INDALÉCIO CUNHA VIEIRA JÚNIOR

EFFICACY OF GAUSSIAN MIXTURE MODELS FOR GENOTYPE SELECTION IN COFFEE BEAN

LAVRAS –MG

2020

INDALÉCIO CUNHA VIEIRA JÚNIOR

EFFICACY OF GAUSSIAN MIXTURE MODELS FOR GENOTYPE SELECTION IN COFFEE BEAN

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, área de concentração em Genética e Melhoramento de Plantas, para a obtenção do título de Doutor.

Orientador

Drª. Flávia Maria Avelar Gonçalves

Coorientador

Dr. Márcio Balestre

LAVRAS –MG
2020

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Júnior, Indalécio Cunha Vieira.
Efficacy of gaussian mixture models for genotype selection in coffee bean / Indalécio Cunha Vieira Júnior. - 2020.
80 p.: il.

Orientador(a): Flávia Maria Avelar Gonçalves.
Coorientador(a): Márcio Balestre.
Tese (doutorado) - Universidade Federal de Lavras, 2020.
Bibliografia.

1. Bienniality. 2. Mixture Models. 3. Genomic Selection. I. Gonçalves, Flávia Maria Avelar. II. Balestre, Márcio. III. Título.

INDALECIO CUNHA VIEIRA JUNIOR

EFFICACY OF GAUSSIAN MIXTURE MODELS FOR GENOTYPE SELECTION IN COFFEE BEAN

EFICÁCIA DOS MODELOS DE MISTURA GAUSSIANO PARA SELEÇÃO DE GENÓTIPOS DE CAFEEIRO

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, área de concentração em Genética e Melhoramento de Plantas, para a obtenção do título de Doutor em Genética e Melhoramento de Plantas.

APROVADA em 28 de fevereiro de 2020.

Dr. Vinicius Quintão Carneiro	UFLA
Dr. Evandro Novaes	UFLA
Dr. Alan carvalho Andrade	EMBRAPA
Dr. André Dominghetti Ferreira	EMBRAPA

Orientador

Drª. Flávia Maria Avelar Gonçalves

Coorientador

Dr. Márcio Balestre

LAVRAS – MG

2020

A Deus, por me guiar nas decisões mais importantes da minha vida e me fazer crer que sempre podemos melhorar.

OFEREÇO

ACKNOWLEDGMENTS

Agradeço a Deus por sempre me abençoar e direcionar meu caminho.

A minha família por sempre me apoiar nas decisões mais difícies e compreender as minhas escolhas.

Ao professor Márcio Balestre, pela paciência, pelo exemplo de cientista e por sempre me ensinar a pensar, não apenas ténicos, mas pricipalmente os da vida real. Certamente a pessoa que mais influeciou a minha maneira de enxergar o mundo. Espero profudamente que nossa jornada continue depois daqui.

A professora Flavia Avelar e ao grupo de mehoramento de perenes, por toda ajuda, apoio e confiança.

Agradeço aos professores Julio Bueno, Daniel Furtado e José Airton por sempre estarem dispostos a me receber e tirar minhas dúvidas.

Aos meus amigos e colegas da UFLA, por tornarem as coisas mais fáceis. Em especial Kaio Olympio, Bruna Line, Samuel Fernades (Prosa), Guilherme de Jong (Gui), Vitor Passos e Mario Henrique (Bola), pelo conhecimento e companheirismo.

To the IPK (LEIBNIZ-INSTITUT FÜR PFLANZENGENETIK UND KULTURPFLANZENFORSCHUNG) where I stayed during some part of my PhD. I am especially grateful for Professor Jochen Reif who trusted on me and gave me this great opportunity. I am also grateful to Dr. Young Jiang for the productive discussions about quantitative genetics and to Fraust Beate for the friendship. And I appreciate the support of the quantitative genetic group.

A Universidade Federal de Lavras e ao programa de pos graduação em genética e melhoramento de plantas pela oportunidade.

O presente trabalho foi realizado com apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

A cada um dos pagadores de impostos que querendo ou não financiar este trabalho o fez pela coerção estatal.

A todos que de alguma maneira participaram para a realização desse trabalho.

“Not only is the Universe stranger than we think, it is stranger than we can think.”

Werner Heisenberg

RESUMO GERAL

Café é uma das principais commodities no mundo. Sabe-se que a produção do cefeeiro apresenta variações bruscas ao longo dos anos, fenômeno chamado de bienalidade. Tal comportamento impõe enormes desafios na análise estatística de dados de produção de café. Nessa espécie os genótipos apresentam comportamento diferencial devido a resposta fisiológica frente as condições ambientais o que sugere a formação de mistura de subpopulações. Estudos prévios propõe algumas soluções, porém eles assumem um único processo estocástico gerando o fenótipo. No primeiro artigo é proposto um modelo de mistura para lidar com o padrão da bienalidade, considerando o fenômeno uma variável latente. Realizou-se análises individuais (por colheita) e de medidas repetidas para modelos mistos padrão e modelo misto de mistura gaussiano. Houve aumento significativo na eficiência das estimativas dos parâmetros e maior ganho genético, sugerindo que na análise de dados de progênies de *C. arabica* exibindo diferentes padrões de bienalidade, modelos mistos de mistura são superiores a modelos mistos e a modelos que estruturam os efeitos da bienalidade com matrizes de covariância. No segundo artigo o modelo misto de mistura é estendido para predição genômica (GMGBLUP) e comparado com um modelo convencional de predição genômica (GBLUP). O objetivo foi verificar a acurácia preditiva quando os efeitos das marcas são corrigidos para a bienalidade. Nos dados reais o GBLUP gerou melhores resultados, entretanto nos dados simulados o GMGBLUP foi superior quando as subpopulações são contrastantes e o parâmetro de mistura é próximo de 0.5. Os resultados GMGBLUP deve ser considerado como uma alternativa para predição genômica em dados do gênero *Coffea*, especialmente em espécies com forte bienalidade.

Palavras-chave: Modelos de mistura, Modelos mistos, Seleção Genômica.

GENERAL ABSTRACT

Coffee is one of the most important traded commodities in the world. It is well known that coffee bean yield is subjected to strong variation through the years in a phenomenon called biennial growth. This behavior has imposed great challenges on statistical analysis of coffee bean yield data. In these species genotypes show a differential biennial behavior due to its physiological response to environmental condition which suggests a mixture of subpopulations. Previous studies have tried to solve the problem, however they assume the presence of only one stochastic process generating the phenotypes. In the first paper it is proposed a finite mixture model to deal with the biennial pattern as hidden variable. Individual (per harvest) and repeated measures analyses were performed using conventional mixed models and Gaussian mixture mixed models. The results showed a great increase on parameter efficiency estimation and lead to greater genetic gain suggesting that for analysis of *C. arabica* progenies exhibiting different biennial patterns, mixture mixed models are superior to traditional mixed models and to models that structure biennial effects using covariance matrices. On the second paper the gaussian mixed mixture model is extended for genomic prediction (GMGBLUP) and compared with a traditional genomic prediction model (GBLUP). The aim was to verify the prediction accuracy when the markers effects are corrected for bias of the biennial growth. For the real data set the GBLUP performed better in all harvests, however the simulated data results showed that the GMGBLUP is superior when the subpopulations means are contrasting and the mixture parameter is close to 0.5. The results suggest that GMGBLUP should be considered as an alternative for genomic prediction in *coffea* genus, especially for species with strong biennial growth behavior.

Keywords: Mixture Model, Mixed Model, Genomic Selection

SUMMARY

FIRST PART

1 GENERAL INTRODUCTION

Coffee is one of the most important commodities traded in the world (Davis et al., 2012; Tran et al., 2016). In 2018 it is estimated that the world produced more than 10M tons of coffee, from this volume Brazil and Vietnam were responsible to approximately 3.5 M and 1.6 M, respectively (FAO, 2018). Thus, coffee bean plays a fundamental role not only on the world economy, but also on cultural and socioeconomic life of developing countries.

Yield is the most important evaluated trait for Coffee bean progenies selection and one of the most complex. This happens not only due to its polygenic nature, but also because of the biennial growth behavior. These aspects affect selection and lead to small genetic gain and they are more problematic for coffee species since their perennial cycle requires long time of evaluation and are very costly.

Biennial growth is a well-known characteristic of coffee species. It causes strong variation on yield through harvest, therefore a genotype which shows a high yield one year, probably will decrease substantially its production on subsequent harvest. This phenomenon occurs because plants exhibiting biennial growth allocate photosynthetic products to fruit formation and growth during years of high production and to vegetative functions during years of low production, causing a pattern of production alternation (Bacha, 1998). Coffee therefore alternates between vegetative growth in one year and fructification in the next.

Among the methods used in the analysis of this type data, the two-year mean is one of the most common (Oliveira et al., 2011). It is an attempt to meet the assumptions required for analysis of variance (ANOVA). However, this strategy considers that the probability of all coffee plants being in the productive or vegetative phase is 1; thus, the production expectation is the arithmetic

mean of two years, and selection is always based on data sets considering the mean of two consecutive years. This approach is not the best way of managing this problem as it assumes homogeneous variance and null covariance between consecutive years, which causes information loss and bias in estimates of variance components.

These aspects require more complex and robust statistical model and tend to result in selection bias when the statistical model is not in accordance with the biological nature of the data (Hu and Spilke, 2011; Piepho and Eckl, 2014). More recently, (Andrade et al., 2016) proposed modeling coffee bean yield using mixed models, with the correct choice of genetic and residual covariance structures in order to capture the serial correlation through harvests. The results obtained by these authors showed improvement in the efficiency of parameter estimation when compared to ANOVA models.

All the afore mentioned methods assume a common stochastic process generating the phenotypes. As mentioned above, an attempt to meet this presupposition is by averaging harvests values of consecutive years and this can generate worse results. It assumes that all genotypes are in the same physiological stage and that this stage follows a (0, 1) sign function (that is, it assumes all genotypes are in the same stage (high or low) of production at a given year), which is not always true. For a given experimental field and harvest, one set of progenies may be in a high-production year and another in a low-production year, generating data overdispersion and a false signal, and compromising the experimental precision. In addition, the alternation between productive and vegetative stages is not always clear and may be triennial (e.g., two years of low production followed by one year of high production), and this particularity must be taken into account in the selection of the best genotypes. Many models used for the analysis of coffee traits assume that the data originate from a single (Gaussian) stochastic process. However, this assumption is not always

accurate because samples may originate from different non-observable processes. In these cases, mixture models may be used to search for latent factors (Murphy, 2012), as these models provide a very flexible tool to work with data having a finite number of unobserved subpopulations. Under the hypothesis of differential biennial growth, overdispersion may be modeled based on the mean of finite mixtures.

In the last years, the technological development allowed significant drop in the cost of genotyping and genome wide selection (GWS) has been successfully applied in breeding programs of different species (Crossa et al., 2014; Xu et al., 2014; Grinberg et al., 2016; Kwong et al., 2017). In coffee beans e other perennial species, some papers have showed great potential to increase the genetic gain per unit of time (Andrade et al., 2017; Ferrão et al., 2017, 2018).

Traditional GWS approaches uses the phenotypic information of the genotyped individuals to estimate the effect of each SNP and then predict the genetic merit other individuals (Meuwissen et al., 2001). There are many methods to estimate the SNP effects on literature (VanRaden, 2008; de los Campos et al., 2013; Azodi et al., 2019) and all of them consider that phenotypes come from a homogenous population, that is, the stochastic process which generates the studied trait is the same through the whole population. As mentioned above this is not true for coffee beans due to biennial bearing. Ignoring this phenomenon for markers estimation effect can cause strong bias and significantly decrease the prediction ability of GWS. In order to address this problem we extended the model proposed by (Vieira Júnior et al., 2019) and create a gaussian mixture GBLUP model, where the markers effects are estimated considering a mixture of two subpopulations. To the best of our knowledge this is the first time on the literature that this class of model is used for genomic prediction.

REFERENCES

Andrade, V.T., F.M.A. Gonçalves, J.A.R. Nunes, and C.E. Botelho. 2016. Statistical modeling implications for coffee progenies selection. Euphytica 207(1): 177–189Available at https://doi.org/10.1007/s10681-015-1561-6.

Andrade, A.C., O.B. da SILVA JUNIOR, F. CARNEIRO, P. Marraccini, and D. Grattapaglia. 2017. Towards GWAS and Genomic Prediction in Coffee: Development and Validation of a 26K SNP Chip for Coffea Canephora. *In* Embrapa Café-Resumo em anais de congresso (ALICE).

Azodi, C.B., A. McCarren, M. Roantree, G. de los Campos, and S.-H. Shiu. 2019. Benchmarking algorithms for genomic prediction of complex traits. bioRxiv: 614479.

Bacha, C.J.C. 1998. A cafeicultura brasileira nas décadas de 80 e 90 e suas perspectivas. Preços agrícolas 7(142): 14–22.

Crossa, J., P. Perez, J. Hickey, J. Burgueno, L. Ornella, J. Cerón-Rojas, X. Zhang, S. Dreisigacker, R. Babu, Y. Li, and others. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity (Edinb). 112(1): 48.

Davis, A.P., T.W. Gole, S. Baena, and J. Moat. 2012. The impact of climate change on indigenous arabica coffee (Coffea arabica): predicting future trends and identifying priorities. PLoS One 7(11): e47981.

F.A.O. 2018. Available online: http://www. fao. org/faostat/en/# data. QC (accessed January 2020).

Ferrão, L.F.V., R.G. Ferrão, M.A.G. Ferrão, A. Fonseca, P. Carbonetto, M. Stephens, and A.A.F. Garcia. 2018. Accurate genomic prediction of Coffea canephora in multiple environments using whole-genome statistical models. Heredity (Edinb).

Ferrão, L.F.V., R.G. Ferrão, M.A.G. Ferrão, A. Francisco, and A.A.F. Garcia. 2017. A mixed model to multiple harvest-location trials applied to genomic prediction in Coffea canephora. Tree Genet. Genomes 13(5): 95.

Grinberg, N.F., A. Lovatt, M. Hegarty, A. Lovatt, K.P. Skøt, R. Kelly, T. Blackmore, D.

Thorogood, R.D. King, I. Armstead, and others. 2016. Implementation of genomic prediction in Lolium perenne (L.) breeding populations. Front. Plant Sci. 7: 133.

Hu, X., and J. Spilke. 2011. Variance--covariance structure and its influence on variety assessment in regional crop trials. F. Crop. Res. 120(1): 1–8.

Kwong, Q. Bin, A.L. Ong, C.K. Teh, F.T. Chew, M. Tammi, S. Mayes, H. Kulaveerasingam, S.H. Yeoh, J.A. Harikrishna, and D.R. Appleton. 2017. Genomic selection in commercial perennial crops: applicability and improvement in oil palm (Elaeis Guineensis Jacq.). Sci. Rep. 7(1): 2872.

de los Campos, G., J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, and M.P.L. Calus. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193(2): 327–345.

Meuwissen, T.H., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4): 1819.

Murphy, K.P. 2012. Machine learning: a probabilistic perspective. Cambridge, MA.

Oliveira, A.C.B. de, A.A. Pereira, F.L. da Silva, J.C. de Rezende, C.E. Botelho, and G.R. Carvalho. 2011. Prediction of genetic gains from selection in Arabica coffee progenies. Crop Breed. Appl. Biotechnol. 11(2): 106–113.

Piepho, H.-P., and T. Eckl. 2014. Analysis of series of variety trials with perennial crops. Grass Forage Sci. 69(3): 431–440.

Tran, H.T.M., L.S. Lee, A. Furtado, H. Smyth, and R.J. Henry. 2016. Advances in genomics for the improvement of quality in coffee. J. Sci. Food Agric. 96(10): 3300–3312.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91(11): 4414–4423.

Vieira, I. Cunha, C. da Silva, J.J. Nuvunga, C.E. Botelho, F.M. Avelar Gonçalves, and M. Balestre. 2019. Mixture Mixed Models: Biennial Growth as a Latent Variable in Coffee Bean Progenies. Crop Sci.

Xu, S., D. Zhu, and Q. Zhang. 2014. Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proc. Natl. Acad. Sci. 111(34): 12456–12461.

SECOND PART

ARTICLE 1 - Mixture mixed models: biennial growth as a latent variable in coffee bean progenies (Coffea arabica L.)

Article published on Crop Science. Doi: 10.2135/cropsci2018.02.0141

Published May 30, 2019

RESEARCH

Mixture Mixed Models: Biennial Growth as a Latent Variable in Coffee Bean Progenies

Indalécio Cunha Vieira Júnior,* Carlos Pereira da Silva, Joel Jorge Nuvunga, César Elias Botelho, Flávia Maria Avelar Gonçalves, and Márcio Balestre

ABSTRACT

Statistical analysis of *Coffea arabica* L. progeny production has been a great challenge. In this species, genotypes may present differential biennial behaviors due to different physiological responses to the environmental conditions, indicating a mixture of two subpopulations in the tested progenies. Previously proposed statistical methods are unable to handle data overdispersion and/or bimodality because they assume the same stochastic process generating different phenotypes. This study proposes a finite mixture mixed model for modeling the biennial patterns. Production data for 21 $S_{0:1}$ progenies, evaluated through eight harvests, were used. Individual (per harvest) and repeated measures analyses were performed using conventional mixed models and Gaussian mixture mixed models. The proposed methodology is also illustrated in a simulation study. On a real dataset, the approximated prediction error variance, CV, and residual variance were drastically reduced using mixture mixed models, resulting in a higher estimated heritability and expected gain from selection. Residual dependence across years was lower for the mixture model, but no differences were observed in genetic correlations. The posterior probability matrix captured the biennial pattern, which indicates the probability of a progeny's physiological stage. The Spearman correlation coefficient (0.87) indicates that selection based on grouped means may not be efficient. In general, the proposed model was more efficient for higher subpopulations means differences. The results suggest that for analysis of *C. arabica* progenies exhibiting different biennial patterns, mixture mixed models are superior to traditional mixed models and to models that structure biennial effects using covariance matrices.

I.C. Vieira Júnior and F.M.A. Gonçalves, Dep. of Biology, Federal Univ. of Lavras, Lavras, Minas Gerais, Brazil; C.P. da Silva and M. Balestre, Dep. of Statistics, Federal Univ. of Lavras, Lavras, Minas Gerais, Brazil; J.J. Nuvunga, Dep. of Agriculture, Eduardo Mondlane Univ., Maputo, Maputo, Monzambique; C.E. Botelho, Empresa de Pesquisa Agropecuária de Minas Gerais-EPAMIG, Unidade Regional do Sul de Minas, Lavras, Minas Gerais, Brazil. Received 27 Feb. 2018. Accepted 25 Mar. 2019. *Corresponding author (indasjunior@hotmail.com). Assigned to Marcio Resende Jr.

Abbreviations: APEV, approximated prediction error variance; BIC, Bayesian information criterion; BLUP, best linear unbiased prediction; EBLUE, estimated best linear unbiased estimation; EBLUP, estimated best linear unbiased prediction; EM, expectation–maximization; GS, gain from selection; REML, restricted maximum likelihood.

Coffee (*Coffea* spp.) is one of the most important commodities in the world (Waller et al., 2007; Davis et al., 2012), and its cultivation is essential to the economy of Brazil and other developing countries (Lewin et al., 2004). It is estimated that, around the world, 125 million people are dependent on coffee for their livelihoods (Osorio, 2002).

An important characteristic of coffee is its marked fluctuation in production between successive years—that is, its alternation between low and high production in consecutive harvests. This phenomenon is called biennial growth (Rodrigues et al., 2014; Andrade et al., 2016) and is highly pronounced in the species *Coffea arabica* L. This biennial alternation results from the physiological nature of coffee plants, which need to vegetate for 1 yr to produce well in the following year (Rena et al., 1986).

Biennial growth is related to the sink–source relationships between leaves and fruit. Leaves are sources of photoassimilates, and growing tissues drain these metabolites. Plants exhibiting biennial growth allocate photosynthetic products to fruit formation and growth during years of high production and to vegetative functions during years of low production, causing a pattern of

Published in Crop Sci. 59:1424–1441 (2019).
doi: 10.2135/cropsci2018.02.0141

production alternation (Bacha, 1998). Coffee therefore alternates between vegetative growth in one year and fructification in the next.

Modeling this type of data is a challenge. Biennial growth may decrease selection efficiency. Furthermore, evidence indicates that biennial growth causes heterogeneity of variances and temporal correlation patterns over multiple harvests (Andrade et al., 2016), and the longitudinal feature of coffee bean yields makes the process of selecting the best progenies harder. Some authors have recommend the use of 2-yr means to decrease the effect of this phenomenon on analyses (de Oliveira et al., 2011) and to meet the assumptions required for ANOVA. This strategy considers that the probability of all coffee plants being in the productive or vegetative phase is 1; thus, the production expectation is the arithmetic mean of 2 yr, and selection is always based on datasets considering the mean of two consecutive years. However, this approach is not the best way of managing this problem, as it assumes homogeneous variance and null covariance between consecutive years, which causes information loss and bias in estimates of variance components. Recently, Andrade et al. (2016) proposed that this type of data should be modeled using mixed models, with the correct choice of genetic and residual covariance structures. The results obtained by these authors showed some increases in the efficiency of estimation compared with the use of ANOVA models.

Genotype evaluation and selection based on 2-yr means assumes that all genotypes are in the same physiological stage and that this stage follows a (0, 1) sign function (i.e., it assumes all genotypes are in the same stage [high or low] of production at a given year), which is not always true. For a given experimental field and harvest, one set of progenies may be in a high-production year and another in a low-production year, generating data overdispersion and a false signal and compromising the experimental precision. In addition, the alternation between productive and vegetative stages is not always clear and may be triennial (e.g., 2 yr of low production followed by 1 yr of high production), and this particularity must be taken into account in the selection of the best genotypes. Many models used for the analysis of coffee traits assume that the data originate from a single (Gaussian) stochastic process. However, this assumption is not always accurate because samples may originate from different unobservable processes. In these cases, mixture models may be used to search for latent factors (Murphy, 2012), as these models provide a very flexible tool to work with data having a finite number of unobserved subpopulations. Under the hypothesis of differential biennial growth, overdispersion may be modeled based on the mean of finite mixtures.

Finite mixture models have been applied in the identification of main-effect quantitative trait loci (QTL) (Fisch et al., 1996; Lynch and Walsh, 1998; Gianola et al., 2004), for selecting haploid seeds based on oil contents (Melchinger et al., 2013) and to select against mastitis in dairy cows using only somatic cell counts (Detilleux and Leroy, 2000; Gianola et al., 2004; Jamrozik and Schaeffer, 2010).

Several approaches have been proposed for indirectly modeling biennial growth in coffee, but finite mixture models can be a better alternative to model this phenomenon directly in tests of progenies. Thus, the aim of the present study was to propose a mixture mixed model to analyze coffee production data and clustering genotypes in different physiological phases.

MATERIALS AND METHODS

The data used in the present study originated from the coffee breeding program coordinated by the Agricultural Research Company of Minas Gerais (Empresa de Pesquisa Agropecuária de Minas Gerais [EPAMIG]). Twenty-one $S_{0:1}$ progenies were tested. The progenies derived from crosses between *C. arabica* cultivars (Mundo Novo × Mundo Novo and Mundo Novo × Bourbon Vermelho) originating from a distinct population developed at in the Agronomy Institute of Campinas (Instituto Agronômico de Campinas). The experiment was set up in the city of Machado, Minas Gerais State (21°40′ S, 45°55′ W), in a randomized complete block design with three replications. Each experimental unit (plot) consisted of eight plants with a spacing of 3.0 m between rows and 1.5 m between plants. Coffee production (kg) was evaluated over eight harvests. The type of soil and climate of the region are described in Andrade et al. (2016). In the present study, the terms "years," "harvests," and "environments" are used synonymously with "crop season."

Modeling

Individual (per harvest) and repeated measures analyses were performed using models with one or two mixture components, corresponding to differential biennial responses. In this case, four models were used: an individual mixed model (M1), an individual mixture mixed model (M2), an unstructured mixed model (M3), and an unstructured mixture mixed model (M4).

The individual analyses using the conventional mixed model (without biennially latent effects) were executed with the PROC MIXED procedure in SAS (SAS Institute, 2009). The remaining analyses were performed using code developed on the R platform. Variance parameters were estimated with the restricted maximum likelihood (REML) function using the expectation–maximization (EM) algorithm (Dempster et al., 1977). For all the analyses, the replicate (block) effect was considered as a fixed effect, and the progeny effect was considered as a random effect. The detailed models are described below.

Gaussian Mixed Model

In this scenario, the latent parameter related to biennial growth is absent, converging for a classical linear mixed model. In this framework, two approaches were adopted: individual analysis (per harvest), and repeated measures analysis with unstructured matrices for residuals and genotypes. The models are described below.

Individual Analyses

Per harvest (year) analyses were performed using the following mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \qquad [1]$$

where $\mathbf{y}_{(n\times 1)}$ is the phenotypic observation vector, $\boldsymbol{\beta}_{(b\times 1)}$ is the fixed effect (block) vector, $\mathbf{u}_{(g\times 1)}$ is the random effect (genotype) vector, and $\mathbf{e}_{(n\times 1)}$ is the residual vector. $\mathbf{X}_{(n\times b)}$ and $\mathbf{Z}_{(n\times g)}$ are the fixed and random effect incidence matrices, respectively, where subscripts n, b, and g are the number of observations, replications, and genotypes, respectively. The following distribution assumptions were made for the random effects:

$$\mathbf{u} \sim N(\mathbf{0}, \sigma_g^2\mathbf{I})$$

$$\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2\mathbf{I})$$

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \sigma_e^2\mathbf{I})$$

where σ_g^2 and σ_e^2 are the genetic and residual variance, respectively, and $\mathbf{I}$ is an identity matrix.

Repeated Measures Mixed Model

The analyses considering all harvests were performed using a model of repeated measures in time with unstructured residual and genetic covariance matrices, which has been proposed to model biennial growth (Andrade et al., 2016). The following repeated measures linear mixed model was used:

$$\mathbf{y}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \mathbf{Z}_k\mathbf{u}_k + \mathbf{e}_k \quad [2]$$

where $\mathbf{y}_{k_{(n\times 1)}}$ is the observation vector, $\boldsymbol{\beta}_{k_{(b\times 1)}}$ is the fixed effect (block) vector, $\mathbf{u}_{k_{(g\times 1)}}$ is the random effect (genotype) vector, and $\mathbf{e}_{k_{(n\times 1)}}$ is the error vector. $\mathbf{X}_{k_{(n\times b)}}$ and $\mathbf{Z}_{k_{(n\times g)}}$ are the fixed and random effect incidence matrices, respectively. The following distribution assumptions were made:

$$\mathbf{u}_k \sim N(0, \mathbf{G}\otimes\mathbf{I})$$

$$\mathbf{e}_k \sim N(0, \mathbf{R}\otimes\mathbf{I})$$

$$\mathbf{y}_k \sim N(\mathbf{X}_k\boldsymbol{\beta}_k + \mathbf{Z}_k\mathbf{u}_k, \mathbf{V})$$

where $\mathbf{G}$ and $\mathbf{R}$ are the unstructured genetic and residual covariance matrices, respectively. All the matrices have $k \times k$ dimensions. We consider $\mathbf{V} = \mathbf{R}\otimes\mathbf{I}$.

The repeated measures model is better visualized as follows:

$$\mathbf{y}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \mathbf{Z}_k\mathbf{u}_k + \mathbf{e}_k$$

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \vdots\\ \mathbf{y}_k\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{O} & \cdots & \mathbf{O}\\ \mathbf{O} & \mathbf{X}_2 & & \vdots\\ \vdots & & \ddots & \mathbf{O}\\ \mathbf{O} & \cdots & \mathbf{O} & \mathbf{X}_3\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\\ \vdots\\ \boldsymbol{\beta}_k\end{bmatrix} + \begin{bmatrix}\mathbf{Z}_1 & \mathbf{O} & \cdots & \mathbf{O}\\ \mathbf{O} & \mathbf{Z}_2 & & \vdots\\ \vdots & & \ddots & \mathbf{O}\\ \mathbf{O} & \cdots & \mathbf{O} & \mathbf{Z}_3\end{bmatrix}\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\\ \vdots\\ \mathbf{u}_k\end{bmatrix} + \begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\\ \vdots\\ \mathbf{e}_k\end{bmatrix} \qquad [3]$$

where each subscript corresponds to a vector or matrix of experimental observations for each evaluated year.

From the matrix system of equations for the mixed model, solutions for $\boldsymbol{\beta}_k$ and $\mathbf{u}_k$ can be obtained as follows:

$$\begin{bmatrix}\hat{\boldsymbol{\beta}}_k\\ \hat{\mathbf{u}}_k\end{bmatrix} = \begin{bmatrix}\mathbf{X}_k'\mathbf{R}^{-1}\mathbf{X}_k & \mathbf{X}_k'\mathbf{R}^{-1}\mathbf{Z}_k\\ \mathbf{Z}_k'\mathbf{R}^{-1}\mathbf{X}_k & \mathbf{Z}_k'\mathbf{R}^{-1}\mathbf{Z}_k + \mathbf{G}^{-1}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_k'\mathbf{R}^{-1}\mathbf{y}\\ \mathbf{Z}_k'\mathbf{R}^{-1}\mathbf{y}\end{bmatrix} \qquad [4]$$

To facility the notation in this work, it will assume that

$$\mathbf{C} = \begin{bmatrix}\mathbf{X}_k'\mathbf{R}^{-1}\mathbf{X}_k & \mathbf{X}_k'\mathbf{R}^{-1}\mathbf{Z}_k\\ \mathbf{Z}_k'\mathbf{R}^{-1}\mathbf{X}_k & \mathbf{Z}_k'\mathbf{R}^{-1}\mathbf{Z}_k + \mathbf{G}^{-1}\end{bmatrix} = \begin{bmatrix}\mathbf{C}_{11} & \mathbf{C}_{12}\\ \mathbf{C}_{21} & \mathbf{C}_{22}\end{bmatrix} \qquad [5]$$

The EM algorithm was used to estimate the best linear unbiased prediction (BLUP) and to obtain the REML solutions for $\mathbf{G}$ and $\mathbf{R}$, as shown below (here, the indices i and j refer to the ith and jth harvest, respectively):

$$\hat{\sigma}_{u_{ij}} = \left[\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j + \mathrm{tr}(\mathbf{C}_{ij}^{-1})\right]\Big/ g \qquad [6]$$

$$\hat{\sigma}_{u_{ij}} = \begin{cases}\hat{\sigma}^2_{u_k} & k \text{ if } i = j\\ \hat{\sigma}_{u_{ij}} & \text{otherwise}\end{cases} \qquad [7]$$

Matrix $\mathbf{C}_{ij}^{-1}$ corresponds to submatrices ij of $\mathbf{C}_{22}$, which is contained in matrix $\mathbf{C}^{-1}$ in Eq. [5]. The residual covariance estimators contained in $\mathbf{R}$ are given as follows:

$$\hat{\sigma}_{e_{ij}} = \left[\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_j + \mathrm{tr}\left(\left[\mathbf{K}\mathbf{C}^{-1}\mathbf{K}^{\mathrm{T}}\right]_{ij}\right)\right]\Big/ n^* \qquad [8]$$

$$\hat{\sigma}_{e_{ij}} = \begin{cases}\hat{\sigma}^2_{e_k} & k \text{ if } i = j\\ \hat{\sigma}_{e_{ij}} & \text{otherwise}\end{cases} \qquad [9]$$

where $\mathbf{K} = \{\mathbf{X}_k, \mathbf{Z}_k\}$, the trace depends on the i and j submatrices of $[\mathbf{K}\mathbf{C}^{-1}\mathbf{K}]$, and $n^\star$ is the length of vector $\{ij\}$ related to the kth harvest.

Gaussian Mixture Mixed Models for Biennial Growth

The individual and repeated measures analyses were performed using a mixture mixed model that includes the biennial effect as a latent effect as described below.

Individual Analyses

Since the physiological state is unknown, it must be modeled using a latent variable s that indicates the production phase of the plant. Classical mixture models using the observed likelihood might address these issues. The justification for using further modeling for biennial effects is presented in the supplemental material. Taking the latent variable as missing information, the following linear model was used for each harvest analysis:

$$\mathbf{y} = \mathbf{J}\boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \qquad [10]$$

where $\mathbf{y}_{(n\times 1)}$ is the phenotypic observation vector, $\boldsymbol{\mu}_{(2\times 1)}$ and $\boldsymbol{\beta}_{(b\times 1)}$ are the fixed effect vectors (general mean related to biennial status and block, respectively). It was imposed side conditions on vector $\boldsymbol{\beta}$ to ensure that $\beta_1 + \beta_2 + \ldots + \beta_b = 0$, as suggested by (Rencher and Schaalje, 2008), $\mathbf{u}_{(g\times 1)}$ is the random

effect (genotype) vector, and $\mathbf{e}_{(n\times 1)}$ is the error vector. $\mathbf{J}_{(n\times 2)}$ is the missing Bernoulli random variable related to the biennial status. Each element of this $\mathbf{J}$ matrix (p_{il}) in this work will be replaced by its expectation (i.e., the probability that the ith observation has been taken from the lth biennial state; Supplemental File S2). $\mathbf{X}_{(n\times b)}$ is the fixed matrix for block, and $\mathbf{X}_{(n\times g)}$ is the random effect incidence matrix (genotypes). Because the matrix for the 2-yr means ($\boldsymbol{\mu}$) is unknown, the expected Bernoulli variable was used as an indicator of the genotype stage in the mixture.

As showed in the Supplemental File S2, since the latent Bernoulli random variable is unknown, it is replaced by its expected value of the complete data likelihood. In other words, $\mathbf{s}_{n\times 1} \sim$ Bernoulli(p_i); therefore, $E\ (s_i = 1) = p_i$, where p_i is the ith element of $\mathbf{J}$ and represents the expectation of an individual assuming any state in the mixture.

The following assumptions were made for random vectors:

$$\mathbf{u} \sim N(\mathbf{0}, \sigma_g^2\mathbf{I})$$

$$\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2\mathbf{I})$$

Given the above assumptions, the observed data likelihood can be given by

$$\begin{aligned}\mathbf{y} \mid \pi, \boldsymbol{\mu}, \mathbf{u}, \boldsymbol{\beta}, \sigma_e^2 \sim\ & \pi N(\boldsymbol{\mu}_1 + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \sigma_e^2\mathbf{I}) \\ & +(1-\pi)N(\boldsymbol{\mu}_2 + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \sigma_e^2\mathbf{I})\end{aligned} \quad [11]$$

However, in an expectation–maximization (EM) algorithm, the observed data (y), and the missing information ($\mathbf{u}$, s) must be jointly modeled using the expectation of the complete log-likelihood (Sorensen and Gianola, 2007), whose objective function is given by

$$\begin{aligned}&E_{\mathbf{u},s|y,\sigma_e^2}[\mathbf{y}, s \mid \pi, \boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{u}, \sigma_e^2] = \\ &\sum_{l=i}^{2} E_{\mathbf{u}|y,\sigma_e^2}\left[\log p(\mathbf{y}|\mathbf{j}_l\boldsymbol{\mu}_l + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \sigma_e^2\mathbf{I}) + I(s_l=1)\log \pi_l\right] \quad [12] \\ &+E_{\mathbf{u}|y,\sigma_e^2} \log p(\mathbf{u} \mid 0, \sigma_g^2\mathbf{I})\end{aligned}$$

where $E_{\mathbf{u}|y,\sigma_e^2}$ is the expectation in relation to the random effect of genotypes, σ_g^2 and σ_e^2 are the genetic and residual variance, respectively, π is the unknown mixture parameter, μ_1 and μ_2 are scalars representing the means related to Physiological State 1 and 2, and $\mathbf{J}$ is the expectation of the $\mathbf{s}_{n\times 1}$ indicator binary vector relating each mean to its subpopulation. In the model described above, π is the a priori probability of genotypes being in the high- or low-production stage, which was assumed to be unknown in the present study. For REML estimates of variance components, the expectation must be taken in relation to random and fixed effects as showed in Supplemental File S2.

Mixture Mixed Model with Repeated Measures

Joint analyses considering all the years evaluated were performed using a mixture mixed model with unstructured matrices, following the same conventional mixed model structure (i.e., considering unstructured residual and genetic covariance matrices). The mixture mixed model with repeated measures using the complete data log-likelihood can be described as follows:

$$\mathbf{y}_k = \mathbf{J}_k\boldsymbol{\mu}_k + \mathbf{X}_k\boldsymbol{\beta}_k + \mathbf{Z}_k\mathbf{u}_k + \mathbf{e}_k \quad [13]$$

where $\mathbf{y}_{k_{(n\times 1)}}$ is the vector for all the observations obtained during the 8 yr of evaluation, $\boldsymbol{\mu}_{k_{(2\times 1)}}$ and $\boldsymbol{\beta}_{k_{(b\times 1)}}$ are the fixed effect vectors (general mean related to biennial status and block, respectively), $\mathbf{u}_{k_{(g\times 1)}}$ is the random effect (genotype) vector, and $\mathbf{e}_{k_{(g\times 1)}}$ is the error vector. $\mathbf{J}_{k(kn\times k2)}$ is the unknown block diagonal posterior matrix related to the biennial status, $\mathbf{X}_{k(n\times b)}$ is the fixed block diagonal matrix for block, and $\mathbf{Z}_{k(kn\times k2)}$ is the random effect (genotype) block diagonal incidence matrix. The following distribution assumptions were made for the random effects:

$$\mathbf{u}_k \sim N(0, \mathbf{G} = \mathbf{A} \otimes \mathbf{I})$$

$$\mathbf{e}_k \sim N(0, \mathbf{R} \otimes \mathbf{I})$$

Given the above assumptions the expectation of the complete (data) log-likelihood (objective function) can be given by (to simplify the representation, it was omitted the subscript k)

$$\begin{aligned}&E_{\mathbf{u},s|y,\sigma_e^2}[\mathbf{y} \mid \pi, \boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{R}] = \\ &\sum_{l=i}^{2} E_{\mathbf{u}|y,\sigma_e^2}\left[\log p(\mathbf{y}|\mathbf{j}_l\boldsymbol{\mu}_l + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R}\otimes\mathbf{I}) + I(s_l=1)\log \pi_l\right] \quad [14] \\ &+E_{\mathbf{u}|y,\sigma_e^2} \log p(\mathbf{u} \mid 0, \mathbf{G} = \mathbf{A} \otimes \mathbf{I})\end{aligned}$$

where $\mathbf{A}$ is the genetic unstructured covariance matrix, and $\mathbf{R}$ is the unstructured residual covariance matrix. We now have vector $\boldsymbol{\pi}$, which was described in the scalar form in the per year analyses, and $\mathbf{j}$ is a vector as described in the individual model.

These equations are better visualized as shown below:

$$\begin{aligned}\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \vdots\\ \mathbf{y}_k\end{bmatrix} =& \begin{bmatrix}\mathbf{J}_1 & \mathbf{O} & \cdots & \mathbf{O}\\ \mathbf{O} & \mathbf{J}_2 & & \vdots\\ \vdots & & \ddots & \mathbf{O}\\ \mathbf{O} & \cdots & \mathbf{O} & \mathbf{J}_k\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}_{l1}\\ \boldsymbol{\mu}_{l2}\\ \vdots\\ \boldsymbol{\mu}_{k2}\end{bmatrix} \\ &+\begin{bmatrix}\mathbf{X}_1 & \mathbf{O} & \cdots & \mathbf{O}\\ \mathbf{O} & \mathbf{X}_2 & & \vdots\\ \vdots & & \ddots & \mathbf{O}\\ \mathbf{O} & \cdots & \mathbf{O} & \mathbf{X}_k\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_1\\ \vdots\\ \boldsymbol{\beta}_k\end{bmatrix} \\ &+\begin{bmatrix}\mathbf{Z}_1 & \mathbf{O} & \cdots & \mathbf{O}\\ \mathbf{O} & \mathbf{Z}_2 & & \vdots\\ \vdots & & \ddots & \mathbf{O}\\ \mathbf{O} & \cdots & \mathbf{O} & \mathbf{Z}_k\end{bmatrix}\begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_1\\ \vdots\\ \mathbf{u}_k\end{bmatrix}+\begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_1\\ \vdots\\ \mathbf{e}_k\end{bmatrix}\end{aligned} \quad [15]$$

where the subscript is a vector or matrix representing the kth year evaluated, and μ_{kl} represents the lth mixture (l = 1, 2) for the kth harvest. The elements of weight matrix $\mathbf{P}$ are the probabilities of each observation belonging to Population (biennial status) 1 or 2, as described for the individual analyses and in the Supplemental File S2.

From the matrix system of equations for the mixed model, solutions for $\boldsymbol{\mu}_k$, $\boldsymbol{\beta}_k$, and $\mathbf{u}_k$ can be obtained as follows:

$$\begin{bmatrix}\hat{\boldsymbol{\mu}}_k\\ \hat{\boldsymbol{\beta}}_k\\ \hat{\mathbf{u}}_k\end{bmatrix} = \begin{bmatrix}\mathbf{P}_k'\mathbf{V}^{-1}\mathbf{P}_k & \mathbf{P}_k'\mathbf{V}^{-1}\mathbf{X}_k & \mathbf{P}_k'\mathbf{V}^{-1}\mathbf{Z}_k\\ \mathbf{X}_k'\mathbf{V}^{-1}\mathbf{P}_k & \mathbf{X}_k'\mathbf{V}^{-1}\mathbf{X}_k & \mathbf{X}_k'\mathbf{V}^{-1}\mathbf{Z}_k\\ \mathbf{Z}_k'\mathbf{V}^{-1}\mathbf{P}_k & \mathbf{Z}_k'\mathbf{V}^{-1}\mathbf{X}_k & \mathbf{Z}_k'\mathbf{V}^{-1}\mathbf{Z}_k + \mathbf{G}^{-1}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{P}_k'\mathbf{V}^{-1}\mathbf{y}\\ \mathbf{X}_k'\mathbf{V}^{-1}\mathbf{y}\\ \mathbf{Z}_k'\mathbf{V}^{-1}\mathbf{y}\end{bmatrix} \quad [16]$$

where $\mathbf{V} = \mathbf{R} \otimes \mathbf{I}$.

From this system, we describe the **C** matrix as

$$\mathbf{C} = \begin{bmatrix} \mathbf{P}_k'\mathbf{V}^{-1}\mathbf{P}_k & \mathbf{P}_k'\mathbf{V}^{-1}\mathbf{X}_k & \mathbf{P}_k'\mathbf{V}^{-1}\mathbf{Z}_k \\ \mathbf{X}_k'\mathbf{V}^{-1}\mathbf{P}_k & \mathbf{X}_k'\mathbf{V}^{-1}\mathbf{X}_k & \mathbf{X}_k'\mathbf{V}^{-1}\mathbf{Z}_k \\ \mathbf{Z}_k'\mathbf{V}^{-1}\mathbf{P}_k & \mathbf{Z}_k'\mathbf{V}^{-1}\mathbf{X}_k & \mathbf{Z}_k'\mathbf{V}^{-1}\mathbf{Z}_k + \mathbf{G}^{-1} \end{bmatrix} \quad [17]$$
$$= \begin{bmatrix} \mathbf{C_{11}} & \mathbf{C_{12}} & \mathbf{C_{13}} \\ \mathbf{C_{21}} & \mathbf{C_{22}} & \mathbf{C_{13}} \\ \mathbf{C_{31}} & \mathbf{C_{32}} & \mathbf{C_{33}} \end{bmatrix}$$

The EM algorithm was used to maximize the restricted likelihood (REML) function, where the random **u** and latent Bernoulli variable **j** were considered as missing information, and their expected values were used.

The Expectation–Maximization Algorithm

Here, we present the EM steps for the mixture mixed model with repeated measures. The others analysis can be derived starting from that point. The EM steps were as follow:

E step (Expectation):

1. Given an initial guess of $\pi = 0.5$, the probability of γ_{ik} clustering in the *l*th group was estimated using the expectation of *s* given by

$$E(\mathbf{s}_{i1k} = 1) = \Upsilon_{ilk} = \frac{\pi P_{i1k}}{\pi P_{i1k} + (1-\pi)P_{i2k}} \quad [18]$$

$$E(\mathbf{s}_{i2k} = 1) \equiv E(\mathbf{s}_{i1k} = 0) = \Upsilon_{i2k} = \frac{\pi P_{i2k}}{\pi P_{i1k} + (1-\pi)P_{i2k}} \quad [19]$$

where $P_{i1k} = p(\gamma_{ik} \mid \boldsymbol{\mu}_{k1} + \mathbf{X}_k\boldsymbol{\beta}_k + \mathbf{Z}_k\mathbf{u}_k,\ \mathbf{R} \otimes \mathbf{I}_1)$, $P_{i2k} = p(\gamma_{ik} \mid \boldsymbol{\mu}_{k2} + \mathbf{X}_k\boldsymbol{\beta}_k + \mathbf{Z}_k\mathbf{u}_k,\ \mathbf{R} \otimes \mathbf{I}_1)$, μ_1 is the mean for the *l*th mixture of observation *i* and year k, and $\theta = \{\boldsymbol{\beta}, \mathbf{u}, \mathbf{V}\}$ is the invariable vector across the mixture components.

M step (Maximization):

2. The joint maximization of fixed effects and the expectation of random effects can be obtained by solving the mixed model equations related to restricted log-likelihood given in Eq. [16] as follows:

$$\begin{bmatrix} \hat{\boldsymbol{\mu}}_k \\ \hat{\boldsymbol{\beta}}_k \\ \hat{\mathbf{u}}_k \end{bmatrix} = \begin{bmatrix} \mathbf{P}_k'\mathbf{V}^{-1}\mathbf{P}_k & \mathbf{P}_k'\mathbf{V}^{-1}\mathbf{X}_k & \mathbf{P}_k'\mathbf{V}^{-1}\mathbf{Z}_k \\ \mathbf{X}_k'\mathbf{V}^{-1}\mathbf{P}_k & \mathbf{X}_k'\mathbf{V}^{-1}\mathbf{X}_k & \mathbf{X}_k'\mathbf{V}^{-1}\mathbf{Z}_k \\ \mathbf{Z}_k'\mathbf{V}^{-1}\mathbf{P}_k & \mathbf{Z}_k'\mathbf{V}^{-1}\mathbf{X}_k & \mathbf{Z}_k'\mathbf{V}^{-1}\mathbf{Z}_k + \mathbf{G}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{P}_k'\mathbf{V}^{-1}\mathbf{y} \\ \mathbf{X}_k'\mathbf{V}^{-1}\mathbf{y} \\ \mathbf{Z}_k'\mathbf{V}^{-1}\mathbf{y} \end{bmatrix} [20]$$

3. Obtaining the maximization REML scale parameters for matrix **G** and **R** as follows:

$$\hat{\sigma}_{g_{ij}} = \left[\hat{\mathbf{u}}_i\hat{\mathbf{u}}_j + \mathrm{tr}(\mathbf{C}_{ij}^{-1})\right] / t \quad [21]$$

$$\hat{\sigma}_{g_{ij}} = \begin{cases} \hat{\sigma}^2_{g_k} & k \text{ if } i = j \\ \hat{\sigma}_{g_{ij}} & \text{otherwise} \end{cases} \quad [22]$$

Matrix $\mathbf{C}_{ij}^{-1}$ corresponds to submatrices *ij* of $\mathbf{C}_{33}$, which is contained in the inverse matrix on Eq. [17]. The residual covariance estimators contained in **R** are given as follows:

$$\hat{\sigma}_{e_{ij}} = \left\{\hat{\mathbf{e}}_i'\hat{\mathbf{e}}_j + \mathrm{tr}\left[\left(\mathbf{K}\mathbf{C}^{-1}\mathbf{K}'\right)_{ij}\right]\right\} / n^* \quad [23]$$

$$\hat{\sigma}_{e_{ij}} = \begin{cases} \hat{\sigma}^2_{e_k} & k \text{ if } i = j \\ \hat{\sigma}_{e_{ij}} & \text{otherwise} \end{cases} \quad [24]$$

where **e** is the residual in the *k*th harvest, $\mathbf{K} = \{\mathbf{X}_k, \mathbf{Z}_k\}$ and the trace depends on the submatrices *i* and *j*. **C** is given in Eq. [17], and the subscripts *i* and *j* depend on submatrices related to the *k*th harvest, and $n^\star$ is the length of vector $\{ij\}$.

4. Estimating the probability of mixture π deriving the objective function as follows:

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n_k} P_{i1k}}{\sum_{i=1}^{n_k} P_{i1k} + P_{i2k}} = \frac{\sum_{i=1}^{n_k} P_{i1l}}{n_k}$$

where π_k is the mixture parameter for the *l*th physiological state in the *k*th year, and n_k is the sample size in the *k*th year. Considering two states (vegetative and productive), only $\hat{\pi}_{k1}$ must be estimated given that $\hat{\pi}_{k2} = \left(1 - \hat{\pi}_{k1}\right)$.

The estimated BLUPs (EBLUPs) were obtained for the genotypic values of each model. Here, "estimated" means that the variance components and latent variable are replaced by their estimations and expectation, respectively. The marginal EBLUPs were considered for the repeated measures models according to (Smith et al., 2007).

Model Selection

To test model fit and the effects of increasing the number of parameters in the mixture model, the Bayesian information criterion (BIC) (Schwarz, 1978) was calculated for the per harvest and repeated measures analyses via the following equation:

$$\mathrm{BIC} = -2\mathrm{Llik} + q[\log(n)]$$

where *q* is the number of free parameters in the model, and *n* is the observation length.

An approximated prediction error variance (APEV) of genotypic values was also estimated for all the analyses using the diagonal of matrix $\mathbf{C}_{22}^{-1}\sigma^2_e$ for the individual analyses and that of matrix $\mathbf{C}_{33}^{-1}$ for the joint analysis. The genetic variance error for the individual analysis was calculated by the negative of the inverse of the expected Fisher information matrix. The observed likelihood and one step in the Fisher scoring algorithm from the EM convergence values were used for these estimates.

Additionally, the CV (%) was estimated for all the models as follows:

$$\text{Mixed models: } \mathrm{CV}_j(\%) = \hat{\sigma}_e / \hat{\mu}_j$$

Mixture mixed models: $CV_j(\%) = \hat{\sigma}_e \left[\hat{\pi}/\hat{\mu}_{1j} + (1-\hat{\pi})/\hat{\mu}_{2j} \right]$

where $\hat{\mu}_j$ is the estimated overall mean for the jth harvest, $\hat{\imath}_{1j}$ and $\hat{\imath}_{2j}$ are the harvest "j" means for Subpopulations 1 and 2, respectively, $\hat{\pi}$ is the mixture parameter, and $\hat{\sigma}_e$ is the residual standard deviation.

For the mixture models, the bimodality was tested according to (Holzmann and Vollmer, 2008) using the following steps:

1. Estimating the distance between the two means through the following calculus:

$$d = \frac{\mu_1 - \mu_2}{2(\sigma_1 \sigma_2)^{0.5}}$$

where σ_1 and σ_2 are the standard deviation of the first and second mixture (in this study, $\sigma_1 = \sigma_2$);

2. The distribution is unimodal if and only if $d \leq 1$, or if $d > 1$ and $|\log(\pi) - \log(1-\pi)| \geq 2\log(d - \sqrt{d^2-1}) + 2d\sqrt{d^2-1}$. Otherwise, there is evidence of bimodality.

To evaluate mixture model usage for the selection of coffee cultivars, the gain from selection (GS) was estimated considering models M3 and M4, according to the estimator

$$GS = \frac{\sum_{i=1}^{n} \hat{g}_i / n}{\bar{Y}_{..}}$$

where $\hat{g}_i$ is the marginal EBLUP of the ith progeny, and $\bar{Y}_{..}$ the phenotypic mean over all harvests. The top five best progenies were considered to apply a selection intensity of 23%.

In the repeated measures analysis, the multivariate heritability and maximum heritability were estimated considering the largest eigenvalue of the (co)heritability **H** matrix (heritability limit considering all the harvests) (Klingenberg and Leamy, 2001; Balestre et al., 2013). In multiple trait selection, the estimated heritability is the maximum heritability of a linear combination of traits and is given by the largest eigenvalue of **H**. In the present study, this value can be interpreted as the linear combination of harvests that would generate the highest possible heritability in the joint analysis. Matrix **H** is given as follows:

$$\mathbf{H} = \mathbf{G}\mathbf{F}^{-1}$$

$$\mathbf{F} = \mathbf{G} + \mathbf{R}$$

where **G**, **R**, and **F** are the genetic, residual, and phenotypic unstructured covariance matrices, respectively.

Simulation Study

In the simulation study, we evaluated the model ability to estimate the mixture parameters. For this, the models M1 and M2 were compared in four scenarios, varying the mixture proportion and mean of each subpopulation. Each scenario was run 1000 times, and the parameters estimate for each model was taken as an average over all runs.

In all scenarios, 100 genotypes were evaluated in a randomized complete block design with two replications. The genetic effects were independently sampled from a Gaussian distribution with mean 0 and variance σ^2_g [i.e., $\mathbf{g} \sim N(0, \mathbf{I}\sigma^2_g)$]. The phenotypic values were simulated as $\mathbf{y} = \mathbf{J}\boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{e}$, where $\boldsymbol{\mu}$ is the vector of means of each subpopulation indicating the physiological state (whose differences varied according to the scenario), **J** is the matrix indicating the physiological state, $\boldsymbol{\beta}$ is the vector of block effects, **g** is the vector genetic values as described above, and e represents the residual effects vector. The residual effects values were sampled from a Gaussian distribution [i.e., $\mathbf{e} \sim N(0, \mathbf{I}\sigma^2_e)$], where $\sigma^2_e = \left[(1-h^2)/h^2 \right] \sigma^2_g$, and h^2 and σ^2_g are the univariate heritability and the genetic variance, respectively.

RESULTS

In this study, alternations between low and high production during the harvests, a feature of biennial growth, were observed (Fig. 1). The estimated variance components varied widely throughout the years (Fig. 2). However, the estimated residual variance and its variation were drastically lower for the mixture models (Fig. 2). For example, the variation range (difference between highest and lowest estimate) of the estimated residual variances was 277.16 for the unstructured mixed model (M3) and 90.3 for the unstructured mixture mixed model (M4) (i.e., the range was 3.06 times lower for model M4 than for M3). A small increase in genetic variance was also observed from the mixture model with repeated measures, which may indicate an increase in the estimated heritability.

The estimated CV showed the same pattern as that for the residual variance (Fig. 2). The CV was lower for individual mixture mixed model (M2) and M4 than for individual mixed model (M1) and M3 and was slightly lower for M4 than for M2. The estimated CV presented a pattern contrary to that of the residual variance over the years, i.e., years with a high mean presented a relatively high residual variance but low CV, indicating lower uncertainty and better experimental precision for these years.

The heatmap showed lower residual correlations for the model M4 than for M3 (Fig. 3A). This result may be related to the modeling of biennial patterns with mixture components. For some year pairs, the estimated residual correlations were much lower in M4, such as for year pairs (2, 3), (3, 4), (4, 8), and (5, 7). This result was confirmed by the fact that the residual covariances between years were always higher for M3 than for M4 (Supplemental Tables S1 and S2). Furthermore, M4 presented less alternation between positive and negative covariance (e.g., the alternation patterns were very different for M3 and M4 in Year 3).

No differences in the estimated genetic correlations were observed between M3 and M4. However, a tendency for relatively high estimated genetic correlations was observed during the first years for M4, namely, during Harvests 1 and 2. In addition, years with a high mean tended to show a high genetic correlation between them, and the same pattern was verified for years with a low mean. (Fig. 3B). Overall, the estimated genetic variance for the years of high production was higher for M4 than for M3, and the opposite

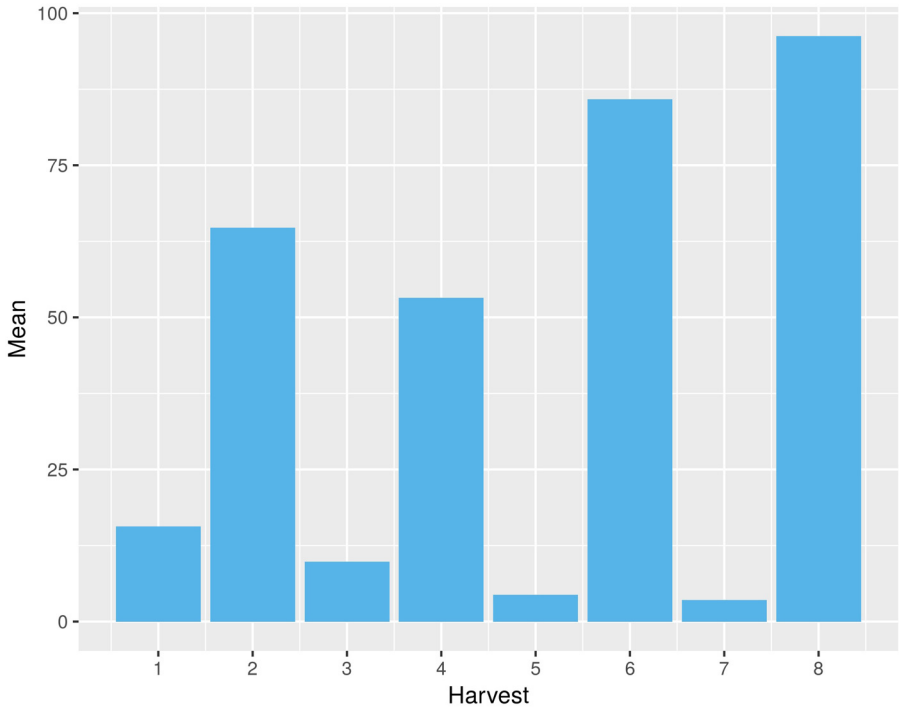

Fig. 1. Mean coffee production (kg) over the evaluated harvests.

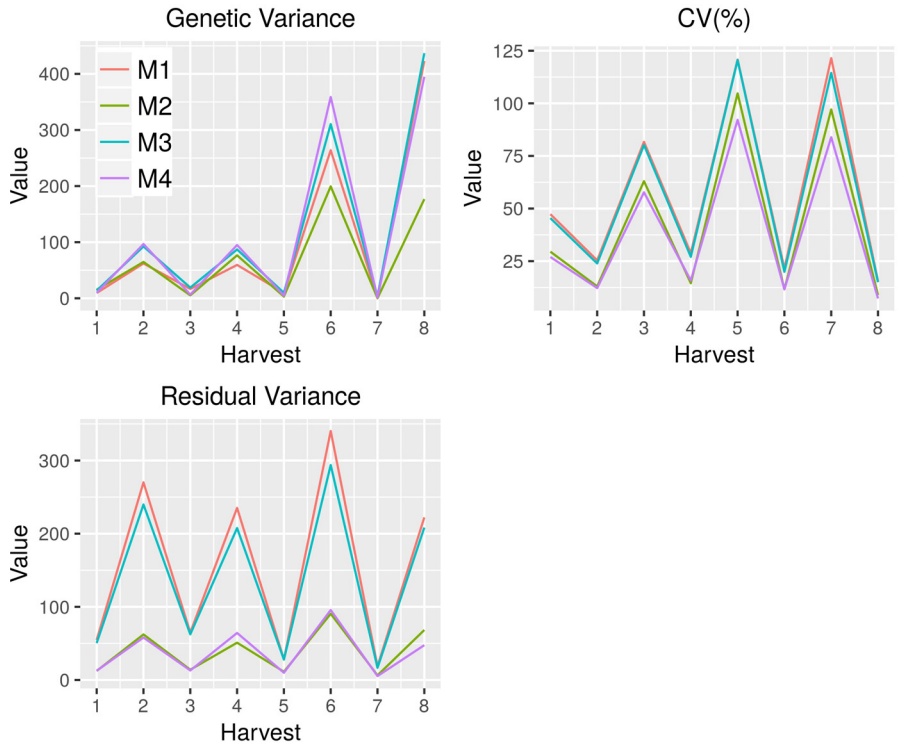

Fig. 2. Estimated genetic variance, CV, and residual variance in individual analyses during each year for the individual mixed model (M1), individual mixture mixed model (M2), repeated measures mixed model (M3), and repeated measures mixture mixed model (M4).

was observed for the years of low production (Supplemental Tables S3 and S4). The genotypic variance was higher for models M3 and M4 than for M1 and M2 (Table 1), and given the drastic decrease in residual variance, the heritability was higher for the mixture models.

The distribution of predicted values of the mixture models for the coffee phenotype data is presented in Fig. 4. These models were better fitted to the data than the traditional mixed models, indicating the existence of different coffee genotypes at different physiological stages. The mixture models were able to capture data overdispersion, predicting even the most extreme values. The prediction curves were "slimmer" for M1 and M3 than for M2 and M4, indicating a concentration of predicted

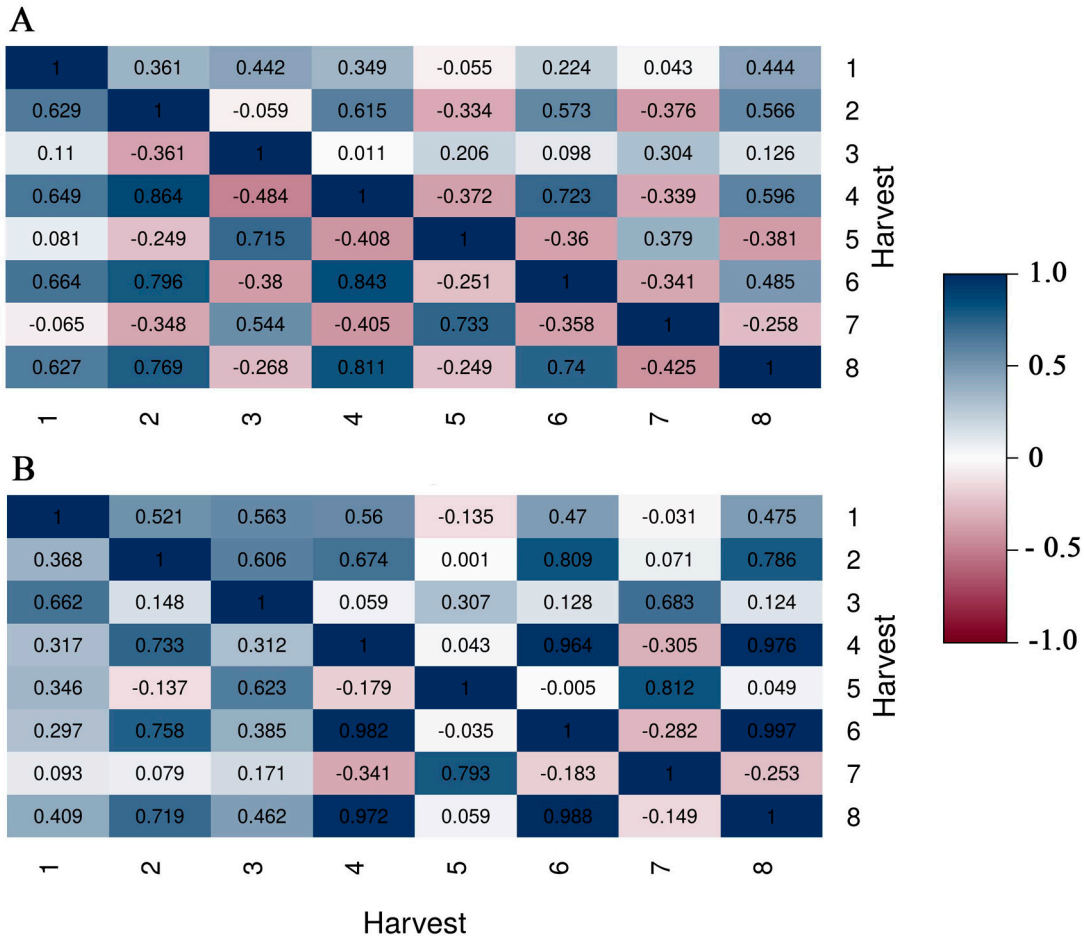

Fig. 3. Heatmap of (A) residual and (B) genetic correlations between harvests for two repeated measures models: the mixed model (below the diagonal) and mixture mixed model (above the diagonal).

values around the overall mean and not effectively accompanying the phenotypic dispersion. As an example, the pattern of maximum and minimum values in the observed value curve was perfectly captured by M2 and M4 during the third year, indicating high predictive efficiency. As expected, the predictions were better for harvests with comparatively high mean production levels (2, 4, 6, and 8), which presented comparatively low CVs.

The APEV of the EBLUP was almost twice as high for M1 than for the remaining models (Fig. 5). The model M3 exhibited drastically decreased APEV. However, model M2 presented a lower prediction error than M3 for most harvests, indicating that biennial growth may be more effectively modeled based on mixtures of means than on (co)variance structures. However, both types of information (latent means plus heterogeneity of variances) in the joint modeling improved the EBLUP estimation, as observed in M4. The APEV value was only slightly lower for M4 than for M2 in Year 8. Because of this lower APEV, the mixture models were better able to detect differences between the EBLUPs.

Despite the evidence of mixture model superiority for all the evaluated scenarios, these models require the estimation of more parameters than the other models, and the effort to increase the number of parameters in the model

Table 1. Estimated genetic variance (GV), standard error associated with genetic variance (SD error), and the *Z* statistic for the individual mixed model (M1) and individual mixture mixed model (M2).

	M1		M2			
Harvest	GV	SD error	GV	SD error	Z1	Z2
1	9.445	9.662	15.384	6.294	1	2.444
2	62.125	52.159	64.843	27.650	1.191	2.345
3	16.730	13.048	5.236	3.352	1.282	1.561
4	59.480	46.994	76.600	30.036	1.265	2.55
5	9.622	6.356	3.111	2.356	1.514	1.32
6	263.870	121.980	199.463	73.377	2.163	2.718
7	1.417	2.796	0.118	0.868	0.507	0.136
8	423.000	158.060	176.779	63.742	2.676	2.773

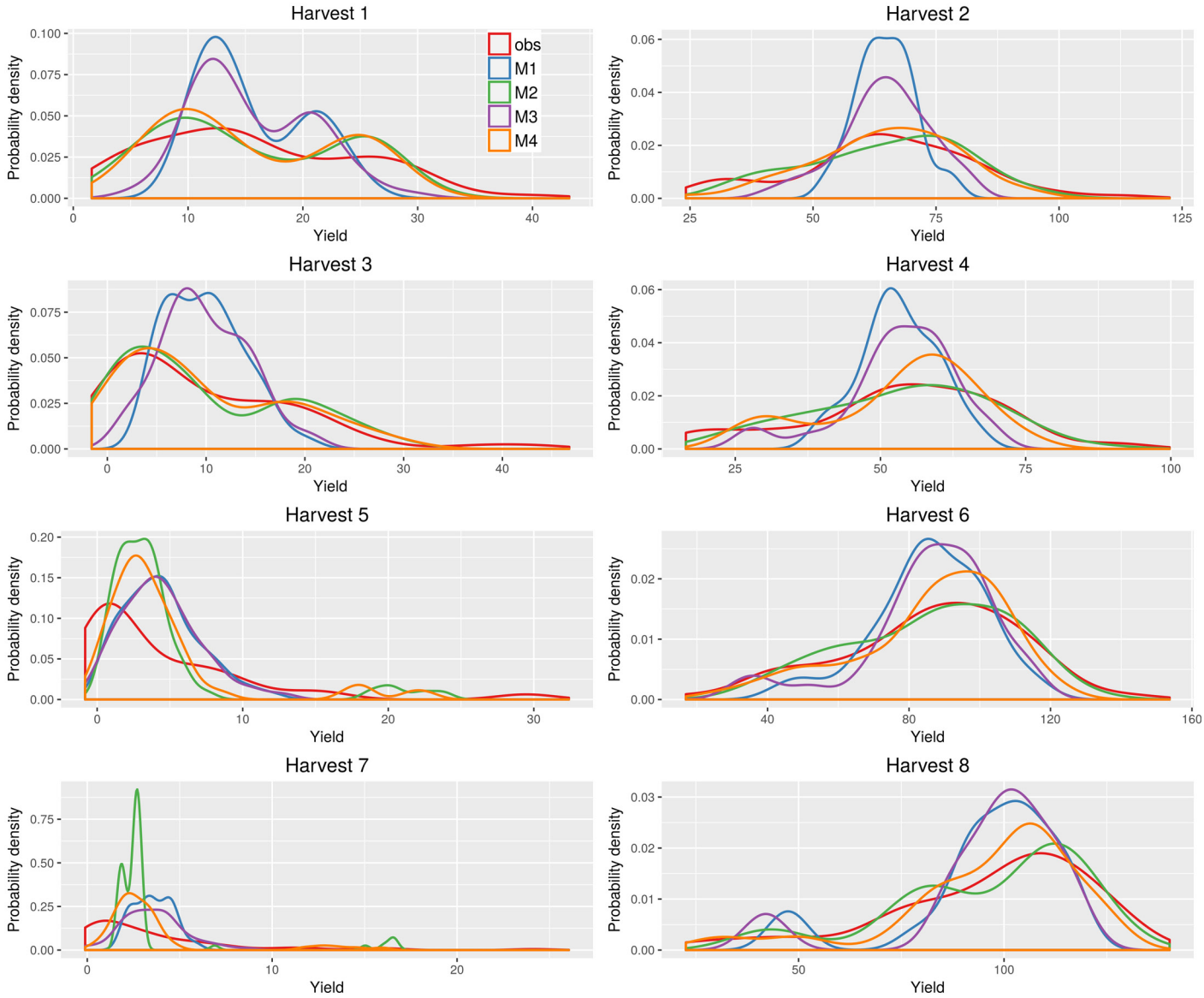

Fig. 4. Probability density for observed values (obs) and values predicted by different models: the individual mixed model (M1), individual mixture mixed model (M2), repeated measures mixed model (M3), and repeated measures mixture mixed model (M4).

may not be compensated by this gain. The BIC among the models was estimated, and the lowest BIC value was adopted as being indicative of the most suitable model.

Compared with the other models, the mixture models presented lower BIC values for all the harvests (Fig. 5). Even with repeated measures analysis, which included an additional 16 parameters in the model (M4), a lower BIC value was observed than in the M3 model, indicating that the mixture models were more informative than the other models in all cases.

The marginal EBLUPs predicted by M3 and M4 are presented in Fig. 6. The differences between the genotypic values were clearer for the mixture model, especially during the high production years. For example, the green and gray lines overlap for M3 but are clearly separated for M4. Furthermore, the genotype corresponding to the yellow line was very close to the *x* axis in the mixed model for all the years but was considerably distant from the axis in the mixture mixed model, which confirms the superiority of mixture mixed models in identifying and separating the best individuals.

Similar to the observed APEV values, the estimated error for the genetic variance component was much smaller for M2 than for M1 (Table 1). However, because the different models presented different genetic variances, directly comparing them to determine which model best estimated the genetic component is difficult. One way to eliminate the scale effect is to use the *Z* statistic, which is an estimate divided by its standard error. Except for Years 5 and 7, *Z* scores were always higher for the mixture model, indicating that the variance component was estimated more accurately in the mixture models (Table 1).

The results of the bimodality test for contrasts M1 to M2 are presented in Table 2. The differences between the

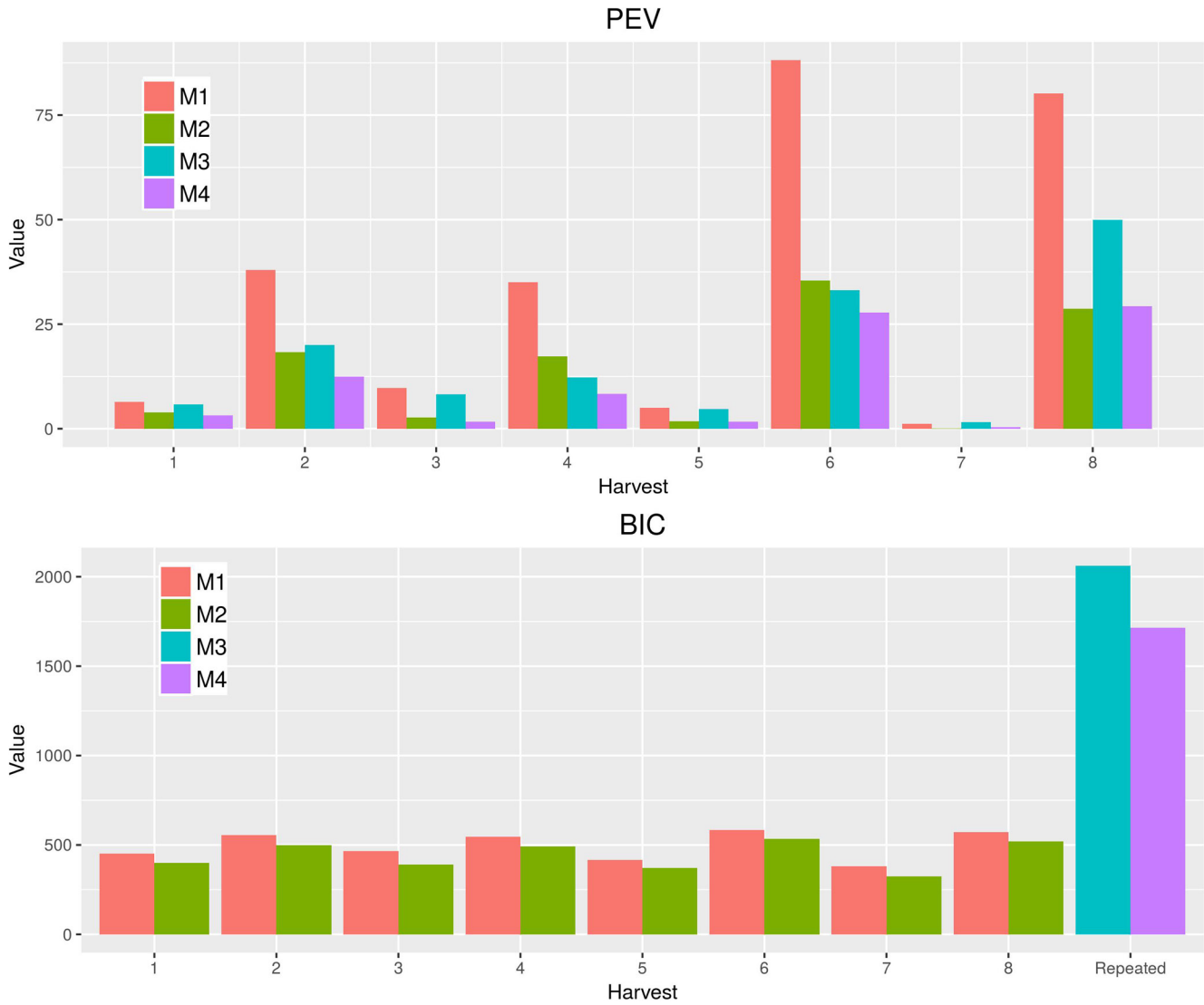

Fig. 5. Bayesian information criterion (BIC) and approximated prediction error variance (PEV), for the individual mixed model (M1), individual mixture mixed model (M2), repeated measures mixed model (M3), and repeated measures mixture mixed model (M4).

mixture means were high for all the harvests, showing strong evidence for a mixture of populations according to (Holzmann and Vollmer, 2008), where the bimodality was verified in all models under different values of π. Another way of verifying the importance of mixtures in models was suggested by Schilling et al. (2002). In their approach, a contrast between means that exceeds twice the residual (equivalent to a test under normality) indicates the presence of two data modes. This criterion was met for all the years evaluated and for both the individual and repeated measures analyses (Table 2). Note that the magnitude of contrasts was lower for M4 than for the individual analysis model (M2).

Graphs describing the probability of each genotype belonging to the high (blue) or low (red) mean production cluster over the harvests are presented in Fig. 7. The formation of subpopulations within each harvest for the different years and a stabilization of the physiological relationship (vegetative and reproductive stage) of the plants were observed. The genotypes indicated by red bars predominate in years of low mean production, and the genotypes indicated by the blue bars predominating in years of high mean production. These changes (as indicated by the different colors) follow a well-defined pattern that is expected to occur in the presence of biennial growth: the genotypes presenting a high probability of belonging to the high-production cluster (blue) in a given year likely belong to the low-production cluster (red) in the following year. These findings confirm that the mixture model was able to capture latent variables, in this case biennial effects.

Progeny Selection

The mixture model was superior regarding the GS because of the high maximum heritability value, indicated by the largest matrix (H^2) eigenvalue. The GS was 1.4% higher for the mixture model than for the M3 model (Table 3). The Spearman correlation coefficient

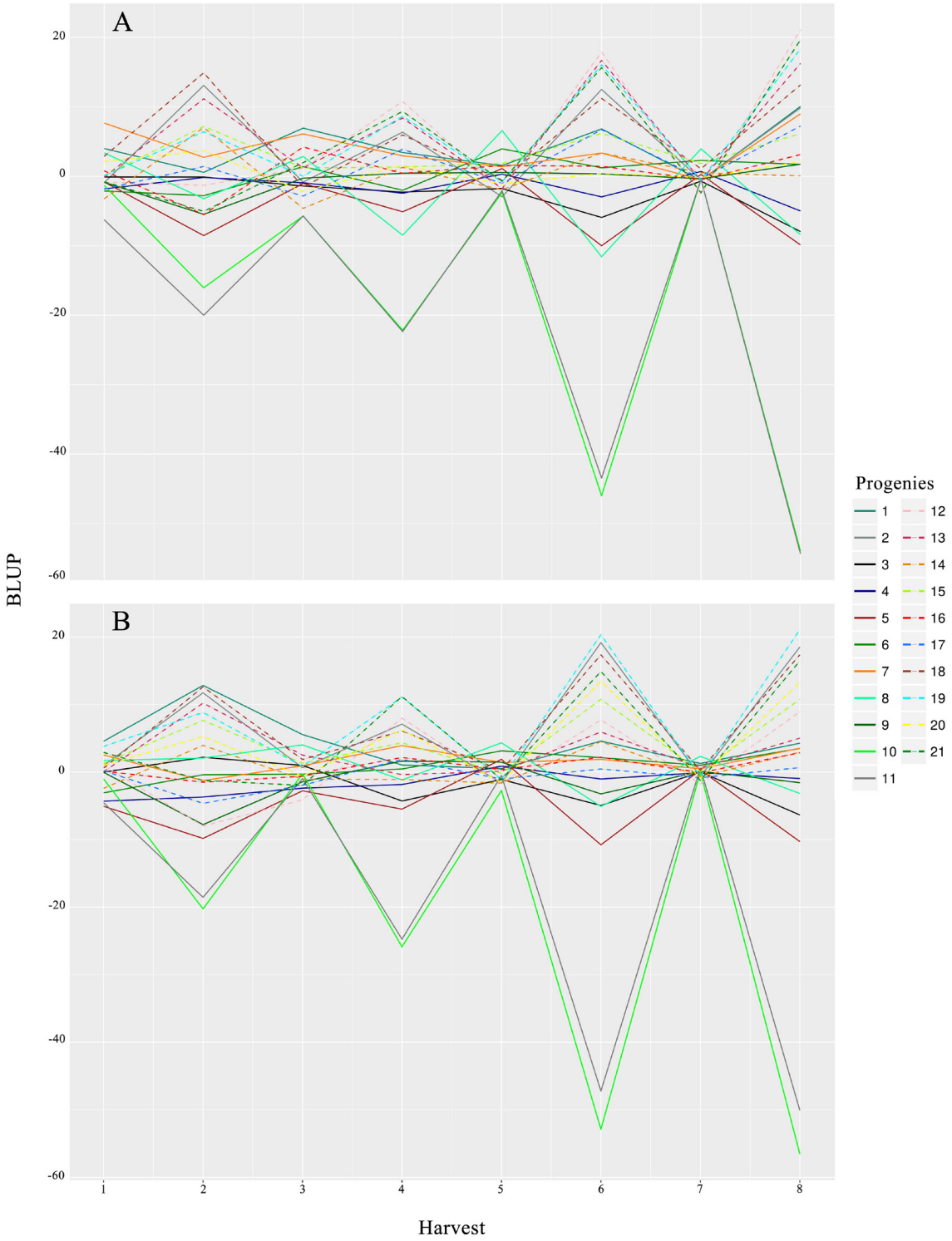

Fig. 6. Marginal best linear unbiased predictions (BLUPs) predicted by models (A) M3 and (B) M4.

between the EBLUPs for both models was 0.87, with some altered ranks caused by the insertion of the mixture component (Table 3). For example, the five best progenies in the mixed model were (in descending order of EBLUP values) Progenies 13, 19, 18, 12, and 21; the best ones in the mixture mixed model were Progenies 19, 18, 2, 21, and 20, with Progenies 20 and 2 discarded by the classical mixed model (without the mixture component).

Simulation

The results of simulation study indicated that EBLUPs were better estimated for model M1 in scenarios where subpopulations had similar means (Scenarios 1 and 2) in other words, when the biennial effects are not evident. In these cases, the heritability and the correlation between the real and estimated BLUPs were greater for this model. However, it is evident that our model presented a better estimation of genetic variance and mixture means. When

Table 2. Estimated means, contrast (difference) between modes, bimodality test, and 2× standard error (2× SD error) statistics for the individual mixture mixed model (M2) and mixture mixed model with repeated measures (M4) for the different evaluated years.

Harvest	Mean 1†	Mean 2†	Difference	*b* test‡	2× SD error
			Univariate		
1	8.33 (0.54)	24.19 (0.46)	15.86	***	7.04
2	44.1 (0.35)	75.82 (0.65)	31.72	***	15.78
3	4.39 (0.67)	20.82 (0.33)	16.43	***	7.50
4	37.25 (0.44)	65.87 (0.56)	28.62	***	14.28
5	2.99 (0.92)	20.76 (0.08)	17.77	***	6.72
6	66.29 (0.47)	103.53 (0.53)	37.24	***	19.02
7	2.41 (0.92)	16.23 (0.08)	13.82	***	5.04
8	76.5 (0.42)	110.69 (0.58)	34.19	***	16.53
			Multivariate		
1	10.09 (0.59)	23.73 (0.41)	13.64	***	7.10
2	48.73 (0.35)	73.31 (0.65)	24.57	***	15.23
3	4.64 (0.67)	20.2 (0.33)	15.56	***	7.22
4	33.71 (0.21)	58.48 (0.79)	24.76	***	16.04
5	3.18 (0.92)	18.2 (0.08)	15.01	***	6.29
6	63.75 (0.22)	91.93 (0.78)	28.17	***	19.56
7	2.53 (0.9)	12.99 (0.10)	10.46	***	4.62
8	82.21 (0.37)	104.59 (0.63)	22.38	***	13.80

† Values in parentheses are the mixture parameters.

‡ Bimodality test. *** Rejected the unimodality.

the subpopulation means were distant or the biennial effect was evident (Scenarios 3 and 4), the model M2 outperformed (Table 4) the classical linear mixed models. In all scenarios, model M2 estimated the genetic variance with greater accuracy.

Changing the mixture proportion (π) did not significantly affect the parameter estimates on Scenarios 1 and 2, independently of the model. For Scenarios 3 and 4 (Table 4), changing it from 0.2 to 0.5 significantly increased the accuracy on parameter estimates by model M2 and decreased the estimates by M1. These results suggest that the difference in subpopulation means has a stronger effect than mixture proportion on parameter estimates, which is in accordance with previous studies (Redner and Walker, 1984; Detilleux and Leroy, 2000).

DISCUSSION

Although modeling individual biennial effects seems difficult, since this information is not available to breeders, the mixture model allows a simple and intuitive way to handle the effects of biennial growth, and these models present some advantages over traditional models. For example, these models allow for computing the posterior probability of a given genotype being in a different physiological stage, which avoids some of the bias present in currently used models where all genotypes are present in identical stages and represent a common mean. Mixture models, on the other hand, assume that both stages can occur in the same year and using probabilistic criteria, allocate each genotype into different groups.

Because the biennial growth pattern modifies the (co)variance pattern of coffee production data, some authors have proposed modeling these data by changing the covariance structure (Andrade et al., 2016). However, this approach does not seem to be the best for directly modeling the differential biennial effects, since it also assumes equal stages for the same year. In addition, it was observed that individual and repeated measures mixed models presented similar precision in parameter estimation. This finding was confirmed by the APEV values and predictive ability related to individual mixture models (per harvest analyses) compared with those of the repeated measures mixed models; even the simplest mixture model generated better estimates than the classical mixed models with complex covariance structures. These results are in disagreement with the suggestions of Andrade et al. (2016). The authors propose modeling biennial patterns using different covariance structure matrices.

According to the hypothesis that there are two different means in coffee progeny tests, and their existence originates from the presence of coffee genotypes at different physiological stages, it is expected that there will be a significant contrast between the two means. Although the Holzmann and Vollmer (2008) test is very attractive since it is based on likelihood ratio, there is currently no consensus on the detection of data bimodality (or the number of different distributions in the mixture), and the use of more than one test is usually recommended. Therefore, the criterion recommended by Schilling et al. (2002) was also applied to verify the presence of bimodality. These statistics agreed in all cases, indicating that the modeling of productivity data

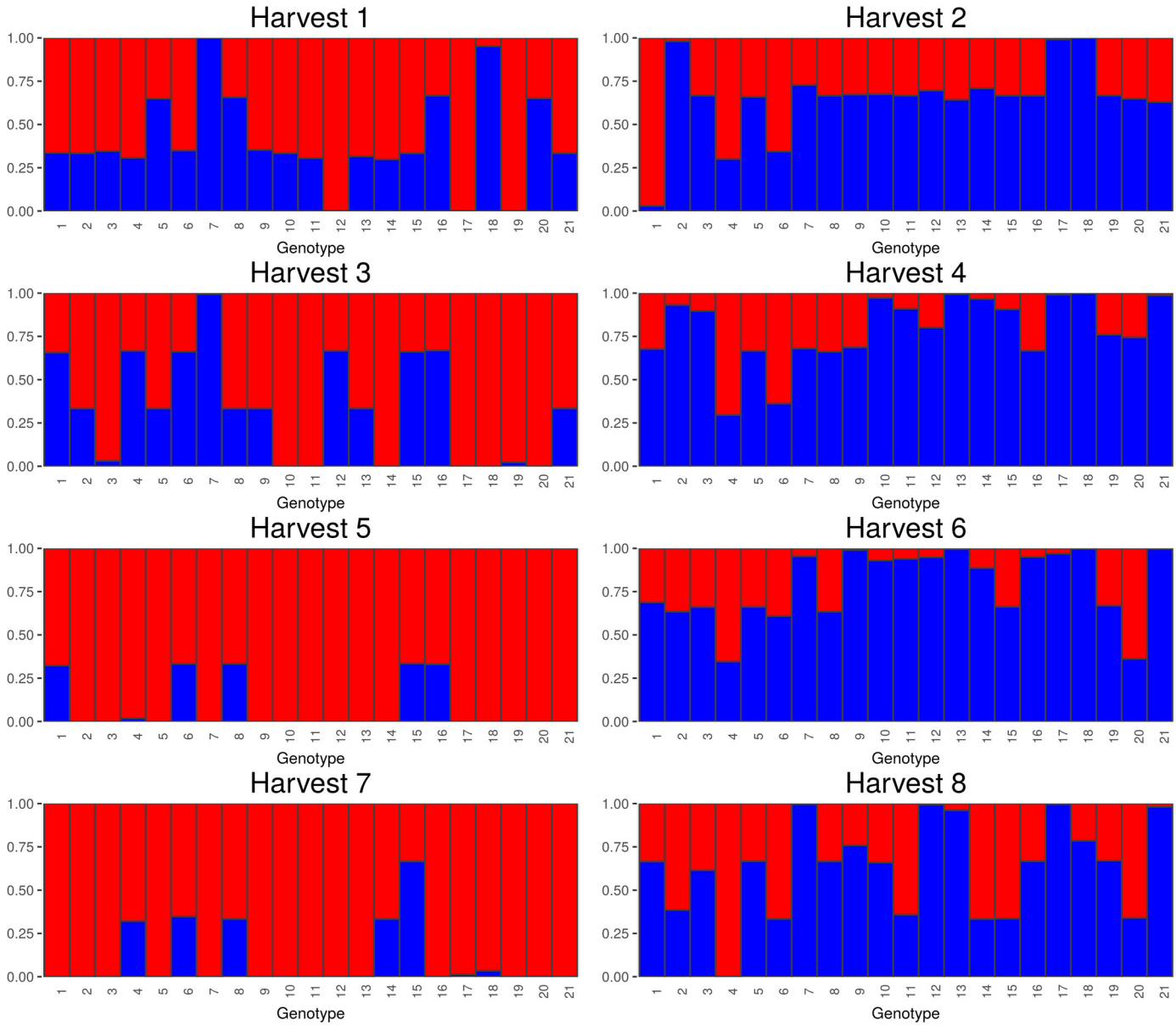

Fig. 7. Probability for each year of the evaluated genotypes being in the year of low (red) or high (blue) production. Values plotted for each genotype were obtained considering the mean probability for the three replicates and estimated using the repeated measures mixture mixed model (M4).

for different coffee progenies is improved by the use of a mixture model with two components that correspond to the physiological stages of the plants.

The BIC values also indicated the importance of including biennial effects in the model (BIC values were always lower). Compared with the other models tested and despite the penalization in the likelihood values due to the higher number of parameters, the mixture models better fit the data.

Correlations between residuals for the different years also indicate the advantage of mixture models (Fig. 3A); thus, ignoring latent variables is especially problematic for longitudinal data because depending on the matrix structure (**R** and **G**), the estimated covariance will directly affect the random effect estimation (Andrade et al., 2016). This effect has been previously observed for mixture models with data presenting latent variables (Detilleux and Leroy, 2000; Jamrozik and Schaeffer, 2010).

Variance is known to be directly affected by the scale of the data. Therefore, comparing residual variance between high- and low-production years is inadequate. In this case, the use of the CV is preferable, since it is a free dimensional-scale estimate (Resende and Duarte, 2007). The CV estimates indicate that the experimental precision of analyzing coffee crops using classical models can be underestimated even when using appropriate experimental designs and good crop management, since the production stage is neglected. As mixture models estimate distinct means for each subpopulation, the CV can be computed without bias, which is another advantage of these models.

Table 3. Estimated gain from selection (GS) for the five best progenies relative to the original mean, potential heritability (H^2), and Spearman correlation between predicted estimated best linear unbiased predictions (EBLUPs) for the mixed model with repeated measures (M3) and mixture mixed model with repeated measures (M4).

Model	GS	H^2	Spearman correlation
	%		
M3	13.710	86.196	0.875
M4	15.114	93.551	

Differences in genetic variance were expected specifically because of the distinct assumptions underlying each model. Despite the fact that genetic variance was considered constant for both subpopulations, the M4 model increased the estimates of this parameter, and the covariances certainly had some influence on these estimates, which explains the lower genetic variance estimates for model M2 in some years.

The present study assessed the different behaviors of progenies within a single experiment. The progenies at different physiological stages within the same experimental field can therefore be differentiated (Fig. 6 and 7; i.e., a progeny group undergoing high production may be distinguished within a specific year from another group undergoing low production). This shows that there is no constant biennial pattern throughout the years for coffee progenies (i.e., the magnitude of biennial effects varies widely among progenies, requiring that the physiological stage of each progeny be modeled separately). This result suggests that grouping data in a biennium is not adequate to model biennial effect since environmental conditions can induce some progenies to repeat a physiological stage for >1 yr. For example, Progeny 7 showed constant production until the fourth evaluation (it was clustered in the group with a high mean physiological state), which violates the assumptions of traditional models about this phenomenon. Only after the fourth harvest was the biennial alternation pattern shown with consistency. This explains why mixture models were superior to the use of grouped data. In addition, grouping means presents a loss of information in which the parameter precision is underestimated and the APEV decreases (Andrade et al., 2016). For some genotypes, an inversion of the EBLUP signal occurred in consecutive years (Fig. 6). For example, the EBLUPs were positive for some genotypes and negative for others in some harvests, and this pattern was reversed in other harvests. This inversion in classification demonstrates the need for classifying the genotypes within the same year and supports the hypothesis that two

Table 4. Simulated and estimated values for models M1 and M2 considering the parameters means of Subpopulation 1 and 2 (μ_1, μ_2), mixture proportion for the lowest mean subpopulation (π), genetic variance (σ^2_g), heritability (h^2), and correlation between simulated and predicted estimated best linear unbiased prediction (EBLUPs) (Cor_blups).

Scenario	Parameter	Simulated value†	M1‡	M2‡
1	μ_1	5	–	5.044 (0.296)
	μ_2	7	–	8.098 (1.136)
	π	0.2	–	0.488 (0.295)
	σ^2_g	1.994	2.634 (0.777)	1.701 (0.484)
	h^2	0.496	0.616 (0.126)	0.749 (0.262)
	Cor_blups	–	0.737	0.598
2	μ_1	5	–	4.443 (0.622)
	μ_2	7	–	7.548 (0.613)
	π	0.5	–	0.499 (0.057)
	σ^2_g	2.011	3.038 (1.155)	1.855 (0.443)
	h^2	0.499	0.633 (0.139)	0.76 (0.268)
	Cor_blups	–	0.704	0.586
3	μ_1	5	–	8.916 (0.4775)
	μ_2	20	–	19.612 (0.737)
	π	0.2	–	0.253 (0.063)
	σ^2_g	1.996	38.383 (36.448)	9.406 (10.008)
	h^2	0.498	0.861 (0.365)	0.764 (0.295)
	Cor_blups	–	0.222	0.298
4	μ_1	5	–	5.002 (0.253)
	μ_2	20	–	19.985 (0.252)
	π	0.5	–	0.5 (0.001)
	σ^2_g	2.01	58.719 (56.775)	2.008 (0.355)
	h^2	0.5	0.884 (0.387)	0.497 (0.062)
	Cor_blups	–	0.177	0.810

† Values for σ^2_g and h^2 are an average over 1000 runs.

‡ Values inside parentheses are standard error considering 1000 runs on the simulation.

subpopulations may occur within the same harvest. This premise is also supported by the year-to-year variation in the probability that a given genotype belongs to the low- or high-production population (Fig. 7).

The results obtained here show that coffee production data present overdispersion due to latent Gaussian mixtures, which is the result of different coffee genotypes that are in different physiological stages. This argument is supported not only by the mixture pattern and mean differences, but also by the predicted genotypic values and the pattern of the variance components. In this case, it could be argued to use a log-transformation on the data to solve overdispersion and distortions on variance components estimates; nevertheless, this approach would require working with log-normal models, which makes it hard to infer about the parameters and extend the model to a more general case (e.g., multivariate model). On the other hand, the proposed model is biologically intuitive and mathematically simple.

Given the difficulty of explicitly modeling individual biennial effects among progenies, the variations between subpopulations are embedded into residuals, inflating or deflating their value depending on the fluctuations in biennial status. In mixture models, the parameter-dependent variance is partly the weighted variance (mean) of groups and partly a measure of dispersion among the group means (subpopulations) (Gianola et al., 2004). When the group means are identical, this term is nullified. Therefore, the variance in this model is also affected by the mean of each group. This circumstance explains not only the high temporal dependence (covariance) between residuals throughout the years, but also the sudden fluctuations in residual variance for the mixed models observed in the present study. Gianola et al. (2007) observed a similar behavior for the error component in mixture models and considered this difference to be attributable to variability between the means for the two mixture components, which is not considered in the conventional mixed model. When a population presents heterogeneity at the genetic or environmental level, genetic parameters based on a theory that derives the estimators assuming that homogeneous distributions can lead to erroneous interpretations (Gianola et al., 2007). Therefore, the estimated variance components in mixed models tend to be biased when latent variables that affect stochastic processes exist. These observations are exemplified by the simulation study (Table 4). In all scenarios, the bias on genetic variance estimates was greater for model M1.

As discussed above, plant breeders commonly use 2-yr means for multienvironment analyses or for calculating overall means for multiple harvests in a single environment. This practice is followed to eliminate the effects of biennial growth, but there is a loss of information regarding the amount of data and covariance structures between environments. In addition, the assumptions that all progenies are at the same physiological stage and that 2-yr means that included a high- and a low-production year can capture the variability present in the field are naive. By contrast, the full analysis using data for multiple years loses power and precision due to high residual variance. However, our results indicate that these progenies vary regarding biennial behavior, and the use of 2-yr means may result in the selection of plants at favorable physiological stages (Fig. 6 and 7).

Figure 6 shows that some progenies were in a favorable stage (positive slope) and others were in an unfavorable stage (negative slope). However, the attributes of several progenies should be highlighted in relation to mixture models: these progenies are fitted according to the physiological stage and, in some cases, present an inverted slope. For example, the slope for Progeny 1 (green curve) indicates a drop in production between Years 1 and 2 in the repeated measures model, but a high production in the mixture model. In addition, this progeny tended not to present wide production fluctuations for the remaining years, indicating that the physiological stage was not sequentially binary.

To verify whether the grouping of 2-yr means eliminates the biennial effect, a joint analysis was performed considering four biennial means (eight harvests), and the results were compared with the mixture mixed model results. The repeated measures mixture model applied to the biennial data resulted in more accurate EBLUP for each pair of years (Supplemental Table S5), higher fitness (lower BICs) estimates, and higher expected GS values. The Spearman correlation between marginal EBLUPs for both models (0.84) indicated changes in genotypic rankings. If the effect of biennial growth had been mitigated by grouping the data by year pairs, the estimates would be very close for both models, but even in this scenario, the mixture model was superior. This pattern shows that the analysis of yield data for coffee genotypes using 2-yr grouped means or the mean of multiple harvests may be inefficient when the different progenies are at different physiological stages. Practical issues also prevent the use of 2-yr means, including the need to discard one evaluation when the number of evaluations is uneven.

Errors in selection are especially problematic in perennial plants because the selection cycles are relatively long and the per plot cost of evaluation is usually higher than that for annual crops. The main *C. arabica* cultivars adapted to Brazilian conditions have a narrow genetic background (Setotaw et al., 2013). Thus, the genetic variance and heritability are relatively low, further complicating the identification of the best progenies and requiring estimates with the lowest possible errors. Good phenotypic evaluations are therefore extremely important but do not guarantee selection efficiency when the models

are inadequate for the data. This situation even compromises the evaluation of experimental precision, giving the false impression that experiments evaluating coffee production have a low experimental precision.

Conventional models for the phenotypic evaluation of coffee may lead to inference errors when fitted to data originating from stochastic processes other than Gaussian ones (Gianola et al., 2007). The production data for the different coffee progenies used in the present study were apparently formed by a mixture of two Gaussian distributions.

Computational demand and convergence are well known problems when fitting mixed models, especially when modeling (co)variance matrices structures. In the proposed model, the computational demand could be increased by estimating the $\mathbf{P}_{n\times 2}$ matrix, which does not require great computational resource given its low dimensionality, even with a very large population size. In this work, we did not observe significant differences on computational time to fit mixture mixed models compared with standard mixed models. The problem of estimates being outside the parameter space can be alleviated by the use of an EM algorithm in the REML function (Dempster et al., 1977). On mixture model, the physiological state of the individual is estimated by a nonlinear function (as pointed out before) given the latent variable. It is also true that if the latent variable is observed, the model becomes linear (Supplemental File S2) and the solution for fixed and random effects becomes asymptotically estimated best linear unbiased estimation (EBLUE) and EBLUP, respectively, such as the classical mixed models on Gaussian distribution. Therefore, these terms (EBLUE and EBLUP) were adopted here assuming that the prefix "E" (estimated or empirical) is related to the asymptotic linearity, unbiasedness, and minimal variance when the true values are replaced by its estimates (Sorensen and Gianola, 2007, p. 212).

Given that EM-REML maximizes the marginal likelihood (free from the nuisance parameters), the resultant model is a residual or (average) of residual likelihoods. Then, in this manuscript, we prefer to use the term REML as "residual' instead "restricted" likelihood, since EM-REML does not poses directly a restriction (contrast) on the fixed effects ($\mathbf{KX}\boldsymbol{\beta} = 0$), as suggested originally by Patterson and Thompson (1971). The mixed mixture models allow the inclusion of kinship information as in traditional mixed models for a prior random effects distribution [i.e., $\mathbf{u} \sim N(\mathbf{0}, \sigma^2_g\mathbf{G})$], where $\mathbf{G}$ is the kinship matrix (from pedigree or molecular markers). Therefore, using molecular markers, the model proposed in this study can be easily adapted to perform genomic selection. Considering multiple mixtures is also possible, which broadens the usage of this model; however, due to agronomical justification for biennially, it was used only two mixtures in this work.

Overall, the conventional and mixture models resulted in different rankings for the EBLUPs of progeny, and the mixture model resulted in a higher GS and a greater confidence in estimates than with the classical mixed model, even when considering the heterogeneity of variance. The mixture model corrected the genotype response according to the year of evaluation, indicating that the effect of the genotype × year interaction on selection may have been reduced.

With the methodology proposed in the present study, the best progenies could be identified in coffee breeding programs during early generations ($S_{0:3}$, $S_{0:4}$), when a stronger confounding biennial effect among progenies occurs compared with that in later generations. The biennial pattern is not stable until the fourth year, as commented on above. Since traditional models do not correct for biennial effects, the predicted breeding values will be confounded by latent variable effects, making them inefficient for use in performing selection. For example, Progeny 7, mentioned above, certainly would be selected if the four first harvests were used; however, after this harvest sequence, the biennial pattern became stable, and its EBLUP values approached zero (Fig. 6A). Many authors have recommended evaluating at least four harvests before selection (Pereira et al., 2013). A mixture model is able to capture the biennial behavior and correct for this hidden variable in the first evaluation, as can be observed in Fig. 7. Therefore, selection would be possible after fewer years of evaluation (2 or 3 yr), similar to what is possible for annual crops (Bernardo, 2002). This reduction could have implications for the way that phenotypic evaluations of coffee plants are conducted in Brazil. For example, the time and cost of launching a new cultivar could be greatly decreased by accelerating recurrent selection cycles and increasing the genetic gain per cycle. Currently, at least 8 yr are needed to complete a full recurrent selection cycle, and most of this time is spent on evaluation (Sera, 2001). We expect that evaluating two harvests before selection would save 4 yr of effort, depending on the breeding strategy adopted.

Simulation results illustrated the properties of the proposed model (Table 4). When the subpopulation means were close to each other, the standard mixed model was better. This behavior was expected, because the mixture parameter is not identifiable in that situation (Aitkin and Wilson, 1980). When this difference was >15, the mixture model was superior, especially in Scenario 4 (Table 4), where the model M2 was more efficient than M1 in estimating the EBLUPs. According to our results, differences in the mean of subpopulations (estimated by model M2) were always large enough to guarantee the superiority of the mixture model (Table 2). Given our results and previous studies, we expect larger differences in subpopulation means being a rule in *Coffea arabica* for the yield trait (Bertrand et al., 2005; Sakai et al., 2015). This reinforces

the justification of modeling population heterogeneity in these data. In addition, it is worth highlighting that even when the difference between the mixtures means are not evident, our REML estimation for variance components and EBLUE for mixture means are robust, showing the model's ability to identify bimodality.

In general, the mixture model was able to identify biennial growth as a latent variable in the yield data for coffee progenies and to cluster the different progenies in a coherent way. This model considerably improved the estimation of the evaluated parameters, showing statistical superiority over the models reported so far to address this agronomic phenomenon. In addition, changes to the EBLUP rankings of the best progenies selected and an increased expected GS indicate that the biennial effects must be taken into account in the evaluation of coffee progenies.

Supplemental Material

Supplemental material is available online for this article.

Conflict of Interest

The authors declare that there is no conflict of interest.

Acknowledgments

The authors thank the National Council for Scientific and Technological Development (CNPq) and Empresa de Pesquisa Agropecuária de Minas Gerais (EPAMIG) for financial support. The authors are also deeply grateful to the anonymous reviewers for their valuable comments and suggestions.

References

Aitkin, M., and G.T. Wilson. 1980. Mixture models, outliers, and the EM algorithm. Technometrics 22:325–331. doi:10.1080/00401706.1980.10486163

Andrade, V.T., F.M.A. Gonçalves, J.A.R. Nunes, and C.E. Botelho. 2016. Statistical modeling implications for coffee progenies selection. Euphytica 207:177–189. doi:10.1007/s10681-015-1561-6.

Bacha, C.J.C. 1998. A cafeicultura brasileira nas décadas de 80 e 90 e suas perspectivas. Preços Agríc. 7:14–22.

Balestre, M., P.P. Torga, R.G. Von Pinho, and J.B. dos Santos. 2013. Applications of multi-trait selection in common bean using real and simulated experiments. Euphytica 189:225–238. doi:10.1007/s10681-012-0790-1.

Bernardo, R. 2002. Breeding for quantitative traits in plants. Stemma Press, Woodbury, MN.

Bertrand, B., H. Etienne, C. Cilas, A. Charrier, and P. Baradat. 2005. *Coffea arabica* hybrid performance for yield, fertility and bean weight. Euphytica 141:255–262. doi:10.1007/s10681-005-7681-7

Davis, A.P., T.W. Gole, S. Baena, and J. Moat. 2012. The impact of climate change on indigenous arabica coffee (*Coffea arabica*): predicting future trends and identifying priorities. PLoS One 7:e47981. doi:10.1371/journal.pone.0047981

de Oliveira, A.C.B., A.A. Pereira, F.L. da Silva, J.C. de Rezende, C.E. Botelho, and G.R. Carvalho. 2011. Prediction of genetic gains from selection in Arabica coffee progenies. Crop Breed. Appl. Biotechnol. 11:106–113. doi:10.1590/S1984-70332011000200002

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B 39:1–38.

Detilleux, J., and P.L. Leroy. 2000. Application of a mixed normal mixture model for the estimation of mastitis-related parameters. J. Dairy Sci. 83:2341–2349. doi:10.3168/jds.S0022-0302(00)75122-8

Fisch, R.D., M. Ragot, and G. Gay. 1996. A generalization of the mixture model in the mapping of quantitative trait loci for progeny from a biparental cross of inbred lines. Genetics 143:571–577.

Gianola, D., P.J. Boettcher, J. Ødegård, and B. Heringstad. 2007. Mixture models in quantitative genetics and applications to animal breeding. Rev. Bras. Zootec. 36:172–183. doi:10.1590/S1516-35982007001000017

Gianola, D., J. Øegård, B. Heringstad, G. Klemetsdal, D. Sorensen, P. Madsen, et al. 2004. Mixture model for inferring susceptibility to mastitis in dairy cattle: A procedure for likelihood-based inference. Genet. Sel. Evol. 36:3. doi:10.1186/1297-9686-36-1-3

Holzmann, H., and S. Vollmer. 2008. A likelihood ratio test for bimodality in two-component mixtures with application to regional income distribution in the EU. AStA Adv. Stat. Anal. 92:57–69. doi:10.1007/s10182-008-0057-2

Jamrozik, J., and L.R. Schaeffer. 2010. Application of multiple-trait finite mixture model to test-day records of milk yield and somatic cell score of Canadian Holsteins. J. Anim. Breed. Genet. 127:361–368. doi:10.1111/j.1439-0388.2010.00875.x

Klingenberg, C.P., and L.J. Leamy. 2001. Quantitative genetics of geometric shape in the mouse mandible. Evolution 55:2342–2352. doi:10.1111/j.0014-3820.2001.tb00747.x

Lewin, B., D. Giovannucci, and P. Varangis. 2004. Coffee markets: New paradigms in global supply and demand. Agric. Rural Dev. Disc. Paper 3. World Bank, Washington, DC.

Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Assoc., Sunderland, MA.

Melchinger, A.E., W. Schipprack, T. Würschum, S. Chen, and F. Technow. 2013. Rapid and accurate identification of in vivo-induced haploid seeds based on oil content in maize. Sci. Rep. 3:2129. doi:10.1038/srep02129

Murphy, K.P. 2012. Machine learning: A probabilistic perspective. MIT Press, Cambridge, MA.

Osorio, N. 2002. The global coffee crisis: A threat to sustainable development. Int. Coffee Org., London. http//www.ico.org/documents/globalcrisise (accessed 6 May 2019).

Patterson, H.D., and R. Thompson. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58:545–554. doi:10.1093/biomet/58.3.545

Pereira, T.B., J.P.F. Carvalho, C.E. Botelho, M.D.V. de Resende, J.C. de Rezende, and A.N.G. Mendes. 2013. Eficiência da seleção de progênies de café F_4 pela metodologia de modelos mistos (REML/BLUP). Bragantia 72:230–236. doi:10.1590/brag.2013.031

Redner, R.A., and H.F. Walker. 1984. Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. 26:195–239. doi:10.1137/1026034

Rena, A.B., M. Maestri, A.B. Rena, E. Malavolta, M. Rocha, and T. Yamada. 1986. Fisiologia do cafeeiro. Inf. Agropecu. 11:26–40.

Rencher, A.C., and G.B. Schaalje. 2008. Linear models in statistics. John Wiley & Sons, Hoboken, NJ.

Resende, M.D., and J.B. Duarte. 2007. Precisão e controle de qualidade em experimentos de avaliação de cultivares. Pesqui. Agropecu. Trop. 37:182–194.

Rodrigues, W.P., H.D. Vieira, D. Barbosa, G.R. Sousa Filho, and F.L. Partelli. 2014. Agronomic performance of arabica coffee genotypes in northwest Rio de Janeiro State. Genet. Mol. Res. 13:5664–5673. doi:10.4238/2014.July.25.22

Sakai, E., E.A.A. Barbosa, J.M. de Carvalho Silveira, and R.C. de Matos Pires. 2015. Coffee productivity and root systems in cultivation schemes with different population arrangements and with and without drip irrigation. Agric. Water Manage. 148:16–23. doi:10.1016/j.agwat.2014.08.020

SAS Institute. 2009. User's guide: Statistics. SAS Inst., Cary, NC.

Schilling, M.F., A.E. Watkins, and W. Watkins. 2002. Is human height bimodal? Am. Stat. 56:223–229. doi:10.1198/00031300265

Schwarz, G. 1978. Estimating the dimension of a model. Ann. Stat. 6:461–464.

Sera, T. 2001. Coffee genetic breeding at IAPAR. Crop Breed. Appl. Biotechnol. 1(2). doi:10.13082/1984-7033.v01n02a08

Setotaw, T.A., E.T. Caixeta, A.A. Pereira, A.C. de Oliveira, C.D. Cruz, E.M. Zambolim, et al. 2013. Coefficient of parentage in *Coffea arabica* L. cultivars grown in Brazil. Crop Sci. 53:1237–1247. doi:10.2135/cropsci2012.09.0541

Smith, A.B., J.K. Stringer, X. Wei, and B.R. Cullis. 2007. Varietal selection for perennial crops where data relate to multiple harvests from a series of field trials. Euphytica 157:253–266. doi:10.1007/s10681-007-9418-2

Sorensen, D., and D. Gianola. 2007. Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer, New York.

Waller, J.M., M. Bigger, and R.J. Hillocks, editors. 2007. Coffee pests, diseases and their management. CABI, Egham, UK. doi:10.1079/9781845931292.0000

File S1: Tables S1 to S5

Table S1. Residual co(variances) estimated using the model M3 (repeated measures mixed models).

Harvest	1	2	3	4	5	6	7	8
1	50.48	69.23	6.17	66.45	3.03	80.84	-1.89	64.26
2	-	239.85	-44.19	192.89	-20.38	211.35	-21.98	171.82
3	-	-	62.44	-55.09	29.89	-51.41	17.52	-30.53
4	-	-	-	207.65	-31.08	208.12	-23.79	168.65
5	-	-	-	-	27.96	-22.71	15.81	-18.99
6	-	-	-	-	-	293.80	-25.01	182.98
7	-	-	-	-	-	-	16.63	-25.05
8	-	-	-	-	-	-	-	208.37

Table S2. Residual co(variances) estimated using the model M4 (repeated measures mixture mixed models).

Harvest	1	2	3	4	5	6	7	8
1	12.61	9.77	5.67	9.95	-0.61	7.79	0.35	10.88
2	-	58.02	-1.62	37.59	-8.01	42.67	-6.62	29.76
3	-	-	13.03	0.32	2.34	3.45	2.54	3.14
4	-	-	-	64.30	-9.37	56.67	-6.28	33.00
5	-	-	-	-	9.89	-11.07	2.75	-8.27
6	-	-	-	-	-	95.66	-7.72	32.72
7	-	-	-	-	-	-	5.35	-4.12
8	-	-	-	-	-	-	-	47.62

Table S3. Genetic co(variances) estimated using the model M3 (repeated measures mixed models).

Harvest	1	2	3	4	5	6	7	8
1	13.71	13.08	10.71	10.95	3.99	19.35	0.65	31.62
2	-	92.39	6.22	65.73	-4.10	128.42	1.42	144.37
3	-	-	19.09	12.70	8.49	29.60	1.40	42.15
4	-	-	-	87.04	-5.19	161.36	-5.97	189.45
5	-	-	-	-	9.73	-1.92	4.64	3.84
6	-	-	-	-	-	310.37	-6.06	363.69
7	-	-	-	-	-	-	3.51	-5.86
8	-	-	-	-	-	-	-	436.86

Table S4. Genetic co(variances) estimated using the model M4 (repeated measures mixture mixed models).

Harvest	1	2	3	4	5	6	7	8
1	9.77	16.03	4.53	17.04	-0.90	27.83	-0.10	29.51
2	-	96.73	15.35	64.56	0.03	150.72	0.75	153.66
3	-	-	6.63	1.47	1.69	6.25	1.88	6.36
4	-	-	-	94.95	0.90	177.96	-3.17	188.99
5	-	-	-	-	4.60	-0.21	1.86	2.11
6	-	-	-	-	-	358.76	-5.71	375.49
7	-	-	-	-	-	-	1.14	-5.36
8	-	-	-	-	-	-	-	394.97

Table S5. Prediction error variance estimates (PEV), expected selection gain for the five best progenies (SG%), Bayesian information criterion (BIC), Spearman correlation between BLUPs estimated using the models M3 (repeated measures mixed models) and M4 (repeated measures mixture mixed models) and means (Mean1 and Mean2) estimated by the model M4 considering the data grouped in biennials.

Biennium	M3	M4		
	PEV	PEV	Mean1	Mean2
1	9.099	6.717	33.545	48.940
2	4.374	4.212	27.591	37.522
3	9.049	7.427	32.024	47.941
4	13.439	7.730	42.839	53.118
SG(%)	13.582	14.460		
BIC	902.730	670.605		
cor_Spearman(%)	0.84			

File S2: Mixture model justification

Consider y_i as the phenotypic observation related to i-th individual whose observed likelihood is given by:

$$L(\pi_1,\theta,\sigma^2|\mathrm{y}_i) = \pi\, p(\mathrm{y}_i \mid \theta_1,\sigma^2) + (1-\pi)\, p(\mathrm{y}_i \mid \theta_2,\sigma^2) = \pi P_1 + (1-\pi)P_2$$

Where $\theta_1 = \boldsymbol{\mu}_1 + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ and $\theta_2 = \boldsymbol{\mu}_2 + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ and $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean parameters related physiological phase, σ^2 is the residual variance and π is the prior probability of y_i being a random realization of P_k. For a vector $\mathbf{y}_{(nx1)}$ the observed-data likelihood is given by: $L(\pi_1,\theta,\sigma^2|\mathbf{y}) = \prod_{i=1}^{n} \pi P_{i1} + (1-\pi)P_{i2}$ whose log-likelihood is $l(\pi_1,\theta,\sigma^2|\mathbf{y}) = \sum \log[\pi P_{i1} + (1-\pi)P_{i2}]$

where the posterior probability of π is obtained taking the derivative of $l(\pi_1,\theta,\sigma^2|\mathbf{y})$ with relation to π which is :

$$\hat{\pi} = \frac{\sum_{i=1}^{n} \Upsilon_{i1}}{\sum_{i=1}^{n} \Upsilon_{i1} + \Upsilon_{i2}} \text{ where } \Upsilon_1 = \frac{\pi P_{i1}}{\pi P_{i1} + (1-\pi)P_{i2}} \text{ and } \Upsilon_2 = \frac{(1-\pi)P_{i2}}{\pi P_{i1} + (1-\pi)P_{i2}}$$

where Υ_1 and Υ_2 are the posterior probability **P** of y_i being classified into the first or second mixture component. However, the EM algorithm makes use of a latent random variable **j** given as missing information. The joint likelihood of observed data **y** and the missing information related to classifier j is called of complete likelihood (Sorensen and Gianola, 2002) where the classifier **j** is a missing Bernoulli random variable. Thus, the complete log-likelihood can be described as following:

$$l(\pi_1,\theta,\sigma^2|\mathbf{y}, j, \mathbf{u}) = \sum_{i=1}^{n} \log[\pi_1 P_{i1}^{\,I(j=1)} \pi_2 P_{i2}^{\,1-I(j=1)}] = \sum_{i=1}^{n}\sum_{k=1}^{2} I(j_k = 1)\log P_{ik} + I(j_k = 1)\log \pi_k$$

Here, it is evident that 1- $I(j_k = 1) = \left[I(j_i = 0) \right]$. Taking the expectation in relation to the missing Bernoulli variable $E\left[I(j_i = 1) \right]$ and the random genetics effects $\mathbf{u}$, the complete-data likelihood becomes:

$$E_{\mathbf{u},\mathbf{j}|\mathbf{y}}\left[l(\pi, \boldsymbol{\theta}, \sigma^2 | \mathbf{y}, \mathbf{j}, \mathbf{u}) \right] = E_{\mathbf{u},\mathbf{j}|\mathbf{y}}\left[p(\mathbf{y}, \mathbf{j} | \pi, \boldsymbol{\theta}, \sigma^2)\, p(\mathbf{u} | \mathbf{y}, \mathbf{j}) \right]$$

In likelihood inference the E-step process can be given using two step: The first one is given using only the joint (complete) likelihood expectation in relation to *j* in which is given by:

$$l(\pi, \boldsymbol{\theta}, \sigma^2 | \mathbf{y}, \mathbf{j}) = \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{k=1}^{2} E\left[(j_k = 1) \right](y_i - \theta_k)^2 + E\left[(j_k = 1) \right]\log \pi_k$$

Given that for y_i the summation of the expectation is equal to 1, the joint likelihood of (y,j) can be rewritten as following:

$$l(\pi, \boldsymbol{\theta}, \sigma^2 | \mathbf{y}, \mathbf{j}) = \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} E\left[(\mathbf{j}_i = 1) \right](y_i - \theta)^2 + \sum_{i=1}^{n}\sum_{k=1}^{2} + E\left[(j_k = 1) \right]\log \pi_k =$$

$$l(\pi, \boldsymbol{\theta}, \sigma^2 | \mathbf{y}, \mathbf{j}) = \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} E_j(y_i - \mathbf{J}\boldsymbol{\mu} - \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u})^2 + \sum_{i=1}^{n}\sum_{k=1}^{2} + \mathbf{P}_{ik}\log \pi_k$$

Thus, $\mathbf{J}_i\mathbf{u}$ describes the linear combination of expectation of i[th] plant be in the j[-th] physiological state with its corresponded mean μ_k where in the i-th line $\mathbf{J}_i\boldsymbol{\mu} = \begin{bmatrix} j & 1-j \end{bmatrix}\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = j\mu_1 + (1-j)\mu_2$.

The **P** elements can be obtained as follow: Assuming j as a Bernoulli random variable indicating the physiological phase of the plants and using the expectation function the complete likelihood in relation to (**y**,**j**) we have:

$E\left[(j_k = 1) \right] = \sum_{j=0}^{1} jBernoulli(j|\rho) = 1 \times \rho = \rho$, where $\rho = \pi_1 P_{i1}$ and

$E\left[(j_k = 0) \right] = \sum_{j=0}^{1} jBernoulli(j|\rho) = 1 \times (1-\rho) = 1-\rho$ where $1-\rho = (1-\pi)P_{i2}$, This result can be used to build the matrix $\mathbf{P}$ as following:

$$\mathbf{P} = \begin{bmatrix} E[(\mathbf{j}_1 = 1)] = \pi P_{11} & E[(\mathbf{j}_1 = 0)] = (1-\pi) P_{12} \\ E[(\mathbf{j}_2 = 1)] = \pi P_{21} & E[(\mathbf{j}_2 = 0)] = (1-\pi) P_{22} \\ E[(\mathbf{j}_3 = 1)] = \pi P_{31} & E[(\mathbf{j}_3 = 0)] = (1-\pi) P_{32} \\ . & . \\ . & . \\ . & . \\ E[(\mathbf{j}_n = 1)] = \pi P_{n1} & E[(\mathbf{j}_n = 0)] = (1-\pi) P_{n2} \end{bmatrix}$$

However, since $\pi_1 P_{i1} + (1-\pi) P_{i2} \neq 1$, the following normalization is necessary to ensure that the sum of probabilities be equal to one: $\Upsilon_1 = \dfrac{\pi P_{i1}}{\pi P_{i1} + (1-\pi) P_{i2}}$ and $\Upsilon_2 = \dfrac{(1-\pi) P_{i2}}{\pi P_{i1} + (1-\pi) P_{i2}}$. Thus, the normalized matrix $\mathbf{P}$ is now given by:

$$\hat{\mathbf{P}} = \begin{bmatrix} E[(\mathbf{j}_1 = 1)] = \Upsilon_{11} & E[(\mathbf{j}_1 = 0)] = \Upsilon_{12} \\ E[(\mathbf{j}_2 = 1)] = \Upsilon_{21} & E[(\mathbf{j}_2 = 0)] = \Upsilon_{22} \\ E[(\mathbf{j}_3 = 1)] = \Upsilon_{31} & E[(\mathbf{j}_3 = 0)] = \Upsilon_{32} \\ . & . \\ . & . \\ . & . \\ E[(\mathbf{j}_n = 1)] = \Upsilon_{n1} & E[(\mathbf{j}_n = 0)] = \Upsilon_{n2} \end{bmatrix}$$

Since the others fixed and random effects are constant across the mixtures, their effects are invariants in the complete-data likelihood. In short, if the biennial stage is known the two mean mixture model can be described by:

$$\mathbf{y} = \mathbf{J}\boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

$$\mathbf{y} = \mathbf{j}\mu_1 + (1-\mathbf{j})\,\mu_2 + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where J_{nx2} is a 0's and 1's matrix relating mean to each mixture component. The **X** and **Z** are incidence matrices for fixed and random effects and $\boldsymbol{\beta}$ and $\mathbf{u}$ fixed and random effects, respectively.

Since the expectation in relation to **j** is obtained, the next step is to take the expectation in relation to $\mathbf{u}$. Then, the objective function **Q** becomes:

$$E_{\mathbf{u},\mathbf{j}|\mathbf{y}}\left[\, p(\mathbf{y},\mathbf{j}\,|\,\pi\ ,\boldsymbol{\theta},\sigma^2)\,p(\mathbf{u}\,|\,\mathbf{y},\mathbf{j})\right] = E_{\mathbf{u}|\mathbf{y}}\left[\begin{array}{l}\dfrac{n}{2}\log(\sigma^2)-\dfrac{1}{2\sigma^2}\displaystyle\sum_{i=1}^{n}E_j(y_i-\mathbf{J}_i\boldsymbol{\mu}-\mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i\mathbf{u}\ )^2+\sum_{i=1}^{n}\sum_{k=1}^{2}+\mathbf{P}_{ik}\log\pi_k \\ -\log N(\mathbf{u}\,|\,0,\sigma_g^2)\end{array}\right]$$

For REML estimates an additional expectation in relation to fixed effects becomes necessary:

$$E_{\mathbf{u},\mathbf{j},\varphi|\mathbf{y}}\left[\, p(\mathbf{y},\mathbf{j}\,|\,\pi\ ,\boldsymbol{\theta},\sigma^2)\,p(\mathbf{u}\,|\,\mathbf{y},\mathbf{j})\right] = E_{\mathbf{u},\varphi|\mathbf{y}}\left[\begin{array}{l}\dfrac{n}{2}\log(\sigma^2)-\dfrac{1}{2\sigma^2}\displaystyle\sum_{i=1}^{n}E_j(y_i-\mathbf{J}_i\boldsymbol{\mu}-\mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i\mathbf{u}\ )^2+\sum_{i=1}^{n}\sum_{k=1}^{2}+\mathbf{P}_{ik}\log\pi_k \\ -\log N(\mathbf{u}\,|\,0,\sigma_g^2)\end{array}\right]$$

where $\boldsymbol{\varphi}=\{\boldsymbol{\mu},\boldsymbol{\beta}\}$. It is evident here that after to take the expectation in relation to Bernoulli random variable the EM step is equal to that used in classical EM-REML mixed models and the parameter derivation becomes equivalent.

Derivation of Mixed model effects

The solution for the fixed and random effects can be done using the objective function:

$$l(\pi\ ,\boldsymbol{\theta},\sigma^2|\mathbf{y},\mathbf{j})=\frac{n}{2}\log(\sigma^2)-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n}\sum_{k=1}^{2}E\left[(j_k=1)\right](y_i-\theta_k)^2+E\left[(j_k=1)\right]\log\pi_k\right\}-\log N(\mathbf{u}\,|\,0,\sigma_g^2)$$

In matrix notation we rewrite the likelihood as:

$$l(\pi\ ,\boldsymbol{\theta},\sigma^2|\mathbf{y},\mathbf{j})=\frac{n}{2}\log(\sigma^2)-\frac{1}{2\sigma^2}\left[\begin{array}{l}E_{j,u}\left\{(\mathbf{y}-\mathbf{j}\mu_1-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})\grave{}(\mathbf{y}-\mathbf{j}\mu_1-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})+\mathbf{j}\log\pi\right\} \\ +E_{[(1-j),u]}\left\{\begin{array}{l}(\mathbf{y}-(\mathbf{1}-\mathbf{j})\mu_2-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})\grave{}(\mathbf{y}-(\mathbf{1}-\mathbf{j})\mu_2-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) \\ +(\mathbf{1}-\mathbf{j})\log(1-\pi)\end{array}\right\}\end{array}\right]$$

$$-E_u\log N(\mathbf{u}\,|\,0,\sigma_g^2)$$

Deriving partially in relation to μ_1 and μ_2 we have we an equivalent least square estimator:

$$\mu_1=E_j(\mathbf{j}\grave{}\mathbf{j})^{-1}E_j(\mathbf{j}\grave{})\left(y-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}E_u(\mathbf{u})\right)$$

$$\mu_2=E_{[(1-j)]}\left[(\mathbf{1}-\mathbf{j})\grave{}(\mathbf{1}-\mathbf{j})\right]^{-1}E_{[(1-j)]}(\mathbf{1}-\mathbf{j})\grave{}(y-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}E_u(\mathbf{u}))$$

As describe above $E[(\mathbf{j}_1 = 1)] = \Upsilon_1 = \frac{\pi P_{i1}}{\pi P_{i1} + (1-\pi)P_{i2}}$ and $E_{[(1-j)]} = \Upsilon_2 = \frac{(1-\pi)P_{i2}}{\pi P_{i1} + (1-\pi)P_{i2}}$. On the other hand, we have:

$$E_j(\mathbf{j}`\mathbf{j}) = \sum_i^n Var(\mathbf{j}) + \left[E_j(\mathbf{j}`)\right]^2 = \sum_i^n \rho_i(1-\rho_i) + \rho_i^{\,2} = \sum_i^n \rho_i = \sum_i^n \frac{\pi P_{i1}}{\pi P_{i1} + (1-\pi)P_{i2}} \text{ and}$$

$$E_{[(1-j)]}\left[(\mathbf{1}-\mathbf{j})`(\mathbf{1}-\mathbf{j})\right] = \sum_i^n \left[\rho_i(1-\rho_i) + (1-\rho_i)^2\right] = \sum_i^n (1-\rho_i)(\rho_i + 1 - \rho_i) = \sum_i^n (1-\rho_i) = \sum_i^n \frac{(1-\pi)P_{i2}}{\pi P_{i1} + (1-\pi)P_{i2}}$$

Now, we can plug it in the mean estimator:

$$\mu_1 = \frac{\Upsilon_1 `\mathbf{y}^*}{\mathbf{j}`\Upsilon_1} = \frac{\sum_{i=1}^n \Upsilon_{i1} y^*_i}{\sum_{i=1}^n \Upsilon_{i1}}$$

$$\mu_2 = \frac{\Upsilon_2 `\mathbf{y}^*}{\mathbf{j}`\Upsilon_1} = \frac{\sum_{i=1}^n \Upsilon_{i2} y^*_i}{\sum_{i=1}^n \Upsilon_{i2}}$$

Where $\mathbf{y}^* = \mathbf{y} - \mathbf{X\beta} - \mathbf{Z}E_u(\mathbf{u})$

In order to facility the notation we used $E_u(\mathbf{u}) = \hat{\mathbf{u}}$.

This result is the same classical EM solution for mixture models found, for example, in (Bishop, 2006) chapter 9. Now, consider that the marginal solution for μ_1 and μ_2 could be given by:

$\boldsymbol{\mu} = E(\mathbf{J}`\mathbf{J})^{-1}E(\mathbf{J}`)(y - \mathbf{X\beta} - \mathbf{Z\hat{u}})$ where $E_j(\mathbf{J}`) = \mathbf{P}$ and

$$E(\mathbf{J}`\mathbf{J}) = \begin{pmatrix} \sum_i^n \Upsilon_1 & 0 \\ 0 & \sum_i^n \Upsilon_2 \end{pmatrix}$$

Here, the off-diagonal elements $E\left[\mathbf{j}\ ,\mathbf{1-j}\right]=\Upsilon_1`\Upsilon_2=0$ in order to ensure the marginal mean allowing the least square estimation since $E\left[\mathbf{j}\ ,\mathbf{1-j}\right]=\rho_i(1-\rho_i)\neq 0$

Proof:

$$\boldsymbol{\mu}=\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}=\begin{pmatrix}\sum_i^n \Upsilon_{i1} & 0\\ 0 & \sum_i^n \Upsilon_{i2}\end{pmatrix}^{-1}\begin{pmatrix}\Upsilon_1`\\ \Upsilon_2`\end{pmatrix}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Z\hat{u}})=\begin{bmatrix}\dfrac{\Upsilon_1`(y-\mathbf{X\beta}-\mathbf{Z\hat{u}})}{\sum_i^n \Upsilon_{i1}}=\dfrac{\Upsilon_1`\mathbf{y}^*}{\sum_i^n \Upsilon_{i1}}\\ \dfrac{\Upsilon_2`(y-\mathbf{X\beta}-\mathbf{Z\hat{u}})}{\sum_i^n \Upsilon_{i2}}=\dfrac{\Upsilon_2`\mathbf{y}^*}{\sum_i^n \Upsilon_{i2}}\end{bmatrix}$$

This is the exactly marginal solution for mixture means in classical likelihood estimation.

In order to facility the notation, it will be adopt that $E_j(\mathbf{J}`\mathbf{J})=\breve{\mathbf{P}}`\breve{\mathbf{P}}$ and $E_j(\mathbf{J}`)=\mathbf{P}$.Then, the marginal solution for $\boldsymbol{\mu}$ is given by:

$$\boldsymbol{\mu}=(\breve{\mathbf{P}}`\breve{\mathbf{P}})^{-1}\mathbf{P}`(y-\mathbf{X\beta}-\mathbf{Z\hat{u}})$$

Since the fixed and random effects $(\boldsymbol{\beta},\mathbf{u})$ are constant across E_j, using the objective function one have:

$$\boldsymbol{\mu}=(\breve{\mathbf{P}}`\breve{\mathbf{P}})^{-1}\mathbf{P}`(\mathbf{y}-\mathbf{X\beta}-\mathbf{Z\hat{u}})$$

$$\boldsymbol{\beta}=(\mathbf{X}`\mathbf{X})^{-1}\mathbf{X}`(\mathbf{y}-\mathbf{P\mu}-\mathbf{Z\hat{u}})$$

And

$$\hat{\mathbf{u}}=\left(\mathbf{Z}`\mathbf{Z}+\frac{\sigma^2}{\sigma_a^2}\right)^{-1}\mathbf{Z}`(\mathbf{y}-\mathbf{P\mu}-\mathbf{X\beta})$$

However, the EM-REML requires the joint solution of fixed and random effects and in order to ensure that the marginal likelihood be free from nuisance parameters. Therefore, the joint solution for μ_1 and μ_2 must be take into account the joint expectation among **j** and (**1-j**). A initial condition is that if the binary indicator $\mathbf{j}=1$ then it is necessary that $(\mathbf{1}-\mathbf{j})=0$. If so, for the i[th] observation we have $\mathbf{j}_i=[1,0]$ with probability [ρ ,$1-\rho$] and $(1-\mathbf{j}_i)=[0,1]$ and probability [ρ , $1-\rho$] if the bienallity state was known. Therefore, the joint expectation for first diagonal element

of $\breve{\mathbf{P}}\grave{}\breve{\mathbf{P}}$ is $E_{j,1-j}\left(j\grave{}j\right)=\sum_{i}^{n}\ \sum j^{2}p(j,1-j)=\sum_{i}^{n}1^{2}\rho^{2}+0^{2}\left(1-\rho\right)^{2}=\sum_{i}^{n}\rho^{2}$, in sense that the pairs [1,1] and [0,0] are not marginally observed. Similarly, for the second diagonal element we have

$$E_{j,1-j}\left(\left(1-j\right)\breve{}\left(1-j\right)\right)=\sum_{i}^{n}\sum\left(1-j\right)^{2}\ p(j,1-j)=\sum_{i}^{n}\left[0^{2}\rho^{2}+1^{2}\left(1-\rho\right)\right]^{2}=\sum_{i}^{n}\left(1-\rho\right)^{2}$$

Note that the joint expectation is very different from the marginal one as given above.

The off-diagonal elements in this scenario can emerge from the assumption that the mixtures are independent. For ensure this, it is necessary that $COV\left[\left(\mathbf{j}\right),\left(\mathbf{1}-\mathbf{j}\right)\right]=0$. This assumption can be obtained using the following equality $E\left(\mathbf{j}\right)\grave{}E\left(1-\mathbf{j}\right)=E_{j,1-j}\left[\left(\mathbf{j}\right)\breve{}\left(\mathbf{1}-\mathbf{j}\right)\right]\neq 0$. In other words, given **j** and **1-j** are replaced by its joint expectations the vectors are not more orthogonal. To see this consider that if $\mathbf{j}_{i}=\left[1,0\right]$, the necessary condition for $\left(1-\mathbf{j}_{i}\right)=\left[0,1\right]$ in the sense that $\mathbf{j}_{i}\grave{}\left(1-\mathbf{j}_{i}\right)=0$. However, since the indicator **j** variable is not observed, but, instead, it is replaced by their expectation, the joint expectation of square form becomes: $E_{j,1-j}\left[\left(\mathbf{j}\right)\breve{}\left(\mathbf{1}-\mathbf{j}\right)\right]=COV\left[\left(\mathbf{j}\right),\left(\mathbf{1}-\mathbf{j}\right)\right]+E\left(\mathbf{j}\right)\grave{}E\left(1-\mathbf{j}\right)=0+\rho(1-\rho)=\rho(1-\rho)$.

Proof:

$$COV\left[\left(\mathbf{j}\right),\left(\mathbf{1}-\mathbf{j}\right)\right]=\rho(1-\rho)\left[1-\rho\right]\left[0-\rho\right]+\rho(1-\rho)\left[0-\rho\right]\left[1-\rho\right]+\rho^{2}\left[1-\rho\right]\left[1-\rho\right]+\left(1-\rho\right)^{2}\left[0-\rho\right]\left[0-\rho\right]=$$
$$-2\rho^{2}(1-\rho)^{2}+2\rho^{2}(1-\rho)^{2}=0$$

Then the new matrix of the joint expectation becomes:

$$E_{j,1-j}(\mathbf{J}\grave{}\mathbf{J})=\begin{pmatrix}E_{j,1-j}(\mathbf{j}\grave{}\mathbf{j}) & E_{j,1-j}\left[\left(\mathbf{j}\right)\breve{}\left(\mathbf{1}-\mathbf{j}\right)\right]\\ E_{j,1-j}\left[\left(\mathbf{1}-\mathbf{j}\right)\breve{}\left(\mathbf{j}\right)\right] & E_{j,1-j}\left[\left(\mathbf{1}-\mathbf{j}\right)\breve{}\left(\mathbf{1}-\mathbf{j}\right)\right]\end{pmatrix}=$$
$$\begin{pmatrix}\sum_{i}^{n}\rho_{i}^{\ 2} & \sum_{i}^{n}\rho_{i}(1-\rho_{i})\\ \sum_{i}^{n}(1-\rho_{i})\rho_{i} & \sum_{i}^{n}(1-\rho_{i})^{2}\end{pmatrix}\grave{}=\begin{pmatrix}\sum_{i}^{n}\Upsilon_{1}\Upsilon_{1} & \sum_{i}^{n}\Upsilon_{i1}\Upsilon_{i2}\\ \sum_{i}^{n}\Upsilon_{i2}\Upsilon_{i1} & \sum_{i}^{n}\Upsilon_{2}\Upsilon_{2}\end{pmatrix}=\begin{bmatrix}\Upsilon_{1}\grave{}\\ \Upsilon_{2}\grave{}\end{bmatrix}\begin{bmatrix}\Upsilon_{1} & \Upsilon_{2}\end{bmatrix}\grave{}=\mathbf{P}\grave{}\mathbf{P}$$

Then, the joint solution for fixed and random effects in EM-REML is given by:

$$\begin{bmatrix}\hat{\boldsymbol{\mu}}\\ \hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}\end{bmatrix}=\begin{bmatrix}\mathbf{P'P} & \mathbf{P'X} & \mathbf{P'Z}\\ \mathbf{X'P} & \mathbf{X'X} & \mathbf{X'Z}\\ \mathbf{Z'P} & \mathbf{Z'X} & \mathbf{Z'Z}+\dfrac{\sigma^2}{\sigma_a^2}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{P'y}\\ \mathbf{X'y}\\ \mathbf{Z'y}\end{bmatrix}$$

where

$$\hat{\mathbf{P}}=\begin{bmatrix}\Upsilon_{11}=\dfrac{\pi P_{11}}{\pi P_{11}+(1-\pi)P_{12}} & \Upsilon_{12}=\dfrac{(1-\pi)P_{12}}{\pi P_{11}+(1-\pi)P_{12}}\\ \Upsilon_{21}=\dfrac{\pi P_{21}}{\pi P_{21}+(1-\pi)P_{22}} & \Upsilon_{22}=\dfrac{(1-\pi)P_{22}}{\pi P_{21}+(1-\pi)P_{22}}\\ \Upsilon_{31}=\dfrac{\pi P_{31}}{\pi P_{31}+(1-\pi)P_{32}} & \Upsilon_{32}=\dfrac{(1-\pi)P_{32}}{\pi P_{31}+(1-\pi)P_{32}}\\ \vdots & \vdots\\ \Upsilon_{n1}=\dfrac{\pi P_{n1}}{\pi P_{n1}+(1-\pi)P_{n2}} & \Upsilon_{n2}=\dfrac{(1-\pi)P_{n2}}{\pi P_{n1}+(1-\pi)P_{n2}}\end{bmatrix}$$

which is equivalent to a non-linear estimator for $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Thus, as described in (Xu, 2013) (chapter 7), the EM likelihood solutions for mixture models could be obtained as regular system of equations as in classical linear models.

An alternative is to use the stochastic EM (SEM) where the **J** matrix is sampled from a Bernoulli distribution with probability ρ and the solutions for fixed and random effects are taken as those in classical mixed models and the orthogonal properties of **j** and **1-j** are ensured . The consistency of the estimator was tested and compared with mixtools library available in R. We showed that our REML method is more efficient that ML methods. As given on tables S1 and S2:

Table S1- Comparing our proposed model with the mixtool considering 1000 runs on data simulation.

	μ_1	μ_2	π	σ^2
REML/Mixture	4.6615	10.1322	0.2986	1.2028
Mixtools	6.4140	9.3386	0.3987	1.8731
True values	5.0000	10.0000	0.3000	1.4142

Table S2- Comparing our proposed model with the mixtool example data available in the library.

	μ_1	μ_2	π	σ^2
REML/Mixture	54.0019	80.4695	0.361667	5.32482
Mixtools	54.6161	80.0917	0.3609	5.869

The R code for the simulation is given below:

```
########## comparing mixtool with our model#####

###### simulation #######

for(i in 1:nsim){


  mix_data <- rnormmix(n=n, lambda=lamb, mu=med, sigma=sig)###generate the data


  ours<-GMM(y=mix_data)


  out<-normalmixEM(mix_data, arbvar = FALSE, epsilon = 1e-04,
                   ECM = FALSE)


  meds[i,]=c(ours[[2]],out$mu)
```

```
  mix[i,]=c(ours[[1]], out$lambda)
  sig2[i,]=c(ours$sigma, out$sigma[1])
}

mmed=matrix(c(apply(meds,2, mean),med), ncol=2, byrow = TRUE)
mixm=matrix(c(apply(mix,2, mean)[c(1,3)],lamb[1]))
sig2m=matrix(c(apply(sig2,2, mean), sig[1]), nrow = 3)
resul=cbind.data.frame(mmed, mixm, sig2m)


colnames(resul)=c("M1","M2","pi","sig2")
rownames(resul)=c("ours","Mixtools", "simu")


write.csv(resul,"mixtool.csv")

####### mixtool data ######
rm(list=ls())
data(faithful)
attach(faithful)

GMM<-function(y, mu, ve, conv.crit){
```

```
y=y

y=as.matrix(y)

n=length(y)


P=matrix(0.5,n,2)


if(missing(ve)){ve=var(y)/4}


if(missing("mu")){

  mu=c(0,0)

  mu[1]=quantile(y,0.4)

  mu[2]=quantile(y,0.6)

}


if(missing("conv.crit")){conv.crit=1e-04}


P1=dnorm(y,mu[1],sqrt(ve))

P2=dnorm(y,mu[2],sqrt(ve))

P[,1]=P1/(P1+P2)

P[,2]=P2/(P1+P2)



iter=0
```

```
maxiter=300

repeat{
  W1=P
  JJ=crossprod(P,P)
  Jy=crossprod(P,y)
  C1=JJ


  C2=solve(C1)
  sol=C2%*%Jy

  m1=sol[1]
  m2=sol[2]


  ma=m1
  mb=m2

  pred=P%*%c(m1,m2)
  e=y-pred

  dd=W1%*%C2%*%t(W1)
```

```
  ve1= (t(e)%*%e+sum(diag(dd*c(ve))))/n ####duvida aqui



  dif=max(abs(ve-ve1))

  ve=ve1

  iter=iter+1

  pi=sum(P[,1])/n

  P1=pi*dnorm(y,ma,sqrt(ve))

  P2=(1-pi)*dnorm(y,mb,sqrt(ve))

  P[,1]=P1/(P1+P2)

  P[,2]=P2/(P1+P2)

  ###print(c(m1,m2,iter))

  ##if (iter>maxiter) dif=0

  if((dif<var(y)*conv.crit)|(iter==maxiter)) break

}

colnames(P)<-c("prob_subP1","prob_subP2")
```

```
  prop<-c(pi, (1-pi))
  med<-c(m1,m2)

  saida<-list(lmabda=unlist(prop), mu=unlist(med), sigma=unlist(sqrt(ve)))
  return(saida)
}

out1<-normalmixEM(waiting, arbvar = FALSE, epsilon = 1e-03)
out2<-GMM(waiting)

summary(out1)
out2
```

REFERENCES

Bishop, C.M. 2006. Periodic Variables. Pattern Recognit. Mach. Learn. 1.

Sorensen, D., and D. Gianola. 2002. Likelihood, Bayesian and MCMC methods in quantitative genetics. Springer-Verlag Inc, Berlin; New York.

Xu, S. 2013. Principles of statistical genomics. Springer.

Preliminary version of the article written according to the standard of the scientific journal Scientific Reports. The editorial board of the journal can recommend changes in order to adjust it to its own style.

Genomic prediction in *Coffea canephora* using Gaussian Mixture Models

Indalécio Cunha Vieira Júnior[1*], Flávia Maria Avelar Gonçalves[1], Alan Carvalho Andrade[2] and Márcio Balestre[3]

[1]Department of Biology, Federal University of Lavras, Lavras, MG, Brazil
[2]Brazilian Agricultural Research Corporation (EMBRAPA), Lavras, MG, Brazil
[2]Department of Statistics, Federal University of Lavras, Lavras, MG, Brazil
*Corresponding author: indasjunior@hotmail.com

Abstract

Recent studies have shown the potential of genomic selection (GS) in increasing the genetic gain per unit of time in *Coffea spp.*. Biennial growth behavior is a well-known characteristic of these species. It imposes great challenges for breeders to select coffee bean progenies and, potentially, it can generate strong bias on estimated breeding values (EBV) for genomic selection. Therefore, the aim of this study was to propose a gaussian mixture genomic selection (GMGBLUP) model and compare its efficiency with the genomic best linear unbiased prediction (GBLUP) model in terms of parameter estimates and prediction accuracy. One thousand three hundred nineteen robusta coffee (*Coffea canephora*) individuals were genotyped using Coffee Axiom chip – 26K and their coffee beans data were evaluated in three harvests or years. The models were also compared using two simulated data sets: data set 1, which considers no genetic control for biennial growth; and data set 2, which assumes that biennial growth is controlled by one gene. Different values of heritability and biennial growth intensity (subpopulation mean difference) were simulated for both data sets. Specifically for data set 1, distinct levels of subpopulation mixture (pi)were also simulated. For real data, the GBLUP model showed higher efficiency for prediction accuracy. However, GMGBLUP generated higher values for heritability estimates. For simulated data sets, the same real data pattern was observed on low biennial intensity scenarios, independently of the mixture parameter (pi). For higher subpopulation mean difference, the GMGBLUP model was superior and it tended to be more efficient on pi values near to 0.5. On

data set 2, GMGBLUP was superior or equal to GBLUP in almost all scenarios. The results suggest that GMGBLUP could be considered as an alternative for genomic prediction in C*offea* genus, especially for species with strong biennial growth behavior.

Keywords: genomic selection, coffee bean breeding, biennial growth.

Introduction

Biennial growth patterns have been frequently reported in coffee beans (Carvalho, 1988; Rodrigues et al., 2014) and other economically important fruit-tree species (Monselise and Goldschmidt, 1982; Guitton et al., 2011; Durand et al., 2013). This phenomenon is characterized by strong variation on the yield of the individuals through the years and in *Coffea* species, it is undesired for farmers and industries due to difficulties in planning the following season and fluctuations on profit. Especially for breeders, these alternations in production impose challenges on progeny selection and on parameter estimates.

Many authors have stressed that this biennial phenomenon affects accuracy to select the best genotypes and some strategies have been proposed to overcome the problem. One attempt is to group the data on biennials (consider the pair of consecutive years) (Oliveira et al., 2011). Accordingly, it would be possible to reduce the discrepancy between the considered years and perform standard statistical procedures. Linear mixed models have become popular in plant breeding to analyze multi-year data and some researchers have proposed to model the residual covariance matrix and the genetic covariance matrix over harvests to capture the serial correlations between successive observation for an individual (Andrade et al., 2016; Pereira et al., 2018). According to them, mixed models were more efficient for parameter estimates, prediction error variance of genotypic values, rankings and coincidence index in selecting the best progenies.

Recently, a different approach considering the biennial growth as latent variable was proposed. In this case, a Gaussian mixture model was used to capture the physiological stage of each individual to correct it in the model (Vieira Junior et al., 2019). The authors modeled biennial growth considering a mixture of two gaussians with different means and the biennial growth as latent variable. They reported a significant increase in the heritability estimates and a higher expected genetic gain using the new approach when compared with previous proposals.

Traditionally, breeders have relied only on phenotypic information to make genetic progress. The technological development in the last years allowed the use of dense genotyping information which opened the doors to genomic selection (GS). Meuwissen et al., 2001 proposed to apply the complete available genomic information to predict the genetic merit of individuals. Nowadays, this is known as GS. Many studies in the literature have addressed the use of this methodology in different plant species (Robertsen et al., 2019; Tsai et al., 2020). Most of them have achieved good results in applying GS techniques to increase the efficiency of plant breeding programs, especially in perennial species, where it is possible to greatly reduce the time per cycle (McClure et al., 2014; Ferrão et al., 2017; Kwong et al., 2017; Sousa et al., 2018; Stejskal et al., 2018).

There are few studies attempting to apply GS in the coffee bean and the results are promising. Genomic selection enabled early selection in *Coffea arabica* progenies allowing a reduction of 50% in the cycle time (Sousa et al., 2018). However, the authors mentioned that most of the analyzed traits showed high complexity and low genomic heritability.

The models tested until now for GS in genus *Coffea* and other species showing biennial growth have not considered this phenomenon for training the markers. For phenotypic data, it was shown that ignoring this phenomenon could cause strong bias on variance components and on BLUPs estimates (Vieira Júnior et al., 2019). Therefore, we hypotheses that ignoring the effects of this phenomenon during the marker training process can impose strong bias and decrease the prediction accuracy of GS. This paper is an extension of our previous work where we proposed gaussian mixture models to circumvent the biennial growth problem in coffee bean progenies (Vieira Júnior et al., 2019). The objective here is to propose and study a gaussian mixture model for genomic prediction in *Coffea canephora* progenies.

Material and Methods

Real data

The genotypes used in this study are from EMBRAPA Cerrados (Planaltina -DF, Brazil). Coffee beans yield from 1319 individuals, measured in liters, was obtained in three consecutive harvests or years (2012, 2013 and 2014). There was no replication in this experiment and each

individual was a plot. Each individual was originated from a "pool" of seeds harvested in a field of open pollination with 48 parents.

Those individuals were genotyped using Coffee Axiom chip – 26K platform developed and adapted for *C. canephora* (Andrade et al., 2017). After the quality control, there were 16685 single nucleotide polymorphic markers (SNPs) with minor allele frequency (MAF) and *call* rate equal to or higher than 1% and 90%, respectively.

Genomic prediction models

Two different genomic prediction models were compared: GBLUP and Gaussian Mixture GBLUP (GMGBLUP).

Gaussian Mixture GBLUP (GMGBLUP)

For this model, the biennial effect (physiological state) was modeled as a latent variable using a gaussian mixture model. Taking the latent variable as missing information, the following linear model was used for each harvest analysis:

$$\mathbf{y} = \mathbf{J}\boldsymbol{\mu} + \mathbf{Z}\mathbf{u} + \mathbf{e} \qquad (1)$$

where $\mathbf{y}_{(n\times 1)}$ is the phenotypic observation vector and $\boldsymbol{\mu}_{(2\times 1)}$ is a fixed effect vector (means related to biennial status). $\mathbf{u}_{(n\times 1)}$ is the random effect (genotype) vector, and $\mathbf{e}_{(n\times 1)}$ is the error vector. $\mathbf{J}_{(n\times 2)}$ is the missing Bernoulli random variable related to the biennial status. Each element of the **J** matrix (p_{il}), in this work, will be replaced by its expectation i.e., the probability that the i^{th} observation has been taken from the l^{th} biennial state as described by Vieira Júnior et al., 2019. $\mathbf{Z}_{(n\times n)}$ is the random effect incidence matrix (genotypes). Because the matrix for the two-year means ($\boldsymbol{\mu}$) is unknown, the expected Bernoulli variable was used as an indicator of the genotype stage in the mixture.

As demonstrated by Vieira Júnior et al., 2019, since the latent Bernoulli random variable is unknown, it is replaced by its expected value of the complete data likelihood. In other words, $\mathbf{s}_{n\times 1} \sim Bernoulli(p_i)$; therefore, $E(s_i = 1) = \rho_i$, where ρ_i is the i^{th} element of J and represents the expectation of an individual assuming any state in the mixture.

The following assumptions were made for random vectors:

$$\mathbf{u} \sim N(\mathbf{0}, \sigma_g^2\mathbf{G})$$

$$\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2\mathbf{I})$$

Where σ_g^2 and σ_e^2 are the genetic and residual variance, respectively. **I** is an identity matrix. **G** is the VanRanden genomic kinship matrix as described as follows:

$$\mathbf{G} = \frac{\mathbf{WW}'}{2\sum p_i(1-p_i)}$$

where **W** is a 0, 2 and 1 minor allele counts centralized matrix with rows as individuals and columns as markers, and p_i is the allele frequency of the i^{th} marker (VanRaden, 2008). For this method, the function "mixed.solve" implemented on rrBLUP package was used (Endelman, 2011).

Given the above assumptions, the observed data likelihood can be given by:

$$\mathbf{y} \mid \pi, \boldsymbol{\mu}, \mathbf{u}, \sigma_e^2 \sim \pi N(\boldsymbol{\mu}_1 + \mathbf{Zu}, \sigma_e^2\mathbf{I}) + (1\text{-}\pi)N(\boldsymbol{\mu}_2 + \mathbf{Zu}, \sigma_e^2\mathbf{I}) \qquad (2)$$

However, in an expectation-maximization (EM) algorithm, the observed data (y) and the missing information (**u**, *s*) must be jointly modeled using the expectation of the complete log-likelihood (Sorensen and Gianola, 2007), whose objective function is given by:

$$E_{\mathbf{u},s|y,\sigma_e^2}[\mathbf{y}, s \mid \pi, \boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{u}, \sigma_e^2] = \sum_{l=i}^{2} E_{\mathbf{u}|y,\sigma_e^2}\left[\log p(\mathbf{y} \mid \mathbf{j}_l\boldsymbol{\mu}_l + \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}, \sigma_e^2\mathbf{I}) + I(s_l = 1)\log \pi_l\right] + E_{\mathbf{u}|y,\sigma_e^2} \log p(\mathbf{u} \mid 0, \sigma_g^2\mathbf{G})$$

(3)

where $E_{\mathbf{u}|y,\sigma_e^2}$ is the expectation in relation to the random effect of genotypes, σ_g^2 and σ_e^2 are the genetic and residual variance, respectively, π is the unknown mixture parameter, μ_1 and μ_2 are scalars representing the means related to physiological states 1 and 2, and J is the expectation of the $\mathbf{s}_{n\times 1}$ indicator binary vector relating each mean to its subpopulation. In the model described above, π is the a priori probability of genotypes being in the high- or low-production stage, which was assumed to be unknown in the present study. For REML estimates of variance

components, the expectation must be taken in relation to random and fixed effects as already shown by Vieira Júnior et al., 2019.

GBLUP

This model was used for each harvest and corresponds to classical GBLUP (Habier et al., 2007; VanRaden, 2008). Therefore, the latent parameter related to biennial growth is absent. It is described as follow:

$$\mathbf{y} = \lambda 1_n + \mathbf{Zu} + \mathbf{e} \qquad (4)$$

where $\mathbf{y}_{(g\times 1)}$ is the phenotypic observation vector, λ is the intercept , $\mathbf{u}_{(n\times 1)}$ is the random effect (genotype) vector, and $\mathbf{e}_{(n\times 1)}$ is the residual vector. $\mathbf{Z}_{(n\times n)}$ is the random effect incidence matrix and 1_n is a vector g-vector of 1´s. The subscript n represents the number of genotypes which is the number of observations. The following distribution assumptions were made for the random effects:

$$\mathbf{u} \sim N(\mathbf{0}, \sigma_g^2\mathbf{G})$$

$$\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2\mathbf{I})$$

$$\mathbf{y} \sim N(\lambda 1_n + \mathbf{Zu}, \sigma_e^2\mathbf{I})$$

Where σ_g^2 and σ_e^2 are the genetic and residual variance, respectively. Matrices **G** and **I** are as described on the previous model.

Prediction accuracy

On simulated data sets, the models were compared between them in terms of prediction accuracy using the k-fold (five-fold) methodology considering the correlation between the estimated BLUP with the simulated breeding values and the simulated phenotype. For the real data set, prediction accuracy was assessed correlating estimated BLUPs with observed phenotypes.

Phenotype Prediction

If the biennial state has genetic control, it is possible to predict this trait and to get a prediction for the phenotype: the subpopulations means plus the genetic effect. Specifically, for the GMGBLUP there was an extra step to get the predicted phenotypes ($\mathbf{y}_{\text{pred}}$). After converging the model (GMGBLUP), the logit of the expected values of the **J** matrix was used as a phenotype and a GBLUP model was run to get an estimate of the biennial status for the unobserved genotypes. It was as follows:

$$\log\left(\frac{\mathbf{p}}{\mathbf{1-p}}\right) = \lambda 1_n + \mathbf{Zl} + \mathbf{e} \quad (5)$$

Where $\mathbf{p}_{gx1}$ is a vector taken from the expectation of the **J** matrix, λ is the intercept, $\mathbf{l}_{nx1}$ is the random effect (genotype) vector, and $\mathbf{e}_{(n\times 1)}$ is the residual vector. $\mathbf{Z}_{(n\times n)}$ is the random effect incidence matrix and 1_n is a vector of 1´s. The subscript n represents the number of genotypes which is the number of observations. The following distribution assumptions were made for the random effects:

$$\mathbf{l} \sim N(\mathbf{0}, \sigma_g^2\mathbf{G})$$

$$\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2\mathbf{I})$$

$$\log\left(\frac{\mathbf{p}}{\mathbf{1-p}}\right) \sim N(\lambda 1_n + \mathbf{Zl}, \sigma_e^2\mathbf{I})$$

where σ_g^2 and σ_e^2 are the genetic and residual variance, respectively. Matrices **G** and **I** are as described on previous models.

The phenotypic prediction ($\mathbf{y}_{\text{pred}}$) was as follow:

$$\mathbf{y}_{\text{pred}} = \mathbf{L\mu} + \mathbf{Zu} \quad (6)$$

Where: $\mathbf{y}_{pred}$ is the vector of predicted phenotypes, L_{nx2} is the matrix of predicted biennial state estimated in the second stage with columns **p** and (**1-p**), $\boldsymbol{\mu}_{2x1}$ is the vector of subpopulation means estimate (estimated with GMGBLUP) and **Z** incidence matrix for random effects, as described before and **u** is the vector of EBLUPs estimated by GMGBLUP model. The vector **p** was obtained as follow:

$$\mathbf{p} = \frac{\mathbf{1}}{\mathbf{1} + e^{-\mathbf{Zl}}} \quad (7)$$

For the GMGBLUP model, the mixture of two subpopulation was evaluated using the bimodality test according to Holzmann and Vollmer (2008) following the same steps described by Vieira Júnior et al. (2019).

Simulation Study

Data set 1 – No genetic control for physiological state

In this data set, sixteen scenarios were evaluated. Each one varying the degree of mixture of the subpopulations, namely the mixture parameter (pi), heritability (h^2) and biennial growth levels. The biennial growth levels were taken as the ratio between the higher and lower subpopulation means. The lower mean was fixed at five and the higher subpopulation mean assumed values: 10; 15; 20 and 30. Thus, the biennial levels were (from softer to stronger) 0.5; 0.33; 0.25 and 0.17. There were two heritability levels: 0.3, 0.8 and two mixture parameters: 0.2 and 0.5.

In all scenarios, the phenotypes of 119 individuals were simulated. The genotypic data was the same used by Ferrão et al. (2018) from the "intermediate population" including the quality control for SNPs. A total of 100 markers were randomly selected as causal *loci* and their effects (l) were independently sampled from a normal distribution, *i.e.* $\mathbf{l} \sim N(0, \mathbf{I}_n)$. Here, **I** is an identity matrix and the subscript "n" is the number of *loci* controlling the trait. From these values, the true parameter values were obtained, namely breeding values (**TBV**) and σ_g^2.

The phenotypic values were simulated as

$$\mathbf{y} = \mathbf{J\mu} + \mathbf{Za} + \mathbf{e} \qquad (8)$$

where $\boldsymbol{\mu}$ is the vector of means of each subpopulation indicating the physiological state (its values varied according to the scenario), $\mathbf{J}$ is the matrix of 0 and 1 indicating the physiological state (it varied according to the scenario in order to meet the mixture parameter), **a** is the vector of **TBV** as described above, and **e** represents the residual effects vector, which were sampled from a Gaussian distribution, *i.e.*, $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, where $\sigma_e^2 = \frac{1-h^2}{h^2}\sigma_g^2$, h^2 is the heritability.

A trait controlled by 100 genes showing only additive effects was assumed. Markers were randomly sampled as causal *loci* and each scenario was run 100x varying the sampled markers. This process was used to compare the models in terms of parameter estimation. Therefore, all individuals were kept on a data set to estimate the genetic variance (gv), heritability (h^2) and the EBLUPs.

In order to verify the prediction ability of the tested models, we performed a cross-validation using the k-fold (five-fold) method (Crossa et al., 2010). For the GMGBLUP, there was a second stage to estimate the biennial state of the individuals belonging to the validation set and obtain the predicted phenotypes, as described for real data.

Data set 2 – Genetic control for physiological state

In order to understand how the prediction prediction accuracy of the models can be influenced in the biennial growth (or physiological state) genetic control, a second data set was simulated assuming that the biennial growth was controlled by one gene. For that purpose, a homozygous marker was chosen along the individuals. The genetic allele effect for it was changed in order to produce a different biennial growth in the same intensity as mentioned before, that is 0.5; 0.33; 0.25 and 0.16. For this data set, a fixed mixture parameter (pi) was considered, equal to 0.4621 and three heritability values: 0.3, 0.5, 0.8. The other simulated parameters such as heritability and genetic variance were the same as in data set 1 and the phenotypes were obtained as described in equation (4).

RESULTS

Real data

Contrasts of subpopulation means, estimated using the GMGBLUP, were significant in all harvests (Table 1) suggesting the presence of biennial growth. As expected the biennial intensity was lower when compared with *Coffea arabica* yield data, as reported previously (DaMatta, 2004). For example the highest mean difference found in is this work is about 14 and for *Coffea arabica* yield data the lowest mean difference was about 13 (Vieira Júnior et al., 2019).

Table 1: Estimates for the means, in liters, of subpopulation 1 (m1) and 2 (m2) and mixture parameter (pi) considering the Mixed Mixture Model and the estimates for genetic variance (σ_g^2), residual variance (σ_e^2) and heritability (h^2) for the two models used in this work. Estimates obtained using the real data.

	Gaussian Mixed Mixture Model - GMGBLUP							**GBLUP**		
Harvest	**m1**	**m2**	**Contrast***	**pi**	σ_g^2	σ_e^2	**h^2**	σ_g^2	σ_e^2	**h^2**
1	7.565	13.809	6.2446^{***}	0.532	3.6939	2.7031	0.577	1.4697	11.7298	0.1113
2	6.963	21.044	14.082^{***}	0.956	3.4211	10.868	0.239	5.1667	16.7653	0.2355
3	5.703	16.684	10.9806^{***}	0.673	11.914	7.0996	0.627	8.8384	32.5877	0.2133

*Bimodality test. ***Rejected the Unimodality.

Comparing the studied models between them, residual variance estimates were lower for GMGBLUP in all years, which explains the same pattern observed for heritability estimates. Genetic variance estimates were higher for GBLUP only on harvest 2, which presented the highest mixture parameter (pi) (0.956) and was closer to a homogeneous population. This elucidates why the heritability estimates were similar for both models in this harvest. For harvests 1 and 3, the mixture parameter estimates were near 0.5 and 0.7, respectively (Table 1), reinforcing the evidence for the presence of a hidden variable (or biennial growth).

We used two measures of prediction ability to compare the models: the correlation of the estimated BLUPs with phenotype and specifically for the GMGLUP model the correlation of the predicted phenotypic value (y_{pred}) with the observed ones (Table 2). In all cases, GMGBLUP

showed the lowest prediction accuracy independently of the year (Table 2) for both prediction accuracy method. As expected, there was a substantial increase when considering the predicted phenotype, since we are including the biennial state prediction. These results suggest that directly correlating EBLUPs with phenotype may not be suitable to verify the prediction accuracy when the mixture of subpopulations is present.

Table 2: Prediction ability for GMGBLUP (Gaussian Mixture Model) and GBLUP, considering the correlation between predicted genomic value on the validation population and observed phenotypic values (rây), the correlation between predicted genomic values and the predicted phenotypes values (rpredy) in the training population. Estimates obtained using the real data set.

GMGBLUP			
	Harvest 1	**Harvest 2**	**Harvest 3**
rây	0.0932	0.3062	0.2357
rpredy	0.1523	0.3676	0.2703
GBLUP			
rây	0.1948	0.3891	0.3025

Simulated data

Data set 1 – No genetic control for physiological state

Overall, the GMGBLUP estimated genetic variance closer to the simulated values and the GBLUP model tended to overestimate it, especially for higher biennial intensity scenarios and mixture parameter (pi) equal to 0.5 (Figure 1). The GMGBLUP tended to overestimate the heritability in scenarios where the simulated value for this parameter was low, which shows that this model tends to underestimate residual variance in these conditions (Figure 1). The GBLUB underestimated it as the ratio lower/higher subpopulation mean decreased. The only exception was for low heritability and pi equal to 0.5 scenario. For higher simulated heritability scenarios, the GMGBLUP generated estimates closer to the simulated value and its efficiency increased for higher biennial intensity scenarios. In those situations, the GBLUP showed high bias for parameter estimates (Figure 1).

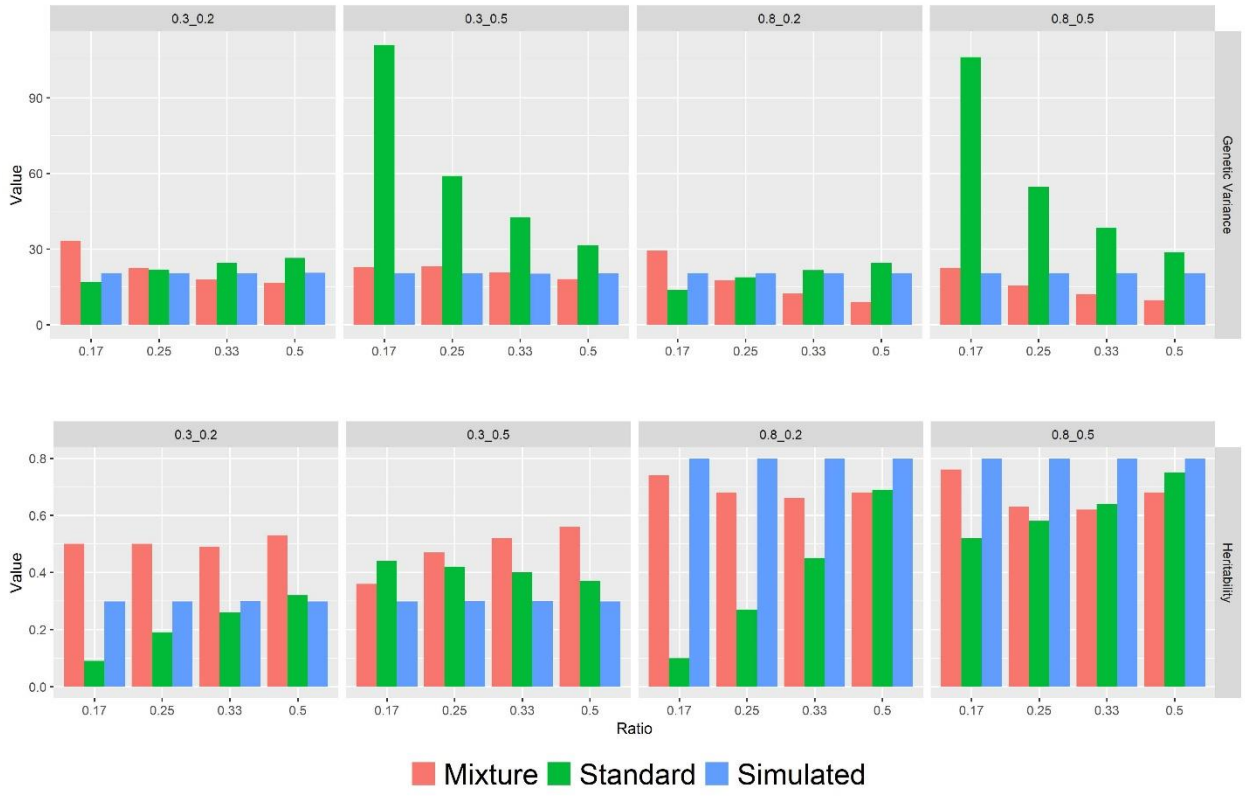

Figure 1: Genetic variance and heritability estimates for simulated data without genetic control of biennial growth (data set 1) considering the models GMGBLUP (Mixture), GBLUP (Standard) and the simulated values (Simulated).

In terms of prediction accuracy for the correlation between the EBLUPs and the simulated phenotype, the GBLUP model showed the highest values in all scenarios (Figure 2). Interestingly the GMGBLUP decreased the prediction accuracy increasing the biennial level. When it is considered the predicted phenotypes in mixture model, accuracy significantly increased and in some scenarios it was near to or higher than GBLUP model (Figure 2).

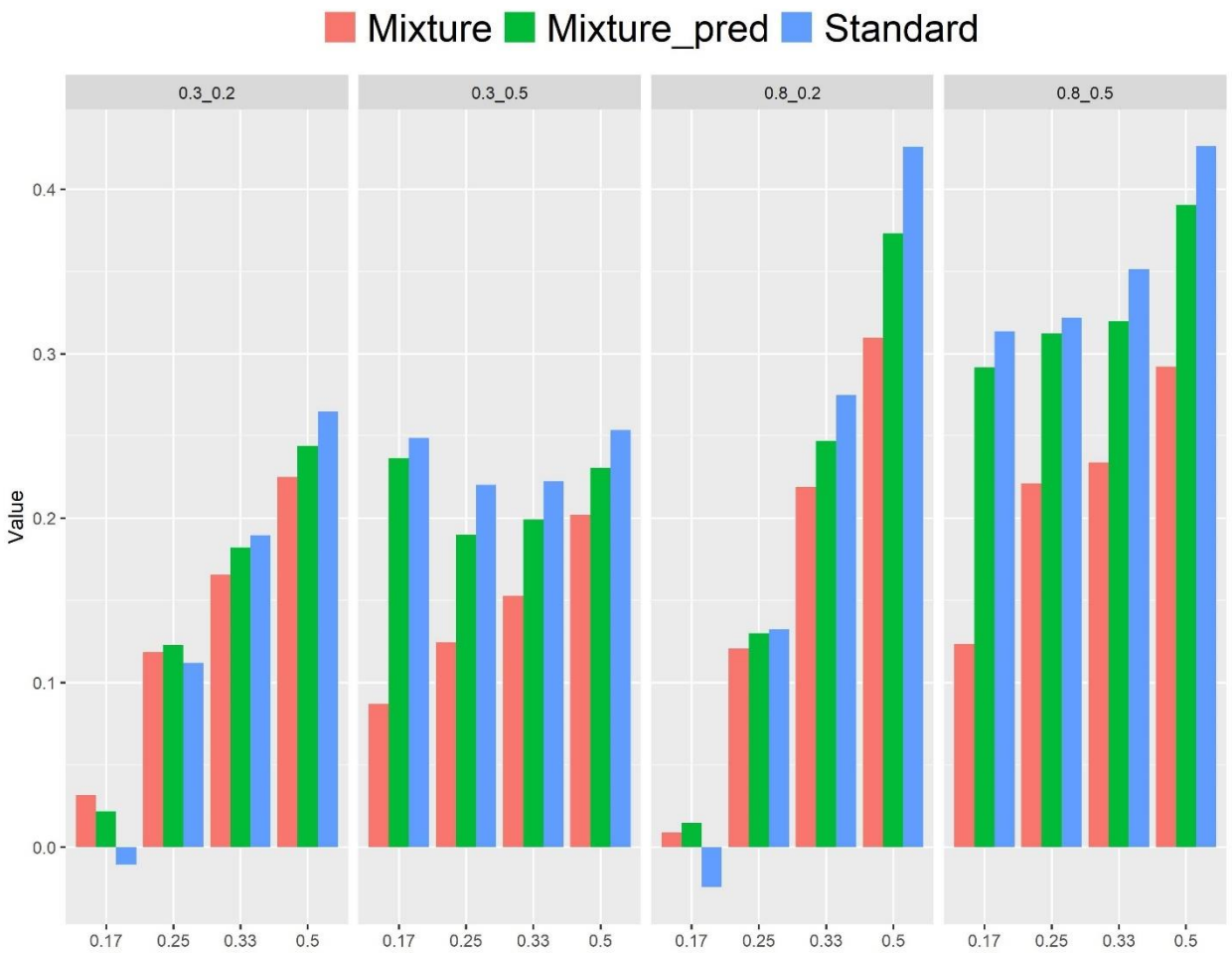

Figure 2: Prediction accuracy considering the correlation of the EBLUPs for GBLUP (Standard), GMGBLUP (Mixture) and the predicted phenotypes by the GMGBLUP (Mixture_pred). Correlation estimated using the simulated phenotypes without genetic control for biennial growth.

When the prediction ability was assessed by correlating the EBLUPs with simulated genetic values (TBV) the GBLUP quickly decreased its prediction accuracy as the biennial intensity increased. In addition, for some scenarios, the GMGBLUP was about three times more efficient in predicting breeding values. This can be observed when the simulated heritability and pi were, respectively equal to 0.8 and 0.5 and low/high mean ratio of 0.17 (Figure 3). These

results suggest that correlating the EBLUPs with phenotypes leads to mistakes when choosing the best model for prediction accuracy.

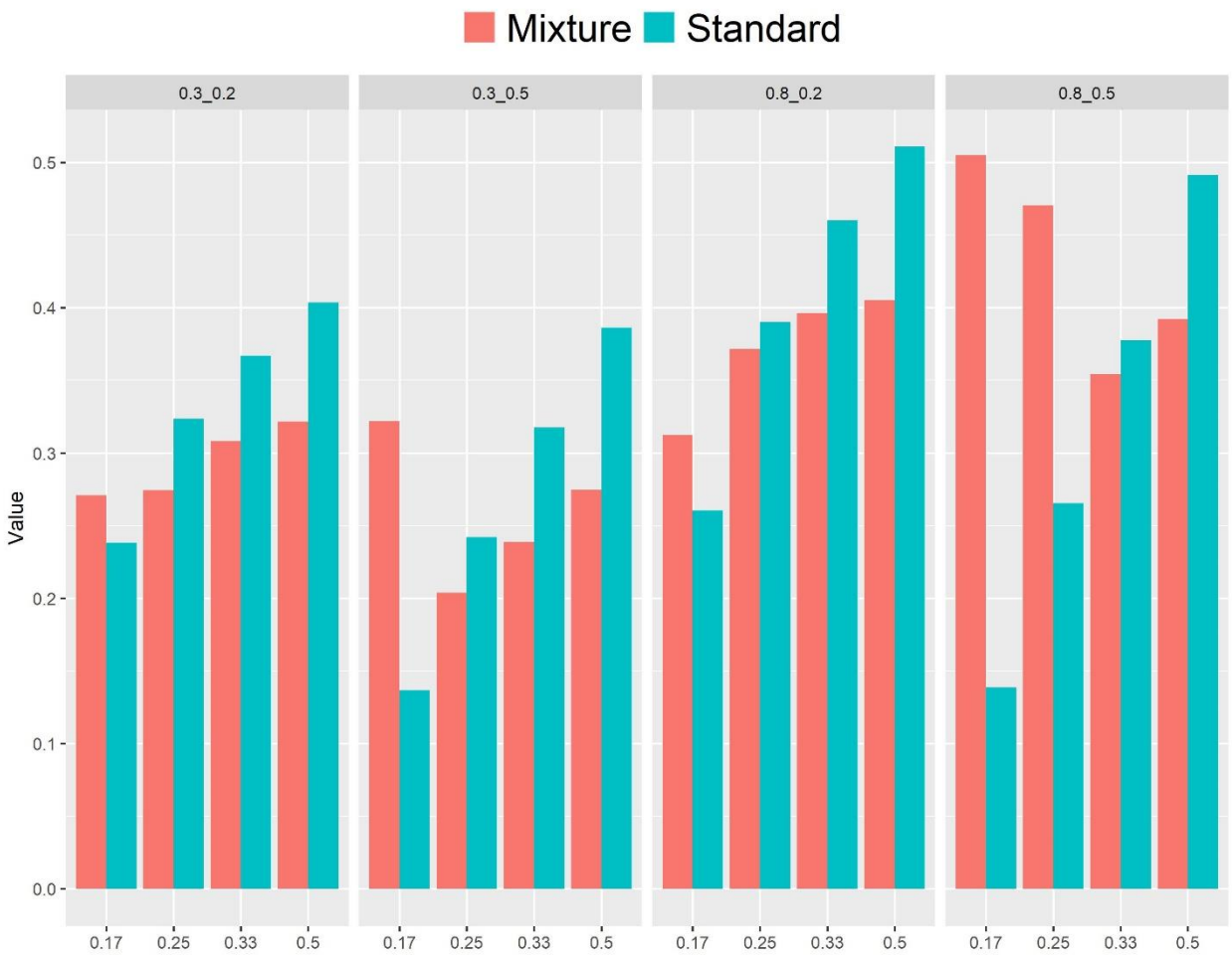

Figure 3: Prediction accuracy considering the correlation of the EBLUPs for GBLUP (Standard) and the GMGBLUP (Mixture). Correlation estimated using the simulated genetic values without genetic control for biennial growth.

Data set 2 – Genetic control for physiological state

In all scenarios, the GBLUP model overestimated the genetic variance. For heritability estimates, the GMGBLUP was more efficient in all cases, except for higher biennial intensity scenarios (Figure 4).

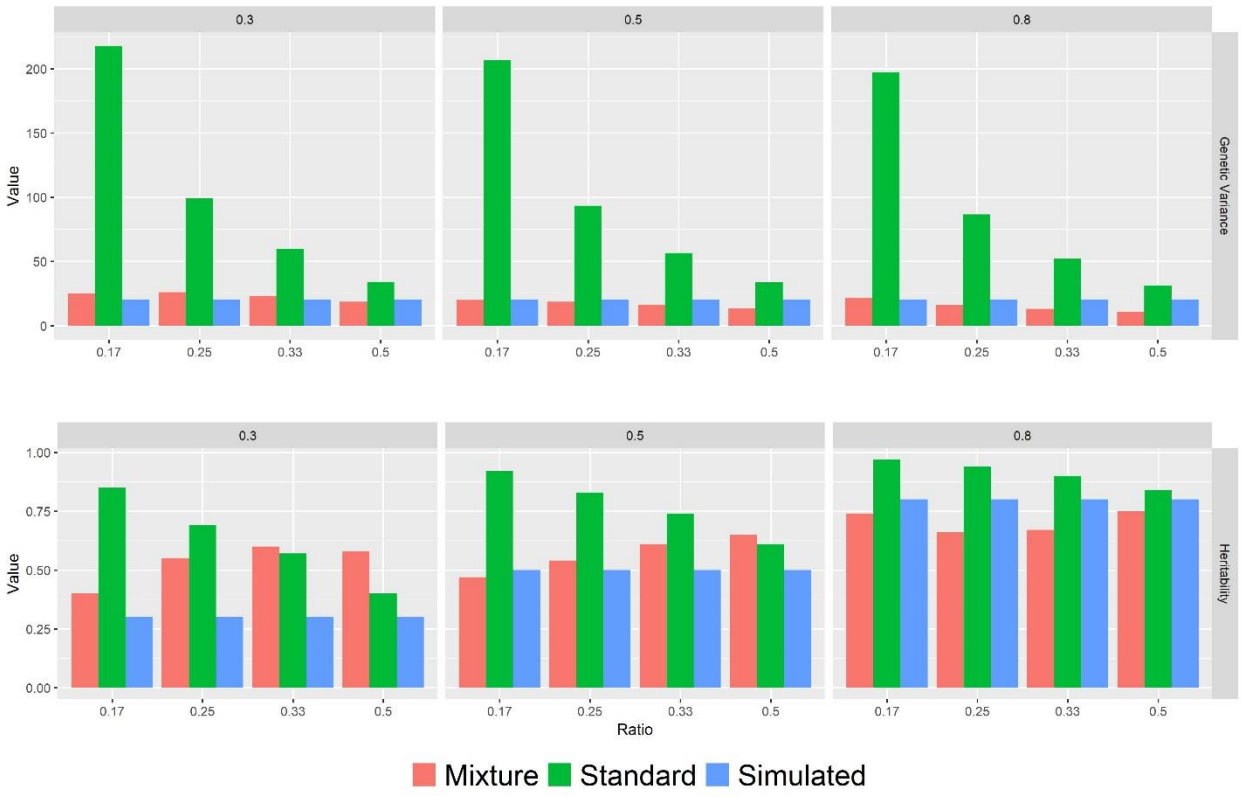

Figure 4: Genetic variance and heritability estimates for simulated data with genetic control of biennial growth (data set 2) considering the models GMGBLUP (Mixture), GBLUP (Standard) and the simulated values (Simulated).

When considering genetic control for biennial status, the GMGBLUP was again worse than the GBLUP for prediction accuracy considering the correlation between simulated phenotypes and EBLUPs. However, for the predicted phenotypes, the GMGLUP was equal to or slightly better than the GBLUP in terms of prediction accuracy. This suggests that the

GMGBLUP was able to predict the biennial status based only on genotypic information (Figure 5) and it also suggests that measuring the performance of mixture models using correlation between phenotypes and EBLUPs may not be the best alternative.

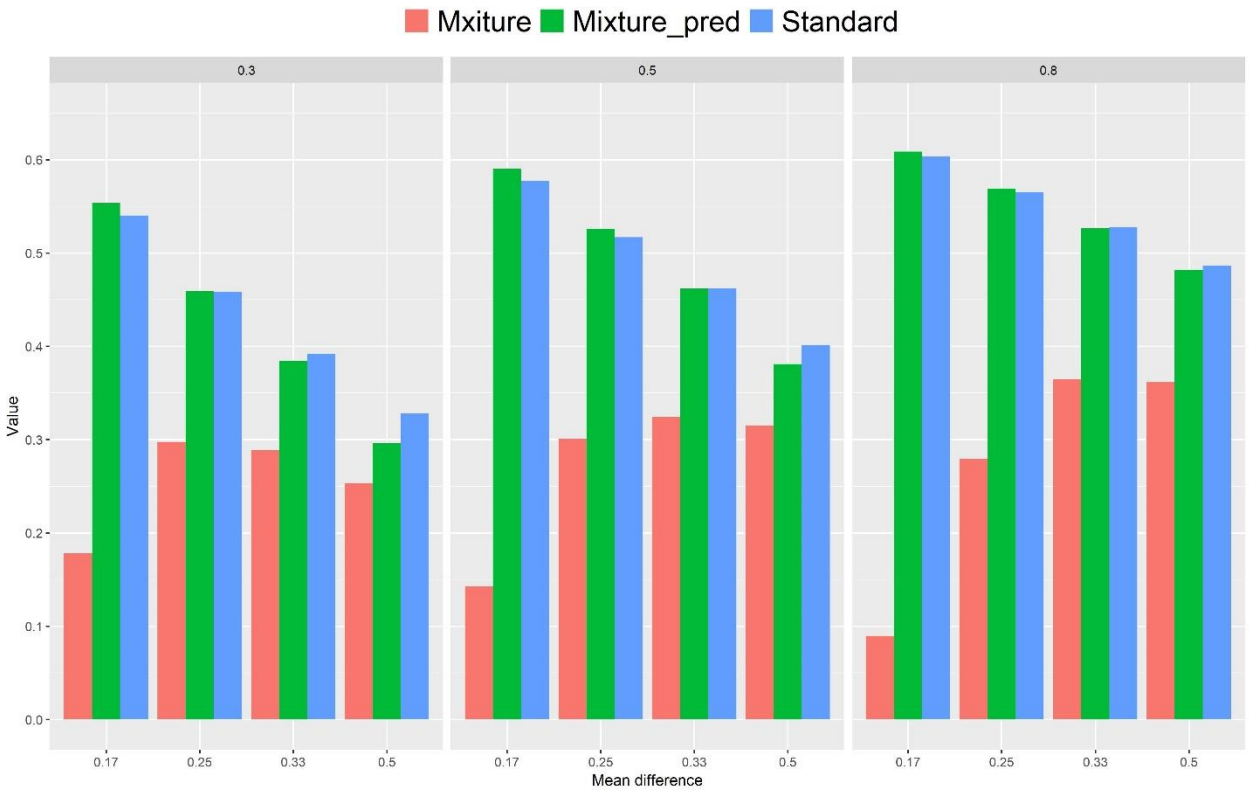

Figure 5: Prediction accuracy considering the correlation of the EBLUPs for GBLUP (Standard), GMGBLUP (Mixture) and the predicted phenotypes by the GMGBLUP (Mixture_pred). Correlation estimated using the simulated phenotypes with genetic control for biennial growth.

For prediction of true genetic value, the GMGBLUP model was more efficient in most of the scenarios and, as expected, its prediction accuracy significantly increased for higher biennial intensity scenarios (Figure 6). The results observed in this data set show that genetic control of

the biennial growth has strong influence on the prediction ability of the genomic prediction models.

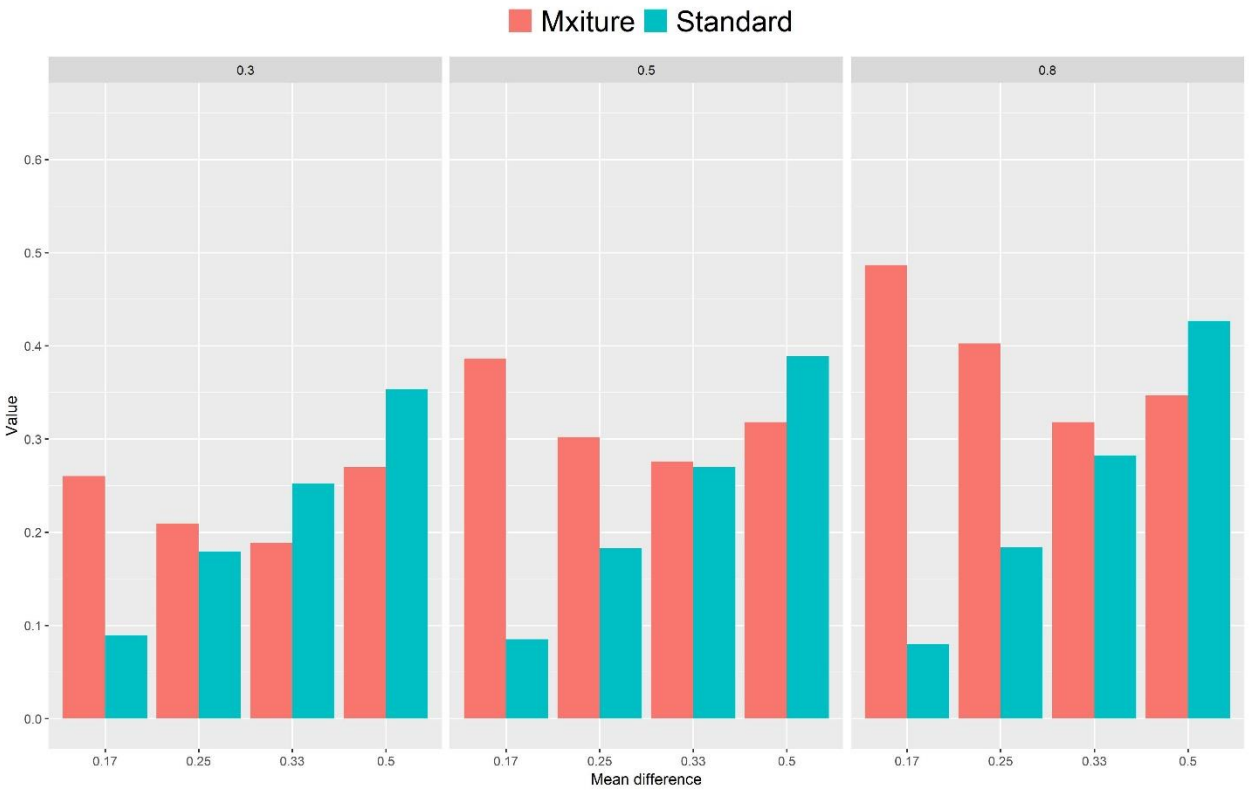

Figure 6: Prediction accuracy considering the correlation of the EBLUPs for GBLUP and the GMGBLUP (Mixture). Correlation estimated using the simulated genetic values with genetic control for biennial growth.

DISCUSSION

According to our hypothesis, the influence of biennial growth in the species of genus *Coffea* can interfere on the prediction accuracy of the genomic selection models. Therefore, the

presence of the phenomenon on the studied population is fundamental. As shown in Table 1, there is strong evidence in favor of the existence of biennial growth in this population.

The GBLUP model was more efficient in predicting phenotypes (Table 2). In all harvests it could deliver the highest correlation between the estimated BLUPS and the phenotypes. These correlations were higher even when the physiological state was considered on the prediction accuracy for the GMGBLUP. Despite the apparent advantage of the GBLUP, the prediction error variance estimates for this model were always higher (Table 2), which implies in higher error for predictions. As shown in previous works (Gianola et al., 2007; Vieira Júnior et al., 2019) , when there is a hidden stochastic process generating the data which is not considered in the model, the variance components are biased and consequently the EBLUPs too. Thus, the error associated with the BLUPs estimates will be higher in the GBLUP model.

It should be emphasized that *Coffea canephora* is an allogamous species that presents gametophytic incompatibility, caused by an S allelic serial in a single locus (Conagin and Mendes, 1961). The individuals from this population were planted on a field without taking this information into account. Probably, some of them were side by side with others that were incompatible between them, which negatively influence the performance of those genotypes. If that is the case, parameter estimates and prediction accuracy will be biased independently of the model.

Gaussian mixture models are able to recognize distinct types of hidden patterns (Murphy, 2012). If this reproductive characteristic is not considered on this model, it will directly influence mixture parameter (pi) and biennial state estimates, which strongly affect the model´s performance and can be more harmful for GMGBLUP comparing with GBLUP.

An important question is if biennial growth (in *Coffea canephora*) has genetic control or its expression is mainly due to environmental variations. As far as we know, there is no research in the literature elucidating the genetic control (if there is any) of this phenomenon on genus *Coffea*. Thus, we simulated two scenarios: no “genetic control” and a single gene controlling the biennial growth. As shown above, for “no genetic control” scenarios the prediction accuracy for Gaussian mixture model (considering the correlation with phenotypes) was lower in all scenarios, except in those with the higher biennial intensity and mixture parameter equal to 0.2 (Figure 2). Correlating EBLUPs with phenotypes has been the standard method to compare genomic selection methods (Montesinos-López et al., 2018; Brauner et al., 2019; Howard et al.,

2019). For gaussian mixture models, this method may not be the best way. As observed before on simulated data, the prediction accuracy estimates for those models is downward biased if the biennial state is not considered (Figures 2 and 5). Correlation estimates presume linearity between two variables (Casella and Berger, 2002), and this is not true when variables come from two different distributions. Actually, the best method to compare prediction ability between models would be correlating EBLUPs with the true breeding value (TBV). However, for real data this is not feasible.

For the simulated data with genetic control, in most of the scenarios, the prediction accuracy for mixture model, considering the predicted values, was equal to or higher than the GBLUP. However, correlations of the EBLUPs estimated by GMGBLUP with the phenotype always generated lower prediction accuracy (Figure 5). These observations suggest that, not only including the effects of a hidden variable (here the biennial status) but also the genetic control of the trait is fundamental for the behavior of the GMGBLUP in terms of prediction and this is in accordance with the genomic prediction theory, since this methodology uses relatedness to predict non-phenotyped individuals (Meuwissen et al., 2001; Habier et al., 2007). The performance of the GMGBLUP was better in most of the scenarios. When correlating the EBLUPs with TBV and for a mean ration equal to or lower than 0.33, it was better than the GBLUP in almost the all the heritability scenarios (Figure 6). When considering no genetic control data set, this happened only with mean ratio of 0.25 or lower, therefore in higher biennial intensity scenarios. These results reinforce the influence of genetic control of biennial growth on prediction accuracy and shows that choice of prediction model is conditioned on the intensity of biennial growth and on its genetic control.

The genetic control of biennial growth is not well understood, we tried to simulate some scenarios in order to verify the behavior of the tested models assuming or not genetic control for this phenomenon. It was expected that, considering a hidden variable on the model, the variance component estimates and the prediction accuracy would be higher independently of the genetic control. However, the results do not confirm that. In order to make good predictions, it is also necessary to predict the biennial state and this task is conditioned to the presence of genetic control of the phenomenon as confirmed by the simulated results.

As pointed out by other authors ignoring that the data originating from a mixture of gaussians can lead to strong bias on the estimates (Henderson, 1975; Gianola et al., 2007; Vieira

Júnior et al., 2019), specially for the variance components. In addition, it is clear from the simulation study that the biennial growth levels (or subpopulation´s means rate) are more important for quality of the estimates (and predictions) than the genetic control of biennial growth by itself.

Previous works have discussed estimation bias and convergence problems of gaussian mixture models under some conditions (Sun and Wang, 2011; Naim and Gildea, 2012; Lourens et al., 2013). Overall, when the mixture components highly overlap themselves themselves and the mixing coefficients (mixture parameters) assume extreme values, the EM algorithm significantly decreases the rate of convergence and tends to stop poor local optima (Xu and Jordan, 1996; Naim and Gildea, 2012). As a consequence, the parameters estimates are biased and the classification accuracy is highly affected. As pointed out by Lourens et al. (2013), if the distributions have large overlap it will be difficult to identify the group membership of observations and to estimate each component´s parameters, in those cases severe bias might result on the estimates.

These observations are in accordance with the obtained results. The GMGBLUP always performed better on extreme biennial intensity scenarios, independently of the genetic control. In those situations, there was the lowest overlapping of the mixture components. Using the index (H) proposed by Hosmer Jr (1973) as a measure of separation between mixture components on simulated data, it is clear that, as we increase the mean difference there, the separation between the two mixtures is larger (Table S1). For example, considering the lowest biennial intensity (mean rate of 0.5) when the heritability is 0.8, H = 2.33 and for the highest, H=11.65. These estimates corroborate the previous results found by other authors about the influence of overlapping in the accuracy of the estimates (Naim and Gildea, 2012; Lourens et al., 2013). In scenarios of higher H values, the GMGBLUP tended to be better.

In the GMGBLUP model, the estimator for random effects is directly dependent on the right classification group, as pointed out in the materials and methods section and showned analytically by Gianola et al. (2007). Intuitively, the bias on model´s estimates becomes larger as this misclassification increases. This is illustrated on figures 1 and 4, the GMGBLUP model estimates more accurately the genetic variance and heritability in scenarios with less overlapping of the two gaussians. Conversely, the GBLUP model consistently increased the estimates bias in those scenarios (Figures 1 and 4) and was the worst model. As discussed before by Gianola et al.

(2007), the regression of genotype on phenotype is not linear on the observations in the presence of the Gaussian mixture. Therefore, when there is reasonable differentiation on the mixture components, standard linear models give less than optimal prediction of genetic effects.

The mean difference is not the only factor that can influence on mixture components overlapping. As the heritability decreases, the residual variance assumes higher proportion on phenotypic variance and the area of gaussians tends to be larger due to increase in parameter´s uncertainty. Mathematically, this can be viewed below by the Hosmer Jr, 1973 equation for mixture disparity:

$$H = \frac{|\mu_1 - \mu_2|}{min(\sigma_1, \sigma_2)}$$

Our results agree with Lourens et al. (2013) and in general for higher heritability scenarios, there were higher prediction accuracy and better parameter estimates for the GMGBLUP .

As shown by other authors, this phenomenon in *Coffea canephora* is not intense as in *Coffea arabica* (Ferrão et al., 2018). As a consequence, the mean of the two subpopulations tends to be closer. This helps to explain the performance of the GMGBLUP compared to GBLUP model for prediction accuracy in the real data. As mentioned before, the information about genetic control of biennial growth is scarce or null in the literature. In any case, our simulated results suggest that independently of the genetic control of this phenomenon, having subpopulations mean ratio equal to or less than 0.15 GMGBLUP model is better, for genomic prediction or genetic parameters estimation. If biennial growth has a genetic control, this information can be recovered by molecular markers and the ability of the model to predict phenotypes is increased (Figure 6).

The advantages of the genomic selection on genus *Coffea* have been studied and discussed before by Ferrão et al. (2018) and Sousa et al. (2018). The biennial growth phenomenon still represents a challenge for breeders and more research is required. Our results support that the GMGBLUP model has a great potential to be applied on species with strong biennial growth behavior, such as *Coffea arabica* and apple trees (Guitton et al., 2011; Durand et al., 2013; Andrade et al., 2017; Vieira Júnior et al., 2019). Thus, considering this phenomenon for genomic prediction in such species is a powerful tool to increase the breeding efficiency. We

believe that more research is required, specially expanding the GMGBLUP to incorporate genotype by environment (GxE) and dominance effects in the case of *Coffea canephora*.

The GMGBLUP model is more efficient than the GBLUP for intermediate to high subpopulation´s mean difference. Specially in this last scenario, it highly improved the prediction accuracy and showed higher statistical efficiency for parameter estimation. Therefore, we believe that it should be considered as an alternative model for genomic prediction in species that show biennial behavior.

REFERENCES

Aitkin, M., and G.T. Wilson. 1980. Mixture models, outliers, and the EM algorithm. Technometrics 22(3): 325–331.

Andrade, V.T., F.M.A. Gonçalves, J.A.R. Nunes, and C.E. Botelho. 2016. Statistical modeling implications for coffee progenies selection. Euphytica 207(1): 177–189Available at https://doi.org/10.1007/s10681-015-1561-6.

Andrade, A.C., O.B. da SILVA JUNIOR, F. CARNEIRO, P. Marraccini, and D. Grattapaglia. 2017. Towards GWAS and Genomic Prediction in Coffee: Development and Validation of a 26K SNP Chip for Coffea Canephora. *In* Embrapa Café-Resumo em anais de congresso (ALICE).

Azodi, C.B., A. McCarren, M. Roantree, G. de los Campos, and S.-H. Shiu. 2019. Benchmarking algorithms for genomic prediction of complex traits. bioRxiv: 614479.

Bacha, C.J.C. 1998. A cafeicultura brasileira nas décadas de 80 e 90 e suas perspectivas. Preços agrícolas 7(142): 14–22.

Balestre, M., P.P. Torga, R.G. Von Pinho, and J.B. dos Santos. 2013. Applications of multi-trait selection in common bean using real and simulated experiments. Euphytica 189(2): 225–238Available at https://doi.org/10.1007/s10681-012-0790-1.

Bernardo, R. 2002. Breeding for Quantitative Traits in Plants. Stemma Press.

Bertrand, B., H. Etienne, C. Cilas, A. Charrier, and P. Baradat. 2005. Coffea arabica hybrid

performance for yield, fertility and bean weight. Euphytica 141(3): 255–262.

Bishop, C.M. 2006. Periodic Variables. Pattern Recognit. Mach. Learn. 1.

Bonomo, P. 2002. Metodologias biométricas para seleção de progênies no melhoramento genético do cafeeiro.

Brauner, P.C., D. Müller, W.S. Molenaar, and A.E. Melchinger. 2019. Genomic prediction with multiple biparental families. Theor. Appl. Genet.: 1–15.

Carvalho, A. 1988. Principles and practice of coffee plant breeding for productivity and quality factors: Coffea arabica. Coffee Agron. 4: 129–165.

Casella, G., and R.L. Berger. 2002. Statistical inference. Duxbury Pacific Grove, CA.

Conagin, C.H., and A.J.T. Mendes. 1961. Pesquisas citológicas e genéticas em três espécies de Coffea: auto-incompatibilidade em Coffea canephora Pierre ex Froehner. Bragantia 20(UNICO): 788–804.

Crossa, J., G. de Los Campos, P. Pérez, D. Gianola, J. Burgueño, J.L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, J. Yan, and others. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186(2): 713–724.

Crossa, J., P. Perez, J. Hickey, J. Burgueno, L. Ornella, J. Cerón-Rojas, X. Zhang, S. Dreisigacker, R. Babu, Y. Li, and others. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity (Edinb). 112(1): 48.

DaMatta, F.M. 2004. Ecophysiological constraints on the production of shaded and unshaded coffee: a review. F. Crop. Res. 86(2–3): 99–114.

Davis, A.P., T.W. Gole, S. Baena, and J. Moat. 2012. The impact of climate change on indigenous arabica coffee (Coffea arabica): predicting future trends and identifying priorities. PLoS One 7(11): e47981.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B: 1–38.

Detilleux, J., and P.L. Leroy. 2000. Application of a mixed normal mixture model for the estimation of mastitis-related parameters. J. Dairy Sci. 83(10): 2341–2349.

Durand, J.-B., B. Guitton, J. Peyhardi, Y. Holtz, Y. Guédon, C. Trottier, and E. Costes. 2013. New insights for estimating the genetic value of segregating apple progenies for irregular bearing during the first years of tree production. J. Exp. Bot. 64(16): 5099–5113.

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4: 250–255.

Ferrão, L.F.V., R.G. Ferrão, M.A.G. Ferrão, A. Fonseca, P. Carbonetto, M. Stephens, and A.A.F. Garcia. 2018. Accurate genomic prediction of Coffea canephora in multiple environments using whole-genome statistical models. Heredity (Edinb).

Ferrão, L.F.V., R.G. Ferrão, M.A.G. Ferrão, A. Francisco, and A.A.F. Garcia. 2017. A mixed model to multiple harvest-location trials applied to genomic prediction in Coffea canephora. Tree Genet. Genomes 13(5): 95.

Fisch, R.D., M. Ragot, and G. Gay. 1996. A generalization of the mixture model in the mapping of quantitative trait loci for progeny from a biparental cross of inbred lines. Genetics 143(1): 571–577.

Gapare, W., S. Liu, W. Conaty, Q.-H. Zhu, V. Gillespie, D. Llewellyn, W. Stiller, and I. Wilson. 2018. Historical Datasets Support Genomic Selection Models for the Prediction of Cotton Fiber Quality Phenotypes Across Multiple Environments. G3 Genes, Genomes, Genet. 8(5): 1721–1732.

Gianola, D., P.J. Boettcher, J. Ødegård, and B. Heringstad. 2007. Mixture models in quantitative genetics and applications to animal breeding. Rev. Bras. Zootec. 36: 172–183.

Gianola, D., J. Øegård, B. Heringstad, G. Klemetsdal, D. Sorensen, P. Madsen, J. Jensen, and J. Detilleux. 2004. Mixture model for inferring susceptibility to mastitis in dairy cattle: a procedure for likelihood-based inference. Genet. Sel. Evol. 36(1): 3.

Grinberg, N.F., A. Lovatt, M. Hegarty, A. Lovatt, K.P. Skøt, R. Kelly, T. Blackmore, D. Thorogood, R.D. King, I. Armstead, and others. 2016. Implementation of genomic prediction in Lolium perenne (L.) breeding populations. Front. Plant Sci. 7: 133.

Guitton, B., J.-J. Kelner, R. Velasco, S.E. Gardiner, D. Chagne, and E. Costes. 2011. Genetic control of biennial bearing in apple. J. Exp. Bot. 63(1): 131–149.

Habier, D., R.L. Fernando, and J.C.M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4): 2389–2397.

Henderson, C.R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics: 423–447.

Holzmann, H., and S. Vollmer. 2008. A likelihood ratio test for bimodality in two-component mixtures with application to regional income distribution in the EU. AStA Adv. Stat. Anal. 92(1): 57–69.

Hosmer Jr, D.W. 1973. A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. Biometrics: 761–770.

Howard, R., D. Gianola, O. Montesinos-López, P. Juliana, R. Singh, J. Poland, S. Shrestha, P. Pérez-Rodr\'\iguez, J. Crossa, and D. Jarqu\'\in. 2019. Joint Use of Genome, Pedigree, and Their Interaction with Environment for Predicting the Performance of Wheat Lines in New Environments. G3 Genes, Genomes, Genet. 9(9): 2925–2934.

Hu, X., and J. Spilke. 2011. Variance--covariance structure and its influence on variety assessment in regional crop trials. F. Crop. Res. 120(1): 1–8.

Jamrozik, J., and L.R. Schaeffer. 2010. Application of multiple-trait finite mixture model to test-day records of milk yield and somatic cell score of Canadian Holsteins. J. Anim. Breed. Genet. 127(5): 361–368.

Klingenberg, C.P., and L.J. Leamy. 2001. Quantitative genetics of geometric shape in the mouse mandible. Evolution (N. Y). 55(11): 2342–2352.

Kwong, Q. Bin, A.L. Ong, C.K. Teh, F.T. Chew, M. Tammi, S. Mayes, H. Kulaveerasingam, S.H. Yeoh, J.A. Harikrishna, and D.R. Appleton. 2017. Genomic selection in commercial perennial crops: applicability and improvement in oil palm (Elaeis Guineensis Jacq.). Sci. Rep. 7(1): 2872.

de los Campos, G., J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, and M.P.L. Calus. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193(2): 327–345.

Lourens, S., Y. Zhang, J.D. Long, and J.S. Paulsen. 2013. Bias in estimation of a mixture of normal distributions. J. Biom. Biostat. 4.

Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Sunderland, MA.

McClure, K.A., J. Sawler, K.M. Gardner, D. Money, and S. Myles. 2014. Genomics: a potential panacea for the perennial problem. Am. J. Bot. 101(10): 1780–1790.

Melchinger, A.E., W. Schipprack, T. Würschum, S. Chen, and F. Technow. 2013. Rapid and accurate identification of in vivo-induced haploid seeds based on oil content in maize. Sci. Rep. 3: 2129.

Meuwissen, T.H., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4): 1819.

Monselise, S.P., and E.E. Goldschmidt. 1982. Alternate bearing in fruit trees. Hortic. Rev. (Am. Soc. Hortic. Sci). 4(1).

Montesinos-López, A., O.A. Montesinos-López, D. Gianola, J. Crossa, and C.M. Hernández-Suárez. 2018. Multi-environment genomic prediction of plant traits using deep learners with dense architecture. G3 Genes, Genomes, Genet. 8(12): 3813–3828.

Murphy, K.P. 2012. Machine learning: a probabilistic perspective. Cambridge, MA.

Naim, I., and D. Gildea. 2012. Convergence of the em algorithm for gaussian mixtures with unbalanced mixing coefficients. arXiv Prepr. arXiv1206.6427.

Oliveira, A.C.B. de, A.A. Pereira, F.L. da Silva, J.C. de Rezende, C.E. Botelho, and G.R. Carvalho. 2011. Prediction of genetic gains from selection in Arabica coffee progenies. Crop Breed. Appl. Biotechnol. 11(2): 106–113.

Patterson, H.D., and R. Thompson. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58(3): 545–554.

Pereira, T.B., J.P.F. Carvalho, C.E. Botelho, M.D.V. de Resende, J.C. de Rezende, and A.N.G. Mendes. 2013. Eficiência da seleção de progênies de café F4 pela metodologia de modelos mistos (REML/BLUP). Bragantia 72(3): 230–236.

Pereira, F.A.C., S.P. de Carvalho, T.T. Rezende, L.L. Oliveira, and D.R.B. Maia. 2018. Selection of Coffea arabica L. hybrids using mixed models with different structures of variance-covariance matrices.

Piepho, H.-P., and T. Eckl. 2014. Analysis of series of variety trials with perennial crops. Grass Forage Sci. 69(3): 431–440.

Redner, R.A., and H.F. Walker. 1984. Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. 26(2): 195–239.

Rena, A.B., M. Maestri, A.B. Rena, E. Malavolta, M. Rocha, and T. Yamada. 1986. Fisiologia do cafeeiro. Inf. agropecuário 11(126).

Rencher, A.C., and G.B. Schaalje. 2008. Linear models in statistics. John Wiley & Sons.

Resende, M.D., and J.B. Duarte. 2007. Precisão e controle de qualidade em experimentos de avaliação de cultivares. Pesqui. Agropecuária Trop. 37(3).

Rodrigues, W.P., H.D. Vieira, D. Barbosa, G.R. Sousa Filho, and F.L. Partelli. 2014. Agronomic performance of arabica coffee genotypes in northwest Rio de Janeiro State. Genet. Mol. Res. 13(3): 5664–5673.

Sakai, E., E.A.A. Barbosa, J.M. de Carvalho Silveira, and R.C. de Matos Pires. 2015. Coffee productivity and root systems in cultivation schemes with different population arrangements and with and without drip irrigation. Agric. water Manag. 148: 16–23.

SAS Institute. 2009. User´s guide: statistics. SAS Institute, Cary.

Schilling, M.F., A.E. Watkins, and W. Watkins. 2002. Is human height bimodal? Am. Stat. 56(3): 223–229.

Sera, T. 2001. Coffee genetic breeding at IAPAR. Crop Breed. Appl. Biotechnol. 1(2).

Setotaw, T.A., E.T. Caixeta, A.A. Pereira, A.C. de Oliveira, C.D. Cruz, E.M. Zambolim, L. Zambolim, and N.S. Sakiyama. 2013. Coefficient of parentage in Coffea arabica L. cultivars grown in Brazil. Crop Sci. 53(4): 1237–1247.

Smith, A.B., J.K. Stringer, X. Wei, and B.R. Cullis. 2007. Varietal selection for perennial crops where data relate to multiple harvests from a series of field trials. Euphytica 157(1–2): 253–

266.

Sorensen, D., and D. Gianola. 2002. Likelihood, Bayesian and MCMC methods in quantitative genetics. Springer-Verlag Inc, Berlin; New York.

Sorensen, D., and D. Gianola. 2007. Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer Science & Business Media.

Sousa, T.V., E.T. Caixeta, E.R. Alkimim, A.C.B. Oliveira, A.A. Pereira, N.S. Sakiyama, L. Zambolim, and M.D.V. Resende. 2018. Early Selection Enabled by the Implementation of Genomic Selection in Coffea arabica Breeding. Front. Plant Sci. 9.

Stejskal, J., M. Lstib\uurek, J. Klápště, J. Čepl, and Y.A. El-Kassaby. 2018. Effect of genomic prediction on response to selection in forest tree breeding. Tree Genet. genomes 14(5): 74.

Sun, H., and S. Wang. 2011. Measuring the component overlapping in the Gaussian mixture model. Data Min. Knowl. Discov. 23(3): 479–502.

Tran, H.T.M., L.S. Lee, A. Furtado, H. Smyth, and R.J. Henry. 2016. Advances in genomics for the improvement of quality in coffee. J. Sci. Food Agric. 96(10): 3300–3312.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91(11): 4414–4423.

Vieira, I. Cunha, C. da Silva, J.J. Nuvunga, C.E. Botelho, F.M. Avelar Gonçalves, and M. Balestre. 2019. Mixture Mixed Models: Biennial Growth as a Latent Variable in Coffee Bean Progenies. Crop Sci.

Waller, J.M., M. Bigger, and R.J. Hillocks. 2007. Coffee pests, diseases and their management. CABI.

Xu, S. 2013. Principles of statistical genomics. Springer.

Xu, L., and M.I. Jordan. 1996. On convergence properties of the EM algorithm for Gaussian mixtures. Neural Comput. 8(1): 129–151.

Xu, S., D. Zhu, and Q. Zhang. 2014. Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proc. Natl. Acad. Sci. 111(34): 12456–12461.

SUPPLEMENTARY MATERIAL

Table S1: Estimates of mixture disparity (**H**) considering different heritabilities and Biennial intensity (mean differences).

Heritability	Biennial intensity	H
0.3	0.5	0.7629
	0.17	3.8147
0.8	0.5	2.3308
	0.17	11.654