LUCAS SOUZA DA SILVEIRA

SELEÇÃO GENÔMICA PARA CARACTERÍSTICAS CATEGÓRICAS EM EUCALIPTO

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

VIÇOSA
MINAS GERAIS – BRASIL
2017

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
S587s
2017

Silveira, Lucas Souza da, 1991-
Seleção genômica para características categóricas em eucalipto / Lucas Souza da Silveira. – Viçosa, MG, 2017.
viii, 41f. : il. ; 29 cm.

Inclui anexo.
Orientador: Sebastião Martins Filho.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Eucalipto. 2. Eucalipto - Doenças e pragas. 3. Genômica. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22 ed. 583.766

LUCAS SOUZA DA SILVEIRA

SELEÇÃO GENÔMICA PARA CARACTERÍSTICAS CATEGÓRICAS EM EUCALIPTO

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

APROVADA: 17 de fevereiro de 2017.

Fabyano Fonseca e Silva

Camila Ferreira Azevedo
(Coorientadora)

Sebastião Martins Filho
(Orientador)

Aos meus pais, Antônio Lúcio e Aparecida,

Aos meus Irmãos Luimar e Luciano,

Aos meus sobrinhos Rafael e Daniel,

Aos meus amigos e demais familiares.

AGRADECIMENTOS

A Deus por ser fonte de inspiração e força nos momentos difíceis e por conceder-me boas oportunidades ao longo da vida.

A Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, por proporcionar a realização do curso.

Aos meus pais, Antônio Lúcio e Aparecida, por todo amor e dedicação aos seus filhos e por me apoiarem em todas as decisões.

À minha família, por estarem sempre torcendo por mim.

Aos meus amigos do PPESTBIO, em especial meus companheiros de estudo Maurício, Leísa e Gabriely. Tenho certeza de que sem vocês a caminhada seria mais cansativa e difícil. Obrigado por toda paciência e atenção que tiveram comigo.

A todos os meus amigos de VRB e Viçosa, em especial, Bruno, Denilson, Eduardo e Higor por estarem sempre à disposição e por serem os irmãos que conheci ao longo da vida.

Ao doutor e orientador Sebastião Martins Filho, pela amizade, pelas orientações e por toda paciência e atenção concedida durante o curso. Muito obrigado.

A doutora e coorientadora Camila Ferreira Azevedo, pela amizade, pelas ajudas durante a realização dos trabalhos, pela atenção, paciência e conselhos. Sou muito grato por tudo.

Ao doutor e coorientador Marcos Deon Vilela de Resende, pelos ensinamentos, disponibilidade, atenção e enriquecimento deste trabalho.

Aos membros da banca examinadora, doutora Camila Ferreira Azevedo e doutor Fabyano Fonseca Silva, pela disponibilidade e pelas sugestões que contribuíram com o engrandecer deste trabalho.

Aos professores do Programa de Pós-Graduação em Estatística Aplicada e Biometria pelos ensinamentos e dedicação ao lecionar.

A secretária do Departamento de Estatística, Anita e ao secretário do PPESTBIO, Júnior, obrigado pela amizade e por estarem sempre disponíveis a nos ajudar.

A CAPES, pela concessão da bolsa de estudos.

Enfim, muito obrigado a todos aqueles que de certa forma contribuíram para o meu crescimento pessoal, profissional e para a realização deste trabalho.

BIOGRAFIA

LUCAS SOUZA DA SILVEIRA, filho de Aparecida Marcelino Souza da Silveira e de Antônio Lúcio da Silveira, nasceu em Visconde do Rio Branco, Minas Gerais, em 15 de maio de 1991.

Em março de 2010, ingressou no curso de Licenciatura em Matemática na Universidade Federal de Viçosa, Viçosa-MG, graduando-se em janeiro de 2015. Em março do mesmo ano, iniciou o curso de mestrado no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa de dissertação em 17 de fevereiro de 2017.

SUMÁRIO

RESUMO

SILVEIRA, Lucas Souza, M.Sc., Universidade Federal de Viçosa, fevereiro de 2017. **Seleção genômica para características categóricas em eucalipto.** Orientador: Sebastião Martins Filho. Coorientadores: Marcos Deon Vilela de Resende e Camila Ferreira Azevedo.

Atualmente muitas metodologias têm sido propostas para melhoria da predição de valores genéticos genômicos, no entanto, muitas delas assumem a pressuposição de que as variáveis respostas possuem distribuição gaussiana. Contudo, existem características como resistência a doença, bifurcação em árvores de eucalipto, estágios de florescimento e acamamento em plantas, entre outras, que são classificadas como categóricas, não possuindo distribuição gaussiana para os dados. Diante do exposto, objetivou-se comparar o modelo linear generalizado com o modelo linear de Gauss-Markov, obtendo os valores genéticos genômicos de indivíduos com fenótipos categóricos referentes a resistência à ferrugem do eucalipto, causada pelo patógeno *Puccinia psidii* Winter. Ambos os modelos foram aplicados quando a característica fenotípica possuía quatro classes de infecção (planta imune ou com reação de hipersensibilidade, pequenas pústulas, pústulas medianas e pústulas grandes) e quando estava categorizada como tipos de reação (resistente ou suscetível). O critério de informação da *deviance* (DIC - *Deviance Information Criterion*) foi utilizado para seleção do modelo adequado para descrever a característica fenotípica. O procedimento de validação cruzada via *Jacknife* foi utilizado para validação das estimativas. A acurácia preditiva e o viés foram utilizados para comparação dos modelos. Quando a característica foi categorizada com quatro classes de infecção, os valores de acurácia foram semelhantes para os dois modelos (diferença menor que 0,03). No entanto, quando a categorização foi realizada com duas classes, estas diferenças foram maiores que 0,03 para apenas um dos estimadores de acurácia. O viés na predição de valores genéticos genômicos foi melhor no modelo linear de Gauss-Markov em ambos os tipos de categorização.

ABSTRACT

SILVEIRA, Lucas Souza, M.Sc., Universidade Federal de Viçosa, February, 2017. **Genomic selection to categorical trait in eucalyptus.** Adviser: Sebastião Martins Filho. Co-advisers: Marcos Deon Vilela de Resende and Camila Ferreira Azevedo.

Currently many methodologies have been proposed for improvement on the prediction of genomic breeding values, but many of them assume that the response variables have Gaussian distribution. However, there are trait such as resistance to disease in plants, bifurcation in eucalyptus trees, flowering of plants and others, which are classified as categorical data. In view of the above, the objective was to compare the use of generalized linear model with the linear Gauss-Markov model to obtain genomic breeding values of the categorical phenotypes related to resistance to rust in eucalyptus caused by *Puccinia psidii* Winter pathogen. Both models were applied when the trait had four infection levels (four classes) and when the trait was classified as reaction types (in this case, having two classes). The DIC (Deviation Information Criterion) was used to choose a model, which the effects explained better the variation of the trait. The cross validation procedure via Jacknife was used to validate the estimates of the models. The predictive accuracy and bias were used to compare the models when the evaluated traits had two and four classes. When trait had four classes, the models had similar accuracy values (difference less than 0.03) and when the trait was classified in two classes, the models presented different accuracy values for only one of the accuracy estimators applied. The bias in the prediction of genomic breeding values was better in the linear Gauss-Markov model.

INTRODUÇÃO GERAL

A Seleção Genômica Ampla (GWS – *Genome Wide Selection*) foi proposta por Meuwissen et al. (2001) e consiste na utilização de marcadores moleculares visando a predição de Valores Genéticos Genômicos (GEBV- *Genomic Estimated Breeding Value*) dos indivíduos candidatos à seleção. Existem diversos tipos de marcadores moleculares e os mais usuais na GWS são os marcadores codominantes SNPs (*Single Nucleotide Polymorphisms*) e os marcadores dominantes DArTs (*Diversity Array Technology*). A abundância de dados moleculares tem contribuído com a pesquisa de métodos e modelos estatísticos que auxiliam na melhoria da predição dos GEBVs. Essa predição é de extrema importância para produtores, agricultores e pesquisadores, pois, por meio de dados genotípicos, é possível selecionar precocemente os indivíduos geneticamente superiores. Tais dados são obtidos no início do ciclo de vida, o que diminui custos de produção e tempo na pesquisa de melhoramento genético.

Embora existam muitas metodologias para predição do GEBV, a maioria avalia características fenotípicas contínuas. No entanto, algumas características fenotípicas, como resistência a doença em plantas, acamamento, bifurcações, geralmente, são definidas como variáveis categóricas dicotômicas ou ordinais necessitando a utilização de metodologias que não utilizam a pressuposição de normalidade para a variável resposta (fenótipo).

Fenótipos categóricos vêm sendo relatada há algum tempo em melhoramento genético e uma alternativa para contornar o problema da não normalidade da variável resposta é o modelo *Threshold* ou modelo de limiar (GIANOLA, 1982; GIANOLA e FOULLEY, 1983; FOULLEY et al., 1987). Esse modelo pertence à uma classe dos modelos lineares generalizados mistos (GLMM- *Generalized Linear Model Mixed*) e ainda é pouco utilizado na GWS (BISCARINI et al, 2014; MONTESINOS-LÓPEZ et al., 2015a; 2015b).

Dessa forma, o objetivo deste trabalho foi comparar o Modelo Linear Generalizado Misto com o Modelo Linear de Gauss-Markov (Modelo Gaussiano) utilizando medidas de acurácia e viés na predição dos valores genéticos genômicos.

CAPÍTULO I
REVISÃO DE LITERATURA

1.1. Seleção genômica Ampla

A Seleção Genômica Ampla (GWS) faz uso de dados de marcadores moleculares visando melhorar geneticamente uma população para determinada característica fenotípica. Esse estudo foi idealizado por Meuwissen et al. (2001) e, segundo os autores, a utilização dos marcadores moleculares surgiu com o intuito de obter um ganho genético mais rápido comparado a seleção baseada apenas em dados fenotípicos. Esse ganho se dá pelo fato de que os dados provenientes do DNA podem ser obtidos no início do ciclo de vida de cada espécie, não sendo necessário à espera do desenvolvimento de cada indivíduo para a coleta do valor fenotípico.

Segundo Hayes et al. (2009), a GWS se baseia na seleção de indivíduos por meio dos valores genéticos genômicos estimados (GEBVs - *genomic estimated breeding values*). Esses valores são obtidos via produto do vetor de efeitos dos marcadores (obtidos em uma geração fenotipada e genotipada G_i) pela matriz de incidência dos marcadores na mesma população. Assim, se os dados fenotípicos forem coletados apenas na geração G_i, o fenótipo dos indivíduos de gerações posteriores (G_{i+j}) podem ser estimados de acordo com os efeitos dos marcadores obtidos na geração G_i, bastando multiplicar a matriz de efeitos de marcadores da população G_i pela matriz de incidência dos marcadores na população G_{i+j}. Esse processo pode agilizar o melhoramento genético, bem como reduzir o custo com animais ou plantas que são geneticamente inferiores para a característica fenotípica desejada.

Para a obtenção dos GEBVs faz-se necessário métodos estatísticos para estimação de parâmetros. Como a GWS utilizam um painel denso de marcadores e, em grande parte

dos casos, possuindo o número de marcadores maior que o número de indivíduos, surge o problema da alta dimensionalidade. Também, se existem muitos marcadores, alguns deles podem possuir perfis genotípico iguais, acrescentado o problema de multicolinearidade. Devido a isso, Meuwissen et al. (2001) mencionaram que métodos tradicionais como o de mínimos quadrados não são adequados, necessitando a utilização de outros métodos que obtém o melhor preditor linear não viciado (BLUP – *Best Linear Unbaised Predictor*).

1.2. Modelagem estatística

1.2.1. Modelos Lineares Generalizados

Os Modelos Lineares Generalizados (GLM - *Generalized Linear Models*) foram propostos por Nelder e Wedderburn (1972) como uma alternativa ao modelo clássico de Regressão (ou Gauss-Markov), isto é, quando a distribuição da variável resposta ou dependente (Y) pertence a família exponencial de distribuições. Para que a distribuição de uma variável aleatória Y pertença à família exponencial uniparamétrica, sua função de probabilidade (no caso discreto) ou sua função densidade de probabilidade (no caso contínuo) devem ser escritas da seguinte forma:

$$f_Y(\boldsymbol{y}|\boldsymbol{\theta}) = h(\boldsymbol{y})\exp\{\eta(\boldsymbol{\theta})t(\boldsymbol{y}) - b(\boldsymbol{\theta})\},$$

em que, $h(\boldsymbol{y}), \eta(\boldsymbol{\theta}), t(\boldsymbol{y})$ e $b(\boldsymbol{\theta})$ são funções que assumem valores reais, $\boldsymbol{\theta} = \left(\theta_1,\ \theta_2, \dots, \theta_p\right)'$ e $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)'$ são, respectivamente, o vetor p-dimensional de parâmetros da distribuição de interesse e o vetor n-dimensional de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.).

Conforme Demétrio (2002), o GLM é formado por um trinômio, definido por: i) Componente Aleatório, ii) Componente Sistemático e iii) Função de Ligação. O Componente Aleatório ($\boldsymbol{Y}$) refere-se ao vetor de variáveis aleatórias $\boldsymbol{Y} = (Y_1, Y_2, \dots Y_n)'$,

independentes e identicamente distribuídas por uma distribuição pertencente à família exponencial. O Componente Sistemático ($\boldsymbol{\eta} = x'\boldsymbol{\beta}$) associa linearmente as variáveis explicativas ou covariáveis ($\boldsymbol{X}$) aos seus respectivos parâmetros, que sob o ponto de vista da estatística clássica ou frequentista, são considerados fixos e desconhecidos. A Função de Ligação (g) relaciona a média do Componente Aleatório ao Componente Sistemático.

Resumindo, o GLM estabelece a seguinte relação:

$$\boldsymbol{g}(\boldsymbol{\mu_Y}) = \boldsymbol{\eta} \qquad \textbf{Erro! Fonte de referência não encontrada.}$$

em que $\boldsymbol{\mu_Y}$ representa a média do componente aleatório, $\boldsymbol{g}(\cdot)$ é a função de ligação que estabelece a relação linear entre $\boldsymbol{\mu_Y}$ e o preditor linear $\boldsymbol{\eta}$ (ou componente sistemático) dado por

$$\boldsymbol{\eta} = \boldsymbol{X'\beta},$$

com matriz de delineamento $\boldsymbol{X}$ e vetor de efeitos $\boldsymbol{\beta}$.

Ao desenvolver a equação (1.1**)**, tem-se que:

$$g(\boldsymbol{\mu_Y}) = \boldsymbol{\eta} = \boldsymbol{X'\beta}$$

$$\boldsymbol{\mu_Y} = g^{-1}(\boldsymbol{\eta}) = g^{-1}(\boldsymbol{X'\beta}).$$

Destaca-se que o GLM é linear na função de ligação, no entanto, pode ou não, ser linear no que se refere à média paramétrica do componente aleatório. Por exemplo, o GLM de uma variável aleatória que tem distribuição de Poisson é linear na função de ligação logarítmica (canônica) e não linear na média paramétrica, ou seja,

$$\log(\boldsymbol{\mu_Y}) = \boldsymbol{X'\beta} \Longrightarrow \boldsymbol{\mu_Y} = exp(\boldsymbol{X'\beta}).$$

Já o modelo GLM Normal (ou modelo regressão clássico de Gauss-Markov), é linear na função de ligação identidade e também na média paramétrica:

$$\mathrm{I}(\boldsymbol{\mu_Y}) = \boldsymbol{X'\beta} \Longrightarrow \boldsymbol{\mu_Y} = \boldsymbol{X'\beta}$$

Na prática, muitas variáveis de interesse (fenótipos, fenômenos e outros eventos) não apresentam distribuição normal, e por isso a metodologia dos GLM tem sido empregada.

Uma aplicação dos GLM na seleção genômica pode ser encontrada em Biscarini et al. (2014). Estes autores utilizaram o GLM com função de ligação *Logit* para avaliar a característica vigor de raiz em beterraba açucareira em um painel de 192 SNPs, cuja a variável resposta foi categorizada como "alta" ou "baixa" e codificada em 1 e 0, respectivamente.

Outra aplicação do GLM na seleção genômica é encontrada em Montezinos-Lopez et al. (2015) para tratamento de dados categóricos ordinais em plantas de milho sob um painel de 46.347 SNPs. Os autores utilizaram a função de ligação *Probit* para modelar a probabilidade de cada indivíduo pertencer a uma categoria com base nos preditores lineares. Uma abordagem bayesiana foi utilizada para predição dos parâmetros do modelo. A característica categórica referia-se à uma infecção que deixa as folhas do milho manchadas de cinza (*Gray leaf spot)*, cujos níveis eram classificados em cinco classes, das quais a primeira indicava que a planta não possuía infecção e a última indicava que a planta estava completamente infectada.

1.2.2. Modelos Lineares Generalizados Mistos (GLMM)

Naturalmente, um Modelo Linear Misto possui parâmetros de efeito fixo e aleatório. Analogamente aos modelos que contém apenas efeitos fixos, os GLMM (Bolker et al., 2009) são uma generalização dos Modelos Lineares Mistos, desenvolvidos devido ao fato da variável resposta não possuir distribuição normal, o que viola a pressuposição de normalidade do Modelo Linear Misto.

1.2.3. Modelo de Limiar

Em análise de dados categóricos, um caso particular de GLMM é o modelo de limiar. Este modelo possui um componente aleatório sob distribuição gaussiana em escala não observável, isto é, subjacente à distribuição categórica dos dados (binomial) e utiliza a função de ligação *Probit* para modelar a probabilidade de um determinado indivíduo pertencer a uma categoria com base nos preditores lineares.

O modelo de limiar para análise de dados categóricos no melhoramento genético é descrito com detalhes em Gianola (1982), Gianola e Foulley (1983), Harville e Mee (1984) e Foulley et al. (1987) e, segundo Gianola (1982), características como dificuldade no parto, resistência a doenças e registros de mortalidade, em criações de animais, geralmente são tratados como características categóricas. No melhoramento vegetal encontra-se características como vigor de raiz em beterraba sacarina (Biscarini et al., 2014), resistência a doença em culturas aquáticas (Villanueva et al., 2011), resistência à *gray leaf spot* em milho (Montesinos-Lopez et al., 2015), entre outras.

Gianola (1982) define π_i como a proporção acima do parâmetro de limiar (ou parâmetro *threshold*) t_i, que atua como um separador de categorias, isto é:

$$\pi_i = \lim_{b\to\infty} \int_{t_i}^{b} f_Y(y)\, dy = \left[1 - \lim_{a\to-\infty} \int_{a}^{t_i} f_Y(y)\, dy\right] = 1 - \Phi(t_i) \qquad (0.1)$$

em que $Y \sim N(0,1)$, $\Phi(\cdot)$ é a função de distribuição acumulada da normal padrão e $t_i = \Phi^{-1}(1-\pi_i)$. Dessa forma, a direita desse parâmetro haverá uma massa de probabilidade π_i e a esquerda $1-\pi_i$.

Portanto, o modelo misto para a estimação de efeitos fixos e predição de valores genéticos, conforme Gianola (1982) é dado por:

$$\ell_i = \boldsymbol{x_i'\beta} + \boldsymbol{z_i'u} + a_i + e_i, \qquad (0.2)$$

em que ℓ_i é a variável latente (não observável) que associa-se a escala categórica conforme a posição em relação ao parâmetro de limiar (t_i) ou conforme o valor de

probabilidade obtido pela função de ligação *probit* ($probit(\ell_i)$), $\boldsymbol{\beta}$ é o vetor de efeitos fixos, $\boldsymbol{u}$ é um vetor de efeitos aleatórios, tal que $u \sim N(0, \boldsymbol{D})$ sendo $\boldsymbol{D} = diag(\sigma_i^2), i \in (1,2,\ldots,n)$, $\boldsymbol{x_i'}$ e $\boldsymbol{z_i'}$ são vetores transpostos de incidência para efeitos fixos e aleatórios, a_i é o valor genético aditivo do indivíduo i e e_i é um erro aleatório com distribuição normal, $\mu = 0$ e $\sigma_e^2 = 1$.

Logo, se $t_{j-1} < \ell_i < t_j$ (t_j obtido por meio da função de ligação *probit*), o indivíduo *i* pertencerá à categoria *j*, ou seja, $y_i = j$. O fenótipo do indivíduo *i* sobre a escala normal subjacente é dado pela combinação linear dos parâmetros de efeitos aleatórios.

Os componentes de variância da equação (1.3) podem ser estimados por máxima verossimilhança restrita (HARVILLE, MEE, 1984) por meio de métodos iterativos de forma a resolver um sistema de equações não lineares descrita em Mrode (1996), ou por uma aproximação bayesiana (LEONARD, 1972).

1.2.4. Função de ligação *Probit*

Dentre as funções de ligação reportadas na literatura, uma das mais utilizadas é a *Probit*. Para exemplificá-la, admita que cada uma das i-ésimas variáveis do vetor aleatório $\boldsymbol{\ell}$ de valores em escala gaussiana se distribua de forma independente com variância comum, isto é, $\ell_i \sim N(\mu_i, \sigma^2)$. A função *Probit* é utilizada quando o objetivo de interesse é modelar a probabilidade (π_i) de que a variável aleatória ℓ_i se associe a uma determinada classe, ou de maneira análoga, $\pi_i$refere-se a probabilidade de $\ell_i$ser menor ou igual ao valor estimado para o parâmetro de limiar (t_c), ou seja,

$$\pi_i = P(Y_i = c) = P(\ell_i \leq t_c).$$

Assim, a função de ligação *Probit* relaciona a probabilidade de ocorrência do evento de interesse (π_i) ao preditor linear η_i por meio da função inversa da função de distribuição acumulada $\Phi^{-1}(\cdot)$, isto é:

$$probit(\pi_i) = \Phi^{-1}(\pi_i) = \eta_i \ \ \forall\, i \in \{1,2,\dots,n\},$$

em que n é o número de indivíduos.

1.3. Métodos Bayesianos

Com os avanços computacionais, a inferência Bayesiana vem ganhando espaço na área de melhoramento genético quando o objetivo é a estimação de parâmetros. Na metodologia frequentista, geralmente, se estima parâmetros maximizando a função de verossimilhança. Na metodologia bayesiana, os parâmetros são estimados maximizando a distribuição *a posteriori* do parâmetro ou definindo e minimizando uma função perda (RESENDE et al., 2014).

A distribuição *a posteriori* é definida como

$$P(\boldsymbol{\theta}|\boldsymbol{Y}) = \frac{P(\boldsymbol{Y}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{Y})},$$

em que $P(\boldsymbol{\theta}|\boldsymbol{Y})$ é a distribuição *a posteriori* ou distribuição condicional dos parâmetros dado as variáveis observadas; $P(\boldsymbol{Y}|\boldsymbol{\theta})$ é a função de verossimilhança dos dados, a qual, em termos práticos, retém toda informação à cerca dos parâmetros; $P(\boldsymbol{\theta})$ é a distribuição *a priori* dos parâmetros, que pode ou não ser informativa, ou seja, assumir uma distribuição conhecida; $P(\boldsymbol{Y})$ é a distribuição marginal dos dados, obtida pela integração da distribuição conjunta das variáveis observadas e dos parâmetros $P(\boldsymbol{Y}, \boldsymbol{\theta})$ no espaço paramétrico dos parâmetros Θ, o que torna $P(\boldsymbol{Y})$ independente de $\boldsymbol{\theta}$. Matematicamente, tem-se

$$P(\boldsymbol{Y}) = \int_{\Theta} P(\boldsymbol{Y}, \boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Quando se deseja estimar os parâmetros pela função perda, existem dois tipos de funções perdas mais comuns, a perda linear e perda quadrática. Para minimizar uma função perda linear, é necessário que se obtenha a moda da distribuição *a posteriori* do

parâmetro, já para minimizar uma perda quadrática, é necessário que se encontre a média da distribuição *a posteriori* do parâmetro (RESENDE et al., 2014).

É interessante ressaltar que, em inferência bayesiana não existe parâmetro fixo, apenas aleatórios. Também, ao contrário da frequentista, em que, os estimadores dos parâmetros seguem uma distribuição de probabilidade, na Inferência Bayesiana, os parâmetros que seguem uma distribuição de probabilidade sendo essa uma das distinções fundamentais entre as duas metodologias (RESENDE et al., 2014).

Atualmente, em seleção genômica, pode-se contar com inúmeros métodos de estimação de parâmetros sob enfoque bayesiano. Esses métodos, diferem entre si em algumas propriedades como *shrinkage*, distribuições *a priori* dos parâmetros, entre outras. Alguns desses métodos são apresentados em Gianola (2013).

Dentre os métodos de estimação de parâmetros, alguns comumente utilizados no melhoramento vegetal para predição de GEBV são o BLASSO (*Bayesian Least Absolute Shrinkage and Selection Operator*) e B-RKHS (*Bayesian Reprocing Kernel Hilbert Space*).

1.3.1. BLASSO

O BLASSO (*Bayesian Least Absolute Shrinkage and Selection Operator*) surgiu a partir de um enfoque bayesiano dado a um método frequentista de regressão penalizada proposto por Tibshirani (1996), o LASSO (*Least Absolute Shrinkage and Selection Operator*). Por vez, o LASSO é um método de estimação de parâmetros de um modelo linear o qual objetiva-se minimizar a soma de quadrados dos resíduos com uma restrição que pode conduzir a redução de parâmetros a partir da soma de valores absolutos dos coeficientes (penalização), ou seja

$$\min\{(\tilde{\boldsymbol{y}} - \boldsymbol{X\beta})^T(\tilde{\boldsymbol{y}} - \boldsymbol{X\beta}) + \lambda \sum_j^p |\beta_j|\},$$

em que $\tilde{\boldsymbol{y}}(n \times 1)$ é o vetor de diferenças entre o vetor de valores observados e a média geral de $\hat{\boldsymbol{y}}(n \times 1)$, $\boldsymbol{X}(n \times p)$ é a matriz de incidência dos coeficientes $\boldsymbol{\beta}(p \times 1)$ na população estudada, $\lambda \geq 0$ é parâmetro desconhecido que controla a redução e β_j é o j-ésimo coeficiente de regressão (PARK e CASELLA, 2008).

A ideia do método BLASSO surge com a obtenção da estimativa da média *a posteriori* dos coeficientes de regressão que, conforme Park e Casella (2008), possui a seguinte distribuição *a priori* para os parâmetros da regressão

$$\pi(\beta|\sigma^2) = \prod_{J=1}^{J} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left\{-\lambda|\beta_j|/\sqrt{\sigma^2}\right\}.$$

Este método tem sido implementado em diversos softwares e segundo Perez e de los Campos (2014), o pacote BGLR do software R dispõe da seguinte distribuição *a priori* para os parâmetros da regressão

$$m_j|\sigma_\epsilon^2, \tau_j, \lambda^2 \sim N\left(0, \tau_j^2 \cdot \sigma_\epsilon^2\right); \tau_j \sim Exp\left(\frac{\lambda^2}{2}\right);\ \lambda^2 \sim gama(r, s); \sigma_\epsilon^2 \sim \chi^{-2}\left(S_\beta\right).$$

1.3.2. RKHS Bayesiano

O RKHS (*Reproducing Kernel Hilbert Space*) é um método semiparamétrico de regressão de *kernel* proposto na GWS por Gianola et al. (2006) e surge como uma alternativa para ajuste de modelos que possuem muitas interações epistáticas e de dominância (RESENDE et al., 2014).

Esse método, em seleção genômica ampla, assume como modelo genético

$$y_i = u + g(\boldsymbol{w_i}) + e_i,$$

em que y_i é o valor do fenótipo, u é a média do caráter em estudo, $g(w_i)$ é uma função caracterizada como reprodução de kernel, a qual associa vetores de valores genotípicos a valores fenotípicos e, conforme Gianola et al. (2006),

$$g(w_i) = E(\boldsymbol{Y}|\boldsymbol{w}) = \frac{\int_{-\infty}^{\infty} \boldsymbol{Y} P(\boldsymbol{Y}, \boldsymbol{w})}{P(\boldsymbol{w})},$$

em que $P(\boldsymbol{Y}, \boldsymbol{w})$ e $P(\boldsymbol{w})$¸ segundo Silverman (1986), são estimados utilizando funções de *kernel* com base em distâncias euclidianas. Por fim, e_i é o erro associado ao modelo.

Segue que, para estimação dos parâmetros deste modelo, busca-se uma função de $g(w_i)$ que minimize a soma de quadrados do resíduo mais uma função de penalização,

$$\hat{\beta} = \arg\min \left\{ \sum_j^N [y_i - u - g(w_i)]^2 + h\left|\left|g(w)\right|\right|_H^2 \right\},$$

em que $h\left|\left|g(w)\right|\right|_H^2$ é a função de penalização, h é o parâmetro de suavização que pode ser obtido via validação cruzada ou abordagem bayesiana (controla a redução de dimensionalidade) e $\left|\left|g(w)\right|\right|_H^2$ é a norma de $g(w_i)$ em um espaço de Hilbert, o qual é um espaço provido de produto interno.

Dessa forma, conforme Resende et al. (2014), a solução para essa minimização é dada por

$$g(w) = \alpha_0 + \sum_j^N \alpha_j k(w - w_i),$$

sendo α_j coeficientes desconhecidos e $k(w - w_i)$ é o kernel de reprodução cuja escolha define o espaço de Hilbert em que se dará a minimização da soma de quadrados.

De los Campos et al. (2009), apresentam um enfoque bayesiano para esse método, no qual deu origem a denominação dada aqui de B-RKHS. Assim, escolhe-se uma distribuição normal *a priori* para os parâmetros do modelo

$$P(\boldsymbol{u}_\ell, \boldsymbol{\varepsilon}) \;=\; N\left(\boldsymbol{u}_\ell \middle| \boldsymbol{0}, \boldsymbol{K}_\ell \boldsymbol{\sigma}^2_{\boldsymbol{u}_\ell}\right) \mathrm{N}(\boldsymbol{\varepsilon}|\boldsymbol{0}, \boldsymbol{I}\sigma_\varepsilon^2)$$

com um *kernel* gaussiano com média zero e variância $K\sigma^2_{\boldsymbol{u}}$ em que $K = \{K(x_i, x_i')\}$, $K(x_i, x_i') = \exp\left\{-h\frac{\sum_{k=1}^p (x_{ik} - x_{i'k})^2}{p}\right\}$ e ℓ é o índice que representa o ℓ-ésimo valor de h obtido via método de *multikernel* (PEREZ, DE LOS CAMPOS, 2014). Logo, o *kernel* utilizado tem como base o quadrado da distância euclidiana entre os marcadores.

1.4. Critério de Informação de *Deviance* (DIC)

O critério de informação de *deviance* (DIC- *Deviance Information criterion*) foi sugerido por Spiegelhalter et al. (2002) como uma medida alternativa para comparação de modelagens bayesianas hierárquicas complexas as quais possuem um número de parâmetros maior que o número de observações, não podendo ser diretamente aplicado o critério de informação bayesiano (BIC- *Bayesian Information Criterion*) (GELFAND E DEY, 1994).

O valor de *deviance,* conforme definido em Carlin e Louis (2008), é dado por

$$D(\boldsymbol{\theta}) = -2\log(f(\boldsymbol{Y}|\boldsymbol{\theta})) + 2\log\big(h(\boldsymbol{Y})\big)\,,$$

com $f(\boldsymbol{Y}|\boldsymbol{\theta})$ sendo a função de verossimilhança para o vetor de dados observados $\boldsymbol{Y}$ dado os parâmetros $\boldsymbol{\theta}$, e $h(\boldsymbol{Y})$ é uma função de parametrização dos dados que apenas padroniza, não exercendo influência na escolha de modelos.

Segundo Carlin e Louis (2008), o DIC é definido como

$$DIC = \overline{D} + p_D,$$

em que $\overline{D}$ é a média a posteriori da *deviance* $\overline{D} = E_{\theta|Y}[D]$ e p_D é uma medida de complexidade do modelo capturada pelo número efetivo de parâmetros podendo ser calculada pela diferença entre a média *a posteriori* da *deviance* ($\overline{D}$) e a *deviance* da média *a posteriori* ($D(\overline{\boldsymbol{\theta}})$), isto é, $p_D = E_{\theta|Y}[D] - D\big(E_{\theta|Y}[\boldsymbol{\theta}]\big) = \overline{D} - D(\overline{\boldsymbol{\theta}})$. Dessa forma, o DIC pode ser reescrito como,

$$DIC = \overline{D} + p_D = 2\overline{D} - D(\overline{\boldsymbol{\theta}}).$$

1.5. Validação Cruzada via procedimento *Jacknife*

O procedimento de validação cruzada é constantemente utilizado na seleção genômica. Esse procedimento visa validar as estimativas obtidas pelo modelo estatístico utilizado dividindo a população avaliada em k grupos. Os indivíduos de k-1 grupos são utilizados como população de estimação e um grupo utilizado como população de validação. Na população de estimação, ajusta-se um modelo estatístico afim de obter as estimativas dos efeitos dos parâmetros contidos nele. Esses efeitos, em seguida, são utilizados na população de validação para obtenção dos valores genéticos genômicos (GEBV). Esse procedimento é executado até que cada um desses k grupos se torne a população de validação uma vez.

1.6. Acurácia

A acurácia é uma das principais medidas para comparação de modelos e métodos na seleção genômica ampla. O estimador tradicional de acurácia da seleção genômica proposto por Legarra et al. (2008) é dado por

$$r_{\hat{y}g} = \frac{r_{\hat{y}y}}{\sqrt{h^2}},$$

em que $r_{\hat{y}y}$ é a capacidade preditiva dada pela correlação de y com $\hat{y}$ e h^2 é a herdabilidade da característica. Em outras palavras, o estimador de acurácia é obtido como a razão entre a capacidade preditiva e a raiz quadrada da herdabilidade da característica. O valor desta medida indica o quão preciso é o modelo em estimar o GEBV.

Azevedo et al. (2016) propuseram um estimador de acurácia que também leva em consideração a herdabilidade genômica h_M^2. Esse estimador é dado por

$$r_{g_M g} = r_{\hat{y}y}\sqrt{\frac{h_M^2}{h^2}}.$$

É interessante notar que este estimador é o mesmo proposto por Legarra et al. (2008) multiplicado pela raiz da herdabilidade genômica, ou seja, $r_{g_M g} = r_{\hat{y}g}\sqrt{h_M^2}$.

2. Referências

ABRAF (ASSOCIAÇÃO BRASILEIRA DE PRODUTORES DE FLORESTAS PLANTADAS). **Anuário estatístico da Abraf 2012**. Disponível em: <http://www.abraflor.org.br/estatisticas/ABRAF08-BR.pdf>. Acesso em: 25 nov. 2016.

BISCARINI, F.; STEVANATO, P.; BROCCANELLO, C.; STELLA, A.; SACCOMANI, M. Genome-enabled predictions for binomial traits in sugar beet populations. **BMC genetics**, v. 15, n. 1, p. 87, 2014.

BOLKER, B. M.; BROOKS, M. E.; CLARK, C. J.; GEANGE, S. W.; POULSEN, J. R.; STEVENS, M. H. H.; WHITE, J. S. S. Generalized linear mixed models: a practical guide for ecology and evolution. **Trends in ecology & evolution**, v. 24, n. 3, p. 127-135, 2009.

BRIER, G.W. Verification of forecasts expressed in terms of probability. **Monthly Weather Review**, v. 78, n. 1, p. 1-3, 1950.

CASELLA, G.; BERGER, R. L. **Inferência estatística (2ª ed.)**. São Paulo: Cengage Learning. v.1, 573 p., 2010.

DE LOS CAMPOS, G.; NAYA, H.; GIANOLA, D.; CROSSA, J.; LEGARRA, A.; MANFREDI, E.; WEIGEL, K.; COTES, J. M. Predicting quantitative traits with regression models for dense molecular markers and pedigree. **Genetics**, v. 182, n. 1, p. 375-385, 2009.

FONSECA, S.; RESENDE, M.; ALFENAS, A.; GUIMARÃES, L.; ASSIS, T.; GRATTAPAGLIA, D. **Manual prático de melhoramento genético do eucalipto**. Editora UFV-Universidade Federal de Viçosa, v. 1, 200 p., 2010.

FOULLEY, J. L.; GIANOLA, D.; IM, S. (1987). Genetic evaluation of traits distributed as Poisson-binomial with reference to reproductive characters. **Theoretical and Applied Genetics**, v. 73, n. 6, p. 870-877, 1987.

GIANOLA, D. Priors in whole-genome regression: the Bayesian alphabet returns. **Genetics**, v. 194, n. 3, p. 573-596, 2013.

GIANOLA, D. Theory and analysis of threshold characters. **Journal of Animal Science**, v. 54, n. 5, p. 1079-1096, 1982.

GIANOLA, D.; FOULLEY, J. L. Sire evaluation for ordered categorical data with a threshold model. **Génétique, Sélection, Évolution**, v. 15, n. 2, p. 1, 1983.

GONZÁLEZ-RECIO, O.; FORNI, S. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. **Genetics Selection Evolution**, v. 43, n. 1, p. 1, 2011.

HARVILLE, D. A.; MEE, R. W. A mixed-model procedure for analyzing ordered categorical data. **Biometrics**, p. 393-408, 1984.

HAYES, B. J.; BOWMAN, P. J.; CHAMBERLAIN, A. J.; GODDARD, M. E. Invited review: Genomic selection in dairy cattle: Progress and challenges. **Journal of dairy science**, v. 92, n. 2, p. 433-443, 2009.

MEUWISSEN, T. H.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, n. 4, p. 1819-1829, 2001.

MONTESINOS-LÓPEZ, O. A.; MONTESINOS-LÓPEZ, A.; PÉREZ-RODRÍGUEZ, P.; DE LOS CAMPOS, G.; ESKRIDGE, K.; CROSSA, J. Threshold models for genome-enabled prediction of ordinal categorical traits in plant breeding. **G3: Genes| Genomes| Genetics**, v. 5, n. 2, p. 291-300, 2015.

LEONARD, T. Bayesian methods for binomial data. **Biometrika,** v. 59, n. 3, p. 581-589, 1972.

MRODE, R. A. **Linear models for prediction of animal breeding values**. Wallingford: CAB international, 187 p., 1996.

PARK, T.; CASELLA, G. The bayesian lasso. **Journal of the American Statistical Association**, v. 103, n. 482, p. 681-686, 2008.

PÉREZ, P.; DE LOS CAMPOS, G. Genome-wide regression & prediction with the BGLR statistical package. **Genetics**, v. 198, n. 2, p. 483, 2014.

RESENDE, M.D.V. (org.); SILVA, F.F (org.); AZEVEDO, C.F. (org.). **Estatística Matemática, Biométrica e Computacional: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência (1ª Ed.)**. Visconde do Rio Branco: Suprema, v.1, 881 p., 2014.

R Development Core Team. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2016. Available: <http://www.R-project.org>.

SOUSA, E. D.; SOARES, N. S.; SILVA, M. D.; VALVERDE, S. R. Desempenho do setor florestal para economia brasileira: uma abordagem da matriz insumo-produto. [Performance of the forest section for brazilian economy: an approach of the input-output matrix]. **Revista árvore-Sociedade de Investigações Florestais** (Viçosa-Brasil). v. 34, n. 6, p. 1129-1138, 2010.

SILVERMAN, B. W. Density estimation for statistics and data analysis. **CRC press**, 1986.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso**. Journal of the Royal Statistical Society. Series B (Methodological)**, v. 58, n. 1, p. 267-288, 1996.

VILLANUEVA, B.; FERNANDEZ, J.; GARCIA-CORTES, L. A.; VARONA, L.; DAETWYLER. H. D.; TORO, M. A. Accuracy of genome-wide evaluation for disease resistance in aquaculture breeding programs. **Journal of Animal Science**, V. 89, N. 11, P. 3433-3442, 2011.

WANG, C. L.; DING, X. D.; WANG, J. Y.; LIU, J. F.; FU, W. X.; ZHANG, Z.; YIN, Z.J.; ZHANG, Q. **Bayesian methods for estimating GEBVs of threshold traits**. Heredity, v. 110, n. 3, p. 213-219, 2013.

CAPÍTULO 2

SELEÇÃO GENÔMICA PARA CARACTERÍSTICAS CATEGÓRICAS EM EUCALIPTO

RESUMO:

O cultivo do eucalipto se apresenta como uma cultura de base florestal e ocupa cerca de cinco milhões de hectares do território brasileiro. Essa proporção é responsável por grande parte da produção de celulose do mundo e por tamanho a proporção, problemas como a proliferação de patógenos começam a surgir nas plantas. Atualmente muitas metodologias têm sido propostas para melhoria da predição de valores genéticos genômicos, no entanto, a maioria delas assumem a pressuposição de que as variáveis respostas possuem distribuição gaussiana. Contudo, característica como a resistência à ferrugem em eucaliptos são classificadas por meio de uma escala categórica ordinal ou como tipos de reação, não possuindo distribuição gaussiana para os dados. Diante do exposto, objetivou-se comparar o uso do modelo linear generalizado com o modelo linear de Gauss-Markov para obtenção de valores genéticos genômicos dos fenótipos categóricos referente a resistência à ferrugem em eucalipto causada pelo patógeno *Puccinia psidii* Winter. Ambos os modelos foram aplicados quando a característica possuía quatro níveis de infecção (quatro classes) e quando a característica estava classificada como tipos de reação (neste caso, possuindo duas classes). O critério de informação de *deviance* (DIC - *Deviance Information Criterion*) foi utilizado para escolha de um modelo cujos efeitos explicassem melhor a variação da característica. O procedimento de validação cruzada via *Jacknife* foi utilizado para validação das

estimativas dos modelos. A acurácia preditiva e viés foram utilizados para comparação dos modelos quando a característica avaliada possuía duas e quatro classes. Quando a característica possuiu quatro classes, os modelos possuíram os valores de acurácia iguais (diferença menor que 0,03) e quando a característica estava classificada em duas classes, os modelos apresentaram os valores de acurácia diferentes para apenas um dos estimadores de acurácia utilizados. O viés na predição de valores genéticos genômicos foi melhor no modelo linear de Gauss-Markov em ambos os casos.

Palavras-chave: GLMM, Modelo Gaussiano, Ferrugem do eucalipto, SNP.

ABSTRACT

Genomic Selection to Categorical Trait in Eucalyptus

The Eucalyptus is a forest-based crop and take up about five million hectares of Brazilian territory. This proportion is responsible for the most part of the cellulose production in the world. Due to the large proportion, some problems like proliferation of pathogens begin to emerge in these plants. Currently, many methodologies have been proposed for improvement on the prediction of genomic breeding values, but many of them assume that the response variables have Gaussian distribution. However, trait as resistance to rust in eucalyptus are classified as categorical. In view of the above, the objective was to compare the use of generalized linear model with the linear Gauss-Markov model to obtain genomic breeding values of the categorical phenotypes related to resistance to rust in eucalyptus caused by *Puccinia psidii* Winter pathogen. Both models were applied when the trait had four infection levels (four classes) and when the trait was classified as reaction types (in this case, having two classes). The DIC (Deviation Information Criterion) was used to choose a model, which the effects explained better the variation of the trait. The cross validation procedure via Jacknife was used to validate the estimates of the models. The predictive accuracy and bias were used to compare the models when the evaluated traits had two and four classes. When trait had four classes, the models had similar accuracy values (difference less than 0.03) and when the trait was classified in two classes, the models presented different accuracy values for only one of the accuracy estimators applied. The bias in the prediction of genomic breeding values was better in the linear Gauss-Markov model.

Keywords: GLMM, Gaussian Model, Rust in eucalyptus, SNP.

1. Introdução

A cultura do eucalipto é de grande importância econômica para o Brasil. Segundo Fonseca et al. (2010) existem mais de 600 espécies de Eucalipto cuja capacidade de produção varia de acordo com o clima e solo. Essa cultura, de base florestal, ocupa mais de cinco milhões de hectares do território nacional (ABRAF, 2013) sendo responsável por grande parte da celulose produzida no mundo.

Um dos grandes problemas oriundos de grandes produções são as doenças e pragas que aparecem nas culturas. Para o Eucalipto, uma das principais doenças é a ferrugem, causado pelo patógeno *Puccinia psidii rust* (*Ppr*) muito comum em arbóreas da família *Myrtaceae* e em plantas jovens de eucalipto (Auer et al., 2010). Esse patógeno pode provocar perdas de até 30% do produto final anual de eucalipto, por afetar o crescimento da planta (Furtado, 2009).

Tendo em vista a necessidade de se obter indivíduos geneticamente resistentes à ferrugem, uma opção em programas de melhoramento é a utilização da teoria de seleção genômica ampla (GWS) proposta por Meuwissen et al. (2001). Esta metodologia se baseia na seleção de indivíduos geneticamente superiores via marcadores moleculares (*SNP- Single-Nucleotide Polymorphism*) que fornecem informações originadas diretamente do DNA. Tais informações podem ser obtidas no início do ciclo de vida, o que diminui custos de produção e tempo na pesquisa de melhoramento genético.

Existem várias metodologias para predição de valores genéticos genômicos (*GEBV-Genomic Estimated Breeding Value*) usados na GWS, sendo que a maioria delas lidam com características fenotípicas contínuas. No entanto, algumas características, como resistência a doença, bifurcação, entre outras, geralmente, são medidas como variáveis categóricas, necessitando a utilização de metodologias específicas para o seu

estudo. Neste caso, faz-se necessário o estudo de uma modelagem estatística para predizer os GEBVs com estimativas acuradas e de menor viés.

2. Material e Métodos

2.1. Descrição dos dados

Os dados utilizados na análise são referentes a 37 famílias F_2 de irmão completos de *Eucalyptus urograndis* oriundo do cruzamento de 10 progênies de híbridos *Eucalyptus grandis x Eucalyptus urophylla*, dos quais 559 árvores possuíam a genotipagem e fenotipagem, designando o estudo a esses indivíduos. O experimento foi conduzido em um delineamento de blocos incompletos e aleatorizados possuindo 36 blocos e 2 tipos de solos (cambissolo e latossolo).

O banco de dados genotípicos possui um painel com 24.806 marcadores SNPs dos quais, previamente, passaram pelo controle de qualidade, com *call rate* > 95% e MAF > 1% (MAF- *Minor Allele Frequency*) conforme descrito por Resende et al. (2016). A resistência a ferrugem foi medida de acordo com a escala de notas proposta por Junghans et al. (2003), a qual atribui valor zero para representar imunidade ou reação de hipersensibilidade do tipo "fleck", o valor um representa a classe de indivíduos que apresentavam pequenas pústulas (< 0,8mm de diâmetro), o valor dois para classe de indivíduos que possuem pústulas medianas (diâmetro entre 0,8 e 1,6mm) e o valor três para representar indivíduos que possuem pústulas grandes (> 1,6 mm de diâmetro), conforme apresentado na Figura 1.

Mais informações sobre a fenotipagem, genotipagem, controle de qualidade dos dados podem ser consultadas em Resende et. al. (2016).

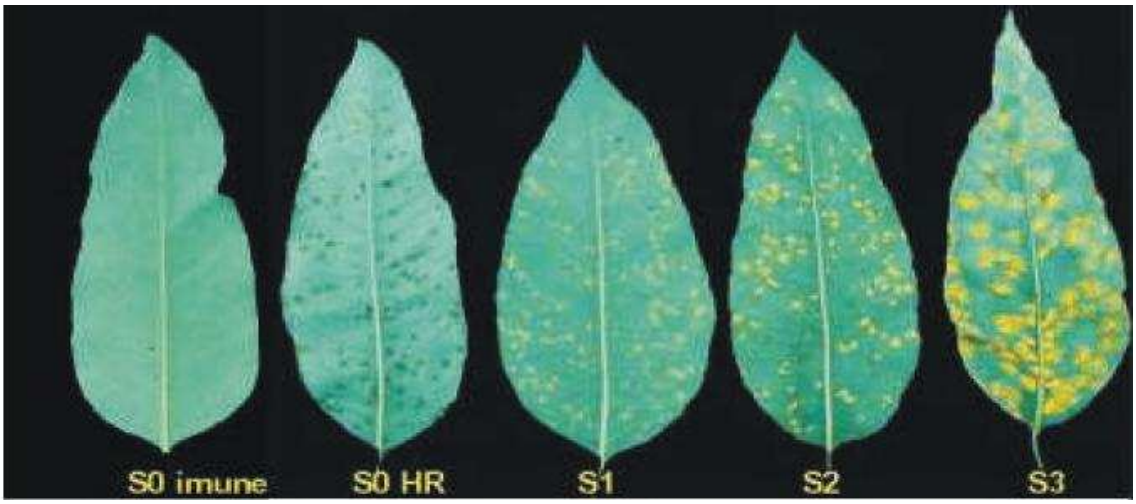

Figura 1. Escala de notas utilizada na avaliação da resistência à ferrugem do eucalipto. Em "S0" tem-se a planta imune (imune) e com reação de hipersensibilidade (HR), para ambas foi designada a nota 0. A nota 1 (S1) quando as folhas apresentavam pequenas pústulas (< 0,8mm de diâmetro) e as notas 2 e 3 (S2 e S3) quando as folhas apresentavam pústulas medianas (diâmetro entre 0,8 e 1,6mm) e pústulas grandes (> 1,6 mm de diâmetro), respectivamente. **Fonte:** Junghans et al. (2003).

2.2. Modelagem

As primeiras modelagens foram feitas sobre as características categóricas ordinais, cujos valores eram classificados em quatro categorias conforme a escala Junghans et al. (2003). Sobre essas respostas, foi ajustado um modelo linear generalizado misto (GLMM – *Generalized Linear Mixed Model*) constituindo na modelagem 1. A modelagem 2 foi executada com as mesmas variáveis respostas sobre o modelo linear de Gauss-Markov (ou modelo gaussiano).

Conforme Junghans et al. (2003), a resistência à ferrugem também pode ser classificada quanto aos tipos de reação (a planta é resistente ou suscetível à doença) de forma que as classes 0 e 1 são consideradas como resistentes e as classes 2 e 3 são consideradas como suscetíveis. Assim, plantas resistentes se encaixaram na classe 0 e plantas suscetíveis na classe 1. Desse forma constituiu-se a modelagem 3, a qual ajustou esses dados dicotômicos sobre o GLMM. A modelagem 4 ajusta os tipos de reação (respostas dicotômicas) sobre o modelo gaussiano.

As modelagens que utilizaram o GLMM possuíram o GEBV predito em escala latente com distribuição gaussiana. Essa distribuição na escala latente é vinculada à variável categórica por meio da função de ligação *probit*. Essa função de ligação modela o valor de probabilidade de uma árvore de eucalipto pertencer a cada uma das categorias de resistência à ferrugem. Dessa forma, realizou-se a categorização considerando a classe de maior probabilidade.

2.3. Seleção de modelos

Devido a necessidade da escolha de efeitos que melhor expliquem a variação genética da população, ajustou-se um conjunto de modelos encaixados em cada uma das quatro modelagens descritas anteriormente, tais modelos são dados por

Modelo 1: $y = X\beta + Z_1 b + Mm_a + e$

Modelo 2: $y = X\beta + Z_1 b + Z_2 f + Mm_a + e$

Modelo 3: $y = X\beta + Z_1 b + Z_2 f + Mm_a + Sm_d + e$

Modelo 4: $y = X\beta + Z_1 b + Z_2 f + Mm_a + Sm_d + Z_3 g_e + e$

Nas modelagens cujas variáveis respostas são ordinais (GLMM), a variável y dos modelos acima é substituída pela variável ℓ indicando que essa resposta é uma variável latente em escala gaussiana cujo valor está ligado a uma variável resposta categórica ordinal.

Nos modelos acima, β é o efeito de solo com matriz de incidência X e $\beta \propto P(\beta) = 1$ (*flat*); b é o efeito de bloco a nível de indivíduo com matriz de incidência $Z_1(559 \times 36)$ e $b \sim N(0, K_b.\sigma_b^2)$; f é o efeito de família a nível de indivíduo com matriz de incidência $Z_2(559 \times x\ 37)$ e $f \sim N(0, K_f.\sigma_f^2)$; m_a é o efeito aditivo dos marcadores com matriz de incidência $M\ (559 \times\ 24806)$ e $m_a \sim N(0, D_a\sigma_{m_a}^2)$; m_d é o efeito devido a dominância dos marcadores com matriz de incidência $S\ (559 \times 24806)$ e

$m_d \sim N(0, D_d \sigma^2_{m_d})$; g_e é o efeito epistático a nível de indivíduos com matriz de incidência Z_3 e $g_e \sim N(0, G_e . \sigma^2_{g_e})$.

As matrizes K_b (559×559) e K_f (559×559) correspondem, respectivamente, a matriz de variâncias e covariância dos efeitos de blocos e famílias e são obtidas pelo produto cruzado das respectivas matrizes de incidência,

$$K_b = Z_1 Z_1' \quad \text{e} \quad K_f = Z_2 Z_2'.$$

Também, as matrizes D_a e D_d corresponde a matrizes diagonais com os componentes de variância dos efeitos aditivos e de dominância, respectivamente. Isto é,

$$D_a = Diag\{\tau^2_{ai}\}_{i=1,2,\ldots,n_{m_a}} \text{ e } D_d = Diag\{\tau^2_{di}\}_{i=1,2,\ldots,n_{m_d}}.$$

A matriz de incidência aditiva dos marcadores M possui elementos 0, 1 e 2 que definem alelos homozigotos recessivos (aa), heterozigotos (Aa) e homozigotos dominantes (AA), respectivamente. Essa matriz foi reparametrizada com a finalidade de se enquadrar na teoria de Genética Quantitativa. Conforme Vitezica et al. (2013) e Azevedo et al. (2015), ela foi obtida a partir das transformações

$$M = \begin{cases} se\ AA, então\ \ 2 - 2p \\ se\ Aa, então\ 1 - 2p \\ se\ aa, então\ 0 - 2p \end{cases}$$

em que p é a frequência alélica de A.

Assim como a matriz de incidência de efeitos aditivos, a matriz de incidência de efeitos de dominância também foi reparametrizada conforme Vitezica et al. (2013) e Azevedo et al. (2015)

$$S = \begin{cases} se\ AA, & então\ \ -2q^2 \\ se\ Aa, & então\ \ \ 2pq \\ se\ aa, & então\ \ -2p^2 \end{cases}$$

em que, p e q são frequências alélicas de A e a, respectivamente.

Para o efeito de epistasia, foi utilizada a matriz de variâncias e covariâncias G_e dada pelo produto de hadamard (multiplicação elemento por elemento de matrizes de mesma ordem cujos elementos estão na mesma posição) da matriz G_a por ela mesma, ou seja, $G_e = G_a \# G_a$(Resende et al., 2014) em que

$$G_a = \frac{MM'}{\sum_{i=1}^{n}[2p_i(1-p_i)]}.$$

As estimativas dos efeitos dos modelos foram preditas por meio dos métodos bayesianos BLASSO e B-RKHS, os quais utilizam um processo iterativo (*Markov chain Monte Carlo* - MCMC) para predição de médias *a posteriori* de cada parâmetro.

A escolha do modelo mais explicativo em cada uma das modelagens foi executada de acordo com o critério de informação de *deviance* (DIC- *Deviance Information criterion*), escolhendo-se o modelo com menor valor de DIC. Tal critério para escolha do melhor modelo foi sugerido por Spiegelhalter et al. 2002 como uma medida alternativa para comparação de modelagens bayesianas hierárquicas complexas as quais possuem um número de parâmetros maior que o número de observações, não podendo ser diretamente aplicado critério de informação bayesiano (BIC- *Bayesian Information Criterion*) (GELFAND E DEY, 1994).

O valor de *deviance*, conforme definido em Carlin e Louis (2008), é dado por

$$D(\boldsymbol{\theta}) = -2\log(f(\boldsymbol{Y}|\boldsymbol{\theta})) + 2\log\big(h(\boldsymbol{Y})\big),$$

com $f(\boldsymbol{Y}|\boldsymbol{\theta})$ sendo a função e verossimilhança para o vetor de dados observados $\boldsymbol{Y}$ dado os parâmetros $\boldsymbol{\theta}$ e $h(\boldsymbol{Y})$ é uma função de parametrização dos dados que apenas padroniza, não exercendo influência na escolha de modelos.

Segundo Carlin e Louis (2008), o DIC é definido como

$$DIC = \overline{D} + p_D \ ,$$

em que $\overline{D}$ é a média a posteriori da *deviance* $\overline{D} = E_{\theta|Y}[D]$ e p_D é uma medida de complexidade do modelo capturada pelo número efetivo de parâmetros podendo ser calculada pela diferença entre a média a posteriori da *deviance* e a *deviance* nas estimativas a posteriori, isto é, $p_D = E_{\theta|Y}[D] - D\left(E_{\theta|Y}[\boldsymbol{\theta}]\right) = \overline{D} - D(\overline{\boldsymbol{\theta}})$. Dessa forma, o DIC pode ser reescrito como,

$$DIC = \overline{D} + p_D = 2\overline{D} - D(\overline{\boldsymbol{\theta}}).$$

2.4. Validação do modelo

O procedimento de validação cruzada consistiu em dividir uma determinada população em k grupos (k = 13). Os indivíduos de k-1 grupos foram utilizados como população de estimação e um grupo utilizado como população de validação. Na população de estimação, ajusta-se um modelo estatístico afim de obter as estimativas dos efeitos dos parâmetros contidos no modelo. Esses efeitos, em seguida, foram utilizados na população de validação para obtenção dos valores genéticos genômicos (GEBV). Esse procedimento é executado até que cada um desses k grupos se torne a população de validação uma vez.

Dessa forma, após escolhido o melhor modelo em cada uma das modelagens conforme o DIC, os modelos foram validados pelo procedimento de validação cruzada via *Jacknife* (13-*fold*). Como haviam 559 indivíduos, a melhor maneira de fazer essa validação de forma a obter grupos com o mesmo número de indivíduos foi dividindo-os em 13 grupos de 43 indivíduos.

Após a validação cruzada os GEBVs foram correlacionados com os valores observados para obter as medidas de capacidade preditiva, acurácia e viés em cada uma das populações de validação.

2.5. Implementação dos Modelos

As estimativas dos parâmetros foram obtidas via análise bayesiana com o auxílio do pacote BGLR (Perez e De Los Campos, 2014) do software R (R Development Core Team, 2016). Nesta análise, utilizou-se 1.050.000 iterações com *burn-in* de 50.000 para retirar o período de aquecimento da cadeia e o *thin* de 10 para eliminar o efeito da autocorrelação. Os parâmetros foram preditos por meio de processo iterativo MCMC e a convergência das cadeias de Markov foram avaliadas por meio do critério de Geweke, o qual indica se a diferença entre a média das n_A primeiras iterações (10% das n iteração) com a média das n_B ultimas iterações (50% das n iterações) segue distribuição normal. A distribuição *a priori* de cada efeito era informativa e fornecia a D.C.C.P. (Distribuição Condicional Completa *a Posteriori*) conhecida, o que implica o uso do algoritmo *Gibbs Sampler*. O efeito de solo foi avaliado com uma distribuição *a priori flat*. Os efeitos de bloco, família e epistasia aditiva (aleatórios) foram preditos pelo método RKHS Bayesiano. Os efeitos de marcadores e dominância aditiva (aleatório) foram preditos pelo método BLASSO.

O método BLASSO, quando avaliados no pacote BGLR do software R, segundo Perez e de los Campos (2014), possui a seguinte distribuição *a priori* para os parâmetros da regressão

$$m_j|\sigma_\epsilon^2, \tau_j, \lambda^2 \sim N\left(0, \tau_j^2 \cdot \sigma_\epsilon^2\right); \tau_j \sim Exp\left(\frac{\lambda^2}{2}\right);\ \lambda^2 \sim gama(r,s); \sigma_\epsilon^2 \sim \chi^{-2}\left(S_\beta\right).$$

O método RKHS Bayesiano mencionado aqui, nada mais é que uma aproximação do método G-BLUP para análise bayesiana, o qual substitui a matriz de *kernel* pela matriz

de correlação no caso dos efeitos de bloco e família e para o efeito de epistasia, a matriz de *kernel* é substituída pela matriz de variâncias e covariâncias G_e. Segundo Perez e de los Campos et al. (2014) este método, quando implementado no pacote BGLR do software R, possui a seguinte distribuição *a priori* para os parâmetros $u|\sigma_u^2 \sim N(0, \boldsymbol{K} \cdot \sigma_u^2)$; $\sigma_u^2 \sim \chi^{-2}(df_u, S_u)$, em que u é o efeito a ser analizado, $\boldsymbol{K}$ é a matriz de kernel que será substituída, σ_u^2 é a variância de u, df_u são os graus de liberdade de u e S_u é a variância amostral de u.

2.6. Herdabilidade

A herdabilidade molecular no sentido amplo, segundo de los Campos et al. (2015), é a proporção da variância fenotípica que pode ser explicada pela regressão linear em um conjunto de marcadores. Assim, esse valor é calculado como $\mathrm{h}_{\mathrm{M}}^2 = \frac{\sigma_{\mathrm{g}}^2}{\sigma_{\mathrm{y}}^2}$, em que σ_g^2 é a variância genética e σ_y^2 é a variância fenotípica.

A variância genética de uma característica em uma população, segundo de los Campos et al. (2015), é a média dos quadrados dos desvios dos valores genéticos para a média de uma população. Também, segundo os mesmos autores, a variância genética aditiva não é apenas uma função das variâncias dos fenótipos e de seus efeitos estimados, mas sim, uma estrutura do desequilíbrio de ligação sobre os genótipos. Sabendo disso, a proporção da variância fenotípica que é explicada pelos efeitos aditivos é conhecida como herdabilidade aditiva h_a^2. Tais valores são calculados como $h_a^2 = \frac{\sigma_a^2}{\sigma_y^2}$, em que $\sigma_a^2 = \sum_i^n \left[2p_i(1-p_i)\sigma_{m_{ai}}^2\right]$ e $\sigma_{m_{ai}}^2$refere-se a variância do efeito aditivo do marcador *i* e σ_y^2 é a variância fenotípica.

A herdabilidade devido a dominância denotada como h_d^2, representa o quanto da variância dos fenótipos é explicada por esse efeito. Esse valor de herdabilidade é

calculado como $h_d^2 = \frac{\sigma_d^2}{\sigma_y^2}$, em que $\sigma_d^2 = \sum_i^n \left[\left(2p_i(1-p_i)\right)^2 \sigma_{m_{di}}^2 \right]$ e $\sigma_{m_{di}}^2$ refere-se a variância do efeito de dominância do marcador *i*.

A herdabilidade devido ao efeito de epistasia é calculado como $h_{ep}^2 = \frac{\sigma_{ep}^2}{\sigma_y^2}$, em que σ_{ep}^2 é a variância desse efeito. Essa medida indica o quanto da variância fenotípica está sendo explicada pelo efeito epistático.

2.7. Acurácia

A acurácia é uma das principais medidas para comparação de modelos e métodos na seleção genômica ampla. O estimador tradicional de acurácia da seleção genômica proposto por Legarra et al. (2008) é dado por

$$r_{\hat{y}g} = \frac{r_{\hat{y}y}}{\sqrt{h^2}},$$

em que $r_{\hat{y}y}$ é a capacidade preditiva dada pela correlação de y com $\hat{y}$ e h^2 é a herdabilidade da característica. Em outras palavras, o estimador de acurácia é obtido como a razão entre a capacidade preditiva e a raiz quadrada da herdabilidade da característica. O valor desta medida indica o quão preciso é o modelo em estimar o GEBV.

Azevedo et al. (2016) propuseram um estimador de acurácia que também leva em consideração a herdabilidade genômica h_M^2. Esse estimador é dado por

$$r_{g_M g} = r_{\hat{y}y} \sqrt{\frac{h_M^2}{h_{trait}^2}}.$$

É interessante notar que este estimador é o mesmo proposto por Legarra et al. (2008) multiplicado pela raiz da herdabilidade genômica, ou seja, $r_{g_M g} = r_{\hat{y}g}\sqrt{h_M^2}$.

2.8. Viés

As estimativas de viés são obtidas a partir do coeficiente de regressão das variáveis respostas observadas em função do GEBV, ou de maneira similar pela equação:

$$b_{y\hat{y}} = \frac{cov(y, \hat{y})}{\sigma^2_{\hat{y}}},$$

em que $b_{y\hat{y}}$ representa o viés e $\sigma^2_{\hat{y}}$ é a variância do GEBV. Segundo Resende et al. (2014) ter uma estimativa não viesada na seleção genômica é importante quando a seleção envolve indivíduos de muitas gerações utilizando os efeitos dos marcadores obtidos em apenas uma geração. Este coeficiente é válido, pois não utilizou testes que exigem normalidade.

3. Resultados e Discussão

3.1. Resultados da seleção de modelos

Na Tabela 1 estão apresentados os resultados do ajuste do modelo em cada uma das quatro modelagens. A escolha do modelo cujos os efeitos explicam a variação dos dados de maneira mais eficiente foi obtido de acordo com o menor valor de DIC.

Em todas as modelagens, o *DIC* do modelo 4 foi menor, indicando que esse modelo é mais adequado para prosseguir com as análises. O valor de p_D do modelo 4 também foi maior que os outros em todas as modelagens, indicando que a complexidade deste modelo, a qual é capturada pelo número efetivo de parâmetros, é maior. Outros trabalhos como Resende et al. (2012) e Lopes et al. (2015) também utilizam os valores de *DIC* para encontrar o modelo mais adequado.

Tabela 1. Resultado do ajuste de modelos contendo os valores de média a posteriori da *Deviance* ($\overline{D}$), *Deviance* avaliado na média a posteriori ($D[\overline{\boldsymbol{\theta}}]$), número efetivo de parâmetros (p_D), Critério de Informação de *Deviance* (DIC).

Modelagem	Modelos	$\overline{D}$	$D[\overline{\boldsymbol{\theta}}]$	p_D	*DIC*
M1	Modelo 1	1157,23	1093,17	64,06	1221,30
M1	Modelo 2	1142,24	1074,27	67,97	1210,22
M1	Modelo 3	1135,53	1063,01	72,53	1208,07
M1	Modelo 4	1038,52	933,05	105,47	1143,99
M2	Modelo 1	1459,49	1377,01	82,47	1541,96
M2	Modelo 2	1434,70	1341,69	93,01	1527,72
M2	Modelo 3	1429,49	1331,33	98,16	1527,66
M2	Modelo 4	1287,12	1097,12	190,00	1477,12
M3	Modelo 1	580,71	535,77	44,94	625,66
M3	Modelo 2	574,28	527,99	46,28	620,57
M3	Modelo 3	571,47	523,16	48,31	619,78
M3	Modelo 4	544,79	483,90	60,89	605,68
M4	Modelo 1	607,27	545,73	61,53	668,81
M4	Modelo 2	587,00	516,14	70,86	657,87
M4	Modelo 3	579,79	502,54	77,25	657,05
M4	Modelo 4	514,08	384,90	129,17	643,26

M1 refere-se a modelagem 1 ajustada sobre o GLMM com a variável resposta contendo 4 categorias; M2 refere-se a modelagem 2 ajustada sobre o GM (*Gaussian Model*) com a variável resposta contendo 4 categorias; M3 refere-se a modelagem 3 ajustada sobre o GLMM com a variável resposta contendo 2 categorias; M4 refere-se a modelagem 4 ajustada sobre o GM com a variável resposta contendo 2 categorias; Modelo 1: $y = X\beta + Z_1 b + Mm_a + e$; Modelo 2: $y = X\beta + Z_1 b + Z_2 f + Mm_a + e$; Modelo 3: $y = X\beta + Z_1 b + Z_2 f + Mm_a + Sm_d + e$; Modelo 4: $y = X\beta + Z_1 b + Z_2 f + Mm_a + Sm_d + Z_3 g_e + e$; $\overline{D} = E_{\theta|Y}[D]; D[\overline{\boldsymbol{\theta}}] = D\left(E_{\theta|Y}[\boldsymbol{\theta}]\right);\ p_D = \overline{D} - D[\overline{\boldsymbol{\theta}}];\ DIC = \overline{D} + p_D$.

3.2. Herdabilidade obtida nas modelagens

Os valores de herdabilidade obtidos no modelo 4 de cada modelagem, estão apresentados na Tabela 2.

Tabela 2. Valores de herdabilidade molecular no sentido amplo (h_M^2), herdabilidade aditiva (h_a^2), herdabilidade devido a dominância (h_d^2) e devido a epistasia (h_{ep}^2).

	h_M^2	h_a^2	h_d^2	h_{ep}^2
Modelagem 1	0.543	0.243	0.001	0.298
Modelagem 2	0.446	0.159	0.002	0.277
Modelagem 3	0.390	0.267	0.004	0.119
Modelagem 4	0.303	0.119	0.013	0.171

Os valores de herdabilidade são obtidos a partir do modelo 4 de cada uma das modelagens. As modelagens 1 e 2 referem-se a modelagem dos dados categóricos 0,1,2,3 sob o GLMM e o GM respectivamente. As modelagens 3 e 4 referem-se a modelagem da codificação 0,1 dada à variável resposta sob os modelos GLMM e GM, respectivamente.

As estimativas de herdabilidade molecular no sentido amplo (h_M^2) foi utilizada no estimador de acurácia proposto por Azevedo et al. (2016) e, neste trabalho, tais estimativas não teve como objetivo a comparação das modelagens, uma vez que, o verdadeiro parâmetro de herdabilidade da população não é conhecido, o que impede a conclusão sobre qual modelo é melhor em predizer os componentes de variância. No entanto, afim de enriquecer o trabalho com algumas informações, é possível perceber que o GLMM produzem estimativas de h_M^2 maiores que as estimativas obtidas nos modelos gaussianos. Essa diferença é de aproximadamente 22% quando a característica possui quatro classes e 30% quando a característica possui duas classes. Tal resultado concorda com Resende et al. (2014) que mostram a relação entre as herdabilidade em escala binomial (para GLMM) e a herdabilidade em escala gaussiana.

Também pode ser verificado na Tabela 2 que o valor de h_a^2 para todas as modelagens foram baixos, comparados aos outros tipos de herdabilidade. Verifica-se ainda, que a herdabilidade aditiva e herdabilidade devido efeito de epistasia explica grande parte da variância fenotípica.

A herdabilidade da característica resistência a ferrugem foi calculada pelo ajuste de um modelo com base no *pedigree* e com os efeitos de bloco e solo, sendo essa medida posteriormente utilizada no cálculo da acurácia dos modelos ($h^2 = \frac{\sigma_g^2}{\sigma_y^2} = \frac{\sigma_g^2}{\sigma_g^2+\sigma_\epsilon^2}$, em que σ_g^2é a variância genética aditiva e σ_ϵ^2é a variância residual). O valor da herdabilidade da característica calculada por meio do GLMM foi de 0,23. Resende et al. (2016) utilizaram o mesmo banco dados, porém sob estudos de GWAS (*genome-wide association study*) e RHM (*Regional heritability mapping*) obteve a herdabilidade da característica resistência à ferrugem de 0,36.

3.3. Comparação entre as modelagens por valores de acurácia e viés

Após fazer a validação cruzada em cada modelagem (sob o modelo 4 selecionado anteriormente), foram obtidas medidas de correlação de Pearson, estimativa de acurácia conforme Legarra et al. (2008), estimativa de acurácia conforme Azevedo et al. (2016) e viés dos modelos. Esses resultados são apresentados na Tabela 3.

Tabela 3. Resultados da capacidade preditiva ($\widehat{r_{\hat{y}y}}$), herdabilidade genômica (h_M^2), acurácia ($\widehat{r_{\hat{y}g}}$) conforme Legarra et al. (2008), acurácia ($\widehat{r_{g_M g}}$) conforme Azevedo et al. (2016) e viés ($\widehat{b_{\hat{y}y}}$) para a avalição da resistência à ferrugem do eucalipto.

	Variável Resposta	Categ[/1]	$\widehat{r_{\hat{y}y}}$[/2]	h_M^2	$\widehat{r_{\hat{y}g}}$	$\widehat{r_{g_M g}}$	$\widehat{b_{\hat{y}y}}$[/2]
Modelagem1 (GLMM)	0,1,2,3	Não	0.33	0.54	0.68	0.50	0.67
		Sim	0.21	0.54	0.45	0.33	0.36
Modelagem2 (GM)	0,1,2,3	Não	0.33	0.446	0.70	0.47	0.94
Modelagem3 (GLMM)	0,1	Não	0.27	0.39	0.56	0.35	0.29
		Sim	0.06	0.39	0.13	0.08	0.11
Modelagem4 (GM)	0,1	Não	0.27	0.303	0.57	0.31	0.98

[/1]Categorização da variável estimada (ou categorização do GEBV); [/2]Valor médio obtido sobre todos os grupos da validação cruzada.

Em cada um dos 13 grupos de população de validação, foi calculado a correlação de Pearson entre o GEBV e o valor observado. Posteriormente, a média dessas correlações foi obtida e utilizada no cálculo das estimativas de acurácia. Essa correlação, na seleção genômica, é conhecida como capacidade preditiva representada por $r_{\hat{y}y}$.

Comparando o GLMM com o GM quanto a capacidade preditiva em cada tipo de codificação da variável resposta (comparação entre os pares de modelagens {1,2}, {3,4}), foi observado que quando o GEBV não era transformado para escala categórica, os valores de capacidade preditiva obtidos no GLMM foram iguais aos do GM. No entanto, quando se categorizava o GEBV (por meio de valores de probabilidades modelados pela

função de ligação *probit*), o valor de capacidade preditiva decrescia. Pela modelagem 3 é possível perceber que quando se reduz o número de categorias, existe um decréscimo nas estimativas de capacidade preditiva, herdabilidade genômica e estimativas de acurácia.

Quando se compara os pares de modelagens {1,2} e {3,4} quanto aos valores de acurácia ($\widehat{r_{\hat{y}g}}$) proposto por Legarra et al. (2008), considerando que diferenças menores ou iguais a 0,03 não são suficientes para dizer que existe diferença entre o GLMM e o GM, pode-se concluir que os modelos possuíram mesmos valores de acurácia. No entanto, se o GEBV das modelagens feitas pelo GLMM forem categorizados, o valor de acurácia deles diminuem (fato explicado pela correlação de Pearson executada sobre duas variáveis categóricas).

Segundo Azevedo et al. (2016), o estimador de acurácia proposto por Legarra et al. (2008), em situações que se tem h^2(herdabilidade da característica) um valor baixo e a capacidade preditiva um valor alto, o estimador fornecerá uma estimativa alta, podendo não pertencer ao espaço paramétrico, além de indicar que a população não possui variabilidade genética. Dessa forma, com a finalidade de corrigir isso, o estimador proposto por Azevedo et al. (2016), leva em consideração a herdabilidade genômica e produz estimativas dentro do espaço paramétrico uma vez que $h_M^2 \leq h^2$, pois, segundo de los Campos (2013) h_M^2 é uma fração de h^2 capturada pelos marcadores.

Ao se comparar o GLMM com o GM pelo valor de acurácia ($\widehat{r_{g_M g}}$) proposto por Azevedo et al. (2016) e considerando que diferenças entre valores de acurácia menores ou iguais a 0.03 não sejam representativas (comparação entre os pares de modelagens {1,2} e {3,4}), apenas a modelagem 3 se mostra mais acurada que a modelagem 4 concordando com Resende et al. (2014), os quais dizem que quanto menos categorias existirem, maior será o ganho em acurácia do GLMM, para o caso de 4 categorias, segundo estes, o GM corresponde a 0.91 do GLMM em acurácia. Acima de 5 categorias

os valores de Acurácia são bem próximos. Ainda, é possível notar que os valores de acurácia proposto por Azevedo et al. (2016) foram menores que os valores de acurácia proposto por Legarra et al. (2008), o que confirma a afirmação feita pelos autores de que o estimador proposto por eles é mais conservador.

Os valores do viés das estimativas em cada modelagem, assim como o valor da capacidade preditiva, foram obtidos em cada grupo da população de validação no procedimento de validação cruzada e posteriormente, obteve-se uma média. O viés de um modelo indica sua tendenciosidade e para que um modelo não seja tendencioso esse valor deve ser igual a 1. Quando o valor de viés é maior que 1, indica que o modelo está subestimando a variável resposta, de forma oposta, quando for menor que 1, tem-se o indício que o modelo está superestimando a variável resposta.

Neste trabalho observa-se que todas as modelagens superestimam a variável resposta. No entanto, nas modelagens que utilizam o GM (modelagens 2 e 4), o viés foi próximo de 1, indicando que, quando respostas categóricas foram avaliadas como respostas gaussiana (ou seja, desconsiderando o caráter categórico da variável), os modelos produziram estimativas menos tendenciosas.

4. Conclusão

Conforme os resultados apresentados, a diferença entre valores de acurácia do modelo linear generalizado misto e do modelo gaussiano aconteceu apenas para um dos estimadores de acurácia quando se tratava de duas classes para a variável resposta. Quando quatro classes foram avaliadas, os valores de acurácia foram semelhantes para os dois tipos de estimadores utilizados. Para qualquer quantidade de classes, o modelo gaussiano produziu estimativas de GEBVs com melhores valores de viés.

5. Referências

ABRAF – Associação Brasileira de Produtores de Florestas plantadas. *Anuário estatístico da ABRAF: ano base 2012*. Brasília: ABRAF, 2013. 142 p. Available at:< http://www.ipef.br/estatisticas/relatorios/anuario-ABRAF13-BR.pdf >

Azevedo, C. F., Resende, M. D. V., Silva, F. F., Viana, J. M. S., Valente, M. S. F., Jr, M. R., & Oliveira, E. J. (2016). New accuracy estimators for genomic selection with application in a cassava (*Manihot esculenta*) breeding program. *Genetics and molecular research: GMR*, *15*(4). DOI: http://dx.doi.org/10.4238/gmr.15048838

Azevedo, C. F., de Resende, M. D. V., e Silva, F. F., Viana, J. M. S., Valente, M. S. F., Resende, M. F. R. e Muñoz, P. (2015). Ridge, Lasso and Bayesian additive-dominance genomic models. *BMC genetics*, 16(1), 1. DOI: 10.1186/s12863-015-0264-2

Auer, C., dos Santos, A. F., & Bora, K. (2010). A ferrugem do eucalipto na região Sul do Brasil. Embrapa Florestas. *Comunicado técnico*. Available at:< http://www.infoteca.cnptia.embrapa.br/bitstream/doc/870855/1/CT252.pdf >

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1-3. DOI: http://dx.doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Casella, G. e Berger, R. L. (2002). *Statistical inference* (Vol. 2). Pacific Grove, CA: Duxbury.

De Los Campos, G., Gianola, D. e Rosa, G. J. M. (2009). Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *Journal of Animal Science*, 87(6), 1883-1887. DOI: 10.2527/jas.2008-1259

De Los Campos, G., Sorensen, D., e Gianola, D. (2015). Genomic heritability: what is it?. *PLoS Genet*, 11(5). DOI: http://dx.doi.org/10.1371/journal.pgen.1005048

Demétrio, C. G. P., (2002). *Modelos lineares generalizados em experimentação agronômica*. (Apostila) Escola Superior de Agricultura Luiz de Queiroz Departamento de Ciências Exatas LCE USP, Piracicaba, SP. Available at: < http://ce.esalq.usp.br/clarice/Apostila.pdf>.

Fonseca, S., Resende, M., Alfenas, A., Guimarães, L., Assis, T. e Grattapaglia, D. (2010). *Manual prático de melhoramento genético do eucalipto*. (pp. 39-42). Editora UFV-Universidade Federal de Viçosa.

Furtado, E.L., Dias, D.C., Ohto, C.T., Rosa, D.D., 2009. *Doenças do Eucalipto no Brasil*, p. 74.

González-Recio, O., D. Gianola, N. Long, K. A. Weigel, G. J. M. Rosa et al., 2008 Non-parametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. *Genetics* 178: 2305–2313. DOI: https://doi.org/10.1534/genetics.107.084293

Junghans, D. T., Alfenas, A. C. e Maffia, L. A. (2003). Escala de notas para quantificação da ferrugem em Eucalyptus. *Fitopatologia Brasileira*, 28(2), 184-188. DOI: http://dx.doi.org/10.1590/S0100-41582003000200012

Legarra, A., Robert-Granié, C., Manfredi, E., & Elsen, J. M. (2008). Performance of genomic selection in mice. *Genetics*, *180*(1), 611-618. DOI: https://doi.org/10.1534/genetics.108.088575.

Lopes, M. S., Bastiaansen, J. W., Janss, L., Knol, E. F., & Bovenhuis, H. (2015). Estimation of additive, dominance, and imprinting genetic variance using genomic data. *G3: Genes| Genomes| Genetics*, 5(12), 2629-2637.

Meuwissen T.H.E., Hayes B.J., Goddard M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics. 157*: 1819-1829. Available at:< http://www.genetics.org/content/157/4/1819 >

Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodríguez, P., De Los Campos, G., Eskridge, K., & Crossa, J. (2015a). Threshold models for genome-enabled prediction of ordinal categorical traits in plant breeding. *G3: Genes| Genomes| Genetics*, *5*(2), 291-300. DOI: 10.1534/g3.114.016188

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Burgueño, J., & Eskridge, K. (2015b). Genomic-enabled prediction of ordinal data with Bayesian logistic ordinal regression. *G3: Genes| Genomes| Genetics*, *5*(10), 2113-2126. DOI: 10.1534/g3.115.021154

Pereira, H. D. (2015). *Predição genômica ampla de híbridos simples de milho considerando modelo não aditivo*. Dissertação (Mestrado em Fitopatologia) – Universidade Federal de Lavras, Lavras, MG. Available at:< http://bdtd.ibict.br/vufind/Record/UFLA_e76767857c2eaa7b480581e61c9ea341 >

Pérez, P.; De Los Campos, G. (2014). Genome-wide regression & prediction with the BGLR statistical package. *Genetics*, genetics-114. DOI: 10.1534/genetics.114.164442

R Development Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, Available at: <http://www.R-project.org>.

Rezende, F. M., Ferraz, J. B. S., Eler, J. P., Silva, R. C. G. D., Mattos, E. C. D., & Ibáñez-Escriche, N. (2012). Study of using marker assisted selection on a beef cattle breeding program by model comparison. *Livestock Science*, 147(1), 40-48.

Resende, M.D.V., Silva, F.F., Azevedo, C.F., (2014). *Estatística Matemática, Biométrica e Computacional: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência*. 1. ed. Visconde do Rio Branco: Suprema, 881 p.

Resende, R. T., Resende, M. D. V., Silva, F. F., Azevedo, C. F., Takahashi, E. K., Silva-Junior, O. B. e Grattapaglia, D. (2016). Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in Eucalyptus. *New Phytologist*. DOI: 10.1111/nph.14266

Vitezica Z. G., Varona L., Legarra A. 2013. On the Additive and Dominant Variance and Covariance of Individuals within the Genomic Selection Scope.*Genetics*, 195(4):1223-30.

ANEXO

```
###   R Script   ###

# Relationhip matrix by pedigree
#================================
library(pedigreemm)
# "pedigree" is the name of the datas, which has pedigree informations
ord <- pedigree[order(pedigree[,1]),]
ped1 <- pedigree(sire=ord[,2],dam=ord[,3],label=ord[,1])
A <- getA(ped1)
A2 <- data.frame(id=ord[,1], as.matrix(A))
A1 = merge(dados[,1:2],A2,by=intersect("id","id"))[,-c(1,2)]
A3 <- data.frame(id=ord[,1], t(as.matrix(A1)))
A4 = merge(dados[,1:2],A3,by=intersect("id","id"))[,-c(1,2)]
A5 <- t(A4)

# Additive incidence matrix
#===========================
gen1<- dados[,-c(1:10)] # "gen1" is the genotype datas
gen1[1:10,1:10]
dim(gen1)
M <- matrix(0,nrow(gen1),ncol(gen1))
#"freq" is the allelic fraquence calculated
for(i in 1:ncol(gen1)){
M[gen1[,i]==2,i] <- 2-2*freq$p[i]
M[gen1[,i]==1,i] <- 1-2*freq$p[i]
M[gen1[,i]==0,i] <- 0-2*freq$p[i]  }

# Dominance incidence matrix
#============================
S <- matrix(0,nrow(gen1),ncol(gen1))
for(i in 1:ncol(gen1)){
S[gen1[,i]==2,i] <- -2*(freq$q[i])^2
S[gen1[,i]==1,i] <- 2*freq$pq[i]
S[gen1[,i]==0,i] <- -2*(freq$p[i])^2     }

# Additive relationship genomic matrix
#======================================
GA <- M%*%t(M)/sum(2*freq$pq)

# Dominance matrix
#==================
GD <- S%*%t(S)/sum((2*freq$pq)^2)

# Epistasys matrix
#=================
library(matrixcalc)
GE=hadamard.prod(GA,GA)

# Incidence matrix for soil, blocks and family effects
#=============================================================
library(matrixcalc)

# incidence matrix and covariance for main eff. of soils.
ZS<-model.matrix(~factor(dados$soil)-1)
KS=tcrossprod(ZS);

# incidence matrix and covariance for main eff. of blocks.
ZB<-model.matrix(~factor(dados$block)-1)
KB=tcrossprod(ZB);
```

```
# incidence matrix and covariance for main eff. of family.
ZF<-model.matrix(~dados$family-1)
KF=tcrossprod(ZF);

# Heritability trait
#===================
# Dados is the datas phenotype and genotype datas where the resistence
to rust is called "PPR".

library(BGLR)

# Model: model with soil effect + block effect(random) + pedigree
effect

ETA3<-list(SOLO=list(X=ZS, model='FIXED'), BLC=list(K=KB,
model='RKHS'),PED=list(K=A5,model='RKHS'))
fm0<-BGLR(y=dados$PPR,response_type='ordinal', ETA=ETA3,
saveAt='FirstModel_', nIter=500000, burnIn=30000,thin = 10)

# Phenotypic variance
varf<- fm0$varE + fm0$ETA$PED$varU + fm0$ETA$BLC$varU

# heritability trait
h2 <- fm0$ETA$PED$varU/varf ;h2

# Ajusting a Generalized Linear Mixed Model
#==========================================
library(BGLR)

# Model 4: y = SNP effect(random) + soil effect(fixed) + block
effect(random) + family effect(random) + dominance effect(Random) +
epistasis effect(random)

set.seed(123)
ETA4<-list(SNP=list(X=M, model='BL'), SOLO=list(X=ZS, model='FIXED'),
BLC=list(K=KB,
model='RKHS'),FML=list(K=KF,model='RKHS'),DMC=list(X=S,model='BL'),EPS
=list(K=GE,model='RKHS'))
fm4<-BGLR(y=dados$PPR,response_type='ordinal', ETA=ETA4, saveAt='M4_',
nIter=200000, burnIn=20000,thin = 10)

# Ajusting a Gaussian model
#============================
library(BGLR)

# Model 4: y = SNP effect(random) + soil effect(fixed) + block
effect(random) + family effect(random) + dominance effect(Random) +
epistasis effect(random)

set.seed(123)
ETA4<-list(SNP=list(X=M, model='BL'), SOLO=list(X=ZS, model='FIXED'),
BLC=list(K=KB,
model='RKHS'),FML=list(K=KF,model='RKHS'),DMC=list(X=S,model='BL'),EPS
=list(K=GE,model='RKHS'))
fm4<-BGLR(y=dados$PPR,response_type='gaussian', ETA=ETA4,
saveAt='M4_', nIter=200000, burnIn=20000,thin = 10)

# Heritabilities
#==============
#additive variance
var.a <- sum(2*freq$pq*fm4$ETA$SNP$tau2*fm4$varE); var.a

#variance due dominance
var.d <- sum(2*freq$pq*fm4$ETA$DMC$tau2)^2; var.d

#variance due epistasys
var.ep<- fm4$ETA$EPS$varU; var.ep
```

```
#genetic variance
var.g = var.a + var.d +var.ep;var.g

#phenotypic variance (or total additive genetic variance)
var.f <- fm4$varE + var.a +var.ep + var.d + fm4$ETA$BLC$varU
+fm4$ETA$FML$varU; var.f

h2   <- var.g/var.f; h2       #genomic heritability
h2.a <- var.a/var.f; h2.a     #heritability due additivity genetic
h2.d <- var.d/var.f; h2.d     #heritability due dominance
h2.ep<- var.ep/var.f; h2.ep   #heritability due epistasys


# Cross Validation
#====================
library(BGLR)

folds<-13
set.seed(123)     #Set seed for the random number generator
sets<-rep(1:13,43) # population 599 indiv divided for ten
sets<-sets[order(runif(nrow(M)))]
system.time(        # checking time
for(fold in 1:folds){
valores <- data.frame()
yNa<- dados$PPR
whichNa<-which(sets==fold)
yNa[whichNa]<-NA
prefix<-paste('PM','_fold_',fold,'_',sep='')
ETA4<-list(SNP=list(X=M, model='BL'), SOLO=list(X=ZS,model='FIXED'),
      BLC=list(K=KB,model='RKHS'),FML=list(K=KF,model='RKHS'),
      DMC=list(X=S,model='BL'),EPS=list(K=GE,model='RKHS'))
set.seed(123)
fm.cv<-BGLR(y=yNa,response_type="ordinal",
                 ETA=ETA4,nIter=200000,burnIn=20000,thin=10)

valores<cbind(dados$id[whichNa],dados$PPR[whichNa],fm.cv$yHat[whichNa]
,fm.cv$probs[whichNa,])colnames(valores)=c("ID","yObs","yHat","0","1",
"2","3")

write.table(valores,paste("valores_",fold,".txt",sep=""),sep=" ",
quote=FALSE,row.names=F)      })
```