Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Aplicação de métodos de previsão e classificação em Seleção Genômica

Ana Gabriela P. de Vasconcelos

Orientador: Profa. Dra. Joanlise Marco de Leon Andrade

Brasília

2018

Ana Gabriela P. de Vasconcelos

Aplicação de métodos de previsão e classificação em Seleção Genômica

Relatório apresentado ao Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Orientador: Profa. Dra. Joanlise Marco de Leon Andrade

Brasília

2018

Ana Gabriela P. de Vasconcelos
Aplicação de métodos de previsão e classificação em Seleção Genômica/ Ana Gabriela P. de Vasconcelos. – Brasília, 2018-
62 p. : il.; 30 cm.

Orientador: Profa. Dra. Joanlise Marco de Leon Andrade

Relatório Final – Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística
Trabalho de Conclusão de Curso de Graduação, 2018.

1. Seleção Genômica. 2. Eucalipto. 3. Melhoramento genético. 4. Machine Learning. 5. Regressão Ridge. 6. Validação cruzada. 7. SVM.

Agradecimentos

Primeiramente não só agradeço, como também dedico este trabalho, aos meus pais, minha motivação para me superar a cada dia e dar um passo de cada vez em direção ao meu sonho e propósito de vida. Gostaria de agradecer ao meu pai, minha eterna luz que acompanha cada movimento meu e se orgulha a cada conquista, e à minha mãe, meu suporte, que sempre me deu forças e me ensinou sobre a vida, sendo responsável pela mulher que me tornei.

Agradeço também aos meus amigos que sempre me apoiaram e comemoraram minhas conquistas, desde a época do INDI, do ballet, do Leonardo da Vinci, até os da faculdade. Ao Lucas, meu companheiro que fez parte de uma fase importante na minha vida e sempre esteve ao meu lado, me tornando uma pessoa melhor a cada dia. Ainda, agradeço aos meus professores do ensino médio que sempre despertaram minha paixão pela matemática, mas principalmente aos do curso de Estatística da UnB que definitivamente me tornaram uma profissional capacitada e preparada para os próximos desafios. E, claro, agradeço também a empresa júnior, a ESTAT, por todos os aprendizados não só profissionais, como também pessoais.

Por fim, agradeço ao professor Bernardo Borba de Andrade não só por todo o auxílio que me deu neste trabalho, mas também pelo seu empenho e dedicação durante suas aulas, que definitivamente foram um diferencial no aprendizado dos alunos. Ao Dr. Dario Grattapaglia por nos ceder os dados, auxiliar em diversos aspectos deste trabalho e agregar valor com seus conhecimentos. Aos pesquisadores Joseane Padilha da Silva, Orzenil Bonfim da Silva Junior, Rafael Tassinari Resende e Bruno Marco de Lima por auxiliarem questões essenciais, além de ajudarem nos pontos principais deste trabalho. Porém nada disso seria possível sem a professora Joanlise Marco de Leon Andrade, que como digo aos meus amigos próximos, tem sido o anjo da minha vida. Serei eternamente grata por todo o seu empenho com minha formação, todas as oportunidades que me forneceu, ao seu carinho, preocupação e apoio a todos os meus planos. Muito obrigada, de coração.

Resumo

Programas de melhoramento genético de árvores de floresta visam aumentar a qualidade e ganho econômico de suas plantações por meio de manipulação genética, porém essa tarefa envolve desafios como longos ciclos de cruzamento e altos custos de coleta de diversos fenótipos para largas populações. Nesse sentido, abordagens que avaliam valores genéticos de árvores jovens, sem a necessidade de fenotipagem, possuem o potencial de superar estes desafios. Uma delas é a Seleção Genômica, que consiste em utilizar informações moleculares para estimar efeitos de marcadores genéticos simultaneamente em todo o genoma da população de melhoramento, com base em um modelo de predição. O modelo, desenvolvido em uma população de treinamento com informações genotípicas e fenotípicas, é utilizado para obter os *Genomic Estimated Breeding Values* (GEBVs) baseados em informações apenas genotípicas de plantas candidatas. A análise destes GEBVs pode auxiliar os pesquisadores no processo de tomada de decisões. Portanto, a escolha do modelo é uma etapa essencial para melhorar o ganho genético e a habilidade preditiva. O presente estudo buscou comparar os modelos mistos de regressão e de máquinas de suporte vetoriais (SVMs) em dados de eucaliptos. Além disso estudou-se também fatores que influenciam as métricas obtidas por tais modelos, como características genéticas, qualidade dos fenótipos e efeitos de parentesco. Notou-se que os modelos para os fenótipos com maiores herdabilidades apresentaram medidas de previsão também superiores. Verificou-se que, de maneira geral, utilizar EBVs em vez de fenótipos como resposta do SVM pode acrescentar informações mais confiáveis, levando até, em alguns casos, a métricas superiores. Ainda foi possível verificar a importância de controlar os efeitos de parentesco por meio da validação cruzada para a obtenção de métricas menos otimistas, uma vez que os modelos serão utilizados com dados de novos indivíduos que não estavam presentes na população de treinamento. Por fim, observou-se que os modelos de regressão e de SVM apresentaram resultados consistentes, os quais evidenciaram que sua escolha deve depender do estudo em questão.

Palavras-chave: Seleção genômica, eucalipto, melhoramento genético, machine learning, regressão ridge, validação cruzada, SVM.

Abstract

Tree improvement programs aim to economically increase forest productivity and quality through genetic manipulation. However, this task involves challenges such as lengthy breeding cycles and high costs of phenotyping large progeny trials for several traits. Thus, approaches that evaluate breeding values of trees early in life, without the need to phenotype, have the potential to help overcome these challenges. One of them is Genomic Selection (GS), which consists in using molecular genetic information to estimate marker effects simultaneously across the whole genome of the breeding population, based on a prediction model. The prediction model, developed in a training sample with both genotype and phenotype data, is then used to calculate Genomic Estimated Breeding Values (GEBV) of selection candidates (based only on genotypes in the testing sample), which can guide the breeders during the decision-making process. Therefore, developing GS models is an essential step to improve the genetic gain and the predictive ability. In this study, Ridge Regression models and Support Vector Machines algorithms were compared using data from 999 Eucalyptus trees sampled from a progeny trial in an elite breeding population. Also, factors that can influence metrics obtained by these models were studied, such as quality of measurements of phenotypes and relationship effects. Models for phenotypes with higher heritability showed better prediction ability. In general, using EBVs instead of deregressed phenotypes as SVM's response variable can add more reliable information, leading, in some cases, to higher metrics. Also, was verified the importance of controlling family effects through cross validation to obtain less optimistic predictive measures, since the models will be used to predict data from new individuals, not present in the training population. Finally, both SVM and regression models showed consistent and similar results, which demonstrated that their choice depends on the study.

Keywords: Genomic selection, eucalyptus, improvement program, machine learning, ridge regression, cross validation, SVM.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

BLUE	Melhor estimador linear não viesado
BLUP	Melhor predito linear não viesado
DAP	Diâmetro à altura do peito
EBV	*Estimated Breeding Value*
GEBV	*Genomic Estimated Breeding Value*
IMA	Incremento médio anual
MAF	Frequência do alelo menos comum
NIRS	Espectrometria de infravermelho próximo
RRBLUP	Regressão Ridge BLUP
SNP	Polimorfismo de nucleotídeo único
SVM	Máquina de Suporte Vetorial

Sumário

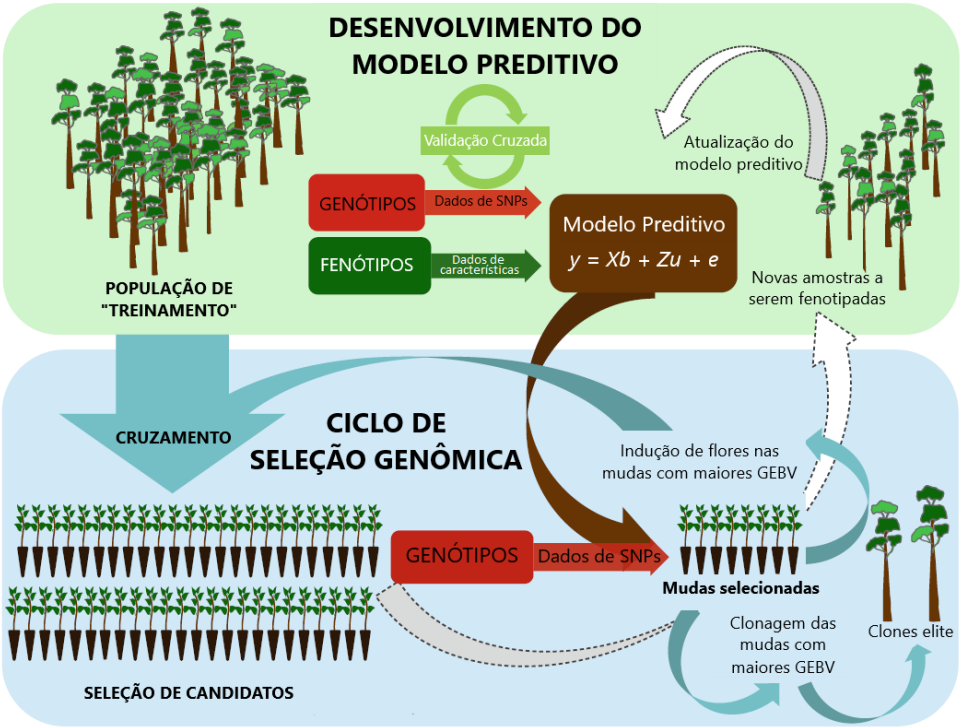
1 Introdução e Justificativa

O melhoramento genético de plantas vem sendo amplamente utilizado no contexto comercial com o objetivo de se obter plantações cada vez mais produtivas, com o menor custo possível. Por muito tempo esse processo foi realizado simplesmente pela observação de características de interesse, seguida da seleção de indivíduos de destaque e da realização de cruzamentos entre eles, em busca de descendentes similares ou ainda melhores. Para tanto, é necessário que as plantas já estejam desenvolvidas a ponto de permitir a comparação das características de interesse, como a altura por exemplo, o que pode levar anos ou até décadas para algumas espécies.

Alguns desafios associados a programas de melhoramento florestal incluem o longo tempo entre ciclos de reprodução e o alto custo e dificuldade de obtenção das características dos indivíduos. Porém, recentemente, novas técnicas de plantio e manutenção permitiram a redução do tempo de coleta dos dados. Além disso, o decaimento do custo de genotipagem de chips de alta densidade, tornaram mais acessível a utilização de modelos de previsão genômica com metodologias estatísticas multivariadas. Com isso, com emprego adequado, a Seleção Genômica (Figura 1) tem o potencial de reduzir o tempo necessário para a obtenção de próximas gerações de plantas, o que reduz custos e recursos envolvidos, principalmente para características de difícil medição.

O processo da Seleção Genômica se inicia com uma população de treinamento composta por árvores adultas, que possuem informação de parte dos seus DNAs obtida por chips de alta densidade e as características observadas (fenótipos) coletadas. Seus dados são utilizados no treinamento de um modelo preditivo baseado em dados genotípicos para prever os *Genomic Estimated Breeding Values (GEBV)*, que representam o quanto da composição genética de um indivíduo contribui para o valor fenotípico da próxima geração. Técnicas de validação cruzada são utilizadas para avaliar o desempenho do modelo. Em seguida no Ciclo de Seleção Genômica, após o cruzamento das plantas adultas, utiliza-se as informações genéticas de plantas jovens na previsão dos GEBVs com base nos modelos obtidos na primeira fase. Os valores preditos são utilizados para ranquear as plantas e as melhores são selecionadas para etapas seguintes do programa de melhoramento. Além disso, após se desenvolverem, pode-se obter as informações fenotípicas das plantas clonadas e adicioná-las à amostra para a atualização do modelo.

Figura 1 – Processo de Seleção Genômica



Fonte – Adaptação de Grattapaglia (2014)

Portanto, a escolha do modelo de predição é uma etapa essencial do processo de melhoramento e deve-se buscar métodos que aperfeiçoem a capacidade preditiva. Diversos métodos de estimação têm sido aplicados nesse contexto incluíndo modelos paramétricos, não paramétricos, clássicos, bayesianos, de regressão ou classificação. Trabalhos como Desta e Ortiz (2014) e Lin, Hayes e Daetwyler (2014) se concentram em comparar diversos modelos e nos fatores que influenciam suas acurácias, como tamanho da população, herdabilidade, número de marcadores e estrutura da população de interesse.

Autores como Ornella et al. (2014) estudaram o comportamento de modelos de Aprendizado de Máquinas como Máquinas de Suporte Vetorial (SVMs) e *Random Forest* em dados de milho e trigo. Nesse artigo, os autores discutem que medidas de avaliação de modelos de regressão não avaliam adequadamente a qualidade do modelo na cauda da distribuição, o que ocorre quando se deseja selecionar os indivíduos com maiores GEBVs. Com isso, mostraram que os algoritmos de classificação apresentam bons resultados em comparação aos de regressão.

O presente trabalho tem por objetivo a implementação de modelos de Regressão Ridge e de classificação SVM para Seleção Genômica em dados de Eucalipto, descritos em Grattapaglia (2014) e Lima (2014). Além disso, foram considerados fatores que podem influenciar na acurácia dos modelos, como a qualidade de coleta dos dados ou a estrutura familiar que faz com que as observações não sejam independentes.

2 Metodologia

2.1 Banco de Dados

Os dados utilizados neste trabalho têm como origem um estudo com delineamento de 8 blocos completos com 5 plantas por parcela, composta por plantas irmãs completas e meio irmãs (LIMA, 2014). Foi selecionada uma subamostra de 1000 árvores pertencentes a 45 famílias, mas como não foi possível obter os dados genéticos de uma delas, a amostra efetiva foi de 999 plantas.

Foram medidos 15 fenótipos (características detectáveis) que podem ser divididos em três grupos: **variáveis de crescimento** como diâmetro à altura do peito (DAP) (cm), altura (m), volume de madeira (m^3) e incremento médio anual (IMA) ($m^3 \cdot ha^{-1}.ano$); **variáveis químicas** como celulose(%), hemicelulose (%), relação Sirigil/Guaiacil (relação S:G), lignina solúvel (%), lignina insolúvel (%) e lignina total (%), e **variáveis físicas** como densidade ($kg \cdot m^{-3}$), ângulo microfibrilar (^{o}), comprimento de fibra (mm), largura de fibra (μm) e rigidez ($mg \cdot 100m^{-1}$).

Devido à complexidade e ao alto custo de coleta dos fenótipos, apenas os de crescimento foram coletados para todas as 999 plantas. As demais variáveis foram obtidas para 350 plantas escolhidas de forma que a variabilidade da amostra seja maximizada (LIMA, 2014). Os fenótipos para as demais 649 foram preditos por modelos de espectroscopia de infravermelho próximo (NIRS), porém tal predição foi bem sucedida apenas para as características químicas e apenas uma física, a densidade. Portanto, as demais variáveis físicas apresentam informação para aproximadamente 350 plantas.

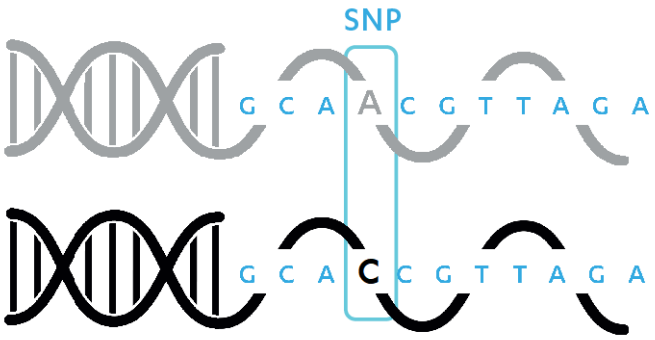
Informações genéticas foram obtidas pela genotipagem de 60.639 marcadores do tipo Polimorfismos de Nucleotídeo Único (*Single Nucleotide Polymorphism* - SNP) utilizando o chip EucHIP60k.br, específico para Eucaliptos (SILVA-JUNIOR et al., 2013) da empresa Illumina. Os SNPs são variações na sequência de DNA em um único nucleotídeo em pelo menos 1% da população e ocorrem em abundância no genoma (Figura 2).

A informação genotípica dos marcadores SNPs é identificada em função de suas bases nitrogenadas (A,T,C e G). Neste trabalho, recodificou-se os genótipos para os valores -1,0 e 1 de acordo com suas frequências em cada marcador. O genótipo homozigoto[1] de menor frequência foi codificado como -1, o genótipo heterozigoto[2] como 0 e o genótipo homozigoto de maior frequência como 1.

[1] Alelos são formas alternativas de um mesmo gene. Um genótipo homozigoto possui ambos os seus alelos iguais.

[2] Um genótipo heterozigoto possui seus alelos diferentes.

Figura 2 – Representação de SNPs



Como medida de controle de qualidade, foram excluídos marcadores com frequência do alelo menos comum (*Minor Allele Frequency* - MAF)[3] inferior a 1% ou com todos os genótipos iguais. Além disso, também foram excluídos marcadores cuja frequência de dados faltantes por inabilidade de genotipagem tenha sido superior a 10%, ou seja aqueles com taxa de genotipagem (*call rate*) inferior a 90%. Por fim, após aplicar o filtro do *call rate*, os demais genótipos faltantes foram imputados utilizando-se o valor médio dos genótipos de cada marcador.

Neste trabalho foram utilizadas variáveis de delineamento, de genótipos e de fenótipos em Modelos Mistos de regressão e no modelo de classificação de Máquinas de Suporte Vetoriais, descritos a seguir.

2.2 Métodos

2.2.1 Modelos Mistos

Modelos mistos apresentam tanto efeitos fixos, relacionados à população toda ou algum nível repetido em fatores de experimento, quanto efeitos aleatórios, associados a unidades de experimentos escolhidas aleatoriamente na população (CAREY; WANG, 2001). Neste trabalho, os efeitos dos marcadores genéticos foram considerados como aleatórios e os efeitos ambientais, ou de delineamento (não genéticos), foram considerados como fixos.

Matricialmente, o modelo misto utilizado para a seleção genômica pode ser descrito pela equação:

[3] As frequências dos alelos podem ser obtidas a partir do percentual dos genótipos, neste caso codificados como -1 e 1 para os homozigotos menos (G_{-1}) e mais comuns (G_1), respectivamente, e 0 para os heterozigoto (G_0). Portanto, a frequência do alelo mais comum (p) se dá pela frequência relativa dos genótipos homozigotos mais comuns somado da frequência dos genótipos heterozigotos divididos por 2, isto é: $p = Fr(G_1) + \frac{Fr(G_0)}{2}$. Já a frequência do alelo menos comum se dá por $q = Fr(G_{-1}) + \frac{Fr(G_0)}{2}$, ou ainda $q = 1 - p$.

$$y = X\beta + Zu + e, \tag{2.1}$$

em que os símbolos são representados a seguir:

- $y_{n\times 1}$: vetor de valores da variável resposta (o fenótipo escolhido);
- $X_{n\times q}$: matriz de incidência de efeitos fixos (informações de blocos e parcelas);
- $\beta_{p\times 1}$: vetor de efeitos fixos;
- $Z_{n\times p}$: matriz de incidência de efeitos aleatórios (informações de marcadores genéticos);
- $u_{p\times 1}$: vetor de efeitos aleatórios;
- $e_{n\times 1}$: vetor de erros residuais;

em que

$$u \sim N(0, \sigma_u^2 \cdot I_{nx1}), \tag{2.2}$$

e

$$e \sim N(0, \sigma_e^2 \cdot I_{nx1}). \tag{2.3}$$

Supondo que $\text{cov}(u_i, e_j) = 0,\ \forall i, j$, tem-se que a variável resposta y segue uma distribuição Normal com média e variância dadas por

$$\mu \equiv \text{E}(y) = X\beta + Z\text{E}(u) + \text{E}(e) = X\beta, \tag{2.4}$$

$$\begin{aligned} V \equiv \text{Var}(y) &= \text{Var}(Zu) + \text{Var}(e) + 2\text{cov}(Zu, e) \\ &= Z\text{Var}(u)Z' + \text{Var}(e) \\ &= Z\sigma_u^2 I Z' + \sigma_e^2 I, \end{aligned} \tag{2.5}$$

isto é,

$$y \sim N(\mu, V). \tag{2.6}$$

Como no Modelo Misto há parâmetros fixos e aleatórios, primeiramente estima-se os efeitos fixos por Máxima Verossimilhança (MV) obtendo-se o Melhor Estimador Linear Não Viesado (*Best Linear Unbiased Estimator (BLUE)*). Em seguida, utiliza-se tais estimativas para prever os parâmetros aleatórios por meio do Melhor Preditor Linear Não Viesado (*Best Linear Unbiased Predictor(BLUP)*).

2.2.1.1 Estimação de efeitos fixos (BLUE) e previsão de efeitos aleatórios (BLUP)

Como descrito anteriormente (relação (2.6)), a variável resposta segue uma distribuição Normal, em que sua função de verossimilhança é dada por

$$L_y(\beta, V) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{1}{2}n}|V|^{\frac{1}{2}}} \exp(-\frac{1}{2}(y_i - X\beta)'V^{-1}(y_i - X\beta)).$$

Ao maximizá-la, obtém-se o Estimador de Máxima Verossimilhança para os efeitos fixos,

$$\hat{\beta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y, \tag{2.7}$$

com sua variância dada por

$$\text{Var}(\hat{\beta}) = (X'\hat{V}^{-1}X)^{-1},$$

em que $\hat{V} = Z\hat{\sigma_u^2}Z' + \hat{\sigma_e^2}I_{1\times n}$ (equação (2.5)), com os Estimadores de Máxima Verossimilhança da variância dos efeitos aleatórios e dos resíduos, $\hat{\sigma_u^2}$ e $\hat{\sigma_e^2}$ respectivamente, obtidos numericamente.

Em seguida, obtém-se as predições para os efeitos aleatórios ao minimizar-se o Erro Quadrático Médio, ou seja

$$\min_{\tilde{u}} E(\tilde{u} - u)^2,$$

em que $\tilde{u}$ é o vetor de efeitos aleatórios predito. Com isso, obtém-se a Melhor Estimativa Linear Não-Viesada dada por

$$\tilde{u} = \sigma_u^2 ZV^{-1}(y - X\hat{\beta}),$$

em que $\hat{\beta}$ é a estimativa obtida na equação (2.7).

Dessa forma as estimativas dos efeitos fixos e aleatórios são encontradas separadamente. Porém, Henderson (HENDERSON, 1963) desenvolveu equações que permitem que tal estimação seja feita simultaneamente como descrito a seguir.

2.2.1.2 Estimação de Henderson

Neste trabalho utilizou-se as equações desenvolvidas por Henderson (1963) que, em contraste com o caso anterior, estima os efeitos fixos e aleatórios simultaneamente ao maximizar a verossimilhança conjunta.

A partir de (2.6), segue que $y|u$ também segue uma distribuição Normal, com média e variância

$$\mathrm{E}(y|u) = X\beta + Zu + \mathrm{E}(e|u) = X\beta + Zu,$$

$$\mathrm{Var}(y|u) = \mathrm{Var}(e|u) = \sigma_e^2 I.$$

Como isto, e como apresentado na equação (2.2), as densidades $f(y|u)$ e $f(u)$ são dadas por

$$f(u) = \frac{1}{(2\pi)^{\frac{1}{2}n}|\sigma_u^2 I|^{\frac{1}{2}}} e^{-\frac{1}{2}y'\sigma_u^{-2}y}, \tag{2.8}$$

$$f(y|u) = \frac{1}{(2\pi)^{\frac{1}{2}n}|\sigma_e^2|^{\frac{1}{2}}} e^{-\frac{1}{2}(y-X\beta-Zu)'\sigma_e^{-2}(y-X\beta-Zu)}, \tag{2.9}$$

a partir das quais obtém-se a função de densidade conjunta

$$f(y,u) = f(y|u)f(u).$$

As equações de estação de máxima verossimilhança são:

$$X'\frac{1}{\sigma_e^2}X\tilde{\beta} + X'\frac{1}{\sigma_e^2}Z\tilde{u} = X'\frac{1}{\sigma_e^2}y$$

e

$$Z'\frac{1}{\sigma_e^2}X\tilde{\beta} + (Z'\frac{1}{\sigma_e^2}Z + \frac{1}{\sigma_u^2}I)\tilde{u} = Z'\frac{1}{\sigma_e^2}y,$$

descritas matricialmente como

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \dfrac{\sigma_e^2}{\sigma_u^2}I \end{bmatrix} \begin{bmatrix} \tilde{\beta} \\ \tilde{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}.$$

Nota-se que, há um deslocamento na diagonal da matriz $Z'Z$ de $\dfrac{\sigma_e^2}{\sigma_u^2}$, o que se assemelha ao que ocorre na estimação dos coeficientes da Regressão Ridge (SEARLE; CASELLA; MCCULLOCH, 2009).

Nesse estudo os modelos mistos foram utilizados nos contextos diferentes apresentados a seguir.

2.2.1.3 Regressão Ridge BLUP e BLUP Fenotípico com matriz A estimada e realizada

A chamada Regressão Ridge BLUP (RRBLUP) utiliza informações genotípicas para prever os GEBVs[4] a partir da equação (2.1), porém considera-se que o vetor de efeitos aleatórios u siga uma distribuição Normal com média zero e variância constante σ_u^2.

Já no modelo BLUP fenotípico, utiliza-se informações apenas de parentesco entre indivíduos para prever os GEBVs. A modelagem do BLUP fenotípico difere da do RRBLUP por ter a matriz de incidência de efeitos aleatórios Z considerada como a identidade e o vetor de efeitos aleatório segue uma distribuição com média zero e variância $\sigma_u^2 A_{n\times n}$, em que A é chamada de matriz de relacionamento aditiva. Esta matriz de relacionamento pode ser descrita de duas formas, realizada e estimada, descritas a seguir.

A matriz de relacionamento realizada estima coeficientes que verificam a relação genética entre os indivíduos a partir dos dados genotípicos. Ela é obtida a partir de (VANRADEN, 2008):

$$G = \frac{WW'}{2\sum_M p_M(1-p_M)}, \tag{2.10}$$

em que $W_{iM} = Z_{iM} - 2p_M + 1$, com Z sendo a matriz dos genótipos e p_M a frequência do alelo mais comum no marcador M.

Já a matriz de relacionamento estimada verifica a relação genética entre os indivíduos a partir da ancestralidade declarada da planta. Para obter tal matriz, inicialmente assume-se que os primeiro descendentes, cujos pais são desconhecidos, foram escolhidos aleatoriamente portanto possuem relação zero. Além disso, a diagonal da matriz é toda igual a 1. Em seguida, como a amostra em estudo é apenas de uma geração, considera-se que um indivíduo possui relacionamento genômico de 0,5 com seus pais e irmãos completos e 0,25 com seus meio-irmãos. O modelo de BLUP fenotípico utilizando esta matriz corresponde ao modelo de melhoramento genético tradicional, que não utiliza informações genotípicas, apenas fenotípicas e de pedigree.

O modelo RRBLUP fornece uma medida que descreve a proporção da variabilidade total fenotípica, devido à variância genética aditiva chamada de herdabilidade, ou seja, quanto do fenótipo é herdável. Esta medida pode ser obtida por

$$h^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_e^2}, \tag{2.11}$$

sendo p_M a frequência do alelo mais comum no marcador M e $\hat{\sigma}_T^2, \hat{\sigma}_g^2$ e $\hat{\sigma}_e^2$ as variâncias estimadas total, genética e dos resíduos, respectivamente. A variância genética se dá por (GIANOLA et al., 2009)

[4] Os GEBVs (*Genomic Breeding Values*) modelam quanto da composição genética de um indivíduo contribui para o valor fenotípico da próxima geração e são obtidos por $GEBV = Z\tilde{u}$.

$$\hat{\sigma}_g^2 = \sum_M 2p_M(1-p_M)\cdot\hat{\sigma}_u^2,$$

em que σ_u^2 no caso de RRBLUP corresponde à variância do vetor de efeitos aleatórios.

A herdabilidade é uma forma de avaliar indiretamente a eficácia do modelo de regressão, porém ela é influenciada pelo tamanho da amostra e estrutura da população de estudo.

Além da herdabilidade, utilizou-se a capacidade preditiva para avaliar o modelo. Ela corresponde à média da correlação de Pearson entre o GEBV obtido a partir da equação (2.1) (GEBV_i) e o fenótipo observado ($y_i^{(\text{obs})}$) em cada grupo dos k grupos da validação cruzada, vistos na seção 2.2.3:

$$r_{y\hat{y}} = \frac{\sum_{i=1}^{k} corr(GEBV_i, y_i^{(\text{obs})})}{k}. \tag{2.12}$$

Nos modelos mistos, descritos até então, são utilizadas informações genéticas para prever os GEBVs, que são variáveis contínuas. Porém outra abordagem possível é a utilização de modelos de classificação, em que os fenótipos são agrupados em dois grupos, quais sejam elite e não elite, e a partir dos dados genotípicos busca-se o ranqueamento dos indivíduos. Neste trabalho, utilizou-se Máquinas de Suporte Vetorial para este fim e sua técnica será descrita a seguir.

2.2.2 Máquinas de Suporte Vetorial (SVMs)

A Máquina de Suporte Vetorial (*Support Vector Machine (SVM)*), desenvolvida por Cortes e Vapnik (1995), é uma técnica de aprendizado de máquinas que permite a classificação de observações a partir de uma regra de decisão. A ideia inicial envolve representar os dados em um espaço em que seja possível encontrar fronteiras lineares que permitam a separação das classes. Esta separação é obtida a partir de um hiperplano ótimo, definido pela função de decisão linear, que maximize a distância entre os vetores de suporte de duas classes, ou seja, maximize a margem.

2.2.2.1 SVM Linear

Inicialmente separa-se os dados em um grupo de teste e outro de treinamento. Tem-se então uma base de treinamento com N observações que possuem um vetor de p variáveis explicativas x_i com $x_i \in R^p$ e uma variável resposta y_i dicotômica, isto é $y_i \in \{-1, 1\}$. Assim como no caso de Modelos Mistos, as variáveis explicativas são os genótipos e a variável reposta é o fenótipo, ou os GEBVs de um modelo de regressão ajustado anteriormente, que deve ser dicotomizado seguindo alguma decisão.

Busca-se separar as observações em dois grupos a partir de um hiperplano ótimo definido por

$$\{x : f(x) = x'\beta + \beta_0 = 0\}. \tag{2.13}$$

Com isso, considera-se que as observações são linearmente separáveis quando o hiperplano é capaz de diferenciá-las a partir de

$$\begin{cases} x'\beta + \beta_0 \geq 1, \ se \ y_i = 1 \\ x'\beta + \beta_0 \leq -1, \ se \ y_i = -1 \end{cases}, \tag{2.14}$$

o que corresponde a $y_i(x'\beta + \beta_0) \geq 1$.

O hiperplano ótimo garante que todos os pontos estão a pelo menos uma distância de $2M$ uns dos outros (como visto na Figura 3a). Busca-se maximizar esta margem obtendo-se os coeficientes β e β_0 que a maximizem. Porém como $M = \dfrac{1}{||\beta||}$ maximizar a largura M corresponde a minimizar $||\beta||$ como em

$$\min_{\beta,\beta_0} \quad \frac{||\beta||^2}{2} \quad \text{sujeito a} \quad y_i(x_i'\beta + \beta_0) > 1, \ \forall i.$$

Esta determinação de valores extremos sujeita à restrições pode ser solucionada com o método dos Multiplicadores de Lagrange (α_i)

$$L_P = \frac{1}{2}||\beta||^2 - \sum_{i=1}^{N} \alpha_i[y_i(\beta x_i + \beta_0) - 1]. \tag{2.15}$$

Ao minimizar-se a equação (2.15), derivando-se em relação à β e à β_0 e igualando-se a zero, obtém-se as relações:

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i, \tag{2.16}$$

$$0 = \sum_{i=1}^{N} \alpha_i y_i. \tag{2.17}$$

Ao se substituir tais condições na função lagrangiana (2.15) tem-se o problema de otimização dual:

$$\begin{aligned} \max_{\alpha} \quad & L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle, \\ s.a. \quad & \sum_{i=1}^{N} \alpha_i y_i = 0, \quad e \quad \alpha_i \geq 0. \end{aligned} \tag{2.18}$$

Nota-se que no problema dual os dados estão representados apenas na forma de um produto escalar, o que facilita o processo de otimização e simplifica o esforço computacional. A solução da equação (2.18) deve satisfazer as condições de Karush-Kuhn-Tucker dadas pelas relações (2.16), (2.17) e

$$\alpha_i[y_i(x_i'\beta + \beta_0) - 1] = 0, \ \forall i. \tag{2.19}$$

Com isso, a estimativa coeficiente β se dá por $\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i$ (equação (2.16)), em que $\hat{\alpha}_i$ é obtido pela solução do dual. Já o coeficiente $\hat{\beta}_0$ é obtido a partir da equação (2.19).

Verifica-se da relação (2.19) que se $\hat{\alpha_i} > 0$ então $y_i(x_i'\hat{\beta} + \hat{\beta}_0) = 1$ o vetor de variáveis explicativas da observação i (x_i) estará exatamente na restrição da margem, também chamados de vetores de suporte. Já os vetores que não pertencem a essa reta possuem o multiplicador de Lagrange estimados $(\hat{\alpha_i})$ nulo.

Dadas as estimativas $\hat{\beta}, \hat{\beta}_0$ e $\hat{\alpha}$, o hiperplano (visto na equação (2.13)) estimado será

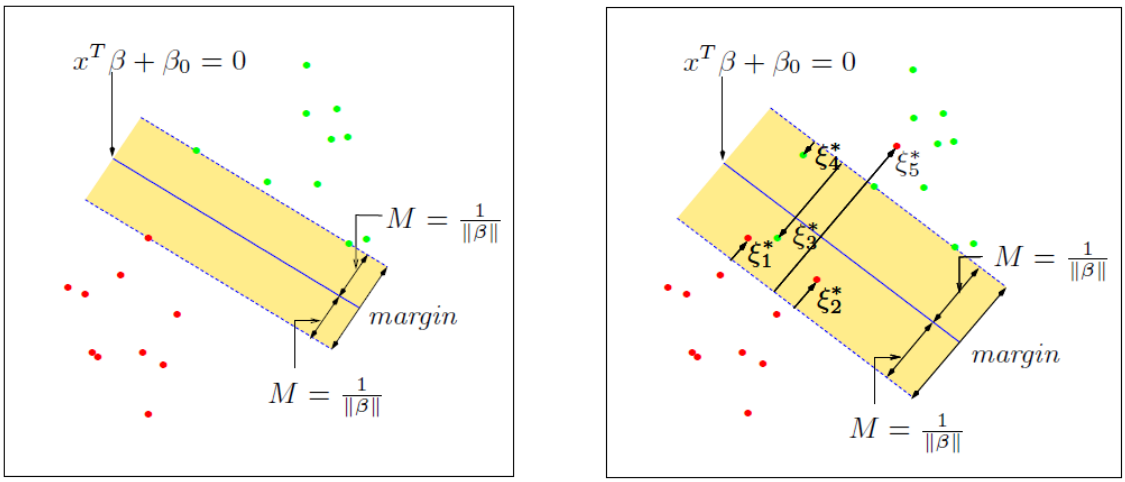
$$\{x : \hat{f(x)} = x'\hat{\beta} + \hat{\beta}_0 = 0\}. \tag{2.20}$$

Com isso tem-se a regra de classificação da observação:

$$\hat{G}(x) = \text{sinal}(\hat{f}(x)) = \begin{cases} 1, \ se \ x'\hat{\beta} + \hat{\beta}_0 > 0 \\ -1, \ se \ x'\hat{\beta} + \hat{\beta}_0 < 0 \end{cases} .$$

Nota-se pela especificação deste problema que nenhum dado de treinamento estará do lado incorreto ou dentro da margem. Porém, ao se aplicar o algoritmo nos dados de teste isso pode ocorrer; por isso busca-se margens o mais largas possível para evitar erros de classificação. Esta especificação é conhecida como SVM com Margens Rígidas (Figura 3a). Entretanto há outra abordagem que permite erros de classificação no treinamento, chamada de SVM com Margens Suaves (Figura 3b).

Figura 3 – Máquinas de Suporte Vetorial (SVM)



(a) Margem Rígida (b) Margem Suave

Fonte – Reprodução de Hastie, Tibshirani e Friedman (2002)

2.2.2.2 SVM Linear com Margens Suaves

O SVM com Margens Suaves utiliza variáveis de folga definidas por $\xi = (\xi_1, \xi_2, ..., \xi_N)$, tal que $\xi_i \geq 0, \forall i$. Neste contexto, a restrição (2.14) passa a ser $y_i(x_i'\beta + \beta_0) > 1 - \xi_i$ e tem-se o problema de otimização, que não só encontra uma solução única e maximiza a margem, como também minimiza a soma dos erros atribuído pelas variáveis de folga (ξ_i) como em:

$$\min_{\beta,\beta_0,\xi} \quad \frac{||\beta||^2}{2} + C\sum_{i=1}^{N}\xi_i,$$
$$s.a. \qquad y_i(x_i'\beta + \beta_0) \geq 1 - \xi_i \ \forall i, \quad e \quad \xi_i \geq 0.$$

A constante de regularização C atribui um peso a essa minimização, pois além de minimizar o erro das classificações incorretas ($x_i > 1$), também minimiza aquelas observações classificadas corretamente mas além da margem ($0 < \xi_i < 1$). Ou seja, valores muito pequenos de C permitem diversos erros de classificação, porém valores muito grandes desta constante geram uma perda de generalização do modelo e aumentam o tempo de processamento. Com isso é necessário se selecionar um valor de C que generalize bem, mas possua um erro de treinamento baixo.

Considerando as variáveis de folga, a função lagrangiana é descrita por

$$L_P = \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i[y_i(\beta x_i + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^{N}\mu_i\xi_i. \tag{2.21}$$

Ao minimizá-la obtém-se as relações

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i, \tag{2.22}$$

$$0 = \sum_{i=1}^{N} \alpha_i y_i, \tag{2.23}$$

$$\alpha_i = C - \mu_i, \ \forall i. \tag{2.24}$$

Substituindo as equações (2.22), (2.23) e (2.24) na função lagrangiana (2.21), o problema de otimização pode ser representado pelo dual:

$$\begin{aligned} \max_{\alpha} \quad & L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\ & s.a. \qquad \sum_{i=1}^{N} \alpha_i y_i = 0 \quad e \quad \alpha_i \geq 0, \end{aligned} \tag{2.25}$$

que deve satisfazer às condições de Karush-Kuhn-Tucker, que correspondem às equações (2.22), (2.23) e (2.24) e às constantes

$$\alpha_i [y_i(\beta x_i + \beta_0) - (1 - \xi_i)] = 0, \tag{2.26}$$

$$y_i(\beta x_i + \beta_0) - (1 - \xi_i) \geq 0, \tag{2.27}$$

$$\mu_i \xi_i = 0 \tag{2.28}$$

O coeficiente $\hat{\beta}$ é obtido a partir da relação (2.22) e $\hat{\beta}_0$ pode ser obtido a partir das equações (2.26) a (2.28). No SVM com margens suaves, os vetores de suporte, que estão exatamente na margem e possuem $\hat{\alpha}_i > 0$ são aqueles em que $y_i(\hat{\beta} x_i + \hat{\beta}_0) = (1 - \hat{\xi}_i)$. Alguns desses pontos de suporte cairão dentro da margem ($\hat{\xi}_i = 0$), então das relações (2.24) e (2.28) tem-se que $0 < \hat{\alpha}_i < C$. Os demais pontos possuem $\hat{\xi}_i > 0$ e $\hat{\alpha}_i = C$

A partir da solução do problema dual, a regra de decisão se dá da mesma forma que o SVM com margens Rígidas:

$$\hat{G}(x) = sinal(\hat{f}(x)) = sinal(x' \hat{\beta} + \hat{\beta}_0).$$

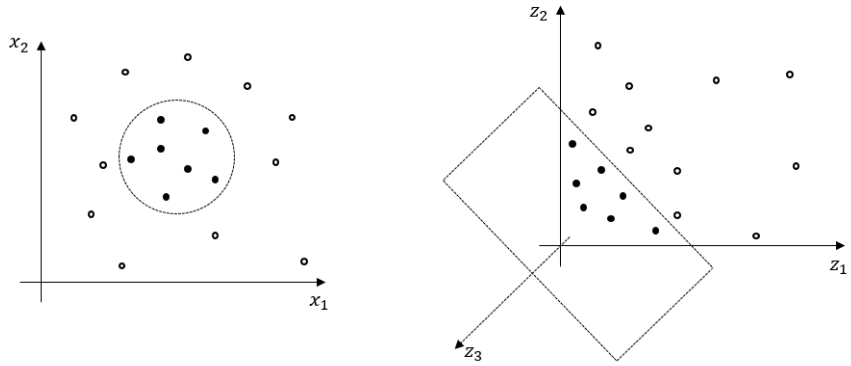
Ambos os casos apresentados até aqui supõem que os dados são linearmente separáveis. Porém em muitas situações os dados não possuem este comportamento e classificações não lineares seriam mais adequadas.

2.2.2.3 SVM Não Lineares

Na teoria do SVM Linear os dados são representados na função pelo produto escalar $\langle x_i, x_j \rangle$, que é linear. Portanto, ao se utilizar uma função $h(x)$ não-linear tem-se uma

classificação também não linear. Com isto, a dimensão é expandida e modificada até que os dados possam ser linearmente separados, como na Figura 4.

Figura 4 – Representação de um Kernel



Para $h(x)$, uma função de transformação realizada em x nos dados de treinamento tem-se a função (não-linear) $\hat{f}(x) = h(x)'\hat{\beta} + \hat{\beta}_0$. Então, o problema dual descrito pela equação (2.18) fica na forma

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle h(x_i), h(x_j) \rangle.$$

Como $\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i h(x_i)$, segue que

$$\hat{f}(x) = h(x)'\hat{\beta} + \hat{\beta}_0 = \sum_{i=1}^{N} \hat{\alpha}_i y_i \langle h(x), h(x_i) \rangle + \hat{\beta}_0.$$

Nota-se então que os dados são representados por meio do vetor escalar de uma função de transformação. Porém, por se tratar de dimensões maiores, até infinita, o cálculo destas funções se torna inviável. Neste caso, não é necessário se conhecer $h(x)$; basta utilizar-se alguma função Kernel dada por:

$$K(x_i, x_j) = \langle h(x_i), h(x_j) \rangle.$$

Neste trabalho, utilizou-se os núcleos

- **Linear**: $K(x_i, x_j) = \langle x_i, x_j \rangle + c$ e
- **Radial**: $K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2)$.

Com isso, o hiperplano ótimo estimado 2.13 pode ser representada por $\hat{f}(x) = \sum_{i=1}^{N} \hat{\alpha}_i y_i K(x_i, x_j) + \hat{\beta}_0$ com a regra de classificação $\hat{G}(x) = sinal(\hat{f}(x))$. Porém, ocorre

que esta regra de classificação não fornece a probabilidade dos dados pertencerem a cada classe, o que muitas vezes é necessário. Wahba et al. (1999) propuseram uma maneira de obter tais probabilidades para cada indivíduo por meio da equação:

$$\hat{P}(Y_i = 1|x_i) = \frac{1}{1 + \exp(-\hat{f}(x_i))}$$

Como o SVM não fornece previsões dos GEBVs, por ser um algoritmo de classificação e não de regressão, utilizou-se como medidas de avaliação o Kappa de Cohen e a Eficiência Relativa.

2.2.2.4 Métrica de Avaliação para modelos de classificação

- **Kappa de Cohen:** verifica a concordância e reprodutibilidade de duas avaliações qualitativas. No contexto de Seleção Genômica, verifica-se a concordância entre a classificação definida como elite e a predita, com base nas probabilidades de pertencerem à classe de plantas elite. Considera-se o grupo das plantas com maiores medidas e o de plantas com menores medidas, definidos a partir de um ponto de corte feito nos fenótipos contínuos. Após a predição e classificação dos indivíduos, os valores preditos e declarados são organizados como na Tabela 1.

Tabela 1 – Tabela de Confusão entre os dois grupos

		Preditos		
		Maiores	**Menores**	**Total**
Definidos	**Maiores**	n_{aa}	n_{ab}	o_a
	Menores	n_{ba}	n_{bb}	o_b
	Total	m_a	m_b	n_{tot}

O coeficiente Kappa é obtido por:

$$K = \frac{P_o - P_e}{1 - P_e},$$

sendo

$$P_o = \frac{n_{aa} + n_{bb}}{n_{tot}} \quad e$$
$$P_e = \frac{m_a}{n_{tot}}\frac{o_a}{n_{tot}} + \frac{m_b}{n_{tot}}\frac{o_b}{n_{tot}}.$$

Com base no valor do coeficiente interpreta-se que:

  - $K < 0$: O predito foi menor do que o esperado;
  - $0 \leq K < 0,2$: Concordância muito ruim;

  - $0,2 \leq K < 0,4$: Concordância fraca;
  - $0,4 \leq K < 0,6$: Concordância moderada;
  - $0,6 \leq K < 0,8$: Concordância boa;
  - $0,8 \leq K < 1$: Concordância ótima;
  - $K = 1$: Concordância perfeita.

- **Eficiência Relativa:** É uma medida *ad hoc* que indica o ganho genético esperado devido à escolha dos indivíduos por Seleção Genômica em relação à escolha feita pelo modelo de BLUP fenotípico.

$$ER = \frac{\mu_{\alpha'} - \mu_{Teste}}{\mu_{\alpha} - \mu_{Teste}},$$

em que μ_{Teste} representa a média do grupo teste em geral e μ_{α} e $\mu_{\alpha'}$ são as médias dos fenótipos observados e preditos dos melhores indivíduos ranqueados, respectivamente. No caso do algoritmo de classificação, o ranqueamento é feito com base na probabilidade do indivíduo pertencer ao grupo de plantas com maiores medidas.

2.2.3 Validação Cruzada

Tanto para os modelos mistos quanto para algoritmos de classificação deve-se utilizar algum tipo de validação cruzada para avaliar a habilidade de previsão do modelo. Neste projeto foi utilizada a metodologia do K-fold, em que o conjunto de dados é separado em K grupos e utiliza-se K-1 para treinamento do modelo e o restante como teste, repetindo-se esse processo até que todos os grupos tenham sido utilizados como teste.

Porém, como em dados de famílias há dependência entre os indivíduos, essa divisão dos K grupos pode influenciar na medida de avaliação gerando métricas muito otimistas, além de violar as suposições de independência dos modelos (ROBERTS et al., 2017). Como posteriormente os modelos serão utilizados para prever informações de novos indivíduos, não presentes da população de treinamento, é importante que o modelo forneça medidas próximas da realidade e que os erros não sejam subestimados. Isto é, busca-se minimizar os efeitos parentais na validação do modelo, para que as métricas reflitam principalmente os efeitos genéticos.

Com isto, utilizou-se duas estratégias de separação dos grupos do K-fold. A primeira foi uma separação aleatória dos indivíduos para cada grupo. A segunda foi a utilização da classificação de cluster hierárquico para separar grupos homogêneos internamente e heterogêneos entre si, seguindo a metodologia descrita a seguir. Espera-se desse modo avaliar-se a capacidade de previsão de fenótipo com base em informações genéticas e não com base em efeitos de parentescos.

2.2.3.1 Análise de Agrupamento Hierárquico

A análise de cluster é uma técnica multivariada que consiste em agregar objetos de acordo com suas relações, de modo que o agrupamento resultante seja homogêneo internamente e heterogêneo entre clusters. Tal agregação é realizada com base em medidas de similaridade ou de distância entre indivíduos (HAIR et al., 2009). Neste trabalho utilizou-se uma matriz de distâncias obtida a partir da matriz de relacionamento realizada G (Equação 2.10). Originalmente a matriz G representa a similaridade genética entre os indivíduos, porém para utilizá-la como uma medida de distância foi feita a transformação $D = 1 - G$.

O Agrupamento hierárquico organiza os dados em uma estrutura a partir das medidas de proximidade e seus resultados são apresentados em forma de dendrogramas. Esta metodologia consiste na realização de uma série de divisões (métodos divisivo) ou aglomerações (método aglomerativo) dos dados.

Nos métodos aglomerativos, utilizados neste trabalho, cada observação pertence a um agrupamento unitário. O algoritmo desses métodos consiste em combinar observações mais parecidas, com base em medidas de similaridade, ou distância, repetidamente até que se tenha apenas um cluster. Para se medir a similaridade entre os agrupamentos são utilizadas diversas metodologias de ligação (RICHARD A. JOHNSON, 2007) descritas a seguir e ilustradas na Figura 5.

- **Ligação Simples:** Realiza agrupamentos a partir da menor distância entre as observações dos grupos. Este método permite a criação de diversos padrões de aglomeração. Porém, isto pode causar problemas em clusters mal delineados, por exemplo com a ligação de dois grupos muito distintos, com duas medidas muito próximas.

  Seu procedimento consiste em encontrar a menor distância na matriz $D = d_{ik}$ e aglomerar os objetos correspondentes. Por exemplo, junta-se os objetos U e V no cluster (UV). Em seguida, a distância entre este cluster e os restantes será a menor entre os dois objetos aglomerados e os demais. Ou seja ,

  $$d_{(UV)W} = \min\{d_{UW}, d_{VW}\},$$

  em que d_{UW} e d_{VW} são as distâncias entre os vizinhos mais próximos dos clusters U e W, e V e W, respectivamente. Repete-se então esse processo até que haja apenas um cluster.

- **Ligação Completa:** É um método similar ao de ligação simples, porém define os clusters a partir da maior distância entre as observações dos dois grupos. Com isso gera-se agrupamentos mais compactos e que possuem distância máxima entre si.

Assim como no caso anterior, encontra-se a menor distância na matriz D e aglomera-se suas observações correspondentes. Porém a distância entre este cluster e os demais será a maior entre os objetos de cada um deles como descrito na equação

$$d_{(UV)W} = \max\{d_{UW}, d_{VW}\},$$

em que d_{UW} e d_{VW} são as distâncias entre os vizinhos mais afastados dos clusters U e W, e V e W, respectivamente.

- **Ligação Média:** Considera-se como medida de dissimilaridade a média das distâncias entre as observações de cada cluster. Desta forma é um método mais robusto em relação à observações extremas, gerando aglomerados com pequena variação interna.

  Novamente busca-se a menor distância na matriz $D = d_{ik}$ para aglomerar os vizinhos mais próximos. Por exemplo, junta-se os objetos U e V no cluster (UV) novamente. A distância entre (UV) e outro cluster W é dada por

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{n_{(UV)} n_W},$$

  em que d_{ik} é a distância entre a observação i do cluster (UV) e a observação k do cluster W e $n_{(UV)}$ e n_W são o número de objetos nos clusters (UV) e W respectivamente.

- **Ligação de Ward:** Define-se como a distância entre clusters, o quanto a soma total de quadrados irá aumentar ao aglomerá-los. Busca-se então grupos que minimizem essa perda de informação. Este método é afetado por valores extremos e tende a produzir aglomerados com aproximadamente o mesmo número de indivíduos.

  Inicialmente, calcula-se a Soma dos Erros Quadráticos de cada aglomerado por

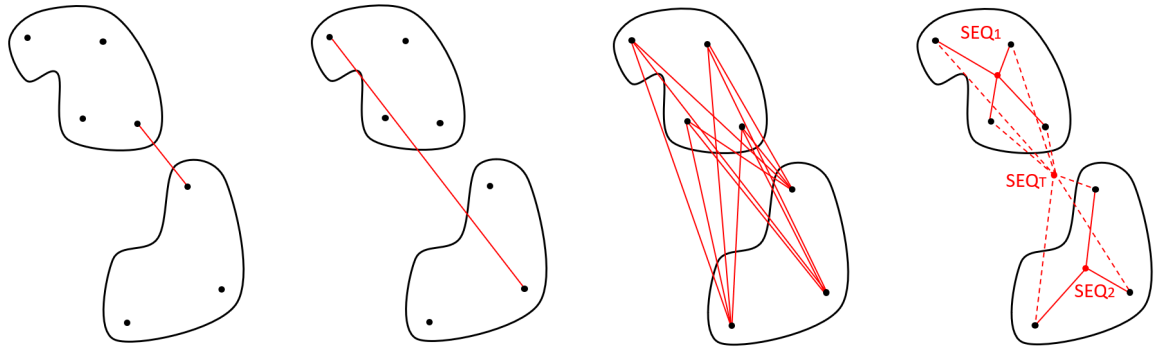
$$SEQ_C = \sum_{i=1}^{n_C} x_i^2 - \frac{1}{n_C}\left(\sum_{i=1}^{n_C} x_i\right)^2,$$

  em que x_i é o valor associado ao i-ésimo objeto e n_C é o número de objetos do cluster C. Em seguida obtém-se a Soma dos Erros Quadráticos Total dada por

$$SEQ_T = SEQ_1 + SEQ_2 + ... + SEQ_K.$$

  Em cada etapa, todas as junções são consideradas e aglomera-se aqueles objetos que apresentarem o menor aumento na SEQ_T.

Figura 5 – Tipos de Ligação



(a) Ligação Simples (b) Ligação Completa (c) Ligação Média (d) Ligação de Ward

2.3 Implementação

A seguir será descrito o processo de implementação das técnicas utilizando o software estatístico Rx64bit - versão 3.4.1 (R CORE TEAM, 2017).

Inicialmente foi REALIZADO o controle de qualidade nos dados genotípicos e os dados fenotípicos foram padronizados. Em seguida, obteve-se a matriz de relacionamento realizada com o comando *A.mat* do pacote *rrBLUP* (ENDELMAN, 2011) e a matriz de relacionamento estimada com auxílio dos pacotes *synbreed* (WIMMER et al., 2012) e *pedigreemm* (BATES; VAZQUEZ, 2014).

Utilizando-se a matriz de relacionamento realizada, implementou-se o algoritmo de clusterização hierárquica para obter os grupos do 5-fold com o comando *hclust*, que necessitou da transformação da matriz de similaridade em matriz de distância devido ao seu padrão.

Os modelos de Regressão Ridge BLUP e BLUP fenotípico com a matriz A estimada e realizada foram ajustados com o comando *mixed.solve* também do pacote *rrBLUP*. Cada modelo foi utilizado em um contexto diferente. O RRBLUP foi usado para prever os GEBVs e para definir grupos de árvores com maiores e menores valores genéticos estimados. Já os modelos de BLUP fenotípico com matriz A estimada representa o melhoramento tradicional, por isso serve para verificar a eficácia relativa dos modelos de seleção genômica. Por fim, os valores preditos pelo BLUP fenotípico com a matriz de relacionamento realizada serão também utilizados como respostas do SVM .

Além dos EBVs do BLUP fenotípico, outra abordagem foi utilizar os fenótipos ajustados pelos efeitos de delineamento, por um modelo de regressão linear, como resposta do SVM. Como em ambos os casos trata-se de variáveis contínuas, foi necessário dicotomizá-las em grupos de plantas com maiores e menores fenótipos, ou GEBVs, a partir de um

ponto de corte. Sabe-se que grupos desbalanceados podem prejudicar o desempenho do algoritmo, por isso testou-se as proporções de elite-não elite de 50-50, 40-60, 30-70, 20-80 e 15-85.

Com o pacote *caret* (JED WING et al., s.d.), modelou-se o SVM considerando-se os Kernels Linear e Radial, com parâmetros de ajuste $C = (2^{-15}, ..., 2^{6})$ e $\gamma = 1.9210^{-5}$ especificado pelo pacote.

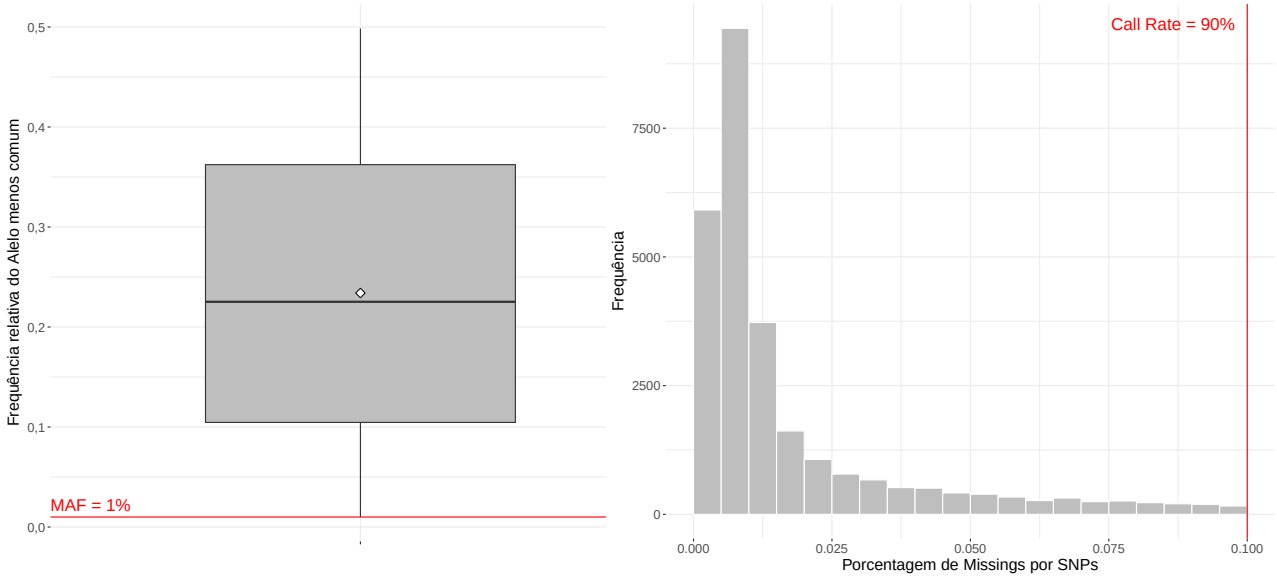
A avaliação do modelo, tanto da Regressão Ridge BLUP quanto do SVM, foi obtida a partir da validação cruzada 5-fold. As medidas de avaliação kappa e eficiência relativa consideram uma porcentagem (5, 10, 20 e 30%) de indivíduos com maiores GEBVs, ou probabilidade de pertencer à classe de elite no caso do SVM, como os melhores indivíduos.

3 Resultados

3.1 Análise Descritiva

Foram utilizados dados da EMBRAPA (GRATTAPAGLIA, 2014), que após passarem por um filtro de dados faltantes menores que 10% e frequência do alelo menos comum (MAF) maior que 1%, resultaram em 27.573 SNPs. A Figura 6 apresenta o boxplot da frequência do alelo menos comum para cada marcador e a porcentagem de dados faltantes por marcador. Nota-se que após a exclusão dos marcadores com *call rate* inferior a 90% grande parte deles possuem menos de 2,5% de dados faltantes.

Figura 6 – Controle de Qualidade dos Genótipos



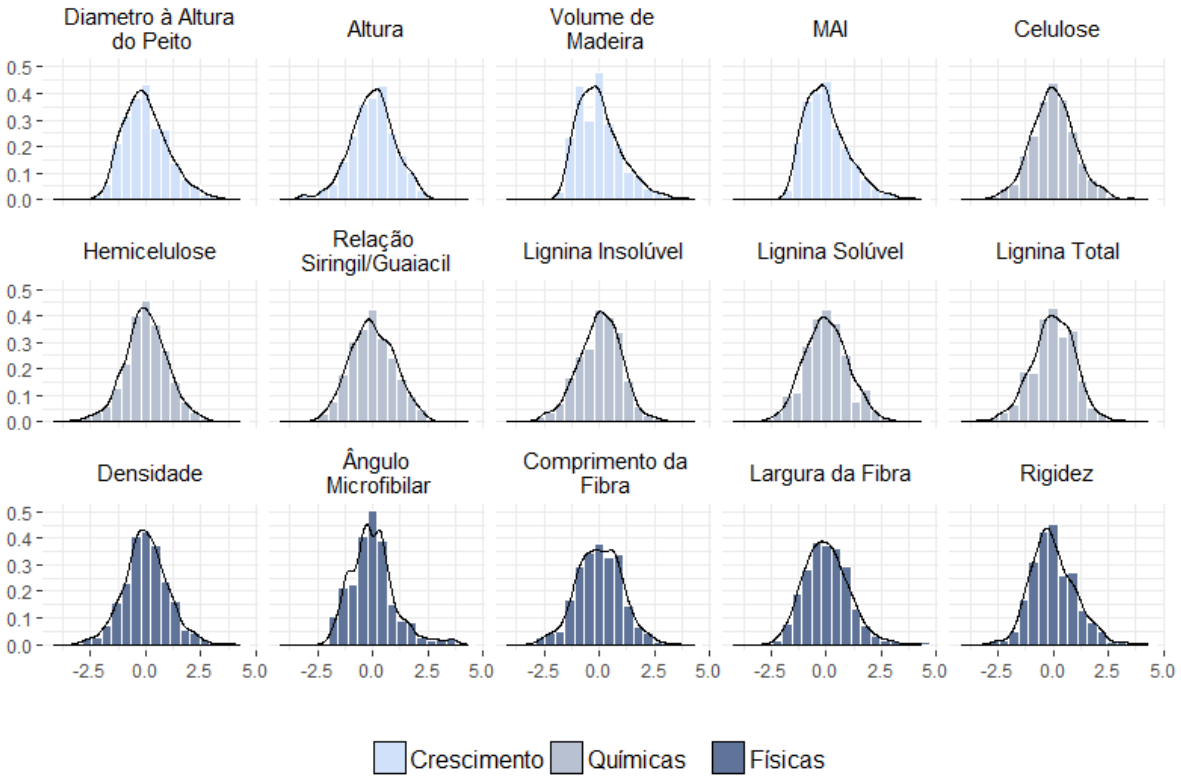
(a) Frequência do alelo mínimo (MAF) (b) Porcentagem de dados faltantes por marcador

Já os fenótipos, todos variáveis contínuas, foram padronizados e centralizados, isto é seus valores foram subtraídos pelas respectivas médias e divididos pelos respectivos desvios-padrão. A Figura 7 apresenta histogramas das 15 características divididas nos 3 grupos. Originalmente, os fenótipos apresentavam coeficientes de variação entre 3,62% (da celulose) até aproximadamente 25% (do volume e IMA) (Tabela 2), ou seja todos apresentam, mesmo padronizados, uma variação desejada. Com isso, tenta-se entender o quanto dessa variação dos fenótipos é explicada pelos marcadores genéticos.

Tabela 2 – Tabela de medidas de fenótipos

Fenótipo	Média	Desvio Padrão	Coeficiente de Variação	Coeficiente de Curtose de Pearson	Coeficiente de Assimetria
DAP	16,60	1,83	11,03	3,22	0,57
Altura	24,19	1,36	5,61	3,28	-0,24
Volume	0,24	0,06	25,49	3,86	0,88
IMA	58,63	14,92	25,45	3,87	0,88
Celulose	48,88	1,77	3,62	3,63	0,06
Hemicelulose	17,27	0,88	5,09	3,59	-0,12
Relação S:G	2,93	0,41	13,94	2,64	0,08
Lignina Insolúvel	25,24	1,11	4,38	3,63	-0,29
Lignina Solúvel	3,55	0,41	11,69	2,97	0,10
Lignina Total	28,78	1,05	3,65	3,51	-0,22
Densidade	512,51	35,90	7,00	3,48	0,13
Ângulo Microfibrilar	12,94	1,22	9,41	4,52	0,88
Comprimento de fibra	0,75	0,06	7,72	3,00	0,01
Largura de fibra	19,84	1,14	5,73	3,68	0,52
Rigidez	7,11	1,02	14,28	3,50	0,49

Figura 7 – Fenótipos Padronizados



Além disso, tem-se que, o comprimento de fibra possui curtose igual à distribuição Normal. A relação S:G e a lignina solúvel apresentam tal medida menor que a Normal, e todos os demais fenótipos apresentaram curtose maior que a distribuição Normal, com o valor máximo de 4,51 do ângulo microfibrilar. Quanto à assimetria, os fenótipos altura, hemicelulose, lignina insolúvel e lignina total se mostraram ligeiramente assimétricos negativamente, chegando até -0,29 para a lignina insolúvel. Já os demais fenótipos apresentam

leve assimetria positiva sendo o máximo de 0,88 do Volume. Em geral todas as variáveis apresentam certa semelhança com a distribuição Normal.

3.2 Validação Cruzada

Como já visto, a avaliação do poder preditivo dos modelos foi realizada através de uma validação cruzada 5-fold. Foram utilizadas duas abordagens diferentes para determinar quais indivíduos pertenceram a cada grupo. Na abordagem aleatória, os indivíduos foram selecionados aleatoriamente para cada grupo, o que resultou, para os fenótipos com 999 plantas, em grupos com 200 plantas e o grupo 2 com 199 (Tabela 3). Já na abordagem baseada em relalacionamento o algoritmo de cluster agrupou as plantas de acordo com a proximidade de parentesco entre elas, utilizando a matriz de relacionamento realizada. Com isso os grupos possuem tamanhos diferentes, variando entre 105 do grupo 4 até 374 indivíduos do grupo 2.

Tabela 3 – Quantidade de indivíduos por grupo de validação cruzada

Grupo	Validação cruzada aleatória	Validação cruzada baseada em relacionamento
1	200	141
2	199	374
3	200	221
4	200	105
5	200	158

Os grupos da validação baseada em relacionamento foram obtidos utilizando a ligação de Ward, pois foi a que forneceu grupos com tamanhos mais próximos e mais balanceados internamente em relação à porcentagem de indivíduos elite e não-elite (Tabelas 4 a 6) para grande parte dos fenótipos. Notou-se um comportamento inverso nos grupos 1 e 5, em que se um tem uma porcentagem maior de árvores com maiores medidas, no outro a porcentagem maior é de árvores com menores medidas. Porém, em geral os grupos se mostraram bem balanceados, com exceção do fenótipo ângulo microfibrilar (Tabela 7), que apresentou grupos desbalanceados, alguns até sem a presença de indivíduos com maiores fenótipos oou GEBVs. Além disso, este tipo de agrupamento só gerou 5 clusters com tamanhos próximos. Ao escolher mais grupos, um deles sempre apresentava uma quantidade bem menor, por isso optou-se por utilizar o 5-fold e não um 10-fold, por exemplo.

Tabela 4 – Distribuição de árvores com menores e maiores medidas, por grupo de validação cruzada baseada em relacionamento, para diâmetro à altura do peito

	50%				30%				15%		
Grupo	n	Menores	Maiores	Grupo	n	Menores	Maiores	Grupo	n	Menores	Maiores
1	141	61,7%	38,3%	1	141	77,3%	22,7%	1	141	83,7%	16,3%
2	374	48,9%	51,1%	2	374	67,6%	32,4%	2	374	79,1%	20,9%
3	221	54,3%	45,7%	3	221	73,8%	26,2%	3	221	89,1%	10,9%
4	105	50,5%	49,5%	4	105	71,4%	28,6%	4	105	92,4%	7,6%
5	158	36,1%	63,9%	5	158	62,7%	37,3%	5	158	89,2%	10,8%

Tabela 5 – Distribuição de árvores com menores e maiores medidas, por grupo de validação cruzada baseada em relacionamento, para relação S:G

	50%				30%				15%		
Grupo	n	Menores	Maiores	Grupo	n	Menores	Maiores	Grupo	n	Menores	Maiores
1	141	42,6%	57,4%	1	141	63,8%	36,2%	1	141	81,6%	18,4%
2	374	51,1%	48,9%	2	374	66,8%	33,2%	2	374	81,8%	18,2%
3	221	56,6%	43,4%	3	221	78,3%	21,7%	3	221	91,0%	9,0%
4	105	63,8%	36,2%	4	105	87,6%	12,4%	4	105	98,1%	1,9%
5	158	36,1%	63,9%	5	158	59,5%	40,5%	5	158	78,5%	21,5%

Tabela 6 – Distribuição de árvores com menores e maiores medidas, por grupo de validação cruzada baseada em relacionamento, para comprimento de fibra

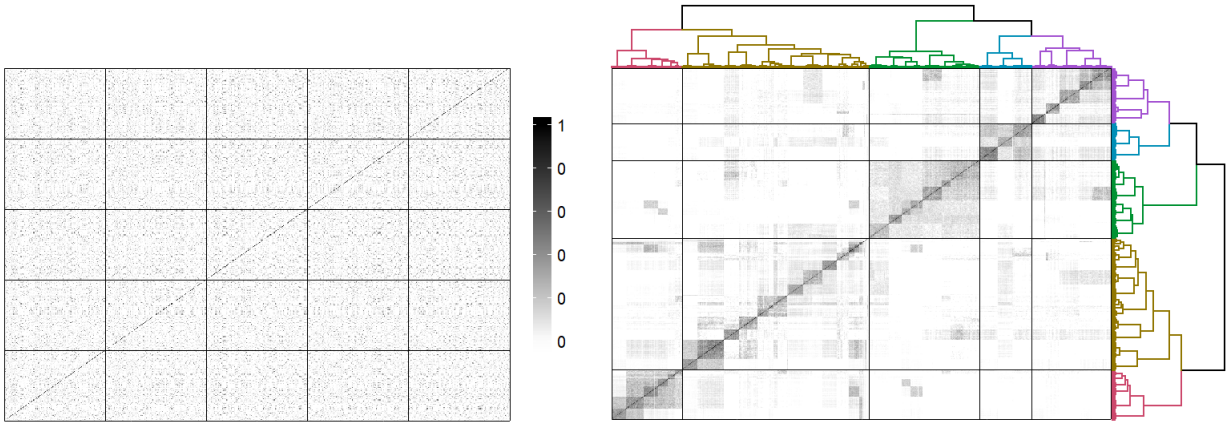
	50%				30%				15%		
Grupo	n	Menores	Maiores	Grupo	n	Menores	Maiores	Grupo	n	Menores	Maiores
1	57	45,6%	54,5%	1	57	70,2%	29,8%	1	57	78,9%	21,1%
2	137	51,1%	48,9%	2	137	69,3%	30,7%	2	137	86,1%	13,9%
3	74	59,5%	40,5%	3	74	77,0%	23,0%	3	74	89,2%	10,8%
4	39	30,8%	69,2%	4	39	64,1%	35,9%	4	39	87,2%	12,8%
5	43	53,5%	46,5%	5	43	65,1%	34,9%	5	43	79,1%	20,9%

Tabela 7 – Distribuição de árvores com menores e maiores medidas, por grupo de validação cruzada baseada em relacionamento, para ângulo microfibrilar

	50%				30%				15%		
Grupo	n	Menores	Maiores	Grupo	n	Menores	Maiores	Grupo	n	Menores	Maiores
1	57	68,4%	31,6%	1	57	96,5%	3,5%	1	57	100,0%	0,0%
2	136	44,1%	55,9%	2	136	61,0%	39,0%	2	136	75,0%	25,0%
3	73	4,1%	95,9%	3	73	32,9%	67,1%	3	73	74,0%	26,0%
4	39	87,2%	12,8%	4	39	97,4%	2,6%	4	39	100,0%	0,0%
5	43	88,4%	11,6%	5	43	100,0%	0,0%	5	43	100,0%	0,0%

Com o agrupamento hierárquico tem-se grupos compostos por indivíduos de parentesco próximo e semelhantes entre si, consequentemente os grupos são pouco relacionados uns com os outros. Isto pode ser observado na figura 8b, em que os locais mais escuros representam proximidade genética maior. Já a escolha aleatória resulta em grupos com indivíduos bastante relacionados a outros dos demais grupos, ou seja, os grupos possuem uma correlação de parentesco entre si maior que na abordagem de grupos formados aleatoriamente.

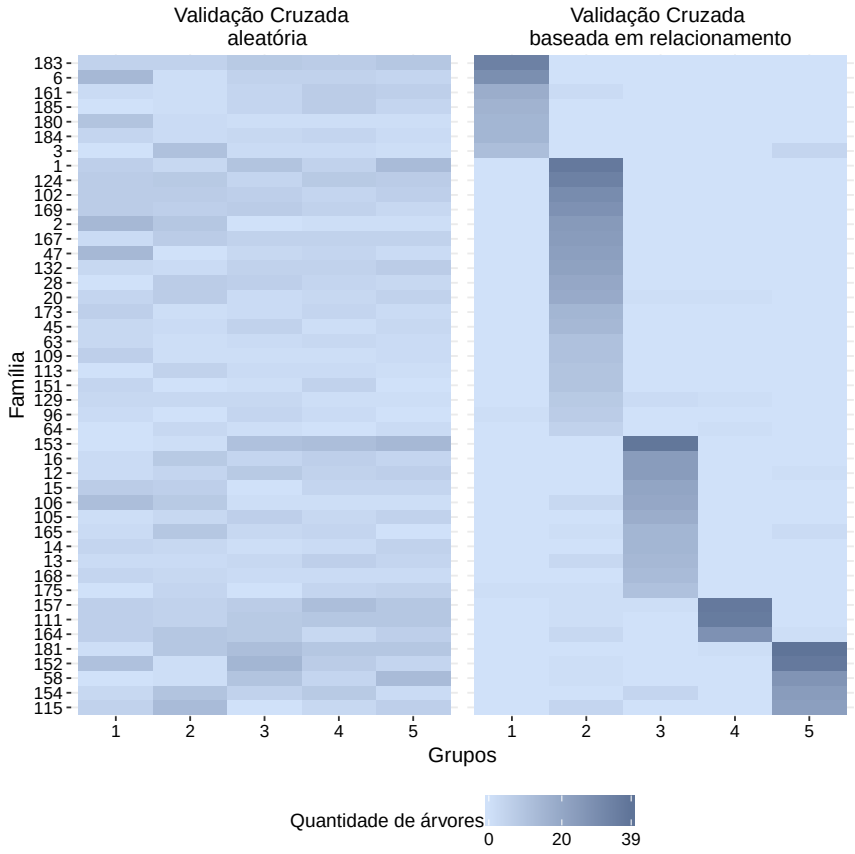
Figura 8 – Matriz de Relacionamento



(a) Grupos obtidos aleatoriamente (b) Grupos obtidos por agrupamento hierárquico

Após a separação dos grupos, comparou-se a frequência de indivíduos ditos de cada família em cada um deles. Nota-se que há uma separação clara entre as famílias nos grupos obtidos pela clusterização, havendo poucos indivíduos classificados diferentemente do resto da família. Já nos grupos aleatórios, tem-se indivíduos de todas as famílias em mais de um grupo.

Figura 9 – Frequência de famílias por grupos de validação cruzada

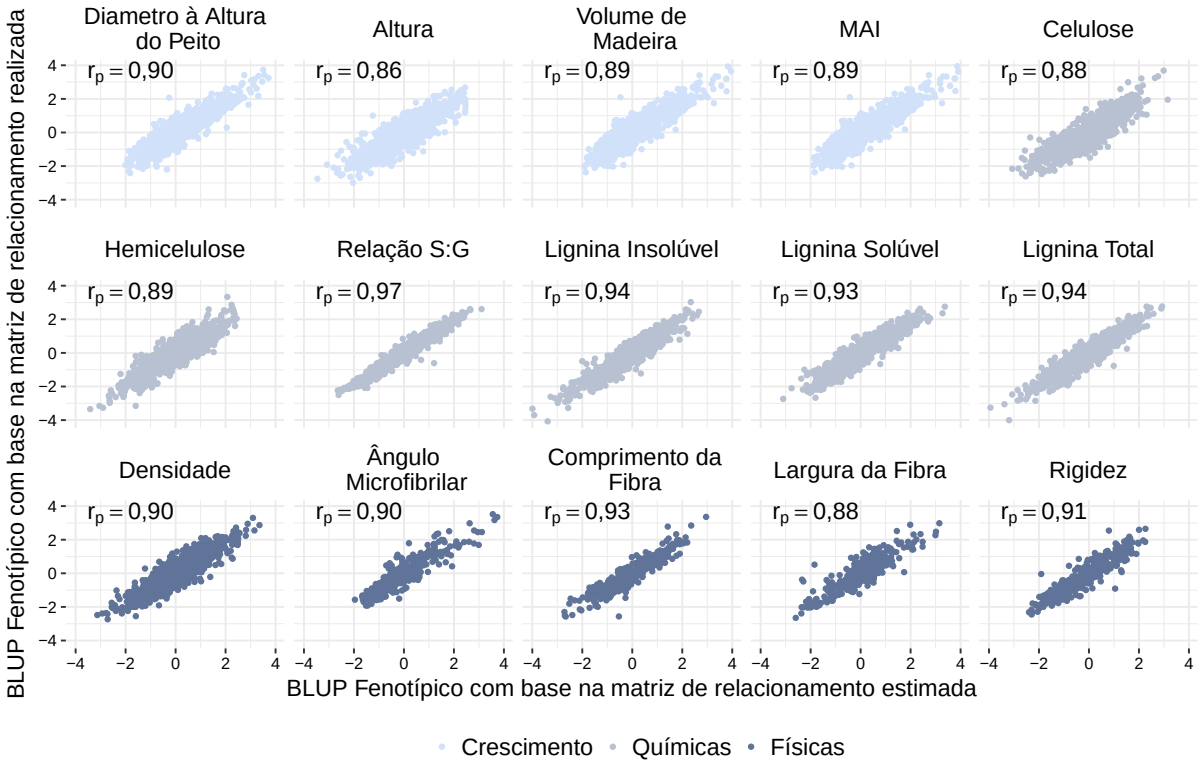


Isto demonstra que grupos obtidos pela clusterização com base em relacionamento possuem pouca relação entre si, portanto durante a validação cruzada o modelo treinado será testado em um conjunto de dados pouco correlacionado com os de treinamento. Já na abordagem aleatória, os agrupamentos de teste e treinamento terão uma correlação de parentesco maior, por se ter indivíduos de diversas famílias em ambos.

3.3 Modelos Mistos

Inicialmente foram ajustados os modelos BLUP fenotípicos com os dois tipos de matriz de relacionamento (estimada e realizada) e obtidos os *Estimated Breeding Values* (EBVs) que serão utilizados posteriormente. A Figura 10 apresenta diagramas de dispersão e coeficientes de correlação de Pearson para os EBVs obtidos com matrizes de relacionamento estimadas (eixo x de cada gráfico) e realizadas (eixo y de cada gráfico) para cada fenótipo. Nota-se que os coeficientes de correlação foram altos (todos acima de 0,8), evidenciando que os valores preditos por ambos os modelos estão relacionados.

Figura 10 – Comparação entre EBVs obtidos por BLUP fenotípico com matriz de relacionamento estimada e realizada



Considerando-se os grupos de validação cruzada, implementou-se o modelo de Regressão Ridge BLUP (RRBLUP). A Tabela 8 apresenta as herdabilidades e capacidades preditivas obtidas para cada fenótipo separadamente. Nota-se que em geral os fenótipos químicos apresentam herdabilidades superiores aos demais, enquanto as variáveis de crescimento apresentam herdabilidades em torno de 0,5. Já para os fenótipos físicos, o

comprimento da fibra e a densidade possuem as maiores herdabilidades, mesmo a primeira tendo medição em apenas 350 plantas. A variável ângulo microfibrilar possui a menor herdabilidade.

A Capacidade preditiva foi obtida por validação cruzada 5-fold, e portanto seus valores foram apresentados para as duas formas de separação dos grupos. Nota-se que para tal medida, os valores para todos os fenótipos foram menores quando os grupos foram divididos utilizando a abordagem baseada em matriz de parentesco.

Em ambos os casos, as variáveis químicas em geral apresentam capacidades preditivas superiores às demais, em que a celulose e a hemicelulose possuem os menores valores. Os fenótipos físicos apresentam capacidades preditivas próximas de 0,4, no caso aleatório e 0,20 no caso baseado em relacionamento, mas a altura apresenta valores menores em relação aos outros fenótipos do mesmo grupo.

Por fim, dentre as variáveis físicas, a densidade apresenta a maior medida, porém é a única do grupo que possui dados para todas as 999 plantas. Contudo, nota-se que mesmo sem informação completa para as árvores, o comprimento de fibra ainda apresenta uma alta herdabilidade e capacidade preditiva. As principais diferenças entre os valores obtidos com as duas abordagens de validação cruzada se dão em fenótipos com cerca de 350 árvores apenas. O ângulo microfibrilar e a largura de fibra passam a ter capacidade preditiva próxima de zero, enquanto a Rigidez apresenta valores superiores aos fenótipos de crescimento, o que não ocorreu na abordagem aleatória.

Tabela 8 – Medidas obtidas para o RRBLUP

Fenótipo	n	Herdabilidade	Capacidade Preditiva	
			Validação cruzada aleatória	Validação Cruzada baseada em relacionamento
DAP	999	0,52	0,41	0,24
Altura	999	0,43	0,33	0,13
Volume	999	0,51	0,40	0,22
IMA	999	0,51	0,40	0,22
Celulose*	999	0,63	0,49	0,32
Hemicelulose*	999	0,72	0,54	0,30
Relação S:G*	999	0,89	0,82	0,68
Lignina Insolúvel*	999	0,74	0,62	0,37
Lignina Solúvel*	999	0,77	0,72	0,58
Lignina Total*	999	0,73	0,60	0,36
Densidade*	999	0,66	0,60	0,49
Ângulo Microfribilar	348	0,14	0,17	-0,02
Comprimento de fibra	350	0,70	0,52	0,36
Largura de fibra	350	0,25	0,19	-0,02
Rigidez	349	0,36	0,35	0,29

*: fenótipos obtidos com auxílio da espectometria NIRS.

A Figura 11 apresenta correlações de Pearson entre valores genéticos estimados pelo BLUP Fenotípico com matriz de relacionamento estimada, e pelo RRBLUP para as duas abordagens de validação cruzada. Novamente as correlações para os grupos obtidos por clusterização hierárquica baseada em relacionamento foram inferiores àquelas dos grupos obtidos aleatoriamente.

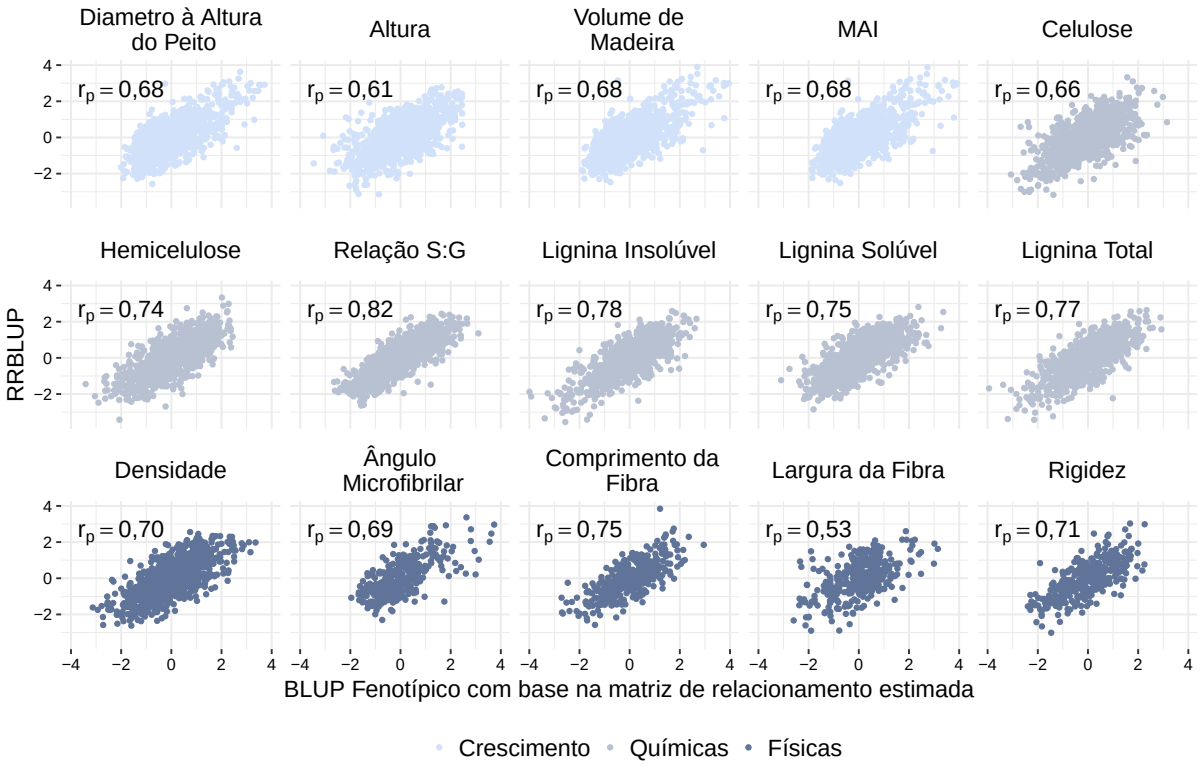
Para os grupos de validação cruzada separados aleatoriamente, os fenótipos químicos apresentam em geral correlações mais altas, ou seja, os valores obtidos pelo RRBLUP estão próximos dos obtidos pelo melhoramento tradicional. A correlação para as variáveis de crescimento foram todas superiores a 0,60 e são inferiores àquelas obtidas para variáveis físicas, com exceção da largura de fibra. Fenótipos que apresentaram herdabilidades baixas, como o ângulo microfibrilar e a rigidez, apresentaram correlações altas, de 0,69 e 0,71 respectivamente.

Já para os grupos obtidos por agrupamento hierárquico, alguns fenótipos químicos que antes apresentaram valores altamente correlacionados, neste caso apresentaram coeficiente próximo a 0,3. Outros, como a relação S:G e a lignina solúvel permaneceram apresentando correlação de 0,6. As variáveis de crescimento permanecem com correlações parecidas, porém agora em torno de 0,3. Por fim, os fenótipos físicos, que possuem apenas dados para 350 árvores, apresentaram correlações dentre as mais altas, como densidade, comprimento da fibra e rigidez; ou apresentam correlações próximas de zero como o ângulo microfibrilar e a largura da fibra. O ângulo microfibrilar, que possui a menor herdabilidade, passou de uma correlação de Pearson de 0,69 nos grupos aleatórios para -0,11 nos grupos obtidos por algoritmos de cluster.

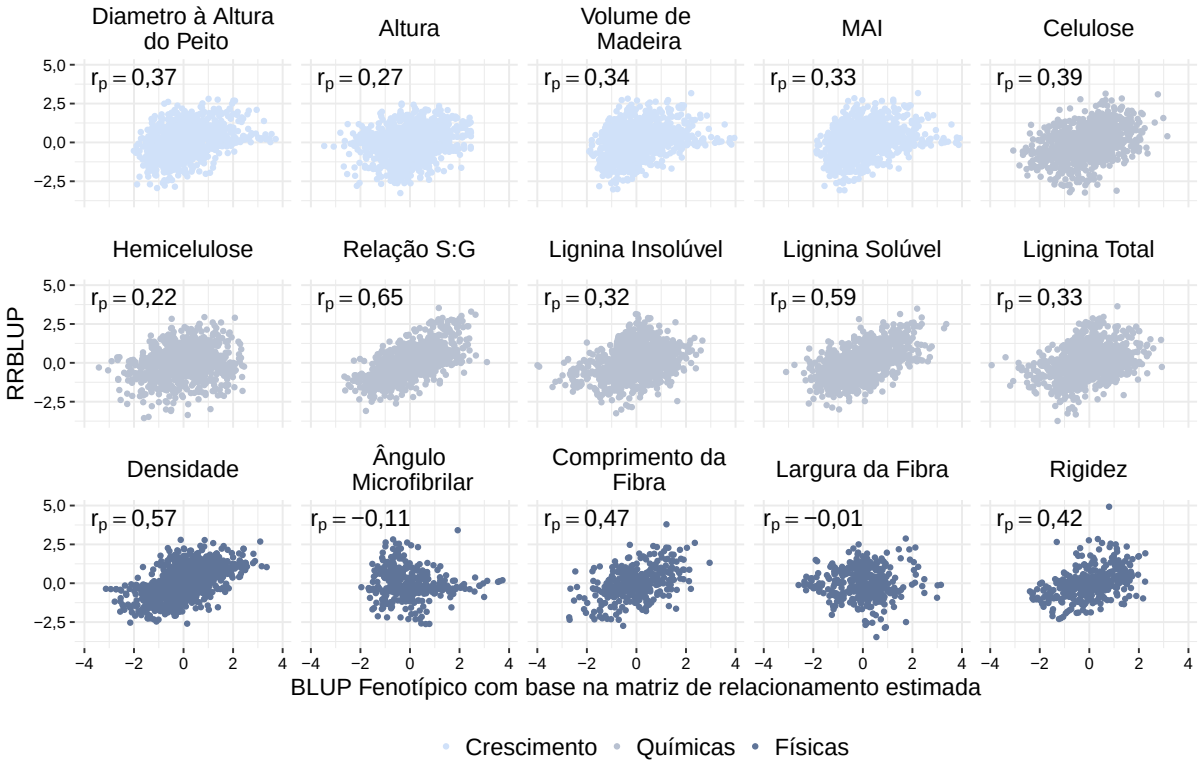
Ainda fazendo um paralelo com o melhoramento tradicional, ordenou-se os GEBVs obtidos no RRBLUP e os EBVs obtidos pelo BLUP fenotípico com matriz estimada e selecionou-se uma certa porcentagem (5, 10, 20 e 30%) dos melhores indivíduos. Em seguida, comparou-se a proporção de plantas que foram classificadas como melhores em ambos os casos para cada um dos 5 folds. Na Figura 12 encontram-se os valores médios dessa proporção bem como a amplitude dos valores, isto é, as faixas em torno das linhas médias representam o intervalo entre o menor e o maior valor obtidos nos folds.

Foram feitas comparações para os fenótipos com as maiores herdabilidade de cada grupo (crescimento, físicas e químicas), que são o diâmetro a altura do peito, a relação S:G e o comprimento de fibra; e com as variáveis que possuem menores herdabilidades, como altura, celulose e ângulo microfibrilar.

Figura 11 – Comparação entre melhoramento tradicional e RRBLUP

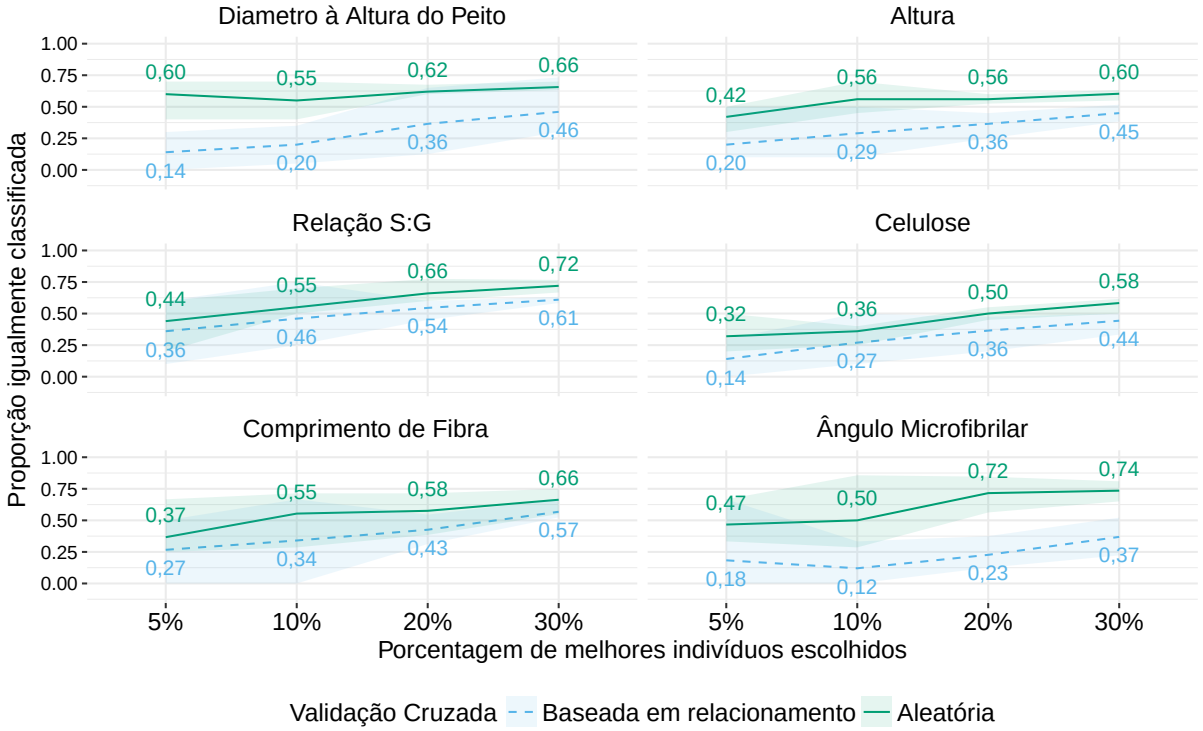


(a) Grupos obtidos aleatoriamente



(b) Grupos obtidos por agrupamento hierárquico

Figura 12 – Porcentagem de indivíduos classificados igualmente entre o BLUP Fenotípico e o RRBLUP



Observa-se que para as variáveis de crescimento, diâmetro à altura do peito e altura, utilizando a validação cruzada aleatória seleciona-se em média em torno de 60% dos indivíduos igualmente com o RRBLUP e o melhoramento tradicional. À medida que mais indivíduos são selecionados há um aumento nas proporções médias e uma diminuição na variação.Já para a validação cruzada baseada em relacionamento há uma variação maior entre os valores obtidos em cada *fold* e, em média, seleciona-se menos indivíduos igualmente, variando entre em torno de 20% até 45%.

As variáveis químicas apresentam proporções igualmente classificadas mais próximas para ambas as validações os demais. A relação S:G, que possui a maior herdabilidade, apresenta uma igualdade média de 44% ao se selecionar 50 árvores com grupos aleatórios e esta porcentagem aumenta gradativamente a medida que seleciona-se mais indivíduos, chegando até 72%.

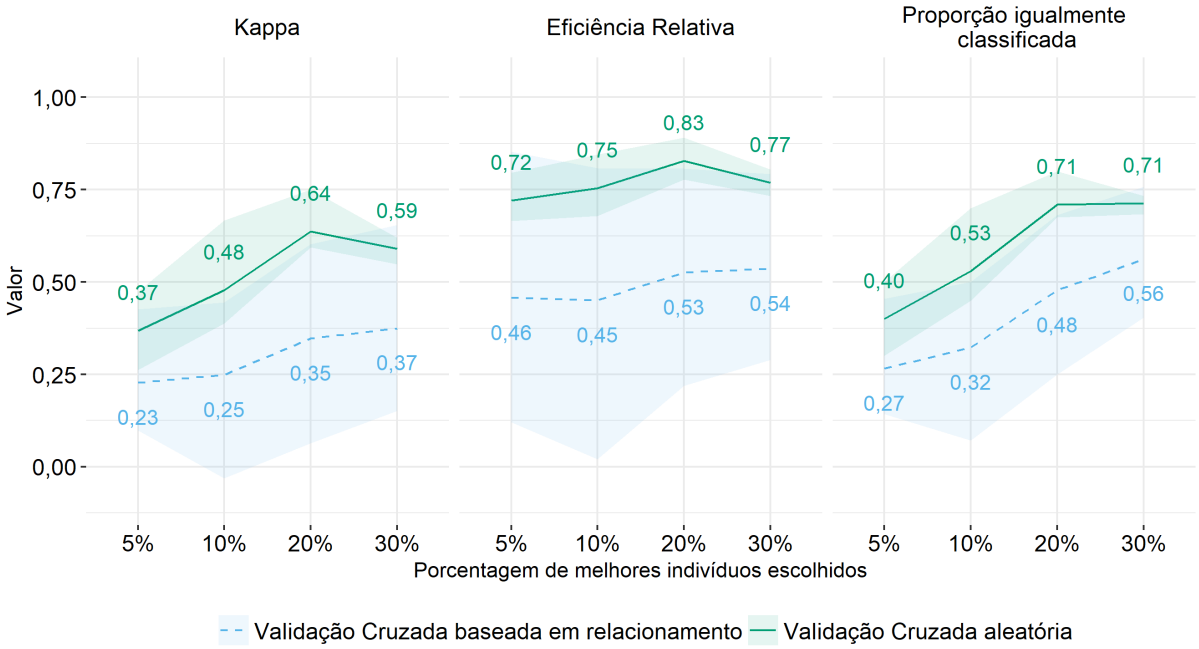
Por fim, para os fenótipos físicos, que possuem dados para apenas 350 plantas, apresentaram variações maiores que os demais fenótipos para ambas as abordagens de validação cruzada. Ainda, o ângulo microfibrilar, que possui a menor herdabilidade dentre todos os fenótipos, apresentou as maiores diferenças entre os valores obtidos por cada validação cruzada.

3.4 Máquinas de Suporte Vetorial

Outro algoritmo utilizado foi a Máquina de Suporte Vetorial. Como se trata de um algoritmo de classificação, a variável resposta deve ser qualitativa; para isso é necessário selecionar pontos de cortes nos fenótipos por serem variáveis contínuas.Nesse caso, selecionou-se cinco separações de grupo elite-não-elite a partir dos percentis, obtendo-se as proporções 50-50, 40-60, 30-70, 20-80 e 15-85. O modelo foi ajustado para todas as separações e para ambos os kernels, o linear e o radial. Optou-se, então, por utilizar o modelo com kernel radial e proporção de corte para classes de 30-70.

Pelo fato dos fenótipos poderem conter erros de coleta , utilizou-se os valores preditos pelo BLUP fenotípico, dicotomizados, como resposta. A matriz de relacionamento realizada foi utilizada na predição de tais valores, pois apresenta informações mais verossímeis e completas do que aquelas da matriz estimada pelo pedigree, por serem obtidas a partir de milhares de genótipos de marcadores. O algoritmo foi ajustado apenas para os fenótipos diâmetro à altura do peito, relação S:G e comprimento de fibra, por serem os que apresentem maior herdabilidade de cada tipo, e para o ângulo microfibrilar por possuir a menor herdabilidade e grupos desbalanceados.

Figura 13 – Médias e amplitudes de medidas de avaliação para diâmetro à altura do peito obtidas por SVM com EBVs como resposta, proporção 30-70 e kernel radial

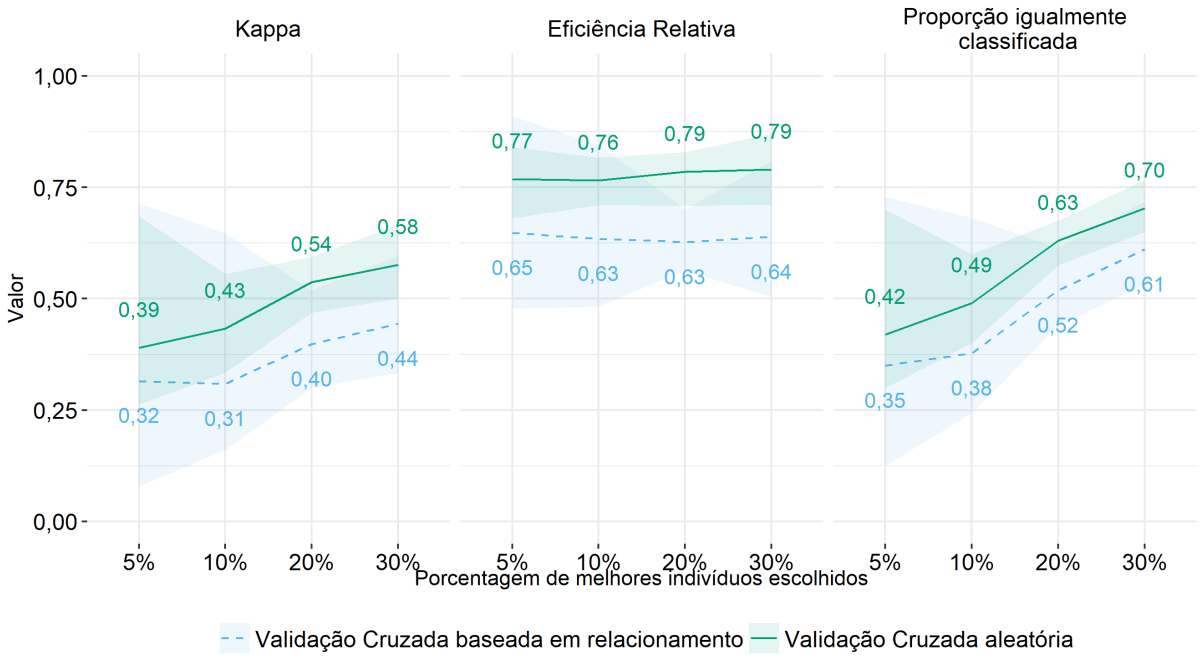


As figuras a seguir (Figuras 13 a 24) apresentam as médias das medidas de avaliação de modelos obtidas em cada *fold* da validação cruzada. Foram calculadas ao se ordenar as probabilidades de pertencer à classe dos melhores. Neste caso considerou-se as porcentagens

dos primeiros indivíduos como sendo os melhores. As faixas em torno dos valores médios representam a amplitude dos valores obtidos, isto é, o intervalo entre o menor e o maior valor obtido entre os *folds*. Esses indivíduos foram comparados com aqueles obtidos pelo ranqueamento dos valores preditos do BLUP fenotípico com matriz estimada, por representarem o melhoramento tradicional.

Para o fenótipo de crescimento de maior herdabilidade (Figura 13), com validação cruzada baseada em relacionamento, nota-se que a medida kappa indica uma concordância fraca ao selecionar 5% dos indivíduos; ainda assim, apresentou eficiência relativa acima de 0,7 e 40% indivíduos classificados igualmente. À medida que um maior número de indivíduos é selecionado, as três medidas crescem e a concordância passa a ser boa, mas ao selecionar cerca de 300 árvores elas se estabilizam. Nota-se que as medidas tem uma variação constante em cada fold obtido aleatoriamente, principalmente a eficiência relativa, que possui uma amplitude menor que as demais. Já para a validação cruzada, utilizando grupos obtidos por clusterização, ainda considerando a Figura 13, as medidas variam bastante em cada fold, e em geral são menos otimistas do que aquelas obtidas pela validação aleatória.

Figura 14 – Médias e amplitudes de medidas de avaliação para relação S:G obtidas por SVM com EBVs como resposta, proporção 30-70 e kernel radial



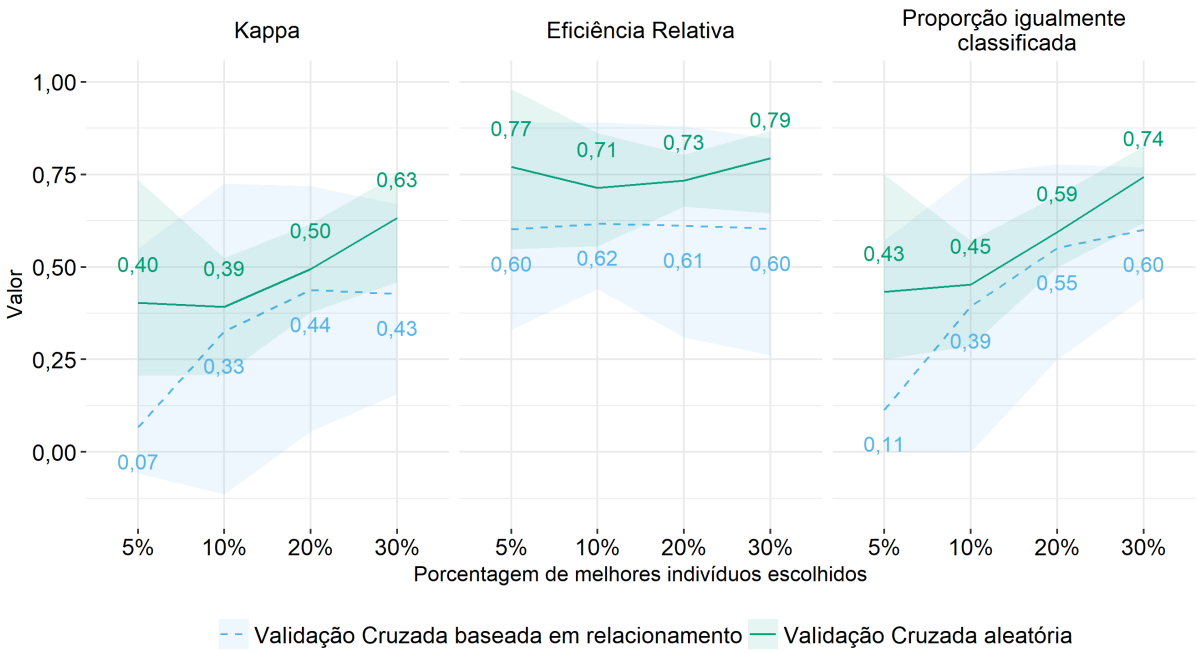
As medidas médias obtidas para a relação S:G (Figura 14) apresentam um comportamento semelhante para as duas formas de validação cruzada, porém nota-se que a amplitude dos valores é bem superior para os grupos obtidos baseados em relacionamento, principalmente ao se selecionar 5 ou 10% dos indivíduos como melhores. Nota-se que, para essas duas seleções, em alguns *folds* da validação cruzada baseada em relacionamento as

métricas foram até superiores à aquelas obtidas pela validação aleatória. Considerando-se os valores médios, o coeficiente kappa de Cohen representa uma concordância moderada para a validação aleatória.

Para a validação baseada em relacionamento, o coeficiente kappa é fraco para as duas primeiras seleções e passa a ser moderado ao selecionar 20% ou 30% dos indivíduos. Entretanto suas amplitudes se sobrepõem, principalmente na seleção de 5%. Mesmo com essa variação na medida kappa, a eficiência relativa se mantém praticamente constante, em torno de 0,8 e 0,65, para validação aleatória e baseada em relacionamento, respectivamente. Já a proporção igualmente classificada apresenta um aumento à medida em que mais plantas são selecionadas, podendo chegar a 7% no caso em que os grupos de validação cruzada são obtidos aleatoriamente.

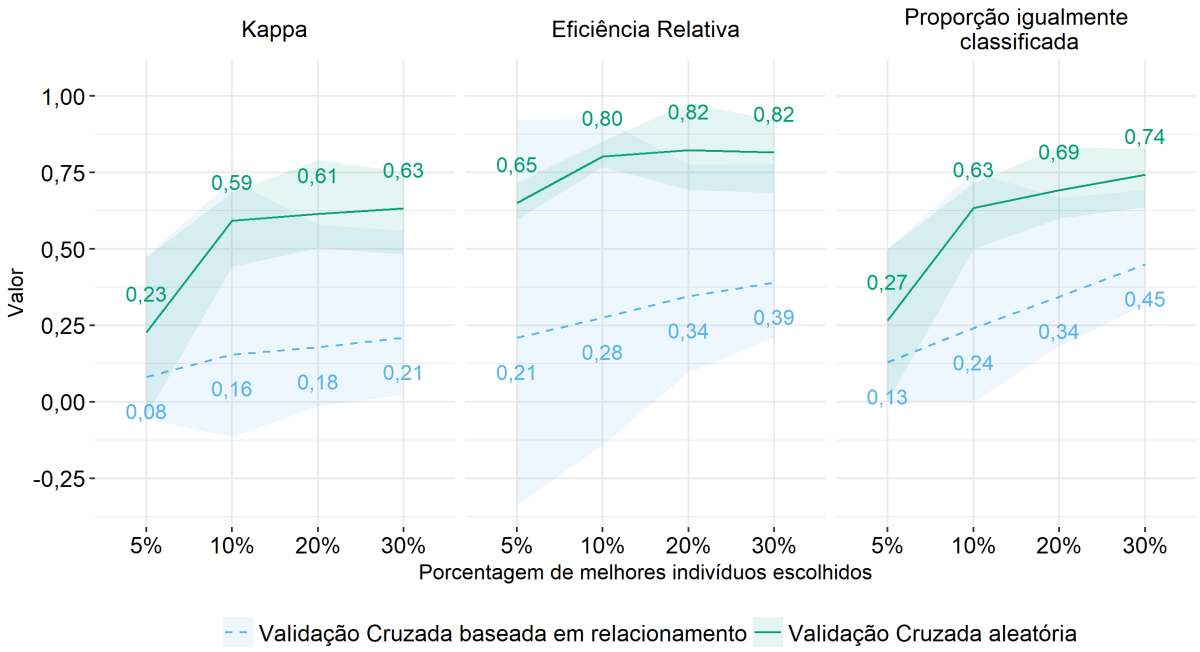
O comprimento de fibra (Figura 15), que contém dados para aproximadamente 350 plantas, apresenta amplitudes maiores para a validação não-estratificada em relação aos fenótipos anteriores. Nota-se que, em alguns casos, a validação estratificada resultou em métricas superiores à não-estratificada em alguns *folds* e a as amplitudes sempre se sobrepõem. Porém, ao se analisar apenas os valores médios, a validação aleatória apresenta medidas mais otimistas. Seus valores médios de eficiência relativa permanecem entre 0,7 e 0,8 e classifica-se igualmente acima de 0,43 dos indivíduos. Com a abordagem de separação dos grupos aleatoriamente, tem-se uma eficiência relativa constante em torno de 0,6, mas a proporção igualmente classificada aumenta de 0,11 até 0,60, ao selecionar se 50 e 300 árvores, respectivamente.

Figura 15 – Médias e amplitudes de medidas de avaliação para comprimento de fibra obtidas por SVM com EBVs como resposta, proporção 30-70 e kernel radial



Por fim, para o ângulo microfibrilar (Figura 16) percebe-se uma diferença grande entre as médias dos valores obtidos pelas duas abordagens de validação cruzada. No caso aleatório há um aumento nas três medidas ao se selecionar de 5% para 10% dos indivíduos. A medida kappa se mostra fraca ao se selecionar 50 árvores e passa a ser constante em torno de 0,5 ao se selecionar mais árvores. A eficiência relativa se mostrou acima de 0,8 para os casos em que mais de 10% das plantas são selecionadas, com proporção igualmente classificada chegando até 0,74. Já para o caso de validação cruzada baseada em relacionamento, o kappa apresenta uma concordância fraca, com valores mínimos negativos nas duas primeiras proporções de seleção, porém suas amplitudes se sobrepõem com as da outra abordagem. A eficiência relativa média obtida foi bem inferior, com seu máximo em 0,39, porém ao selecionar 5% e 10% em algum *fold* foram obtidas medidas superiores à da abordagem aleatória. A proporção igualmente classificada média acompanha os baixos valores, com um crescimento gradativo e novamente com amplitude se sobrepondo à outra.

Figura 16 – Médias e amplitudes de medidas de avaliação para ângulo microfibrilar obtidas por SVM com EBVs como resposta, proporção 30-70 e kernel radial



Nos casos anteriores a variável resposta utilizada foi o EBV obtido pelo BLUP fenotípico com matriz de parentesco realizada. Porém, para comparações, foram ajustados os modelos de SVM de classificação utilizando-se os resíduos de uma regressão em que os fenótipos são ajustados pelos efeitos de delineamento. As métricas obtidas em ambos os casos serão apresentadas a seguir.

Figura 17 – Comparação de médias e amplitudes de medidas obtidas por SVM com EBVs e fenótipos como resposta, com proporção 30-70 e kernel radial para diâmetro à altura do peito

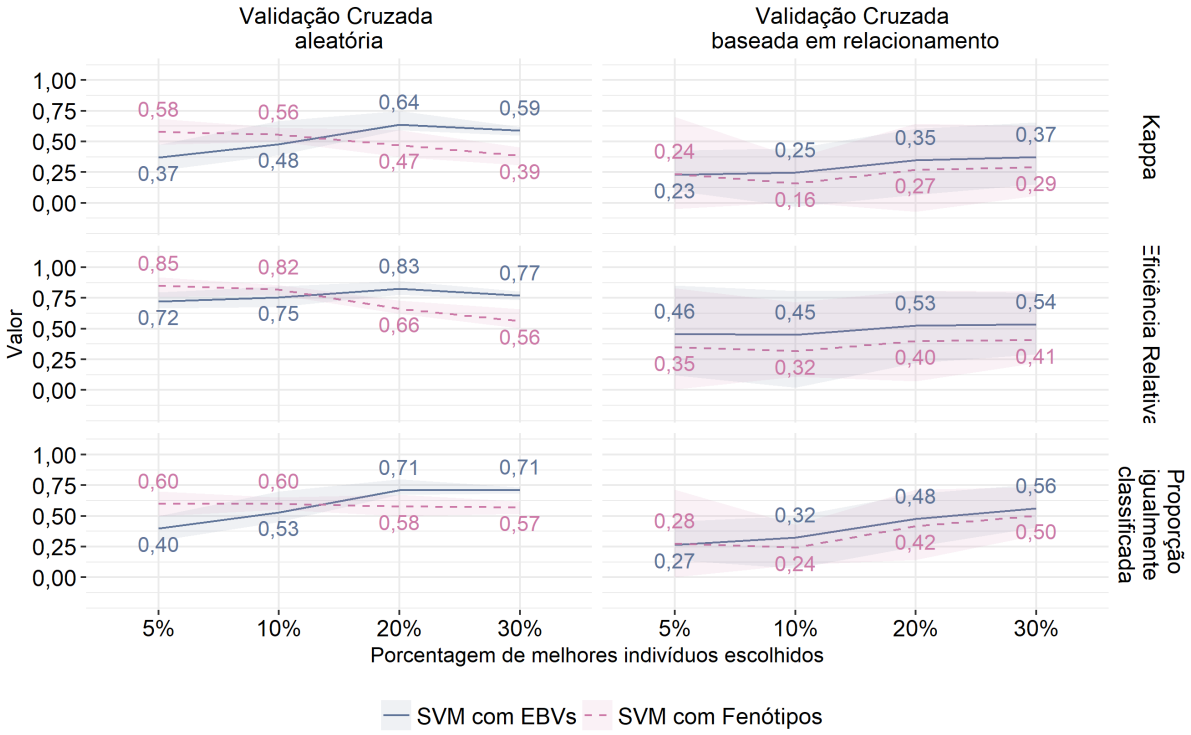
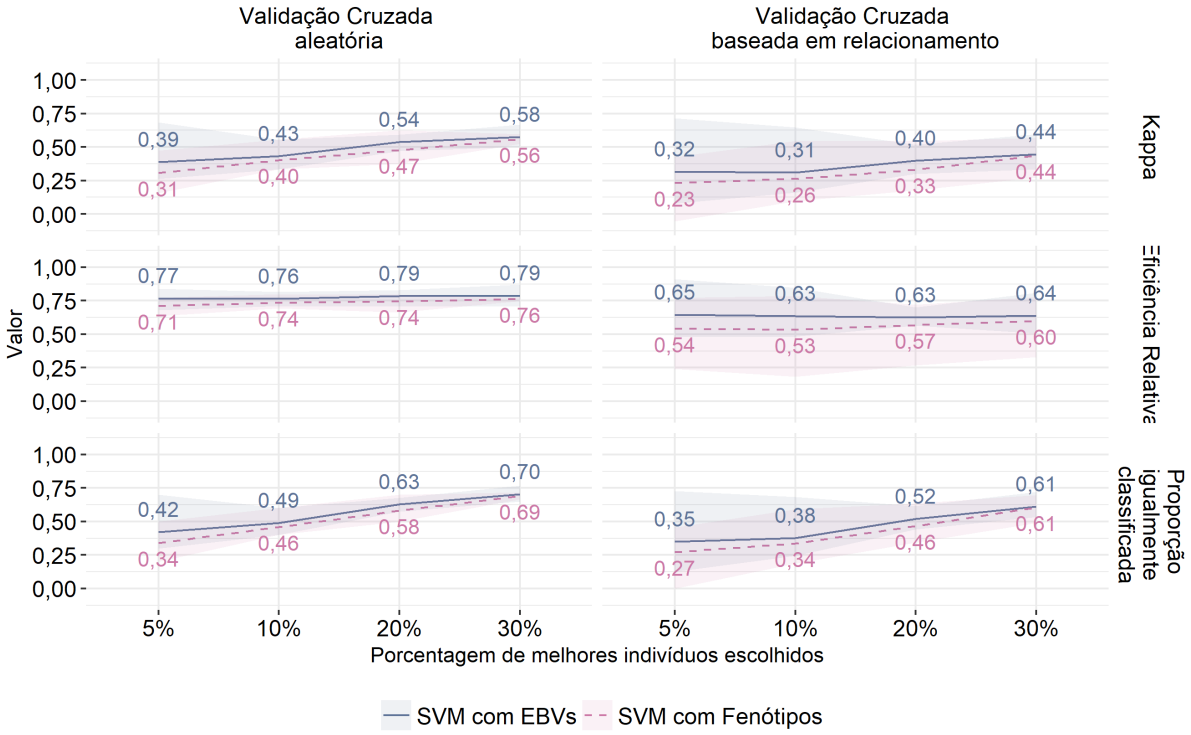


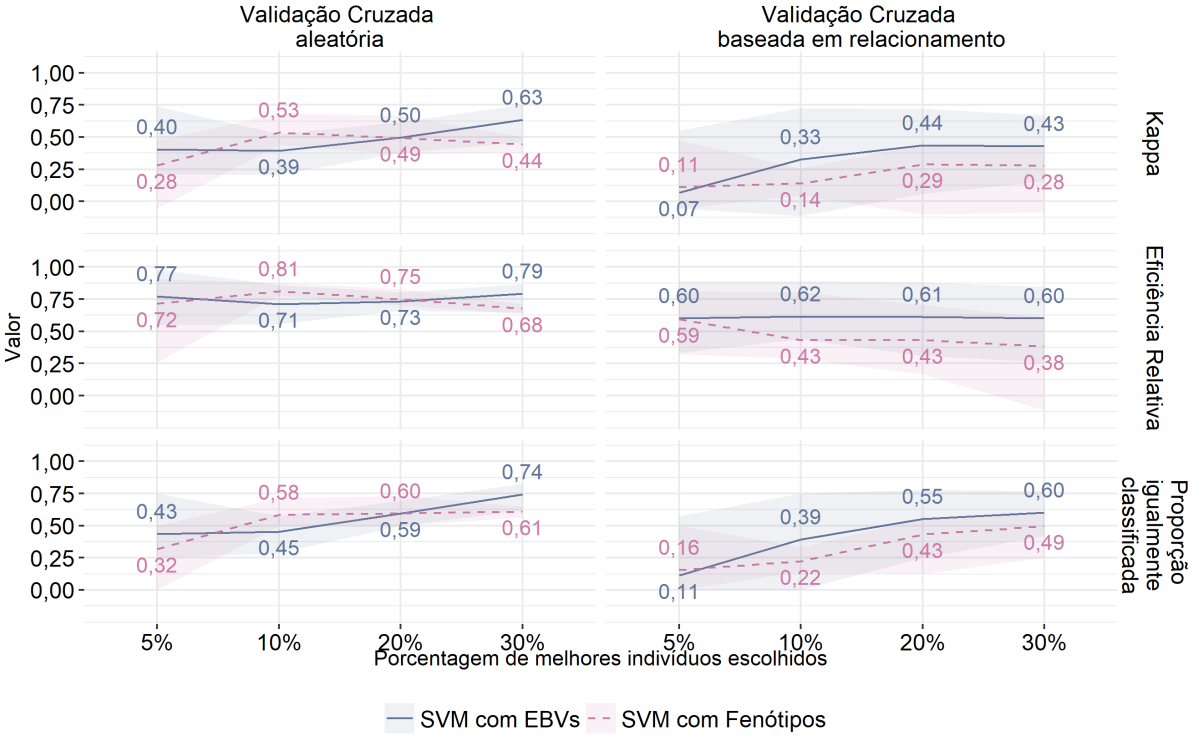
Figura 18 – Comparação de médias e amplitudes de medidas obtidas por SVM com EBVs e fenótipos como resposta, com proporção 30-70 e kernel radial para relação S:G



Nota-se que as métricas para o fenótipo de crescimento diâmetro à altura do peito (Figura 17) apresenta valores próximos para ambos os SVMs. Na maioria dos casos o SVM com EBVs como resposta apresentaram médias ligeiramente superiores. Contrastando ambas as formas de validação cruzada observa-se que a baseada em relacionamento apresenta métricas com médias inferiores, porém que variam mais.

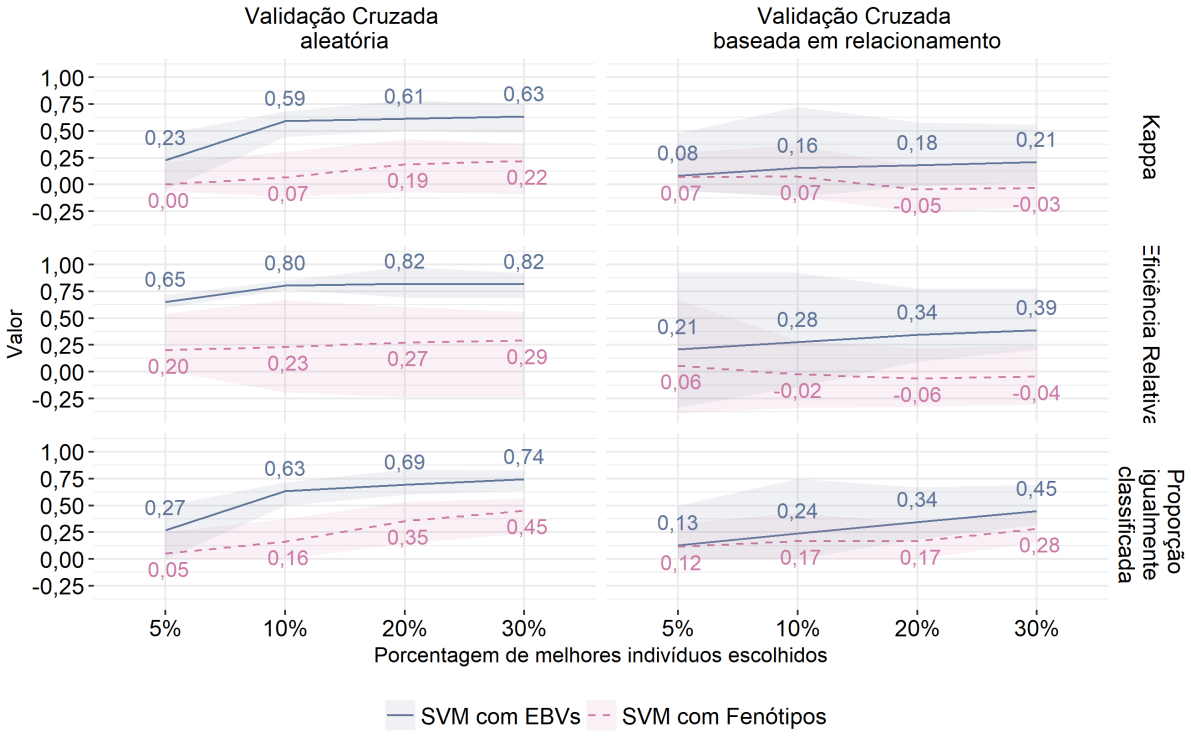
Para o fenótipo químico relação S:G (Figura 18), este comportamento permanece, diferenciando no fato de que as métricas dos SVMs são mais próximas. Enquanto que para o fenótipo físico com maior herdabilidade, comprimento de fibra (Figura 19), há uma troca entre qual SVM se mostrou em média superior na validação cruzada aleatória. Na outra abordagem de validação o SVM com EBVs apresentou médias superiores, com uma diferença entre os dois métodos mais evidenciada em relação aos fenótipos anteriores, porém com amplitudes maiores neste caso.

Figura 19 – Comparação de médias e amplitudes de medidas obtidas por SVM com EBVs e fenótipos como resposta, com proporção 30-70 e kernel radial para comprimento da fibra



A característica com menor herdabilidade (Figura 20) foi a que apresentou as maiores diferenças entre as duas abordagens de ajuste do SVM, independente da forma de validação cruzada. Ao se selecionar 50 árvores as métricas ainda são próximas, mas nos demais são bastante superiores no caso do SVM com EBVs. Nota-se principalmente na eficiência relativa que o SVM com fenótipos apresenta valores negativos com ambas as abordagens de validação cruzada para todas as proporções de seleção.

Figura 20 – Comparação de médias e amplitudes de medidas obtidas por SVM com EBVs e fenótipos como resposta, com proporção 30-70 e kernel radial para ângulo microfibrilar



3.5 Comparação entre RRBLUP e SVM

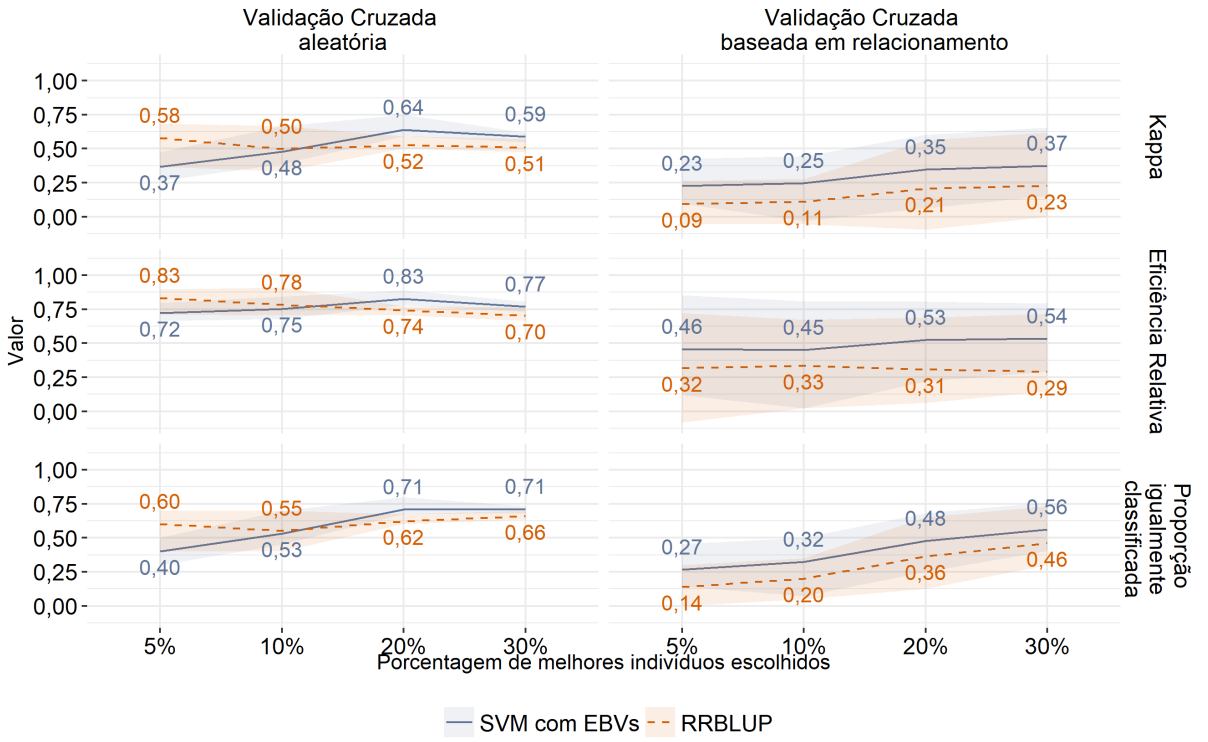
Os modelos RRBLUP e o SVM com EBVs como resposta foram ajustados e as métricas de avaliação do modelo foram obtidas considerando-se as duas abordagens de validação cruzada. Mesmo que as medidas kappa e eficiência relativa não sejam utilizadas para variáveis contínuas, e sim dicotômicas, ainda é possível obtê-las para o caso do RRBLUP, em que os indivíduos são ordenados a partir dos seus GEBVs.

Ao observar o fenótipo diâmetro à altura do peito (Figura 21), nota-se que as métricas do SVM e do RRBLUP foram próximas, sempre com suas amplitudes se sobrepondo.

Ao observar o fenótipo diâmetro à altura do peito (Figura 21), nota-se que ao utilizar a validação cruzada aleatória e selecionando-se apenas 5% das plantas, o modelo de regressão apresentou métricas ligeiramente superiores. Já ao selecionar mais indivíduos as métricas do SVM se tornam superiores, porém em todos os casos as amplitudes se sobrepõem.

Na abordagem de validação baseada em relacionamento o comportamento médio muda: em todas as medidas os valores médios obtidos pelo algoritmo de classificação foram superiores. Ao se considerar a amplitude dos valores obtidos nos *folds* tem-se que as métricas foram semelhantes, visto que eles se sobrepõem.

Figura 21 – Comparação entre médias e amplitudes de medidas obtidas com RRBLUP e SVM com EBVs para diâmetro à altura do peito considerando a proporção 30-70 e kernel radial



Para a relação S:G (Figura 22) as métricas médias obtidas pelo RRBLUP se mostraram ligeiramente superiores em praticamente todos os casos, com uma diferença de no máximo 10 pontos percentuais. Entretanto suas amplitudes se mostraram bem próximas em ambas as abordagens de validação cruzada, indicando uma semelhança entre os métodos.

Os fenótipos físicos (Figuras 23 e 24) apresentam comportamentos semelhantes. Na abordagem de validação cruzada aleatória as amplitudes indicam que os valores para os dois algoritmos foram próximas, em que para cada porcentagem de indivíduos selecionados as médias do SVM e do RRBLUP alternam em qual é superior. Já na validação cruzada baseada em relacionamento as métricas médias do SVM em geral se mostraram ligeiramente superiores, porém sua variação também é consideravelmente maior que a do RRBLUP. Novamente suas amplitudes se sobrepõem em ambas as abordagens de validação cruzada, o que evidencia uma semelhança entre as métricas obtidas pelo SVM e pelo RRBLUP.

Figura 22 – Comparação entre médias e amplitudes de medidas obtidas com RRBLUP e SVM com EBVs para relação S:G considerando a proporção 30-70 e kernel radial

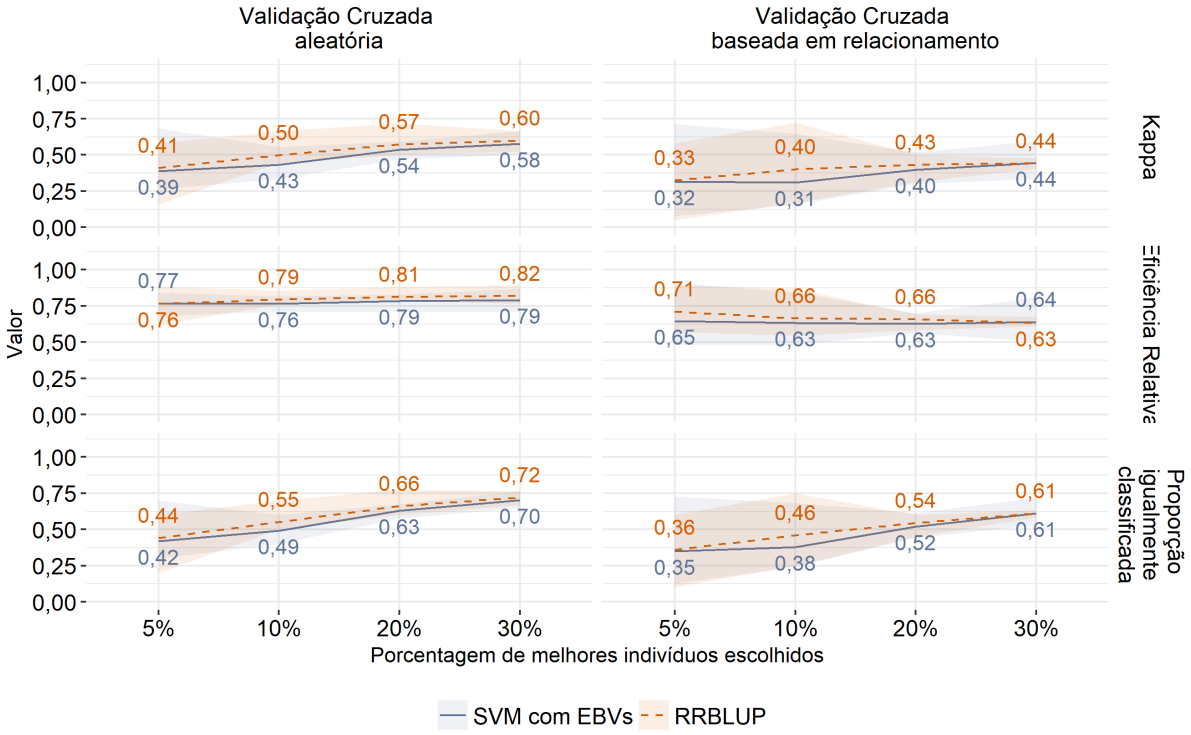


Figura 23 – Comparação entre médias e amplitudes de medidas obtidas com RRBLUP e SVM com EBVs para comprimento da fibra considerando a proporção 30-70 e kernel radial

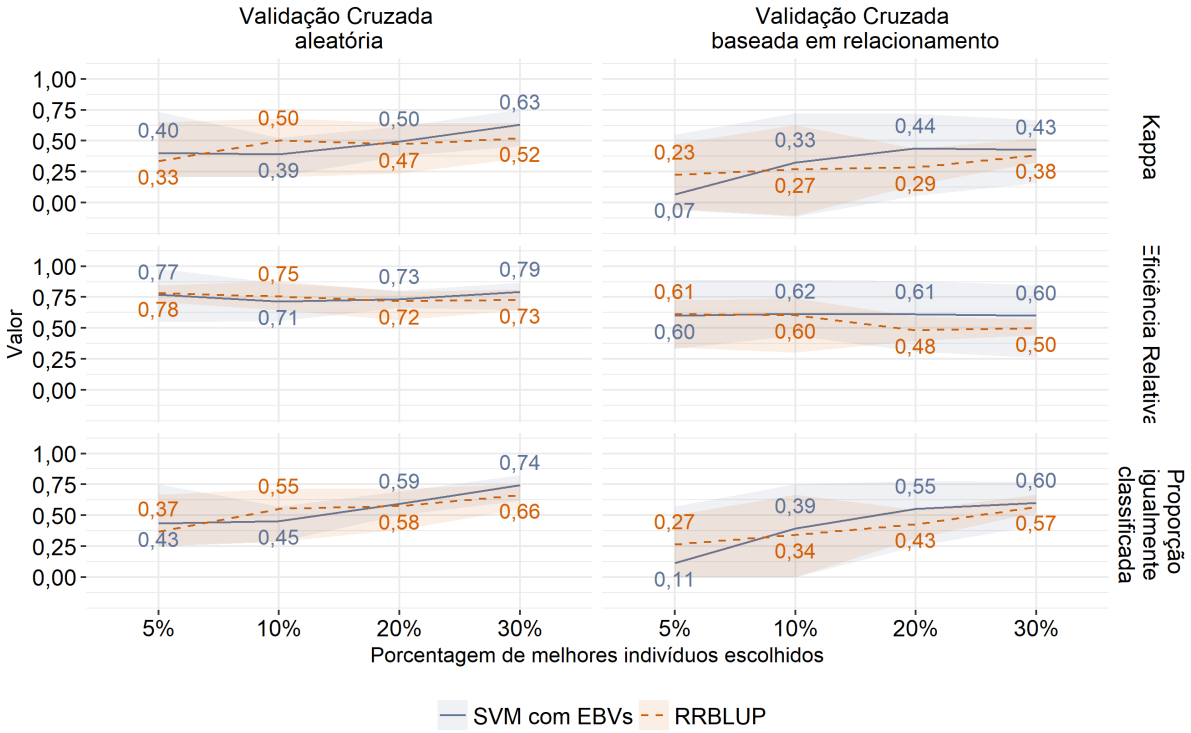
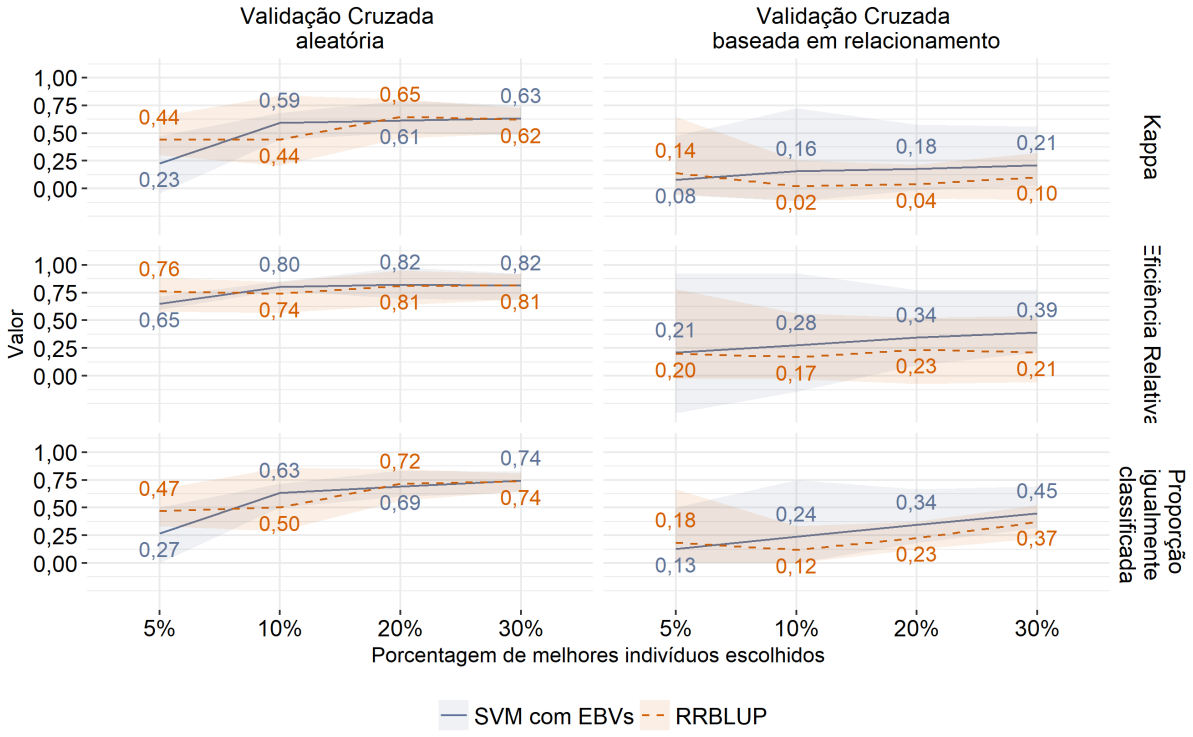


Figura 24 – Comparação entre médias e amplitudes de medidas obtidas com RRBLUP e SVM com EBVs para ângulo microfibrilar considerando a proporção 30-70 e kernel radial



4 Discussão e Conclusão

Neste trabalho utilizou-se modelos mistos para a previsão de 15 fenótipos de crescimento, químicos e físicos em dados de eucalipto. Implementou-se ainda um algoritmo de aprendizado de máquinas de classificação com duas resposta diferentes (fenótipos e EBVS), buscando métricas melhores ao se minimizar os erros de coleta fenotípica. Ambos são comparados com o modelo de melhoramento genético tradicional, buscando modelos com alto poder preditivo, que permita diminuir o tempo entre ciclos reprodutivos e diminuir os custos envolvidos. Isto é importante principalmente em relação a fenótipos complexos que demoram um tempo maior para serem obtidos, além da dificuldade de medição e altos investimentos.

Os modelos implementados possibilitam a obtenção de GEBVs ou a classificação de indivíduos com base em suas informações genéticas. Porém,informações parentais são incluídas, pelo fato de se ter uma estrutura familiar. Isto influencia nos coeficientes obtidos nos modelos, bem como nas métricas de avaliação. Buscou-se controlar o efeito parental por meio da divisão dos grupos de validação cruzada, em que grupos de treinamento e teste são pouco correlacionados.

Espera-se com isto que o pressuposto de independência permaneça válido e as estimativas sejam menos otimistas. Assim como observado em Roberts et al. (2017) e Resende et al. (2017), as métricas obtidas utilizando-se os grupos de validação cruzada baseada em relacionamento variavam mais e foram menores em média que com a validação cruzada com grupos separados aleatoriamente. Em alguns traços esta diminuição foi maior, provavelmente por se tratar de fenótipos mais influenciados por efeitos parentais; em outros não houve tanta diferença entre eles, como para a relação S:G, por exemplo. Entretanto, na prática, esse efeito familiar está presente, pois as plantas utilizadas no modelo serão descendentes da população de treinamento. Todavia, podem haver cruzamentos entre famílias e espécies diferentes que não foram observados e previstos anteriormente. Então, com as duas formas de validação cruzada é possível se observar o comportamento do modelo nos dois extremos, um provavelmente superestimado, considerando as estruturas de famílias específicas, e outro provavelmente subestimado, sem considerá-las.

Neste estudo optou-se por utilizar apenas uma partição aleatória possível. Porém, outras abordagens poderiam ser utilizadas, como a validação cruzada *leave one out*, diversas partições aleatórias, outros valores de K para o K-*fold*, entre outros. Além disso, outros modelos como Lasso Bayesiano, Bayes A, Bayes B Random Forest ou Redes Neurais poderiam ser utilizados.

Na Regressão Ridge BLUP as capacidades preditivas apresentaram uma relação

positiva com as herdabilidades, isto é , fenótipos que apresentam herdabilidade mais altas também possuem capacidade preditiva alta. Isso pode se dar pelo fato de que variáveis com alta herdabilidade possuem uma alta correspondência entre genótipo e fenótipo. Com isso os dados genéticos são capazes de predizer os GEBVs com maior precisão. Nota-se também que há correlações de Spearman altas entre os valores obtidos com o RRBLUP e o melhoramento tradicional.

O ajuste do algoritmo de classificação do SVM foi realizado apenas para alguns fenótipos, devido ao custo computacional. Foram escolhidos aqueles fenótipos com maior herdabilidade de cada grupo: crescimento, químicos e físicos, o diâmetro à altura do peito, relação S:G e comprimento de fibra com herdabilidades 0,52, 0,89 e 0,70 respectivamente. Além disso ajustou-se também o modelo para a característica ângulo microfibrilar, que possui herdabilidade de 0,14, dados apenas para 348 árvores e seus grupos de validação cruzada baseada em relacionamento são desbalanceados (Figura 7).

Ao se transformar o fenótipo, ou GEBVs, em uma variável dicotômica, um valor de corte deve ser escolhido a partir do qual definiu-se grupos com as maiores e menores medidas. Essa escolha pode influenciar na habilidade preditiva do modelo, pois pode tornar os grupos mais ou menos desbalanceados. Com isso selecionou-se as proporções 50-50, 40-60, 30-70, 20-80 e 25-85 e optou-se por utilizar a proporção 30-70 por se aproximar do que ocorre na prática no melhoramento. Modelos considerando kernels linear e radial foram utilizados, mas por fim o radial foi escolhido. Tais escolhas podem influenciar nas métricas obtidas e por isso as decisões devem considerar o estudo em questão. Além disso, os parâmetros dos modelos podem variar de acordo com a sua implementação e podem afetar nas medidas de avaliação do modelo.

Tanto para o modelo de regressão ridge BLUP quanto para o SVM, há uma diferença entre as métricas obtidas nas diferentes abordagens de validação cruzada. Nota-se que, em média, as obtidas pela validação estratificada são menores que as obtidas pela validação aleatória, porém variam mais entre cada *fold*. Como futuramente estes modelos serão utilizados para classificar indivíduos não pertencentes à população de treinamento, as medidas obtidas com grupos aleatoriamente selecionados são otimistas; porém por minimizar o efeito de parentesco, os grupos obtidos por clusterização hierárquica apresentam-se pessimistas. Com isso consegue-se obter uma noção mais real das métricas e do poder preditivo real do modelo, ao não se basear apenas em medidas superestimadas.

Outro questionamento trazido neste trabalho é a utilização de GEBVs do BLUP fenotípico com a matriz realizada como variável resposta do SVM, em vez de apenas os fenótipos ajustados por efeitos de delineamento. O GEBV representa um valor de maior interesse que o fenótipo nesse contexto, pois o fenótipo está sujeito a outras fontes de variação, como efeitos ambientais, erros de coleta e mensuração, entre outros. Ao comparar os resultados obtidos no SVM utilizando os GEBVs como resposta e os fenótipos ajustados

pelos efeitos de delineamento, observou-se que os fenótipos de crescimento e químicos apresentaram métricas mais próximas. Já os fenótipos físicos, que possuem dados apenas para cerca de 350 árvores o SVM com EBVs apresentou medidas superiores. A diferença entre eles é ainda mais evidenciada na validação cruzada baseada em relacionamento.

Por fim, medidas de avaliação dos modelos RRBLUP e SVM com GEBVs, nota-se que para ambos elas são bem próximas para todos os fenótipos, independente da validação cruzada utilizada. Ou seja, os resultados obtidos por ambos são consistentes e a escolha do modelo a ser utilizado deve ser feita considerando-se o fenótipo em questão, a quantidade de árvores que serão selecionadas, o custo computacional e assim por diante.

Traçando-se um paralelo com outros artigos que comparam metodologias de Seleção Genômica, tanto em eucaliptos quanto em outras plantas, nota-se algumas diferenças. Artigos como Ornella et al. (2014) fazem comparações entre as métricas obtidas por diversos modelos, porém utilizam diretamente os fenótipos obtidos como resposta e utilizam uma abordagem de validação cruzada, separando aleatoriamente os indivíduos entre os grupos. Neste exemplo citado, o esquema de validação cruzada considera diversas repetições de grupos; porém a estrutura familiar não foi avaliada, o que pode ter gerado superestimação das métricas. Em compensação, são realizados testes que verificam a significância da diferença entre as métricas de diversos modelos, algo que não foi realizado no presente trabalho.

Há muito ainda para ser explorado no contexto estatístico de Seleção Genômica para plantas. Além de testar outras abordagens de aprendizado de máquinas como *Random Forest*, Redes Neurais, *Support Vector Regression*, ainda é possível se explorar técnicas de diminuição de dimensão visando melhorar o modelo. Outra abordagem envolve a criação de um índice de "planta elite"que agrupe diversos fenótipos, com técnicas multivariadas por exemplo, permitindo a escolha das melhores plantas a partir desta medida global, uma vez que em geral é feito um modelo para cada fenótipo independentemente. Ou seja, ainda são necessários diversos estudos futuros para entender e aprimorar modelos de previsão, no contexto genômico de plantas de floresta.

Bibliografia

BATES, Douglas; VAZQUEZ, Ana Ines. *pedigreemm: Pedigree-based mixed-effects models*. [S.l.], 2014. R package version 0.3-3. Disponível em: <`https://CRAN.R-project.org/package=pedigreemm`>.

CAREY, VJ; WANG, You-Gan. *Mixed-effects models in S and S-PLUS*. [S.l.]: Taylor & Francis, 2001.

CORTES, Corinna; VAPNIK, Vladimir. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995.

DESTA, Zeratsion Abera; ORTIZ, Rodomiro. Genomic selection: genome-wide prediction in plant improvement. *Trends in Plant Science*, v. 19, n. 9, p. 592–601, 2014. ISSN 1360-1385. DOI: `http://dx.doi.org/10.1016/j.tplants.2014.05.006`. Disponível em: <`http://www.sciencedirect.com/science/article/pii/S1360138514001411`>.

ENDELMAN, J. B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome*, v. 4, p. 250–255, 2011.

GIANOLA, Daniel et al. Additive genetic variability and the Bayesian alphabet. *Genetics*, Genetics Soc America, v. 183, n. 1, p. 347–363, 2009.

GRATTAPAGLIA, D. *Breeding forest trees by genomic selection: current progress and the way forward. In "Genomics of Plant Genetic Resources Vol 1 pp 651-682. eds R. Tuberosa, A. Graner & E. Frison.* [S.l.], 2014.

HAIR, Joseph F et al. *Análise multivariada de dados*. [S.l.]: Bookman Editora, 2009.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Biometrics*, 2002.

HENDERSON, Charles R. Selection index and expected genetic advance. *Statistical genetics and plant breeding*, Washington, DC, v. 982, p. 141–163, 1963.

JED WING, Max Kuhn. Contributions from et al. *caret: Classification and Regression Training*. [S.l.]. R package version 6.0-78. Disponível em: <`https://github.com/topepo/caret/`>.

LIMA, Bruno Marco de. *Bridging genomics and quantitative genetics of Eucalyptus: genome-wide prediction and genetic parameter estimation for growth and wood properties using high-density SNP data*. 2014. Tese (Doutorado) – Escola Superior de Agricultura "Luiz de Queiroz".

LIN, Z; HAYES, BJ; DAETWYLER, HD. Genomic selection in crops, trees and forages: a review. *Crop and Pasture Science*, CSIRO, v. 65, n. 11, p. 1177–1191, 2014.

ORNELLA, L et al. Genomic-enabled prediction with classification algorithms. *Heredity*, Nature Publishing Group, v. 112, n. 6, p. 616, 2014.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2017. Disponível em: <https://www.R-project.org>.

RESENDE, RT et al. Assessing the expected response to genomic selection of individuals and families in Eucalyptus breeding with an additive-dominant model. *Heredity*, Nature Publishing Group, v. 119, n. 4, p. 245, 2017.

RICHARD A. JOHNSON, Dean W. Wichern. *Applied Multivariate Statistical Analysis (6th Edition)*. 6. ed. [S.l.]: Prentice Hall, 2007. ISBN 0131877151,9780131877153. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=43C0D654A6EC0D23AA6697C7619DAF27>.

ROBERTS, David R et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, Wiley Online Library, v. 40, n. 8, p. 913–929, 2017.

SEARLE, Shayle R; CASELLA, George; MCCULLOCH, Charles E. *Variance components*. [S.l.]: John Wiley & Sons, 2009. v. 391.

SILVA-JUNIOR, OB et al. Eucalyptus genotyping taken to the next level: development of the"EucHIP60k. br"based on large scale multi-species SNP discovery and ascertainment, pp. In: IUFRO Tree Biotechnology Conference 2013. [S.l.: s.n.], 2013.

VANRADEN, Paul M. Efficient methods to compute genomic predictions. *Journal of dairy science*, Elsevier, v. 91, n. 11, p. 4414–4423, 2008.

WAHBA, Grace et al. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. *Advances in Kernel Methods-Support Vector Learning*, v. 6, p. 69–87, 1999.

WIMMER, Valentin et al. synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics*, v. 28, n. 15, p. 2086–2087, 2012.