MAURÍCIO DE OLIVEIRA CELERI

SPLINES DE REGRESSÃO ADAPTATIVA MULTIVARIADA NA PREDIÇÃO GENÔMICA

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

Orientador: Moysés Nascimento
Coorientadoras: Ana Carolina C. Nascimento
Camila Ferreira Azevedo

VIÇOSA – MINAS GERAIS
2020

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

T

C392s
2020

Celeri, Maurício de Oliveira, 1996-
Splines de regressão adaptativa multivariada na predição genômica / Maurício de Oliveira Celeri. - Viçosa, MG, 2020.
32 f. : il. ; 29 cm.

Orientador: Moysés Nascimento.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Melhoramento genético - Métodos estatísticos. 2. Aprendizado do computador - Métodos estatísticos. 3. Análise de regressão. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.536

MAURÍCIO DE OLIVEIRA CELERI

SPLINES DE REGRESSÃO ADAPTATIVA MULTIVARIADA NA PREDIÇÃO GENÔMICA

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

APROVADA: 27 de outubro de 2020

Assentimento:

Maurício de Oliveira Celeri
Autor

Moysés Nascimento
Orientador

Aos meus Pais, Raquel e Paulo Cezar.

AGRADECIMENTOS

Agradeço aos meus pais Raquel e Paulo Cezar por sempre estarem acreditarem em mim e me apoiarem durante minha trajetória. Aos meus irmãos Alexsandro e Alessandra pelo companheirismo.

Aos amigos de longa data do Instituto Federal do Espírito Santo, em especial Caroline, Vinício e Victor por fazerem parte da minha caminhada na UFV e por todo carinho.

Aos amigos da Matemática – UFV Aline, Fernanda, Patrícia B., Rogério, Ronaldo e Tainara por tornarem a caminhada até o mestrado mais leve e me aguentarem neste novo período da minha vida.

À Patrícia Leite por ser uma amiga que sempre pude contar em todos os momentos. Ao Guilherme por ser o maior presente do 1722 e por sempre acreditar e me incentivar com suas conversar filosóficas que nunca entendia. À Valcívia pelo convívio diário, carinho e cuidado.

Ao grupo de apoio (Amana, Ana Paula, Felipe, Josiane, Josué, Thamires e William) pelos melhores cafés da tarde que um estudante da pós-graduação precisaria. Em especial ao William por ser um verdadeiro irmão que Viçosa me deu e ao Felipe pela convivência diária e pelo carinho e à Josi pelas infinitas risadas.

Aos funcionários do Departamento de Estatística, em especial ao Junior, por ser sempre tão solícito e disposto em ajudar.

Aos docentes do Programa de Pós-graduação em Estatística Aplicada e Biometria pelos ensinamentos.

Aos integrantes dos Laboratório de Inteligência Computacional e Aprendizado Estatístico e Laboratório de Bioinformática pelo companheirismo, amizade, apoio e pelo aprendizado.

Ao meu orientador, Moysés Nascimento, por todo ensinamento, paciência e confiança nessa caminhada nem sempre fácil.

Às minha coorientadoras Ana Carolina Campana Nascimento e Camila Azevedo Ferreira por toda a dedicação e ensinamentos. Aos professores Cosme Damião Cruz e Laís M. Azevedo Barroso pelas contribuições dadas.

A todos que de alguma forma contribuíram para este trabalho.

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

"Os erros são quase sempre de uma natureza sagrada. Nunca tente corrigi-los. Pelo contrário: racionalize-os, compreenda-os a fundo. Depois disso, lhe será possível sublimá-los."

Salvador Dalí

RESUMO

CELERI, Maurício de Oliveira, M.Sc., Universidade Federal de Viçosa, outubro de 2020. **Splines de regressão adaptativa multivariada na predição genômica**. Orientador: Moysés Nascimento. Coorientadoras: Ana Carolina Campana Nascimento e Camila Ferreira Azevedo.

A Seleção Genômica Ampla (SGA), proposta em 2001, é uma variação da Seleção Assistida por Marcadores (SAM) que prediz o valor genético genômico com base em marcadores distribuídos ao longo de todo o genoma. Desde sua proposição vários métodos estatísticos vêm sendo propostos para SGA, dentre os quais podemos destacar o Melhor Preditor Linear não Viesado Genômico (G-BLUP), uma das abordagens mais amplamente difundida devido a sua simplicidade e por permitir considerar efeitos não aditivos em seu ajuste. Uma metodologia ainda não explorada em SGA é a Splines de Regressão Adaptativa Multivariada (MARS). A MARS modela efeitos individuais e possíveis interações entre variáveis preditoras, podendo ser particularmente útil para SGA considerando efeitos não aditivos. Diante disso o objetivo deste trabalho é avaliar a utilização da MARS em SGA, considerando características quantitativas com efeitos não aditivos. Para isso, fez-se uso de um conjunto de dados simulados de 1000 indivíduos com 4010 marcadores SNPs e 12 cenários considerando efeitos não aditivos, definidos pela combinação de oito, 40, 80 ou 120 locus controladores e herdabilidade 0,3, 0,5 ou 0,8. Confrontou-se os resultados da MARS contra os resultados obtidos com o modelo ajustado G-BLUP considerando efeitos aditivos, aditivos e devido a dominância e aditivos e devido a epistasia aditiva-aditiva no que se refere à capacidade preditiva e ao coeficiente kappa de Cohen para seleção de indivíduos superiores. MARS apresentou resultados de capacidade preditiva superior a todos os modelos G-BLUP ajustado para os cenários de herdabilidade 0,3 e resultados semelhantes nos demais casos. Quanto ao coeficiente kappa de Cohen, MARS foi superior aos métodos G-BLUP em 11 dos 12 cenários avaliados. Portanto, MARS é uma metodologia que apresenta potencial para estudos em SGA.

Palavras-chave: Regressão. Aprendizado estatístico. Seleção genômica ampla. Efeitos não aditivos.

ABSTRACT

CELERI, Maurício de Oliveira, M.Sc., Universidade Federal de Viçosa, October, 2020. **Multivariate adaptive regression splines in genome prediction**. Adviser: Moysés Nascimento. Co-advisers: Ana Carolina Campana Nascimento and Camila Ferreira Azevedo.

Proposed in 2001 the Genomic Wide Selection (GWS) is a variation of the Marker Assisted Selection (MAS) that predicts the genomic breeding values based on markers distributed throughout the wide genome. Since its proposition, several statistical methods have been proposed for GWS, among which we can cite the Genomic Best Linear Unbiased Prediction (G-BLUP), one of the most widespread approaches due to its simplicity and because it allows considering non-additive effects in its adjustment . A methodology not yet explored in GWS is the Multivariate Adaptive Regression Splines (MARS). The MARS modeling individual effects and possible interactions between predictor variables and can be particularly useful for GWS considering non-additive effects. Therefore, the aim of this work is to evaluate the use of MARS in GWS, considering quantitative characteristics with non-additive effects. We used a simulated data set of 1000 individuals with 4010 SNPs markers and 12 scenarios considering non-additive effects, defined by the combination of 08, 40, 80 or 120 QTLs and heritability 0.3, 0.5 or 0.8. The results of MARS were compared with the results obtained with the adjusted model G-BLUP considering additive effects, additive and dominance and additives and additive-additive epistasis regarding the predictive capacity and Cohen's kappa coefficient for selection of superior individuals. The MARS presented results of predictive capacity superior to all the G-BLUP models adjusted for the heritability scenarios 0.3 and similar results in the other cases. For Cohen's kappa coefficient, MARS was superior to the G-BLUP methods in 11 of the 12 scenarios evaluated. Therefore, MARS is a methodology that presents potential for studies on GWS.

Keywords: Regression. Statistical learning. Genome Wild Selection. Non-additives effects.

SUMÁRIO

1. **INTRODUÇÃO** 9
2. **REVISÃO DE LITERATURA** 11
2.1.Seleção Genômica Ampla 11
2.2.Melhor Preditor Linear não Viesado Genômico (G-BLUP) 12
2.3.Splines de Regressão Adaptativa Multivariado (MARS) 14
2.4.Referências Bibliográfica 16
3. **PREDIÇÃO GENÔMICA DE CARACTERÍSTICAS NÃO ADITIVAS POR MEIO SPLINES DE REGRESSÃO ADAPTATIVA MULTIVARIADA** 20
Resumo 20
3.1.Introdução 21
3.2.Material e Métodos 22
3.2.1. Dados Simulados 22
3.2.1.1.Genoma simulado 22
3.2.1.2.Cenários avaliados 22
3.2.1.3.Fenótipo simulado 22
3.2.1.4.Recursos computacionais 23
3.2.2. Melhor Preditor Linear Não Viesado Genômico 23
3.2.3. Splines de Regressão Adaptativa Multivariada 24
3.2.4. Comparação de Metodologias 25
3.3.Resultados e Discussão 25
3.3.1. Seleção do modelo MARS 26
3.3.2. Capacidade preditiva e coeficiente kappa de Cohen 26
3.4.Conclusão 29
3.5.Referências Bibliográficas 29

1. INTRODUÇÃO

Segundo ONU (2019) a população mundial aumentará para, aproximadamente 10 bilhões de pessoas até 2050. Frente a isso a produção agrícola precisará aumentar em cerca de 50%, o que é um grande problema, visto que as taxas de crescimento da produção agrícola são baixas e existem áreas com produção estagnadas (HU; LI, 2021; ARAUS; CAIRNS, 2014). Uma das formas de melhorar o rendimento da agricultura é o uso de plantações geneticamente superiores, no entanto o grande gargalo do melhoramento genético é o tempo gasto para o desenvolvimento de variedades melhoradas, tempo de que pode chegar a 20 anos (VOSS-FELS; STAHL; HICKEY, 2019).

No início dos anos 1990 é proposta a Seleção Assistida (MAS) por Marcadores por Lande e Thompson (1990) baseada na existência do desequilíbrio de ligação entre marcadores de DNA e os genes de interesse. No entanto a MAS se mostrou útil para características oligogênicas e de alta de herdabilidade (ALKIMIM et al., 2020) diferentemente das características de interesse agronômico, que, em geral, são complexas e determinadas por um grande número de genes (WANG et al., 2018).

A Seleção Genômica Ampla (SGA), proposta por Meuwissen, Hayes e Goddard (2001), por outro lado, é uma variação da MAS que prediz o efeito de todos os marcadores simultaneamente (RESENDE et al., 2008). A SGA é uma técnica especialmente útil para características de baixa herdabilidade (CROSSA et al., 2017) e para culturas perenes (ALKIMIM et al., 2020). Dentre as vantagens da SGA pode-se destacar o potencial para aumentar o ganho genético por unidade de tempo (GOIS et al., 2016), acelerar os ciclos de reprodução (DESTA; ORTIZ, 2014) e reduzir o custo de desenvolvimento de linhas e híbridos (CROSSA et al., 2017).

Desde a idealização da SGA diversos métodos estatísticos vêm sendo propostos. Um dos métodos mais amplamente utilizados em SGA é o G-BLUP (Melhor Preditor Linear não Viesado Genômico), principalmente vantagem quanto à eficiência computacional (HERNANDEZ; WYATT; MAZOUREK, 2020). Varona et al. (2018) destacam o fato de a maior parte dos métodos consideram apenas efeitos aditivos associados aos marcados, além disso chamam a atenção para a facilidade de incorporar tais efeitos no modelo G-BLUP. Uma alternativa para a incorporação de efeitos não aditivos nas metodologias propostas para SGA são as abordagens baseadas em aprendizado de máquina e aprendizado estatístico (SOUZA et al., 2021).

Uma metodologia alternativa e ainda não explorada em SGA é a Splines de Regressão Adaptativa Multivariada (MARS). A MARS é uma técnica de regressão não paramétrica

proposta para a solução de problemas de alta dimensionalidade, que usa funções base para ajustar a relação existente entre a variável dependente e seus preditores (FRIEDMAN, 1991). Além disso, MARS modela a contribuição de cada variável individualmente e, também, as possíveis interações entre elas (LI et al., 2019). Segundo Motsinger, Ritchie e Reif (2007), MARS apresenta grande potencial para tratar dados discretos e contínuos e identificar interações gene-gene e gene-ambiente. Cook, Zee e Ridker (2004), estudaram a interação gene-gene para AVC isquêmico e observaram que MARS oferece maior flexibilidade ao lidar com efeitos não aditivos quando comparado com as Árvores de Regressão.

Diante disso o objetivo deste trabalho é avaliar a utilização da MARS em SGA, considerando características quantitativas com efeitos não aditivos simuladas com diferentes níveis de herdabilidade e número de QTLs. Os resultados obtidos com a MARS serão comparados com os resultados obtidos na abordagem G-BLUP quanto à capacidade preditiva e o coeficiente kappa de Cohen.

2. REVISÃO DE LITERATURA

2.1. Seleção Genômica Ampla

No início dos anos 1990 Lande e Thompson (1990) propõem a Seleção Assistida por Marcador (MAS, do inglês *Marker-Assisted Selection*) com base no desequilíbrio de ligação existente entre marcadores e genes de interesse para características de interesse agronômico. A MAS mostrou-se eficiente para características oligogênicas com herdabilidade alta ou monogênicas, este fato comprometem a eficiência da MAS para características de interesse agronômico, que são, em geral, quantitativas (ALKIMIM et al., 2020) e possuem baixa herdabilidade (XAVIER; MUIR; RAINEY, 2016) inviabilizando o uso da MAS. Além disso MAS exige o conhecimento prévio da associação entre marcadores e QTLs fato que se torna de difícil acesso em características quantitativas, já que estas são influenciadas pelo ambiente e existem poucos locus de grande efeitos (RESENDE; SILVA; AZEVEDO, 2014).

A Seleção Genômica Ampla (SGA) foi proposta por Meuwissen, Hayes e Goddard (2001) e é uma variação da MAS que prediz o efeito de todos os marcadores simultaneamente (RESENDE et al., 2008). Segundo Resende, Silva e Azevedo (2014), devido aos custos elevados de genotipagem e a restrição de marcadores na época a SGA teve poucos avanços por cerca de seis anos. No entanto, dados a queda dos custos de genotipagem a o aumento dos custos de fenotipagem, a SGA passou a ganhar força (JANNINK; LORENZ; IWATA, 2010). Os marcadores mais usados atualmente são os Polimorfismos de Nucleotídeos Únicos (SNPs) (CROSSA et al., 2017).

A SGA tem grande importância, principalmente, para espécies perenes, devido ao fato do potencial desta técnica de aumentar o ganho genético por unidade de tempo (ALKIMIM et al.,2020; GOIS et al., 2016). Entre outras vantagens da SGA podemos citar a capacidade de selecionar indivíduos mais cedo (JANNINK; LORENZ; IWATA, 2010), é especialmente eficaz para características de herdabilidade baixa (RESENDE; SILVA; AZEVEDO, 2014; GOIS et al., 2016; CROSSA et al., 2017), acelera os ciclos de reprodução (DESTA; ORTIZ, 2014) e reduz o custo de desenvolvimento de linhas e híbridos (CROSSA et al., 2017).

O modelo geral de SGA (MEUWISSEN; HAYES; GODDARD, 2001) é definido por:

$$y_i = \mu + \sum_{j=1}^{p} g_j x_{ij} + \epsilon_i.$$

Onde y_i é o valor fenotípico do indivíduo i, i=1, 2, ..., n, μ é a média geral da características em estudo, g_j é efeito do marcador j, j=1, 2, ..., p, x_{ij} é o elemento da matriz de incidência do marcador j e indivíduo i e ϵ_i é o erro aleatório da observação i.

Como já destacado a SGA prevê, simultaneamente, o efeito de todos os marcadores, entretanto, Desta e Ortiz (2014) atentam ao fato da alta dimensionalidade, isto é, p muito maior do que n, o que pode acarretar problemas de multicolinearidade. Para contornar esse tipo de problema, vários métodos vem sendo propostos (JANNINK; LORENZ; IWATA, 2010).

Azodi et al. (2019) destacam alguns desses métodos, como por exemplo, rrBLUP (ridge regression Best Linear Unbhetic Prediction), BayesA e BayesB (sendo estes três propostos por Meuwissen, Hayes e Goddard, (2001)), LASSO, Rede Elástica e métodos bayesianos, como, por exemplo, LASSO Bayesiano, Baeys Cπ e BaeysDπ. Hernandez, Wyatt e Mazourek (2020) destacam o uso do G-BLUP (Melhor Preditor Linear não Viesado Genômico), uma das abordagens estatísticas mais utilizadas até então, possuindo a grande vantagem de eficiência computacional.

Mais recentemente pode-se notar o uso expressivo de trabalhos relacionados a SGA baseados em aprendizado de máquina e aprendizado estatístico, como por exemplo, redes neurais artificiais (SANT'ANNA et al., 2020) e árvores de regressão e seus refinamentos (SOUZA et al., 2021). Souza et al. (2021) destacam as técnicas de aprendizagem estatística e de máquina para captura de efeitos complexos, como dominância e epistasia. Este fato vai de encontro ao discutido por Varona et al. (2018), que chamam atenção ao fato de poucos estudos levarem em consideração efeitos não aditivos na SGA. Para os autores a inserção de efeitos não aditivos é interessante pois aumentam a precisão da previsão dos valores genéticos, contribuem para definir-se os companheiros entre candidatos para seleção e podem aumentas a variação não aditiva via cruzamentos favoráveis.

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

G-BLUP é umas das técnicas estatísticas mais usadas em SGA (HERNANDEZ; WYATT; MAZOUREK, 2020) e a única diferença desta para o modelo mista padrão é a substituição da matriz de relacionamento do numerador pela matriz de parentesco genômico, além de ainda herdar os fundamentos do modelo infinitesimal do BLUP tradicional (VERONA et al., 2019). Esta metodologia é indicada especialmente para características poligênicas e com resultados contínuos (ALKIMIM et al., 2020).

Segundo Varona et al. (2019) são facilmente incorporadas no procedimento G-BLUP. O modelo completo, considerando efeitos aditivos e não aditivos, para o G-BLUP é dado por:

$$\mathbf{y}=\mathbf{X}\beta+Z\mathbf{u_a}+Z\mathbf{u_d}+Z\mathbf{u_e}+\epsilon,$$

em que **y** (N×1, onde N representa o número de indivíduos) é o vetor de observações fenotípicas, **X** (N×p, p é o número de efeitos fixos) é a matriz de incidência para β, β ($N^{x}1$) é o

vetor de efeitos fixos, $\mathbf{u_a}$ (N×1), $\mathbf{u_d}$ (N×1) e $\mathbf{u_e}$ (N×1) são os efeitos aditivos, devido a dominância e epistático aditivo×aditivo dos indivíduos, respectivamente, $\boldsymbol{\epsilon}$ (N×1) é o vetor de erros aleatórios e Z (N×N) é a matriz de incidência. A estrutura de variância do modelo é dada por $u_a \sim N(0, G_a\sigma^2_{u_a})$, $u_d \sim N(0, G_d\sigma^2_{u_d})$, $u_e \sim N(0, G_e\sigma^2_{u_e})$ e $\epsilon \sim N(0, I\sigma^2_e)$, 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, G_a, G_d e G_e (N×N) são as matrizes de relacionamento genômico para efeitos aditivos, de dominância e de epistasia aditiva×aditiva, respectivamente.

Para a construção das matrizes de relacionamento genômico utilizadas no modelo, considere M_{ij} como sendo a incidência do número de alelos da marca j e do indivíduo i e p_j a frequência do alelo dominante A na marca j. Desta forma, pode-se definir as matrizes W e S dadas por (ZHANG et a., 2019):

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

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

Dessa forma, obtém-se

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

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

$$G_e=G_a\#G_a.$$

Sendo # o operador do produto de Hadamard.

As equações de modelos mistos (RESENDE; SILVA; AZEVEDO, 2014) para o modelo completo é dada por:

$$\begin{bmatrix} X'X & X'Z & X'Z & X'Z \\ Z'X & Z'Z+A^{-1}\lambda_1 & Z'Z & Z'Z \\ Z'X & Z'Z & Z'Z+D^{-1}\lambda_2 & Z'Z \\ Z'X & Z'Z & Z'Z & Z'Z+(A\#A)^{-1}\lambda_3 \end{bmatrix}\begin{bmatrix} \hat{\beta} \\ \hat{u}_a \\ \hat{u}_d \\ \hat{u}_e \end{bmatrix}=\begin{bmatrix} X'y \\ Z'y \\ Z'y \\ Z'y \end{bmatrix}$$

onde $\lambda_1=\frac{\sigma^2_e}{\sigma^2_{u_a}}$, $\lambda_2=\frac{\sigma^2_e}{\sigma^2_{u_d}}$ e $\lambda_3=\frac{\sigma^2_e}{\sigma^2_e}$ e as variâncias são estimadas pelo Método da Máxima Verossimilhança Restrita (REML) pelo algoritmo *Expactation Maximization*.

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. Splines de Regressão Adaptativa Multivariado (MARS)

A *Splines* de Regressão Adaptativa Multivariada (MARS – do inglês *Multivariate Adaptive Regression Splines)* é uma técnica de regressão não paramétrica proposta para a solução de problemas de alta dimensionalidade, que usa funções base para ajustar a relação existente entre a variável dependente e seus preditores (FRIEDMAN, 1991; HUANG et al., 2020). Além disso, MARS modela a contribuição de cada variável individualmente e, também, as possíveis interações entre elas (LI et al., 2019).

As funções de base são funções lineares por partes do tipo $(x\text{-}t)_+$ ou $(t\text{-}x)_+$, definidas por

$$(x\text{-}t)_+ = \text{máx}(0,x\text{-}t) = \begin{cases} x\text{-}t, \text{ se } x>t \\ 0, \text{ caso contrário} \end{cases};$$

$$(t\text{-}x)_+ = \text{máx}(0,t\text{-}x) = \begin{cases} t\text{-}x, \text{ se } x<t \\ 0, \text{ caso contrário} \end{cases}.$$

Estas duas funções, $(x\text{-}t)_+$ ou $(t\text{-}x)_+$, são chamadas de par reflexivo e t é chamado de nó (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). Considerando $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ as N variáveis observadas e que, para cada variável X_j, j=1, 2, …, N, tenha-se n observações x_{j1}, x_{j2}, …, x_{jn}, é possível construir Nn pares reflexivos: $(X_j\text{-}x_{jk})_+$ e $(x_{jk}\text{-}X_j)_+$, com k=1, 2, …, n.

O modelo contendo M termos proposto por Friedman (1991) é

$$\hat{f}(\mathbf{X}) = c_0 + \sum_{i=1}^{M} c_i B_i(\mathbf{X}) + \epsilon.$$

Onde c_0 é o intercepto, $B_i(\mathbf{X})$ é uma função de base, ou produto de funções de base, c_i é o coeficiente de B_i, M é o número de funções base, ou produtos de funções de base, do modelo, definido automaticamente pelo algoritmo MARS[1] (AL-SUDANI et. al, 2019) e ϵ é o erro aleatório.

O grau do modelo é o número máximo de funções de base que permitimos no produto, assim, um modelo de grau dois permite, no máximo, o produto, ou interação, de duas funções de base.

[1] Milborrow (2018) sugere um limite máximo igual a min{200, max{20,n.col(**X**)}}+1, onde n.col(**X**) representa o número de colunas na matriz de dados X.

A estimação dos coeficientes c_0 e c_i baseia-se em minimizar a soma de quadrados do resíduo (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

O algoritmo MARS (FRIEDMAN, 1991; MILBORROW, 2018) é composto por duas por duas etapas: fase *forward* e fase *backward*. Na fase *forward* ocorre o processo de inserção dos pares de funções de base reflexiva. O par reflexivo inserido no modelo é aquele que maximiza a redução da soma de quadrado do resíduo (PARK; KIM, 2018; HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

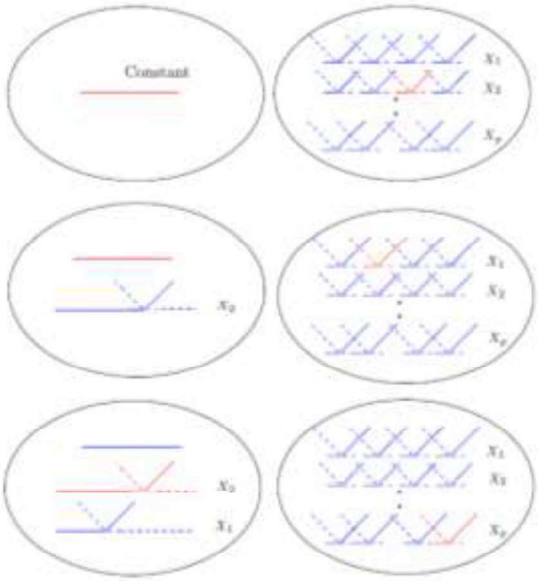

Figura 1. *Etapas da fase forward do modelo MARS (HASTIE; TIBSHIRANI; FRIEDMAN, 2008)*

Para a construção de uma modelo de grau 1 inicialmente consideramos apenas pela constante. O primeiro passo é determinar um par simétrico que minimiza a soma de quadrados do resíduo de ajuste, após a inserção deste par no modelo escolhe-se outro par para ser inserido no modelo. Este processo continua até que seja alcançada uma das condições de parada seja atingida.

Em um modelo de grau superior, a mudança ocorre a partir do segundo passo: após a inserção do primeiro par simétrico as próximas inserções podem ser, tanto de pares simétricos, quanto de produtos de funções de base não presentes no modelo com funções de base que já estão inseridas no modelo. Ao final desta fase é comum que o modelo ajustado apresente *overfits*, o que justifica a fase seguinte.

Milborrow (2018) lista as cinco condições de paradas do algoritmo: (i) número máximo de termos é atingido; (ii) a adição de novos termos altera o R^2 em menos que 0,001; (iii) atingir R^2 maior ou igual a 0,999; (iv) GRSq menor do que -10; ou, (v) nenhum novo termo altera o R^2.

Na fase *backward* é aplicado um processo de exclusão das funções de base que consiste em remover a cada passo a função de base menos importante, isto é, aquela que causa menor incremento no erro quadrático do resíduo, até que o modelo ajustado consista apenas no intercepto, c_0 (LI et al., 2019). MARS usa a Validação Cruzada Generalizada (VCG) para apresentar o melhor modelo de tamanho λ e o modelo final apresentado pelo algoritmo MARS é aquele que minimiza o valor de VCG.

O critério VCG é dado por

$$VCG(\lambda)=\frac{SQR}{\left(1-\frac{N(\lambda)}{n}\right)^2}.$$

Onde SQR é a Soma de Quadrado dos Resíduos, N(λ) é o número efetivo de parâmetros do modelo e n é o número de observações. Hastie, Tibshirani e Friedman (2008) propõem N(λ)= r+ dK, onde r é número de funções de base linearmente independente presente no modelo e d é o fator penalização, de forma geral d=3 em modelos que consideram interações entre variáveis (HASTIE; TIBSHIRANI; FRIEDMAN, 2008; MILBORROW, 2018), e K é o número de nós presente no modelo.

A MARS, além do coeficiente de determinação (R^2), apresenta o coeficiente de determinação generalizado (GRSq), dados por:

$$GRSq=1-\frac{VCG}{VCG_0},$$

em que VCG_0 representa o VCG de um modelo contendo apenas o intercepto.

2.4. Referências Bibliográficas

ALKIMIM, E. R.; CAIXETA, E. T.; SOUSA, T. V.; RESENDE, M. D. V.; DA SILVA, F. L.; SAKIYAMA, N. S.; ZAMBOLIM, L. Selective efficiency of genome-wide selection in Coffea canephora breeding. **Tree Genetics & Genomes**, v. 16, n. 3, p. 1-11, 2020.

AL-SUDANI, Z. A.; SALIH, S. Q.; SHARAFATI, A.; YASEEN, Z. M. Development of multivariate adaptive regression spline integrated with differential evolution model for streamflow simulation. **Journal of Hydrology**, v. 573, p. 1-12, 2019.

ARAUS, J. L.; CAIRNS, J. E. Field high-throughput phenotyping: the new crop breeding frontier. **Trends in plant science**, v. 19, n. 1, p. 52-61, 2014.

AZODI, C. B.; BOLGER, E.; MCCARREN, A.; ROANTREE, M.; DE LOS CAMPOS, G.; SHIU, S. H. Benchmarking parametric and Machine Learning models for genomic prediction of complex traits. **G3: Genes, Genomes, Genetics**, v. 9, n. 11, p. 3691-3702, 2019.

COOK, N. R.; ZEE, R. Y.; RIDKER, P. M. Tree and spline based association analysis of gene–gene interaction models for ischemic stroke. **Statistics in medicine**, v. 23, n. 9, p. 1439-1453, 2004.

CROSSA, J.; PÉREZ-RODRÍGUEZ, P.; CUEVAS, J.; MONTESINOS-LÓPEZ, O.; JARQUÍN, D.; DE LOS CAMPOS, G.; BURGUEÑO, J.; GONZÁLEZ-CAMACHO, J. M.; PÉREZ-ELIZALDE, S.; BEYENE, Y.; DREISIGACKER, S.; SINGH, R.; ZHANG, X.; GOWDA, M.; ROORKIWAL, M.; RUTKOSKI, J.; VARSHNEY, R. K. Genomic selection in plant breeding: methods, models, and perspectives. **Trends in plant science**, v. 22, n. 11, p. 961-975, 2017.

DESTA, Z. A.; ORTIZ, R. Genomic selection: genome-wide prediction in plant improvement. **Trends in plant science**, v. 19, n. 9, p. 592-601, 2014.

FRIEDMAN, J. H. Multivariate Adaptive Regression Splines (with discussion). **The Annals of Statistics**, Beachwood, OH, v. 19, n. 1, p. 1-67, mar. 1991.

GOIS, I. B.; BORÉM, A.; CRISTOFANI-YALY, M.; DE RESENDE, M. D. V.; AZEVEDO, C. F.; BASTIANEL, M.; NOVELLI, V.M.; MACHADO, M. A. Genome wide selection in citrus breeding. **Genetics and Molecular Research**, v. 15, n. 4. DOI http://dx.doi.org/10.4238/gmr15048863.

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

HERNANDEZ, C. O.; WYATT, L.E.; MAZOUREK, M. R. Genomic Prediction and Selection for Fruit Traits in Winter Squash. **G3: Genes, Genomes, Genetics**, v. 10, n. 10, p. 3601-3610, 2020.

HUANG, H.; JI, X.; XIA, F.; HUANG, S.; SHANG, X.; CHEN, H.; ZHANG, M.; DAHLGREN, R. A.; MEI, K. Multivariate adaptive regression splines for estimating riverine constituent concentrations. **Hydrological Processes,** v. 34, n. 16, p. 1213–1227, jul. 2020.

JANNINK, J.; LORENZ, A. J.; IWATA, H. Genomic selection in plant breeding: from theory to practice. **Briefings in functional genomics**, v. 9, n. 2, p. 166-177, 2010.

LANDE, R.; THOMPSON, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. **Genetics**, v. 124, n. 3, p. 743-756, 1990.

LI, D. H.; CHEN, W.; LI, S.; LOU, S. Estimation of hourly global solar radiation using Multivariate Adaptive Regression Spline (MARS)–A case study of Hong Kong. **Energy**, v. 186, p. 115857, 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.

MILBORROW, S. Derived from mda:mars by T. Hastie and R. Tibshirani. earth: Multivariate Adaptive Regression Splines, 2011. R package.

MILBORROW, S. **earth: Multivariate Adaptive Regression Splines**. 2018. R package. Disponível em: https://CRAN.R-project.org/package=earth. Acesso em: 08 novembro 2019.

MONTSINGER, A. A.; RITCHIE, M. D.; REIF, D. M. Novel methods for detecting epistasis in pharmacogenomics studies. **Future Medicine**, v. 8, n. 9, 2007.

MOURA, E. G.; PAMPLONA, A. K. A.; BALESTRE, M. Functional models in genome-wide selection. **PloS one**, v. 14, n. 10, p. e0222699, 2019.

ONU. **População mundial deve ter mais 2 bilhões de pessoas nos próximos 30 anos**. Disponível em: https://news.un.org/pt/story/2019/06/1676601#:~:text=A%20popula%C3%A7%C3%A3o%20mundial%20deve%20aumentar,9%2C7%20bilh%C3%B5es%20em%202050.. Acesso em: 25 fevereiro 2021.

PARK, J.; KIM, J. Defining heatwave thresholds using an inductive machine learning approach. **PLoS ONE**, v. 13, n. 11, nov. 2018. https://doi.org/10.1371/journal.pone.0206872. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0206872. Acesso em: 23 jul. 2020.

RESENDE, M. D. V.; LOPES, P. S.; SILVA, R. L.; PIRES, I. E. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa florestal brasileira**, n. 56, p. 63-63, 2008.

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, MG: UFV, 2014.

SANT'ANNA, I. C.; NASCIMENTO, M.; SILVA, G. N.; CRUZ, C. D.; AZEVEDO, C. F.; GLORIA, L. S.; E SILVA, F. F. Genome-enabled prediction of genetic values for using radial basis function neural networks. **Functional Plant Breeding Journal**, v. 1, n. 2, 2020.

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.

VARONA, L.; LEGARRA, A.; TORO, M. A.; VITEZICA, Z. G. Non-additive effects in genomic selection. **Frontiers in genetics**, v. 9, n. 78, 2018.

VOSS-FELS, K. P.; STAHL, A.; HICKEY, L. T. Q&A: Modern crop breeding for future food security. **BMC biology**, v. 17, n. 1, p. 1-7, 2019.

WANG, X.; XU, Y.; HU, Z.; XU, C. Genomic selection methods for crop improvement: Current status and prospects. **The Crop Journal**, v. 6, n. 4, p. 330-340, 2018.

XAVIER, A.; MUIR, W. M.; RAINEY, K. M. Assessing predictive properties of genome-wide selection in soybeans. **G3: Genes, Genomes, Genetics**, v. 6, n. 8, p. 2611-2616, 2016.

U, H.; LI, J. Short-and long-term challenges in crop breeding. **National Science Review**, v. 8, n. 2, p. nwab002, 2021.

ZHANG, W.; WU, C.; LI, Y.; WANG, L.; SAMUI, P. Assessment of pile drivability using random forest regression and multivariate adaptive regression splines. **Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards**, p. 1-14, 2019. Disponível em: https://www.tandfonline.com/doi/full/10.1080/17499518.2019.1674340. Acesso em: 30 jul. 2020.

3. PREDIÇÃO GENÔMICA DE CARACTERÍSTICAS NÃO ADITIVAS POR MEIO SPLINES DE REGRESSÃO ADAPTATIVA MULTIVARIADA

Resumo: A seleção genômica ampla utiliza marcadores distribuídos por todo o genoma para predizer o valor genético genômico de indivíduos. Esta abordagem possibilita acelerar o processo de melhoramento e aumenta a precisão de predição dos valores genéticos genômicos. Diversas técnicas estatísticas, de aprendizado de máquina e aprendizado estatístico vem sendo propostas para a seleção genômica, dentre elas G-BLUP, BLASSO, redes neurais artificiais e árvores de regressão, uma metodologia ainda não explorada para predição genômica é a splines de regressão adaptativa multivariada (MARS). MARS modela, de forma automática, não linearidades e interações entre as variáveis preditores de um modelo e, além disso, apresenta um modelo ajustado simples e de fácil interpretação. O objetivo deste trabalho é avaliar o uso da MARS para a predição genômica de características não aditivas. Para isso, fez-se uso de um conjunto de dados simulados de 1000 indivíduos com 4010 marcadores SNPs e 12 cenários considerando efeitos não aditivos, definidos pela combinação de oito, 40, 80 ou 120 locus controladores e herdabilidade 0,3, 0,5 ou 0,8. Confrontou-se os resultados da MARS contra os resultados obtidos com o modelo ajustado G-BLUP considerando efeitos aditivos, aditivos e devido a dominância e aditivos e devido a epistasia aditiva-aditiva no que se refere à capacidade preditiva e ao coeficiente kappa de Cohen para seleção de indivíduos superiores. Para isso, fez-se uso de um conjunto de dados simulados de 1000 indivíduos com 4010 marcadores SNPs e 12 cenários considerando efeitos não aditivos, definidos pela combinação de oito, 40, 80 ou 120 locus controladores e herdabilidade 0,3, 0,5 ou 0,8. Confrontou-se os resultados da MARS contra os resultados obtidos com o modelo ajustado G-BLUP considerando efeitos aditivos, aditivos e devido a dominância e aditivos e devido a epistasia aditiva-aditiva no que se refere à capacidade preditiva e ao coeficiente kappa de Cohen para seleção de indivíduos superiores, para obtenção dos resultados utilizou-se a validação cruzada 5-*fold*. Quanto à capacidade preditiva, MARS apresentou resultados superiores a todos os modelos G-BLUP ajustado para os cenários de herdabilidade 0,3 e resultados semelhantes nos demais casos. Já para o coeficiente kappa de Cohen, MARS foi superior aos métodos G-BLUP em 11 dos 12 cenários avaliados. Assim, MARS é uma metodologia que apresenta potencial para estudos em SGA.

Palavras-chave: MARS, Seleção Genômica Ampla, Capacidade preditiva, Aprendizado estatístico, Efeitos não aditivos.

3.1.Introdução

A seleção genômica ampla (SGA), proposta por Meuwissen et al. (2001), utiliza marcadores moleculares distribuídos ao logo de todo o genoma a fim de predizer o mérito genético de plantas e animais. Tal abordagem, 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 a concepção da SGA, várias técnicas estatísticas têm sido propostas como, por exemplo, o *Random Regression Best Linear Unbiased Predictior* (RR-BLUP), Bayes A e Bayes B (MEUWISSEN et al., 2001). Um dos métodos mais amplamente utilizados em SGA é o G-BLUP (Melhor Preditor Linear não Viesado Genômico), principalmente vantagem quanto à eficiência computacional (HERNANDEZ; WYATT; MAZOUREK, 2020). Varona et al. (2018) destacam o fato de a maior parte dos métodos consideram apenas efeitos aditivos associados aos marcados, além disso chamam a atenção para a facilidade de incorporar tais efeitos no modelo G-BLUP.

Recentemente, o interesse em métodos para predição genômica baseados em inteligência computacional e aprendizado de máquinas (AZODI et al., 2019), como por exemplo redes neurais artificiais (SANT'ANNA et al., 2020) e árvores de regressão (SOUZA et al., 2021), vem aumentando. Diferentemente dos métodos usuais de predição genômica, tais metodologias não requerem suposições a priori sobre a relação entre as entradas e a saída. Essa característica permite grande flexibilidade para lidar com diferentes tipos de efeitos não aditivos complexos, como por exemplo, dominância e epistasia (ZINGARET et al., 2020).

Além destas, uma metodologia alternativa e que também não requer a definição sobre a relação funcional entre as entradas (marcadores SNPs) e a saída (valores fenotípicos e genotípicos preditos) é a splines regressão adaptativa multivariada (MARS - *Multivariate Adaptive Regression Splines*) (FRIEDMAN, 1991). A MARS, além de modelar automaticamente não linearidades e interações complexas entre as variáveis de entrada (ADNAN et al., 2020), apresenta, ao final do processo, o modelo ajustado. Neste sentido, a MARS possivelmente aumenta a acurácia de predição para características que apresentem diferentes tipos de efeitos não aditivos complexos e permite ao pesquisador obter informações sobre a arquitetura genética da característica.

Segundo Motsinger, Ritchie e Reif (2007), MARS apresenta grande potencial para tratar dados discretos e contínuos e identificar interações gene-gene e gene-ambiente. Cook, Zee e Ridker (2004), estudaram a interação gene-gene para AVC isquêmico e observaram que MARS

oferece maior flexibilidade ao lidar com efeitos não aditivos quando comparado com as Árvores de Regressão.

Diante do exposto, objetivo deste trabalho é avaliar o uso da *splines* de regressão adaptativa multivariada (MARS) para predição do mérito genético de característica não aditivas. Os resultados obtidos serão comparados com aqueles provenientes G-BLUP.

3.2. Material e Métodos

3.2.1. Dados Simulados

3.2.1.1. Genoma simulado

Foi simulada uma população F2 de uma espécie diploide (2n = 20), com 1000 indivíduos. Considerou-se 4010 marcadores (locos) codominantes de polimorfismos de nucleotídeo único bi alélicos (SNPs), distribuídos igualmente e de forma equidistante em 10 grupos de ligação (cromossomos) com tamanho de 200 cM cada, sendo que os dois últimos grupos não possuíam efeito significativo sobre o fenótipo.

3.2.1.2. Cenários avaliados

Foram simulados 12 cenários com o número de genes controladores da característica iguais a 08, 40, 80 ou 120, distribuídos igualmente entre os oito primeiros grupos de ligação, e herdabilidades de 0,3, 0,5 ou 0,8, conforme apresentado na Tabela 1. Os dois últimos grupos de ligação, contendo 802 dos 4010 marcadores utilizados, não tem influência direta e nem por desequilíbrio de ligação sobre as características e servirá como controle para qualidade dos modelos avaliados.

Tabela 1 – Cenários avaliados

Herdabilidade (h^2)	Número de locus controladores da característica			
	8	40	80	120
0,3	C1	C4	C7	C10
0,5	C2	C5	C8	C11
0,8	C3	C6	C9	C12

3.2.1.3. Fenótipo simulado

As características fenotípicas dos 12 cenários foram simuladas considerando a média (μ) igual a 100 e coeficiente de variação de 10%, com nível de dominância igual a 0,5 e controladas por um modelo de efeito epistático:

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

em que: $\alpha_j = a_i + d_i$ e $d_i/a_i = 0,5$, sendo $\mu + a_i$, $\mu + d_i$ e $\mu - a_i$ os valores genotípicos associados às classes AA, Aa e aa, respectivamente. Para variação do resíduo será considerado a equação, $v_e=((1-h^2)*v_g)/h^2$.

3.2.1.4. Recursos computacionais

A simulação do genoma, cenários e fenótipos foi realizada no software Genes (CRUZ, 2016).

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

O modelo completo, considerando efeitos aditivos e não aditivos, para o G-BLUP é dado por:

$$\mathbf{y}=\mathbf{1}b+Z\mathbf{u_a}+Z\mathbf{u_d}+Z\mathbf{u_e}+\boldsymbol{\epsilon},$$

em que $\mathbf{y}$ (N×1, onde N representa o número de indivíduos) é o vetor de observações fenotípicas, $\mathbf{1}$ (N×1) é um vetor com todos os elementos iguais a 1, b é a média geral, $\mathbf{u_a}$ (N×1), $\mathbf{u_d}$ (N×1) e $\mathbf{u_e}$ (N×1) são os efeitos aditivos, devido a dominância e epistático aditivo×aditivo dos indivíduos, respectivamente, $\boldsymbol{\epsilon}$ (N×1) é o vetor de erros aleatórios e Z (N×N) é a matriz de incidência. A estrutura de variância do modelo é dada por $u_a \sim N(0, G_a\sigma^2_{u_a})$, $u_d \sim N(0, G_d\sigma^2_{u_d})$, $u_e \sim N(0, G_e\sigma^2_{u_e})$ e $\epsilon \sim N(0, I\sigma^2_e)$, 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, G_a, G_d e G_e (N×N) são as matrizes de relacionamento genômico para efeitos aditivos, de dominância e de epistasia aditiva^{x}aditiva, respectivamente.

Para a construção das matrizes de relacionamento genômico utilizadas no modelo, considere M_{ij} como sendo a incidência do número de alelos da marca j e indivíduo i e p_j a frequência do alelo dominante A na marca j. Desta forma, pode-se definir as matrizes W e S dadas por (ZHANG et a., 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

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

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

$$G_e=G_a\#G_a.$$

Sendo que # denota o produto de Hadamard.

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

Neste trabalho consideraremos três modelos reduzidos do modelo apresentado acima: G-BLUP-A (considerando apenas o componente aditivo), G-BLUP-AD (considerando os componentes aditivo e devido a dominância) e G-BLUP-AE (considerando os componentes aditivo e devido à epistasia aditiva×aditiva).

Todos os modelos foram ajustados utilizando o software GenomicLand (AZEVEDO et al., 2019).

3.2.3. Splines de Regressão Adaptativa Multivariada (MARS)

A Splines de Regressão Adaptativa Multivariada é uma técnica de regressão não paramétrica proposta para a solução de problemas de alta dimensionalidade que usa funções base para ajustar a relação existente entre a variável dependente e seus preditores (FRIEDMAN, 1991; HUANG et al., 2020). Além de modelar a contribuição de cada variável individualmente e, também, as possíveis interações entre elas, quando requeridas (AL-SUDANI et al., 2019).

O modelo contendo M termos proposto por Friedman (1991) é dado por:

$$\hat{f}(\mathbf{X})=c_0+\sum_{i=1}^{M} c_i B_i(\mathbf{X})+\epsilon,$$

em que c_0 é o intercepto, $B_i(\mathbf{X})$ é uma função de base, ou produto de funções de base, **X** é o conjunto referente aos dados de marcadores, c_i é o coeficiente de B_i,, com i=1, ..., M, M é o número de termos do modelo, definido automaticamente pelo algoritmo MARS (AL-SUDANI et. al, 2019) e ϵ é o erro. A estimação dos coeficientes c_0 e c_i baseia-se em minimizar a soma de quadrados do resíduo (HASTIE; TIBSHIRANI; FRIEDMAN, 2008).

O algoritmo MARS (FRIEDMAN, 1991; MILBORROW, 2018) é composto por duas etapas: fase *forward* e fase *backward*. Na fase *forward* ocorre o processo de inserção dos pares de funções de base reflexiva. O par reflexivo inserido no modelo é aquele que maximiza a redução da soma de quadrado do resíduo (PARK; KIM, 2018; HASTIE; TIBSHIRANI;

FRIEDMAN, 2008). Na fase *backward* é aplicado um processo de exclusão das funções de base. A MARS usa a Validação Cruzada Generalizada (VCG) para apresentar o melhor modelo de tamanho λ, o modelo final apresentado pelo algoritmo MARS é aquele que minimiza o valor de VCG.

O critério VCG é dado por:

$$VCG(\lambda)=\frac{SQR}{\left(1-\frac{N(\lambda)}{n}\right)^2},$$

em que SQR é a Soma de Quadrado dos Resíduos, N(λ) é o número efetivo de parâmetros do modelo e n é o número de observações. Hastie, Tibshirani e Friedman (2008) propõem N(λ)= r+ cK, onde r é número de funções de base linearmente independente presente no modelo e c é o fator penalização (c=3) e K é o número de nós presente no modelo.

A MARS, além de apresentar o coeficiente de determinação (R^2), apresenta o coeficiente de determinação generalizado (GRSq), dados por:

$$GRSq=1-\frac{VCG}{VCG_0},$$

em que VCG_0 representa o VCG de um modelo contendo apenas o intercepto.
Toda a análise para a *Splines* de Regressão Adaptativa Multivariada foi feita no software R (R Core Team, 2020), versão 4.0.2, com o pacote *earth* (MILBORROW, 2011).

3.2.4. Comparação de Metodologia

A avaliação da predição genômica foi feita usando a abordagem da validação cruzada 5-*fold*. Para cada um dos cinco *folds* foram calculados a capacidade preditiva dos quatro modelos de seleção genômica (MARS, G-BLUP-A, G-BLUP-AD e G-BLUP-AE). A capacidade preditiva é definida como o coeficiente de correlação Pearson entre os valores genéticos simulados dos indivíduos e os preditos pelos modelos ajustados.

Além da capacidade preditiva, o coeficiente kappa de Cohen (COHEN, 1960) foi usado para calcular a porcentagem de concordância dos 10% indivíduos com maiores GEBVs estimados, para cada um dos quatro modelos ajustados, em comum com os 10% indivíduos com maiores GEBVs simulados. O coeficiente kappa de Cohen é definido por:

$$\kappa=\frac{p_0-p_e}{1-p_e},$$

em que p_0 é a concordância observada e p_e é a concordância devida ao acaso.

3.3. Resultados e Discussão

3.3.1. Seleção do modelo MARS

Na Figura 1 são apresentados os valores do coeficiente de determinação generalizado (GRSq) juntamente com o erro padrão associado, para os modelos aditivos e não aditivos. O modelo aditivo apresentou GRSq variando de 0,048 a 0,406, enquanto o modelo não aditivo apresentou GRSq entre 0,088 e 0,550, sendo estes superiores ao do modelo aditivo em todos os cenários avaliados. Dessa forma, selecionou-se o modelo MARS que considera efeitos aditivos e não aditivos para estimação dos valores genéticos genômicos.

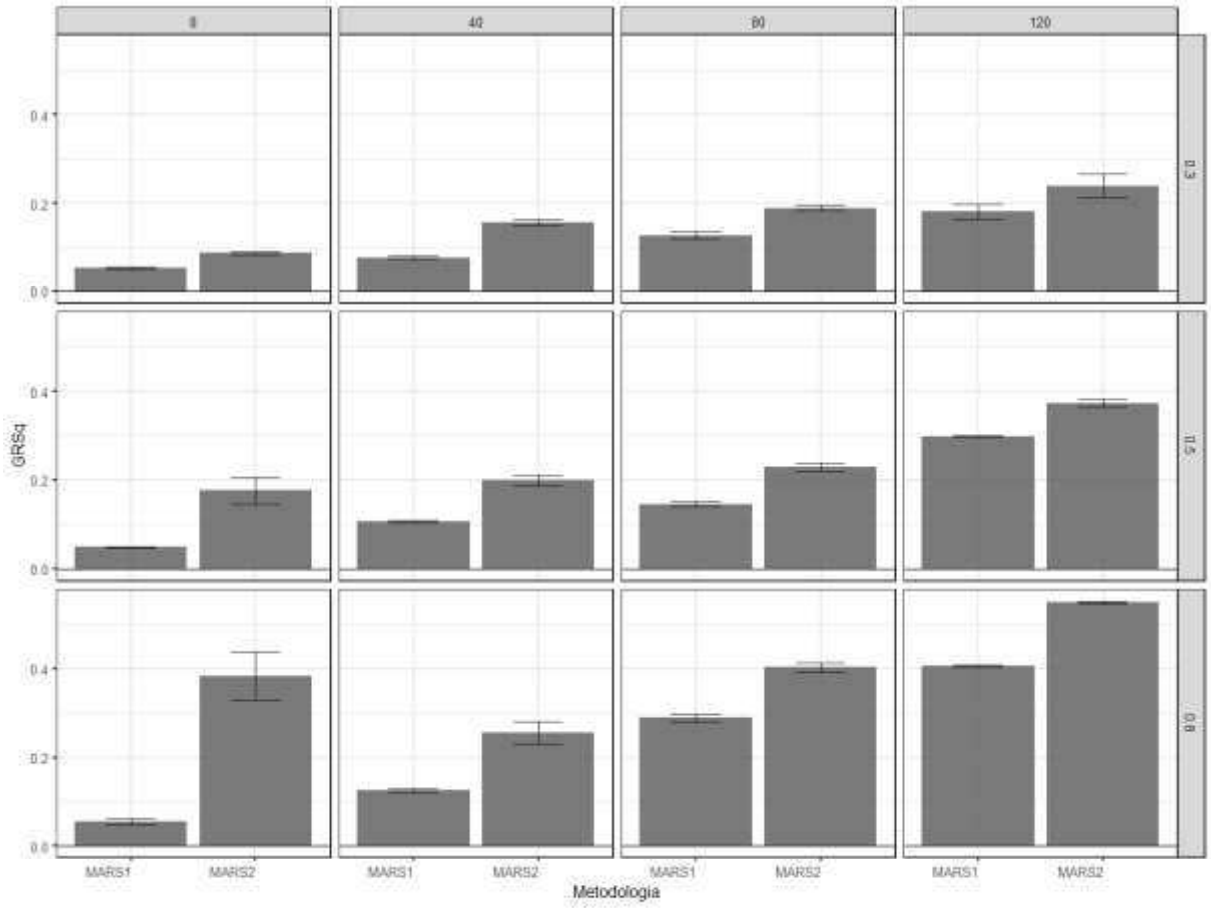

Figura 2 – GRSq dos modelos splines de regressão adaptativa multivariada de grau 1 (MARS1) e splines de regressão adaptativa multivariada de grau 2 (MARS2), para as combinações do número de genes controladores da característica fenotípica (8, 40, 80 e 120) e herdabilidade (0,3, 0,5 e 0,8).

Tal resultado já era esperado, visto o conjunto de dados simulados não considerou nenhuma caraterística fenotípica com arquitetura genética aditiva. Na presença de dominância considerar um modelo estritamente aditivo pode reduzir a capacidade preditiva, afetando assim a classificação dos valores genéticos (TORO; VERONA, 2010).

3.3.2. Capacidade preditiva e coeficiente kappa de Cohen

As estimativas da capacidade preditiva para os métodos G-BLUP-A, G-BLUP-AD, G-BLUP-AE e MARS juntamente com os respetivos erros-padrão, podem ser observados na Figura 2.

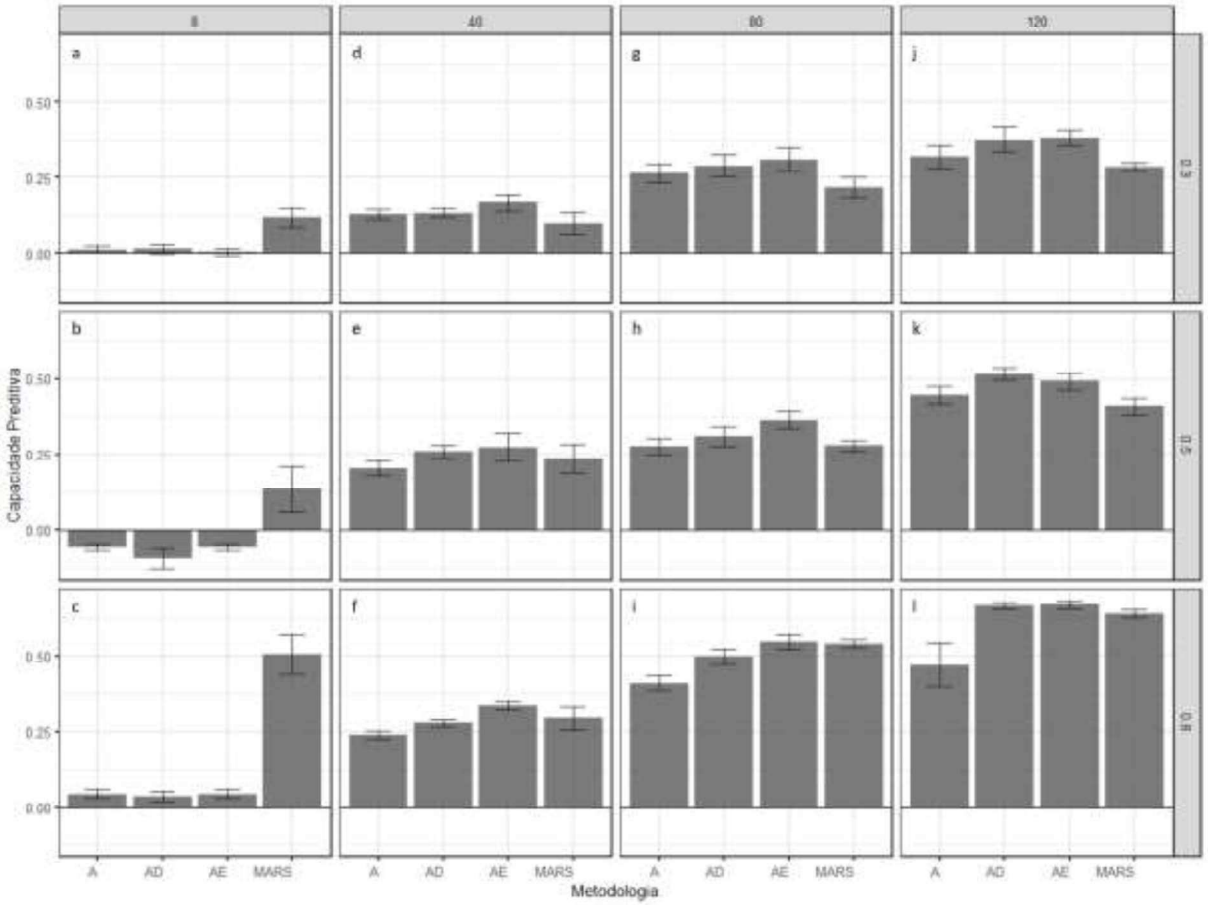

Figura 3 – Estimativa capacidade preditiva dos modelos Melhor Preditor Linear não viesado Genômico com efeito Aditivo (A), Melhor Preditor Linear não viesado Genômico com efeitos Aditivo e Dominante (AD), Melhor Preditor Linear não viesado Genômico com efeitos Aditivo e Epistático (AE) e *Splines* de Regressão Adaptativa Multivariada (MARS) para as combinações do número de genes controladores (8, 40, 80 e 120) e herdabilidade (0,3, 0,5 e 0,8).

De maneira geral, a MARS apresentou capacidade preditiva superior aos modelos ajustados G-BLUP's nos cenários em que foram considerados apenas 8 QTL no controle da característica (Figuras 2 a, b e c). Já para os demais cenários, em que as características eram controladas por pelo menos 40 QTLs, a MARS apresentou resultados semelhantes aqueles obtidos pelos ajustes dos modelos G-BLUP (Figuras 2 d, e, f, g, h, i, j, k e l).

A melhora do desempenho do G-BLUP pode ser explicada, pois em sua construção, é considerado o modelo infinitesimal (Henderson, 1985), ou seja, muito genes controlando a característica igualmente. Por outro lado, nos cenários em que a característica é controlada por poucos QTLs (Figura 2 a, b e c) o ajuste por meio do MARS apresenta maiores valores de

capacidade preditiva visto que a pressuposição do modelo infinitesimal não é plenamente atendida.

De forma geral, baseado no coeficiente kappa de Cohen (Tabela 2), G-BLUP e MARS apresentaram coeficientes variando de não significativo (0-0,20) a fraca (0,40-0,59) (MCHUGH, 2012). No entanto, MARS apresentou coeficiente kappa de Cohen superior aos métodos G-BLUP em quase todos os cenários avaliados, exceto no cenário (h^2=0,5 e 40 QTL), neste caso específico o modelo G-BLUP com efeitos aditivo e epistático apresentou resultado superior. À medida que o número de QTL aumenta para 40, 80 e 120, temos uma tendência de que os métodos G-BLUP sejam comparáveis ou superiores à MARS.

Tabela 2 – Coeficiente kappa de Cohen para seleção de indivíduos

Cenário	h^2	Número de QTL	MARS	G-BLUP-A	G-BLUP-AD	G-BLUP-AE
1	0,3	8	**0.14**	0.04	0.04	0.04
2	0,5	8	**0.23**	0.08	0.08	0.08
3	0,8	8	**0.47**	0.07	0.08	0.07
4	0,3	40	**0.31**	0.20	0.20	0.20
5	0,5	40	0.28	0.27	0.27	**0.29**
6	0,8	40	**0.46**	0.31	0.32	0.31
7	0,3	80	**0.37**	0.27	0.27	0.29
8	0,5	80	**0.32**	0.21	0.20	0.20
9	0,8	80	**0.57**	0.33	0.32	0.31
10	0,3	120	**0.30**	0.21	0.21	0.21
11	0,5	120	**0.44**	0.39	0.40	0.38
12	0,8	120	**0.59**	0.42	0.44	0.48

Splines de Regressão Adaptativa Multivariada (MARS), Melhor Preditor Linear não viesado Genômico com efeito Aditivo (G-BLUP-A), Melhor Preditor Linear não viesado Genômico com efeitos Aditivo e Dominante (G-BLUP-AD) e Melhor Preditor Linear não viesado Genômico com efeitos Aditivo e Epistático (G-BLUP-AE).

Quando compara-se a MARS com o G-BLUP em cenários com maior quantidade de QTL, tem-se uma semelhança expressiva, principalmente no que diz respeito à capacidade preditiva. Este fato pode ser explicado pois, segundo Liu et al. (2019), incluir efeitos não aditivos no modelo de seleção genômica não proporcionaria uma melhora relevante na capacidade preditiva. Ao mesmo tempo em que o aumento no número de QTL implica em uma diminuição significativa da variação devido a epistasia mantendo a variação devido a dominância (VARONA et al., 2018). A MARS pode ter apresentado resultado semelhantes em

capacidade preditiva e superiores no coeficiente kappa de Cohen, neste contexto, devido ao fato de conseguir captar relações complexas com um modelo mais simples (ZHANG et al., 2019).

De forma geral, os métodos de aprendizado de máquina, como a MARS, são interessantes para ajuste de modelos para a predição genômica. Esta classe de métodos permite capturar estruturas complexas, como dominância e epistasia, diretamente do conjunto de dados disponíveis (SOUZA et al., 2021; SANT'ANNA et al., 2019), diferentemente das metodologias usuais, onde a arquitetura genética deve ser informada *a priori*. Além disso, metodologias baseadas em aprendizado de máquina não fazem suposições a respeito da distribuição dos valores fenotípicos observados.

A Predição genômica baseada em métodos de aprendizado de máquina, como por exemplo, redes neurais artificiais, métodos baseados em árvore de decisão e seus refinamentos e máquina de vetor suporte, tem obtido bons resultados para características complexas (AZODI et al., 2019). Até o momento não há relatos sobre o uso da MARS para predição genômica. Deve-se atentar ao fato de que ainda hoje os métodos de predição genômica encontram grandes desafios na detecção de efeitos epistáticos, mesmo os métodos baseados em aprendizado de máquina (MATHEW et al., 2018).

3.4. Conclusão

A MARS se mostrou eficiente para a detecção de efeitos genéticos não aditivos. Além disso, apresentou resultados superiores aos modelos G-BLUP quanto a capacidade preditiva para características controladas por poucos locus e resultados semelhantes para características controladas por mais de 40 locus. Em geral, MARS obteve maior coeficiente kappa de Cohen para seleção dos 10% melhores indivíduos, se mostrando eficaz para seleção genômica.

3.5. Referências Bibliográficas

ADNAN, R. M.; LIANG, Z.; HEDDAM, S.; ZOUNEMAT-KERMANI, M.; KISI, O.; LI, B. Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs. **Journal of Hydrology**, v. 586, 2020.

AL-SUDANI, Z. A.; SALIH, S. Q.; SHARAFATI, A.; YASEEN, Z. M. Development of multivariate adaptive regression spline integrated with differential evolution model for streamflow simulation. **Journal of Hydrology**, v. 573, p. 1-12, 2019.

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.

AZODI, C. B.; BOLGER, E.; MCCARREN, A.; ROANTREE, M.; DE LOS CAMPOS, G.; SHIU, S. H. Benchmarking parametric and Machine Learning models for genomic prediction of complex traits. **G3: Genes, Genomes, Genetics**, v. 9, n. 11, p. 3691-3702, 2019.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and psychological measurement**, v. 20, n. 1, p. 37-46, 1960.

COOK, N. R.; ZEE, R. Y.; RIDKER, P. M. Tree and spline based association analysis of gene–gene interaction models for ischemic stroke. **Statistics in medicine**, v. 23, n. 9, p. 1439-1453, 2004.

CRUZ, C. D. Genes Software - extended and integrated with the R, Matlab and Selegen. **Acta Scientiarum. Agronomy**. vol.38, n.4, 2016.

FRIEDMAN, J. H. Multivariate Adaptive Regression Splines (with discussion). **The Annals of Statistics**, Beachwood, OH, v. 19, n. 1, p. 1-67, mar. 1991.

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

HENDERSON, C. R. Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. **Journal of animal science**, v. 60, n. 1, p. 111-117, 1985.

HUANG, H.; JI, X.; XIA, F.; HUANG, S.; SHANG, X.; CHEN, H.; ZHANG, M.; DAHLGREN, R. A.; MEI, K. Multivariate adaptive regression splines for estimating riverine constituent concentrations. **Hydrological Processes,** v. 34, n. 16, p. 1213–1227, jul. 2020.

JIANG, Y.; REIF, J. C. Modeling epistasis in genomic selection. **Genetics**, v. 201, n. 2, p. 759-768, 2015.

LI, D. H.; CHEN, W.; LI, S.; LOU, S. Estimation of hourly global solar radiation using Multivariate Adaptive Regression Spline (MARS)–A case study of Hong Kong. **Energy**, v. 186, p. 115857, 2019.

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

MATHEW, B.; LÉON, J.; SANNEMANN, W.; & SILLANPÄÄ, M. J. Detection of epistasis for flowering time using Bayesian multilocus estimation in a barley MAGIC population. **Genetics**, v. 208, n. 2, p. 525-536, 2018.

MCHUGH, M., L. Interrater reliability: the kappa statistic. **Biochemia medica: Biochemia medica**, v. 22, n. 3, p. 276-282, 2012.

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.

MILBORROW, S. Derived from mda:mars by T. Hastie and R. Tibshirani. earth: Multivariate Adaptive Regression Splines, 2011. R package.

MILBORROW, S. **earth: Multivariate Adaptive Regression Splines**. 2018. R package. Disponível em: https://CRAN.R-project.org/package=earth. Acesso em: 08 novembro 2019.

MONTSINGER, A. A.; RITCHIE, M. D.; REIF, D. M. Novel methods for detecting epistasis in pharmacogenomics studies. Future Medicine, v. 8, n. 9, 2007.

PARK, J.; KIM, J. Defining heatwave thresholds using an inductive machine learning approach. **PLoS ONE**, v. 13, n. 11, nov. 2018. https://doi.org/10.1371/journal.pone.0206872. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0206872. Acesso em: 23 jul. 2020.

SANT'ANNA, I. C.; NASCIMENTO, M.; SILVA, G. N.; CRUZ, C. D.; AZEVEDO, C. F.; GLORIA, L. S.; E SILVA, F. F. Genome-enabled prediction of genetic values for using radial basis function neural networks. **Functional Plant Breeding Journal**, v. 1, n. 2, 2020.

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

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.

TORO, M. A.; VARONA, L. A note on mate allocation for dominance handling in genomic selection. **Genetics Selection Evolution**, v. 42, n. 1, 2010.

VARONA, L.; LEGARRA, A.; TORO, M. A.; VITEZICA, Z. G. Non-additive effects in genomic selection. **Frontiers in genetics**, *9*, 78, 2018. doi: 10.3389/fgene.2018.00078.

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. Disponível em: https://www.frontiersin.org/articles/10.3389/fgene.2019.00189/full. Acesso em 28 jul. 2020.

ZHANG, W.; WU, C.; LI, Y.; WANG, L.; SAMUI, P. Assessment of pile drivability using random forest regression and multivariate adaptive regression splines. **Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards**, p. 1-14, 2019. Disponível em: https://www.tandfonline.com/doi/full/10.1080/17499518.2019.1674340. Acesso em: 30 jul. 2020.

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.