PEDRO THIAGO MEDEIROS PAIXÃO

ANÁLISE DE FATORES APLICADA EM ESTUDOS DE SELEÇÃO GENÔMICA NO MELHORAMENTO DE *Coffea canephora*

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

Orientadora: Ana Carolina Campana Nascimento

Coorientadores: Camila Ferreira Azevedo
Moysés Nascimento

VIÇOSA - MINAS GERAIS
2020

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

T
P149a
2020

Paixão, Pedro Thiago Medeiros, 1994-
Análise de fatores aplicada em estudos de seleção genômica no melhoramento de *Coffea canephora* / Pedro Thiago Medeiros Paixão. – Viçosa, MG, 2020.
35 f. : il. ; 29 cm.

Orientador: Ana Carolina Campana Nascimento.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.31-35.

1. Análise Multivariada. 2. Predição. 3. Genômica. 4. Melhoramento Genético. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22 ed. 519.535

PEDRO THIAGO MEDEIROS PAIXÃO

ANÁLISE DE FATORES APLICADA EM ESTUDOS DE SELEÇÃO GENÔMICA NO MELHORAMENTO DE *Coffea canephora*

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 20 de fevereiro de 2020.

Assentimento:

Pedro Thiago Medeiros Paixão
Autor

Ana Carolina Campana Nascimento
Orientador

AGRADECIMENTOS

A Deus pela imensa bondade em minha vida, por estar sempre presente ao meu lado durante esta trajetória;

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, pela oportunidade concedida para realização do curso.

À Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) Código de Financiamento 001, pela concessão da bolsa de estudos.

À D. Sc. Ana Carolina Campana Nascimento, pela orientação, confiança, pela oportunidade, paciência, amizade e pelos ensinamentos transmitidos;

Aos meus coorientadores: Camila Ferreira Azevedo e Moysés Nascimento, pela orientação, amizade, conhecimentos e sugestões valiosas;

Aos pesquisadores D. Sc.s: Eveline Teixeira Caixeta, Isabela de Castro Sant'Anna e Renato Domiciano Silva Rosado, por aceitarem participar da minha banca de defesa e pelas sugestões no trabalho;

A minha mãe Naudir e minha avó Cândida, a quem sou grato pelos incentivos e orações. A minha família por todo apoio e torcida, em especial a Tia Maria (em memória);

Aos meus amigos, que se tornaram minha família: Gabi, Cris, Taiana e Fred, por todo companheirismo, carinho, conselhos e risadas;

Aos amigos que fiz em Viçosa, em especial a Jakeline, Fernanda, Gisele, Matias, Gabriely, Roberta, Marcinha, Talita, Katia, Isabella, Mayara, Poliana, Geise, Beatriz, Wanessa, Luma, e Thamiris pelo carinho e bons momentos;

Aos meus amigos de Janaúba e da graduação pela amizade, em especial a Gabriel, Francielle, Higor, Gabriela, Zenóbia, Débora, Helena, Josi e Ciene por estarem sempre comigo;

Aos amigos do PPESTBIO, LICAE e BIOINFOMÁTICA pelos momentos de descontração, companheirismo, pela troca de experiência e incentivo.

Agradeço a todos os professores e funcionários do Programa de Pós-Graduação em Estatística Aplicada e Biometria em especial ao Junior, Anita e Conceição.

Muito obrigado a todos que de alguma forma participaram desta etapa.

RESUMO

PAIXÃO, Pedro Thiago Medeiros, M.Sc., Universidade Federal de Viçosa, fevereiro de 2020. **Análise de fatores aplicada em estudos de seleção genômica no melhoramento de *Coffea canephora***. Orientadora: Ana Carolina Campana Nascimento. Coorientadores: Camila Ferreira Azevedo e Moysés Nascimento.

O Brasil se destaca em âmbito mundial na produção de café. Os incrementos observados em sua produtividade é resultado do aprimoramento de diversas metodologias. Dentre elas, destacam-se os métodos preditivos de valor genético. Estes contribuem significativamente na seleção de genótipos superiores, de forma a aumentar o ganho genético por unidade de tempo. Neste contexto, a seleção genômica ampla (GWS) é uma ferramenta que se destaca, uma vez que permite predizer o fenótipo futuro de um indivíduo baseado apenas em informações de marcadores moleculares. Realizar a seleção de maneira simultânea para várias características é o interesse da maioria dos programas de melhoramento, e a análise de fatores (AF) tem sido utilizada para auxiliar neste fim. A utilização de fatores se justifica devido a existência de correlações genéticas entre as características, as quais podem ser atribuídas aos QTL que têm efeitos pleiotrópicos ou aos QTL estreitamente ligados. Dessa forma, o objetivo deste trabalho foi de avaliar o uso da AF no contexto de GWS, em genótipos de *Coffea canephora*. Os resultados obtidos da seleção baseada nos fatores foram comparados, por meio da capacidade preditiva, acurácia e do coeficiente de Cohen's Kappa, com aqueles advindos da análise das variáveis individuais. Para isso, foram utilizados dados fenotípicos e genotípicos de populações compostas por clones dos grupos varietais Conilon e Robusta e por híbridos originados de cruzamentos entre estes grupos, avaliados durante três anos consecutivos (2014 a 2016), e uma densidade de 18111 marcadores SNPs identificados. A partir dos resultados observados, verificou-se que a AF foi eficiente para elucidar as relações entre as características e originar novas variáveis. Os fatores formados são interessantes em termos de seleção, pois além de permitirem interpretações conjuntas, apresentam boas estimativas de capacidade preditiva, herdabilidade e acurácia. Ademais observou-se alta concordância entre os indivíduos selecionados com base nos fatores e aqueles selecionados considerando as variáveis individuais. Entretanto, cabe destacar que, a seleção baseada nos fatores conseguiu selecionar indivíduos de porte mais adequado.

Palavras-chave: Predição Genômica. Análise Multivariada. Melhoramento Genético.

ABSTRACT

PAIXÃO, Pedro Thiago Medeiros, M.Sc., Universidade Federal de Viçosa, February, 2020. **Factor analysis applied in genomic selection studies to improvement *Coffea canephora.*** Adviser: Ana Carolina Campana Nascimento. Co-advisers: Camila Ferreira Azevedo and Moysés Nascimento.

Brazil stands out worldwide in the production of coffee. The increases observed in its productivity are the result of the improvement of several methodologies. Among them, the predictive methods of genetic value stand out. These contribute significantly to the selection of superior genotypes, in order to increase the genetic gain per unit of time. In this context, broad genomic selection (GWS) is a tool that stands out, since it allows predicting the future phenotype of an individual based only on information from molecular markers. Performing the selection simultaneously for various characteristics is the interest of most breeding programs, and factor analysis (AF) has been used to assist in this end. The use of factors is justified due to the existence of genetic correlations between the characteristics, which can be attributed to the QTL that have pleiotropic effects or to the closely linked QTL. Thus, the objective of this work was to evaluate the use of AF in the context of GWS, in genotypes of *Coffea canephora.* The results obtained from the selection based on the factors were compared, through predictive capacity, accuracy and the Cohen's Kappa coefficient, with those derived from the analysis of the individual variables. For this, phenotypic and genotypic data from populations composed of clones of the varietal groups Conilon and Robusta and hybrids originated from crosses between these groups, evaluated for three consecutive years (2014 to 2016), and a density of 18111 identified SNPs markers were used. From the observed results, it was found that AF was efficient in elucidating the relationships between the characteristics and originating new variables. The factors formed are interesting in terms of selection, because in addition to allowing for joint interpretations, they present good estimates of predictive capacity, heritability and accuracy. Furthermore, there was a high agreement between the individuals selected based on the factors and those selected considering the individual variables. However, it is worth noting that, the selection based on the factors managed to select individuals of more appropriate size.

Keywords: Genomic Prediction. Multivariate Analysis. Breeding Genetic.

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

O café é a *commodity* agrícola tropical mais comercializada no mundo (TRAN *et al*., 2016), possuindo impacto considerável na geração de empregos e tributos nos países produtores, o que reflete de forma relevante para a formação de receita (FASSIO e SILVA, 2007). No cenário mundial, o Brasil se destaca como o maior produtor e exportador de café e o segundo maior consumidor da bebida (EMBRAPA, 2019).

O cafeeiro, membro da família *Rubiaceae*, pertence ao gênero *Coffea* que compreende mais de 100 espécies. Dentre elas, as duas espécies exploradas comercialmente no Brasil, são *Coffea arabica* (alotetraploide, isto é, possui quatro conjuntos do número básico de cromossomos do gênero (n=11), totalizando 44 cromossomos) e *Coffea canephora* (diploide, portanto apresenta duas cópias do número básico de cromossomos, perfazendo um total de 22 cromossomos por núcleo celular) (HENDRE *et al.*, 2008). *C. arabica* é responsável pelo aroma e sabor adocicado, proporcionando bebida de melhor qualidade. *C. canephora* apresenta maiores quantidades de cafeína, sólidos solúveis e oferece corpo à bebida (BABOVA *et al.,* 2016).

Em razão da importância da cultura do café, os programas de melhoramento visam obter cultivares mais produtivas e adaptadas (SOUSA *et al.*, 2017). Neste contexto, métodos para predição de valores genéticos contribuem significativamente para aumentar a eficiência na seleção de genótipos superiores. E, a seleção genômica ampla (GWS) é uma ferramenta que se destaca, uma vez que permite predizer o fenótipo futuro de um indivíduo baseado apenas em informações de marcadores moleculares. Este processo permite a redução do tempo e do custo para a seleção, que são princípios condicionantes de sucesso no melhoramento genético, especialmente tratando-se de espécies perenes, em que seu processo de melhoramento está associado a um maior dispêndio de recursos necessários para realizar avaliações ao longo do tempo.

O *C. canephora* é um bom ponto de partida para o desenvolvimento de estudos de GWS por razões econômicas e genéticas, incluindo ploidia e ampla variabilidade genética, quando comparado ao *C. arabica.* Ambas as características tornam o processo de genotipagem e modelagem estatística mais viável do que em *C. arabica*, que é alotetraploide e possui uma base genética estreita (FERRÃO *et al*., 2017). Do ponto de vista econômico, constitui matéria-prima básica na indústria de café solúvel e como componente importante na composição dos "blends" com *C. arabica* (IVOGLO *et al*., 2008).

Vários estudos mostraram a alta acurácia da GWS (ALKIMIM, 2017; CAVALCANTI *et al*. 2012; RESENDE *et al.,* 2008; SILVA *et al*. 2013; TEIXEIRA *et al.,* 2016, SOUSA *et al.,* 2019). No que tange às culturas perenes, como o café, em que características de interesse são expressas tardiamente devido ao longo período juvenil da cultura (FERRÃO *et al.,* 2009), a GWS assume importância acentuada. Nesse caso, essa metodologia tem o potencial de aumentar o ganho genético por unidade de tempo, além de melhorar a seleção para características poligênicas e características com baixa herdabilidade, de alto custo de avaliação e de difícil mensuração (RESENDE *et al.,* 2014; SOUSA *et al.,* 2019).

Os principais métodos estatísticos usados em GWS (G-BLUP, RR-BLUP, BLASSO, Bayes B, dentre outros), apresentam a limitação de obterem conclusões direcionadas apenas para uma única variável. Assim, estudos que permitem analisar mais de uma característica de interesse podem ser úteis, fornecendo resultados válidos para um conjunto de características representadas por uma única variável latente (TEIXEIRA *et al.,* 2015). Do ponto de vista prático, os agricultores geralmente preferem lidar com um índice combinado em vez de várias características únicas (GLANTZ *et al*., 2011).

No melhoramento do cafeeiro, em que várias características são consideradas, elevados graus de correlação entre essas características são observados (FERREIRA *et al*., 2005). Para aproveitar a estrutura de correlações entre características e aumentar a precisão na predição do mérito genético, a utilização de metodologias multivariadas torna-se necessária (PAIVA *et al.*, 2019). Com esse intuito, uma técnica de análise multivariada utilizada no estudo da estrutura de correlação envolvendo várias características simultaneamente é a análise de fatores (AF) (RESENDE, 2015).

AF visa sintetizar as informações contidas em um conjunto de n variáveis observadas buscando um novo conjunto de p variáveis $(p < n)$, denominadas fatores latentes comuns (MACCIOTTA *et al*., 2012). Dessa forma, as variáveis latentes (fatores comuns) conseguem representar simultaneamente um conjunto de características, sem a perda de significado biológico, de modo que, os escores das novas variáveis formadas podem ser tratados como novos fenótipos (PAIVA *et al.,* 2019).

No melhoramento, a utilização de fatores se justifica devido a existência de correlações genéticas entre as características, as quais podem ser atribuídas aos QTL (locos de características quantitativas) que têm efeitos pleiotrópicos ou aos QTL estreitamente ligados. Desta forma, tais fatores podem ser utilizados em análises posteriores tais como em estudos de predição, de modo que a seleção dos indivíduos será realizada de maneira simultânea para estas

características. Uma vez que os fatores contêm informação de um conjunto de caracteres, é possível que os valores genéticos genômicos sejam preditos com maior acurácia.

AF vem sendo utilizada com sucesso em alguns estudos no melhoramento, como nos estudos com bovinos leiteiros de Macciotta *et al*. (2006, 2012) e com búfalo leiteiro Aspiculeta-Borquis *et al.* (2012), visando obter estimativas de parâmetros genéticos entre fatores latentes comuns. Paiva *et al.* (2019), utilizaram variáveis latentes formadas, identificadas como pseudo-fenótipos, na avaliação genética de frangos de corte sob uma estrutura bayesiana.

Especificamente em estudos de GWS, a utilização de AF foi proposta e avaliada por Teixeira *et al.* (2016). Esses autores obtiveram variáveis latentes, que permitiram estudar simultaneamente um conjunto de caracteres importantes e semelhantes em suínos, para posterior utilização em análises de GWS. Os fatores criados foram avaliados quanto à eficácia na predição genômica dos indivíduos e os autores observaram resultados satisfatórios em termos de acurácia, demonstrando que a utilização de variáveis latentes em estudos de GWS é uma abordagem interessante e promissora. Entretanto, apesar da importância e considerando o melhor de nosso conhecimento, o número de estudos no gênero *Coffea* envolvendo AF ou GWS ainda é reduzido.

Este trabalho tem por objetivo avaliar o uso da Análise de Fatores (AF) na obtenção de variáveis latentes (fatores comuns), que representem um conjunto de caracteres economicamente importantes em *C. canephora*. As variáveis latentes obtidas serão, posteriormente, utilizadas em estudos de seleção genômica (GWS), visando estimar os valores genéticos genômicos em acessos de *C. canephora.* Os resultados obtidos da seleção baseada nos fatores serão comparados, por meio da capacidade preditiva, acurácia e do coeficiente de Cohen's Kappa, com aqueles advindos da análise das variáveis individuais.

2. REFERENCIAL TEÓRICO

2.1. Análise de Fatores

A análise de fatores (AF) é um método para modelar a covariância entre um conjunto de variáveis observadas em função de uma ou mais construções latentes ou fatores. Tais fatores são considerados latentes por não serem diretamente observáveis, e podem ser usados para explicar o relacionamento entre um conjunto de variáveis (CORRAR; PAULO; DIAS FILHO, 2007). Assim, a AF é uma solução para reduzir dimensionalidade e permitir interpretações conjuntas das variáveis envolvidas no estudo.

Considerando o vetor aleatório p-dimensional $\mathbf{Y} = [\mathbf{Y_1}, \mathbf{Y_2}, ..., \mathbf{Y_p}]^{\mathbf{T}}$ com média $\boldsymbol{\mu}$ (p×1) e matriz de variâncias e covariâncias $\boldsymbol{\Sigma}$(p × p), o modelo fatorial é definido por:

$$\mathbf{Y} - \boldsymbol{\mu} = \boldsymbol{\Gamma}\mathbf{F} + \boldsymbol{\epsilon},$$

em que $\boldsymbol{\Gamma} = [\boldsymbol{\gamma_{ij}}]$ é uma matriz (p × m) de coeficientes, conhecidos por cargas fatoriais, de posto m ≤ p, $\mathbf{F}$ é um vetor aleatório (m × 1) de fatores comuns latentes não observáveis e $\boldsymbol{\epsilon}$ é um vetor (p × 1) de erros aleatórios denominados de fatores específicos (FERREIRA, 2018).

Algumas suposições adicionais a respeito dos fatores comuns, fatores específicos e variáveis originais são necessárias para a estimação do modelo, tais como: $\mathbf{E(Y)} = \boldsymbol{\mu}$, $\mathbf{E(F)} = \mathbf{E}(\boldsymbol{\epsilon}) = \mathbf{0}$, $\mathbf{Cov(F)} = \mathbf{I}_{\mathrm{m}}$, $\mathbf{Cov(Y)} = \boldsymbol{\Sigma}$, $\mathbf{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\psi}$ e $\mathbf{Cov}(\mathbf{F}, \boldsymbol{\epsilon}) = \mathbf{0}$ (m × p), sendo $\boldsymbol{\psi}$ dada por uma matriz diagonal com $\boldsymbol{\psi_i} > 0, \forall i = 1,2, \ldots p$. A matriz de covariâncias de $\mathbf{Y}$ pode ser decomposta em comunalidades e fatores específicos, isto é, $\mathbf{Cov(Y)} = \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathbf{T}} + \boldsymbol{\psi}$ (FERREIRA, 2018). Dessa forma, para operacionalizar a análise fatorial na prática, primeiramente, utiliza-se mecanismos para estimar o número de fatores $\boldsymbol{m}$, e a partir do valor estimado de $\boldsymbol{m}$ podemos então estimar as matrizes $\boldsymbol{\Gamma_{pxm}}$ e $\boldsymbol{\psi_{pxp}}$ (MINGOTI, 2005).

Com este intuito, o primeiro passo para conduzir a análise fatorial é estimar a matriz de correlação ($\widehat{\boldsymbol{\rho}} = \mathbf{R}_{\mathrm{pxp}}$) ou de covariâncias ($\widehat{\boldsymbol{\Sigma}} = \mathbf{S}_{\mathrm{pxp}}$). Para a estimação do número de fatores $\boldsymbol{m}$, temos que extrair os autovalores da matriz $\mathbf{R}$ ou $\mathbf{S}$ e ordená-los em ordem decrescente. Observam-se, então, quais autovalores são mais importantes em termos de magnitude e, permanecem na análise, aqueles autovalores que representarem maiores proporções da variância total. Portanto, o valor de $\boldsymbol{m}$ será igual ao número de autovalores retidos (MINGOTI, 2005).

Determinando-se o valor de $\boldsymbol{m}$, é possível estimar as matrizes $\boldsymbol{\Gamma_{pxm}}$ e $\boldsymbol{\psi_{pxp}}$. Essa estimação pode ser realizada pelo método dos componentes principais, que se baseia na decomposição espectral da matriz de correlação ou covariância, que é dada por:

$$\hat{\mathbf{\Gamma}} = \left[\sqrt{\hat{\boldsymbol{\lambda}}_1\hat{\boldsymbol{e}}_1}, \sqrt{\hat{\boldsymbol{\lambda}}_2\hat{\boldsymbol{e}}_2}, \ldots, \sqrt{\hat{\boldsymbol{\lambda}}_m\hat{\boldsymbol{e}}_m}\right] \text{ e } \widehat{\boldsymbol{\psi}} = \mathbf{diag}\left(\mathbf{R} - \hat{\mathbf{\Gamma}}\hat{\mathbf{\Gamma}}^{\mathbf{T}}\right),$$

em que $\hat{\boldsymbol{\lambda}}_{\boldsymbol{i}}$ $(\boldsymbol{i} = \mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{m})$ é o i-ésimo autovalor da matriz de correlações (**R**) ou de covariâncias (**S**) (FERREIRA, 2018; MINGOTI, 2005).

Em alguns casos, a interpretação dos fatores originais é dificultada devido à formação de coeficientes de grandeza numérica similar, e não desprezível, em diferentes fatores (MINGOTI, 2005). Neste caso, e para dar maior capacidade de interpretação dos fatores, é recomendado utilizar a técnica de rotação fatorial. Essa metodologia permite rearranjo dos autovetores, não alterando o total da variância explicada na etapa anterior. O tipo de rotação mais utilizado é a *varimax* (CORRAR; PAULO; DIAS FILHO, 2007).

Kaiser (1958) propôs como medida de rotação, a soma das variâncias das cargas fatoriais quadráticas dentro de cada coluna da matriz de cargas $\hat{\mathbf{\Gamma}}$. O valor de v é definido por:

$$v = \frac{1}{p^2}\sum_{j=1}^{m}\left[p\sum_{i=1}^{p}x_{ij}^4 - \left(\sum_{i=1}^{p}x_{ij}^2\right)^2\right],$$

em que $x_{ij} = \frac{\gamma_{ij}}{\sqrt{\sum_{j=1}^{m}\gamma_{ij}^2}}$ é a ij-ésima carga dividida pela raiz quadrada de sua respectiva comunalidade, que é a fração da variância explicada pelos fatores comuns (FERREIRA,2018).

Haja vista que ao realizar uma análise fatorial deseja-se reduzir um conjunto de variáveis correlacionadas a um conjunto de menor dimensão de fatores comuns não-correlacionados, com a intenção de utilizar essas novas variáveis em análises posteriores, é necessário predizer seus valores para cada unidade amostral. Esses valores são denominados escores fatoriais. Para tanto, pode-se fazer uso do método da regressão, em que os escores fatoriais para o j-ésimo indivíduo são preditos por:

$$\hat{\mathbf{F}}_{\mathbf{j}} = \hat{\mathbf{\Gamma}}^{\mathbf{T}}\left(\hat{\mathbf{\Gamma}}\hat{\mathbf{\Gamma}}^{\mathbf{T}} + \widehat{\boldsymbol{\psi}}\right)^{-\mathbf{1}}\left(\mathbf{Y}_{\mathbf{j}} - \overline{\mathbf{Y}}\right),$$

sendo $\widehat{\mathbf{\Gamma}}_{\mathbf{pxm}}$ a matriz das cargas fatoriais, $\widehat{\boldsymbol{\psi}}$ a matriz dos fatores específicos, $\mathbf{Y}_{\mathbf{j}}$ o vetor referente a j-ésima unidade amostral e $\overline{\mathbf{Y}}$ o vetor de médias referente às variáveis fenotípicas avaliadas (FERREIRA, 2018).

O ajuste de um modelo de análise fatorial aos dados pressupõe que as variáveis sejam correlacionadas entre si. Desse modo, é interessante verificar a adequação dos dados a este tipo de análise. Com esse objetivo, podemos destacar o coeficiente Kaiser-Meyer-Olkin (KMO), proposto inicialmente por Kaiser (1970) e o teste de esfericidade de Bartlett (RENCHER, 2002; JOBSON, 1996). O coeficiente KMO é definido por (MINGOTI, 2005):

$$KMO = \frac{\sum_{i \neq j} R_{ij}^2}{\sum_{i \neq i} R_{ij}^2 + \sum_{i \neq j} Q_{ij}^2},$$

em que R_{ij} é a correlação amostral entre as variáveis X_i e X_j, e Q_{ij} é a correlação parcial entre X_i e X_j. A correlação parcial entre duas variáveis é a correlação que existe entre elas quando todas as outras (p-2) variáveis são consideradas como constantes. O índice KMO varia entre zero e 1 e quanto maior o valor, melhor será a adequabilidade dos dados.

O teste de esfericidade de Bartlett tem por objetivo verificar se a matriz de correlação é próxima ou não da identidade. As hipóteses testadas são:

H_0: $\boldsymbol{\rho}_{pxp} = \mathbf{I}_{pxp}$ contra H_a: $\boldsymbol{\rho}_{pxp} \neq \mathbf{I}_{pxp}$,

em que $\mathbf{I}_{pxp}$ é a matriz identidade e $\boldsymbol{\rho}_{pxp}$ é a matriz de correlação das p-variáveis. A estatística de teste,T, é definida por:

$$T = -[n - 1/6(2p + 11)] \left[\sum_{j=1}^{p} \ln \hat{\lambda}_i\right],$$

em que ln(.) denota a função logaritmo neperiano e $\hat{\lambda}_i$, $i = 1, 2\ldots, p$ são autovalores da matriz de correlação amostral $\mathbf{R}_{pxp}$, n é o tamanho amostral e p é o número de variáveis. Sob a hipótese nula e n grande, a estatística T tem uma distribuição aproximadamente qui-quadrado com $\frac{1}{2}p(p-1)$ graus de liberdade. Para que o modelo de análise fatorial possa ser ajustado, o teste de Bartlett deve rejeitar a hipótese nula (MINGOTI, 2005).

2.2. Seleção Genômica Ampla

A identificação de genótipos superiores é um componente crítico da maioria dos programas de melhoramento, em razão do alto custo da fenotipagem e da pesquisa experimental (SANT'ANNA *et al.,* 2019). Em culturas perenes como o café, que demandam mais de trinta anos para obtenção de uma nova cultivar (CARVALHO *et al.*, 2011), acentua-se a importância da utilização de métodos de seleção acurados.

O advento dos marcadores moleculares ofereceu oportunidade para obter ganhos genéticos mais rápidos (LANDE e THOMPSON, 1990). Meuwissen *et al.* (2001) propuseram a Seleção Genômica Ampla (*Genome Wide Selection* - GWS), baseada no uso de marcadores SNPs (polimorfismo de um único nucleotídeo*,* ou *single nucleotide polymorphism*) para predição do fenótipo, por meio de metodologias estatísticas capazes de analisar um grande número de marcadores, independentemente da significância dos seus efeitos (SILVA *et al.,* 2013).

O SNP é a variação genética mais abundante no genoma (GANAL *et al*., 2009). A utilização de grande quantidade de marcas aumenta a probabilidade de capturar uma maior

proporção da variação genética, facilitando a seleção de genótipos superiores (MEUWISSEN; HAYES; GODDARD, 2001). Devido à complexidade e a arquitetura genética poligênica da maioria das características agronômicas do café, estudos de GWS são promissores, pois permitem estimar os efeitos de todos os locos que explicam a variação genética (SOUSA *et al*., 2019).

A soma dos efeitos dos marcadores (que se relacionam com os alelos presentes no genótipo de um indivíduo) é uma estimativa do desempenho fenotípico, e é conhecido como valor genético genômico (GEBV). Este, quando predito, ordenado e comparado é usado para selecionar os melhores indivíduos (OAKEY *et al.,* 2016). Entretanto, na prática, há fatores que fazem com que a soma de todas as marcas não representem o valor genético verdadeiro, consequência disso, o sucesso da GWS, depende da herdabilidade, da arquitetura genética das características, bem como da disponibilidade de desequilíbrio de ligação (LD) entre marcadores e QTL (locos de características quantitativas) (SANT'ANNA *et al.,* 2019).

Dessa forma, o método ideal para a GWS deve considerar a arquitetura genética do caráter em termos de genes de pequenos e grandes efeitos e suas distribuições; realizar a regularização do processo de estimação em presença de multicolinearidade, além de poder realizar a seleção de marcadores (RESENDE *et al.,* 2014). A GWS tem sido estudada em várias culturas para aumentar as taxas de ganho genético e reduzir a duração dos ciclos de melhoramento. Apesar de sua relevância, há apenas um número modesto de estudos aplicados ao gênero *Coffea* (ALKIMIM, 2017; FERRÃO *et al*., 2017; SOUSA *et al.,*2019)*,* por isso sua implementação efetiva nesta cultura depende da capacidade de considerar métodos que consigam capturar mais informação da sua constituição genotípica (FERRÃO *et al*., 2017).

O modelo geral de seleção genômica proposto por Meuwissen *et al*. (2001) é definido como:

$$\mathbf{y} = \mathbf{1}\boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

em que $\mathbf{y}$ é o vetor de fenótipos, $\mathbf{1}$ é um vetor com elementos iguais a um, $\mathbf{X}$ é matriz de incidência de marcadores; $\boldsymbol{\beta}$ é o vetor de efeitos genéticos aditivos desconhecidos de marcadores e $\boldsymbol{\epsilon}$ representa o vetor de erros aleatórios, com $\boldsymbol{\epsilon} \sim N\ (0, {\sigma^2}_e)$.

No contexto da GWS, os GEBVs dos indivíduos podem ser preditos por diferentes metodologias estatísticas, em que os principais métodos são classificados como métodos de regressão explícita, divididos em dois grupos: métodos de estimação penalizada *Least Absolute Shrinkage and Selection Operator* (LASSO), *Random Regression/Ridge Regression Best Linear Unbiased Predictor* (RR-BLUP), *Genomic Best Linear Unbiased Predictor* (G-BLUP),

dentre outros e os métodos de estimação bayesiana (Bayes A, Bayes B, *Bayesian LASSO* (BLASSO), entre outros).

Os métodos bayesianos se diferem devido ao tipo e extensão do *shrinkage*; capacidade de aprender com os dados e pela influência da distribuição *a priori*. Na estimação bayesiana, o encurtamento das estimativas dos efeitos do modelo é controlado pela distribuição *a priori* assumida para esses efeitos. Diferentes distribuições *a priori* induzem a diferentes encurtamentos. Já os estimadores penalizados são obtidos como solução para um problema de otimização, em que, a função cujo valor é minimizado é definida pelo balanço entre a soma dos quadrados dos resíduos e o componente de penalização. Os métodos de estimação penalizada diferem de acordo com as funções de penalização usadas, as quais produzem diferentes graus de encurtamento (RESENDE *et al.,* 2014).

2.3. G-BLUP

O Melhor Preditor Genômico Imparcial Linear G-BLUP (*Genomic Best Linear Unbiased Predictor*) tem sido extensivamente aplicado ao GWS e recomendado para predições de valor genômico (LIMA *et al.,*2019). Isto se deve ao fato deste apresentar as vantagens de relativa simplicidade, tempo reduzido de computação e propriedade de modelos mistos para a seleção. Nesse modelo, a predição de efeitos aleatórios, pode ser realizada na presença de efeitos fixos, pelos BLUPs (Melhor Preditor Linear Não-Viesado) (FERNANDO e GIANOLA, 1986).

No G-BLUP, os marcadores são usados para estimar as relações entre os indivíduos. Essa informação é usada ainda em uma análise de modelo misto para predizer o desempenho de indivíduos observados e não observados (FERNANDO e GIANOLA, 1986), caracterizando o seguinte modelo linear misto em nível de indivíduos para valores genéticos aditivos individuais (modelo G-BLUP):

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zu_a} + \boldsymbol{\epsilon},$$

em que $\boldsymbol{y}$ é o vetor de fenótipos (Nx1, em que N é o número de indivíduos); $\boldsymbol{b}$ é o vetor de efeitos fixos (px1, em que p é o número de efeitos fixos), $\boldsymbol{u_a}$ é o vetor de efeitos genéticos aditivos dos indivíduos (Nx1, aleatório), sendo as estruturas de variâncias dadas por $\boldsymbol{u_a} \sim N(0, G_a\sigma_a^2)$ em que σ_a^2 é a variância aditiva do caráter e $\boldsymbol{G_a}$ (N×N) é a matriz de parentesco genômica entre indivíduos para os efeitos aditivos; $\boldsymbol{\epsilon}$ é o vetor de efeitos residuais aleatório (Nx1) $com\ \boldsymbol{\epsilon} \sim N(0, I\sigma_e^2)$ em que σ_e^2 é a variância residual; $\boldsymbol{X}$ é a matriz de incidência para os efeitos fixos (Nxp); $\boldsymbol{Z}$ é a matriz de incidência para os efeitos aleatórios (NxN).

As equações de modelos mistos para predição de $\boldsymbol{u_a}$ por meio do método G-BLUP, equivalem a:

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z + G_a^{-1}\frac{\sigma_e^2}{\sigma_a^2}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\hat{b}} \\ \boldsymbol{\hat{u}_a} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \end{bmatrix},$$

em que a matriz de parentesco genômica aditiva **G** deve ser positiva definida, e os componentes de variância (σ_e^2 *e* σ_a^2) são estimados via REML (*Restricted Maximum Likelihood*) por produzir estimativas não viesadas. A matriz **G** é dada por:

$$\boldsymbol{G_a} = \frac{\mathbf{ww'}}{\sum_{i=1}^{n} \boldsymbol{2p_iq_i}},$$

em que **w** contém os valores 0, 1 e 2 para o número de alelos do marcador em um indivíduo diploide, e p_i e q_i são as frequências alélicas. Os efeitos de marcadores podem ser obtidos pela expressão: $\boldsymbol{\hat{m}_a} = (\mathbf{W'\ W})^{-1}\mathbf{W\hat{u}_a}$ (RESENDE *et al.,* 2014).

2.4. Correlação e Capacidade Preditiva

A correlação avalia o grau de relacionamento linear entre duas variáveis, seus valores estão entre -1 e 1. A correlação fenotípica é estimada diretamente de medidas fenotípicas, sendo resultante, portanto, de causas genéticas e ambientais. A correlação genotípica, que corresponde à porção genética da correlação fenotípica, é mais usada para orientar programas de melhoramento, por ser de natureza herdável (FERREIRA *et al*., 2003). As principais causas da correlação genética são decorrentes da pleiotropia, causa principal, e a outra ocasionada pela ligação gênica, causa temporária (CRUZ e REGAZZI, 1997).

Para verificar a precisão na predição dos valores genéticos genômicos, há necessidade de avaliá-los por meio da estimativa de parâmetros genéticos. A capacidade preditiva (CP) é uma medida prática para verificar o quanto o valor predito por um determinado modelo se aproxima do valor fenotípico observado. Logo, quanto maior a capacidade preditiva do modelo, maior é a confiança na avaliação e no valor genético predito dos indivíduos (RESENDE, 2015). A CP é obtida por meio da correlação entre os valores genômicos preditos e os valores fenotípicos observados, $r_{y,\hat{y}}$, isto é:

$$r_{y,\hat{y}} = \frac{Cov(y,\hat{y})}{\sqrt{Var(y).Var(\hat{y})}},$$

em que $Cov(y,\hat{y})$é a covariância entre o valor fenotípico observado e o valor genômico predito, $Var(y)$ é a variância dos valores fenotípicos observados, $e\ Var(\hat{y})$ é a variância dos valores genéticos genômicos preditos.

2.5. Herdabilidade e Acurácia

Uma maneira de estudar um caráter, de forma que se possa inferir sobre o seu potencial genético para fins de seleção, é utilizando a herdabilidade. Esta mede a proporção da variação fenotípica na população atribuída à causa genética. A herdabilidade será igual a 1 quando toda a variação expressa for de natureza genética e, será zero quando, a variação entre os indivíduos for unicamente de natureza ambiental. Sendo representada pelo símbolo h^2, pode ser estimada por meio da razão entre a variância genotípica v_g e a variância fenotípica v_{fen} (CRUZ, 2005):

$$h^2 = \frac{v_g}{v_{fen}}.$$

Outra medida para avaliar o ajuste do modelo é a acurácia. Esta depende da herdabilidade da característica e da capacidade preditiva (CP), em que, quanto maior o seu valor, melhor preditor do valor genético é o método de seleção. A qual pode ser obtida pelo seguinte estimador (RESENDE, 2015):

$$\mathrm{r}_{\mathrm{q},\hat{\mathrm{q}}} = \mathrm{r}_{\mathrm{y},\hat{\mathrm{y}}}/\sqrt{\mathrm{h}^2},$$

em que $\mathrm{r}_{\mathrm{y},\hat{\mathrm{y}}} = \mathrm{CP}$ e h^2 é a herdabilidade.

2.6. Coeficiente Cohen's Kappa

Para verificar a concordância entre duas avaliações independentes, ou seja, a semelhança entre os indivíduos selecionados por meio de diferentes abordagens, pode-se utilizar o coeficiente Cohen's Kappa (k) proposto por Cohen (1960). Este índice leva em conta a probabilidade de a concordância ter ocorrido ao acaso, o que o torna uma medida mais precisa. Este índice pode ser medido pela equação:

$$\hat{\mathrm{k}} = \frac{\mathrm{Pr(a)} - \mathrm{Pr(e)}}{1 - \mathrm{Pr(e)}},$$

em que $\mathrm{Pr(a)} - \mathrm{Pr(e)}$ representa a proporção de observações em que a concordância ocorreu além do que se esperava aleatoriamente e $1 - \mathrm{Pr(e)}$, a proporção de observações que não ocorreu concordância. Essa medida varia de zero a 1 e quanto maior o índice, maior a concordância entre os grupos (TEIXEIRA *et al*., 2016).

3. MATERIAIS E MÉTODOS

3.1. Dados fenotípicos e informações genotípicas

Para verificar a viabilidade da utilização das variáveis latentes em GWS, foram utilizadas populações compostas por clones dos grupos varietais Conilon e Robusta e por híbridos originados de cruzamentos entre estes grupos.

Os dados de Conilon são provenientes do Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural (INCAPER) e o material de Robusta foi obtido do Centro Agronómico Tropical de Investigación y Enseñanza (CATIE). Esses genótipos compõem o programa de melhoramento da Empresa de Pesquisa Agropecuária de Minas Gerais (Epamig) em parceria com a Universidade Federal de Viçosa (UFV) e a Empresa Brasileira de Pesquisa Agropecuária – Café (Embrapa Café).

Os grupos varietais Conilon e Robusta são constituídos por 51 e 32 genótipos, respectivamente. Além desses genótipos, 82 híbridos interpopulacionais foram obtidos por meio de cruzamentos artificiais entre cinco genótipos do grupo Conilon e cinco do grupo Robusta. Foram avaliadas as seguintes características durante três anos consecutivos (2014 a 2016): Vigor vegetativo (**Vig**), avaliado pelo aspecto geral da planta, utilizando-se escala de notas de 1 a 10, sendo a nota 1 atribuída às plantas totalmente depauperada e 10 às plantas altamente vigorosas; Altura da planta (**Apl**), medida em centímetros (cm), do nível do solo até o último ponto apical do cafeeiro com o auxílio de uma trena métrica; Diâmetro da projeção da copa (**Dco**), determinado em centímetros (cm) por meio de régua no sentido perpendicular à linha de plantio; e a Produção por planta (**Prod**), avaliada colhendo todos os frutos presentes em um genótipo e mensurado o volume total em litros de café, recém colhido. Detalhes sobre as características podem ser consultados em ALKIMIM (2017).

Os dados fenotípicos observados foram corrigidos para efeitos ambientais de anos e blocos, utilizando o *software* Selegen REML/BLUP (RESENDE, 2016). O modelo utilizado equivale a:

$$\boldsymbol{y = Xu + Tc + Wf + Zm + Qs + Sb + e,}$$

em que as matrizes de incidência dos efeitos são representadas pelas letras maiúsculas, $\boldsymbol{y}$ é o vetor de dados fenotípicos; $\boldsymbol{u}$ é o vetor de efeitos de médias de anos (assumidos como fixos) somados à média geral; $\boldsymbol{c}$ é o vetor de efeitos da capacidade específica de combinação entre os genitores Conilon e Robusta (assumidos como aleatórios); $\boldsymbol{f}$ é o vetor de efeitos da capacidade geral de combinação do parental Robusta (assumidos como aleatórios); $\boldsymbol{m}$ é o vetor de efeitos da capacidade geral de combinação do parental Conilon (assumidos como aleatórios); $\boldsymbol{s}$ é o

vetor de efeitos de ambiente permanente de indivíduos (assumidos como aleatórios); $\boldsymbol{b}$ é o vetor de efeitos de ambiente permanente de blocos (assumidos como aleatórios); $\boldsymbol{e}$ é o vetor de resíduos (assumidos como aleatórios).

Além disso, foi extraído DNA de folhas jovens e totalmente expandidas dos mesmos 165 cafeeiros fenotipados, por meio da metodologia descrita por Diniz *et al*. (2005). As amostras foram enviadas para a Rapid Genomics (Florida/EUA) para a identificação dos marcadores moleculares. Um total de 18111 marcadores SNPs foram identificados para as populações em estudo, estes, foram submetidos à análise de qualidade implementada no *software* GenomicLand (AZEVEDO *et al.,*2019).

O controle de qualidade foi realizado por meio da eliminação de locos pouco polimórficos (MAF– *minor allele frequency* = 5%); e/ou dos locos com baixa taxa de atendimento na genotipagem dos indivíduos (*call rate* = 95%). A densidade de SNP antes e após o controle de qualidade foi de 18111 a 14387 SNPs respectivamente, representando uma redução do conjunto inicial de marcadores em 20,56%.

3.2. Análise Fatorial

Com o objetivo de estudar várias características simultaneamente, estimou-se a correlação do conjunto de dados fenotípicos corrigidos e aplicou-se a análise de fatores na matriz resultante. Tal metodologia tem como objetivo resumir grandes grupos de dados, reduzindo sua dimensionalidade e permitindo interpretações conjuntas das variáveis envolvidas no estudo, de modo que essas novas variáveis (interpretáveis e não observáveis) sejam capazes de explicar a maior parte da variação total (TEIXEIRA *et al.,* 2015).

Considerando o vetor aleatório p-dimensional $\mathbf{Y} = [\mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_p}]^{\mathbf{T}}$ com média $\boldsymbol{\mu}$(p×1) e matriz de variâncias e covariâncias $\boldsymbol{\Sigma}$(p × p), o modelo fatorial utilizado é definido como segue:

$$\mathbf{Y} - \boldsymbol{\mu} = \boldsymbol{\Gamma}\mathbf{F} + \boldsymbol{\epsilon},$$

em que $\boldsymbol{\Gamma} = [\boldsymbol{\gamma_{ij}}]$ é uma matriz (p × m) de coeficientes conhecidos por cargas fatoriais de posto m ≤ p, **F** é um vetor aleatório (m × 1) de fatores comuns latentes não observáveis e $\boldsymbol{\epsilon}$ é um vetor (p × 1) de erros aleatórios denominados de fatores específicos (FERREIRA, 2018).

A adequação da estrutura de correlação foi avaliada usando o método de Kaiser-Meyer-Olkin critério (KMO) e teste de Bartlett (MINGOTI, 2005). O número de fatores **m** foi determinado conforme Ferreira (2011), isto é, considerando um percentual de explicação de pelo menos 70% da proporção da variância total das características. Para maximizar a variabilidade das cargas fatoriais e facilitar a interpretação da distribuição das variáveis nos

respectivos fatores, foi utilizada a rotação Varimax. Desta forma, a alocação das variáveis em cada fator foi feita por meio das cargas fatoriais, que consistem na correlação entre cada variável e os respectivos fatores. Assim, as variáveis que fizeram parte do fator são aquelas mais correlacionadas a este.

Posteriormente, os escores fatoriais utilizados foram obtidos pelo método de regressão, dado por:

$$\hat{\mathbf{F}}_{\mathbf{j}} = \hat{\boldsymbol{\Gamma}}^{\mathbf{T}}\left(\hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Gamma}}^{\mathbf{T}} + \widehat{\boldsymbol{\Psi}}\right)^{-\mathbf{1}}\left(\mathbf{Y}_{\mathbf{j}} - \overline{\mathbf{Y}}\right),$$

em que $\widehat{\boldsymbol{\Gamma}}_{\mathrm{pxm}}$ é a matriz das cargas fatoriais, $\widehat{\boldsymbol{\Psi}}$ a matriz dos fatores específicos, $\mathbf{Y}_{\mathbf{j}}$ o vetor referente a j-ésima unidade amostral e $\overline{\mathbf{Y}}$ o vetor de médias referente às variáveis fenotípicas avaliadas (FERREIRA, 2018).

3.3. Seleção Genômica

Os escores fatoriais calculados para cada genótipo foram utilizados como pseudo-fenótipos (características) juntamente com os 14387 marcadores SNPs em análise de GWS, a fim de estimar os valores genéticos genômicos (GEBVs) dos 165 indivíduos. Para tanto, utilizou-se o modelo linear misto G-BLUP, para estimar os valores genéticos aditivos individuais, conforme a expressão:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zu_a} + \boldsymbol{\epsilon},$$

em que **y** é o vetor de pseudo-fenótipos ($N \times 1$, em que N é o número de indivíduos); **b** é o vetor de efeitos fixos ($p \times 1$, em que p é o número de efeitos fixos), $\boldsymbol{u_a}$ é o vetor de efeitos genéticos aditivos dos indivíduos ($N \times 1$, aleatório), sendo as estruturas de variâncias dadas por $\boldsymbol{u_a} \sim N(0, G_a\sigma_a^2)$ em que σ_a^2 é a variância aditiva do caráter e $\boldsymbol{G_a}$ ($N \times N$) é a matriz de parentesco genômica entre indivíduos para os efeitos aditivos; $\boldsymbol{\epsilon}$ é o vetor de efeitos residuais aleatório $(N \times 1)\, com\ \boldsymbol{\epsilon} \sim N(0, I\sigma_e^2)$ em que σ_e^2 é a variância residual; $\boldsymbol{X}$ é a matriz de incidência para os efeitos fixos ($N \times p$); $\boldsymbol{Z}$ é a matriz de incidência para os efeitos aleatórios ($N \times N$) (RESENDE *et al.,* 2014). Além disso, análises considerando as características individualmente foram realizadas para fins de comparação.

Após a estimativa do mérito genético (GEBV) para cada fator interpretável e para cada característica individualmente, estimou-se a correlação genética entre as características. Os modelos preditivos também foram comparados por meio da capacidade preditiva, herdabilidade e acurácia.

3.3.1. Capacidade Preditiva, Herdabilidade e Acurácia

A capacidade preditiva (CP) foi estimada a partir da correlação entre os valores fenotípicos observados para cada indivíduo (y) e seus respectivos valores genéticos genômicos estimados (GEBVs) ($\hat{y}$), isto é: $r_{y,\hat{y}} = \frac{Cov(y,\widehat{y})}{\sqrt{Var(y).Var(\widehat{y})}}$.

Para que a CP não seja superestimada, foi utilizada a técnica de validação cruzada. Este método permite avaliar a capacidade de generalização de um modelo preditivo em um conjunto de dados (SOUSA *et al.,* 2019). O método de validação cruzada empregado foi o *K-Folds*, sendo considerado nesse trabalho *K* igual a 5 folds.

A herdabilidade (h^2), que mede a proporção da variação fenotípica na população atribuída à causa genética, é dada por meio da razão entre a variância genotípica v_g e a variância fenotípica v_{fen} (CRUZ, 2005): $h^2 = \frac{v_g}{v_{fen}}$.

A acurácia depende da h^2obtida e da CP, sendo uma medida que está associada à precisão na seleção, e é obtida pelo estimador (RESENDE, 2015): $r_{q,\hat{q}} = r_{y,\hat{y}}/\sqrt{h^2}$, em que $r_{y,\hat{y}} = CP$ e h^2 é a herdabilidade.

3.4. Cohen's Kappa

Com base no mérito genético, foram selecionados 10% dos melhores indivíduos, para cada abordagem utilizada. De posse destes indivíduos, foi calculada a concordância entre os indivíduos selecionados, com base em cada característica individualmente e pelos fatores, utilizando-se o coeficiente Cohen's Kappa (Cohen, 1960). Este índice pode ser medido por:

$$\hat{k} = \frac{Pr(a) - Pr(e)}{1 - Pr(e)},$$

em que $Pr(a) - Pr(e)$ representa a proporção de observações em que a concordância ocorreu além do que se esperava aleatoriamente e $1 - Pr(e)$, a proporção de observações que não ocorreu concordância. Essa medida varia de zero a 1, e quanto maior o índice, maior a concordância acordo entre os grupos formados.

3.5. Ferramentas computacionais

A análise de fatores foi realizada no *software* R (R CORE TEAM, 2019), utilizando a função *principal* por meio pacote do *psych* (REVELLE, 2019). As análises estatísticas genômicas foram realizadas no *software* GenomicLand (AZEVEDO *et al.,*2019).

4. RESULTADOS E DISCUSSÃO

4.1. Correlação fenotípica

A análise de correlações permite determinar o grau de relacionamento entre duas variáveis. Na Tabela 1 estão apresentados os coeficientes de correlação fenotípica entre as características agronômicas avaliadas. De acordo com o teste *t,* a 1% de probabilidade, foram identificadas seis associações positivas e significativas, com estimativas variando entre 0,21 a 0,84 (Tabela 1).

Tabela 1. Correlação de Pearson entre as características fenotípicas avaliadas.

	Vig	**Apl**	**Dco**	**Prod**
Vig	1,00	0,56**	0,84**	0,21**
Apl		1,00	0,81**	0,30**
Dco			1,00	0,28**
Prod				1,00

Vig - vigor vegetativo; Apl- altura da planta; Dco - diâmetro da projeção da copa; Prod - produção por planta.
**= Significativo a 1% de probabilidade de erro pelo teste T.

O conhecimento das correlações entre as características é útil no processo de seleção, principalmente quando há necessidade de ser feita seleção simultânea de caracteres (BENIN *et al.,*2005). Este permite avaliar o quanto da alteração de um caráter pode afetar as demais características (SILVA *et al.,* 2009). Dessa forma, no contexto de melhoramento do gênero *Coffea*, a presença de correlações entre as características representa um indicativo de que estas podem ser melhoradas conjuntamente, devido ao comportamento associativo entre elas. Permitindo assim, a opção por métodos de seleção mais adequados.

As características vigor vegetativo (**Vig**), altura da planta (**Apl)** e diâmetro da copa (**Dco**) por se correlacionarem de forma positiva e significativa (valor p<0,01), é indicativo de que para estas, não é possível realizar seleção baseada em uma única característica sem resultar em alterações nas demais. Nesse caso, ainda indica que existe uma tendência de plantas mais altas apresentarem maior vigor e diâmetro de copa.

De acordo com Lopes *et al*. (2002), existe uma tendência de se valorizar mais o sinal (positivo ou negativo) e a magnitude dos valores na interpretação aplicada das correlações, valorizando as estimativas abaixo de -0,5 e acima de 0,5. Dentre as características avaliadas, a produção (**Prod)** é aquela que apresenta menor associação com as demais, apesar de significativas, cujas estimativas variaram entre 0,21 e 0,30 (Tabela 1). A baixa correlação da

produção com as demais características estudadas é um indicativo de que não é recomendado, aplicar a seleção conjunta para produção com as demais características avaliadas neste estudo, pois estas não são fortemente associadas.

A estimativa das correlações entre as características estudadas permitiu verificar as associações entre elas, o que pode subsidiar o planejamento de estratégias de seleção para o melhoramento genético do *C. canephora*.

4.2. Análise de fatores das características avaliadas

Para verificar a adequação do conjunto de dados para análise de fatores, foi utilizado o índice proposto por Kaiser (KMO). Sendo baseado na comparação entre os valores de correlação fenotípica de Pearson e correlações parciais. O valor obtido pela estatística KMO foi de 0,58, acima do valor proposto para a adequabilidade do método ao conjunto de dados (0,5) sugerido por Hair *et. al* (2005). Além disso, o teste de esfericidade de Bartlett foi estatisticamente significativo ($valor\ p < 0{,}01$), demonstrando a adequação da análise fatorial.

Na Tabela 2, estão apresentados os autovalores da matriz de correlação fenotípica com as respectivas porcentagens de variação total explicada. Utilizando-se o critério de escolha do número *m* de fatores que expliquem pelo menos 70% da variância total (FERREIRA, 2011), tem-se o valor estimado *m*= 2, que, neste estudo, foi capaz de explicar 87,25% da variância total (Tabela 2).

Tabela 2. Autovalores da matriz de correlação ($\widehat{\lambda_i}$) e porcentagem de variância explicada.

Ordem (i)	$\widehat{\boldsymbol{\lambda_i}}$	**Variância Explicada (%)**	**Variância Acumulada (%)**
1	2,61	65,24	65,24
2	0,88	22,01	87,25
3	0,43	10,80	98,05
4	0,08	1,944	100,0

Considerando o critério acima citado, verificou-se a formação de dois fatores latentes (Tabelas 2 e 3). Os dois fatores obtidos apresentam interpretação biológica, atestando que é possível elucidar a estrutura de relações entre as características analisadas. Esse resultado sugere que pode ser incluída como meta em programas de melhoramento de café, a seleção com base em complexos de variáveis simplificadas pela análise fatorial. Ferreira *et al.* (2005), ressaltaram que a seleção de genótipos com base em um conjunto de caracteres de importância é necessária, visando ganhos simultâneos em várias características de interesse agronômico.

Tabela 3. Fatores observados (Fi), comunalidades (h2) e unicidades (ψ).

Variáveis	F1	F2	h2	ψ
Vig	**0,89**	0,04	0,80	0,20
Apl	**0,83**	0,23	0,75	0,25
Dco	**0,97**	0,14	0,95	0,05
Prod	0,15	**0,98**	0,99	0,01

Vig - vigor vegetativo; Apl- altura da planta; Dco - diâmetro da projeção da copa; Prod - produção por planta.

Por meio dos coeficientes (cargas fatoriais) relacionados com cada fator para cada característica avaliada (Tabela 3), e sabendo que estes coeficientes representam a correlação entre o fator e a característica, observou-se que o primeiro fator (F1) que explicou cerca de 65% da variância original (Tabela 2), agrupou características associadas à morfologia do indivíduo (Vig, Apl e Dco), sendo interpretado como "fator vigor". Todas estas características estão positivamente e altamente correlacionadas com o fator. Logo, maiores valores para estas características estarão associados à maiores valores do "fator vigor".

O segundo fator (F2), que explica cerca de 22% da variação (Tabela 2), é altamente correlacionado apenas com a característica de produção (Tabela 3), sendo este nomeado como "fator produção". Tal resultado já era esperado, visto a estrutura de correlação das variáveis, que demonstram o elevado grau de associação das características que compõe o F1 e à baixa correlação destas com a variável produção (Tabela 1). Dessa forma, para a característica de produção, pode ser mais interessante analisá-la individualmente, sem estar ligada a um complexo fatorial. Este resultado corrobora com os obtidos por Ferreira *et al.* (2005) e Barbosa *et al.* (2019), que avaliaram a possibilidade do emprego de índices de seleção em complexos fatoriais, em caracteres de genótipos de *C. canephora* e *C. arabica* respectivamente. Estes autores também não agruparam em nenhum fator a característica de produção.

Verificou-se que as variáveis pertencentes aos fatores apresentaram um valor aceitável para as comunalidades ($h2 > 0,50$), superiores a 0,75, indicando que mais de 75% da variabilidade dessas características foi explicada pelos dois fatores formados (Tabela 3). Esse resultado indica que a variabilidade presente na população proporciona subsídio para predizer os ganhos genéticos e o possível sucesso no programa de melhoramento.

É interessante destacar que para as características avaliadas, poucos fatores foram suficientes para explicar quase 90% da variabilidade dos dados. Tal resultado pode ser atribuído principalmente à forte correlação entre esses caracteres (Tabela 1) e a quantidade de variáveis

avaliadas neste trabalho. Resultados semelhantes foram obtidos por Ferreira *et al.* (2005) que estudaram 14 características em 40 genótipos de *C. canephora*. Esses autores mencionaram que quatro fatores explicaram um mínimo de 70% da variabilidade total dos dados. Segundo Barbosa *et al.* (2019), esse tipo de resultado demonstra que as variáveis avaliadas no melhoramento de cafeeiros apresentam certo padrão de correlação e podem ser resumidas por meio de fatores comuns.

Após predição dos valores para cada unidade amostral (escores para cada fator), as variáveis latentes puderam ser tratadas como pseudo-fenótipos e utilizadas em estudos de GWS.

4.3. Análise de GWS

A predição de valores genéticos genômicos (GEBV's) foi realizada por meio da metodologia G-BLUP, para os dois fatores formados e, para as quatro características avaliadas individualmente, para fins de comparação. Foram obtidos os seguintes parâmetros genéticos: herdabilidade, capacidade preditiva média, correlação entre os GEBV's e a acurácia do modelo.

A herdabilidade refere-se à proporção relativa das influências genéticas e ambientais na manifestação fenotípica das características e indica o grau de facilidade ou dificuldade para melhorar determinados caracteres (BOURDON, 2000). Características com herdabilidade mais baixa são geralmente controladas por mais genes e, portanto, a seleção é mais complexa (SOUSA *et al.,* 2019). Sua compreensão possibilita a execução dos procedimentos e estratégias a serem adotadas nas etapas do desenvolvimento de um genótipo superior (FALCONER, 1987; RESENDE, 2002).

De acordo com a classificação proposta por Resende (1997), a herdabilidade pode ser classificada como baixa (menor que 0,15), média ou moderada (entre 0,15 e 0,50) ou alta magnitude (maior que 0,50). Neste estudo, observou-se características classificadas em todos os níveis, uma vez que as herdabilidades variaram de 0,0034 a 0,6029 (Tabela 4). Com exceção da produção por planta e sua correspondente variável latente, a magnitude de herdabilidade foi de moderada a alta, indicando que as características são herdáveis, com a possibilidade de sucesso na seleção com base no "fator vigor", além de demonstrar um bom controle genético dos caracteres.

Os resultados são corroborados por aqueles obtidos por Alkimim (2017), utilizando uma versão bayesiana do método G-BLUP e densidade de marcadores igual a 14429 SNPs em genótipos de *C.canephora*, em que as estimativas de herdabilidade variaram de 0,15 para a característica produção a 0,53 para a característica diâmetro da copa.

Tabela 4. Herdabilidade (h^2), capacidade preditiva e acurácia para as características e os fatores formados.

Características	h^2	Capacidade Preditiva	Acurácia
Vig	0,4899	0,3964	0,5663
Apl	0,4522	0,3722	0,5535
Dco	0,6029	0,5732	0,7382
Prod	0,0034	0,0052	0,0892
F1	0,5568	0,5054	0,6773
F2	0,0087	-0,0298	-0,3195

Vig - vigor vegetativo, Apl- altura da planta, Dco - diâmetro da projeção da copa, Prod - produção por planta, F1- "fator vigor", F2- "fator produção".

Observa-se que o valor da CP referente ao "fator vigor" (0,51) é maior quando comparado as variáveis Apl (0,37), Vig (0,39), e se aproxima da CP da característica Dco (0,57) (Tabela 4), indicando capacidade em predizer o fenótipo futuro dos indivíduos por meio do "fator vigor". Teixeira *et al.* (2016) também observaram que os valores obtidos de CP para o fator são, em geral, semelhantes ou superiores, àqueles obtidos pelas análises individuais. Alkimim (2017), também observou boa capacidade preditiva para as características Vig (0,44), Apl (0,41) e Dco (0,58).

Observa-se que o valor da acurácia referente ao "fator vigor" (0,68) é maior quando comparado as características Apl (0,55), Vig (0,57), que foram analisadas individualmente pelo método G-BLUP, e inferior quando comparado a variável Dco (0,74) (Tabela 4). A acurácia é classificada em muito alta, quando maior que 0,90; alta, quando entre 0,70 e 0,90; moderada, entre 0,50 e 0,70 e baixa, quando menor que 0,50 (RESENDE e DUARTE, 2007). Dessa forma, para essas características e a variável latente, a acurácia é classificada como moderada, indicando boa precisão na estimativa do mérito genético, uma vez que as estimativas de herdabilidade foram consideradas de moderada a alta.

Sabendo-se que, devido a particularidades agronômicas do cafeeiro (período juvenil longo e acentuada oscilação bianual de produção), são necessários vários anos para avaliar a precocidade de produção e longevidade da planta para realizar a seleção (PETEK; SERA; FONSECA, 2008). Assim, métodos preditivos de GWS combinados com a seleção conjunta de características, apresentam-se como de grande importância. Dessa forma, a partir dos resultados observados, em termos de seleção podemos verificar que a variável latente torna-se interessante, pois além de permitir interpretações conjuntas, apresenta boas estimativas de CP e acurácia.

Na Tabela 5 observa-se as estimativas de correlação dos GEBVs entre as características analisadas e os fatores formados, que variaram entre –0,15 a 0,99. A correlação entre os valores genéticos genômicos pode ser considerada como uma estimativa da correlação genética. Quando a correlação é positiva, é possível obter ganhos para uma característica selecionando outra. E, quando negativa, a seleção para uma característica irá contribuir negativamente em outra (RESENDE, 2002).

Tabela 5. Estimativas da correlação entre os valores genéticos genômicos para as características avaliadas e os fatores identificados (triangular superior) e coeficientes Cohen´s Kappa (triangular inferior)

	Vig	**Apl**	**Dco**	**Prod**	**F1**	**F2**
Vig	1,00	0,72**	0,90**	-0,07^{ns}	0,93**	-0,15^{ns}
Apl		1,00	0,88**	0,15^{ns}	0,90**	0,09^{ns}
Dco			1,00	0,08^{ns}	0,98**	0,00^{ns}
Prod				1,00	0,00^{ns}	1,00**
F1	0,74^{MB}	0,34^{R}	0,74^{MB}		1,00	-0,07^{ns}
F2				0,87^{E}		1,00

Vig - vigor vegetativo; Apl- altura da planta; Dco - diâmetro da projeção da copa; Prod - produção por planta; F1– fator 1; F2 – fator 2. **= Significativo a 1% de probabilidade de erro pelo teste T; ns = não significativo. E= Excelente, MB = Muito Bom e R= Razoável.

Correlações positivas e significativas (valor $p < 0,01$) dos GEBVs (Tabela 5), foram verificadas entre as características vigor vegetativo e diâmetro da projeção da copa (0,90), entre altura da planta e diâmetro da copa (0,88) e entre vigor vegetativo e altura da planta (0,72). Como o "fator vigor" representa este conjunto das características, maiores valores de correlação são observados entre elas (valor $p < 0,01$). Além disso, esta correlação genética favorável indica que é possível alcançar ganhos para uma característica por meio da seleção indireta da outra característica (RIBEIRO *et al.,* 2001). Esse resultado sugere também que alguns genes que influenciam a expressão desses fenótipos podem ser os mesmos ou a ocorrência da pleiotropia. De forma a pormenorizar esse resultado, podemos citar o estudo de associação genômica em genótipos de *C. canephora* de Silva (2018). Utilizando densidade de 17885 marcadores, observou a presença de SNP com associações significativas, nos genes para essas três características.

Verificou-se que as correlações genéticas (valor $p < 0,01$) apresentaram magnitudes superiores às suas correspondentes correlações fenotípicas (Tabela 1 e 5). Esse resultado é um

indicativo de que os fatores genéticos tiveram maior influência que os de ambiente, possibilitando, assim, a seleção simultânea de várias características, uma vez que o interesse do melhorista recai, quase sempre, em um conjunto delas (FERRÃO *et al.*, 2007). Além disso, as diferenças verificadas nas estimativas de correlação entre as abordagens genômicas e fenotípicas, são justificadas devido a maior quantidade de informações que são usadas em GWS. Outro aspecto refere-se à capacidade dos SNPs de capturar variantes causais associadas às características (SOUSA *et al.,* 2019).

Coeficientes de concordância entre os 10% melhores indivíduos selecionados com base em cada variável individualmente, e com base nos fatores estão apresentados na Tabelas 5. Na tabela verificam-se os valores obtidos pelo coeficiente Cohen's Kappa e sua classificação de acordo com Landis e Koch (1977). Os valores estimados de Cohen's Kappa variaram de 0,34 até 0,87. Especificamente, considerando o primeiro fator (Vigor) apenas a característica altura de plantas teve um valor estimado de Cohen's Kappa considerado ruim (0,34) (Landis e Koch, 1977). Todos os outros valores estimados de Cohen's Kappa foram superiores a 0,50 indicando uma classificação muito boa (F1 x Vig = 0,74; F1 x Dco = 0,74) e excelente (F2 x Prod = 0,87) (Landis e Koch, 1977).

No contexto do melhoramento de *C. canephora*, objetiva-se a identificação de genótipos mais vigorosos e de maior diâmetro de copa, características que são positivamente correlacionadas a maior altura de planta. Este comportamento constitui uma problemática/entrave na seleção dessa cultura, uma vez que do ponto de vista fitotécnico, cultivares de café de porte baixo são melhores, pois, permitem o adensamento de plantas na área, proporcionando maiores produtividades, além de reduzir a suscetibilidade das plantas a ventos frios, facilitando a colheita e os tratos culturais (CARVALHO, 2008).

Entretanto, observa-se que os genótipos selecionados por meio do "fator vigor" apresentam uma menor concordância com aqueles selecionados considerando a característica de altura da planta (Tabela 5). Esse resultado mostra-se interessante do ponto de vista fitotécnico, pois como discutido anteriormente, cultivares de café de porte baixo são melhores por permitirem o adensamento de plantas na área.

Em termos de valores fenotípicos, a seleção com base no "fator vigor" permitiu selecionar indivíduos de menor porte, com uma redução de 26,46% em média na altura. Já para as outras características (Dco = 2,45% e Vig = 8,62%), redução percentual foi menor quando comparada com aquele obtida para a característica Apl (Tabela 6).

Tabela 6. Média, mediana, desvio padrão (DP) e coeficiente de variação (CV), para os 10% melhores genótipos selecionados para cada característica individualmente e por meio do "fator vigor".

Parâmetro	Apl	Apl*	Dco	Dco*	Vig	Vig*
Média	338,8	249,2	264,5	258,0	15,20	13,89
Mediana	362,5	254,3	266,0	266,0	14,11	13,93
DP	93,94	95,95	57,95	64,65	4,52	5,44
CV	27,73%	38,50%	21,91%	25,06%	29,73%	39,14%

Vig - vigor vegetativo; Apl- altura da planta; Dco - diâmetro da projeção da copa; Vig* - vigor vegetativo selecionada com base no "fator vigor"; Apl*- altura da planta selecionada com base no "fator vigor"; Dco* - diâmetro da projeção da copa selecionada com base no "fator vigor".

O bom desempenho da AF e da GWS destaca a importância da utilização dessas ferramentas no gênero *Coffea*. Em *C. Canephora*, o desenvolvimento de novas cultivares pode levar décadas, mas isso pode ser acelerado com a incorporação dessas metodologias. Nesse sentido, associar a análise de fatores a informações genômicas, incluindo estes em estudos de GWS, permite melhor manejo da quantidade de variação presente em acessos avaliados, de forma que se apresenta como uma boa opção para obter maiores ganhos genéticos em prazos mais curtos, o que é potencialmente útil para programas de melhoramento cafeeiro.

5. CONCLUSÃO

A análise de fatores foi eficiente, pois conseguiu elucidar a estrutura de relações das características avaliadas e formar novas variáveis, interpretadas como “fator vigor” e “fator produção”. Os fatores formados apresentaram um percentual satisfatório de variabilidade explicada e podem que ser tratados como novos fenótipos em estudos de seleção genômica.

As características vigor vegetativo, altura da planta e diâmetro da copa em *C. canephora* podem ser melhoradas conjuntamente devido ao seu elevado grau de associação.

O uso de fatores mostrou-se uma abordagem promissora associada a GWS para o melhoramento de *C. canephora*, pois além de permitir interpretações conjuntas, apresenta boas estimativas de capacidade preditiva, herdabilidade e acurácia.

Ademais, a concordância entre os 10% melhores indivíduos selecionados com base no “fator vigor” e em cada variável individualmente, foi considerada satisfatória, com o incremento da vantagem do “fator vigor” selecionar indivíduos de porte mais adequado.

6. REFERÊNCIAS

ASPILCUETA-BORQUIS, R. R., *et al*. Genetic parameters of total milk yield and factors describing the shape of lactation curve in dairy buffaloes. **Journal of Dairy Research**, v. 79, p. 60-65, 2012.

ALKIMIM, E. R. **Diversidade genética, ganhos com seleção e seleção genômica ampla na espécie *coffea canephora.*** 2017. 105p. Tese (Doutorado em Genética e Melhoramento) – Universidade Federal de Viçosa, Viçosa, 2017.

AZEVEDO, C. F., *et al*. GenomicLand: Software for genome-wide association studies and genomic prediction. **Acta Scientiarum. Agronomy**, v. 41, 2019.

BABOVA, O.; OCCHIPINTI, A.; MAFFEI, M. E. Chemical partitioning and antioxidant capacity of green coffee (*Coffea arabica* and *Coffea canephora*) of different geographical origin. **Phytochemistry**, v. 123, p. 33-39, 2016.

BARBOSA, I. P., *et al*. Recommendation of Coffea arabica genotypes by factor analysis. **Euphytica**, 2019.

BENIN, G., *et al.* Estimativas de correlações genotípicas e de ambiente em gerações com elevada freqüência de heterozigotos. **Ciência Rural**, v.35, p.523-529, 2005

BOURDON, G. E. P. **Understanding animal breeding**. Upper Saddle River: Prendice-Hall, 2000. 538 p.

CAVALCANTI, J. J. V., *et al*. Predição simultânea dos efeitos de marcadores moleculares e seleção genômica ampla em cajueiro. **Rev. Bras. Frutic**., v. 34, p. 840-846, 2012.

CARVALHO, C. H. S. (ed.). **Cultivares de café**. Brasília: Embrapa, 2008. 247p.

CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. **Análise Multivariada para os Cursos de Administração, Ciências Contábeis e Economia**. 1. ed. São Paulo: Atlas S. A., 2007. 541p.

COHEN, J. A coeficiente of agrément for nominal scales. **Educational and Psychological Measurement,** v. 20, p. 37-46, 1960.

CRUZ, C.D.; REGAZZI, A.J. **Modelos biométricos aplicados ao melhoramento genético**. Viçosa: UFV, 1997. 390 p.

CRUZ, C. D. **Princípios de genética quantitativa.** 1. Ed. Viçosa: UFV, 2005. 394p.

DINIZ, L. E. C., *et al.* Analysis of AFLP markers associated to the Mex-1 resistance locus in Icatu progenies. **Crop Breed Appl Biotechnol**, v. 05, 2005.

EMBRAPA. **Consumo interno dos Cafés do Brasil representa 13% da demanda mundial**. 12 fev. 2019. Disponível em: <https://www.embrapa.br/busca-de-noticias/-/noticia/41277124/consumo-interno-dos-cafes-do-brasil-representa-13-da-demanda-mundial.> Acesso em : 22 mar. 2019.

FALCONER, D. S. **Introdução à genética quantitativa**. Viçosa, UFV: Impr. Univ., 1987. 279p.

FASSIO, L. H.; SILVA, A. E. S. da. **Importância econômica e social o café Conilon**. 2007. Disponível em: <https://biblioteca.incaper.es.gov.br/digital/bitstream/item/694/1/livro2007cafeconilon1.pdf> Acesso em: 22 mar. 2019.

FERNANDO, R. L.; GIANOLA, D. Optimal properties of the conditional mean as a selection criterion. **Theoretical and applied genetics**, p. 822-825, 1986.

FERREIRA, M. A. J. F., *et al*. Correlações genotípicas, fenotípicas e de ambiente entre dez caracteres de melancia e suas implicações para o melhoramento genético. **Horticultura Brasileira**, v. 21, p. 438-442, 2003.

FERREIRA, A., *et al.* Seleção simultânea de por meio da combinação de análise de fatores e índices de seleção. **Pesq. agropec. bras.,** v.40, p.1189-1195, 2005.

FERREIRA, D. F. **Estatística Multivariada**. 2.ed. Lavras: Ed. UFLA, 2011. 675p.

FERREIRA, D. F. **Estatística multivariada**. 3.ed. Lavras: Ed. UFLA, 2018. 624p.

FERRÃO, M. A. G., *et al*. Genetic divergence in conilon coffee revealed by RAPD markers. **Crop Breed Appl Biotechnol,** 2009.

FERRÃO, L. F. V**.,** *et al.* A mixed model to multiple harvest-location trials applied to genomic prediction in *Coffea canephora*. **Tree Genetics & Genomes.** 2017

GANAL, M. W., *et al*. SNP identification in crop plants. **Curr. Opin. Plant** Biol. V. 12, p.211–217. 2009

GLANTZ, M., *et al*. Genomic selection in relation to bovine milk composition and processability. **J. Dairy Res**. 2011.

HAIR, Jr., *et al.* **Multivariate Data Analysis**. 6. ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006. 889p.

HENDRE, P.S., *et al*. Development of new genomic microsatellite markers from robusta coffee (*Coffea canephora* Pierre ex A. Froehner) showing broad cross-species transferability and utility in genetic studies. **BMC plant biology**, 2008.

IVOGLO, M. G., *et al.* Divergência genética entre progênies de café robusta. **Bragantia**, v. 67, p. 823-831, 2008.

JOBSON, J. D. **Applied multivariate data analysis**. V. I e II. New York: Springer Verlag, 1996. 731p.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. **Psychometrika,** v. 23, p. 187-200, 1958.

KAISER, H. An index of factor simplicity. **Psychometrika,** p. 31–36. 1974.

LANDE, R., e THOMPSON, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. **Genetics** v.124, p. 743–756, 1990.

LANDIS, J. R., E KOCH, G. G. The measurement of observer agreement for categorical data. **biometrics**, p. 159-174, 1977.

LIMA, L. P, *et al.* New insights into genomic selection through population-based non-parametric prediction methods. **Scientia Agricola,** v. 76, p. 290-298, 2019.

LOPES, A. C. de A., *et al.* Variabilidade e correlações entre caracteres em cruzamentos de soja. **Scientia Agrícola**, v. 59, p. 341-348, 2002.

MACCIOTTA, N. P. P.; VICARIO, D.; CAPPIO-BORLINO, A. Use of multivariate analysis to extract latent variables related to level of production and lactation persistency in dairy cattle. **Journal of dairy science**, v. 89, p. 3188-3194, 2006.

MACCIOTTA, N. P. P.; CECCHINATO, A.; MELE, M.; BITTANTE, G. Use of multivariate factor analysis to define new indicator variables for milk composition and coagulation properties in Brown Swiss cows. **Journal of dairy science**, v. 95, p. 7346-7354, 2012.

MINGOTI, S. A. **Análise de dados através de estatística multivariada: uma abordagem aplicada**. Belo Horizonte. Ed. UFMG, 2005. 297 p.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using Genome-Wide dense marker maps. **Genetics Society of America**. v.157, p. 1819-1829, 2001.

OAKEY, H., *et al.* Genomic selection in multi-environment crop trials. **G3: Genes, Genomes, Genetics,** p. 1313-1326, 2016.

PAIVA, J. T., *et al.* "Genetic evaluation for latent variables derived from factor analysis in broilers." **British Poultry Science**, 2019.

PETEK, M. R.; SERA, T.; FONSECA, I. C. de B. Predição de valores genéticos aditivos na seleção visando obter cultivares de café mais resistentes à ferrugem. ***Bragantia*****,** v. 67, p.133-140, 2008.

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>. 2019.

RENCHER, A. C. **Methods of multivariate analysis**. New York: John Wiley, 2002.

RESENDE, M. D. V. de. Avanços da genética biométrica florestal. In: ENCONTRO SOBRE TEMAS DE GENÉTICA E MELHORAMENTO, 14., 1997, Piracicaba. **Anais**... Piracicaba: Esalq, 1997. 20-46 p.

RESENDE, M. D. V. de. **Genética biométrica e estatística no melhoramento de plantas perenes**. Embrapa Informação Tecnológica. Colombo: Embrapa Florestas, Brasília, 2002. 975p.

RESENDE, M.D.V. **Matemática e estatística na análise de experimentos e no melhoramento genético**. Colombo: Embrapa Florestas, 2007. 561 p.

RESENDE, M. D. V e DUARTE, J. B. Precisão e controle de qualidade em experimentos de avaliação de cultivares. **Pesquisa Agropecuária**, p.182–194, 2007.

RESENDE, M. D. V. de., *et al*. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, p.63-77, 2008.

RESENDE, M. D. V de. Software Selegen-REML/BLUP: a useful tool for plant breeding. **Crop Breed Appl Biotechnol**, 2016.

RESENDE, M. D. V.; SILVA, F. F.; AZEVEDO, C. F. **Estatística matemática, biométrica e computacional: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência**. Viçosa: Suprema, 881p. 2014.

RESENDE, M. D. V. de. **Genética quantitativa e de populações.** – Viçosa, MG: Ltda, 2015. 463p.

REVELLE, W. psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, <https://CRAN.R-project.org/package=psych Version = 1.9.12>. 2019.

RIBEIRO, N. D., *et al.* Correlações Genéticas de Caracteres Agromorfológicos e suas Implicações na Seleção de Genótipos. **Rev. Bras. de AGROCIÊNCIA,** v.7, p.93-99, 2001.

SANT'ANNA, I. C., *et al*. Multigenerational prediction of genetic values using genome-enabled prediction. **PloSone**, 2019.

SILVA, M. A., *et al.* Análise de trilha para caracteres morfológicos do feijão-bravo (*Capparis flexuosa*) no cariri paraibano. **Archivos de Zootecnia**, v.58, p.121-124. 2009.

SILVA, F. F., *et al.* Seleção genômica ampla para curvas de crescimento. **Arq. Bras. Med. Vet. Zootec.,** v.65, p.1519-1526, 2013.

SILVA, L. de F. **Estudo de associação genômica ampla (GWAS) em *Coffea canephora.*** 2018. 32p. Dissertação (Mestrado em Genética e Melhoramento) – Universidade Federal de Viçosa, Viçosa, 2018.

SOUSA, T. V., *et al.* Molecular markers useful to discriminate *Coffea arabica* cultivars with high genetic similarity. **Euphytica**, v. 213, p. 75, 2017.

SOUSA, T. V., *et al.* Early Selection Enabled by the Implementation of Genomic Selection in *Coffea arabica* Breeding. **Frontiers in Plant Science**, v. 9, p. 1-12, 2019.

TEIXEIRA, F. R. F., *et al*. Determinação de fatores em características de suínos. **Rev. Bras. Biom.**, v.33, p.130-138, 2015

TEIXEIRA, F. R. F., *et al*. Factor analysis applied to genome prediction for high-dimensional phenotypes in pigs. **Genetics and Molecular Research- GMR**, 2016.

TRAN, H. T., *et al*. Advances in genomics for the improvement of quality in coffee. **Journal of the Science of Food and Agriculture**, p. 3300-3312, 2016.