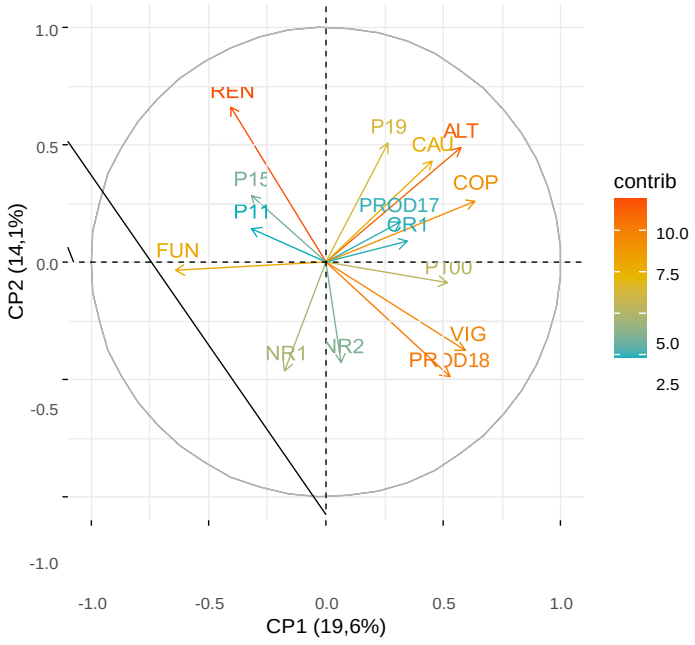


Figura 3. Contribuição das variáveis produção de 2017 e 2018 (PROD17 e PROD18), vigor (VIG), altura (ALT), diâmetro da copa (COP), diâmetro do caule (CAU), comprimento de ramos primários (CR1), número de ramos primários (NR1), número de ramos secundários (NR2), rendimento (REN), peneiras 19, 15 e 11 (P19, P15 e P11), fundo de peneira (FUN) e peso de 100 grãos (P100), para a variação contida nos componentes principais 1 (CP1) e componentes principais 2 (CP2).

A matriz de distância Euclidiana foi obtida para os valores de BLUP calculados via MLM, uma vez que foram obtidas baixas correlações entre as variáveis. O método de agrupamento UPGMA apresentou um coeficiente de correlação cofonética de 0,78. O método de Mojena (1977) indicou seis grupos como ideais para os agrupamentos das variedades (Figura 4), divergindo dos resultados de agrupamentos fornecidos pelo SOM (Figura 2c).

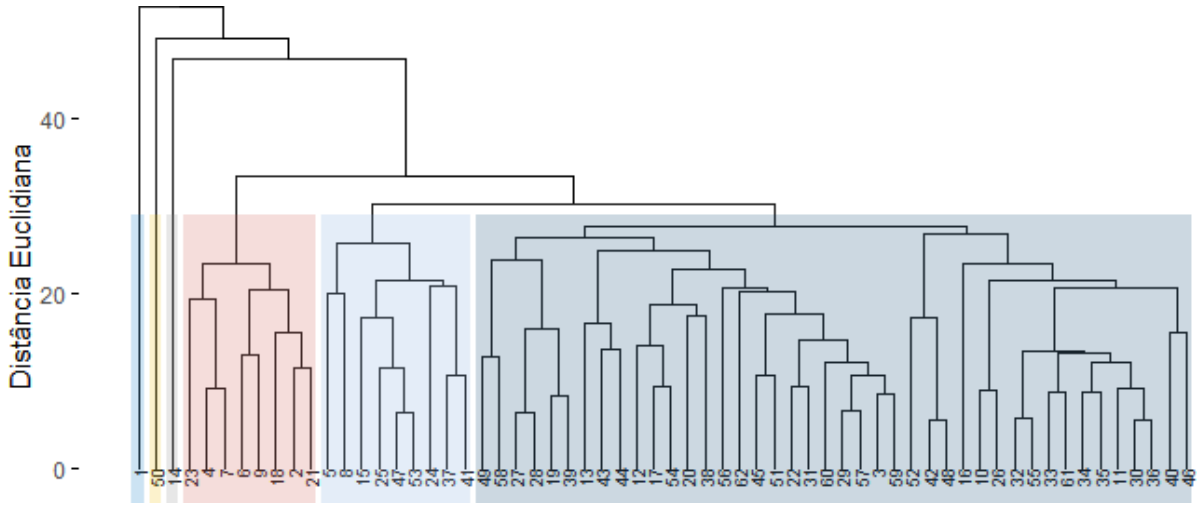


Figura 4. Dendrograma obtido a partir da matriz de dissimilaridade de distância Euclidiana e método UPGMA de agrupamento para 62 variedades de *C. arabica*. Os 6 clusters foram determinados pelo método de Mojena (1977).

4. DISCUSSÃO

Alguns dos principais objetivos dos melhoristas são acessar a diversidade genética contida nas variedades em teste, analisar correlações entre as variáveis de interesse e escolher genótipos candidatos para os cruzamentos para estabelecer estratégias eficientes de seleção. Frequentemente análises de CP, cálculo de matrizes de similaridade e dissimilaridade e métodos de agrupamento são utilizados para os estudos de diversidade genética no café (GUEDES et al., 2013; TEIXEIRA et al., 2013; RODRIGUES et al., 2016; MACHADO et al., 2017; DUTRA GILES et al., 2019;). Entretanto, é importante que métodos mais eficientes para acessar a complexidade das relações contidas nos dados sejam investigados. Neste estudo, foram explorados os níveis de informações que podem ser obtidos pelas redes neurais tipo SOM para o estudo da diversidade genética em variedades de *C. arabica*.

As variáveis biométricas avaliadas neste estudo são frequentemente utilizadas em estudos de diversidade genética de diferentes variedades de café (GUEDES et al., 2013; RESENDE et al, 2001; RODRIGUES et al., 2016; SOUSA et al., 2019; TASSONE et al., 2019; TOUNEKTI et al., 2017) . Portanto, a falta da detecção dos efeitos de genótipos para algumas dessas variáveis podem indicar uma variabilidade genética relativamente baixa para as variedades em estudo. Entretanto, foi detectada variabilidade por meio das variáveis de peneira avaliadas (P19, P15, P11 e FUN), foram obtidos valores altos de H^2 para P19 e P100, além de ter sido detectadas algumas correlações entre as peneiras e outras variáveis importantes de crescimento e desenvolvimento. As classificações do café através das peneiras são muito importantes, uma vez que seja um critério fundamental na comercialização do café, por garantir a torração homogênea dos grãos (LOPES et al., 2013). Portando, com base nestes resultados, recomenda-se a avaliação da quantidade de grãos retidas por diferentes peneiras para estudos de diversidade e auxiliar na seleção de genótipos superiores em programas de melhoramento de café.

As correlações genéticas entre as variáveis foram predominantemente baixas e as correlações observadas pela análise de CP não foram muito informativas a respeito dos dados, uma vez que o percentual de explicação das componentes foi muito baixo (33,7%). Por outro lado, os resultados obtidos através dos mapas auto-organizáveis de Kohonen, mais especificamente, o do tipo *codes* possibilitou a identificação de diferentes padrões e correlações entre variáveis importantes. Foi observado que valores altos de produção, independente do ano, estão correlacionados, principalmente, com o VIG, ALT, COP e CAU em determinadas unidades no mapa. Além disso, também foi identificado quais são os genótipos com maiores REN, P19 e P100 simultaneamente, traços que são alvo de seleção para o melhoramento de café. Este tipo de resultado é interessante por permitir o acesso de um volume informações que não são possíveis através das análises de CP, permitindo uma seleção mais eficiente de características e germoplasmas elites nos programas de melhoramento (PARK et al., 2004; IBRAHIM et al., 2016; SPANOGHE et al., 2020).

Nove grupos foram naturalmente determinados pelo SOM (Figura 2c), divergindo dos seis *clusters* determinados pelos agrupamentos UPGMA (Figura 4). A grande diferença que pode ser observada para os grupos formados pelos SOM é que além dos genótipos que compõem cada grupo (Figura 2b), é possível identificar quais são os padrões dos valores

genotípicos para as variáveis em comum para cada agrupamento (Figura 2a). Além do número de grupos, os genótipos agrupados por cada método também foram diferentes. Resultados divergentes foram obtidos para o estudo da diversidade genética de arroz (*Oryza sativa* L.) (SANTOS et al., 2019), em que os genótipos foram agrupados de forma semelhante pelos SOM e pelo método hierárquico UPGMA. Entretanto, foram observadas diferenças nos agrupamentos de variedades de batatas e de trigo (IBRAHIM et al., 2016; SPANOGUE et al., 2020) entre o método SOM e agrupamento hierárquico de Ward. Nestes casos os agrupamentos gerados pelo SOM foram considerados mais adequados, sendo mais fiéis às informações de *pedrigree*, origens, resistências a doenças e outros traços agronômicos das variedades.

As divergências entre os resultados destes métodos estão relacionadas, principalmente pelo maior refinamento e complexidade das redes neurais para organização de dados, capturando variações em diferentes dimensões contidas nos dados (PARK et al., 2004; MELO; SOUSA, 2011; IBRAHIM et al., 2016; SPANOGHE et al., 2020), tornando inviável a comparação entre os resultados de redes neurais SOM e métodos de agrupamento clássicos (OLIVEIRA; SANTOS; CRUZ, 2020). Além de não fornecer informações acerca dos grupos formados, os métodos de agrupamentos clássicos são sensíveis a presença de *outliers*, aspecto que não interfere muito nos resultados dos SOM, em que apenas o peso das unidades que contém os *outliers* é afetado, sem interferir nas outras unidades do mapa. Portanto os grupos de genótipos aqui gerados a partir dos mapas obtidos pelas redes neuras SOM podem ser considerados mais próximos aos agrupamentos reais em relação aos grupos obtidos pelo método UPGMA.

5. CONCLUSÕES

A análise de CP não foi eficiente para explicar a variação dos dados a partir das variáveis avaliadas. Os mapas de Kohonen obtidos permitiram identificar genótipos superiores para diferentes características de interesse para o melhoramento do café, determinar e caracterizar os diferentes grupos de genótipos obtidos, e observar correlações entre diferentes variáveis, confirmando a aplicabilidade das redes neurais SOM para os programas de melhoramento genético de plantas. Além disso os agrupamentos formados pelos SOM para os genótipos devem ser considerados mais reais devido ao refinamento das análises de redes neurais.

6. REFERÊNCIAS

BARBOSA, C. D. et al. Artificial neural network analysis of genetic diversity in Carica papaya L. **Crop Breeding and Applied Biotechnology**, v. 11, n. 3, p. 224–231, 2011.

CARVALHO, C. H. S. Cultivares de café: origem, características e recomendações. **Cultivares de café**, p. 334, 2008.

DA SILVA, D. O. et al. Genetic progress with selection of coffea canephora clones of superior processed coffee yield. **Ciencia Rural**, v. 48, n. 3, p. 1–7, 2018.

RESENDE, M. et al. Estimativas de parâmetros genéticos e predição de valores genotípicos no melhoramento do cafeeiro pelo procedimento REML/BLUP. **Bragantia**. v. 60, n. 1, p. 185–193, 2001.

GILES, J. A. et al. Divergence and genetic parameters between coffea sp. genotypes based in foliar morpho-anatomical traits. **Scientia Horticulturae**, v. 245, n. May 2018, p. 231–236, 2019.

GUEDES, J. M. et al. Divergência genética entre cafeeiros do germoplasma Maragogipe. **Bragantia**, v. 72, n. 2, p. 127–132, 2013.

IBRAHIM, O. M. et al. Evaluating the Performance of 16 Egyptian Wheat Varieties Using Self-Organizing Map (SOM) and Cluster Analysis. **Journal of Applied Sciences**, v. 16, n. 2, p. 47–53, 2016.

KOHONEN, T. Selg organized formation of topologically correct feature maps. **Biological Cybernetics**, v. 69, p. 59–69, 1982.

KOHONEN, T. **MATLAB Implementations and Applications of the Self-Organizing Map**. Unigraphia Oy: Helsinki, Finland, 2014.

LOPES, L. et al, Avaliação de cultivares de *Coffea arabica* L. através da classificação por peneira. In: Simpósio de Pesquisa dos Cafés do Brasil e Workshop Internacional de Café & Saúde. Anais. Brasília, DF: Embrapa Café. P.220-221, 2003.

MACHADO, C. M. S. et al. Genetic diversity among 16 genotypes of Coffea arabica in the Brazilian cerrado. **Genetics and Molecular Research**, v. 16, n. 3, p. 1–13, 2017.

MELO, B.; SOUSA, L. B. Biologia da reprodução de Coffea arabica L . e Coffea canephora Pierre. **Revista verde de agroecologia e desenvolvimento sustentavelerde de agroecologia e desenvolvimento sustentavel**, v. 5 (3), p. 05–11, 2011.

MOJENA, R. Hierarquical grouping methos and stopping rules: An evaluation. **The Computer Journal**, v. (20)4, n. 2, p. 359–363, 1977.

OLIVEIRA, M.; DOS SANTOS, I. G.; CRUZ, C. D. Self-organizing maps: a powerful tool for capturing genetic diversity patterns of populations. **Euphytica**, v. 216, n. 3, 2020.

PARK, Y. S. et al. Hierarchical community classification and assessment of aquatic ecosystems using artificial neural networks. **Science of the Total Environment**, v. 327, n. 1–3, p. 105–122, 2004.

RICHARDSON, A. J.; RISI EN, C.; SHILLINGTON, F. A. Using self-organizing maps to identify patterns in satellite imagery. **Progress in Oceanography**, v. 59, n. 2–3, p. 223–239, 2003.

RODRIGUES, W. P. et al. Assessment of genetic divergence among coffee genotypes by Ward-MLM procedure in association with mixed models. **Genetics and Molecular Research**, v. 15, n. 2, 2016.

ROUX, O. et al. ISSR-PCR: Tool for discrimination and genetic structure analysis of Plutella xylostella populations native to different geographical areas. **Molecular Phylogenetics and Evolution**, v. 43, n. 1, p. 240–250, 2007.

R DEVELOPMENT CORE TEAM. R. a language and environment for statistical computing. R foundation for Statistical Computin, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

SANTOS, I. G. et al. Self-organizing maps in the study of genetic diversity among irrigated rice genotypes. **Acta Scientiarum - Agronomy**, v. 41, n. 1, p. 1–9, 2019.

SETOTAW, T. et al. Coefficient of parentage in Coffea arabica L. cultivars grown in Brazil. **Crop Science**, v. 53, n. 4, p. 1237–1247, 2013.

SMITH, A. B.; CULLIS, B. R.; THOMPSON, R. The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. **Journal of Agricultural Science**, v. 143, n. 6, p. 449–462, 2005.

SOUSA, T. V. et al. Early selection enabled by the implementation of genomic selection in coffea arabica breeding. **Frontiers in Plant Science**, v. 9, n. January, p. 1–12, 2019.

SPANOGHE, M. C. et al. Genetic patterns recognition in crop species using self-organizing map: the example of the highly heterozygous autotetraploid potato (Solanum tuberosum L.). **Genetic Resources and Crop Evolution**, v. 67, n. 4, p. 947–966, 2020.

TASSONE, G. A. T. et al. Simultaneous selection in coffee progenies of mundot novo by selection indices. **Coffee Science**, v. 14, n. 1, p. 83–92, 2019.

TEIXEIRA, A. L. et al. Principal component analysis on morphological traits in juvenile stage arabica coffee. **Coffee Science**, v. 8, n. 2, p. 205–211, 2013.

TOUNEKTI, T. et al. Genetic Diversity Analysis of Coffee (<i>Coffee arabica</i> L.) Germplasm Accessions Growing in the Southwestern Saudi Arabia Using Quantitative Traits. **Natural Resources**, v. 08, n. 05, p. 321–336, 2017.

VOLSI, B. et al. The dynamics of coffee production in Brazil. **PLoS ONE**, v. 14, n. 7, p. 1–15, 2019.

WEHRENS, R. <Kohonen-Manual.Pdf>. **JSS Journal of Statistical Software**, v. 21, n. 5, 2007.

ZHAO, N. et al. Microsatellites assessment of Chinese sturgeon (Acipenser sinensis Gray) genetic variability. **Journal of Applied Ichthyology**, v. 21, n. 1, p. 7–13, 2005.

ZHOU, X. et al. Genetic diversity patterns in Japanese soybean cultivars based on coefficient of parentage. **Crop Science**, v. 42, n. 4, p. 1331–1342, 2002.

CONCLUSÕES GERAIS

Comparar métodos estatísticos para estudos de parâmetros genéticos em variedades de espécies de importância econômica é fundamental para auxiliar o desenvolvimento dos programas de melhoramento genético. Baseando-se nos resultados dos resíduos de Pearson calculados, os modelos lineares generalizados mistos (MLGM) ajustaram melhor os dados obtidos pelas avaliações das variáveis de importância agronômica em *C. arabica*. As grandes diferenças entre os MLM e MLGM se devem principalmente pela adição de um preditor linear, uma função de ligação e à diversificação de escalas adicionados aos MLGM. Mesmo que não tenha sido observada diferença entre o ranqueamento dos genótipos para os modelos em estudo, atribuir diferentes distribuições aos dados e determinar funções de ligação para os modelos influenciam diretamente na estimação dos componentes de variância e consequentemente nos parâmetros genéticos. Devido à complexidade dos MLGM, ainda não há um consenso fundamentada pela literatura a respeito das interpretações dos parâmetros genéticos estimados, principalmente para modelos gama, como observado no presente trabalho, esse aspecto dificulta a aplicabilidade em programas de melhoramento, mesmo que sejam demonstrados ajustes mais verossímeis para os dados.

O estudo da aplicabilidade dos mapas auto-organizáveis de Kohonen para análises de diversidade genética em variedades de *C. arabica* demonstrou que as redes neurais podem ser utilizadas para o reconhecimento de correlações genéticas, em diferentes níveis dos dados, entre características agronômicas importantes, detecção de genótipos superiores e agrupamentos entre genótipos. É importante que ferramentas mais robustas e informativas sejam exploradas para estudos de diversidade genética, uma vez que as análises de componentes principais (CP) usualmente utilizadas não sejam muito eficientes para explicar variações nos dados e os métodos de agrupamentos clássicos tenham algumas limitações de processamento dos dados e apresentação dos resultados.