CRISTIANE BOTELHO VALADARES

RANDOM FOREST QUANTÍLICO APLICADO EM ESTUDOS DE SELEÇÃO GENÔMICA

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

Orientador: Moysés Nascimento

Coorientadores: Ana Carolina C. Nascimento
Camila Ferreira Azevedo
Weverton Gomes da Costa

VIÇOSA – MINAS GERAIS
2022

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

T
V136r
2022

Valadares, Cristiane Botelho, 1985-
Random Forest Quantílico em estudos de seleção genômica: método de *machine learning* / Cristiane Botelho Valadares. – Viçosa, MG, 2022.
1 tese eletrônica (58 f.): il. (algumas color.).

Orientador: Moysés Nascimento.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Estatística, 2022.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2022.684
Modo de acesso: World Wide Web.

1. Análise de regressão. 2. Aprendizado do computador. 3. Mapeamento cromossômico - Métodos estatísticos. 4. Melhoramento genético. I. Nascimento, Moysés, 1979-. II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 519.536

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

CRISTIANE BOTELHO VALADARES

RANDOM FOREST QUANTÍLICO APLICADO EM ESTUDOS DE SELEÇÃO GENÔMICA

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

APROVADA: 04 de novembro de 2022.

Assentimento:

Cristiane Botelho Valadares
Autora

Moysés Nascimento
Orientador

DEDICO

A Deus e minha família, eles são tudo pra mim, minha vida e meu combustível diário. Tudo por eles e para eles.

AGRADECIMENTOS

A Deus por todo cuidado comigo, por iluminar minhas escolhas, apontar meus caminhos e dar força e segurança para superar os desafios da vida.

Ao meu marido Josiel por todo carinho, cuidado e amor, por ser meu suporte, meu maior incentivador e por acreditar mais em mim do que eu mesma.

Aos meus lindos filhos Laura, Isabela e Mateus, meus maiores tesouros, privados muitas vezes da minha presença em feriados e finais de semana.

A minha mãe Maria Lúcia e minha irmã Gleiciane por serem meu socorro, sempre presentes na hora da angústia.

Ao meu irmão Fabiano por ser meu incentivador e fonte de inspiração.

As queridas Yarítissa e Natalina por serem meu suporte para que esse trabalho pudesse ser concretizado.

Aos queridos amigos que incentivaram, oraram e acreditaram em mim.

Ao meu orientador professor Moysés Nascimento pelos ensinamentos, oportunidades, apoio, confiança e amizade, contribuindo de forma significativa para que pudesse chegar até aqui.

Ao meu coorientador Weverton, pelos ensinamentos, amizade e valiosas contribuições para realização desse trabalho.

Aos professores Ana Carolina, Camila, Laís e Filipe, pela disponibilidade em participar das bancas de qualificação e defesa de tese, contribuindo para melhoria deste trabalho.

Aos amigos do Laboratório de Inteligência Computacional e Aprendizado Estatístico (LICAE), pela união, companheirismo, ensinamentos e momentos de descontração e alegria.

A querida amiga Gabi que sempre esteve presente desde o início, ajudando e me apoiando.

Ao secretário Júnior por ser sempre solícito e disposto a ajudar.

Ao Programa de Pós-Graduação em Estatística Aplicada e Biometria pela oportunidade.

Aos amigos do Departamento de Matemática – DMA/UFV que apoiaram e incentivaram a minha saída para o doutorado.

À Universidade Federal de Viçosa por ouvir a minha história e me dar todo o suporte necessário para a minha capacitação.

Por fim, a todos os que contribuíram direta ou indiretamente para a concretização deste trabalho.

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

MUITO OBRIGADA!

BIOGRAFIA

CRISTIANE BOTELHO VALADARES, filha de José Vitor Botelho (*in memorian*) e Maria Lúcia de Lima Botelho, nasceu em Caratinga, Minas Gerais, em 01 de dezembro de 1985. Em março de 2004, ingressou no curso de Matemática na Universidade Federal de Viçosa, Viçosa - MG, graduando-se em janeiro de 2008, nas modalidades Bacharelado e Licenciatura em Matemática. Em março do mesmo ano, iniciou o curso de Mestrado em Economia Matemática no Instituto de Matemática Pura e Aplicada (IMPA), no Rio de Janeiro, concluindo o mesmo em março de 2010. Em abril de 2010 foi aprovada em concurso público na Universidade Federal de Viçosa e assumiu, em julho de 2010, a função de professora de matemática da instituição. Em março de 2019, iniciou o curso de Doutorado no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa de tese em 04 de novembro de 2022.

“A fé é o combustível que torna nossos sonhos realidade” – Autor desconhecido

“O êxito da vida não se mede pelo caminho que você conquistou, mas sim pelas dificuldades que superou no caminho” – Booker T. Washington.

“Tudo tem o seu tempo determinado, e há tempo para todo o propósito debaixo do céu.” – Eclesiastes 3:1

RESUMO

VALADARES, Cristiane Botelho, D.Sc., Universidade Federal de Viçosa, novembro de 2022. ***Random Forest* Quantílico aplicado em estudos de seleção genômica**. Orientador: Moysés Nascimento. Coorientadores: Ana Carolina Campana Nascimento, Camila Ferreira Azevedo e Weverton Gomes da Costa.

A seleção genômica ampla (GWS) utiliza marcadores distribuídos por todo o genoma para predizer o valor genético genômico de indivíduos. Esta abordagem possibilita acelerar o processo de melhoramento a partir de seleção precoce e aumentar a precisão de predição dos valores genéticos genômicos. Diversas técnicas estatísticas usadas para predição genômica, tais como RR-BLUP, G-BLUP, Bayes A e Bayes B são baseados em erros e, consequentemente, valores fenotípicos com pressupostos de normalidade. Técnicas de aprendizado de máquina tais como Bagging (BA), Random Forest (RF) e Random Forest Quantílico (QRF) aparecem como modelos alternativos já que não requerem suposições a priori sobre a relação funcional entre marcadores e os valores fenotípicos, sem a necessidade de atender pressuposições sobre as distribuições dos dados e dos resíduos. O QRF, metodologia ainda não explorada no contexto de seleção genômica, é um algoritmo não paramétrico que combina as vantagens do *Random Forest* (RF) e da Regressão Quantílica (QR). O método determina a distribuição de probabilidade de uma variável resposta e extrai informações de diferentes quantis e não apenas prevê a média. Neste trabalho propõe-se a avaliação do uso do QRF na predição genômica e a comparação de seus resultados com outras técnicas que já vem sendo exploradas em GWS. Neste trabalho dois artigos foram desenvolvidos com essa proposta. No primeiro deles, o objetivo foi avaliar o desempenho do QRF (nos quantis 0,1; 0,3; 0,5; 0,7 e 0,9) na predição dos valores genéticos genômicos para características com arquitetura genética não aditiva (epistasia e dominância). Adicionalmente, as acurácias obtidas foram comparadas com aquelas advindas do G-BLUP (G-BLUP aditivo, G-BLUP aditivo dominante e G-BLUP aditivo epistático). Foi simulada uma população F2 com 1.000 indivíduos genotipados para 4.010 marcadores SNP. Além disso, doze características foram simuladas a partir de um modelo considerando efeitos aditivos e não aditivos, com número de QTL (*Quantitative trait loci*) variando de oito a 120 e três níveis de herdabilidade (0,3, 0,5 ou 0,8). Em todos os cenários, os resultados da capacidade preditiva do QRF foram iguais ou superiores ao G-BLUP e mostrou ser uma ferramenta alternativa para predizer valores genéticos em características complexas. No segundo trabalho o objetivo foi avaliar o uso do QRF na predição genômica para três características de *Coffea arábica* e comparar as suas capacidades preditivas com metodologias de *machine learning*

(*Bagging* e *Random Forest*), métodos bayesianos (Bayes Cπ e Bayes Dπ) e o G-BLUP. Foram utilizadas as características bicho mineiro, cercosporiose e produção de grãos referentes à 195 indivíduos genotipados com 20.477 marcadores moleculares SNP, resultantes do cruzamento entre Catuaí e Híbrido de Timor, contrastantes em relação à ferrugem do cafeeiro. Os métodos bayesianos apresentaram melhor desempenho para a produção, já o QRF foi igual ou superior aos outros métodos para as características bicho mineiro e cercosporiose, com tempo de processamento muito inferior comparado ao Bayes Cπ e Bayes Dπ. O QRF surge, então, como um algoritmo promissor para predição possibilitando, em alguns cenários, predições mais acuradas de GWS.

Palavras-chave: Predição Genômica. Simulação de Dados. Melhoramento Genético do Cafeeiro. Métodos Bayesianos. G-BLUP. Aprendizado de Máquinas.

ABSTRACT

VALADARES, Cristiane Botelho, D.Sc., Universidade Federal de Viçosa, November, 2022. **Quantile Random Forest applied in genomic selection studies**. Adviser: Moysés Nascimento. Co-advisors: Ana Carolina Campana Nascimento, Camila Ferreira Azevedo and Weverton Gomes da Costa.

Genome Wide Selection (GWS) uses markers distributed throughout the genome to predict the genomic genetic value of individuals. This approach makes it possible to accelerate the improvement process from early selection and increase the prediction accuracy of genomic genetic values. Several statistical techniques used as RR-BL, such as RR-BL, G-BLUP, Bayes A and Bayes are calculated on errors and, consequently, as phenopic values with normality values. Machine learning techniques such as Bagging (BA), Random Forest (RF) and Random Forest Quantile (QRF) appear as alternative models since they do not require a priori assumptions about the functional relationship between markers and phenotypic values, without the need to meet assumptions about the distributions of data and residuals. QRF, a methodology not yet explored in the context of genomic selection, is a non-parametric benefit that combines with Random Forest (RF) and Quantile Regression (QR). This approach can explore nonlinear functions by determining the probability distribution of a response variable extracting information from different quantiles and not just predicting the mean. In this work, it is proposed to evaluate the use of QRF in genomic prediction and compare its results with other techniques that have already been explored in GWS. In this work two articles were developed within this proposal. In the first one, the goal was to evaluate the performance of the QRF (in the quantiles 0,1; 0,3; 0,5, 0,7 and 0,9) in predicting the genomic genetic value for traits with non-additive genetic architecture (epistasis and dominance). Additionally, the achieved accuracy was compared with those from G-BLUP (G-BLUP additive, G-BLUP dominant additive and G-BLUP epistatic additive). An F2 population was simulated with 1.000 genotyped individuals for 4.010 SNP markers. In addition, twelve characteristics were simulated from a model considering additive and non-additive effects, with the number of QTLs (Quantitative trait loci) ranging from 8 to 120 and and three levels of heritability (0,3, 0,5, or 0,8). In all scenarios, the results of the predictive capacity of the QRF were equal to or superior to the G-BLUP and proved to be an alternative tool to predict genetic values in complex characteristics. In the second work the objective was to evaluate the use of QRF in genomic prediction for three Coffea arabica traits and compare their predictive capabilities with machine learning methodologies (Bagging and Random Forest), Bayesian methods (Bayes Cπ and Bayes

Dπ) and G -BLUP. The traits leaf miner, cercosporiosis, and grain production yield for 195 individuals genotyped with 20,477 SNP molecular markers were genotyped, resulting from the crossing between Catuaí and Timor Hybrid, contrasting concerning coffee rust. The objective was to evaluate the use of QRF in genomic prediction. The Bayesian methods showed better performance for the production, while the QRF was equal to or superior to the other methods for the characteristics of leaf miner and cercosporiosis, with a much lower processing time than Bayes Cπ and Bayes Dπ. The QRF then emerges as a promising algorithm for prediction, enabling, in some scenarios, more accurate predictions of GWS.

Keywords: Genomic Prediction. Data Simulation. Genetic Improvement of Coffee. Bayesian Methods. G-BLUP. Machine Learning.

SUMÁRIO

1. INTRODUÇÃO GERAL

A necessidade em se obter maior ganho genético em um menor tempo levou diversos pesquisadores a buscarem estratégias de melhoramento que atendessem essa demanda. Uma delas, conhecida como seleção genômica ampla (GWS – *Genome Wide Selection*) e proposta por Meuwissen et al. (2001), utiliza marcadores moleculares do tipo SNPs (*Single Nucleotide Polymorphisms*), distribuídos ao logo de todo o genoma a fim de predizer o mérito genético de plantas e animais. Tal abordagem tem ganhado destaque pois, além de apresentar alta precisão de predição dos valores genéticos (SINGH et al., 2019), possibilita acelerar o processo de melhoramento (LIU et al., 2019).

Desde o surgimento da GWS, várias técnicas estatísticas têm sido propostas como por exemplo RR-BLUP, Bayes A e Bayes B (MEUWISSEN et al., 2001).Como métodos alternativos, destacam-se as técnicas de *machine learning*, como *Random Forest*, *Bagging*, *Random Forest* Quantílico (BARBOSA et al., 2021; SOUSA et al. 2021; COSTA et al. 2022), dentre outras. Tais metodologias não requerem suposições a priori sobre a relação entre as entradas (marcadores SNPs) e a saída (valores fenotípicos). Isso permite uma certa flexibilidade para lidar com diferentes tipos de efeitos não aditivos complexos, como por exemplo, dominância e epistasia (ZINGARETTI et al., 2020).

O *Random Forest* Quantílico (QRF – *Quantile Random Fo*rest) foi desenvolvido para aproveitar as vantagens da Regressão Quantílica (QR – *Quantile Regression*) e *Random Forest* (RF) juntos. Essa técnica fornece uma análise não paramétrica e um método para estimar quantis condicionais para variáveis preditoras de alta dimensão. (MEINSHAUSEN, 2006).

O QRF tem sido objeto de estudo em diversas áreas do conhecimento e tem mostrado alto poder preditivo (JUBAN et al., 2007; VOLZ et al.., 2016; VAYSSE; LAGACHERIE, 2017; FANG et al., 2018; HERMAN; SCHUMACHER, 2018; LIND; ANDERSON, 2019; KHAN et al*.,* 2019; TAILLARD et al., 2019; ROHMER; IDIER; PEDREROS, 2020). Este algoritmo foi classificado dentre aqueles de melhor desempenho no desafio de predição de sensibilidade a drogas no tratamento de câncer (FANG et al., 2018; LIND; ANDERSON, 2019). Também foi eficaz para a entrega de estimativas confiáveis de incerteza em produtos de mapeamento digital de solo. Neste caso, o QRF superou a Krigagem de Regressão, modelo geralmente aplicado em mapeamento digital de solo operacional, na entrega de estimativas de incerteza (VAYSSE; LAGACHERIE, 2017).

O método também foi proposto para a predição de ondas de calor no Paquistão, o desempenho do QRF foi comparado com o RF e os resultados mostraram desempenho melhor do QRF na detecção de ondas de calor. (KHAN et al., 2019). Volz et al. (2016) fizeram um comparativo entre QR e QRF na predição do comportamento da travessia de pedestres em faixas de pedestres urbanas e o QRF foi superior na predição. A técnica também já foi utilizada com êxito por Juban et al. (2007) para predição de energia eólica e por Zamo et al. (2014) para produção de eletricidade fotovoltaica. Podemos citar ainda o bom desempenho do mesmo na predição de inundações marinhas (ROHMER; IDIER; PEDREROS, 2020), predição de temperatura e vento (TAILLARD et al., 2019); precipitação (HERMAN; SCHUMACHER, 2018).

Apesar de apresentar resultados satisfatórios no contexto dos trabalhos citados, o QRF ainda não foi explorado na GWS. Devido a sua natureza não paramétrica que combina as vantagens dos modelos de RF e QR, essa técnica surge como um algoritmo promissor para predição. Nesse sentido, o objetivo deste trabalho foi avaliar o desempenho do *Random Forest* Quantílico para predição genômica e também compará-lo com outras técnicas já consolidadas no contexto de GWS (G-BLUP, métodos bayesianos, *Random Forest* e *Bagging*). Tais metodologias foram aplicadas em dados simulados de características complexas (epistasia e dominância) e dados reais provenientes do programa de melhoramento de *Coffea arabica* da Empresa de Pesquisa Agropecuária de Minas Gerais (Epamig) em parceria com a Universidade Federal de Viçosa (UFV) e a Empresa Brasileira de Pesquisa Agropecuária (Embrapa Café).

O trabalho foi dividido da seguinte forma: No capítulo 2 foi feito um levantamento bibliográfico das técnicas citadas e utilizadas. No capítulo 3 foram apresentadas as referências utilizadas nos dois primeiros capítulos. Já nos capítulos 4 e 5 foram produzidos dois artigos, frutos dos resultados desta tese, comparando a técnica do *Random Forest* Quantílico com metodologias de *machine learning* (*Bagging* e *Random Forest*), métodos bayesianos (Bayes Cπ e Bayes Dπ) e o G-BLUP , sendo o primeiro artigo desenvolvido com os dados simulados e o segundo com dados provenientes do programa de melhoramento de café.

2. REVISÃO DE LITERATURA

2.1. Seleção genômica ampla (GWS)

A seleção genética tem sido baseada na melhor predição linear não viesada (BLUP), usando dados fenotípicos avaliados a campo (RESENDE et al., 2008). Lande e Thomson (1990) realizaram uma proposição para aumentar a eficiência desse procedimento, baseado em dados fenotípicos, que foi descrita por meio da seleção auxiliada por marcadores moleculares (MAS), a qual usa simultaneamente dados fenotípicos e moleculares.

Os programas de melhoramento genético buscam identificar genótipos superiores e a obtenção de combinações melhoradas por meio de cruzamento entre os indivíduos elite. A necessidade de se obter maior ganho genético em um menor intervalo de tempo levou Meuwissen et al. (2001) a propor um novo método de seleção denominado seleção genômica ampla (GWS – *Genome Wide Selection*), que permite estimar o valor genético genômico dos indivíduos de fenótipos coletados. Com os dados fenotípicos é possível estimar os efeitos dos marcadores de uma população conhecida. Feito isso com uma população de estimação ou treinamento, estes efeitos são testados em uma população de validação (RESENDE et al, 2014). Tal metodologia se manteve discreta por cerca de cinco anos devido os marcadores moleculares serem caros e restritos, mas foi capaz de se desenvolver devido aos diversos investimentos em genotipagem e com o desenvolvimento dos marcadores SNPs.

A GWS é baseada em marcadores que tiveram os efeitos genéticos estimados dos dados fenotípicos em uma população de estimação ou treinamento. Depois de estimados, os efeitos genéticos são testados em uma população de validação e, então, selecionados os marcadores que explicam a variância genética da característica em estudo (RESENDE et al., 2010). Essa metodologia é indicada para caracteres de baixa herdabilidade, apresenta alta acurácia seletiva para seleção e, consequentemente, oferece alta eficiência e rapidez nos ganhos genéticos, com seleção de baixo custo em comparação com a seleção baseada em dados fenotípicos (RESENDE, 2014).

Desde o surgimento da GWS, várias técnicas estatísticas têm sido propostas como, por exemplo, RR-BLUP, G-BLUP, Bayes A e Bayes B (MEUWISSEN et al., 2001). Métodos alternativos são as técnicas de *machine learning*, como Árvores de Decisão e seus refinamentos, dentre outras, que tem apresentado um bom desempenho para predição de valores genéticos

com controle gênico epistático em características com diferentes graus de herdabilidade e diferentes números de genes controladores (BARBOSA et al., 2021; COSTA et al., 2022).

2.2. Melhor Preditor Linear Não Viesado Genômico (G-BLUP)

O G-BLUP (VANRADEN et al., 2009) é um dos modelos comumente utilizado na seleção genômica, utilizando a relação entre os indivíduos estimados pelos marcadores moleculares para predizer o valor genético genômico. É um método muito utilizado no melhoramento animal e vegetal, podendo ser aplicado para caracteres de herança poligênica e que apresentam normalidade dos dados (RESENDE et al., 2008).

O modelo completo para o G-BLUP considerando efeitos aditivos e não aditivos (epistasia e dominância) é dado por:

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

em que $\mathbf{y}$ (N×1, onde N representa o número de indivíduos) é o vetor de observações fenotípicas, $\mathbf{b}$ (p×1, p o número de efeitos fixos) é o vetor de efeitos fixos, $\mathbf{X}$ (N×p) é a matriz de incidência para $\mathbf{b}$, $\mathbf{u_a}$ (n×1, em que n é o número de marcadores moleculares), $\mathbf{u_d}$ (n×1) e $\mathbf{u_e}$ (n×1) são os efeitos aditivos, de dominância e epistático aditivo×aditivo dos indivíduos, respectivamente, $\boldsymbol{\epsilon}$ (N×1) é o vetor de erros aleatórios e $\mathbf{Z}$ (N×n) é a matriz de incidência que relaciona os efeitos dos marcadores aos fenótipos (ou fenótipos corrigidos) contidos em $\mathbf{y}$, esses valores dependem do tipo de marcador e da codificação utilizada. A estrutura de variância do modelo é dada por $\mathbf{u_a}\sim N(0, G_a\sigma^2_{u_a})$, $\mathbf{u_d}\sim N(0, G_d\sigma^2_{u_d})$, $\mathbf{u_e}\sim N(0, G_e\sigma^2_{u_e})$ e $\boldsymbol{\epsilon}\sim N(0, I\sigma^2_e)$, em que $\sigma^2_{u_a}$ é a variância aditiva, $\sigma^2_{u_d}$ é a variância da dominância, $\sigma^2_{u_i}$ é a variância da epistasia, $\mathbf{G_a}$, $\mathbf{G_d}$ e $\mathbf{G_e}$ (N×N) são as matrizes de relacionamento genômica para efeitos aditivos, de dominância e de epistasia, respectivamente.

Para a construção das matrizes de relacionamento genômico utilizadas no modelo, considere M_{ij} como sendo a combinação de alelos da marca j e indivíduo i e p_j a frequência do alelo dominante A na marca j. Pode-se definir as matrizes $\mathbf{W}$(N×N) e $\mathbf{S}$(N×N) dadas por (ZHANG et al., 2019):

$$W_{ij}=\begin{cases}2-2p_j, \text{ se } M_{ij}=AA\\ 1-2p_j, \text{ se } M_{ij}=Aa\\ -2p_j, \text{ se } M_{ij}=aa\end{cases};$$

$$S_{ij}=\begin{cases}-2\left(1-p_j\right)^2, \text{ se } M_{ij}=AA\\ 2p_j(1-p_j), \text{ se } M_{ij}=Aa\\ -2p_j^2, \text{ se } M_{ij}=aa\end{cases}.$$

Dessa forma, obtém-se

$$\mathbf{G_a}=\frac{\mathbf{WW'}}{\sum_{j=1}^{n} 2p_i(1-p_i)};$$

$$\mathbf{G_d}=\frac{\mathbf{SS'}}{\sum_{j=1}^{n}\left[2p_j\left(1-p_j\right)\right]^2};$$

$$\mathbf{G_e}=\mathbf{G_a}\#\mathbf{G_a}.$$

sendo que # denota o produto de Hadamard.

As equações do modelo misto para o modelo completo são dadas por (RESENDE et al., 2014):

$$\begin{bmatrix}\boldsymbol{X'X} & \boldsymbol{X'Z} & \boldsymbol{X'Z} & \boldsymbol{X'Z}\\ \boldsymbol{Z'X} & \boldsymbol{Z'Z}+\boldsymbol{G_a^{-1}}\lambda_1 & \boldsymbol{X'Z} & \boldsymbol{X'Z}\\ \boldsymbol{Z'X} & \boldsymbol{X'Z} & \boldsymbol{Z'Z}+\boldsymbol{G_d^{-1}}\lambda_2 & \boldsymbol{X'Z}\\ \boldsymbol{Z'X} & \boldsymbol{X'Z} & \boldsymbol{X'Z} & \boldsymbol{Z'Z}+\boldsymbol{G_{epi}^{-1}}\lambda_3\end{bmatrix}\begin{bmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{u}}_{\boldsymbol{a}}\\ \widehat{\boldsymbol{u}}_{\boldsymbol{d}}\\ \widehat{\boldsymbol{u}}_{\boldsymbol{epi}}\end{bmatrix}=\begin{bmatrix}\boldsymbol{X'y}\\ \boldsymbol{Z'y}\\ \boldsymbol{Z'y}\\ \boldsymbol{Z'y}\end{bmatrix}$$

onde $\lambda_1=\frac{\sigma_\varepsilon^2}{\sigma_{u_a}^2}$, $\lambda_2=\frac{\sigma_\varepsilon^2}{\sigma_{u_d}^2}$ e $\lambda_3=\frac{\sigma_\varepsilon^2}{\sigma_{u_{epi}}^2}$ e as variâncias foram estimadas pelo Método da Máxima Verossimilhança Restrita (REML).

O Valor Genético Genômico Estimado do indivíduo k ($\widehat{GEBV_k}$) é definido como $\widehat{GEBV_k}=\sum_{j=1}^{N} z_{jk}\widehat{u_{aj}}+\sum_{j=1}^{N} z_{jk}\widehat{u_{dj}}+\sum_{j=1}^{N} z_{jk}\widehat{u_{ej}}$.

2.3. Métodos Bayesianos

Meuwissen et al. (2001) introduziu os métodos bayesianos na seleção genômica ampla a partir dos métodos Bayes A e Bayes B. O modelo geral para seleção genômica é dado por:

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

em que $\boldsymbol{y}$ (N×1, N o número de indivíduos) é o vetor que contém os valores da variável Y para cada indivíduo; $\mathbf{1}$ é o vetor de 1's(N×1); μ é a média de Y; $\boldsymbol{\beta}$ (n×1, n o número de marcadores) é o vetor com os efeitos dos marcadores SNPs; $\boldsymbol{X}$ (N×n) é a matriz de incidência que relaciona efeitos de SNPs aos valores da característica $\boldsymbol{y}$ e $\boldsymbol{e}$ (N×1) é o vetor dos erros.

Na prática, a principal diferença entre os métodos bayesianos em SWG está nas suposições a respeito das distribuições à priori dos efeitos de marcadores (m_i) e das suas respectivas variâncias ($\sigma^2_{m_i}$). A seguir são descritas algumas metodologias e suas conjunturas.

2.3.1. Bayes A

No método Bayes A, proposto por Meuwissen et al. (2001), as variâncias dos marcadores são heterogêneas. Os efeitos dos marcadores (m_i), estimados via função de verossimilhança junto com as distribuições à priori para as variâncias, tem distribuição normal, média zero e a variância de cada marcador $\sigma^2_{m_i}$ dada por uma distribuição qui-quadrada inversa e escalonada (RESENDE et al., 2014). O método Bayes A apresenta as seguintes distribuições a *priori*:

$$m_i|\sigma^2_{m_i} \sim N(0, \sigma^2_{m_i});$$
$$\sigma^2_{m_i} \sim \chi^{-2}(\nu, S^2_m);$$
$$\sigma^2_e \sim \chi^{-2}(\nu_e, S^2_e);$$
$$\sigma^2_a = 2\sum_{i=1}^{m} p_i(1-p_i)\,\sigma^2_{m_i},$$

em que σ^2_a é a variância aditiva, σ^2_e a variância do erro, ν são os graus de liberdade, S^2_m é o parâmetro da escala de distribuição (S^2_m pode ser obtido considerando a esperança de uma variável que segue a distribuição Qui-quadrado invertida escalada), p_i e $(1-p_i)$ representam

as frequências alélicas. Segundo Meuwissen et al. (2001) deve-se considerar $\nu = 4{,}012$ ou 4,2 e $S_m^2 = 0{,}002$ ou 0,0429, $\nu_e \cong -2$ e $S_e^2 = 0$. De acordo com os autores, ao assumir tais valores a distribuição Qui-quadrado invertida escalada produz valores de uma distribuição Uniforme.

2.3.2. Bayes B

No método Bayes B alguns marcadores são considerados nulos (proporção π) e os demais marcadores tem uma variância individual, considerando as mesmas distribuições a *priori* usadas no método Bayes A (CRUZ; SALGADO; BHERING, 2013). A dificuldade do método é escolher o melhor valor para π, pois é necessário conhecer o genótipo analisado ou ter uma metodologia específica para estimá-lo. Este procedimento se equivale ao Bayes A quando $\pi = 0$ (RESENDE et al., 2014). O Bayes B tem as seguintes distribuições a *priori*:

$$m_i|\pi, \sigma^2_{m_i} \sim (1-\pi)N\left(0, \sigma^2_{m_i}\right) + \pi N\left(0, \sigma^2_{m_i} = 0\right);$$
$$\sigma^2_{m_i} \sim \chi^{-2}(\nu, S_m^2);$$
$$\sigma_e^2 \sim \chi^{-2}(\nu_e = -2, S_e^2),$$

em que ν e S_m^2 são os mesmos valores do Bayes A (MEUWISSEN et al., 2001), $\sigma^2_{m_i}$ é a variância de cada marcador e p_i e $(1-p_i)$ são as frequências alélicas.

2.3.3. Bayes Cπ e Bayes Dπ

As diferenças entre os métodos de regressão via abordagem bayesiana aplicados em SWG são devidas principalmente a escolha das distribuições a priori atribuídas aos efeitos dos marcadores. Gianola et al. (2009) fizeram uma análise crítica dos métodos Bayes A e Bayes B em relação as formulações em termos dos hiperparâmetros que propiciam variâncias específicas para cada marcador. De acordo com os autores, nenhum dos dois métodos permite aprendizado bayesiano de maneira adequada, visto que o fator de *shrinkage* (encolhimento) depende demasiadamente da definição de S_m^2. Outra desvantagem do Bayes B é que a probabilidade π é considerada conhecida. Assim, o fator de *shrinkage* para os efeitos dos marcadores também é afetado por π.

Visando resolver esse problema, Habier et al. (2011), propuseram os métodos Bayes Cπ e Bayes Dπ visando limitar a influência de S_m^2 no fator de *shrinkage* para os efeitos dos marcadores (m_i) e tratar π (fração dos marcadores que possuem as mesmas distribuições a priori) como um parâmetro desconhecido do modelo.

As distribuições a *priori* assumidas para os efeitos dos marcadores caracterizam as diferenças entre os modelos bayesianos mencionados (BARILI et al., 2018). De maneira específica, o Bayes Cπ assume as seguintes distribuições a priori (HABIER et al., 2011):

$$m_i|\pi, \sigma_m^2, \sigma_e^2 \sim (1-\pi)N(0, \sigma_m^2) + \pi N(0, \sigma_m^2 = 0);$$
$$\sigma_m^2|\pi, m_i, \sigma_e^2 \sim \chi^{-2}(v,\ S_m^2);$$
$$\pi|\ m_i, \sigma_m^2, \sigma_e^2 \sim U[0,1];$$

em que S_m^2 é derivado como apresentado nos métodos Bayes A e Bayes B.

Já o Bayes Dπ considera as seguintes distribuições a priori (HABIER et al., 2011):

$$m_i|\pi \sim (1-\pi)N\left(0, \sigma_{m_i}^2\right) + \pi N\left(0, \sigma_{m_i}^2 = 0\right);$$
$$\sigma_{m_i}^2|\pi \sim \chi^{-2}(v,\ S_m^2);$$
$$S_m^2 \sim Gamma(\alpha, \beta);$$
$$\pi|\ m_i, \sigma_m^2, \sigma_e^2,\ S_m^2 \sim U[0,1].$$

As variâncias genéticas aditiva para os métodos do Bayes Cπ e Bayes Dπ são dadas, respectivamente por:

$$\sigma_a^2 = \sigma_m^2 \sum_{i=1}^{n} 2p_i(1\text{-}2p_i) \text{ e } \sigma_a^2 = \sum_{i=1}^{n} 2p_i(1\text{-}2p_i)\ \sigma_{m_i}^2.$$

2.4. Regressão Quantílica

Um método estatístico muito utilizado é a regressão linear, que estabelece uma relação funcional entre a variável resposta Y e a variável preditora X. No modelo de regressão linear simples a relação entre essas variáveis é descrita pela equação da reta:

$$Y = \beta_0 + \beta_1 X + \ e,$$

onde β_0 e β_1 são parâmetros desconhecidos e e o erro aleatório, sendo $e \sim N(0, \sigma^2)$ e σ^2 variância desconhecida.

No modelo acima deve-se ter normalidade e homocedasticidade dos erros para que a estimação seja confiável. Koenker e Bassett (1978) propuseram a Regressão Quantílica (QR – *Quantile Regression*), para contornar essas limitações. A técnica é baseada no método dos erros absolutos ponderados que estimam apenas um modelo para o conjunto de dados e permite ajustar modelos de regressão ao longo de toda a distribuição da variável dependente.

A Regressão Quantílica é um tipo especial de análise de regressão, porém, em vez de estimar a média condicional da variável resposta Y, o objetivo é estimar os quantis condicionais a partir dos dados e a estimação dos coeficientes da QR é feita através do Método Simplex (HAO e NAIMAN, 2007).

O modelo da QR permite estudar a influência de um conjunto de variáveis explicativas em qualquer quantil da variável resposta. A relação funcional entre a variável dependente e a variável independente é descrito como:

$$Y_i = \beta_{0i}(\tau) + \beta_{1i}(\tau)X_i + e_i(\tau),$$

em que $\beta_{0i}(\tau)$ é a constante da regressão; $\beta_{1i}(\tau)$ é o coeficiente da regressão; $e_i(\tau)$ são os erros aleatórios independentes e identicamente distribuídos com quantil de ordem τ igual a zero; X_i é variável independente e τ refere se ao quantil assumido ($\tau \in [0,1]$).

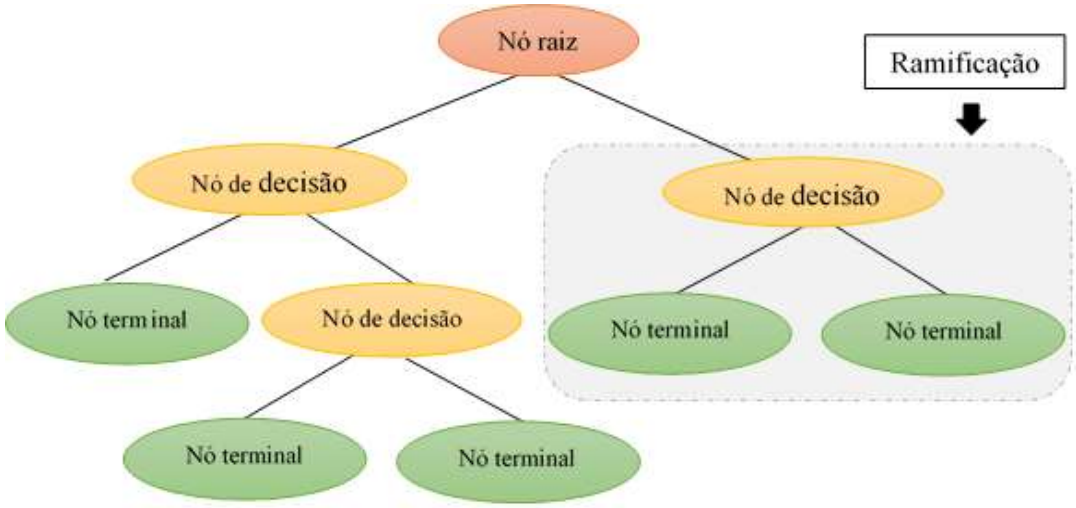
Na QR podem ser estimadas retas para cada quantil de interesse, desta maneira se torna mais adequado à interpretação dos resultados para o conjunto de dados com presença de assimetria, pois através dela é possível traçar a relação em regiões centrais, através da mediana, e nas caudas da distribuição condicional de acordo com o interesse do pesquisador (HAO; NAIMAN, 2007).

No contexto da GWS, a QR permite ajustar modelos em diferentes quantis ao longo de toda a distribuição dos valores fenotípicos e obter os valores genéticos genômicos nos diferentes quantis de interesse (NASCIMENTO et al., 2018). É possível ajustar modelos para todas as partes de uma distribuição de probabilidade permitindo um estudo mais informativo da relação entre variáveis (NASCIMENTO et al., 2017).

2.5. Árvores de Decisão

A árvore de decisão é uma técnica estatística não paramétrica que visa subdividir o conjunto de observações várias vezes de forma que os subgrupos subsequentemente formados sejam cada vez mais homogêneos (SOUSA et al., 2021). O método pode ser aplicado a problemas de regressão e classificação. A árvore tem uma estrutura simples (Figura 1), com uma raiz, ramos, nós e folhas, em que primeiro nó é uma raiz, os nós não terminais (nós de decisão) representam testes em um ou mais atributos e os nós terminais (folhas refletem os resultados da decisão (JAMES et al., 2013).

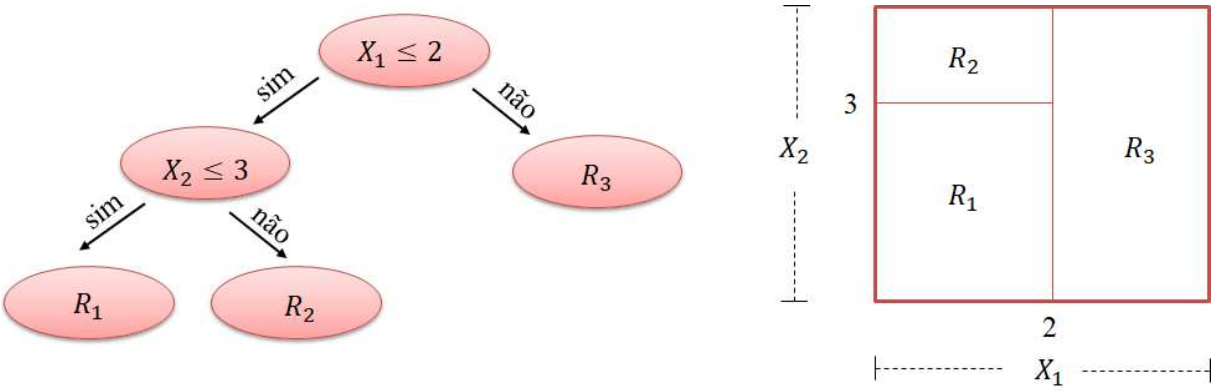
Figura 1. Árvore de decisão com seus elementos: raiz, ramos, nós não terminais (de decisão) e nó terminal (folha).



Fonte: A autora.

Cada nó de uma árvore corresponde a uma certa característica e os ramos correspondem a uma gama de valores (Figura 2). Na figura abaixo temos uma árvore de decisão (à esquerda) e a partição do espaço das preditoras nas regiões correspondentes (à direita da Figura 2).

Figura 2. Árvore de decisão (à esquerda) com as características avaliadas e as regiões *Rj* correspondentes a cada folha (à direita).



Fonte: A autora.

Os métodos baseados em árvore são simples e de fácil interpretação (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; JAMES et al., 2013). Contudo, normalmente não são competitivos como outras abordagens de aprendizagem supervisionada; combinar um grande número de árvores pode muitas vezes resultar em grandes melhorias na precisão da predição, às custas de alguma perda na interpretação (SOUSA et al., 2021).

2.6. Bagging

É conhecido o problema da alta variância das árvores de decisão (SOUSA, 2021). Utilizando o fato em que a média de um conjunto de observações reduz a variância, uma maneira natural de fazer isso e, portanto, aumentar a acurácia da predição, é considerar muitos conjuntos de treinamento da população e construir um modelo de predição separado usando cada um desses conjuntos e depois calcular a média das predições resultantes (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). Esse procedimento é chamado de *Bagging*.

Embora o *Bagging* possa melhorar as predições para muitos métodos de regressão, o método é particularmente útil para árvores de decisão. De acordo com James et al. (2013), para se aplicar o *Bagging* a árvores de regressão, construímos B árvores de regressão usando B conjuntos de treinamento obtidos por *bootstrap*, que consiste em obter B amostras com reposição da amostragem disponível, obtendo assim, B modelos $\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x)$ e calculamos a média das predições resultantes.

$$\hat{f}_{medio}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{b}(x)$$

Cada árvore individual tem alta variância, mas baixo viés e a média dessas B árvores reduz a variância. Para problemas de classificação podemos fazer o seguinte: para uma observação teste dada, podemos registrar a classe prevista por cada uma das B árvores e pegar a maioria dos votos. A predição geral é a classe que ocorre com mais frequência entre as B predições (JAMES et al., 2013). Para evitar *overfitting* o *Bagging* constrói diferentes estimadores considerando diferentes conjuntos de dados de modo que o estimador não seja enviesado. A quantidade de árvores utilizadas no *Bagging* não é um parâmetro que irá resultar num super-ajustamento do modelo, na prática é utilizado uma quantidade onde o erro tenha se estabilizado (JAMES et al., 2013).

Cada árvore usa, em média, cerca de dois terços das observações em subconjuntos obtidos via *bootstrap*. O restante (1/3 dos dados) não é utilizado em cada modelo criado, sendo estes chamados de observações OOB (*out-of-bag*). Segundo James et al. (2013), podemos predizer a resposta para a observação i, usando cada uma das árvores na qual essa observação foi OOB. Isso resultará em B/3 predições. A fim de obter uma única predição para a i-ésima observação, podemos calcular a média dessas respostas previstas (regressão) ou obter uma votação majoritária (classificação). Isso leva a uma única predição OOB para a observação i. Podemos fazer isso para cada uma das n observações, de modo que o erro quadrático médio (regressão) ou erro de classificação (classificação) pode ser calculado. O erro OOB resultante é uma estimativa válida do erro do teste, uma vez que a resposta para cada observação é predita usando apenas as árvores que não foram ajustadas com aquela observação.

2.7. Random Forest

O *Random Forest* é uma combinação de árvores preditoras, de modo que cada árvore depende dos valores de um vetor aleatório amostrado de forma independente e com a mesma distribuição para todas as árvores (BREIMAN, 2001).

As árvores binárias usadas no *Bagging* são construídas a partir dos mesmos dados, então não são estatisticamente independentes e a variância de sua média não pode ser diminuída indefinidamente. Para tornar as árvores mais independentes, Breiman (2001) propôs adicionar outra etapa de randomização ao *Bagging*, cada divisão de cada árvore é construída em um

subconjunto aleatório de variáveis preditoras. Como no *Bagging*, o problema de *overfitting* é resolvido ajustando-se o número de árvores (TAILLARD, 2016).

O RF segue a mesma ideia do *Bagging*, alterando somente o número de variáveis preditoras m usadas em cada partição, Hastie; Tibshirani; Friedman (2009) sugerem que o número de variáveis preditora utilizadas em cada partição seja $m = \sqrt{p}$ para árvore de classificação e $m = p/3$ para árvores de regressão, sendo p o número total de variáveis preditoras.

2.8. *Random Forest* Quantílico

O *Random Forest* Quantílico (QRF), proposto por Meinshausen (2006), é uma generalização do *Random Forest* que combina as vantagens do RF e QR, apresenta uma abordagem não paramétrica e é de fácil interpretação. O método consiste na construção do RF a partir das árvores de decisão de classificação e regressão e, do mesmo modo que na QR, pode obter informações de diferentes quantis, no lugar de prever apenas a sua média (MEINSHAUSEN, 2006). Devido a sua natureza não paramétrica, o modelo QRF pode ser aplicado a qualquer tipo de distribuição de probabilidade, ou seja, a forma /natureza da lei probabilística é obtida a partir dos dados.

Para cada nó, em cada árvore, o RF mantém apenas a média das observações que caem neste nó e negligencia todas as outras informações. Em contraste, o QRF preserva o valor de todas as observações neste nó (MEINSHAUSEN, 2006). Enquanto o RF se aproxima da média condicional, o QRF fornece uma aproximação da distribuição condicional completa. O QRF, portanto, considera a propagação da variável resposta e é considerado mais robusto do que o RF (ROY; LAROCQUE, 2012). Estas informações podem ser usadas para construir intervalos de predição e detectar *outliers* nos dados.

Da mesma forma que o RF, o QRF é um conjunto de árvores de regressão binárias. Mas, para cada folha final de cada árvore, não se calcula a média dos valores do preditor, mas sim sua função de distribuição acumulada empírica (FDA). Uma vez que o RF é construído, determina-se, para um novo vetor de preditores, sua folha associada em cada árvore seguindo a divisão binária. Então, a predição final é a FDA calculada pela média da FDA de todas as árvores. Assim, os quantis preditivos são obtidos diretamente da FDA. Por construção, a FDA

final é limitada entre o valor mais baixo e o mais alto da amostra de aprendizagem (TAILLARD et al., 2016).

Seja Y uma variável resposta de valor real e X uma variável preditora, tal que $X: \Omega \rightarrow \mathcal{B} \subseteq \mathbb{R}^p$. Seguindo a notação de Breiman (2001), considere o vetor aleatório θ que determina como uma árvore de decisão é construída (por exemplo, quais variáveis serão consideradas para os pontos de divisão em cada nó). A árvore correspondente é denotada por $T(\theta)$ (maiores detalhes sobre árvore de decisão podem ser encontrados em Hastie; Tibshirani; Friedman, 2008). Para cada folha $l_j(x, \theta)$ de uma árvore $T(\theta)$ considere o subespaço retangular $R_{l_j} \subseteq \mathcal{B}$, $j = 1, \ldots, L$. Considere $(X_i, Y_i), i = 1, \ldots, n$, n observações independentes. A predição de uma árvore $T(\theta)$, de acordo com Meinshausen (2006), para um novo ponto $X = x$, é dada pela média ponderada das observações Y_i:

$$\hat{\mu}(x) = \sum_{i=1}^{n} \boldsymbol{w_i}(x, \theta\,) Y_i$$

sendo $\boldsymbol{w_i}(x, \theta)$ o vetor peso definido por:

$$\boldsymbol{w_i}(x, \theta) = \frac{1_{\left\{X_i \in R_{l_j(x,\theta)}\right\}}}{\#\left\{m: X_m \in R_{l_j(x,\theta)}\right\}},$$

$$\text{com} \sum_{i=1}^{n} \boldsymbol{w_i}(x, \theta\,) = 1.$$

Segundo Meinshausen (2006), o RF cria um conjunto de k árvores individuais, para essas n observações independentes. A predição por RF é, então, dada por:

$$\hat{\mu}(x) = \sum_{i=1}^{n} \boldsymbol{w_i}(x) Y_i$$

$$\text{onde } \boldsymbol{w_i}(x) = \frac{1}{k} \sum_{t=1}^{k} \boldsymbol{w_i}(x, \theta_t), \qquad \text{sendo } \theta_t \text{ i. i. d}, \qquad t = 1, \ldots, k.$$

Por fim, para fazer a predição através do *Random Forest* Quantílico (QRF), de acordo com Meinshausen (2006), deve-se calcular os pesos médios $\boldsymbol{w_i}(x)$ de cada observação *i* sobre todas as árvores do RF. Esses pesos são usados para calcular a estimativa da função de distribuição $\hat{F}$, definida como:

$$\hat{F}(y|X = x) = \sum_{i=1}^{n} \boldsymbol{w_i}(x) 1_{\{Y_i \leq y\}}$$

Calcula-se então a estimativa do quantil condicional $Q_\alpha(x)$:

$$Q_\alpha(x) = \inf\left\{y: \hat{F}(y|X = x) \geq \alpha\right\},$$

para algum $\alpha,\ 0 < \alpha < 1$, utilizando os mesmos pesos $w_i(x)$ do modelo RF.

A principal diferença entre QRF e RF é que, para cada nó, em cada árvore, o RF mantém apenas a média das observações que caem nesse nó e negligenciam qualquer outra informação. Por outro lado, o QRF mantém o valor de todas as observações do nó (não apenas a média) e avalia a distribuição condicional com base nessas informações.

3. REFERÊNCIAS BIBLIOGRÁFICAS

BARBOSA, I. P.; SILVA, M. J.; COSTA, W. G.; SANT'ANNA, I. C.; NASCIMENTO, M.; CRUZ, C. D. Genome-enabled prediction through machine learning methods considering different levels of trait complexity. **Crop Science**, v.61: p. 1890–1902, 2021.

BARILI, L. D.; VALE, N. M. do; SILVA, F., F.; CARNEIRO, J. E. S; OLIVEIRA, H. R. de; VIANELLO, R. P.; VALDISSER, P. A. M. R.; NASCIMENTO, M. Genome prediction accuracy of common bean via Bayesian models. **Ciência Rural**, v.48, n. 8, 2018.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, p. 5–32, 2001.

COSTA, W. G. da; CELERI, M. O.; BARBOSA, I. P.; SILVA, G., N.; AZEVEDO, C., F.; BOREM, A.; NASCIMENTO, M.; CRUZ, C. D. Genomic prediction through machine learning and neural networks for traits with epistasis. **Computational and Structural Biotechnology Journal**, v. 20, p. 5490 – 5499, 2022.

CRUZ, C. D.; SALGADO, C. C.; BHERING, L. L. **Genômica Aplicada**. Visconde do Rio Branco, MG: Ed. Suprema, 424p., 2013.

FANG Y.; XU P.; YANG J.; QIN Y. A quantile regression forest based method to predict drug response and assess prediction reliability. **PLoS ONE,** v. 13, n.10, 2018.

GIANOLA, D. et al. Additive genetic variability and the Bayesian alphabet. **Genetics**, v. 183, p. 347-363. 2009.

HABIER, D. et al. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics, 12:186. 2011.

HAO, L.; NAIMAN, D. Q. **Quantile Regression.** Sage publications. 126p. 2007.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statiscal Learning: Data Mining, Inference, and Prediction**. Springer, 2. ed., 745p, New York, NY, USA, 2008.

HERMAN, G. R.; SCHUMACHER, R. S. Money Doesn't Grow on Trees, but Forecasts Do: Forecasting Extreme Precipitation with Random Forests. **Monthly Weather Review**, v. 146, n.5, p. 1571–1600, 2018.

JAMES, G.; WITTEN D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning with Applications in R.** Springer, 426p., New York, NY, USA, 2013.

JUBAN , J.; FUGON, L.; KARINIOTAKIS, G. Probabilistic short-term wind power forecasting based on kernel density estimators. **European Wind Energy Conference and Exhibition**, EWEC, Milão, Itália, 2007.

KHAN, N.; SHAHID, S.; JUNENG, L.; AHMED, K.; ISMAIL, T.; NAWAZ, N. Prediction

of heat waves in Pakistan using quantile regression forests. **Atmospheric Research**, 2019.

KOENKER , R.; BASSETT JR, G. Regression Quantiles. **Econometrica**, v. 46, n.1, p. 33-50, 1978.

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

LIND A.P.; ANDERSON, P. C. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. **PLoS ONE,** v.14, n.7, 2019.

LIU, X.; WANG, H.; HU, X.; LI, K.; LIU, Z.; WU, Y.; Huang, C. Enhancing genomic selection with quantitative trait loci and nonadditive effects revealed by empirical evidence in maize. **Frontiers in plant science**, v. 10, 2019.

MEUWISSEN, T. H. E., HAYES, B. J., GODDARD, M.E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. **Genetics**, v. 157, 2001.

MEINSHAUSEN, N., Quantile regression forests. **Journal of Machine Learning Research**, v. 7, p. 983–999, 2006.

NASCIMENTO, M. et al. Regularized quantile regression applied to genome-enabled prediction of quantitative traits. **Genetics and Molecular Research**, Ribeirão Preto, v. 16, n. 1, p. 1-12, 2017.

NASCIMENTO, M. et al. Quantile regression for genome-wide association study of flowering time-related traits in common bean. **PloS one**, Bethesda, v. 13, n. 1, 2018.

PORTNOY, S.; KOENKER, R. The gaussian hare and the laplacian tortoise: Computability of squared-error versus absolute-error estimates. **Statistical Science,** v. 12, p. 279–300, 1997.

RESENDE, M. D. V. et al. Seleção Genômica Ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, v. 56, p. 63-77, 2008.

RESENDE, M. D. V. de; RESENDE JÚNIOR, M. F. R. R.; AGUIAR A. M.; ABAD, J. I. M.; MISSIAGGIA A. A.; SANSALONI, C.; PETROLI, C.; GRATTAPAGLIA, D. **Computação da seleção genômica ampla (GWS)**. Colombo, Embrapa Florestas. 79p., 2010.

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

ROHMER, J.; IDIER, D.; PEDREROS, R. A nuanced quantile random forest approach for fast prediction of a stochastic marine flooding simulator applied to a macrotidal coastal site. **Stoch Environ Res Risk Assess**, v. 34, p. 867–890, 2020.

ROY, M. H.; LAROCQUE, D. Robustness of random forests for regression. **Journal of Nonparametric Statistics**, v. 24, n. 4, p. 993–1006, 2012.

SINGH, B.; MAL, G.; GAUTAM, S. K.; MUKESH, M.Whole-Genome Selection in Livestock. In: **Advances in Animal Biotechnology**. Springer, p. 349-364, 2019.

SOUSA, I. C. D.; NASCIMENTO, M.; SILVA, G. N.; NASCIMENTO, A. C. C.; CRUZ, C. D.; ALMEIDA, D. P. D.; PESTANA, K. N.; AZEVEDO, C. F.; ZAMBOLIM, L.; CAIXETA, E. T. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. **Scientia Agricola**, v. 78, n.4, 2021.

TAILLARD, M.; MESTRE, O.; ZAMO M.; NAVEAU P. Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics. **Monthly Weather Review**, v. 144, n.6, p. 2375–2393, 2016.

VANRADEN, P. M. et al. Invited review: Reliability of genomic predictions for North American Holstein bulls. **Journal of dairy science**, v. 92, n. 1, p. 16-24, 2009.

VAYSSE K.; LAGACHERIE P. Using quantile regression forest to estimate uncertainty of digital soil mapping products, **Geoderma**, v. 291, p. 55-64, 2017.

VÖLZ B.; H. MIELENZ, R; Siegwart e J. Nieto, "Predicting pedestrian crossing using Quantile Regression woods", **2016 IEEE Intelligent Vehicles Symposium (IV)**, Gotemburgo, Suécia, p. 426-432, 2016.

ZAMO , M.; MESTRE, O.; ATBOGASTO, P.; PANNEKOUCKE, O. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part I: Deterministic forecast of hourly production. **Solar Energy**, v. 105. p. 792–803, 2014.

ZHANG, H.; YIN, L.; WANG, M.; YUAN, X.; LIU, X. Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. **Frontiers in genetics**, v. 10, 2019.

ZINGARETTI, L. M.; GEZAN, S. A.; FERRÃO, L. F. V.; OSORIO, L. F.; MONFORT, A.; MUÑOZ, P. R.; WHITAKER, V. M.; PÉREZ-ENCISO, M. Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. **Frontiers in plant science**, v. 11, 2020.

4. ARTIGO 1

QUANTILE RANDOM FOREST FOR GENOMIC PREDICTION IN COMPLEX CHARACTERISTICS

Abstract

Quantile Random Forest (QRF) is a non-parametric methodology that combines the advantages of *Random Forest* (RF) and Quantile Regression (QR). Specifically, this approach can explore non-linear functions, determining the probability distribution of a response variable and extracting information from different quantiles instead of just predicting the mean. The objective of this work was to evaluate the performance of the QRF in the genomic prediction for complex traits (epistasis and dominance). In addition, compare the accuracies obtained with those derived from the G-BLUP. The simulation created an F2 population with 1.000 individuals and genotyped for 4.010 SNP markers. Besides, twelve traits were simulated from a model considering additive e non-additive effects, QTL (*Quantitative trait loci*) numbers ranging from eight to 120, and heritability of 0,3, 0,5, or 0,8. For training and validation, the 5-fold cross-validation approach was used. For each fold, the accuracies of all the proposed models were calculated: QRF in five different quantiles and three G-BLUP models (additive effect, additive and epistatic effects, additive and dominant effects). Finally, the predictive performance of these methodologies was compared. In all scenarios, the QRF accuracies were equal to or greater than the methodologies evaluated and proved to be an alternative tool to predict genetic values in complex traits.

Keywords: Genomic selection, accuracy, epistasis, dominance, prediction.

Resumo

Random Forest Quantílico (QRF) é uma metodologia não paramétrica, que combina as vantagens do *Random Forest* (RF) e da Regressão Quantílica (QR). Especificamente, essa abordagem pode explorar funções não lineares, determinando a distribuição de probabilidade de uma variável resposta e extraindo informações de diferentes quantis em vez de apenas prever a média. O objetivo deste trabalho foi avaliar o desempenho do QRF em predizer o valor genético genômico para características com arquitetura genética não aditiva (epistasia e

dominância). Adicionalmente, as acurácias obtidas foram comparadas com aquelas advindas do G-BLUP. A simulação criou uma população F2 com 1.000 indivíduos genotipados para 4.010 marcadores SNP. Além disso, doze características foram simuladas a partir de um modelo considerando efeitos aditivos e não aditivos, com número de QTL (*Quantitative trait loci*) variando de oito a 120 e herdabilidade de 0,3, 0,5 ou 0,8. Para treinamento e validação foi usada a abordagem da validação cruzada 5-fold. Para cada um dos folds foram calculadas as acurácias de todos os modelos propostos: QRF em cinco quantis diferentes e três modelos do G-BLUP (com efeito aditivo, aditivo e epistático, aditivo e dominante). Por fim, o desempenho preditivo dessas metodologias foi comparado. Em todos os cenários, as acurácias do QRF foram iguais ou superiores às metodologias avaliadas e mostrou ser uma ferramenta alternativa para predizer valores genéticos em características complexas.

Introduction

Genomic wide selection (GWS) presents high accuracy in predicting genomic breeding values, accelerating the process of genetic improvement (SINGH et al., 2019; LIU et al., 2019). Statistical methodologies generally used for genomic prediction, such as RR-BLUP, G-BLUP, Bayes A, and Bayes B (MEUWISSEN et al., 2001), are based on errors and, consequently, phenotypic values normality assumptions. The employment of computational intelligence-based methods to predict genomic breeding values is increasing (SOUSA et al., 2021; KUJAWA; NIEDBAŁA, 2021). Compared to statistical methods for predicting genomic values, such methodologies do not require assumptions about the model, making them more flexible for a wide range of problems (ROSADO et al., 2022). Specifically, such flexibility allows one to deal naturally with different types of non-additive genetic effects, like dominance and epistasis.

Among computational intelligence methodologies, *Random Forest* (RF) has proven to be an interesting alternative for genomic prediction; in addition to possibly increasing the predictive performance of the method, it reduces, through the selection of variables, problems related to correlated variables (BREIMAN, 2001). Such methodology has been successfully employed to predict genomic breeding values, like the study by Barbosa et al. (2021) that used RF in simulated populations with different levels of heritability and loci numbers of quantitative characteristics (QTL) in the presence of dominant and epistatic effects. These authors have

found that machine-learning methodologies, such as RF, are powerful tools to predict genomic breeding values for traits that comprise non-additive genetic architecture. In addition, Sousa et al. (2021) successfully applied the methodology in predicting rust resistance in *Coffea arabica*. This study has verified that the RF presented higher results, in terms of apparent error rate, compared to generalized Bayesian LASSO.

Another method used in GWS that is robust to the break of assumptions and allows one to adjust models along the entire probability distribution of the characteristic of interest is quantile regression (QR) (KOENKER; BASSET, 1978). The QR allows the investigation of possible issues related to asymmetry and heteroscedasticity present in data sets (NASCIMENTO et al., 2019). Concerning GWS, such a method was used by Nascimento et al. (2017) that employed QR to estimate genomic breeding value from simulated data in different asymmetry scenarios and observed that the technique provided better results regarding accuracy in the presence of asymmetric phenotypic values. In addition, Oliveira et al. (2021) evaluated the use of QR considering simulated data of autogamous plants with oligogenic characteristics and observed better or equal results to those obtained by RR-BLUP and BLASSO.

A methodology that seizes qualities from both QR and RF is the *Quantile Random Forest* (QRF) (MEINSHAUSEN, 2006). This approach combines the best explanation of a phenomenon obtained through QR and increases predictive power by using the RF.

The QRF was ranked among those with the best performance in the challenge of predicting drug sensitivity in cancer treatment (FANG et al., 2018; LIND; ANDERSON, 2019), efficiently predicted heat waves in Pakistan (KHAN et al., 2019) and marine flooding (ROHMER; IDIER; PEDREROS, 2020).

Despite its potential, the QRF has not yet been used in the context of GWS, so this study aimed: i) to propose and evaluate the use of QRF in genomic prediction for complex characteristics (inclusion of dominance and epistasis effects); ii) to compare the accuracy of the QRF with those resulting from the G-BLUP methodology.

Materials and Methods

Experimental data

An F2 population of a diploid species (2n = 20) was simulated, containing 1.000 individuals. A co-dominant 4.010 markers (locus) of bi-allelic single nucleotide polymorphisms (SNPs) were considered, distributed equally and equidistantly into 10 binding groups

(chromosomes) with a size of 200 cm each (401 markers on each chromosome). With the simulated genotypic data of the F2 population, 12 scenarios (C1 to C12) were considered, with the number of controlling genes (QTLs) equal to 8, 40, 80 or 120, distributed equally among the first eight linkage groups and heritabilities of 0,3, 0,5 or 0,8.

Table 1. Evaluated scenarios

Heritability (h^2)	number of controlling genes (QTLs)			
	8	40	80	120
0.3	C1	C4	C7	C10
0.5	C2	C5	C8	C11
0.8	C3	C6	C9	C12

Fonte: A autora.

The phenotypic characteristics of the 12 scenarios were simulated considering the mean (μ) equal to 100 and, coefficient of variation of 10%, average degree of dominance (d_i) equal to 0,5 and controlled by a model that also includes epistatic effect:

$$Y_i = \mu + \sum_j \alpha_j + \sum_j \sum_{j'} \alpha_j \alpha_{j'} + \varepsilon_i,$$

in which Y_i is the phenotypic value for the *i* observation; μ the overall mean; α_j is the effect of the favorable allele on the j locus and assume the values $u + a_i$, $u + d_i$ and $u - a_i$ the genotypic values associated with the classes AA, Aa, and aa, respectively, with u as the mean between the dominant homozygous (AA) and the recessive homozygous (aa); $\alpha_j\alpha_{j'}$ represents the interaction between favorable alleles in different loci (epistasis). For the variance of errors, we have the errors vector $\varepsilon \sim N(0, I\, V_\varepsilon)$, in which $V_\varepsilon=[(1-h^2)V_g]/h^2$ with V_ε as the residual variance, V_g the genotypic variance and h^2 the heritability.

Quantile Random Forest (QRF)

For the construction of the QRF, it is necessary to obtain T regression trees generated from *bootstrap* samples considering subsets of markers under study, i.e., the construction of trees based on *Random Forest* (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). Later, for each

generated tree (T_t), conditional distribution is obtained by weighting the observed values of the studied characteristic. Specifically, given an observation, $X = x$, it is defined for each terminal node (adjusted tree leaf), $F(x, T_{tf})$, the following weighting factor: $w_i(x, T_{tf}) = \frac{I_{\{x \in F(x,T_{tf})\}}}{\#\{m: X_m \in F(x,T_{tf})\}}$, with $\sum_{i=1}^{n} w_i(x, T_{tf}) = 1$, $I_{\{X_i \in F(x_i,T_{tf})\}}$ an indicator variable stating that the observed value ($X = x$) belongs to f-th leaf and $\#\{m: X_m \in F(x, T_{tf})\}$ represents the number of observations on the f-th leaf.

The prediction of a tree T_t, according to MEINSHAUSEN (2006), for a new point, $X = x_{new}$, is given by the weighted average of the observations Y_i, that is, $\hat{\mu}(x_{new}) = \sum_{i=1}^{n} w_i(x, T_{tf}) Y_i$. In this way, the prediction for a given observation, $X = x$, after the construction of T trees is given by: $\hat{\mu}_{RF}(x) = \sum_{i=1}^{n} w_i(x) Y_i$, in which: $w_i(x) = \frac{1}{T} \sum_{t=1}^{T} w_i(x, T_{tf})$. Taking into consideration that the estimated cumulative distribution function is given by: $\hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(x) I_{\{Y_i \leq y\}}$, in which $I_{\{Y_i \leq y\}}$ is an indicator function, the predicted value for the τ-th quantile is given by $Q_\tau(x) = \inf \{y: \hat{F}(y|X = x) \geq \tau\}$, for any τ, $0 < \tau < 1$.

The main difference between QRF and RF is that for each node in each tree, the RF maintains only the average of the observations that fall into that node and neglects any other information. On the other hand, the QRF maintains the value of all node observations (not just the average) and evaluates conditional distribution based on this information (MEINSHAUSEN, 2006).

Genomic best linear unbiased predictor (G-BLUP)

In order to compare the results obtained by *Quantile Random Forest*, the adjustment of three G-BLUP models was considered: i. G-BLUP-A (only the additive component), ii. G-BLUP-AD (additive component and due to the dominance effect); iii. G-BLUP-AE (additive component and due to epistasis additive×additive). The description of such models can be seen in detail in RESENDE et al. (2014).

Comparison of methodologies

In order to access the prediction quality of the evaluated models, accuracy was used. This measure is defined as Pearson's correlation coefficient between the individuals' simulated genetic values and those predicted by the adjusted model. The higher the accuracy value, the better the model is in terms of prediction capacity, and it can be used in the selection phase of

new individuals. In this work, the accuracy of the QRF was calculated in the quantiles 0,1, 0,3, 0,5, 0,7, and 0,9 and compared to the accuracy of the adjusted G-BLUPs models.

For training and validation, 5-fold cross-validation was performed. The data set is divided into 5 populations. At the k-th fold ($k = 1,..,5$), the k-th population is used as a validation population. The remaining populations were used as a validation population. For each of the five folds, the accuracy of all the proposed genomic selection models was calculated: QRF at quantiles 0,1, 0,3, 0,5, 0,7 e 0,9, G-BLUP-A, G-BLUP-AD e G-BLUP-AE and at the end, the mean and the standard error between the folds was estimated.

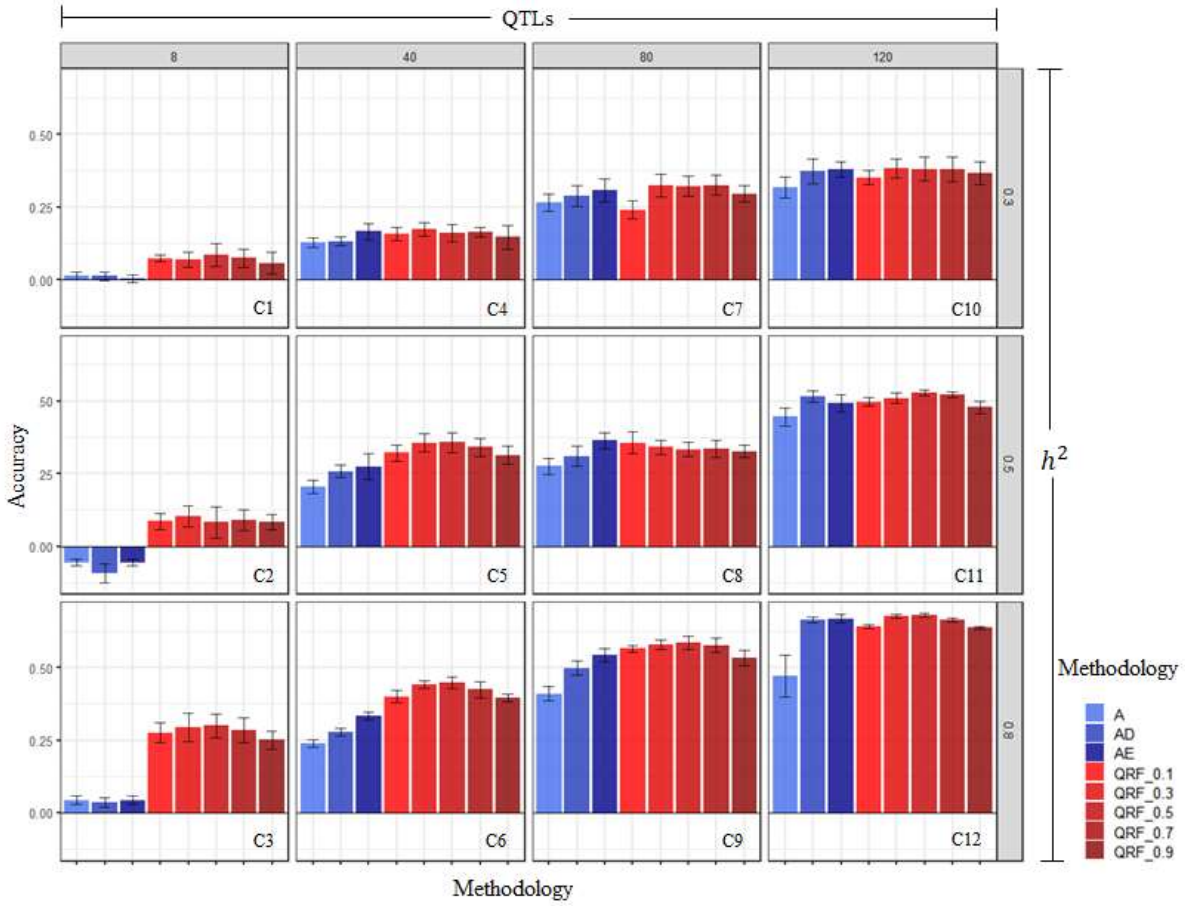
Computational Resources

Data simulation was performed on the Genes software (CRUZ, 2016). All G-BLUP models (with additive, additive and epistatic, additive and dominant effect) were adjusted using the GenomicLand software (AZEVEDO et al., 2019). To run the QRF, it was used the "quantregForest" package (MEINSHAUSEN, 2017) of the R software (R Core Team et al., 2020).

Results and Discussion

Overall, among the three evaluated G-BLUPs models (Additive - A, Additive and Dominant - AD, Additive and Epistatic - AE), those that have non-additive effects in their adjustment, i.e., GBLUP - AD and GBLUP - AE, were the ones that presented the best accuracy values for all evaluated scenarios (Figure 1). This result is reasonable since the data used in this study consider dominance and epistasis effects in its simulation, indicating that the adjustment of non-additive effects is an essential factor to be considered in the modeling. Similar results have been reported in the literature (CALLEJA-RODRIGUES et al., 2021; YADAV et al., 2021). Specifically, Calleja-Rodrigues et al. (2021) showed that genomic prediction is more accurate when considering the inclusion of non-additive effects in *Pinus sylvestris*. Yadav et al. (2021) also obtained results that indicate the superiority of the adjustment of models with non-additive effects in sugarcane.

Figure 1. Estimation of the accuracy of the QRF's models (QRF_0,1, QRF_0,3, QRF_0,5, QRF_0,7 e QRF_0,9), G-BLUP-A (only the additive component), G-BLUP-AD (additive component and due to the dominance effect); G-BLUP-AE (additive component and due to epistasis additive×additive) for combinations of the number of controlling genes (8, 40, 80 and 120) and heritability (0,3, 0,5 and 0,8), represented by scenarios C1 to C12.



Fonte: A autora

Another result that can be highlighted is that, with the increase in the number of QTLs, there was a significant improvement in the accuracy of G-BLUPs methods. This result is justified once G-BLUP considers the infinitesimal model (AZEVEDO et al., 2015), that is, many genes of little effect controlling the characteristic. Thus, genomic predictions are based on the kinship obtained from all markers and, thus, when more markers have a genetic effect, the accuracy of the prediction increases (WANG et al., 2018; ZHANG et al., 2019).

Considering, as a basis for comparison the G-BLUP models with non-additive effects, the construction of models based on the QRF showed greater accuracy in the scenarios in which the characteristics were controlled by 8 and 40 QTLs (Figure 1, Scenarios C1 to C6). In the other scenarios, in which at least 80 QTLs control the characteristics, the adjustment through

the QRF presented results similar to those obtained by adjusting the G-BLUP models considering non-additive effects (Figure 1, Scenarios C7 to C12). Barbosa et al. (2021), considering simulated data, observed that for characteristics with the lowest number of QTLs, the multiplicative effects of the controlling genes (epistasis) may be more important since the individual effect of each gene is more significant than in the characteristics controlled by a larger number of QTLs. In Sousa et al. (2021), *machine learning* methodologies were applied for genomic prediction of rust resistance in *Coffea arabica*. Such a study verified that such methodologies, including the RF, presented higher results concerning apparent error rate compared to generalized Bayesian LASSO.

An interesting feature of the QRF is that, like QR, this method enables the adjustment of models for all parts of the probability distribution of the characteristic, allowing the conditional quantile that "better" describes the functional relationship between dependent and independent variables to be used for prediction (NASCIMENTO et al., 2019). In addition, QR does not require assumptions as to probability distribution and is robust to outliers (OLIVEIRA et al., 2021). In the present study, the quantiles 0,1, 0,3, 0,5, 0,7, and 0,9 were considered for the construction of the models. It was possible to observe, especially for a smaller number of QTLs (8 and 40 QTLs), that the highest accuracy values were observed when considering the models QRF_0,3 (quantile 0,3) and QRF_0,5 (quantile 0,5).

Conclusion

The QRF proved to be able to predict genetic values with epistatic gene control in characteristics with different degrees of heritability and different numbers of QTLs. It was equal to or superior to the G-BLUP methodology in all evaluated scenarios, presenting higher accuracy even when the characteristic is of low heritability.

References

AZEVEDO, C. F.; NASCIMENTO, M.; FONTES, V. C.; RESENDE, M. D. V. D.; CRUZ, C. D. GenomicLand: Software for genome-wide association studies and genomic prediction. **Acta Scientiarum. Agronomy**, v. 41, 2019.

AZEVEDO, C. F.; RESENDE, M. D. V.; SILVA, F. F.; VIANA, J. M. S.; VALENTE, M. S.; RESENDE, M. F. R.; MUNOZ, P. Ridge, Lasso and Bayesian additive-dominance genomic models. **BMC Genetics** (Online), v. 16, p. 105, 2015.

BARBOSA, I. P.; SILVA, M. J.; COSTA, W. G.; SANT'ANNA, I. C.; NASCIMENTO, M.; CRUZ, C. D. Genome-enabled prediction through machine learning methods considering different levels of trait complexity. Crop Science, v.61: p. 1890–1902, 2021.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, p. 5–32, 2001.

CALLEJA-RODRIGUEZ A.; CHEN Z.; SUONTAMA M.; PAN J.; WU HARRY X. Genomic Predictions With Nonadditive Effects Improved Estimates of Additive Effects and Predictions of Total Genetic Values in Pinus Sylvestris. **Frontiers in Plant Science**, v. 12, 2021. CRUZ, C. D. Genes Software - extended and integrated with the R, Matlab and Selegen. **Acta Scientiarum. Agronomy**. v. 38, n. 4, 2016.

FANG Y.; XU P.; YANG J.; QIN Y. A quantile regression forest based method to predict drug response and assess prediction reliability. **PLoS ONE**, v. 13, n.10, 2018.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statiscal Learning: Data Mining, Inference, and Prediction.** Springer, 2. ed., 745p, New York, NY, USA, 2008.

KHAN, N.; SHAHID, S.; JUNENG, L.; AHMED, K.; ISMAIL, T.; NAWAZ, N. Prediction of heat waves in Pakistan using quantile regression forests. **Atmospheric Research**, 2019.

KOENKER, R.; BASSET, G. Regression Quantiles. **Econometrica**, v. 46, p. 33–50, 1978. KUJAWA, S.; NIEDBAŁA, G. Artificial Neural Networks in Agriculture. **Agriculture,** v. 11, p. 497, 2021.

LIND A.P.; ANDERSON, P. C. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. **PLoS ONE**, v.14, n.7, 2019.

LIU, X.; WANG, H.; HU, X.; LI, K.; LIU, Z.; WU, Y.; Huang, C. Enhancing genomic selection with quantitative trait loci and nonadditive effects revealed by empirical evidence in maize. **Frontiers in plant science**, v. 10, 2019.

MEUWISSEN, T. H. E., HAYES, B. J., GODDARD, M.E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. **Genetics**, v. 157, 2001.

MEINSHAUSEN, N. Quantile regression forests. **Journal of Machine Learning Research**, v. 7, p. 983–999, 2006.

MEINSHAUSEN, N. quantregForest: Quantile Regression Forests. R package version 1.3-7. https://cran.r-project.org/web/packages/quantregForest/quantregForest.pdf, 2017. (Último acesso em 21 de julho de 2021).

NASCIMENTO, A.C.; NASCIMENTO, M.; AZEVEDO, C.; SILVA, F.; BARILI, L.; VALE, N.; CARNEIRO, J.E.; CRUZ, C.D.; CARNEIRO, P.C.; SERÃO, N. Quantile Regression

Applied to Genome-Enabled Prediction of Traits Related to Flowering Time in the Common Bean. **Agronomy**, v.9, n.12, 2019.

NASCIMENTO, M.; SILVA, F.F.; RESENDE, M.D.V.; CRUZ, C.D.; NASCIMENTO, A.C.C.; VIANA, J.M.S.; AZEVEDO, C.F.; BARROSO, L.M.A. Regularized quantile regression applied to genome-enabled prediction of quantitative traits. **Genetics and Molecular Research**, Ribeirão Preto, v. 16, n. 1, p. 1-12, 2017.

OLIVEIRA G.F.; NASCIMENTO A.C.C.; NASCIMENTO M.; SANT'ANNA I. de C.; ROMERO J.V. et al. Quantile regression in genomic selection for oligogenic traits in autogamous plants: A simulation study. **PLoS ONE**, v.16, n.1, 2021.

R Core Team, R.F. for S. Computing, and R.C. Team. 2020. R: A Language and Environment for Statistical Computing. https://www.r-project.org/. (accessed 1 July 2021).

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

ROHMER, J.; IDIER, D.; PEDREROS, R. A nuanced quantile random forest approach for fast prediction of a stochastic marine flooding simulator applied to a macrotidal coastal site. **Stoch Environ Res Risk Assess**, v. 34, p. 867–890, 2020.

ROSADO, R. D. S.; PENSO, G. A.; SERAFINI, G. A. D.; SANTOS, C. E. M.; PICOLI E. A.T.; CRUZ C. D.; BARRETO, C. A. V.; NASCIMENTO, M.; CECON, P. R. Artificial neural network as an alternative for peach fruit mass prediction by non-destructive method. **Scientia Horticulturae**, v. 299, 2022.

SINGH, B.; MAL, G.; GAUTAM, S. K.; MUKESH, M.Whole-Genome Selection in Livestock. In: **Advances in Animal Biotechnology**. Springer, Cham, p. 349-364, 2019.

SOUSA, I. C. de; NASCIMENTO, M.; SILVA, G. N.; NASCIMENTO, A. C. C.; CRUZ, C. D.; SILVA, F. F.; ALMEIDA, D. P. de; PESTANA, K. N.; AZEVEDO, C. F.; ZAMBOLIM, L.; CAIXETA, E. T. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. **Scientia Agricola**, v. 78, p. 1, 2021.
WANG J.; ZHOU Z.; ZHANG Z. et al. Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. **Heredity**, v.121, p. 648–662, 2018.

YADAV, S.; WEI, X.; JOYCE, P. et al. Improved genomic prediction of clonal performance in sugarcane by exploiting non-additive genetic effects. **Theoretical and Applied Genetics**, v. 134, p. 2235–2252, 2021.

ZHANG, H.; YIN, L.; WANG, M.; YUAN, X.; LIU, X. Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. **Frontiers in genetics**, v. 10, 2019.

5. ARTIGO 2

RANDOM FOREST QUANTÍLICO PARA PREDIÇÃO GENÔMICA EM GENÓTIPOS DE *COFFEA ARABICA*

Resumo

O *Random Forest* Quantílico é uma metodologia não paramétrica, que combina as vantagens dos modelos *Random Forest* e Regressão Quantílica, capaz de determinar a distribuição de probabilidade completa de uma variável resposta e com isso extrair as informações de diferentes quantis e não apenas prever a média. Neste trabalho foram avaliadas três características de *Coffea arabica*: bicho mineiro, cercosporiose e produção de grãos. Foram genotipados 195 indivíduos com 20.477 marcadores moleculares SNP, resultantes do cruzamento entre Catuaí e Híbrido de Timor, contrastantes em relação à ferrugem do cafeeiro. O objetivo foi avaliar o uso do *Random Forest* Quantílico na predição genômica para estas características e comparar as suas capacidades preditivas com metodologias de *machine learning* (*Bagging* e *Random Forest*), métodos bayesianos (Bayes Cπ e Bayes Dπ) e o G-BLUP. Os métodos bayesianos apresentaram melhor capacidade preditiva para a característica produção, já o QRF foi igual ou superior aos outros métodos para as características bicho mineiro e cercosporiose e tempo de processamento inferior comparado ao Bayes Cπ e Bayes Dπ.

Palavras-chave: Bicho mineiro, Cercosporiose, Produção, Capacidade preditiva, *machine learning*

5.1. Introdução

O Brasil é o maior produtor e exportador de café mundial (CONAB, 2021), cujo faturamento bruto em 2021 foi de R$ 42,59 bilhões, onde o *Coffea arabica* corresponde por 75,5% desse montante e ocupa o quarto lugar no ranking das lavouras do país (EMBRAPA, 2022), sendo de fundamental importância para a economia brasileira.

Embora haja um grande potencial para a cultura do café (SANTANA et al., 2018; VELOSO et al., 2020; NASCENTES et al., 2021), a produção dos frutos e a qualidade dos

grãos são fortemente influenciadas por doenças e pragas (VERHAGE et al., 2017; HARVEY et al., 2018). A praga do bicho mineiro (*Leucoptera coffeella*) e a Cercosporiose (*Cercospora coffeicola*) são responsáveis por comprometer a qualidade dos grãos e de gerar perdas na produção do café (BOTELHO et al., 2017; CHAVES et al., 2018; SILVA et al., 2019).

O desenvolvimento de cultivares resistentes ou tolerantes são alternativas para o controle de doenças e pragas, no entanto pelo melhoramento clássico este processo é ainda demorado pois a planta é perene e possui ciclo reprodutivo longo. Desta forma, o melhoramento genético do cafeeiro demanda agilidade e eficácia no lançamento de novas cultivares (RESENDE et al., 2022). Para tanto, pesquisadores têm adotado a seleção genômica (GWS) para acelerar o processo de desenvolvimento de novas cultivares (LIU et al., 2019).

As metodologias estatísticas comumente utilizadas em GWS, tais como RR-BLUP, G-BLUP, Bayes Cπ e Bayes Dπ (MEUWISSEN et al., 2001; HABIER et al., 2011), são baseadas em pressuposições de normalidade dos valores fenotípicos, onde na maioria das ocasiões não se aplicam à dados genômicos. Para contornar essa limitação, existem os métodos estatísticos não paramétricos e algoritmos de *machine learning,* tais como *Bagging* (BA), *Random Forest* (RF) e *Random Forest* Quantílico (QRF) que não requerem suposições a priori sobre a relação funcional entre marcadores e os valores fenotípicos, sem a necessidade de atender pressuposições sobre as distribuições dos dados e dos resíduos.

O BA e o RF já se mostraram eficazes em GWS (BARBOSA et al., 2021; SOUSA et al., 2021, COSTA et al., 2022). Até onde se sabe, não há relatos na literatura sobre o uso do QRF para predição genômica em dados reais, mas o método já foi utilizado em dados simulados para características complexas e, em todos os cenários, as acurácias do QRF foram iguais ou superiores às metodologias avaliadas (VALADARES et al., no prelo). Em outras diversas áreas do conhecimento o QRF também vem sendo aplicado com êxito, confirmando seu alto poder preditivo (TAILLARD et al., 2016; VÖLZ et al., 2016; FANG et al., 2018; KHAN et al., 2019; LIND; ANDERSON, 2019; ROHMER; IDIER; PEDREROS, 2020).

Neste sentido, o objetivo deste trabalho foi: i. propor e avaliar o uso do QRF na predição genômica para características de *Coffea arabica*; ii. comparar as capacidades preditivas e tempo de processamento do QRF com metodologias de *machine learning* (BA e RF), métodos bayesianos (Bayes Cπ e Bayes Dπ) e o G-BLUP.

5.2. Materiais e Métodos

Foram utilizados dados provenientes do programa de melhoramento de *Coffea arabica* da Empresa de Pesquisa Agropecuária de Minas Gerais (Epamig) em parceria com a Universidade Federal de Viçosa (UFV) e a Empresa Brasileira de Pesquisa Agropecuária (Embrapa Café). Os dados foram coletados nos anos 2014, 2015 e 2016, resultantes do cruzamento entre três genitores Catuaí e três genitores Híbrido de Timor, contrastantes em relação à característica ferrugem do cafeeiro. No total, foram consideradas 13 progênies, que são gerações de retrocruzamentos resistentes, retrocruzamentos suscetíveis e F2. Em cada progênie foram selecionados 15 genótipos, totalizando 195 indivíduos que foram genotipados para 20.477 marcadores moleculares SNP (selecionados após análise de qualidade). Mais informações podem ser encontradas em Souza et al. (2019).

5.2.1. Dados fenotípicos

A avaliação de 3 características – 1 contínua (produção de frutos) e 2 categóricas (incidência de cercosporiose e infestação de bicho mineiro) – foi realizada nos 195 indivíduos de *Coffea arabica*. A cercosporiose e o bicho mineiro foram avaliados por notas de 1 a 5, em que se atribuíram a nota 1 aos genótipos assintomáticos e a nota 5 aos genótipos altamente suscetíveis. A produção de frutos foi medida através da quantidade de litros de cerejas frescas colhidas por plantas.

5.2.2. Análises fenotípicas

As análises dos dados fenotípicos foram realizadas considerando modelos lineares mistos (procedimento REML/BLUP). Os parâmetros genéticos foram estimados pela análise individual das 3 características, de acordo com o seguinte modelo estatístico:

$$\boldsymbol{y = Xu + Zg + Wp + Vr + Tb + Ri + \varepsilon},$$

em que**:** $\boldsymbol{y}(195\times1)$ é o vetor de dados; $\boldsymbol{u}(195\times1)$ é o vetor da média geral em cada ano de avaliação; $\boldsymbol{g}(195\times1)$ é o vetor de efeitos de progênie (efeito aleatório); $\boldsymbol{p}$ (195×1) é o efeito permanente entre plantas (efeito aleatório); $\boldsymbol{r}(195\times1)$ é o efeito entre os tipos de população

(efeito aleatório); $\boldsymbol{b}(195\times1)$ é os efeitos entre parcelas (efeito aleatório); $\boldsymbol{i}(195\times1)$ é o efeito da interação progênies x anos (efeito aleatório); $\boldsymbol{\varepsilon}(195\times1)$ é o vetor de resíduos (efeito aleatório). As letras maiúsculas representam as matrizes de incidência para esses efeitos.

5.2.3. Métodos de seleção genômica

Bagging (BA) e Random Forest (RF)

De acordo com James et al. (2013), para se aplicar o *Bagging* (BA) a árvores de regressão, construímos B árvores usando B conjuntos de treinamento obtidos por *bootstrap* e calculamos a média das predições resultantes de modo que isso reduz a variância da predição. O *Random Forest* (RF) segue a mesma ideia, porém, além do conjunto de observações, também altera o número de variáveis preditoras m usadas em cada partição ($m = p/3$ para árvore de regressão, sendo $p = 20.477$ o número total de variáveis preditoras). Em ambos os modelos foram consideradas 500 árvores de regressão (*ntree* = 500).

***Random Forest* Quantílico (QRF)**

O QRF é derivado da combinação entre o RF e a Regressão Quantílica. Para sua construção é necessário a obtenção de T árvores de regressão com base na construção do RF (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). Posteriormente, para cada árvore gerada (T_t), a distribuição condicional é obtida por meio de uma ponderação dos valores observados da característica em estudo. Mais especificamente, dada uma observação, $X = x$, define-se para cada nó terminal (folha da árvore ajustada), $F(x, T_{tf})$, o seguinte fator de ponderação:

$$w_i(x, T_{tf}) = \frac{I_{\{x \in F(x, T_{tf})\}}}{\#\{m: X_m \in F(x, T_{tf})\}}, \text{ sendo } \sum_{i=1}^{n} w_i(x, T_{tf}) = 1,$$

em que $I_{\{X_i \in F(x_i, T_{tf})\}}$ é uma variável indicadora informando que o valor observado ($X = x$) pertence a f-ésima folha e $\#\{m: X_m \in F(x, T_{tf})\}$ representa o número de observações na f-ésima folha.

A predição de uma árvore T_t, de acordo com Meinshausen (2006), para um novo ponto, $X = x_{novo}$, é dada pela média ponderada das observações Y_i, ou seja,

$$\hat{\mu}(x_{novo}) = \sum_{i=1}^{n} w_i(x, T_{tf})Y_i.$$

Desta forma, a predição para uma determinada observação, $X = x$, após a construção de T árvores é dada por:

$$\hat{\mu}_{RF}(x) = \sum_{i=1}^{n} w_i(x)Y_i, \text{ em que } w_i(x) = \frac{1}{T}\sum_{t=1}^{T} w_i(x, T_{tf}).$$

Considerando que a função de distribuição acumulada estimada é dada por:

$$\hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(x) I_{\{Y_i \leq y\}},$$

em que $I_{\{Y_i \leq y\}}$ é uma função indicadora e o valor predito para o τ-ésimo quantil é dado por:

$$Q_\tau(x) = \inf\{y: \hat{F}(y|X = x) \geq \tau\}, \text{ para algum } \tau, \ 0 < \tau < 1.$$

A principal diferença entre QRF e RF é que, para cada nó, em cada árvore, o RF mantém a média das observações que caem nesse nó e negligenciam qualquer outra informação. Já o QRF utiliza os quantis, mantendo o valor de todas as observações do nó e avalia a distribuição condicional com base nessas informações (MEINSHAUSEN, 2006).

Genomic Best Linear Unbiased Prediction - G-BLUP

O modelo para o método REML/G-BLUP, incluindo efeitos aditivos, é dado por:

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

em que $\boldsymbol{y}(195\times1)$ é o vetor de observações fenotípicas; $\boldsymbol{b}(1\times1)$ é o vetor de efeitos fixos com matriz de incidência $\boldsymbol{X}(195\times1)$; $\boldsymbol{u_a}(20.477\times1)$ é o vetor de valores genéticos de aditivos, $\boldsymbol{Z}(195\times20.477)$ é a matriz de incidência para esse vetor e $\boldsymbol{\varepsilon}(195\times1)$ é o vetor de erro aleatório. A estrutura de variância do modelo é dada por $\boldsymbol{u_a}$~N(0, $\boldsymbol{G_a}\sigma^2_{u_a}$), em que $\sigma^2_{u_a}$ é a variância aditiva

e $\boldsymbol{G_a}$(195×195) a matriz de relacionamento genômica para efeitos aditivos. Maiores detalhes da construção de $\boldsymbol{G_a}$ podem ser encontrados em Resende et al. (2014).

Métodos Bayesianos

Habier et al. (2011), propuseram os métodos Bayes Cπ e Bayes Dπ visando limitar a influência de S_m^2 (parâmetro de escala) no fator de *shrinkage* (encolhimento) para os efeitos dos marcadores e tratar π (fração dos marcadores que possuem as mesmas distribuições a *priori*) como um parâmetro desconhecido do modelo, com distribuição a *priori* uniforme (0,1). No Bayes Cπ uma variância comum é especificada para todos os locos, já o Bayes Dπ mantém variâncias específicas para cada loco (RESENDE et al., 2014).

De maneira específica, o Bayes Cπ assume as seguintes distribuições a priori (HABIER et al., 2011):

$$m_i|\pi, \sigma_m^2, \sigma_e^2 \sim (1-\pi)N(0, \sigma_m^2) + \pi N(0, \sigma_m^2 = 0);$$
$$\sigma_m^2|\pi, m_i, \sigma_e^2 \sim \chi^{-2}(v,\ S_m^2);$$
$$\pi|\ m_i, \sigma_m^2, \sigma_e^2 \sim U[0,1],$$

em que m_i são os efeitos dos marcadores (i = 1, ... , 20.477) e são assumidos como amostras de uma distribuição normal com média zero e variância comum σ_m^2 para todos os 20.477 marcadores, ν representa os graus de liberdade, S_m^2 o parâmetro da escala de distribuição (S_m^2 pode ser obtido considerando a esperança de uma variável que segue a distribuição Qui-quadrado invertida escalada) e σ_e^2 a variância do erro.

Já o Bayes Dπ considera uma variância $\sigma_{m_i}^2$ para cada marcador e tem as seguintes distribuições a priori (HABIER et al., 2011):

$$m_i|\pi \sim (1-\pi)N\left(0, \sigma_{m_i}^2\right) + \pi N\left(0, \sigma_{m_i}^2 = 0\right);$$
$$\sigma_{m_i}^2|\pi \sim \chi^{-2}(v,\ S_m^2);$$
$$S_m^2 \sim Gamma(\alpha, \beta);$$
$$\pi|\ m_i, \sigma_{m_i}^2, \sigma_e^2,\ S_m^2 \sim U[0,1].$$

As variâncias genéticas aditiva (σ_a^2)para os métodos do Bayes Cπ e Bayes Dπ são dadas, respectivamente por: $\sigma_a^2 = \sigma_m^2 \sum_{i=1}^{20.477} 2p_i(1\text{-}2p_i)$ e $\sigma_a^2 = \sum_{i=1}^{20.477} 2i(1\text{-}2p_i)\ \sigma_{m_i}^2$, em que, p_i e $(1 - p_i)$, i = 1, ..., 20.477, representam as frequências alélicas.

5.2.4. Avaliação das metodologias

Para obtenção da capacidade preditiva (CP) (coeficiente de correlação de Pearson entre o valor ajustado e o observado) foi realizada a validação cruzada 5-fold. A população, composta por 195 indivíduos, foi dividida em 5 folds, 156 indivíduos foram usados para treinamento e 39 indivíduos foram utilizados para validação. Esse processo foi repetido 5 vezes para que cada parte fosse utilizada uma vez como conjunto de validação. Para cada um dos cinco folds foi calculado a capacidade preditiva de todos os modelos de seleção genômica propostos (BA, Bayes Cπ e Dπ, G-BLUP, QRF nos quantis 0,1, 0,3, 0,5, 0,7 e 0,9 e RF). No final, foi estimado a média e o erro padrão das capacidades preditivas dos cinco folds. Foi também avaliado o tempo de processamento computacional de cada um dos métodos utilizados. Finalmente, estimou-se a concordância entre os métodos considerando 10% dos indivíduos superiores identificados por cada método. No caso de doenças e pragas, considera-se indivíduos superiores os genótipos que são assintomáticos, ou seja, aqueles que apresentaram menores notas; já para a produção, são considerados superiores aqueles que produzem uma maior quantidade de frutos.

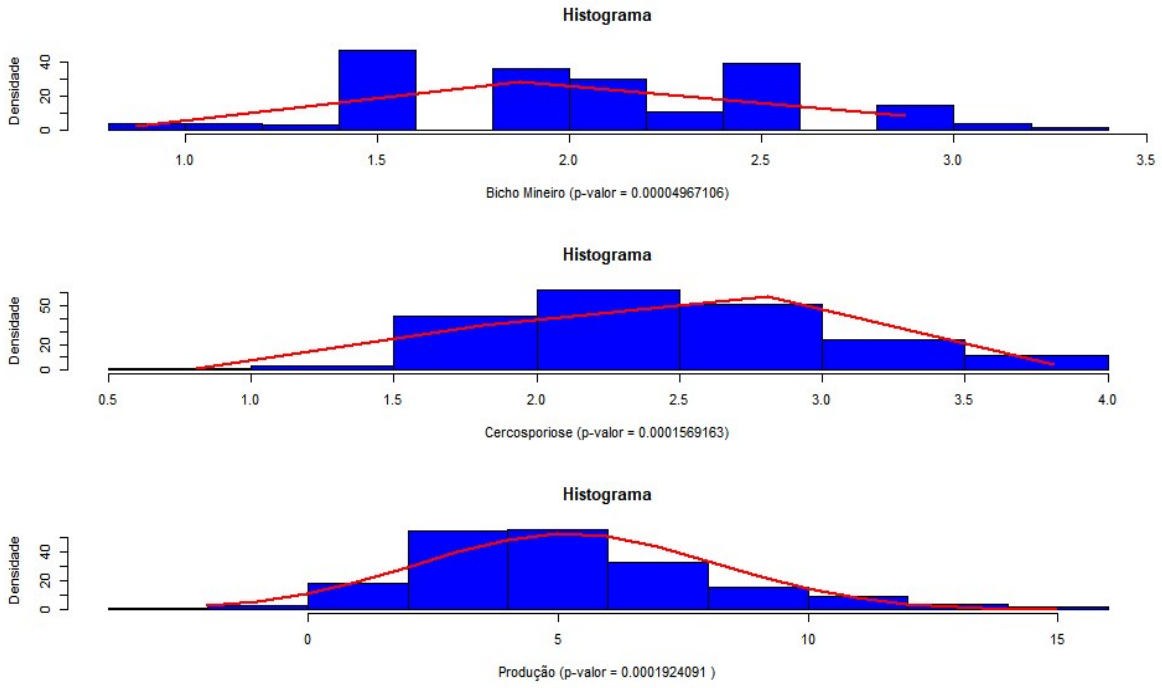
5.2.5. Recursos Computacionais

Os modelos Bayes Cπ e Dπ e o G-BLUP (aditivo) foram ajustados utilizando o software GenomicLand (AZEVEDO et al., 2019). Para os métodos bayesianos foram consideradas 200.000 iterações, dos quais 10.000 foram descartadas (*burn-in*), com intervalo de amostragem igual a 10 (*thin*). As demais análises foram realizadas no software R (R Development Core Team, 2021). Para executar o BA e RF foi utilizado o pacote *randomForest* (LIAW; WIENER, 2002) e para o QRF, ajustado nos quantis 0,1, 0,3, 0,5, 0,7 e 0,9, foi utilizado o pacote *quantregForest* (MEINSHAUSEN, 2017).

5.3. Resultados e Discussões

A Figura 1 apresenta o histograma dos valores fenotípicos ajustados das características em estudo (bicho mineiro, cercosporiose e produção). De acordo com o teste de Shapiro Wilk (SHAPIRO; WILK, 1965), os valores fenotípicos ajustados de nenhuma das características em estudo apresentaram distribuição normal. Nestes cenários, métodos que não necessitam de pressuposições referentes as distribuições de probabilidade dos erros ou dados podem ser interessantes. Nesse sentido, os métodos de aprendizado de máquinas se apresentam como ferramentas alternativas para predição sem a necessidade de se adicionar pressupostos sobre distribuição ou sobre a relação funcional entre entradas e saídas (SANT'ANNA et al., 2020; BARBOSA et al., 2021; SOUSA et al., 2021; COSTA et al., 2022).

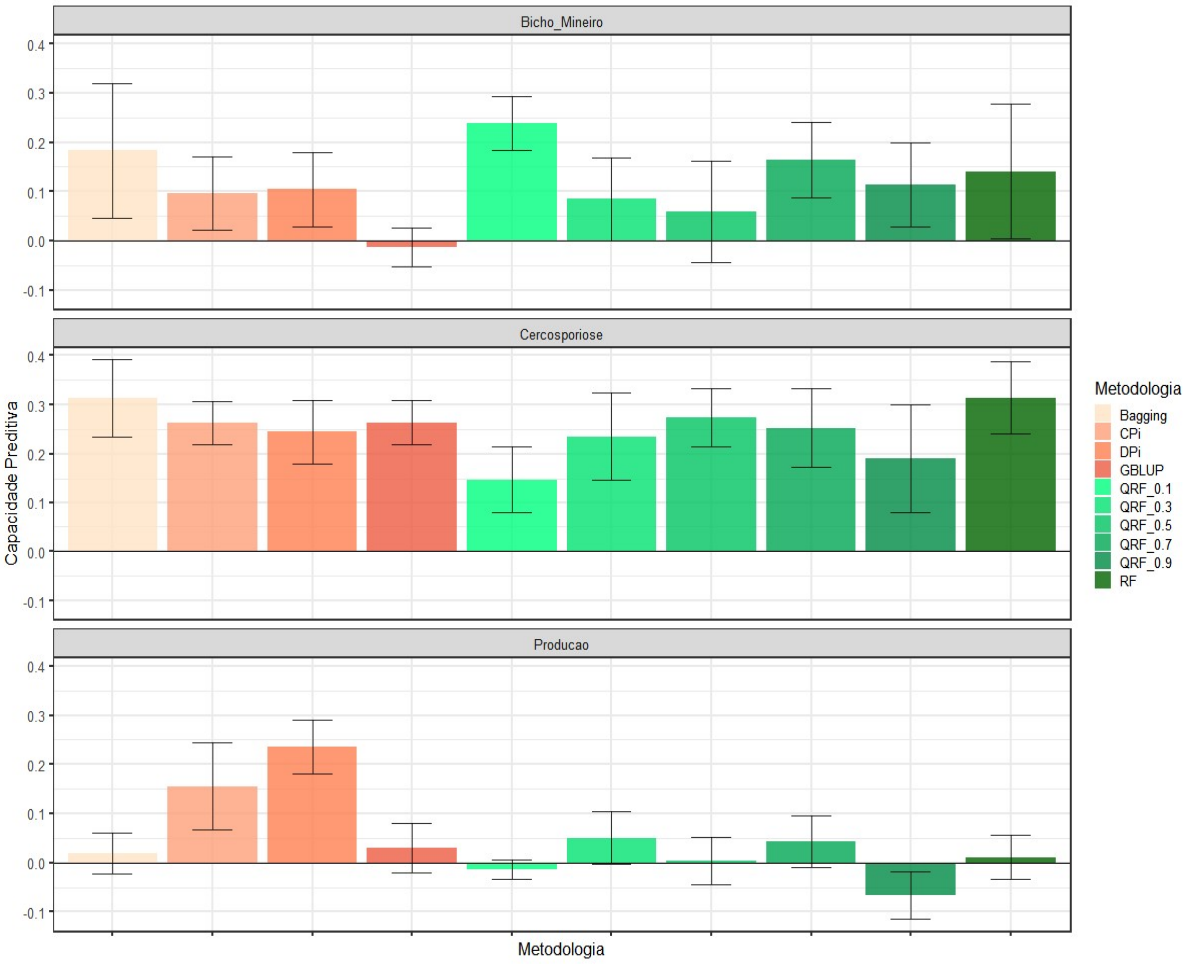
Figura 1: Histograma das características bicho mineiro, cercosporiose e produção de grãos dos dados de *Coffea arabica*, com os respectivos p-valores dos testes de Shapiro Wilk.



Fonte: A autora.

A capacidade preditiva dos métodos de GWS avaliados para as três características de *Coffea arabica,* é apresentada na Figura 2.

Figura 2. Estimativa da capacidade preditiva dos modelos BA, Bayes Cπ (CPi), Bayes Dπ (DPi), G-BLUP com efeito aditivo, QRF nos quantis 0,1, 0,3, 0,5, 0,7 e 0,9 e RF, para três características de *Coffea arabica*: bicho mineiro, cercosporiose e produção.



Fonte: A autora.

De acordo com a Figura 2 (a), o método que apresentou maior CP média para a característica bicho mineiro foi o QRF no quantil 0,1 (0,24). Na Regressão Quantílica, a capacidade de escolher a "melhor" relação entre o fenótipo e os marcadores pode, em algumas situações, aumentar o desempenho preditivo do modelo (NASCIMENTO, 2019). O mesmo se aplica para o QRF já que tem, também, a possiblidade de se estudar toda a distribuição da variável resposta. Portanto, para melhorar a capacidade preditiva do QRF, uma estratégia eficaz é avaliar todas as distribuições fenotípicas para escolher o melhor modelo de ajuste de quantil. Vale ressaltar ainda que a CP do QRF foi mais alta onde os dados fenotípicos estão mais concentrados (Figura 1 e Figura 2(a)). Neste cenário o G-BLUP apresentou capacidade preditiva negativa (Figura 2 (a)).

Para a incidência de Cercosporiose (Figura 2 – (b)), a capacidade preditiva foi similar em todas as metodologias considerando o erro padrão. O BA e RF apresentaram CP maior que 0,3 e, no caso do QRF, a melhor CP foi encontrada no quantil q = 0,5 (0,27). Para esta característica a CP do QRF também foi melhor onde os dados estão mais concentrados e seu gráfico (Figura 2(b)) foi similar ao histograma com seus dados fenotípicos (Figura 1).

A produção de grãos é uma característica controlada por muitos genes e, portanto, a seleção é mais complexa (SOUSA et al., 2019; PAIXÃO et al., 2022). Seu desempenho preditivo não foi bom para as metodologias de *machine learning* e G-BLUP (Figura 2 (c)), apresentando valores de CP próximos a zero e até mesmo negativos. O Bayes Dπ, que apresentou a maior CP neste cenário, é um método que apresenta variância distinta para cada um dos efeitos dos marcadores (HABIER et al., 2011), isso sugere que essa característica é governada por muitos genes com efeitos distintos.

A Tabela 1 apresenta a concordância entre os métodos avaliados (BA, Cπ, Dπ, G-BLUP, QRF e RF) para escolha dos 10% indivíduos superiores. Para o Bicho mineiro, o QRF no quantil q = 0,1 (maior CP) foi mais concordante com o quantil q = 0,3 (71,8%) dentre os quantis avaliados. Em relação as demais metodologias o QRF_0,1 teve uma melhor concordância com o *Bagging* e *Random Forest* (46,1% segundo a Tabela 1) e não foi nada concordante com a metodologia G-BLUP.

Para a característica Cercosporiose, o QRF_0,3, QRF_0,5, QRF_0,7 e QRF_0,9 atingiram concordância entre 76,9% e 87,2% (Tabela 1), sendo estes os quantis nos quais os dados fenotípicos estão mais concentrados (Figura 1). Os métodos Bayes Cπ, Bayes Dπ e G-BLUP tiveram altos valores concordância (entre 66,7% e 82,1%) com o QRF, exceto para o quantil q = 0,1 (Tabela1). Em relação a produção, o maior valor de concordância do método QRF foi entre QRF_0,3 e QRF_0,5 (51,3%) e quando comparado com as demais metodologias, a concordância foi inferior a 30,8%. Isso sugere que as metodologias estão selecionando indivíduos diferentes.

Para os demais métodos avaliados, o BA e RF foram aqueles que apresentaram maiores concordâncias entre os 10% melhores indivíduos para cada característica, segundo a Tabela 1. Especificamente, os valores de concordância foram iguais à 92,3, 87,2 e 92,3 para, respectivamente, bicho mineiro, cercosporiose e produção. Já o G-BLUP apresentou, em geral, menores valores de concordância entre os indivíduos selecionados considerando todas as metodologias.

O tempo de processamento (em segundos) para ajuste/treinamento dos modelos é apresentado na diagonal da Tabela 1. Para todas as características observou-se um tempo de processamento relativamente alto dos métodos Bayesianos, quando comparados ao G-BLUP e QRF.

Tabela 1. Tempo de processamento, em segundos, para ajuste/treinamento (diagonal) e concordância (%) (acima da diagonal) entre os métodos *Bagging* (BA), Bayes Cπ, Bayes Dπ, G-BLUP (GB), *Random Forest* Quantílico nos quantis 0,1, 0,3, 0,5, 0,7, 0,9 (Q_0,1, Q_0,3, Q_0,5, Q_0,7 e Q_0,9, respectivamente) e o *Random Forest* (RF) para os 10% indivíduos superiores, das características bicho mineiro, cercosporiose e produção.

Bicho mineiro										
	BA	Cπ	Dπ	GB	Q_0,1	Q_0,3	Q_0,5	Q_0,7	Q_0,9	RF
BA	2245	10,3	5,1	0,0	46,2	51,3	56,4	35,9	41,0	92,3
Cπ		11212	0,0	46,2	10,3	15,4	15,4	15,4	15,4	10,3
Dπ			12945	15,4	5,1	5,1	5,1	0,0	5,1	5,1
GB				9	0,0	0,0	0,0	5,1	5,1	0,0
Q_0,1					202	71,8	46,2	35,9	61,5	46,2
Q_0,3						202	61,5	46,2	61,5	51,3
Q_0,5							202	61,5	56,4	56,4
Q_0,7								202	51,3	35,9
Q_0,9									202	41,0
RF										715
Cercosporiose										
	BA	Cπ	Dπ	GB	Q_0,1	Q_0,3	Q_0,5	Q_0,7	Q_0,9	RF
BA	2232	41,0	46,2	51,3	20,5	46,2	46,2	41,0	46,2	87,2
Cπ		11046	87,2	66,7	25,6	76,9	76,9	82,1	71,8	41,0
Dπ			14269	71,8	30,8	71,8	71,8	76,9	66,7	51,3
GB				4	25,6	66,7	66,7	71,8	71,8	51,3
Q_0,1					195	30,8	35,9	35,9	41,0	25,6
Q_0,3						195	76,9	82,1	82,1	41,0
Q_0,5							195	87,2	82,1	46,2
Q_0,7								195	82,1	41,0

Q_0,9									195	41,0
RF										986
Produção										
	BA	Cπ	Dπ	GB	Q_0,1	Q_0,3	Q_0,5	Q_0,7	Q_0,9	RF
BA	1888	10,3	10,3	0,0	5,1	10,3	10,3	10,3	10,3	92,3
Cπ		10373	51,3	25,6	20,5	15,4	20,5	20,5	20,5	10,3
Dπ			14471	15,4	10,3	25,6	30,8	20,5	20,5	10,3
GB				62	10,3	20,5	20,5	15,4	15,4	0,0
Q_0,1					174	30,8	15,4	10,3	10,3	5,1
Q_0,3						174	51,3	30,8	25,6	10,3
Q_0,5							174	46,2	25,6	10,3
Q_0,7								174	41,0	10,3
Q_0,9									174	10,3
RF										1004

Fonte: A autora.

Os métodos bayesianos se destacarem para a característica produção, porém apresentaram desvantagem em relação ao tempo de processamento computacional. Em um estudo sobre seleção genômica de ovelhas leiteiras Lacaune, Duchemin et al. (2012), mostraram que o Bayes Cπ obteve acurácia máxima para produção de leite, teor de gordura e escores de células somáticas, porém, o tempo de processamento foi dispendioso. O QRF, independente do quantil, apresentou tempo de processamento bem menor em relação aos métodos bayesianos, podendo ser até 59 vezes inferior em relação ao Bayes Cπ e até 83 vezes inferior ao Bayes Dπ (Tabela 1 – característica produção).

Em síntese, a capacidade preditiva do QRF foi melhor ou igual em relação as demais metodologias avaliadas para o bicho mineiro e cercosporiose, coincidindo com a distribuição dos dados fenotípicos. As maiores concordâncias do QRF ocorreram entre diferentes quantis e para as demais metodologias as técnicas de *machine learning* foram as mais concordantes com o QRF. Além disso, o tempo de processamento computacional foi vantajoso em relação aos métodos bayesianos.

De modo geral, pelos resultados apresentados, é possível concluir que a escolha do método pode levar a diferentes resultados (Tabela 1). Não existe consenso sobre a melhor abordagem (WANG et al., 2019) e a arquitetura genética da característica, herdabilidade, densidade de marcadores desempenham um papel importante na tomada de decisões (LORENZANA; BERNARDO, 2009; DAETWYLER et al., 2010; CLARK et al., 2011; HOWARD et al., 2014).

5.4. Conclusão

O desempenho dos métodos estatísticos utilizados dependeu da característica analisada. O QRF foi igual ou superior para as características bicho mineiro e cercosporiose. O bicho mineiro, por exemplo, apresentou melhor desempenho no quantil q = 0,1. Além disso, o QRF comparado com os métodos bayesianos demandam menos recurso computacional. Baseado nos resultados apresentados, o QRF pode ser, então, uma ferramenta alternativa a ser explorada no contexto de seleção genômica.

5.5. Referências

AZEVEDO, C. F.; NASCIMENTO, M.; FONTES, V. C.; RESENDE, M. D. V. D.; CRUZ, C. D. GenomicLand: Software for genome-wide association studies and genomic prediction. **Acta Scientiarum. Agronomy**, v. 41, 2019.

BARBOSA, I. P.; SILVA, M. J.; COSTA, W. G.; SANT'ANNA, I. C.; NASCIMENTO, M.; CRUZ, C. D. Genome-enabled prediction through machine learning methods considering different levels of trait complexity. Crop Science, v.61: p. 1890–1902, 2021.

BOTELHO, D. M. S.; RESENDE M. L. V.; ANDRADE, V. T.; PEREIRA, A. A.; PATRICIO, F. R. A. Cercosporiosis resistance in coffee germplasm collection. **Euphytica**, v. 213, n. 6, 2017.

CHAVES, E.; POZZA, E. A.; NETO, H. S.; VASCO, G. B.; DORNELAS, G. A. Temporal analysis of brown eye spot of coffee and its response to the interaction of irrigation with phosphorous levels. **J Phytopathol**, v. 166, n. 9, p. 613–622, 2018.

CLARK, S.A.; HICKEY, J.M.; VAN DER WERF, J.H. Different models of genetic variation and their effect on genomic evaluation**. Genetics Selection Evolution**, v. 43, n.18, 2011.

COSTA, W. G. da; CELERI, M. O.; BARBOSA, I. P.; SILVA, G., N.; AZEVEDO, C., F.; BOREM, A.; NASCIMENTO, M.; CRUZ, C. D. Genomic prediction through machine learning and neural networks for traits with epistasis. **Computational and Structural Biotechnology Journal**, v. 20, p. 5490 – 5499, 2022.

CONAB - Companhia nacional de abastecimento. Conjunturas da Agropecuária. Café – Conjuntura Semanal, Brasília, p. 1, agosto 2021, disponível também em: <https://www.conab.gov.br/info-agro/analises-do-mercado-agropecuario-e-extrativista/analises-do-mercado/historico-de-conjunturas-de-cafe/item/download/38698_a63bf49547c465c11cad849018614fcc >. Acesso em: 01 de Set. de 2022.

DAETWYLER, H.D.; PONG-WONG, R.; VILLANUEVA, B.; Woolliams, J.A. The impact of genetic architecture on genome-wide evaluation methods. **Genetics** v. 185, p. 1021–1031, 2010.

DUCHEMIN, S. I.; COLOMBANI, C.; LEGARRA, A.; BALOCHE, G.; LARROQUE, H.; ASTRUC, J. M.; BARILLET, F.; ROBERT-GRANIÉ, C.; MANFREDI, E. Genomic selection in the French Lacaune dairy sheep breed. **Journal of Dairy Science**, v. 95, n. 5, p. 2723–2733, 2012.

EMBRAPA. Agroindústria: Cafés do Brasil geram receita cambial de US$ 5,64 bilhões no ano safra 2016/2017. Disponível em: < https://www.embrapa.br/busca-de-noticias/-/noticia/25326094/cafes-do-brasil-geram-receita-cambial-de-us-564-bilhoes-no-ano-safra-20162017>. Acesso em: 5 Set. 2022.

FANG Y.; XU P.; YANG J.; QIN Y. A quantile regression forest based method to predict drug response and assess prediction reliability. **PLoS ONE**, v. 13, n.10, 2018.

HABIER, D.; FERNANDO, R. L.; KIZILKAYA, K.; GARRICK, D. J. Extension of the bayesian alphabet for genomic selection. **BMC Bioinformatics**, v. 12, n. 186, 2011.

HARVEY, C.A.; SABORIO-RODRÍGUEZ, M.; MARTINEZ-RODRÍGUEZ, M. R.; VIGUERA, B.; CHAIN-GUADARRAMA, A. Climate change impacts and adaptation among smallholder farmers in Central America. **Agriculture & Food Security**, v. 7, n. 57, p. 1-20, 2018.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statiscal Learning: Data Mining, Inference, and Prediction.** Springer, 2. ed., 745p, New York, NY, USA, 2008.

HOWARD, R.; CARRIQUIRY, A.L.; BEAVIS, W.D. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. **G3: Genes, Genomes, Genetics**, v. 4, p. 1027–1046, 2014.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning: with Applications in R.** Springer, New York, NY, USA, 2013.

KHAN, N.; SHAHID, S.; JUNENG, L.; AHMED, K.; ISMAIL, T.; NAWAZ, N. Prediction of heat waves in Pakistan using quantile regression forests. Atmospheric Research, 2019.

LEGARRA, A.; ROBERT-GRANIÉ, C.; CROISEAU, P.; GUILHAUME, F.; FRITZ, S. Improved Lasso for genomic selection. **Genetics Research**, v. 93, p. 77 – 87, 2011.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. R News 2(3), 18—22, 2002.

LIND A. P.; ANDERSON, P. C. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. **PLoS ONE**, v.14, n.7, 2019.

LIU, X.; WANG, H.; HU, X.; LI, K.; LIU, Z.; WU, Y.; Huang, C. Enhancing genomic selection with quantitative trait loci and nonadditive effects revealed by empirical evidence in maize. **Frontiers in plant science**, v. 10, 2019.

LORENZANA, R. E.; BERARDO, R. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. **Theoretical and Applied Genetics,** v.120, p. 151–161, 2009.

MEUWISSEN, T. H. E., HAYES, B. J., GODDARD, M.E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. **Genetics**, v. 157, 2001.

MEINSHAUSEN, N. Quantile regression forests. **Journal of Machine Learning Research**, v. 7, p. 983–999, 2006.

MEINSHAUSEN, N. quantregForest: Quantile Regression Forests. R package version 1.3-7. https://CRAN.R-project.org/package=quantregForest, 2017.

NASCENTES, T. F.; NASCIMENTO, L. G.; FERNANDES, M. I. S.; DUTRA, W. B.; CARVALHO, F. J.; ANDALÓ, V.; GALLET, D. S.; DIAS, W. P.; ASSIS, G. Condições climáticas na incidência de cercosporiose (Cercospora coffeicola) e bicho-mineiro (Leucoptera coffeella) em cultivares de cafeeiros em Monte Carmelo, Minas Gerais, Brasil. **Research, Society and Development**, v. 10, n. 3, 2021.

NASCIMENTO, A.C.; NASCIMENTO, M.; AZEVEDO, C.; SILVA, F.; BARILI, L.; VALE, N.; CARNEIRO, J.E.; CRUZ, C.D.; CARNEIRO, P.C.; SERÃO, N. Quantile Regression Applied to Genome-Enabled Prediction of Traits Related to Flowering Time in the Common Bean. **Agronomy**, v.9, n.12, 2019.

PAIXÃO, P. T. M.; NASCIMENTO, A. C. C.; NASCIMENTO, M.; AZEVEDO, C. F.; OLIVEIRA, G. F.; SILVA, F. L.; CAIXETA, E. T. Factor analysis applied in genomic selection studies in the breeding of *Coffea canephora*. **Euphytica**, v. 218, n.42, 2022.

R Core Team. A Language and Environment for Statistical Computing; Foundation for Statistical Computing: Vienna, Austria, 2021.

RESENDE, M. D. V. de; OLIVEIRA, A. C. B. de; CAIXETA, E. T.; ALKIMIM, E. R.; SOUSA, T. V.; PEREIRA, A. A.; ALVES, R. S.; AZEVEDO, C. F. Tamanho amostral e detecção de genes via GWAS em características quantitativas do cafeeiro. Embrapa, 2022.

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

ROHMER, J.; IDIER, D.; PEDREROS, R. A nuanced quantile random forest approach for fast prediction of a stochastic marine flooding simulator applied to a macrotidal coastal site. **Stoch Environ Res Risk Assess**, v. 34, p. 867–890, 2020.

SANT'ANNA, I. de C.; NASCIMENTO, M.; SILVA, G. N.; CRUZ, C. D.; AZEVEDO, C. F.; GLORIA, L. S.; SILVA, F. F. Genome-enabled prediction of genetic values for using radial basis function neural networks. **Functional Plant Breeding Journal** [Internet], v. 1, p. 2595–9433, 2020.

SANTANA, M. F.; ZAMBOLIM, E. M.; CAIXETA, E. T.; ZAMBOLIM, L. Population genetic structure of the coffee pathogen Hemileia vastatrix in Minas Gerais. Brazil**. Tropical Plant Pathology**, v. 43, n. 5, p. 473–476, 2018.

SHAPIRO, S. S.; WILK, M. B. An Analysis of Variance Test for Normality (Complete Samples). **Biometrika**, v. 52, n. 3/4, pp. 591-611, 1965.

SILVA, M. G.; POZZA, E. A.; CHAVES, E.; NETO H. S.; VASCO, G. B. Spatiotemporal aspects of brown eye spot and nutrients in irrigated coffee. **European Journal of Plant Pathology**, v. 153, n. 3, p. 931–946, 2019.

SOUSA, T. V.; CAIXETA, E. T.; ALKIMIM, E. R.; OLIVEIRA, A. C. B.; PEREIRA, A. A.; SAKIYAMA, N. S.; ZAMBOLIM, L.; RESENDE, M. D. V. Early Selection Enabled by the Implementation of Genomic Selection in Coffea arabica Breeding. **Frontiers in Plant Science**, 2019.

SOUSA, I. C. de; NASCIMENTO, M.; SILVA, G. N.; NASCIMENTO, A. C. C.; CRUZ, C. D.; SILVA, F. F.; ALMEIDA, D. P. de; PESTANA, K. N.; AZEVEDO, C. F.; ZAMBOLIM, L.; CAIXETA, E. T. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. **Scientia Agricola**, v. 78, p. 1, 2021.

SU G.; CHRISTENSEN O. F.; OSTERSEN, T.; HENRYON, M.; LUND, M. S. Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. **PLoS ONE**, v. 7, n. 9, 2012.

TAILLARD, M.; MESTRE, O.; ZAMO M.; NAVEAU P. Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics. **Monthly Weather Review**, v. 144, n. 6, p. 2375–2393, 2016.
VALADARES, C. B.; NASCIMENTO, M.; CELERI, M. O.; NASCIMENTO, A. C. C.; BARROSO, L. M. A.; SANT'ANNA, I. C.; AZEVEVEDO, C. F. Quantile Random Forest for genomic prediction in complex characteristics. **Ciência Rural**, no prelo.

VANRADEN, P. M. Efficient Methods to Compute Genomic Predictions. **Journal of Dairy Science**, v. 91, p. 4414– 4423, 2012.

VERHAGE, F. Y. F; ANTEN, N. P. R; SENTELHAS, P. C. Carbon dioxide fertilization offsets negative impacts of climate change on Arabica coffee yield in Brazil. **Climatic Change**, v.144, n.4, p. 671–685, 2017.

VELOSO, T. G. R.; SILVA, M. C. S. da; CARDOSO, W. S.; GUARÇONI, R. C.; KASUYA, M. C. M.; PEREIRA, L. L. Effects of environmental factors on microbiota of fruits and soil of Coffea arabica in Brazil. **Scientific Reports**, v.10, 2020.

VÖLZ B.; H. MIELENZ, R; Siegwart e J. Nieto, "Predicting pedestrian crossing using Quantile Regression woods", IEEE Intelligent Vehicles Symposium (IV), Gotemburgo, Suécia, p. 426-432, 2016.

WANG X.; MIAO J.; CHANG T.; XIA J.; AN B.; LI, Y.; XU, L.; ZHANG, L.; GAO, X.; LI, J.; GAO, H. Evaluation of GBLUP, BayesB and elastic net for genomic prediction in Chinese Simmental beef cattle. **PLoS ONE**, v. 14, n.2, 2019.

YADAV, S.; WEI, X.; JOYCE, P. et al. Improved genomic prediction of clonal performance in sugarcane by exploiting non-additive genetic effects. **Theoretical and Applied Genetics**, v. 134, p. 2235–2252, 2021.

6. CONCLUSÕES GERAIS

De modo geral, não existe consenso sobre a melhor abordagem para predição de valores genéticos em seleção genômica. Na realidade, a arquitetura genética, herdabilidade, densidade de marcadores desempenham um papel importante na tomada de decisões. Nos dados avaliados neste trabalho o desempenho dos métodos estatísticos utilizados dependeu da característica analisada. O *Random Forest* Quantílico se destacou em diversos cenários, com alguns resultados superiores a metodologias já consolidadas em GWS. Neste sentido o QRF pode ser uma ferramenta alternativa no cenário de GWS.

A principal contribuição científica deste trabalho refere-se à ampliação da base de conhecimento relacionada ao comportamento e potencialidade do *Random Forest* Quantílico em relação aos diferentes cenários de complexidade de características fenotípicas e o desempenho destas em comparação com metodologias já consolidadas no contexto da seleção genômica.

Os conhecimentos gerados por este trabalho poderão ser utilizados como literatura para ajudar no desenvolvimento de pesquisadores em diversas áreas do conhecimento, como na genética e estatística, podendo ajudar a reduzir esforços e recursos financeiros com a aplicação de técnicas mais adequadas e eficientes. Fornece ainda, oportunidades para novas perguntas e questionamentos que devem ser resolvidos em pesquisas futuras a fim de obter um ganho contínuo no conhecimento científico.