ISSN: 2317-0840

Aplicação de Modelos Mistos e SVM em Seleção Genômica de árvores de Eucalipto

Ana Gabriela P. Vasconcelos[1†], Joanlise M. L. Andrade[2], Bernardo B. Andrade[3], Dario Grattapaglia[4], Bruno M. Lima[5]

[1] *Instituto de Matemática e Estatística - Universidade de São Paulo*
[2] *Departamento de Estatística - Universidade de Brasília. E-mail:* joanlise@unb.br.
[3] *Departamento de Estatística - Universidade de Brasília. E-mail:* bbandrade@unb.br.
[4] *Embrapa Recursos Genéticos e Biotecnologia. E-mail:* dario.grattapaglia@embrapa.br.
[5] *Centro de Tectonologia - Fibria. E-mail:* bruno.lima@fibria.com.br.

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 fenótipos. 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 se utilizar informações moleculares para se estimar efeitos de marcadores genéticos com base em um modelo de predição. O modelo, desenvolvido em uma população de treinamento com informações genotípicas e fenotípicas, é utilizado para se obter valores genéticos baseados em dados genotípicos de plantas candidatas. Portanto, a escolha do modelo é uma etapa essencial. Este estudo compara modelos mistos e SVMs em dados de eucaliptos, além de avaliar fatores que influenciam as métricas obtidas, 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 superiores. Ainda foi possível verificar a importância do controle dos efeitos de parentesco por meio da validação cruzada para a obtenção de métricas menos otimistas, uma vez que os modelos são utilizados com dados de indivíduos não incluídos 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; melhoramento genético; regressão ridge; validação cruzada, SVM.

Abstract: *This study aims to compare Ridge Regression models and SVM focusing on factors that influence perdition metrics, such as quality of data collected and relationship effects. The importance of controlling family effects through cross validation was also verified. SVM and regression models showed consistent and similar results.*

Keywords: Genomic selection; improvement program; ridge regression; cross validatio;, SVM.

[†]Autora correspondente: anagabipv@gmail.com.

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. Entretanto, o melhoramento tradicional consiste na seleção de indivíduos com base na observação de características de interesse, como a altura, por exemplo. Para tanto, é necessário que as plantas já estejam desenvolvidas a ponto de permitir a comparação de tais características, 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.

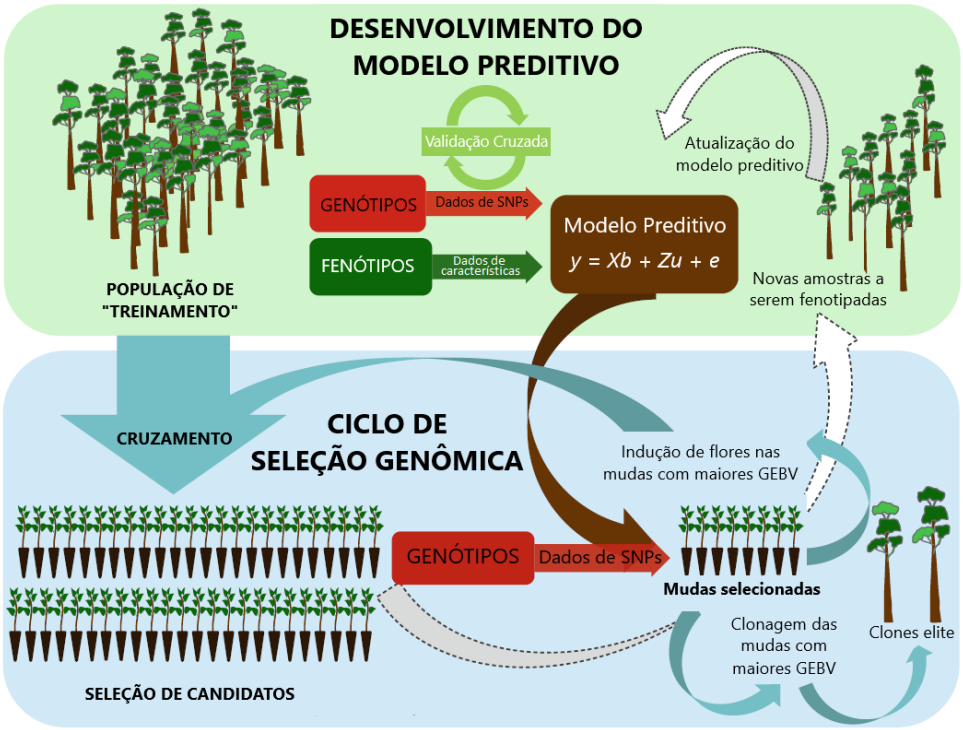


Figura 1: Processo de Seleção Genômica
Fonte: Adaptação de Grattapaglia (2014)

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.

Portanto, a escolha do modelo de predição é uma etapa essencial do processo de melhoramento e deve-se buscar maneiras de aperfeiçoar 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 têm como origem um estudo contendo 2784 plantas de eucalipto pertencentes à 58 famílias de irmãs completas e meio irmãs, cada uma com 48 indivíduos. O delineamento utilizado envolveu 8 blocos com 6 indivíduos de cada família. Após seleção de uma subamostra com base em volume de madeira obteve-se genótipos de 999 árvores pertencentes à 45 famílias. Com tal seleção cada bloco incluiu até 6 indivíduos de cada família, sendo que algumas acabaram não estando representadas em todos os blocos.

Foram medidos 15 fenótipos (características detectáveis) químicos, físicos e de crescimento. As informações genéticas foram obtidas pela genotipagem de 60.639 marcadores do tipo SNP utilizando um chip específico para Eucaliptos (SILVA-JUNIOR et al.,2013). Após procedimentos de controle de qualidade, tais como exclusão de marcadores com frequência do alelo menos comum inferior a 1%, com genótipos constantes ou cuja frequência de dados faltantes tenha sido superior a 10%, selecionou-se 27.573 marcadores (variáveis explicativas) para as análises.

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 Modelos de Regressão

Utilizou-se o modelo misto dado por

$$y = Zu + X\beta + e, \tag{1}$$

para o qual supõe-se que $u \sim N(0, \sigma_u^2 \cdot I)$ e $\sim N(0, \sigma_e^2 \cdot I)$, em que u e β são os vetores de efeitos aleatórios e fixos, Z e X as matrizes de incidência de efeitos aleatórios e fixos, respectivamente, e é o erro residual, K é uma matriz de variância-covariância dos efeitos aleatórios e y representa a variável resposta.

A partir de (1) derivam-se três modelos conhecidos. O RRBLUP (*Ridge Regression Best Linear Unbiased Prediction*) utiliza informações genotípicas para prever os GEBVs[ii]. Neste modelo a matriz Z representa os dados genéticos e a matriz K é a identidade, isto é, os efeitos aleatórios apresentam variância constante σ_u^2. Seus efeitos são estimados conforme a metodologia proposta por Henderson (1963),dada por

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

que se assemelha à estimação dos coeficientes da Regressão Ridge (SEARLE; CASELLA; MCCULLOCH, 2009). Trata-se, portanto, de um modelo de regularização que assume que todos os marcadores contribuem homogeneamente e em pequena escala para os valores preditos. No PBLUP (*Phenotypic Best Linear Unbiased Prediction*) são utilizadas informações apenas de parentesco entre indivíduos para prever os GEBVs (neste caso chamados de EBVs). A matriz de incidência de efeitos aleatórios Z é considerada como a identidade e o vetor de efeitos aleatório segue uma distribuição com média zero e variância $\sigma_u^2 K$, em que K é a matriz que verifica a relação genética entre os indivíduos a partir da ancestralidade declarada da planta. Este modelo corresponde ao modelo de melhoramento tradicional, uma vez que não utiliza informações genéticas de marcadores, apenas fenotípicas e de parentesco. Já o terceiro modelo, GBLUP (*Genetic Best Linear Unbiased Prediction*), também possui a matriz de incidência Z como a identidade, porém $K = G$ é a matriz de relacionamento estimada, obtida por (VANRADEN, 2008):

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

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.

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

$$h^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_e^2}, \tag{3}$$

sendo $\hat{\sigma}_T^2, \hat{\sigma}_g^2$ e $\hat{\sigma}_e^2$ as variâncias estimadas total, genética e dos resíduos, respectivamente. A variância genética se dá por (GIANOLA et al.,2009)

[ii] 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 p_M é a frequência do alelo mais comum no marcador M e 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 (1) (GEBV_i) e o fenótipo observado ($y_i^{(\text{obs})}$) em cada grupo dos k grupos da validação cruzada, vistos na seção 2.4:

$$r_{y\hat{y}} = \frac{\sum_{i=1}^{k} corr(GEBV_i, y_i^{(\text{obs})})}{k}. \tag{4}$$

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 ideia será descrita a seguir.

2.3 SVM

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 (Figura 2) 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. Porém, para considerar funções de decisão além de lineares é possível utilizar diferentes Kernels, como o polinomial ou radial.

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:

- **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.

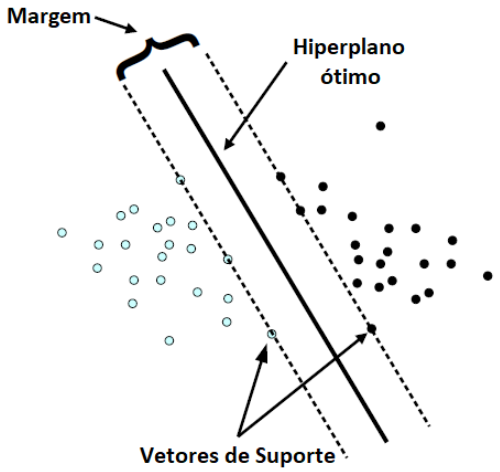


Figura 2: Representação de SVM
Fonte: Adaptação de Meyer e Wien (2001)

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

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

- **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.

2.4 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 com base na matriz de relacionamento realizada. 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.

2.5 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.

Visando controlar os efeitos de delineamento, os fenótipos foram ajustados pelos efeitos de blocos, por um modelo de efeitos aleatórios, e estes foram utilizados como variável resposta dos modelos preditivos. 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, assim como feito nos modelos de regressão, 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 $\gamma = (2^{-20}, 2^{-15}, 2^{-10})$.

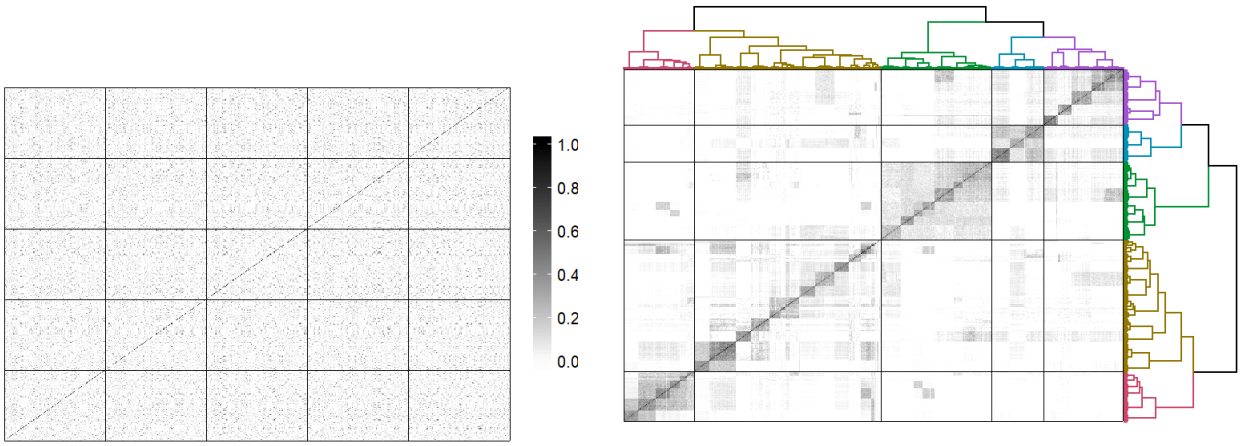
A avaliação do modelo, tanto da Regressão Ridge BLUP quanto do SVM, foi obtida a partir da validação cruzada 5-fold. As medidas de avaliação kappa e eficiência relativa consideram uma porcentagem (5, 10, 20 e 30%) de indivíduos com maiores GEBVs, ou probabilidade de pertencer à classe de elite no caso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

3 Resultados

3.1 Validação Cruzada

Como já mencionado, a avaliação do poder preditivo dos modelos foi realizada através de uma validação cruzada 5-fold com duas abordagens diferentes. Na aleatória, os indivíduos foram selecionados aleatoriamente para cada grupo, o que resultou em grupos de mesmo tamanho. Já na abordagem baseada os grupos possuem tamanhos diferentes, variando entre 105 e 374 indivíduos.

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 3b, 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.



(a) Grupos obtidos aleatoriamente

(b) Grupos obtidos por agrupamento hierárquico

Figura 3: Matriz de Relacionamento

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. 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 envolver indivíduos de diversas famílias em ambos.

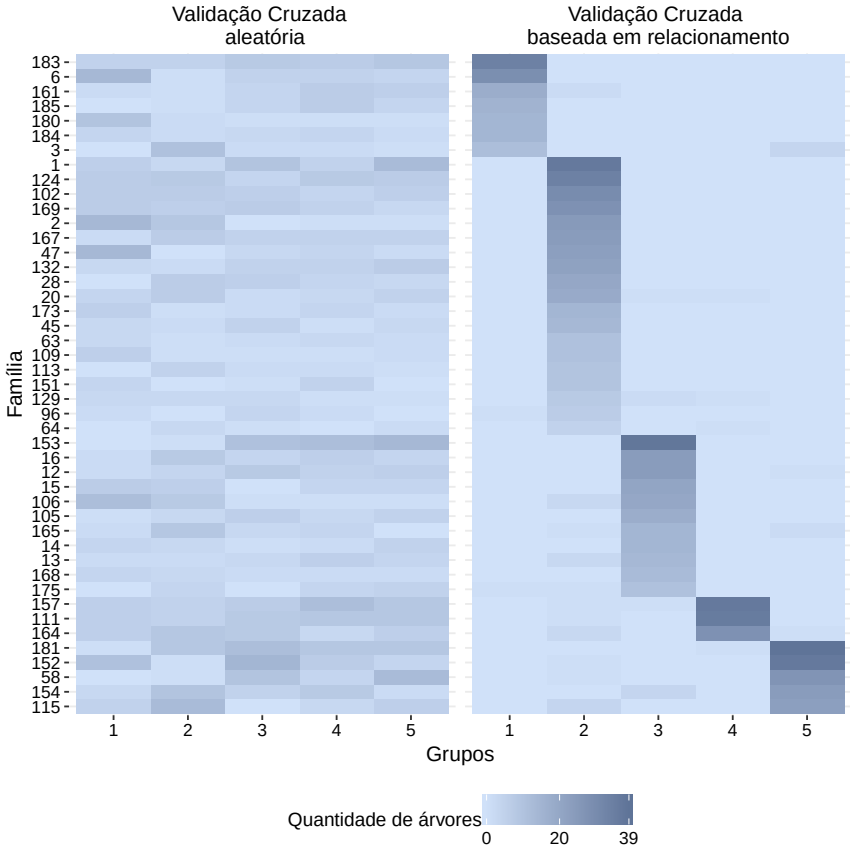


Figura 4: Frequência de famílias por grupos de validação cruzada

3.2 Modelos Mistos

Considerando-se os grupos de validação cruzada, implementou-se o modelo de Regressão Ridge BLUP (RRBLUP). A Tabela 2 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 0,35. 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 largura de fibra 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 de crescimento apresentam capacidades preditivas próximas de 0,6, no caso aleatório e 0,30 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 maiores medidas, 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 passa a ter capacidade preditiva negativa, enquanto a Rigidez apresenta valores superiores aos fenótipos de crescimento, o que não ocorreu na abordagem aleatória.

Tabela 2: Medidas obtidas para o RRBLUP

Fenótipo	n	Herdabilidade	Capacidade Preditiva	
			Validação cruzada aleatória	Validação Cruzada baseada em relacionamento
DAP	999	0,42	0,68	0,39
Altura	999	0,32	0,61	0,26
Volume	999	0,41	0,68	0,37
IMA	999	0,41	0,67	0,37
Celulose*	999	0,55	0,74	0,39
Hemicelulose*	999	0,63	0,69	0,33
Relação S:G*	999	0,84	0,89	0,75
Lignina Insolúvel*	999	0,65	0,81	0,40
Lignina Solúvel*	999	0,69	0,87	0,71
Lignina Total*	999	0,65	0,79	0,37
Densidade*	999	0,57	0,82	0,66
Ângulo Microfribilar	348	0,12	0,39	0,24
Comprimento de fibra	350	0,61	0,61	0,42
Largura de fibra	350	0,10	0,29	-0,12
Rigidez	349	0,27	0,61	0,47

*: fenótipos obtidos com auxílio da espectometria NIRS.

3.3 SVM

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 reali-

zada 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 uma das menores herdabilidades e grupos desbalanceados.

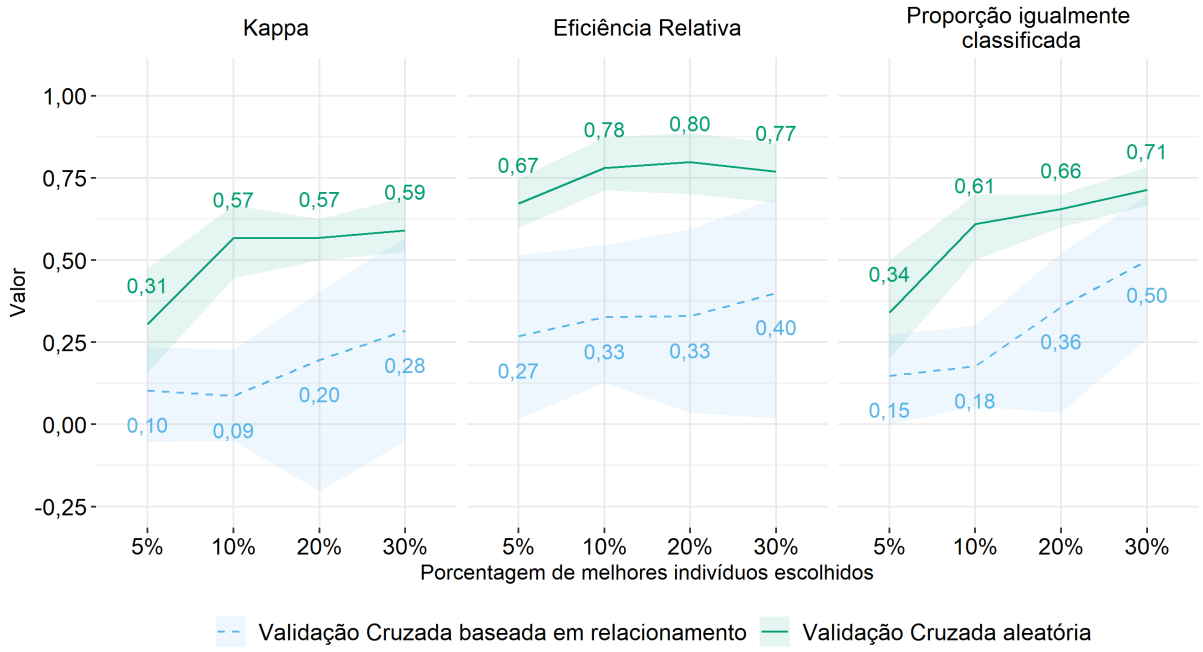


Figura 5: Médias e amplitudes de medidas de avaliação para **diâmetro à altura do peito** obtidas por SVM com EBVs como resposta, proporção 30-70 e kernel radial

As Figuras 5 a 16 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 5), com validação cruzada aleatória, nota-se que a medida kappa indica uma concordância fraca ao selecionar 5% dos indivíduos; ainda assim, apresentou eficiência relativa próxima de 0,7 e 34% 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. Já para a validação cruzada utilizando grupos obtidos por clusterização, ainda considerando a Figura 5, as medidas variam bastante em cada *fold*, e em geral são menos otimistas do que aquelas obtidas pela validação aleatória.

As medidas médias obtidas para a relação S:G (Figura 6) apresentam um comportamento médio semelhante para as duas formas de validação cruzada, porém nota-se que a

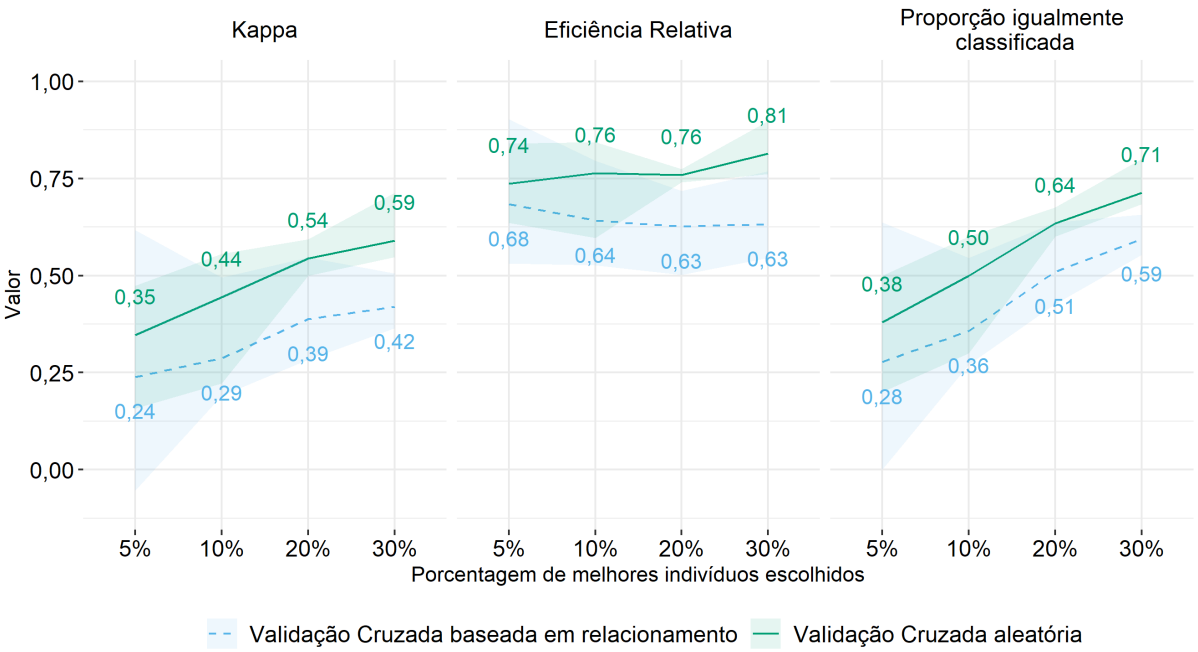


Figura 6: Médias e amplitudes de medidas de avaliação para **relação S:G** obtidas por SVM com EBVs como resposta, proporção 30-70 e kernel radial

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 70% no caso em que os grupos de validação cruzada são obtidos aleatoriamente.

O comprimento de fibra (Figura 7), que contém dados para aproximadamente 350 plantas, apresenta amplitudes maiores para a validação aleatória em relação aos fenótipos anteriores. Nota-se que, em alguns casos, a validação baseada em relacionamento resultou em métricas superiores à aleatória 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 0,85 e classifica-se igualmente acima de 0,38 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 0,91 até 0,63, ao selecionar se 50 e 300 árvores, respectivamente.

Por fim, para o ângulo microfibrilar (Figura 8) percebe-se uma diferença grande entre

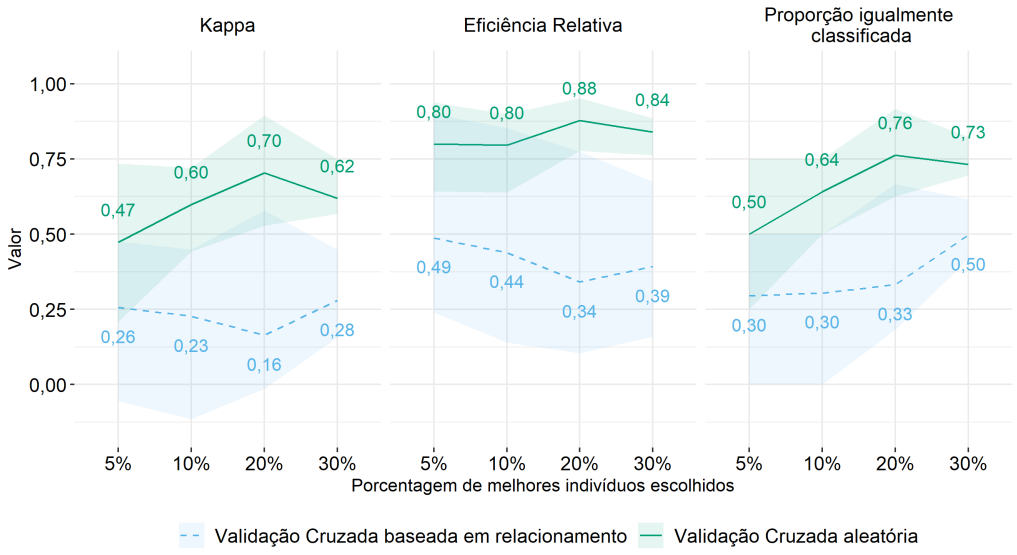


Figura 7: Médias e amplitudes de medidas de avaliação para **ângulo microfibrilar** obtidas por SVM com EBVs como resposta, proporção 30-70 e kernel radial

as médias dos valores obtidos pelas duas abordagens de validação cruzada. A medida kappa se mostra moderada ao se selecionar 50 árvores e passa a ser boa ao se selecionar mais árvores. A eficiência relativa se mostrou acima de 0,8 para todos os casos, com proporção igualmente classificada chegando até 0,76. 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 0,49. A proporção igualmente classificada média acompanha os baixos valores, com um crescimento gradativo e novamente com amplitude se sobrepondo à outra.

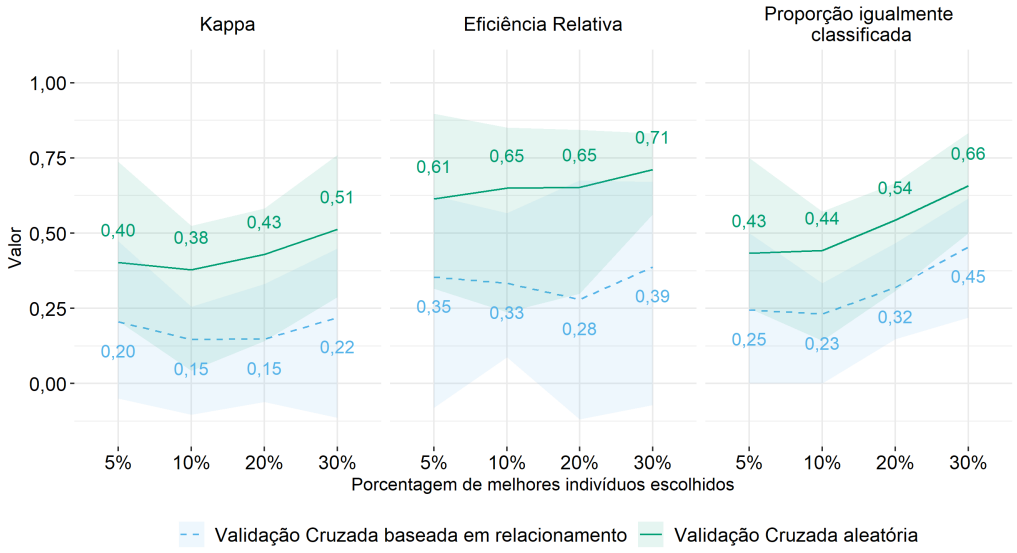


Figura 8: Médias e amplitudes de medidas de avaliação para **largura de fibra** obtidas por SVM com EBVs como resposta, proporção 30-70 e kernel radial

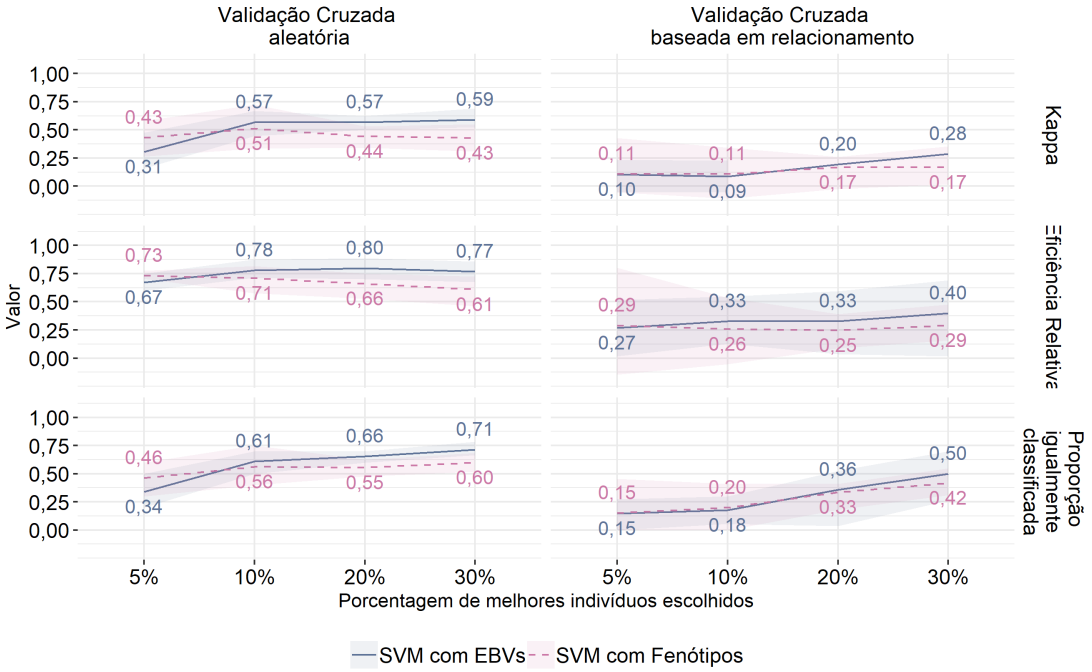


Figura 9: Comparação de médias e amplitudes de medidas obtidas por SVM com EBVs e fenótipos como resposta, com proporção 30-70 e kernel radial para **diâmetro à altura do peito**

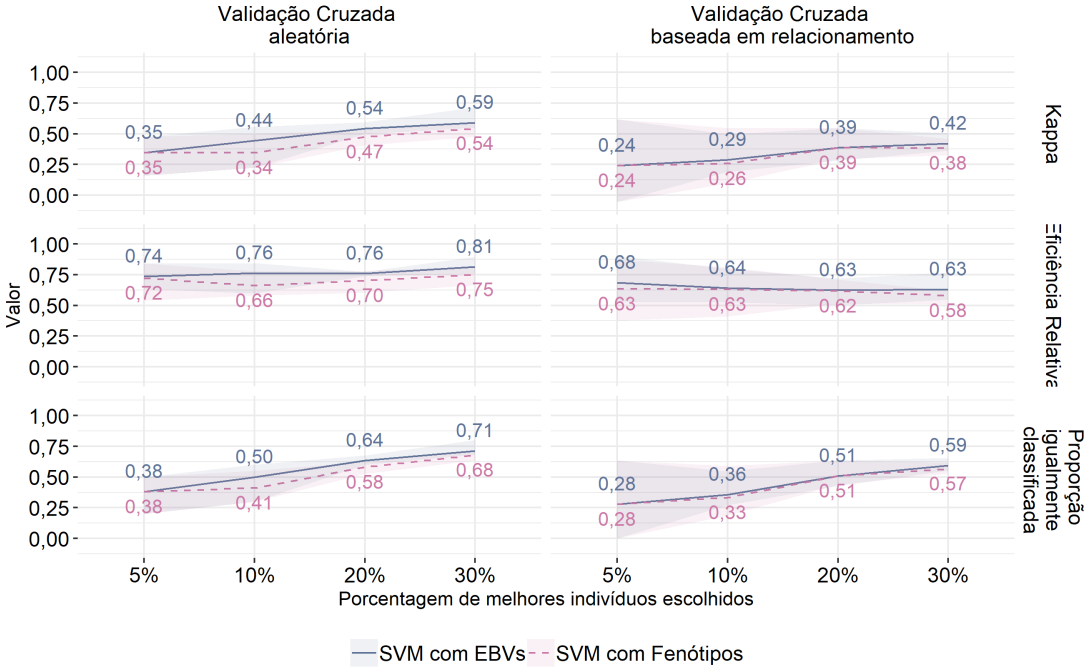


Figura 10: Comparação de médias e amplitudes de medidas obtidas por SVM com EBVs e fenótipos como resposta, com proporção 30-70 e kernel radial para **relação S:G**

Nota-se que as métricas para o fenótipo de crescimento diâmetro à altura do peito (Figura 9) 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 10), 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 11), 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, em ambas as abordagens de validação.

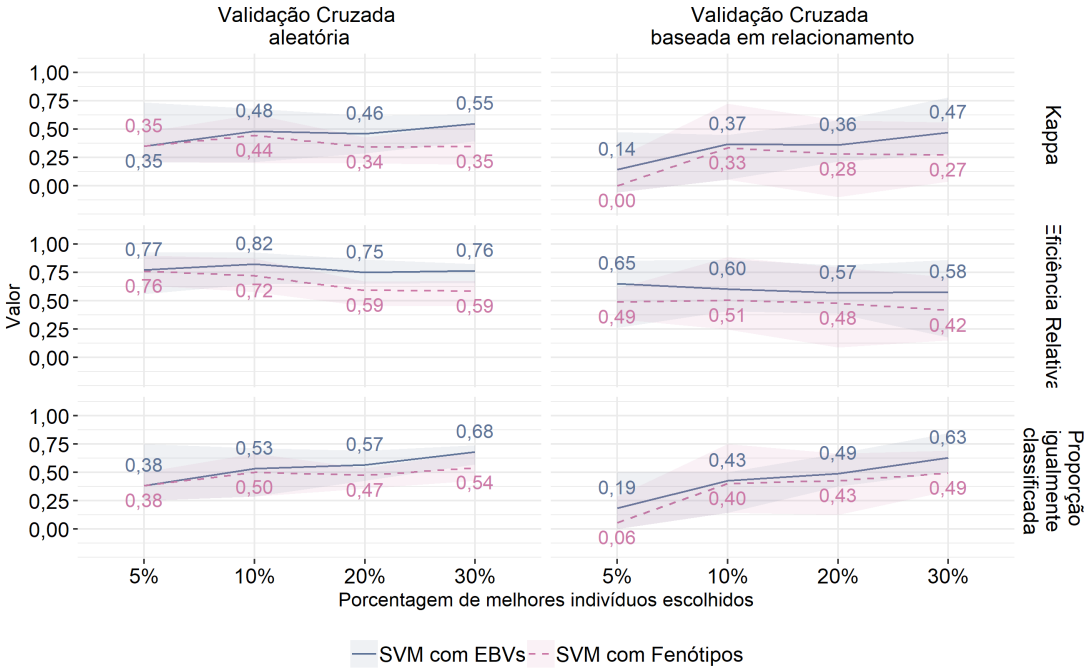


Figura 11: Comparação de médias e amplitudes de medidas obtidas por SVM com EBVs e fenótipos como resposta, com proporção 30-70 e kernel radial para **comprimento da fibra**.

A característica com menor herdabilidade (Figura 12) foi a que apresentou as maiores diferenças entre as duas abordagens de ajuste do SVM, independente da forma de validação cruzada. 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.

3.4 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.

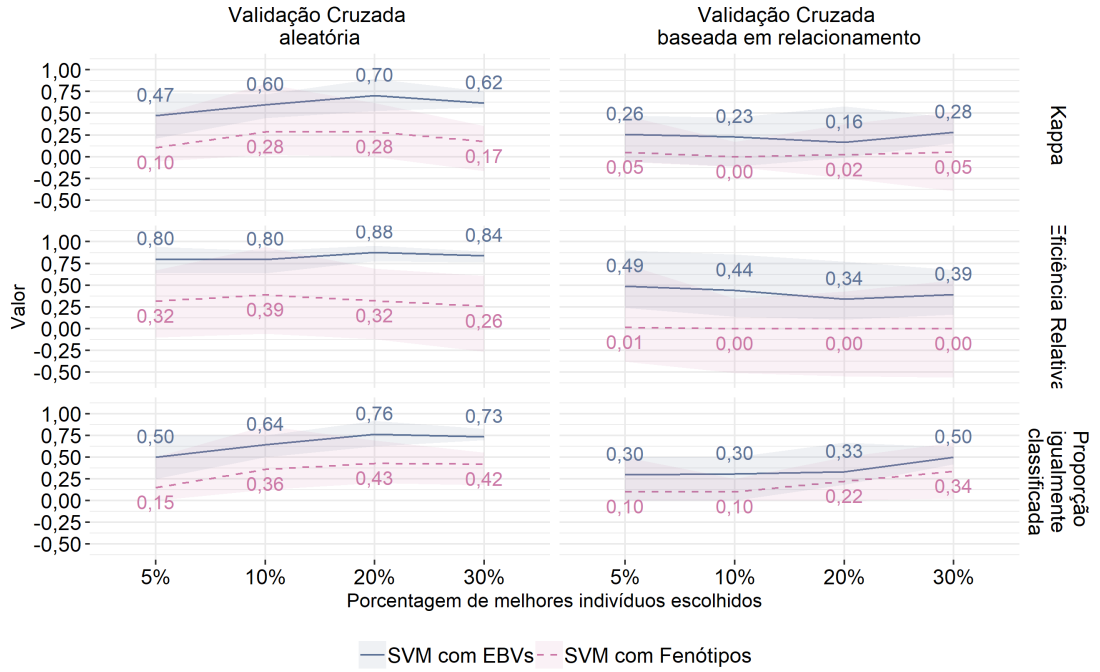


Figura 12: Comparação de médias e amplitudes de medidas obtidas por SVM com EBVs e fenótipos como resposta, com proporção 30-70 e kernel radial para **ângulo microfibrilar**

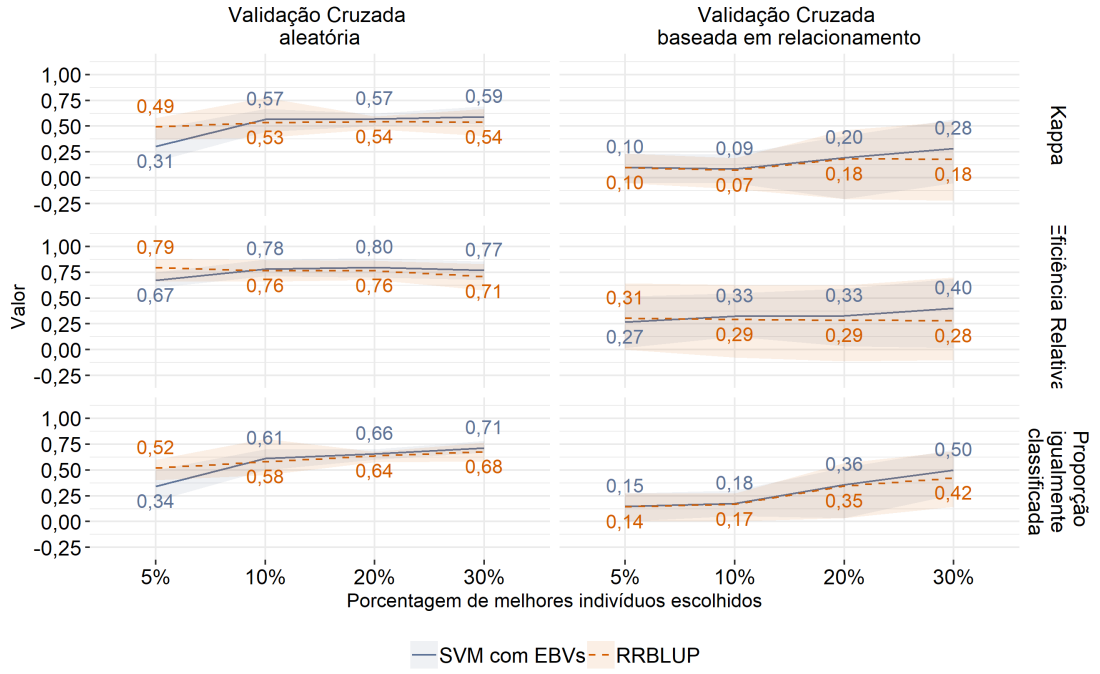


Figura 13: Comparação entre médias e amplitudes de medidas obtidas com RRBLUP e SVM com EBVs para **diâmetro à altura do peito** considerando a proporção 30-70 e kernel radial

Ao observar o fenótipo diâmetro à altura do peito (Figura 13), 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.

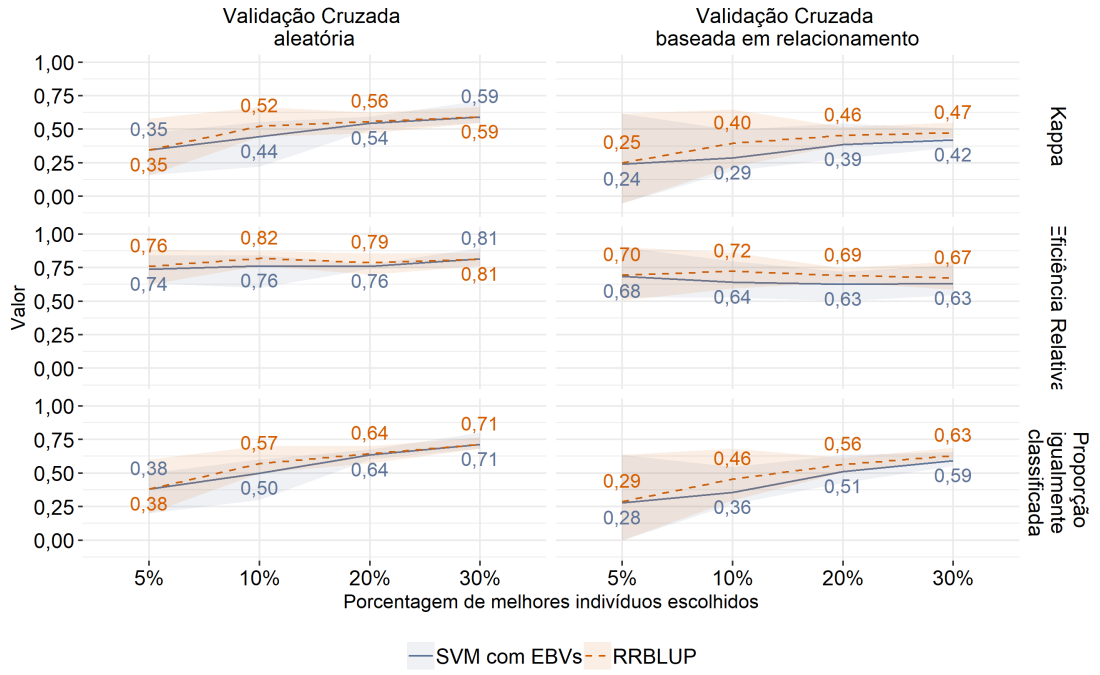


Figura 14: Comparação entre médias e amplitudes de medidas obtidas com RRBLUP e SVM com EBVs para **relação S:G** considerando a proporção 30-70 e kernel radial

Para a relação S:G (Figura 14) as métricas médias obtidas pelo RRBLUP se mostraram ligeiramente superiores em praticamente todos os casos, com uma diferença de no máximo 11 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 15 e 16) 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.

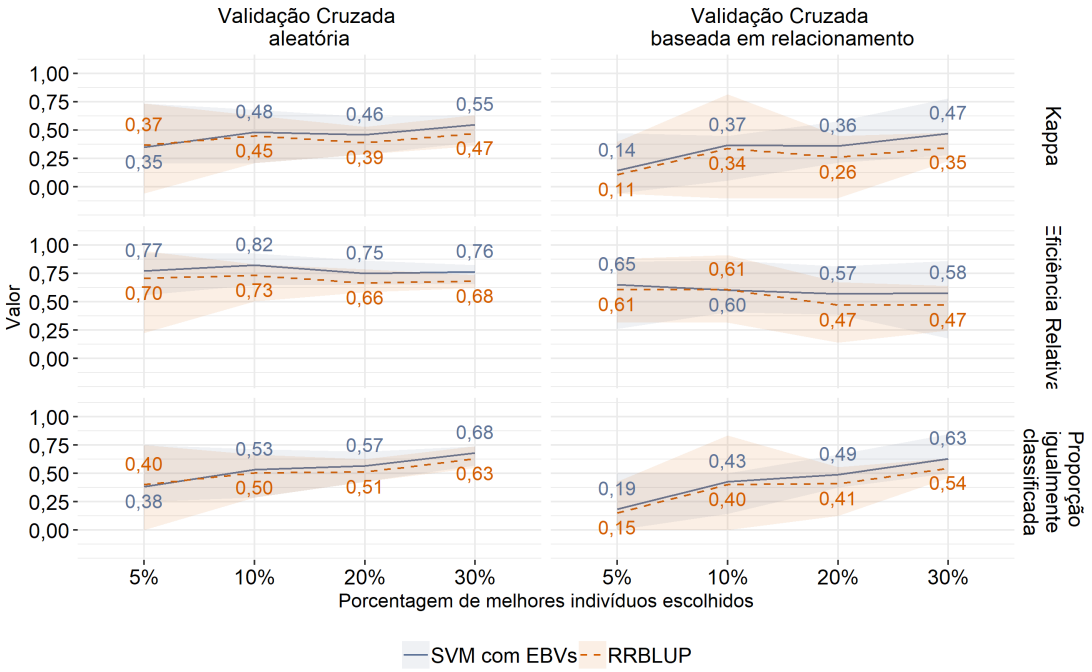


Figura 15: Comparação entre médias e amplitudes de medidas obtidas com RRBLUP e SVM com EBVs para **comprimento da fibra** considerando a proporção 30-70 e kernel radial.

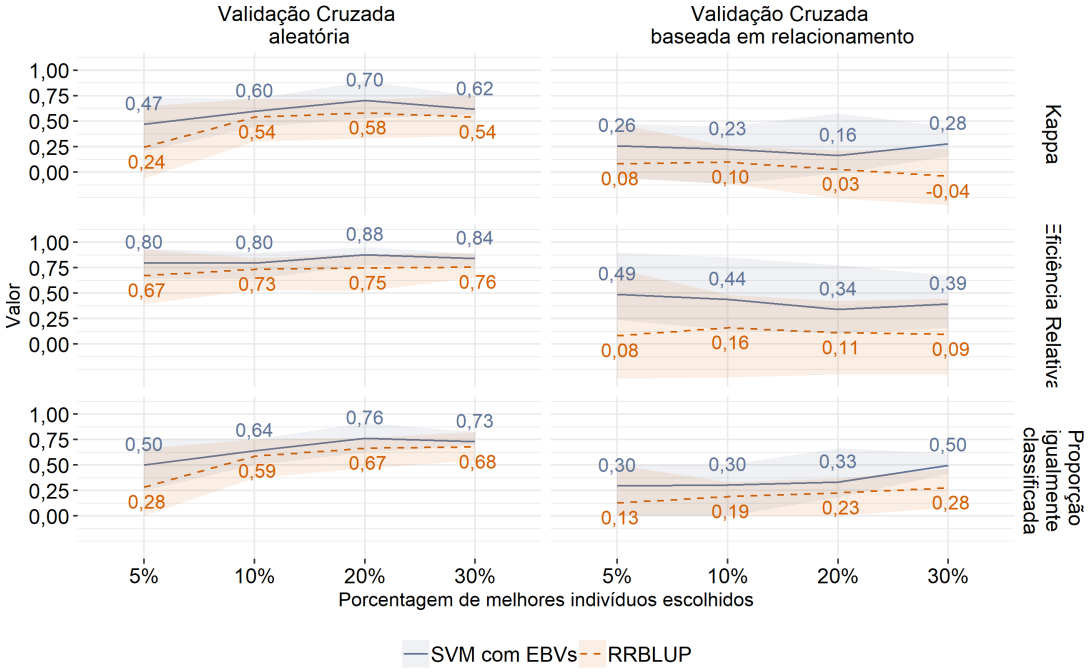


Figura 16: Comparação entre médias e amplitudes de medidas obtidas com RRBLUP e SVM com EBVs para **ângulo microfibrilar** considerando a proporção 30-70 e kernel radial

4 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.

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 se 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. Seria interessante ainda a investigação de outros fatores que possam influenciar modelos de previsão tais como efeitos ambientais, de delineamento e de estrutura de populações. Ou seja, ainda são necessários estudos adicionais, no contexto de Seleção Genômica de plantas de floresta, para a avaliação e aprimoramento de modelos de previsão sob diversos aspectos.

5 Referencias Bibliográficas

BATES, Douglas; VAZQUEZ, Ana Ines. *pedigreemm: Pedigree-based mixed-effects models.* [S.l.], 2014. R package version 0.3-3.

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.

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*, v. 1, p. 651-682, 2014.

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.

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.

MEYER, David; WIEN, FH Technikum. Support vector machines. *R News*, v. 1, n. 3, p. 23-26, 2001.

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.

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.

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, 2013.

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

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.