GABI NUNES SILVA

PREDIÇÃO DE VALORES GENÉTICOS POR ABORDAGENS DE SELEÇÃO GENÔMICA AMPLA E DE INTELIGÊNCIA COMPUTACIONAL

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

VIÇOSA
MINAS GERAIS - BRASIL
2018

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

T
S586p
2018

Silva, Gabi Nunes, 1986-
Predição de valores genéticos por abordagens de seleção genômica ampla e de inteligência computacional / Gabi Nunes Silva. – Viçosa, MG, 2018.
xii, 108f. : il. (algumas color.) ; 29 cm.

Inclui anexos.
Orientador: Cosme Damião Cruz.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Redes neurais (Computação). 2. Seleção genômica. 3. Biometria. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 006.3

GABI NUNES SILVA

PREDIÇÃO DE VALORES GENÉTICOS POR ABORDAGENS DE SELEÇÃO GENÔMICA AMPLA E DE INTELIGÊNCIA COMPUTACIONAL

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

APROVADA: 01 de fevereiro de 2018.

Leonardo Siqueira Glória

Fabyano Fonseca e Silva

Camila Ferreira Azevedo
(Coorientadora)

Moysés Nascimento
(Coorientador)

Cosme Damião Cruz
(Orientador)

A DEUS,

Aos meus avós, Alfredo e Valdeci (in memoriam),

Aos meus pais, Fátima e Edinilson,

À minha irmã Bianca,

Dedico.

A Lista

Faça uma lista de grandes amigos,
quem você mais via há dez anos atrás...
Quantos você ainda vê todo dia?
Quantos você já não encontra mais?
Faça uma lista dos sonhos que tinha...
Quantos você desistiu de sonhar?
Quantos amores jurados pra sempre...
Quantos você conseguiu preservar?
Onde você ainda se reconhece,
na foto passada ou no espelho de agora?
Hoje é do jeito que achou que seria?
Quantos amigos você jogou fora...
Quantos mistérios que você sondava,
quantos você conseguiu entender?
Quantos defeitos sanados com o tempo,
era o melhor que havia em você?
Quantas mentiras você condenava,
quantas você teve que cometer?
Quantas canções que você não cantava,
hoje assobia pra sobreviver...
Quantos segredos que você guardava,
hoje são bobos ninguém quer saber ...
Quantas pessoas que você amava,
hoje acredita que amam você?
(Oswaldo Montenegro)

AGRADECIMENTOS

Agradeço a Deus por ter me guiado em mais uma jornada e em todas outras pelas quais tenho passado durante minha vida pessoal e acadêmica. O Senhor tem sido minha força e me agraciado com uma família e amigos mais que especiais.

Agradeço aos meus pais, Fátima e Edinilson, e à minha irmã Bianca, por estarem presentes em todos os momentos da minha vida, sejam eles de alegria, tristeza, desafio ou ausência física. Com vocês eu pude aprender o mais sincero e profundo significado das palavras amor e amizade.

À toda minha família, incluindo tios, primos e avós, que sempre me apoiou e se esforçou para estar por perto.

À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, por me proporcionar o privilégio de me tornar mestre e agora doutora em Estatística Aplicada e Biometria em um curso de excelência.

Ao professor Cosme Damião Cruz, que além de orientador sempre foi um grande amigo. Obrigada por ser um exemplo de dedicação, paciência, trabalho em equipe, humildade, bom humor e sabedoria para todos nós.

Ao professor Moysés Nascimento, pela amizade, ensinamentos e conselhos durante o desenvolvimento desse projeto.

Ao professor Matias Kirst por me conceder a oportunidade de cursar um ano do meu doutorado no exterior. Foi de grande valia essa experiência.

Agradeço aos professores dos Departamentos de Estatística e de Biologia Geral pelo saber transmitido, fosse por meio de disciplinas, palestras ou cursos. Essa contribuição foi muito importante para o meu crescimento pessoal e profissional.

Agradeço a todos os meus amigos – que graças a Deus são muitos – pelos momentos de descontração e de estudo, pelas risadas e pelas lágrimas compartilhadas, pelas ligações, festas e viagens. Amigos de curso, amigos de trabalho, amigos do muay thai, amigos de infância, amigos da vida. Agradeço às Serenas lindas por fazerem da nossa república uma grande família. Um agradecimento especial à Lala, nossa amizade começou na graduação e tem uma vida inteira pela frente. Obrigada por tudo!

Agradeço a todos os amigos do laboratório de Bioinformática que fizeram meu ambiente de trabalho se tornar tão especial, divertido e produtivo, com cafés, trabalhos, simpósios, cursos e *happy hours*, e em especial à Isa, obrigada pelas conversas, discussões e amizade que nasceu em meio a projetos acadêmicos em comum e que com certeza perdurará para sempre.

Aos membros da banca examinadora, Prof. Doutor Cosme Damião Cruz, Profª. Doutora Camila Ferreira Azevedo, Prof. Doutor Fabyano Fonseca e Silva, Prof. Doutor Leonardo Siqueira Glória e Prof. Doutor Moyses Nascimento, pela disponibilidade e pelas valiosas sugestões para o enriquecimento deste trabalho.

À CAPES, pelo suporte financeiro para o desenvolvimento deste trabalho.

Enfim, a todos que direta ou indiretamente colaboraram de alguma forma, fosse com um abraço ou com uma palavra de incentivo, para o sucesso deste trabalho.

MEU MUITO OBRIGADA!!!

"O amigo: um ser que a vida não explica
Que só se vai ao ver outro nascer. E o espelho de minha alma multiplica."

Vinicius de Moraes

BIOGRAFIA

GABI NUNES SILVA, filha de Edinilson Santos Silva e Maria de Fátima Nunes da Silva, nascida em 5 de novembro de 1986, em Ipatinga, no estado de Minas Gerais.

Cursou o ensino fundamental e médio na rede pública na cidade de Ipatinga/MG. Em 2005 iniciou o curso de Licenciatura em Matemática, na Universidade Federal de Viçosa – UFV, graduando-se em julho de 2011.

Em fevereiro de 2014 concluiu o Mestrado *Stricto Sensu* no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, Minas Gerais.

Em março de 2014 iniciou o doutorado no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa de tese em 20 de fevereiro de 2018.

SUMÁRIO

RESUMO

SILVA, Gabi Nunes, D.Sc., Universidade Federal de Viçosa, fevereiro de 2018. **Predição de valores genéticos por abordagens de Seleção Genômica Ampla e de Inteligência Computacional**. Orientador: Cosme Damião Cruz. Coorientadores: Camila Ferreira Azevedo e Moysés Nascimento.

Os programas de melhoramento genético existem com dois objetivos principais: identificação de genótipos superiores e a obtenção de combinações melhoradas por meio de cruzamento entre esses indivíduos elite. Os mais diversos ramos da genética, estatística e biometria contribuíram para o estabelecimento de diferentes estratégias de melhoramento para seleção de genótipos superiores. Em particular, metodologias baseadas em seleção genômica ampla tem apresentado grande destaque dentre os estudos mais recentes de seleção. A seleção genômica ampla (*Genome Wide Selection*), envolve estudos biométricos e une genética de populações, genética molecular e a genética quantitativa. A maior motivação para tais estudos consiste na possibilidade de utilizar genotipagem em grande escala e incorporar informações genômicas no processo de predição, de modo a aumentar a eficiência seletiva, obter ganhos genéticos de forma mais ágil e diminuir os custos. Nos modelos de genética, as variações fenotípicas dos indivíduos consistem na variância genotípica dos mesmos agregando variâncias devido a dominância, variância ambiental e também epistasia. No entanto, os modelos de GWS, de modo geral, negligenciam a influência de dominância e epistasia, levando em consideração apenas os efeitos aditivos das características. Além disso, a alta densidade de marcadores moleculares pode levar a problemas de dimensionalidade e multicolinearidade. Neste contexto, o uso de estratégias de redução de dimensionalidade e de metodologias baseadas em inteligência computacional que abordem mais adequadamente a inclusão de tais efeitos em estudos de seleção e predição constituem a proposta neste trabalho. O trabalho visa abordar três tópicos principais: o capítulo 1 propõe avaliar a eficiência do RR-BLUP para predição de valores genéticos de uma população simulada com 12 características complexas que contemplavam efeitos de dominância, epistasia e efeitos ambientais. No capítulo 2 propõe-se a aplicação dos Métodos de Regressão Stepwise e da Sonda para redução de dimensionalidade a fim de aumentar a eficiência preditiva do método RR-BLUP aplicado na mesma população considerada no capítulo 1. Finalmente, o capítulo 3 visa avaliar a eficiência das metodologias de inteligência computacional baseadas em Redes Neurais Artificiais de Redes Perceptron Multicamadas e as Redes de Função de Base Radial para predição dos valores genéticos da população simulada abordada nos capítulos anteriores. Os resultados

indicaram que o uso de metodologias de redução de dimensionalidade contribui para o aumento da eficiência do método RR-BLUP. No entanto, também evidenciaram a deficiência desse método para predizer valores genéticos de populações que incluam efeitos de dominância e epistasia no controle gênico das características de interesse. As metodologias de Redes Neurais Multicamadas e as Redes de função de Base Radial propostas apresentaram acurácia preditiva, expressa pelo erro quadrático médio, superior à apresentada pelo RR-BLUP, demonstrando que as metodologias de inteligência computacional foram mais eficientes que a Seleção Genômica Ampla para o estudo de características complexas com controle gênico envolvendo efeitos aditivos, dominantes e epistáticos.

ABSTRACT

SILVA, Gabi Nunes, D.Sc., Universidade Federal de Viçosa, February, 2018. **Prediction of genetic values by Genome Wide Selection and Computational Intelligence approaches.** Adsiver: Cosme Damião Cruz. Co-advisers: Camila Ferreira Azevedo and Moysés Nascimento.

Genetic breeding programs exist with two main objectives: to identify superior genotypes and to obtain improved combinations through cross-breeding among these elite individuals. The most diverse branches of genetics, statistics and biometry contributed to the establishment of different breeding strategies for selecting superior genotypes. In particular, methodologies based on genomic selection have shown great prominence among the most recent selection studies. Genome Wide Selection, involves biometric studies and gathers genetic of populations, molecular genetics and quantitative genetics. The greatest motivation for such studies is the possibility of using large-scale genotyping and incorporating genomic information into the prediction process, in order to increase selective efficiency, obtain genetic gains and reduce costs. In genetic models, the phenotypic variations of the individuals consist in the genotypic variance including variances due to dominance, environmental variance and also epistasis. However, the GWS models generally neglect the influence of dominance and epistasis, taking into consideration only the additive effects of the characteristics. In addition, the high density of molecular markers can lead to problems of dimensionality and multicollinearity. In this context, the use of dimensionality reduction strategies and methodologies based on computational intelligence that more adequately address the inclusion of effects due to dominance and epistasis in a selection and prediction study are the proposal in this work. The aim of this work is to address three main topics: Chapter 1 proposes to evaluate the efficiency of RR-BLUP for predicting genetic values of a simulated population with 12 complex traits that included effects of dominance, epistasis and environmental effects. In Chapter 2 we propose the application of the Stepwise Regression and Sonda methods to reduce dimensionality in order to increase the predictive efficiency of the RR-BLUP method applied in the same population considered in chapter 1. Finally, chapter 3 aims to evaluate the efficiency of computational intelligence methodologies based on Artificial Neural Networks of Multilayer Perceptron and the Radial Basis Function Neural Networks to predict the genetic values of the simulated population discussed in previous chapters. The results indicated that the use of dimensionality reduction methodologies contribute to increase the efficiency of the RR-BLUP method. However, they also

evidenced the deficiency of this method to predict genetic values for populations that include effects of dominance and epistasis in the gene control of the characteristics of interest. The methodologies of Multilayer Perceptron and Radial Basis Function Neural Networks proposed presented predictive accuracy, expressed by the mean square error, higher than that presented by the RR-BLUP, demonstrating that the computational intelligence methodologies were more efficient than the Genome Wide Selection for the study of complex characteristics with gene control involving additive, dominant and epistatic effects.

INTRODUÇÃO GERAL

Os avanços do melhoramento genético fizeram com que a ciência passasse a desempenhar papel fundamental no desenvolvimento de plantas com grande produtividade agrícola associada à melhoria na qualidade nutricional. Para tanto, diversas estratégias têm sido empregadas a fim de identificar o real potencial genético dos indivíduos ou populações, para assim aprimorá-los. Nass et al. (2001) destacaram que ramos da genética, envolvendo estudos de marcadores moleculares de DNA e princípios de Seleção Genômica Ampla (GWS), são de extrema importância para o melhoramento. O uso de informações de marcadores possibilita agregar informações de DNA na seleção de genótipos superiores, além de proporcionar maiores ganhos genéticos com maior eficácia e menor custo (BORÉM, 1997; RESENDE et al., 2008).

Princípios de seleção genômica têm sido adotados para o desenvolvimento de métodos precisos de predição de fenótipos de plantas e também para inferir sobre seus reguladores (MEUWISSEN et al., 2001; RESENDE et al., 2012; WESTBROOK et al. 2013). No entanto, apesar de eficientes, alguns desafios são rotineiramente enfrentados pelos métodos baseados em GWS. O primeiro deles se refere às pressuposições que devem ser assumidas *a priori*, além disso, o pesquisador deve ser muito cauteloso com relação à dimensão do modelo adotado; possível presença de multicolinearidade entre os marcadores e também com relação à complexidade dos caracteres quantitativos em estudo. Visando solucionar tais limitações e tornar as análises mais viáveis e rápidas, diversos autores têm proposto o uso de metodologias de redução de dimensionalidade da matriz de marcas e demonstrado que métodos de redução propiciam melhora na acurácia preditiva dos modelos (AZEVEDO et al., 2014).

Outro desafio enfrentado pela GWS se refere ao fato de que esta abrange, na maioria das vezes, os modelos aditivos de predição, uma vez que a maioria dos modelos estatísticos adotados estimam somente a variância aditiva associada à matriz de marcas (CRUZ et al., 2013). Possíveis efeitos devido a interações intra e inter-alélicas não são detectados e inflacionam o resíduo associado ao modelo de GWS adotado. Neste contexto, metodologias de redes neurais artificiais como as Redes Perceptron Multicamadas (MLP) e Redes de função de base radial (RBF) constituem novo paradigma que tem sido empregado nos programas de melhoramento genético (GIANOLA et al., 2011; BARROSO et al., 2013; NASCIMENTO et al., 2013; BEAM et al., 2014; SILVA et al., 2014; BHERING et al., 2015; SANT'ANNA et al., 2015; CARNEIRO et al., 2017),

e diferentemente das modelagens estocásticas convencionais utilizadas até então, se baseiam nos princípios de aprendizado e de inteligência computacional (SILVA, 2014).

Diante do exposto, a chave do sucesso dos programas de melhoramento genético consiste no uso de metodologias mais eficazes para a predição do mérito genético dos indivíduos. Assim, esse trabalho de simulação apresenta duas abordagens para a predição de valores genéticos, aliadas a métodos de redução de dimensionalidade (Regressão Stepwise e Sonda). A primeira abordagem apresentada no capítulo 1 foi feita por meio do método RR-BLUP, tradicionalmente aplicado à seleção genômica ampla, de modo a se avaliar a eficiência da metodologia diante do estudo de características de controle gênico complexo dada a inclusão de efeitos de dominância, epistasia e efeito ambiental. O capítulo 2 visou avaliar a influência da dimensionalidade na acurácia do RR-BLUP por meio do uso de duas metodologias de redução de dimensionalidade (Regressão *Stepwise* e Método da Sonda). No terceiro capítulo, esta tese também propõe e avalia a eficiência de outra abordagem, baseada em inteligência computacional, por meio dos métodos RBF e MLP de redes neurais para a predição de características de controle gênico complexo, incluindo efeitos de dominância, epistasia e ambiental.

REVISÃO DE LITERATURA

1. Princípios de genética quantitativa no estudo de populações

O uso do melhoramento genético tem sido um grande aliado dos pesquisadores pois permite a manipulação de caracteres quantitativos de interesse de modo a selecionar e recombinar as formas genéticas mais adaptadas, de melhor qualidade e mais eficientes (CRUZ, 2012). Para tanto, é necessário que estudos acerca dos parâmetros populacionais envolvidos sejam realizados, assim como o controle da forma de acasalamento, da endogamia ou da seleção. Estudos desse tipo caracterizam a genética quantitativa.

A genética quantitativa, área da genética que estuda a herança e a variação dos caracteres quantitativos, constitui uma ciência de grande importância e aplicação para o melhoramento genético de plantas e animais.

Esses caracteres quantitativos, ou de herança complexa, são controlados por um número elevado de genes com forte influência ambiental sobre o valor fenotípico, enquanto os caracteres qualitativos são controlados por um, ou poucos genes, e têm pouca, ou nenhuma, influência do ambiente (FALCONER & MACKAY, 1996). Por essa razão, para o estudo de caracteres de heranças quantitativas, as informações com base somente no indivíduo têm pouco valor, devido à influência do ambiente, o que obriga o pesquisador a estender seus estudos a populações, adotando-se ainda, um modelo biométrico apropriado. Assim sendo, de modo a minimizar os efeitos ambientais de uma característica, conjuntos populacionais devem ser considerados, tomando-se os valores médios e medidas apropriadas de variância da característica de interesse (SILVA, 2014).

Mais especificamente em estudos de populações, o conhecimento da estrutura genética, que envolve parâmetros genéticos e estatísticos, é indispensável ao melhorista. Deste modo, neste trabalho serão utilizados os conceitos de genética quantitativa, a fim de caracterizar a constituição genética dos indivíduos das populações em estudo.

1.1 Modelo para estudo de caracteres quantitativos

Estudos genéticos envolvendo caracteres quantitativos adotam o modelo básico descrito a seguir:

$$F = G + E \quad (1)$$

Em que,

F é o valor fenotípico, medido dos indivíduos;

G é o valor genotípico resultante da ação de cada gene determinante do genótipo do indivíduo;
E é o desvio causado pelo ambiente.

Valor fenotípico consiste no valor observado quando se tem uma característica quantitativa mensurada no indivíduo. Segundo Falconer (1987), observações como média e variância devem ser baseadas nessa medida.

Como visto, o valor fenotípico do indivíduo pode ser decomposto em componentes devido ao genótipo e na fração influenciada pelo ambiente. Em contrapartida, genótipo expressa o conjunto particular de genes do indivíduo e o ambiente caracteriza toda a influência não genética sobre o fenótipo e, por definição, somente estes dois componentes determinam o valor fenotípico do indivíduo. Se o pesquisador ou melhorista pudesse duplicar um particular genótipo em um número de indivíduos e medi-los, sob condições normais de ambiente para a população, o desvio médio ocasionado pelo ambiente, seria zero e seu valor fenotípico médio, por conseguinte, igual ao valor genotípico desse particular genótipo. Esse é o significado do valor genotípico de um indivíduo (FALCONER, 1987).

O valor genotípico pode ser desdobrado em três partes: a fração herdável, também denominada valor genético aditivo, e a fração não herdável por processos sexuados, correspondente aos desvios de dominância devido às interações intra-alélicas e aos efeitos epistáticos devido às interações interalélicas (ALLARD, 1964). A partir destes valores, pode-se estimar a variância genotípica (equação 2).

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 \quad (2)$$

em que σ_A^2 é a variância aditiva; σ_D^2 é a variância atribuída aos desvios de dominância devido às interações intra-alélicas e σ_I^2 é a variância atribuída aos efeitos epistáticos devido às interações interalélicas.

O valor fenotípico total, no modelo aditivo-dominante, expresso por um determinado indivíduo pertencente à população pode ser estimado a partir da expressão (3) e o valor fenotípico total, no modelo epistático, expresso por um determinado indivíduo pertencente à população pode ser estimado a partir da expressão (4).

$$Y_i = \mu + \sum_{j=1}^{n} p_j \alpha_j + E_i \quad (3)$$

$$Y_i = \mu + \sum_{j=1}^{n} p_j \alpha_j + \sum_{j=1}^{n-1} p_j \alpha_j \, \alpha_{j+1} + E_i \quad (4)$$

Em que:

$\alpha_j = a_i + d_i$ e d_i/a_i = gmd (grau médio da dominância), sendo $\mu + a_j$, $\mu + d_j$ e $\mu - a_j$ os valores genotípicos associados as classes AA, Aa e aa, respectivamente que são identificadas pela codificação 1, 0 ou -1, respectivamente, e p_j a contribuição do loco j para a manifestação da característica.

Uma vez mensurados os valores fenotípicos, torna-se possível estimar a variância fenotípica (σ_F^2) e, sob determinadas condições, desdobrá-la nos componentes de variância genética ou genotípica (σ_G^2) e em variância ambiental (σ_E^2). Portanto, tem-se a equação (5) a seguir:

$$\sigma_F^2 = \sigma_G^2 + \sigma_E^2 \quad (5)$$

1.2 Estimativas da média e da variância fenotípica

A média é uma das principais medidas de posição, também chamada de estatística de primeira ordem. A variância, por sua vez, caracteriza uma medida de dispersão das informações acerca dos caracteres quantitativos dos indivíduos, também chamada de estatística de segunda ordem (RESENDE, 2007).

As estimativas da média ($\bar{X}$) e da variância ($\hat{\sigma}^2$) fenotípica são dadas pelas equações (6) e (7) a seguir:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad (6)$$

$$\hat{\sigma}^2 = \frac{\left[\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right]}{n-1} \quad (7)$$

1.3 Estimativa dos parâmetros genéticos

Com as estimativas de média e de variância, torna-se possível a obtenção de parâmetros genéticos, úteis para avaliar o potencial genético do genótipo para a prática do melhoramento para posterior seleção (CRUZ, 2012). Dentre os parâmetros genéticos

mais utilizados em programas de melhoramento, podem-se destacar a herdabilidade (h^2), a correlação (r) e a endogamia.

A herdabilidade consiste em um parâmetro que expressa a proporção da variabilidade observada devido aos efeitos aditivos dos genes, ou seja, representa a proporção herdável da variabilidade total (BORÉM, 2001). Essa fração da variância é expressa pela razão entre a variância genética e a variância fenotípica, de acordo com a equação (8) abaixo:

$$h^2 = \frac{\sigma_G^2}{\sigma_F^2} \tag{8}$$

Toda a avaliação do potencial genético dos indivíduos ou populações é feita com base nos valores fenotípicos, isso porque o valor genotípico é desconhecido. Se ambos esses valores fossem conhecidos, seria possível o cálculo da correlação entre os mesmos, e esta seria dada pela equação (9):

$$r = \frac{cov(F,G)}{\sqrt{\sigma_G^2.\sigma_F^2}} = h \tag{9}$$

Em que:

$cov(F, G)$: é a covariância entre os valores genotípicos e os valores fenotípicos

σ_G^2: é a variância genotípica

σ_F^2: é a variância fenotípica

Ao se admitir que os efeitos do ambiente atuam de forma aleatória e tendo como base a equação $F = G + E$, tem-se:

$$cov(F, G) = \ cov(G + E, G) = cov(G, G) + \ cov(E, G) = \ \sigma_G^2$$

E daí se obtém a equação (10):

$$r = \sqrt{\frac{\sigma_G^2}{\sigma_F^2}}\ e, portanto,\ r^2 = \ h^2 \tag{10}$$

O parâmetro r obtido acima refere-se à correlação entre o valor genotípico verdadeiro do tratamento genético e aquele estimado ou predito a partir das informações dos experimentos. Segundo Henderson (1984), este parâmetro estatístico também conhecido como acurácia seletiva ($\hat{r}_{FG}$), é o parâmetro mais importante e de maior relevância para a avaliação genotípica. Observando-se a equação que determina a herdabilidade, chega-se que se esta for alta, existirá uma alta correlação entre o valor fenotípico e o valor genotípico, ou seja, a acurácia na predição será alta (CRUZ, 2012).

Outro parâmetro genético de grande importância para o estudo de populações é a endogamia. A endogamia é um fenômeno ocasionado pelo acasalamento entre indivíduos aparentados, e pode ter consequências sobre a média da população, afetando a similaridade das linhas derivadas. Este parâmetro é determinado por meio do coeficiente de endogamia (F). Esse coeficiente refere-se à probabilidade de que os alelos de um loco de um indivíduo sejam idênticos por ascendência. Esses alelos são idênticos quando derivam ou são cópias de um alelo comum, encontrado nos ancestrais daquele indivíduo (CRUZ et al., 2012).

Numa população, podem-se encontrar homozigotos com alelos idênticos por ascendência ou idênticos apenas pela função que exercem. Assim, para um indivíduo I de genótipo A_pA_m, define-se o coeficiente de endogamia por meio de (11):

$$F = P(A_p \equiv A_m) \tag{11}$$

2. A genômica no melhoramento genético

A necessidade de se obter maior ganho genético com menor intervalo de tempo levou os pesquisadores a formularem diferentes estratégias de melhoramento de forma que a observação fenotípica pudesse ser mais acurada. Nas metodologias tradicionais, que se baseiam nos princípios de genética quantitativa, a seleção é praticada fundamentalmente com base na identificação de indivíduos superiores, testes e recombinações para gerar novas populações de desempenho superior. No entanto, quando os estudos envolvem o controle de características complexas, esses métodos tradicionais se tornam por diversas vezes ineficazes. Outra problemática envolvida está ligada ao fato de que com o desenvolvimento de populações superiores, fica cada vez mais difícil identificar novos genótipos ou indivíduos ainda mais superiores (CRUZ et al., 2013).

Nesse sentido, aliar o uso da biotecnologia às metodologias disponíveis tem-se mostrado uma estratégia bastante eficaz. Dentre as ferramentas de biotecnologia no

melhoramento genético destaca-se o uso de marcadores moleculares. Marcadores moleculares são sequências de DNA que revelam polimorfismos entre indivíduos geneticamente relacionados, e a identificação de marcadores ligados a genes controladores de características oligogênicas e, ou, quantitativas tem sido de grande valia para os programas de melhoramento genético.

Características como maior produtividade agrícola associada à melhoria na qualidade nutricional das populações, resistência e melhor adaptação à ambientes desfavoráveis representam pontos decisivos nos programas de melhoramento (SILVA, 2014), por isso, uma ferramenta disponível e cada vez mais útil aos melhoristas, é a genômica. Genômica é a ciência que estuda o genoma aplicado a sistemas biológicos. Cruz et al. (2013) afirmam que ao trabalhar com genômica, o pesquisador na verdade lança mão de três grandes áreas: Genômica Clássica, pois seus estudos envolvem marcadores genéticos e mapeamentos; Bioinformática, com o uso de extensos bancos de dados e finalmente Análise de sequências de DNA.

3. A seleção genômica ampla (GWS) no melhoramento genético

Outra técnica eficaz para seleção de indivíduos com maior desempenho, usando a informação de marcadores moleculares é a seleção genômica ampla (*Genome Wide Selection - GWS*). Tal metodologia, proposta por Meuwissen et al. (2001), permite incorporar informações moleculares diretamente na predição do mérito genético dos indivíduos. Essa nova metodologia de seleção conseguiu se desenvolver graças aos amplos investimentos em genotipagem em grande escala e com o desenvolvimento dos marcadores tipo SNP (*Single Nucleotide Polymorphism*).

O maior atrativo da GWS em benefício do melhoramento genético consiste na possibilidade de utilização direta da informação de DNA na seleção, podendo ser em todas as famílias em avaliação e apresentando alta acurácia seletiva, além de não exigir prévio conhecimento dos mapas associados à seleção, e ainda, a GWS permite a predição de valores genômicos e é excelente para características de baixa herdabilidade (RESENDE et al., 2014).

Existem diversas metodologias de GWS utilizadas nas áreas do melhoramento. Métodos baseados em modelos mistos (HENDERSON, 1973), como RR-BLUP (*Ridge Regression-Best Linear Unbiased Prediction*) e LASSO (*Least Absolute Shrinkage and Selection Operator*); métodos bayesianos, como BayesA, BayesB e BGLR (*Bayesian*

Generalized Linear Regression), dentre outros. No decorrer desse trabalho abordaremos o RR-BLUP.

3.1 RR-BLUP (*Ridge Regression-Best Linear Unbiased Prediction*)

O método RR-BLUP consiste em um método de regressão aleatória ou regressão de cumeeira que estima simultaneamente os efeitos de todos os marcadores. Para tanto, utilizam-se os denominados Modelos Mistos (HENDERSON, 1973), com estimação e predição realizadas via BLUP (*Best Linear Unbiased Prediction*) (ROBINSON, 1991). Basicamente, no RR-BLUP os marcadores são considerados variáveis regressoras e como efeitos aleatórios no modelo (MEUWISSEN et al., 2001). Nessa metodologia, as predições lineares assumem que todas as marcas que possuem mesma frequência alélica contribuem igualmente para a variação genética.

A fim de resolver o paradoxo de ter o número de marcadores (n) muito maior que o número de indivíduos genotipados (N), o RR-BLUP utiliza em seu procedimento um método de relugarização ou *shrinkage* (CRUZ et al., 2013). Os estimadores associados ao modelo RR-BLUP produzem o denominado *shrinkage* λ (parâmetro de penalização), definido por:

$$\lambda = {\sigma_e^2}\Big/{\sigma_{gi}^2} = {\sigma_e^2}\Big/{(\sigma_g^2/n_Q)} \tag{12}$$

Em que:

σ_{gi}^2: variância genética aditiva do i-ésimo loco;

σ_g^2: variância genética aditiva do caráter

σ_e^2: variância residual;

n_Q: número de segmentos cromossômicos de QTL, obtido por meio da seguinte expressão: $n_Q = 2\sum_i^n p_i(1-p_i)$.

Obtidos os marcadores – que podem ser do tipo SNP, microssatélites ou DArT –, os efeitos são estimados com base nas informações fenotípicas de uma população conhecida. Assume-se um modelo linear misto geral, conforme Resende et al. (2008), tal como descrito a seguir na Equação 13 abaixo:

$$y = Wb + Xm + e \tag{13}$$

Matricialmente, as equações de predição do método RR-BLUP podem ser reescritas como se segue (RESENDE et al., 2014):

$$\begin{bmatrix} W'W & W'X \\ X'W & X'X + I\frac{\sigma_e^2}{(\sigma_g^2/n_Q)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} W'y \\ X'y \end{bmatrix} \quad (14)$$

Em que y: vetor de observações fenotípicas; *b*: vetor de efeitos fixos (média geral) com matriz de incidência W; *m*: vetor dos efeitos aleatórios dos marcadores com matriz de incidência X e *X* é matriz de incidência composta pelos valores -1, 0 e 1 para o número de alelos do marcador dos genótipos mm, Mm e MM, respectivamente, com $m \sim N(0, I\sigma_g^2)$; *e*: refere-se ao vetor de resíduos aleatórios com $e \sim N(0, I\sigma_e^2)$ com σ_e^2 variância do erro; σ_g^2: variância genética e $n_Q = \sum_{j=1}^{j} 2p_j(1-p_j)$ e p_j é a frequência alélica do marcador j.

4. Inteligência artificial no melhoramento genético

Os estudos envolvendo a inteligência computacional se iniciaram a partir do momento em que as comunidades científicas e pesquisadores reconheceram que a maneira que o cérebro humano processa as informações e identifica padrões de aprendizado difere, em muitos aspectos, da maneira como os computadores convencionais o fazem (SILVA, et al., 2010). A inteligência computacional consiste em uma área da ciência da computação que surgiu para fins de simulação – por meio de máquinas – da capacidade humana de solucionar problemas e realizar tarefas que envolvam generalização e aprendizado (NORVIG & RUSSELL, 2013).

Segundo Bishop (2007), o uso de inteligência computacional é importante porque permite a captação de relações não lineares entre variáveis preditoras e respostas, e ainda podem aprender sobre formas funcionais de forma adaptativa, pois uma série de transformações, chamada genericamente de função de ativação, é conduzida por parâmetros. Essa não linearidade e grande capacidade de generalização fez com que metodologias baseadas em inteligência computacional – como técnicas de Aprendizado de Máquina, Redes Neurais Artificiais e Lógica *Fuzzy* – se consolidassem em estudos da ciência da computação, planejamento autônomo, jogos, resolução de problemas, dentre outros (FERNANDES, 2003).

O grande sucesso obtido ao utilizar técnicas de inteligência computacional, aliado à busca constante por metodologias mais eficientes que possam ser aplicadas nos programas de melhoramento genético, tem levado diversos pesquisadores a aplicar, com sucesso, as metodologias de inteligência computacional baseadas em Redes Neurais Artificiais como as Redes Perceptron Multicamadas (MLP) e Redes de Base Radial (RBF) em diversos problemas ligados a genética (CARNEIRO et al., 2017; SILVA et al., 2017; SANT'ANNA et al., 2015; SILVA et al., 2014; BARROSO et al., 2013; NASCIMENTO et al., 2013; GONZÁLEZ-CAMACHO et al., 2012; VENTURA et al., 2012; GIANOLA et al., 2011).

Buscando estender a aplicação de metodologias de inteligência computacional a estudos que envolvam predição e seleção genômica, neste trabalho abordaremos a metodologia de Redes Neurais Artificiais com enfoque nas Redes Perceptron Multicamadas (MPL) e nas Redes de funções de base radial (RBF).

4.1 Redes Neurais Artificiais (RNA)

As Redes Neurais Artificiais (RNA) consistem em modelos de processamento de dados que emulam uma rede de neurônios biológicos, capazes de recuperar rapidamente uma grande quantidade de dados e reconhecer padrões baseados na experiência, ou seja, tentam reproduzir as funções das redes biológicas, buscando implementar seu comportamento funcional e sua dinâmica (HAYKIN, 2001).

No modelo neural artificial, o desempenho está diretamente ligado às conexões entre os elementos que o compõe. Este modelo é composto por três elementos principais: um conjunto de sinapses, um somatório e uma função de ativação (HAYKIN, 2001), e o processamento de uma RNA envolve três etapas primordiais: treinamento, aprendizado e validação, aliadas à escolha de uma arquitetura apropriada que possua funções de ativação eficientes, número de camadas ocultas e número de neurônios por camadas satisfatórios (SILVA et al., 2014).

Basicamente, um modelo de neurônio artificial é uma simplificação de um neurônio biológico. Desse modo, para representar o comportamento das sinapses, os m terminais de entrada do neurônio artificial possuem pesos w_1, w_2, ..., w_m, cujos valores podem ser positivos ou negativos, de acordo com o sinal sináptico correspondente – inibitório ou excitatório (SILVA et al., 2010). O neurônio artificial é constituído por *m* entradas com os pesos ponderados, um limiar de disparo conhecido como *bias* associado à função de ativação, e finalmente, um terminal de saída *y*. A função de ativação é

responsável por gerar a saída y do neurônio a partir das somas ponderadas recebidas pelo neurônio, e é escolhida de acordo com o problema a ser resolvido (SILVA, 2014).

De modo geral, podem-se considerar duas classes principais de arquiteturas de rede: redes alimentadas por uma única camada intermediária de neurônios e redes alimentadas diretamente com múltiplas camadas (HAYKIN, 2001). Dentre as classes de redes perceptron alimentadas com múltiplas camadas se destacam as redes neurais multicamadas.

4.1.1 Redes Perceptron Multicamadas (MLP)

A Rede Perceptron Multicamadas ou *Multilayer Perceptron* (MLP) consiste numa extensão do perceptron simples (HAYKIN, 2001) e pode ser limitada por funções de ativação parcialmente ou totalmente diferenciáveis (SILVA et al., 2010).

A arquitetura de rede utilizada – escolha do algoritmo de treinamento e das funções de ativação, tal como o estabelecimento do número de camadas ocultas e número de neurônios que se deve considerar em cada camada – é responsabilidade do pesquisador. A decisão deve ser tomada com base no tipo de problema avaliado e de acordo com a complexidade do mesmo.

Tendo a arquitetura de rede e as funções de ativação bem definidas, o procedimento primordial para o bom funcionamento de uma rede neural artificial consiste no aprendizado. A capacidade de aprender a partir do ambiente e de melhorar o seu desempenho caracterizam a propriedade mais importante da RNA (HAYKIN, 2001). Portanto, o algoritmo de aprendizado adotado é fator determinante para treinar a rede. A escolha da arquitetura, pesos, número de neurônios e método de aprendizado para treinamento é feita de modo empírico e leva em conta a complexidade do modelo em estudo (SILVA et al., 2010).

O processo de treinamento da rede MLP abordada nesse estudo é realizado por meio do algoritmo de retropropagação de erro (*backpropagation)* (SILVA, et al., 2010). Nesse algoritmo, a rede é alimentada para frente (*forward*) e para trás (*backward*). Na etapa *forward*, os pesos sinápticos $w_{(t)}$ permanecem inalterados e os sinais funcionais da rede neural são calculados para cada neurônio até que seja produzida a saída desejada na camada de saída. A etapa *backward*, por sua vez, se inicia na camada de saída da rede, passando os sinais de erro para as camadas anteriores, de modo que os pesos sinápticos sejam recalculados de acordo com a regra Delta (equação 15) até que se retorne à primeira camada oculta da rede (HAYKIN, 2001).

$$\Delta w_{(t)} = \mu \Delta w_{(t-1)} + \eta \delta_{(t)} y_{(t)} \quad (15)$$

Em que μ é a constante de *momentum* com 0< μ < 1; δ é o gradiente local; η é a taxa de aprendizagem; y é a saída desejada.

O termo de *momentum* consiste em um parâmetro ponderador das matrizes sinápticas, a partir do qual pode-se calcular qual alteração entre duas iterações anteriores ou consecutivas ocorreu nessas matrizes. À medida que a variação do erro quadrático entre duas iterações se torna muito pequeno, o valor de α também diminui e o processo de convergência da rede se tona mais eficiente (SILVA, et al., 2010).

4.1.2 Redes de funções de base radial (RBF)

Os primeiros estudos envolvendo redes de base radial foram desenvolvidos por Moody & Darken, (1989). As RBF (*radial basis function*) apresentam uma estrutura mais simples que as redes de múltiplas camadas, sendo constituídas por uma camada de entrada, apenas uma camada intermediária e uma camada de saída de neurônios, que é alimentada para frente – procedimento chamado de *feedforward.* Segundo Hecht-Nielsen (1989), com apenas uma camada intermediária na rede neural já é possível se calcular uma função arbitrária qualquer a partir de dados fornecidos. Para esses autores, a camada oculta deve ter por volta de (2i+1) neurônios, onde *i* é o número de variáveis de entrada, no entanto, Haykin (2001) afirma que esse tipo de estruturação é capaz de resolver problemas multivariáveis, mas possui algumas restrições no que se refere a problemas complexos.

A topologia da RBF é composta por funções de ativação de base radial em sua camada intermediária. De modo geral, essas funções retornam valores cada vez menores à medida que a distância entre o ponto observado e centro da função aumentam. Dentre as funções de ativação disponíveis para redes RBF, destacam-se as multiquadráticas, multiquadráticas inversas e as funções gaussianas. É procedência comum empregar as funções do tipo gaussianas para a camada oculta. Na camada de saída são utilizadas funções lineares (PARK & SANDBERG, 1991).

Tal como para a MLP, nas redes RBF a escolha da arquitetura a ser utilizada fica a cargo do pesquisador. O algoritmo de aprendizado adotado é fator determinante para treinar a rede. Nas redes RBF, o treinamento é realizado em duas etapas, motivo pelo qual permite classificar as redes RBF como híbridas (PARK & SANDBERG, 1991). A

primeira etapa adota um método de aprendizagem auto-organizado ou não supervisionado no qual o objetivo principal é formar grupos de indivíduos semelhantes para posterior obtenção dos pesos das funções de base radial. Na segunda etapa, o treinamento é feito com base na regra delta generalizada, de modo similar ao utilizado quando se adota uma arquitetura de rede de múltiplas camadas (SILVA et al., 2010).

A etapa de treinamento não supervisionada da RBF é realizada com o auxílio de métodos de agrupamento de otimização provenientes da estatística multivariada, como o método de *k-means*, por exemplo (SILVA et al., 2010).

a) Treinamento da RBF: Etapa I

Nessa etapa é realizado todo o ajuste dos neurônios da camada intermediária. Para tanto, funções de ativação do tipo radiais gaussianas são utilizadas. Basicamente, essa etapa é realizada com o intuito de transformar um conjunto de entradas não linearmente separáveis em um conjunto linearmente separável (BRAGA et al., 2011).

A idéia da RBF é ajustar os parâmetros *c* e σ^2, em que *c* corresponde ao centro da função gaussiana e σ^2 corresponde à variância da mesma. Assim sendo, após definir o número de neurônios a ser adotado na camada oculta da rede, o parâmetro *c* estará diretamente relacionado aos pesos, de modo que a entrada $u_j^{(1)}$ de cada peso será o próprio vetor de entrada x, que representa os *n* sinais sinápticos. A saída de cada neurônio j da camada intermediária é dada pela equação 16:

$$g_j^{(1)}(u_j^{(1)}) = g_j^{(1)}(x) = e^{\frac{(x_i - w_{ij}^{(1)})^2}{2{\sigma_j}^2}} \tag{16}$$

em que $u_j^{(1)}$ é a entrada da rede.

A diferença básica entre a filosofia da MLP e da RBF é que a primeira utiliza a combinação de hiperplanos enquanto que na RBF os problemas de classificação são resolvidos por meio de campos hiperesféricos (SILVA et al., 2010). A Figura 3 ilustra de forma bem clara essa diferença de filosofia entre a MLP e a RBF para um problema de classificação com duas entradas x1 e x2.

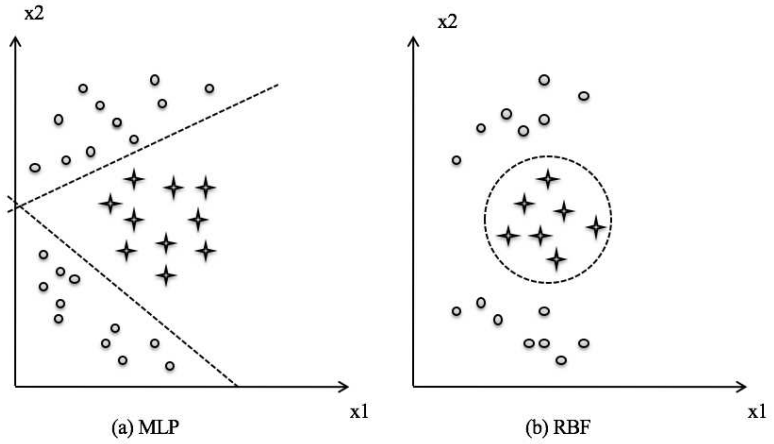

Figura 3. Fronteiras de separabilidade: (a) MLP e (b) RBF

Na Figura 3, os círculos representam a classe que chamaremos aqui de classe 1 e as estrelas representam o que chamaremos de classe 2. É possível observar que utilizando somente um plano hiperesférico (circunferência) já foi possível realizar uma separação eficaz das classes observadas, o que significa que um único neurônio foi considerado na camada oculta. Para a MLP, dois hiperplanos foram considerados (duas retas), indicando que foram necessários dois neurônios para a solução do mesmo problema. Esse exemplo simples nos ajuda a entender melhor a função dos neurônios na camada intermediária de uma rede RBF: posicionar os centros das funções gaussianas utilizadas do modo mais apropriado possível. Para tanto, adota-se o método de *k-means* (BEZDEK et al., 1984), com o propósito de posicionar os centros das k-gaussianas em regiões nas quais as entradas associadas aos pesos tenderão a se agrupar (DUDA et al., 2001).

b) Treinamento da RBF: Etapa II

Após finalizar a primeira etapa de treinamento da rede RBF, na segunda etapa, os passos de ajuste dos pesos dos neurônios da camada de saída serão executados.

O conjunto de treinamento para os neurônios da camada de saída será constituído por pares de entrada e saída desejada. As entradas consistem nas respostas produzidas pelas funções de ativação gaussianas dos neurônios da camada intermediária. Matematicamente:

$$u_j^{(2)} = \sum_{i=1}^{n_1} w_{ij}^{(2)} \cdot g_i^{(1)}\left(u_i^{(1)}\right) - \theta_j,\ j = 1, \ldots, n_2 \qquad (17)$$

em que $w_{ij}^{(2)}$ são os pesos referentes aos neurônios da camada de saída;

θ_j: limiares referentes aos neurônios da camada de saída.

Procedendo com o treinamento, adotando-se uma função linear para ativar a camada de saída, os neurônios da camada de saída serão responsáveis pela combinação linear das funções de ativação gaussianas produzidas (Equação 4) pelos neurônios da camada anterior (SILVA et al., 2010). Matematicamente, a resposta produzida pelo k-ésimo neurônio da camada de saída da RBF será dada pela expressão abaixo:

$$y_j = g_j^{(2)}\left(u_j^{(2)}\right) = u_j^{(2)},\ j = 1, \ldots, n_2 \tag{18}$$

As redes RBF já foram aplicadas em diversos trabalhos que envolvem aproximações de funções e classificação de padrões (CHO & WANG, 1996; CASTRO, 2001; MULGREW, 1996). No entanto, poucos são os estudos que avaliam a aplicação dessa metodologia nos programas de melhoramento genético, de cunho animal ou vegetal (GIANOLA et al., 2011; GONZÀLEZ-CAMACHO ET AL., 2012).

5. Uso de simulação para estudos de genética

A busca incessante pela obtenção de cultivares melhorados a fim de aumentar a produtividade e praticar a seleção de modo mais eficaz consiste no desafio diário dos programas de melhoramento genético. Para tanto, diversos métodos já preconizados de seleção podem ser adotados, como métodos de seleção massal, seleção via progênies, seleção genômica, etc. Apesar das mais diversas metodologias de seleção disponíveis, esta não é uma tarefa fácil porque a grande maioria delas envolve situações complexas e demandam a realização de experimentos, para os quais a obtenção de um banco de dados sob a instalação de delineamentos demanda alto custo e tempo tornando-se, muitas vezes, inviável (SILVA, 2014).

Levando em conta tais dificuldades, as facilidades proporcionadas pelos recursos computacionais disponíveis para os pesquisadores representam um grande aliado na compreensão de problemas genéticos e estatísticos e na obtenção de bancos de dados confiáveis para estudos ligados ao melhoramento genético, com menor tempo e custo e maior capacidade de generalização de situações complexas (BAKER, 1995).

Dentre os recursos computacionais, a simulação computacional tem sido de grande valia em estudos genéticos sob vários contextos, como o de populações, do indivíduo ou do próprio genoma (GURGEL, 2004). Quando se recorre ao uso de modelos simulados, demanda-se dos geneticistas a obtenção de modelos mais simples do que os reais e que retratem da melhor maneira possível os fenômenos de interesse. Dos programadores, esperam-se as rotinas para o processamento adequado, isso porque é por meio da linguagem de programação que se pode estabelecer comunicação com o computador, de tal forma que os dados sejam convenientemente analisados. A linguagem consiste num conjunto de códigos, regras e vocabulários, que fará com que o computador entenda a instrução (CRUZ, 2001).

De modo bem simples pode-se definir simulação como sendo o processo de imitar, por meio de recursos computacionais, o comportamento de um sistema real (DACHS, 1988), englobando certos tipos de modelos lógicos que permitam descrever o sistema natural (NAYLOR, 1971). Os primeiros processos envolvendo simulação de dados surgiram com a utilização do Método de Monte Carlo, por Von Neuman, em 1940, com blindagem de reatores nucleares (MORGAN, 1995) e no Brasil, seus primeiros indícios foram encontrados no trabalho pioneiro de Alain Pierre Clanet, que formou engenheiros de produção com conhecimentos nessa área (DACHS, 1988).

Os processos de simulação podem ser constituídos por modelos físicos, matemáticos, biológicos, computacionais, dentre outros e estão distribuídos em duas linhas de pesquisa: a primeira delas envolve a resolução de problemas matemáticos e a outra utiliza simulações de ensaios baseadas em conceitos probabilísticos (CRUZ, 2006).

Os primeiros trabalhos envolvendo procedimentos de simulação em estudos genéticos foram desenvolvidos por Fraser (1957). Nestes trabalhos se avaliou o efeito da ligação nas taxas de ganho via seleção massal e também nos avanços dos ganhos genéticos. A partir daí diversos outros autores também adotaram a simulação em estudos de genética e melhoramento de plantas (VEIGA et al., 2000; FERREIRA, 2001; GURGEL, 2004; SILVA, et al., 2014; SILVA, et al., 2016; SANT`ANNA, et al., 2015; GUIMARÃES, 2016).

É importante salientar, no entanto, que alguns aspectos devem ser considerados ao se trabalhar com simulação. Um deles se refere à eficiência do processo adotado, a fim de evitar problemas como tamanho amostral, o uso de distribuições inadequadas ou altas taxas de erro. A avaliação da eficiência pode ser realizada por meio de processos de validação. Para validar um sistema simulado, o pesquisador pode fazer com que a simulação opere sob as mesmas condições do sistema real, e verificar – por meio de testes

de hipóteses e outras análises estatísticas relevantes ou comparações com situações já avaliadas anteriormente – se os resultados observados na simulação estão de acordo com os observados no sistema real (CRUZ, 2001).

Os procedimentos de simulação de dados e de populações podem ser realizados em diversos softwares. Como exemplo tem-se o Diallel (BUROW & COORS, 1994), MENDEL (EUCLYDES, 1996), GENES (CRUZ, 2016) e GQMOL (UFV, 2004).

Neste trabalho daremos enfoque ao software GENES, uma vez que este foi o programa adotado para a realização do mesmo. O programa GENES é amplamente adotado em análises de modelos aplicados ao melhoramento genético animal e de plantas pois permite que sejam geradas informações sobre genomas, genótipos de genitores, cruzamentos entre populações e estabelecimento de valores fenotípicos de variáveis quantitativas, dentre outros, por meio do processamento de dados com o auxilio de modelos biométricos adequados. O programa disponibiliza ao pesquisador procedimentos uni e multivariados, com ênfase na estimação de parâmetros genéticos.

O programa GENES está disponível para download, gratuitamente e sem restrições de uso ou divulgação no endereço ftp://ftp.ufv.br/dbg/biodata/, além disso, o programa conta com o auxílio de uma página online no facebook – https://www.facebook.com/GenesNews/ – por meio da qual o usuário pode manter contato direto com professores e alunos do laboratório de Bioinformática da Universidade Federal de Viçosa a fim de tirar dúvidas relacionadas à instalação do software assim como dúvidas relacionadas a procedimentos a serem realizados.

6. Referências

ALLARD, R. W. **Princípios de melhoramento genético das plantas**. São Paulo: Edgard Blücher, 381p, 1971.

BAKER, R. J. **Selection indices in plant breeding**. Boca Raton, Florida: CRC, 1995. 218p.

BEZDEK, J.; EHRLICH, R.; FULL, W.F.C.M. The fuzzy c-means clustering algorithm. **Comp Geosci**. v10, 1984, p:191-203.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. Singapore: Springer. p.738. 2007.

BORÉM, A. **Melhoramento de plantas**. 3.ed. Viçosa: UFV, 2001. 500p.

BRAGA, A. de P.; CARVALHO, A.P. de L. e de.; LUDERMIR, T.B. **Redes neurais artificiais: teoria e aplicações**. 2.ed. Rio de Janeiro: LTC, 2011.

BUROW, M. D.; COORS, J. G. Diallel: a microcomputer program for the simulation and analysis of diallel crosses. **Agronomy Journal**, Madison, v. 86, n. 1, p. 154-158, Jan./Feb. 1994.

CASTRO, M.C.F. **Predição não-linear de Séries temporais usando Redes Neurais RBF por Decomposição em Componentes Principais.** Tese (Doutorado em Engenharia Elétrica) – Universidade Estadual de Campinas, 2001.

CHO, K. B.; WANG, B. H. **Radial basis function based adaptive fuzzy systems and their applications to system identification and prediction**. Fuzzy Sets and Systems, v. 83, n. 3, p. 325-339, 1996.

CRUZ, C.D. Genes Software – extended and integrated with the R, Matlab and Selegen. **Acta Scientiarum. Agronomy**. Maringá, v. 38, n. 4, p. 547-552, Oct.-Dec., 2016.

CRUZ, C.D. **Princípios de genética quantitativa**. Viçosa: Ed. da UFV, 2012. 394p.

CRUZ, C.D.; REGAZZI, A.J.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético**. UFV, Viçosa, 2012.

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

CRUZ, C.D. **Programa genes: análise multivariada e simulação**. Viçosa: UFV, 2006. 175p.

CRUZ, C. D. A informática no melhoramento genético. In: NASS, L. L. et al. (Eds). **Recursos Genéticos e Melhoramento de Plantas**. Rondonópolis: Fundação-MT, 2001. p. 1086-1118.

DACHS, J. N. W. **Estatística computacional: uma introdução ao Turbo Pascal**. Rio de Janeiro: Livros Técnicos e Científicos, 1988. 236p.

DUDA, R.O.; HART, P.E.; STORK, D.G. **Pattern Classification**. 2ed. Wiley Interscience, 2001.

EUCLYDES, R. F. **Uso de sistemas Genesys na avaliação de métodos de seleção clássicos e associados a marcadores moleculares**. 1996. 135 p. Tese (Doutorado) – Universidade Federal de Viçosa, Viçosa, MG

FALCONER, D.S.; MACKAY, T.F.C. **Introduction to Quantitative Genetics**, Ed 4. Longmans Green: Harlow, Essex, UK, 1996, 464p.

FALCONER, D.S. **Introdução à genética quantitativa**, Trad. SILVA, M.A. & SILVA, J.C. Viçosa, UFV. Imprensa Universitária, 1987. 279p.

FERREIRA, D. F. Uso de simulação no melhoramento. In: NASS, L. L.; VALOIS, A. C. C.; MELO, I. S. de; VALADARES-INGLIS, M. C. (Ed). **Recursos genéticos e melhoramento de plantas**. Rondonópolis: Fundação-MT, 2001. p:1119-1141.

FRASER, A. S. Simulation of genetics systems by automatic digital computers. I: Introduction. **Australian Journal of Biological Science**, Melboourne, v. 10, p. 484-491, 1957a.

GONZÀLEZ-CAMACHO, J.M.; DE LOS CAMPOS, G.; Pérez, P.; GIANOLA, D.; CAIRNS, J.E.; MAHUKU G., BABU, R.; CROSSA, J. Genome-enabled prediction of genetic values using radial basis function neural networks. **Theor. Appl. Genet.** 125:759–771, 2012.

GIANOLA, D.; OKUT, H.; KENT A WEIGEL, K.A.; ROSA, G J.M. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. **BMC Genetics.** 12:87, 2011.

GUIMARÃES, J.F.R. **Efeito da interação dominância x ambiente na habilidade de predição genômica**. Tese (Doutorado em Genética e Melhoramento) – Universidade Federal de Viçosa, 30p, 2016.

GURGEL, F.L. **Simulação computacional no melhoramento genético de plantas**. Tese (Doutorado em Agronomia) – Universidade Federal de Lavras, 174p, 2004.

HAYKIN, S. **Redes neurais: princípios e prática**. 2ed. Porto Alegre: Bookman, 2001.

HECHT-NIELSEN, R. **Theory of the backpropagation neural network**. In: Neural Networks, 1989. IJCNN., International Joint Conference on. IEEE, 1989. p. 593-605.

HENDERSON, C. R. **Applications of Linear models in Animal Breeding**. Univ. of Guelph (in press), 1984.

HENDERSON, C. R. **Sire evaluation and genetic trends**. In: ANIMAL BREEDING AND GENETICS SYMPOSIUM, 10., 1973, Champaign. Proceedings Champaign: American Society of Animal Science, 1973. p:10-41.

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

MOODY, J.; DARKEN, C. Fast learning in networks of locally-tuned processing units. **Neural Comput**. v1, 1989, p:281-294.

MORGAN, B. J. T. **Elements of simulation**. London: Chapman & Hall, 1995. 351p.

MULGREW, B. Applying Radial Basis Functions. **IEEE Signal Processing Magazine**, p:50-65, 1996.

NAYLOR, Thomas H. **Computer simulation experiments with models of economic systems.** New York: John Wiley & Sons, 1971. 502p.

PARK, J.; SANDBERG, I.W. Universal approximation using radial basis function networks. **Neural Comput.** 3ed, v2, 1991, p:246–259.

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

RESENDE, M.D.V.; LOPES, P.S.; SILVA, R.L.; PIRES, I.E. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, Colombo, n.56, p.63-77, jan./jun. 2008.

RESENDE, M.D.V. **Matemática e estatística na análise de experimentos e no melhoramento genético**. Colombo: Embrapa Florestas, 2007. 561p.

ROBINSON, D.L. That BLUP is a good thing: the estimation of random effects. **Statistical Science**, Hayward, v.6, p.15-32, 1991.

SANT'ANNA, I.C.; TOMAZ, R.S.; SILVA, G.N.; NASCIMENTO, M.; BHERING, L.L.; CRUZ, C.D. Superiority of artificial neural networks for a genetic classification procedure. **Genetics and Molecular Research**, v.14, p.9898-9906, 2015. DOI: 10.4238/2015. August.19.24.

SILVA, G.N.; TOMAZ, R.S.; SANT'ANNA, I.C.; CARNEIRO, V.Q.; CRUZ, C.D.; NASCIMENTO, M. Evaluation of the efficiency of artificial neural networks for genetic value prediction. **Genetic Molecular Research**, v.15, p.1-11, 2016. DOI: 10.4238/ gmr.15017676, 2016.

SILVA, G.N.; TOMAZ, R.S.; SANT'ANNA, I. de C.; NASCIMENTO, M.; BHERING, L.L.; CRUZ, C.D. Neural networks for predicting breeding values and genetic gains. **Scientia Agricola**, v.71, p.494-498. DOI: 10.1590/0103- 9016-2014-0057. 2014.

SILVA, G.N. **Redes neurais artificiais: novo paradigma para a predição de valores genéticos**. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, 105p, 2014.

SILVA, I. N.; SPATTI, H. D.; FLAUZINO, R. A. **Redes Neurais Artificiais: para engenharia e ciências aplicadas**. São Paulo: Artliber, 2010. 399p.

VEIGA, R. D.; FERREIRA, D. F.; RAMALHO, M. A. P. Eficiência dos dialelos circulantes na escolha de genitores. **Pesquisa Agropecuária Brasileira**, Brasília, v. 35, n. 7, jul. 2000.

CAPÍTULO 1

SELEÇÃO GENÔMICA NA PREDIÇÃO DE VALORES GENÉTICOS EM UMA POPULAÇÃO SIMULADA

Resumo
O objetivo deste trabalho foi avaliar o uso da seleção genômica (SG) na predição de valores genéticos para doze características quantitativas, que contemplavam diferentes estruturações quanto à modelo (aditivo, aditivo-dominante e epistáticos), graus médios de dominância e herdabilidade. Foram utilizados 500 indivíduos provenientes de uma população simulada F1, oriunda do cruzamento de duas populações contrastantes, também obtidas via simulação. Os 500 indivíduos foram genotipados com 1000 marcadores. Realizaram-se análises com seleção genômica ampla por meio da metodologia de *Ridge Regression-Best Linear Unbiased Prediction* – RR-BLUP. A metodologia de RR-BLUP foi capaz de realizar a predição, no entanto, a existência de epistasia e de efeitos de dominância resultaram em acréscimos no erro quadrático médio na fase de validação. O elevado número de marcadores moleculares também se apresentou como fator perturbador para realizar-se a predição, dada a alta dimensionalidade e existência de multicolinearidade entre as marcas. Os resultados demonstraram que o modelo convencional empregado na abordagem RR-BLUP não apresenta boa parametrização para predizer características que apresentem a presença de efeitos intra e inter alélicos. Demonstraram ainda que a redução de dimensionalidade da matriz de marcas se apresenta como alternativa para aumentar a capacidade preditiva dessa metodologia.
Termos para indexação: Interações alélicas, marcadores moleculares, redução de dimensionalidade, predição.

Abstract
The objective of this work was to evaluate the use of Genome Wide Selection to predict genetic values for twelve quantitative traits, which included different modeling (additive, additive-dominant and epistatic), average degrees of dominance and heritability. We used 500 individuals from a simulated F1 population, coming from the crossing of two contrasting populations, also obtained through simulation. The 500 individuals were genotyped with 1000 markers. Analyzes with genome wide selection were carried out using the *Ridge Regression-Best Linear Unbiased Prediction* – RR-BLUP. methodology. The RR-BLUP was able to perform the prediction, however, the presence of epistasis and dominance effects increased the mean squared error in the validation step. The high number of molecular markers also acted as a disturbing factor to carry out the prediction, given the high dimensionality and existence of multicollinearity between the markers. The results demonstrated that the RR-BLUP method does not present a good parameterization to predict characteristics that present the presence of intra and inter-allelic effects. They also demonstrated that reduce the dimensionality of the markers matrix represents as a good alternative to increase the predictive capacity of this methodology.
Index terms: Allelic interactions, molecular markers, dimensionality reduction, prediction.

INTRODUÇÃO

A necessidade de se obter maior ganho genético por unidade de tempo levou os pesquisadores a formularem diferentes estratégias de melhoramento de forma que a observação fenotípica pudesse ser mais acurada. Nas metodologias tradicionais – que se

baseiam nos princípios de genética quantitativa –, a seleção é praticada fundamentalmente com base no estabelecimento de uma população segregante, resultante do intercruzamento de genitores favoráveis, seguido da escolha fenotípica dos melhores indivíduos e/ou, famílias. No entanto, quando os estudos envolvem o controle de características complexas – afetadas por fatores perturbadores como diferentes níveis de dominância, presença de epistasia e forte influência ambiental –, esses métodos tradicionais se tornam por diversas vezes ineficazes mesmo com aprimoramento das técnicas experimentais e com a inclusão de alguma informação genética adicional como parentesco ou caracteres correlacionados. Outra problemática envolvida está ligada ao fato de que com o desenvolvimento de populações já com desempenho superiores, fica cada vez mais difícil identificar novos genótipos que permitam ir além do patamar estabelecido pela população melhorada já em uso (CRUZ et al., 2013).

Nesse sentido, aliar o uso da biotecnologia às metodologias disponíveis tem se mostrado estratégia bastante eficaz. Dentre as ferramentas de biotecnologia no melhoramento genético destaca-se o uso de marcadores moleculares. Marcadores moleculares são sequências de DNA que revelam polimorfismos entre indivíduos geneticamente relacionados (CRUZ et al., 2013), e a identificação de marcadores ligados a genes controladores de características oligogênicas e/ou poligênicas de interesse tem sido de grande valia para os programas de melhoramento genético.

Características como maior produtividade agrícola associada à melhoria na qualidade nutricional das populações, resistência e melhor adaptação à ambientes adversos, dentre outras, representam pontos decisivos nos programas de melhoramento (SILVA, 2014). Visando melhorar tais características, uma área com um conjunto de ferramentas biométricas diferenciadas, disponíveis e cada vez mais úteis aos melhoristas é a genômica. Genômica é a ciência que estuda o genoma aplicado a sistemas biológicos (LANDER & WEINBERG, 2000).

As análises de marcadores moleculares, inicialmente relacionadas à construção de mapas genéticos, representam uma das principais contribuições da genômica para os programas de melhoramento genético animal e vegetal e vem sendo utilizada de forma crescente por diversos autores (FALEIRO et al., 2003; CABRAL, 2001; BRONDANI, 2000; KY et al., 2000; MELO, 2000; TANKSLEY, 1994; PAILLARD et al., 1996). O uso dessas informações genéticas por meio de informações de marcadores moleculares se consolidou como técnica eficaz para seleção de indivíduos com maior desempenho quando, em 2001, Meuwissen propôs a Seleção Genômica Ampla (*Genome Wide Selection - GWS*) cuja eficácia se fundamenta no desequilíbrio de fase gamética que

contempla vários fatores além da ligação fatorial entre marcas e genes controladores das características quantitativas. Tal metodologia permite incorporar informações moleculares diretamente na predição do mérito genético dos indivíduos. Dessa forma, utiliza-se tanto a informação fenotípica do indivíduo quanto os dados de marcadores moleculares para identificar e selecionar genótipos superiores (CRUZ et al., 2013). Essa nova metodologia de seleção conseguiu se desenvolver graças aos amplos investimentos em genotipagem em grande escala com o desenvolvimento dos marcadores tipo SNP (*Single Nucleotide Polymorphism*).

O maior atrativo da GWS consiste na possibilidade de utilização direta da informação de DNA na seleção, podendo ser em todas as famílias em avaliação e propiciando alta acurácia seletiva, além de não exigir prévio conhecimento dos mapas associados à seleção. A GWS permite ainda a predição de valores genômicos e tem sua eficácia comprovada para características de baixa herdabilidade (RESENDE et al., 2014). Com base no exposto, o objetivo deste capítulo foi avaliar a GWS, por meio da metodologia de RR-BLUP, como alternativa de predição em população simulada considerando diferentes cenários que representam situações de dificuldade para o melhoramento fundamentado em reprodução sexuada tais como a presença de dominância, epistasia e de efeitos ambientais.

MATERIAL E MÉTODOS

1. Simulação das populações

Inicialmente foram simuladas dez populações, todas elas constituídas por 500 indivíduos e genotipadas em relação a 1000 marcadores moleculares codominantes. Um estudo prévio de diversidade genética – por meio do método de Identidade de Nei 1972 – foi realizado afim de que se determinassem as duas populações mais divergentes dentre as 10 simuladas (Tabela 1). A condição básica para que exista eficiência nas técnicas fundamentadas na GWS é que exista desequilíbrio de ligação entre o loco controlador da característica quantitativa (QTL) e o marcador genético.

Vários fatores podem contribuir para este desequilíbrio e, segundo Cruz et al. (2013), populações advindas da hibridação entre populações contrastantes oferecem as condições mais apropriadas para ajustes de modelos e obtenção de preditores mais acurados. Assim sendo, a necessidade de se gerar desequilíbrio de ligação, estimar os efeitos e o mapeamento é que justificam a escolha das populações mais contrastantes, que no presente trabalho foram as populações P7 e P8 (Tabela 1), a partir das quais foi gerada

a população F1, cujas informações genotípicas e fenotípicas foram objetos de estudo deste trabalho.

Tabela 1. Matriz de dissimilaridade – Identidade de Nei 1972 – das dez populações.

P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
	0,2807	0,2894	0,2989	0,3083	0,3093	0,2703	0,3094	0,2914	0,2896
		0,2889	0,2986	0,3116	0,2902	0,2727	0,3107	0,29	0,3058
			0,293	0,2971	0,2981	0,284	0,2893	0,2778	0,2863
				0,3408	0,2798	0,2888	0,2971	0,2726	0,2835
					0,3259	0,3161	**0,3474**	0,2869	0,3177
						0,2822	0,3062	0,2855	0,2994
							0,3137	0,2838	0,2717
								0,2932	0,3156
									0,2772

Os dados fenotípicos foram simulados adotando-se dois modelos: modelo aditivo (Equação 2) e modelo epistático (Equação 3) e adotando-se três níveis de grau médio de dominância (gmd): (i) ausência de dominância (gmd=0); (ii) dominância parcial (gmd=0,6); (iii) sobredominância (gmd=1,2). Foi pressuposto que estes caracteres simulados eram controlados por 100 locos, com importância diferenciada sobre a expressão fenotípica com pesos estabelecidos por uma família de probabilidade resultante da distribuição binomial com parâmetros $p = q = 0,5$ e $n = 99$. Também foi considerada a influência ambiental de forma que as herdabilidades apresentassem magnitude de $h^2 = 35\%$ e $h^2 = 70\%$.

Assim, combinando os modelos, tal como níveis de dominância e herdabilidades adotados, o trabalho contemplou o estudo de doze características quantitativas. Para facilitar a apresentação dos resultados e posteriores discussões, as características serão denotadas como se segue na Tabela 2.

Tabela 2. Características avaliadas no estudo com seus respectivos valores de herdabilidade, modelo adotado e grau médio de dominância (gmd).

Característica	Herdabilidade (%)	Modelo	gmd
V1 - D0H35Ad	35	aditivo	0
V2 - D0H35Ep	35	epistático	0
V3 - D0H70Ad	70	aditivo	0
V4 - D0H70Ep	70	epistático	0
V5 - D60H35Ad	35	aditivo-dominante	0,6
V6 - D60H35Ep	35	epistático	0,6
V7 - D60H70Ad	70	aditivo-dominante	0,6
V8 - D60H70Ep	70	epistático	0,6
V9 - D120H35Ad	35	aditivo-dominante	1,2
V10 - D120H35Ep	35	epistático	1,2
V11 - D120H70Ad	70	aditivo-dominante	1,2
V12 - D120H70Ep	70	epistático	1,2

Os valores fenotípicos, F_i, dos indivíduos foram gerados segundo a equação (1) abaixo.

$$F_i = G_i + E_i \tag{1}$$

Em que:

G_i: efeito genético dado pelo somatório dos efeitos genéticos em cada loco;

E_i: efeito ambiental

O valor fenotípico total, no modelo aditivo-dominante e no modelo epistático, expresso por um determinado indivíduo pertencente à população foi estimado a partir das expressões (2) e (3), respectivamente.

$$Y_i = \mu + \sum_{j=1}^{100} p_j \alpha_j + E_i \tag{2}$$

$$Y_i = \mu + \sum_{j=1}^{100} p_j \alpha_j + \sum_{j=1}^{99} p_j \alpha_j \, \alpha_{j+1} + E_i \tag{3}$$

Em que:

$\alpha_j = a_i + d_i$ e d_i/a_i = gmd, sendo $\mu + a_j$, $\mu + d_j$ e $\mu - a_j$ os valores genotípicos associados as classes AA, Aa e aa, respectivamente que são identificadas pela codificação 1, 0 ou -1, respectivamente, e p_j a contribuição do loco j para a manifestação da característica considerada estabelecida por meio dos elementos representativos da

probabilidade gerada a partir de uma distribuição binomial, com parâmetros n = 99 e p = 0.5.

2. Método estatístico empregado para fins de predição por seleção genômica ampla

Quando se adota um método de GWS, espera-se que três atributos básicos sejam satisfeitos: (i) a arquitetura genética do caráter em estudo deve ser mantida; (ii) deve-se proceder à regularização do processo de estimação caso haja multicolinearidade e grande número de marcadores, e finalmente, (iii) deve-se identificar os marcadores mais importantes por meio de seleção de marcadores (RESENDE et al., 2014). Na literatura estão disponíveis diversos métodos estatísticos para seleção genômica ampla que incluem metodologias de regressão explícita, de regressão implícita e de regressão Kernel não paramétrica (RESENDE et al., 2014). Neste trabalho adotamos o método de regressão explícita de RR-BLUP aditivo (sem a inclusão da matriz de efeitos de dominância) seguindo o modelo proposto por Meuwissen et al., (2001):

$$y = Wb + Xm + e \tag{5}$$

em que y é o vetor de observações fenotípicas; *b* é vetor de efeitos fixos (média geral) com matriz de incidência W; *X* é a matriz de incidência composta pelos valores -1, 0 e 1 para o número de alelos do marcador dos genótipos mm, Mm e MM, respectivamente; *m* é vetor dos efeitos aleatórios dos marcadores com matriz de incidência X com $m \sim N(0, I\sigma_g^2)$; *e* refere-se ao vetor de resíduos aleatórios com $e \sim N(0, I\sigma_e^2)$, σ_e^2 é a variância do erro.

O RR-BLUP utiliza preditores do tipo BLUP e assume que os efeitos dos marcadores são covariáveis de efeito aleatório. Sua predição baseia-se na seguinte equação de modelo misto:

$$\begin{bmatrix} W'W & W'X \\ X'W & X'X + I\frac{\sigma_e^2}{(\sigma_g^2/n_Q)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} W'y \\ X'y \end{bmatrix} \tag{6}$$

em que b é o vetor de efeitos fixos; m é vetor dos efeitos aleatórios dos marcadores com matriz de incidência X; σ_g^2 é a variância genética; $n_Q = \sum_{j=1}^{j} 2p_j(1-p_j)$ e p_j é a

frequência alélica do marcador j.

2.1 Populações de treinamento e validação

A população de treinamento, também chamada por alguns autores de população de estimação, consiste em uma população relativamente grande, que terá seus fenótipos avaliados para as características de interesse (RESENDE et al., 2014). A população de validação por sua vez, consiste em um conjunto de dados, que geralmente é menor que a população de treinamento, e envolve indivíduos genotipados e fenotipados para a característica de interesse (RESENDE et al., 2014). Geralmente, após a estimação do modelo de predição, este é aplicado a uma população de validação. Na prática, essas populações podem ser definidas de diferentes maneiras, de modo que uma mesma população pode ser particionada em dois conjuntos (treinamento e validação), ou em três conjuntos (treinamento, validação e seleção). Pode-se considerar ainda a utilização de duas ou três populações distintas para o desenvolvimento das etapas apresentadas acima. Em estudos GWS, é comum utilizar uma mesma população para exercer ambas as funções (RESENDE et al., 2014).

Neste trabalho, a mesma população F1 foi utilizada para a realização de ambas as fases. Na literatura, quando uma mesma população é particionada a fim de se obterem populações de treinamento e de validação, recomenda-se adotar um procedimento denominado validação cruzada (LEGARRA et al., 2008) para posterior estimação das acurácias.

2.2 Validação cruzada (*k-fold*)

A validação cruzada consiste em um método empregado para avaliar a capacidade de generalização de um modelo preditivo de modo a contornar eventuais problemas de superparametrização (MEUWISSEN, 2007; RESENDE et al., 2014), uma vez que ao final das análises, a média de resultados será considerada em detrimento ao uso da técnica em um conjunto único de treinamento e de validação. Neste trabalho foi adotada uma validação cruzada (*k-fold*) com k = 5 partições, isto é, a população de tamanho n = 500 foi particionada em cinco subconjuntos mutuamente exclusivos e a cada rodada quatro desses subconjuntos constituíram a população de treinamento (totalizando 80% dos indivíduos) e o subconjunto restante constituiu a população de validação (20% da população total).

Para calcular a acurácia da GWS, utilizou-se o valor genético genômico estimado (GEBV) do indivíduo j via RR-BLUP por meio da expressão que se segue:

$$GEBV = \hat{y}_j = \hat{\mu} + \sum_i x_{ij}\hat{m}_i \quad (7)$$

em que X_i equivale a $-1, 0$ ou 1 para os genótipos mm, Mm e MM, respectivamente, para o marcador i. O componente x_{ij} é o elemento i da linha j da matriz X, referente ao indivíduo j.

2.3 Medidas de eficiência do procedimento biométrico

Para avaliar a acurácia, confiabilidade e eficiência do método de RR-BLUP empregado, o trabalho contemplou estudo de algumas estatísticas que informam sobre as acurácias seletivas e preditivas do modelo. A acurácia seletiva é dada pela correlação entre o valor fenotípico observado (Y) e o valor genético genômico estimado (GEBV), refletida nas estatísticas: r_t^2 e r_v^2, que correspondem ao quadrado da correlação entre Y e GEBV para as fases de treinamento, respectivamente; CP, que fornece a capacidade preditiva do modelo e o viés (β), que consiste no coeficiente da regressão estimada.

A acurácia preditiva por sua vez, consiste na habilidade que o modelo possui de predizer corretamente o valor verdadeiro esperado e pode ser medida por meio do erro quadrático médio (EQM) ou por sua raiz ($REQM$), definidas a seguir.

a) O quadrado da correlação (r^2)

A confiabilidade da metodologia proposta pode ser expressa pelo quadrado da correlação entre os parâmetros estimados (efeito estimado dos marcadores) e os parâmetros observados expressos pelo valor fenotípico dos indivíduos ($\hat{Y}$ e Y), em analogia ao quadrado da correlação entre os valores genotípicos e valores fenotípicos dos indivíduos, que em genética quantitativa, expressam a herdabilidade da característica (CRUZ, 2005). Diversos autores têm utilizado esse parâmetro a fim de verificar a eficiência de metodologias que envolvam problemas de predição ou classificação de populações (SILVA et al., 2014; SANT'ANNA et al., 2015; SILVA et al., 2016).

A correlação consiste na contabilização da relação linear existente entre duas variáveis de interesse. Desse modo, o valor de r^2 pode ser definido como se segue:

$$r^2 = (cor(GEBV, Y))^2 = (cor(\hat{Y}, Y))^2 \quad (8)$$

b) Raiz do erro quadrático médio (REQM)

Para avaliar o desempenho de determinada metodologia, o pesquisador precisa optar por aquela que lhe confira a possibilidade de quantificar o quão próximo os valores preditos estão do valor verdadeiro por ele esperado. Para ajustes de modelos de regressão, a literatura propõe o uso do erro quadrático médio (EQM) como medida mais adequada para exercer tal papel (JAMES et al., 2013). O EQM consiste na esperança do quadrado da diferença entre o valor do estimador e do parâmetro ao quadrado. Matematicamente define-se:

$$EQM = E(GEBV - Y)^2 = E(\hat{Y} - Y)^2 = \frac{\sum(\hat{Y}-Y)^2}{n} \quad (9)$$

James et al., (2013) afirmam que os valores do EQM serão menores à medida que as respostas do estimador se aproximarem das respostas do parâmetro (respostas verdadeiras). Se $\hat{Y}$ e Y diferirem substancialmente, o pesquisador terá indícios de que a metodologia avaliada apresenta problemas de ajuste.

Adicionalmente ao EQM, a raiz do erro quadrático médio (REQM) é comumente adotada para expressar a acurácia preditiva dos modelos pois apresenta a vantagem de apresentar os valores do erro na mesma escala da variável de interesse. Desse modo, adotou-se o REQM para esse estudo.

$$REQM = \sqrt{\frac{\sum(\hat{Y}-Y)^2}{n}} \quad (10)$$

c) Capacidade preditiva (CP)

Quando a amostra de validação não é envolvida na predição de efeitos dos marcadores, a correlação entre o efeito estimado dos marcadores e o valor fenotípico dos indivíduos é predominantemente de natureza genética (RESENDE et al., 2014). Nesse caso, a correlação passa a ser definida como sendo a capacidade preditiva da GWS e é dada pela equação abaixo:

$$CP = cor(GEBV, Y) = cor\left(\hat{Y}, Y\right) \quad (11)$$

d) Viés (β)

O viés foi avaliado tendo como referência o coeficiente da regressão estimada (β) para os dados centrados na média. Os valores de β iguais a 1 representam, teoricamente, que não há viés na predição. O viés é definido como:

$$\hat{\beta} = \frac{cov(\hat{Y}, Y)}{var(Y)} \quad (14)$$

Uma vez comprovada a acurácia dos modelos preditos – por meio dos parâmetros apresentados acima –, estes poderão ser utilizados posteriormente para predição dos VGGs de indivíduos em gerações futuras da população na qual está sendo praticado o melhoramento, sem que seja necessário novo processo de fenotipagem para o caráter de interesse econômico (CRUZ et al., 2013). Nessa fase, a estimativa da acurácia é obtida de modo análogo ao processo realizado na população de validação (GODDARD & HAYES, 2007; RESENDE, 2008).

Esse trabalho contemplou somente a avaliação do método RR-BLUP nas populações simuladas para as etapas de treinamento e validação, sem estendê-lo para gerações futuras.

3. Recursos Computacionais

Para a etapa de simulação das populações e análises de diversidade genética das mesmas, o software GENES foi utilizado (CRUZ, 2013). Para avaliar a metodologia de RR-BLUP proposta, o software GENES também foi utilizado, porém no módulo de integração com o software R (R Core Team, 2018).

RESULTADOS E DISCUSSÃO

Como apresentado anteriormente, as informações acerca de X e y são suficientes para o estabelecimento de inferências a respeito das análises efetuadas, uma vez que o RR-BLUP, para GWS, contempla em seu modelo: o número de observações fenotípicas; a contagem de ‘doses’ associados aos marcadores moleculares; os componentes de

variância (herdabilidade associada a estratégia da seleção) e o número de locos cromossômicos (RESENDE et al., 2014).

Os principais desafios enfrentados pelas metodologias estatísticas de GWS consistem nos cuidados a serem tomados com alguns atributos. O primeiro deles está ligado à acomodação da arquitetura genética em termos dos efeitos dos genes e suas distribuições e realização da seleção de covariáveis (marcadores moleculares) importantes para o caráter em análise (RESENDE et al., 2014).

Outro problema enfrentado pela GWS é que, na maioria das vezes, o experimento a ser avaliado envolve número parâmetros a ser estimado muito maior que o número de indivíduos genotipados, levando a problemas de dimensionalidade. E ainda, o elevado número de marcas moleculares pode levar também a problemas de multicolinearidade, devido ao desequilíbrio de ligação e alta correlação entre os marcadores (RESENDE et al., 2012). Uma maneira de tentar contornar tais problemas consiste em adotar os métodos de regressão explícita ou então assumir os efeitos de marcadores como sendo aleatórios ao invés de fixos, de modo que os efeitos de todos os marcadores possam ser estimados simultaneamente (RESENDE et al., 2012). Essa é a filosofia geral adotada pelo método RR-BLUP/GWS. A avaliação da eficiência do RR-BLUP nos vários cenários estudados neste capítulo será discutida nos tópicos a seguir, tomando por base algumas considerações genéticas e estatísticas importantes.

a) Influência da dominância sob a eficiência da GWS

A dominância consiste no efeito de toda e qualquer interação intra-alélica que possa existir em uma característica quantitativa de interesse. A presença de tal efeito tem sido considerada fator perturbador para os programas de melhoramento quando o intuito do pesquisador é distinguir um homozigoto dominante de um heterozigoto, uma vez que dificulta a identificação de genótipos superiores em populações segregantes. A variância atribuída aos desvios de dominância por sua vez, corresponde à fração não herdável da variância genotípica, advinda da soma de quadrados dos desvios do modelo aditivo-dominante, de modo que os valores genotípicos dos indivíduos são expressos em função não somente de cada alelo individualmente, mas também considerando o efeito de interação dos mesmos (CRUZ & CARNEIRO, 2003; CRUZ, 2005).

A eficiência do RR-BLUP em cenários complexos pode ser confirmada pelos resultados descritos na Tabela 3.

Tabela 3. Estimativas obtidos pelo RR-BLUP com cinco validações cruzadas.

Característica		r_t^2	r_v^2	$REQM_t$	$REQM_v$	CP	Viés
V1 - D0H35Ad	Md	**0,4887**	**0,0513**	**105,002**	**105,606**	**0,2159**	**0,1793**
	dv	0,0386	0,033	1,39	1,50229	0,0768	0,033
V2 - D0H35Ep	Md	**0,4869**	**0,0181**	**152,35**	**164,93**	**0,1296**	**0,1567**
	dv	0,0078	0,0105	1,39	1,50229	0,0401	0,008
V3 - D0H70Ad	Md	**0,8783**	**0,2442**	**99,3829**	**99,9704**	**0,4885**	**0,6198**
	dv	0,0149	0,0818	4,7894	8,79578	0,0832	0,0307
V4 - D0H70Ep	Md	**0,5665**	**0,0727**	**129,85**	**140,382**	**0,2462**	**0,2507**
	dv	0,0734	0,0625	2,9971	3,05234	0,1229	0,0664
V5 - D60H35Ad	Md	**0,5905**	**0,0737**	**140,792**	**141,481**	**0,2639**	**0,2452**
	dv	0,079	0,0422	3,1413	3,6502	0,0713	0,0654
V6 - D60H35Ep	Md	**0,5533**	**0,0475**	**311,655**	**350,52**	**0,2073**	**0,225**
	dv	0,0647	0,0256	21,8313	17,488	0,0752	0,0493
V7 - D60H70Ad	Md	**0,8815**	**0,2788**	**132,231**	**132,977**	**0,5272**	**0,6398**
	dv	0,0091	0,0351	3,14723	1,29534	0,0329	0,0161
V8 - D60H70Ep	Md	**0,715**	**0,1469**	**236,89**	**271,211**	**0,3696**	**0,3953**
	dv	0,0385	0,0763	27,3487	21,3052	0,1131	0,0546
V9 - D120H35Ad	Md	**0,6635**	**0,1248**	**169,171**	**170,42**	**0,3446**	**0,3011**
	dv	0,0345	0,0589	4,0344	5,07289	0,0872	0,0385
V10 - D120H35Ep	Md	**0,384**	**0,025**	**930,306**	**968,754**	**0,1295**	**0,1021**
	dv	0,0539	0,028	26,434	70,2143	0,1014	0,0325
V11 - D120H70Ad	Md	**0,7949**	**0,2057**	**170,65**	**171,203**	**0,4439**	**0,5009**
	dv	0,0297	0,0927	3,8723	2,7558	0,1043	0,0507
V12 - D120H70Ep	Md	**0,7032**	**0,14**	**730,146**	**792,085**	**0,3729**	**0,3788**
	dv	0,0298	0,0253	51,6098	44,1653	0,0345	0,026

r_t^2 e r_v^2: referem-se aos coeficientes de determinação obtidos no treinamento e na validação, respectivamente; CP: capacidade preditiva; $REQM_t$ e $REQM_v$: raiz do erro quadrático médio para as fases da treinamento e validação, respectivamente.

É importante que a interpretação dos resultados apresentados na Tabela 3 seja feita sobre duas óticas, sendo a primeira relativa ao impacto da presença da dominância sobre a confiabilidade da seleção ou acurácia seletiva, expressa pelo r_v^2 e a segunda pela acurácia de predição expressa pela raiz do erro quadrático médio, denotado por $REQM_v$.

Analisando os valores do r_v^2, e parâmetros associados tais como capacidade preditiva e viés, verifica-se que a presença da dominância tem pequeno impacto ou até melhora a acurácia seletiva. Considerando os conjuntos (V1, V5, V9) ou (V2, V6, V10) ou (V3, V7, V11), ou (V4, V8, V12) geralmente há, com a inclusão de níveis mais elevados da dominância, sentido de acréscimo no valor do parâmetro estudado (r_v^2). Levando em consideração as definições de dominância e variância devido aos desvios de dominância explicitados anteriormente, e recorrendo novamente ao modelo do RR-BLUP

para GWS apresentado em tópicos anteriores, nota-se que ele contempla somente os efeitos aditivos dos indivíduos, uma vez que para o cálculo do parâmetro de penalização λ somente variâncias genéticas aditivas são consideradas com incidência sobre a matriz *X*. No entanto, para estudo de herança e variação de características quantitativas, Allard (1971) reconhece que a variância genotípica dos indivíduos pode ser particionada em três componentes: variância aditiva; variância atribuída aos desvios de dominância devido às interações intra-alélicas e variância atribuída aos efeitos epistáticos devido às interações interalélicas.

O modelo proposto no RR-BLUP aditivo não reproduz todos os componentes da variância genotípica, uma vez que possíveis efeitos devido às interações intra-alélicas são negligenciados pela metodologia de RR-BLUP. Entretanto, a variância aditiva é função do efeito médio de substituição gênica que, por sua vez, é impactado pelo valor genotípico do heterozigoto de forma que redução sobre a variação herdável, na presença de dominância, pode não ser detectada como verificado no presente estudo.

Por outro lado, reanalisando os conjuntos (V1, V5, V9) ou (V2, V6, V10) ou (V3, V7, V11), ou (V4, V8, V12), mas enfatizando a acurácia preditiva refletida nos resultados obtidos para o $REQM_v$, constata-se aumento indesejável sobre estes valores. Percebe-se que a eficiência de predição do RR-BLUP é maior quando a característica estudada se estabelece via um modelo aditivo, com ausência de dominância, independentemente da herdabilidade considerada (características V1 e V3). Especificamente para essas características, encontraram-se $REQM_v$ = 105,606 e $REQM_v$ = 99,9704, para a etapa de validação de V1e V3, respectivamente. O erro quadrático mais elevado encontrado para V1 evidenciou que baixas herdabilidades também atuam como fatores perturbadores em problemas de predição.

A inclusão da dominância para as características estabelecidas ainda sob o modelo aditivo – gmd=0,6 para características (V5) e (V7); gmd=1,2 para as características (V9) e (V11) – resultaram em erros quadráticos médios maiores. O $REQM_v$ nesse caso variou entre 132,977 e 171,203. Os resultados evidenciaram que a inclusão de efeitos de dominância prejudicou a eficiência do RR-BLUP no contexto de reduzir a acurácia de predição. A ineficiência da GWS na presença de efeitos de dominância já foi relatada por alguns autores (RESENDE et al., 2008; ALMEIDA FILHO et al., 2016). Resende et al., 2008 afirmaram que havendo efeitos de dominância e epistasia em níveis consideráveis, os efeitos de marcadores devem ser re-estimados a fim de manter a acurácia da GWS.

Assim, o impacto da dominância foi no sentido de reduzir a acurácia preditiva que mede a aproximação do valor predito com o valor real, mas não compromete a acurácia

seletiva que mede a superioridade do valor genético aditivo de um indivíduo em relação a outro. Cruz & Carneiro (2003) destacaram a importância de se estudar a variância relacionada efeitos aditivos e não-aditivos em caráter de interesse, de modo a determinar a fração herdável – variância aditiva – e também a fração não herdável – devida aos desvios de dominância, e posteriormente, facilitar a tomada de decisão sobre o método de melhoramento mais eficiente, mas lembram que em um modelo biométrico relacionando o valor genotípico com doses de alelos favoráveis, o coeficiente de regressão é uma medida do efeito médio de substituição alélica que depende das frequências alélicas e dos valores genotípicos *a* e *d*, associados a homozigotos e heterozigotos.

Considerando uma abordagem estatística, a não inclusão de uma matriz de incidência associado aos efeitos de dominância no modelo RR-BLUP reflete diretamente na soma de quadrados do resíduo do modelo. Isso ocorre porque todos os efeitos excluídos do modelo são contabilizados no resíduo, resultando em um quadrado médio de resíduos maior e em decréscimos consideráveis na acurácia preditiva (*REQM*) da metodologia, uma vez que o erro quadrático médio do modelo fica inflacionado (RESENDE et al., 2014).

b) Influência da epistasia sob a eficiência da GWS

Para melhor entendimento da influência dos efeitos epistáticos concentraremos nossa interpretação nos pares de variáveis (V1 e V2), (V3 e V4), (V5 e V6), (V7 e V8), (V9 e V10) e (V11 e V12). Esta análise permite evidenciar o impacto negativo deste fator genético tanto na acurácia preditiva ($REQM$) quanto na acurácia seletiva (r_v^2, CP e Viés) e que a abordagem RR-BLUP adotado foi incapaz de captar a presença destes efeitos dentro do modelo. Deve-se ter em mente que uma fração, que concentra também fatores não herdáveis, da variância genotípica de uma população decorre da expressão, em um dado indivíduo, de efeitos provenientes da combinação de dois ou mais locos, que se refere à interação entre alelos de genes diferentes e constitui a denominada variância epistática (CRUZ & CARNEIRO, 2003). Assim, a epistasia consiste em toda e qualquer interação interalélica. Retomando ao modelo do RR-BLUP/GWS, percebe-se que além de efeitos devidos ao desvio de dominância, as variâncias atribuídas à epistasia também são negligenciadas pela metodologia, uma vez que o modelo contempla somente o termo Xm, que em sua equação representa o espaço de incidência dos efeitos de dose dos marcadores estudados. Mais uma vez concluiu-se que, ao retirar (ou não incluir) esses

efeitos do modelo, os mesmos foram contabilizados no resíduo, resultando em maiores erros quadráticos médios e ineficiência da acurácia preditiva do RR-BLUP.

Verificamos na Tabela 3, e considerando agora as características pré-estabelecidas segundo um modelo epistático, independentemente dos graus médios de dominância e valores de herdabilidade – características (V2), (V4), (V6), (V8), (V10) e (V12) –, a ineficiência da técnica pôde mais uma vez ser observada tendo em vista os valores bem inferiores aos obtidos para as características (V1), (V3), (V5), (V7), (V9) e (V11). Além de quedas significativas nos valores de r^2 e na capacidade preditiva, os valores de $REQM_v$ foram inflacionados ainda mais, e nesse caso variaram entre 99,9704, no melhor cenário (modelo aditivo, sem dominância e com alta herdabilidade) e 968,754 no cenário mais crítico (modelo epistático, com dominância e com baixa herdabilidade).

c) Influência da herdabilidade na eficiência da GWS

Nos programas de melhoramento animal ou de plantas, o desejo maior dos melhoristas consiste em descobrir a variação genética resultante da segregação e ou recombinação dos indivíduos e da interação gênica de modo a determinar a sua influência na próxima geração, no entanto, é o valor fenotípico que pode ser mensurado (CRUZ, 2012). O parâmetro a ser avaliado, nesse caso, deve ser a herdabilidade (h^2), uma vez que o h^2 consiste no parâmetro genético que expressa a variabilidade genética existente na população F1 avaliada, ou seja, a intensidade com que o fenótipo expressa o genótipo (BUENO et al., 2001). A herdabilidade, em seu sentido amplo, pode ser descrita pela equação abaixo:

$$h^2 = \frac{\sigma_G^2}{\sigma_F^2} \tag{15}$$

A herdabilidade é útil também como medida que possibilita auxiliar a percepção dos efeitos ambientais sob a característica de interesse. O aumento da variância ambiental resulta em diminuição da herdabilidade da característica, e ao diminuir a variância ambiental, a herdabilidade aumenta (BUENO et al., 2001).

Para melhor entendimento da influência desses efeitos ambientais sobre a população em estudo, concentraremos nossa interpretação nos pares de características (V1 e V3), (V2 e V4), (V5 e V7), (V6 e V8), (V9 e V11) e (V10 e V12). O primeiro fato a ser considerado é que a abordagem da GWS possibilita obter, de modo geral, valores de r_t^2 superiores aos valores das herdabilidades paramétricas indicando o bom retorno em

aplicação de recursos em genotipagem para obtenção de valores preditos de maior confiabilidade seletiva, entretanto tais aspectos nem sempre resultam em vantagens quando se pensa na necessidade de extrapolação de resultados em populações de validação (Tabela 3). Na a etapa de validação, para nenhuma das características avaliadas foi possível se recuperar os valores das herdabilidades paramétricas avaliadas.

Para h^2=35%, independente dos níveis de dominância e da existência ou não de epistasia, os valores de r_v^2 não superaram 12,5%, ou seja, a metodologia foi capaz de recuperar menos da metade da herdabilidade quando o estudo foi estendido para a população de validação. Mesmo comparando duas características advindas dos cenários mais simples – modelos aditivos, sem presença de dominância – e diferindo apenas segundo a herdabilidade – h^2=35% para V1 e h^2=70% para V3 – nota-se que o aumento de h^2 não foi suficiente para superar a influência do ambiente na confiabilidade seletiva, que nesse caso foi de 5,1% e 24,4%, para V1 e V3, respectivamente. Para o cenário mais complexo – modelo epistático com sobredominância – diferindo apenas segundo a herdabilidade – h^2=35% para V10 e h^2=70% para V12 – nota-se uma queda ainda mais drástica na confiabilidade seletiva, que foi de 2,5% e 14%, para V10 e V12, respectivamente.

Diversos estudos já têm salientado as dificuldades enfrentadas pelos programas de melhoramento quando os caracteres de interesse possuem baixas herdabilidades. No melhoramento animal, por exemplo, Hayes et al. (2009) relataram que o estudo de características de baixa herdabilidade, para levarem a acurácias elevadas, devem ser realizados com grandes populações. Goddard (2009) destacaram que o GEBV está diretamente relacionado ao número de indivíduos da população e à herdabilidade do caráter em estudo.

Para se ter uma ideia comparativa da influência dos três fatores abordados nesse estudo – dominância, epistasia e efeitos ambientais –, tomemos os cenários mais adversos avaliados, considerando a estatística média das características segundo o nível de dominância (ausência, parcial ou sobredominância); epistasia (presença ou ausência) e níveis de herdabilidades (alta ou baixa). Avaliando-se as médias de r^2, $REQM$, CP e Viés obtidas ao considerar os grupos citados (Tabela 4), algumas ressalvas podem ser feitas. A primeira delas refere-se à dominância. Nota-se que, em média, a inclusão de efeitos de dominância não influencia tanto a confiabilidade seletiva, que nessa situação apresenta ligeira melhora, o r_v^2 aumenta de 0,0966 para 0,1239. No entanto, acurácia preditiva sofre grande impacto, e o $REQM$ se torna quatro vezes maior, passando de 127,72 para 525,62.

Tabela 4. Comparação entre os valores médios obtidos via RR-BLUP agrupando as características segundo o nível de dominância; presença ou ausência de epistasia e segundo às duas herdabilidades consideradas, h^2=35% (baixo) e h^2=70% (alto).

Cenário	r_t^2	r_v^2	$REQM_t$	$REQM_v$	CP	Viés
Sem Dominância	0.6051	0.0966	121.65	127.72	0.2701	0.3016
Dominância parcial	0.6851	0.1367	205.39	224.05	0.34199	0.3763
Sobredominância	0.6364	0.1239	500.07	525.62	0.3227	0.3207
Sem Epistasia	0.7162	0.1631	136.20	136.94	0.3807	0.4144
Com Epistasia	0.5682	0.0750	415.20	447.98	0.2425	0.2514
h2 Alto	0.7566	0.1814	249.86	267.97	0.4081	0.4642
h2 Baixo	0.5278	0.0567	301.55	316.95	0.2151	0.2016

r_t^2 : quadrado da correlação de treinamento; r_v^2 : quadrado da correlação de validação; $REQM_t$: erro quadrático médio de treinamento; $REQM_v$: erro quadrático médio de validação; CP: capacidade preditiva

Ao avaliar os efeitos da inclusão da epistasia, percebe-se que o r_v^2 cai aproximadamente pela metade – passando de 0,1631 para 0,0750 –, e o $REQM_v$ triplica – passando de 136,94 para 447,98 –, demonstrando a forte influência que esses efeitos exercem sobre a acurácia do modelo RR-BLUP/GWS adotado.

Na literatura existem trabalhos que destacam as vantagens que a inclusão de efeitos de dominância e epistasia nos modelos estatísticos que envolvam o estudo de caracteres complexos trazem para os programas de melhoramento, tanto para estudos com dados reais (LOPES et al., 2014; MUÑOZ et al., 2014), quanto para estudos com populações simuladas (TORO & VARONA, et al., 2010; ZENG et al., 2013; ALMEIDA FILHO et al., 2016). Almeida Filho et al. (2016) desenvolveram um trabalho para estudar a contribuição da dominância para a predição fenotípica em populações simuladas e concluíram que, à medida que a dominância aumenta, a acurácia preditiva da seleção genômica ampla se torna menos adequada, uma vez que o modelo adotado por essa metodologia não contempla a matriz de efeitos devidos ao desvio de dominância.

Finalmente, observando os valores médios obtidos para as características com altas e baixas herdabilidades, nota-se uma queda significativa na acurácia seletiva do modelo RR-BLUP, uma vez que os valores de r_v^2 foram de 0,1814 e 0,0567 para as herdabilidades de 70% e 35%, respectivamente. A acurácia preditiva também foi afetada com a queda da herdabilidade, resultando no aumento do $REQM_v$.

De modo geral, o parâmetro que exerceu maior influência na acurácia seletiva do RR-BLUP foi o efeito ambiental, uma vez que os menores valores médios de r_v^2 e CP foram obtidos ao considerar as características de baixa herdabilidade (V1, V2, V5, V6, V9 e V10). As interações intralélicas (V9, V10, V11 e V12) e interalélicas (V2, V4, V6,

V8, V10 e V12), por sua vez, acarretaram em quedas bastante significativas na acurácia preditiva, dados os elevados valores de $REQM_v$ (Tabela 4).

d) Possível influência da dimensionalidade na eficiência da GWS

Todos os fatores mencionados anteriormente que proporcionam diferenças na predição de valores genéticos tem direta, ou indiretamente, influência de fatores genéticos. A solução para alguns destes problemas pode vir de algumas ações na condução experimental, na parametrização do modelo, no tipo de família utilizado na genotipagem, dentre outros (CRUZ et al., 2012). Dois fatores adicionais importantes, de fundamentação analítica, devem ser mencionados e relacionam-se às questões da dimensionalidade do estudo e a abordagem estocástica empregada.

Quanto ao problema da dimensionalidade das matrizes envolvidas, deve ser destacado que a população estudada era constituída de 500 indivíduos genotipados com 1000 marcadores moleculares. Note que, nesse caso, o número de marcas era maior que o número de indivíduos avaliados, implicando em problemas de estimação, pela dimensionalidade da matriz Xm, além de possibilitar a existência de marcadores altamente correlacionados – o que caracteriza a existência de multicolinearidade (AZEVEDO et al., 2014). Embora o RR-BLUP considere o uso do parâmetro de penalização λ em sua modelagem, este não foi eficiente para contornar o problema de dimensão do trabalho, dados os baixos valores de r_v^2 e CP, e elevados valores de erro quadrático médio encontrados (Tabela 3). Problemas relacionados à dimensionalidade em seleção genômica ampla já foram relatados por diversos autores (AZEVEDO, 2012; JAMES et al., 2013; AZEVEDO et al., 2014). Azevedo et al. (2014) afirmaram que o elevado número de marcadores utilizados para estimar o mérito genético de indivíduos representa um grande desafio prático dada a dificuldade que de se encontrar um modelo funcional adequado para a estimação dos efeitos de marcas. Para solucionar tal problema, esses autores propuseram o uso de métodos de redução de dimensionalidade para praticar a seleção genômica em características de carcaça em porcos. James et al. (2013) discutiram os problemas advindos das altas dimensionalidades consideradas em seleção genômica ampla – tais como viés da variância, *overfitting* e multicolinearidade – e destacaram a possibilidade de utilização de procedimentos de seleção de subamostras, métodos de penalização ou os métodos de redução propriamente ditos, sugerindo algumas metodologias, tais como: Regressão *Stepwise*, Componentes Principais Parciais, Quadrados Mínimos Parciais e Lasso Bayesiano como alternativas.

Quanto ao problema referente a abordagem estocástica empregada devemos destacar que existem inúmeros modelos, alguns baseados em modelos lineares para os quais efeitos de marcas são assumidos como advindos de uma distribuição normal – como no caso do BLUP e do RR-BLUP – outros com enfoque bayesiano, como BayesA, por exemplo, para o qual o modelo é não linear e efeitos de marcas são regressados com variâncias específicas (RESENDE et al., 2012), de modo a se adequar a diferentes situações referentes a pressupostos de ação gênica para prover melhores resultados.

Além de recorrer a outras metodologias de GWS, há também que se considerar o uso de novos paradigmas tais como empregados em Redes Neurais Artificias, na busca de melhores soluções. Estudos de aplicação das Redes Neurais Artificiais no melhoramento já tem demonstrado o grande potencial dessa metodologia em estudos de predição (SILVA et al., 2014; SILVA et al., 2016); classificação (SANT'ANNA et al., 2015; CARNEIRO et al., 2017); estabilidade e adaptabilidade (NASCIMENTO et al., 2013), e até mesmo estudos de seleção genômica (GIANOLA et al., 2011; SILVA et al., 2017).

Os resultados obtidos nesse estudo atentam também para o desafio enfrentado pela Seleção Genômica Ampla ao estudar populações que possuem um elevado número de marcadores moleculares – excedendo assim o número de indivíduos da população. Essa poderia ser uma justificativa para os baixos valores de acurácia seletiva (r^2 e CP) e elevados valores de erro quadrático médio obtidos. Tais resultados revelam a necessidade de estudos posteriores envolvendo técnicas de seleção de variáveis e redução de dimensionalidade conjuntamente com a GWS de modo a contribuir para o aumento da acurácia da mesma.

CONCLUSÕES

A inclusão de marcadores moleculares e a abordagem fundamentada em GWS proporcionaram valores preditos com acurácia seletiva superior a herdabilidade em conjuntos de dados de treinamento, porém a validação do modelo apresenta considerável redução de eficiência de predição.

Foi evidenciado o efeito perturbador da dominância sobre a acurácia preditiva, mas sem comprometimento sobre a acurácia seletiva.

O modelo usual de GWS, baseado em RR-BLUP, não foi capaz de captar efeitos epistáticos e teve eficiência comprometida com a presença de tais efeitos.

Os efeitos elevados do ambiente, reduzindo a herdabilidade do caráter, é o fator de influência na expressão genética que mais reduz a eficiência da GWS.

REFERÊNCIAS

AKAIKE, H. A new look at the statistical model identification **IEEE Transaction on Automatic Control**. v.19, p.716-723, 1974.

ALLARD, R. W. **Princípios de melhoramento genético das plantas**. São Paulo: Edgard Blücher, 381p, 1971.

ALMEIDA FILHO, J.E. de; GUIMARÃES, J.F.R.; SILVA, F.F.; RESENDE, M.D.V.; Muñoz, P.; KIRST, M.; and RESENDE Jr, M.F.R. The contribution of dominance to phenotype prediction in a pine breeding and simulated population. **Heredity**, v.117, p.33-41, 2016.

AZEVEDO, C.F.; SILVA, F.F.; RESENDE, M.D.V.; LOPES, M.S.; DUIJVESTEIJN, N.; GUIMARAES, S.E.F.; LOPES, P.S.; KELLY, M.J.; VIANA, J.M.S.; KNOL, E.F. Supervised independent component analysis as an alternative method for genomic selection in pigs. **J. Anim. Breed. Genet**. 131, 452–461, 2014.

BUENO, L.C.S.; MENDES, A.N.G.; CARVALHO, S.P. **Melhoramento genético de plantas: princípios e procedimentos**. Lavras: UFLA, 2001. 282p.

CRUZ, C.D. Genes Software – extended and integrated with the R, Matlab and Selegen. **Acta Scientiarum. Agronomy**. Maringá, v. 38, n. 4, p. 547-552, Oct.-Dec., 2016.

CRUZ, C.D. **Princípios de genética quantitativa**. Viçosa: Ed. da UFV, 2012. 394p.

CRUZ, C.D.; REGAZZI, A.J.; CARNEIRO, P.C.S. (2012). **Modelos biométricos aplicados ao melhoramento genético**. UFV, Viçosa.

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

GIANOLA, D.; OKUT, H.; KENT A WEIGEL, K.A.; ROSA, G J.M. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. **BMC Genetics.** 12:87, 2011.

GODDARD, M. Genomic selection: prediction of accuracy and maximization of long term response. **Genetica**, v. 136, n. 2, p. 245–57, 2009.

HAYES, B. J.; BOWMAN, P. J.; CHAMBERLAIN, A. J.; GODDARD, M. E. Genomic selection in dairy cattle: progress and challenges. **Journal Dairy Science**, v. 92, n. 1, p. 433–443, 2009.

JAMES, G.; WITTEN, D.; HASTIE; T.; TIBSHIRANI, R. **An Introduction to Statistical Learning with Applications in R**. 2013.

LONG, N.; GIANOLA, D.; ROSA, G.J.; WEIGEL, K.A. Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. **J Anim. Breed. Genet.** Aug; 128(4): 247-57, 2011.

LOPES M.S.; BASTIAANSEN J.W.M.; HARLIZIUS B.; KNOL E.F.; BOVENHUIS H. A genome-wide association study reveals dominance effects on number of teats in pigs. **PLoS One** 9: e105867, 2014

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

MUÑOZ PR, RESENDE MFR, GEZAN SA, RESENDE MDV, DE LOS CAMPOS G, KIRST M et al. (2014). Unraveling additive from non-additive effects using genomic relationship matrices. **Genetics** 198: 1759–1768.

R CORE TEAM. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2017. Available at: <https://www.R-project.org/>.

RESENDE, M.D.V.; SILVA, F.F.; AZEVEDO, C.F. **Estatística matemática, biométrica e computacional: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção**

Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição Sobrevivência. Viçosa: Suprema, 881p. 2014.

RESENDE, M.D.V.; LOPES, P.S.; SILVA, R.L.; PIRES, I.E. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, Colombo, n.56, p.63-77, jan./jun. 2008.

SANT'ANNA, I.C.; TOMAZ, R.S.; SILVA, G.N.; NASCIMENTO, M.; BHERING, L.L.; CRUZ, C.D. Superiority of artificial neural networks for a genetic classification procedure. **Genetics and Molecular Research**, v.14, p.9898-9906, 2015. DOI: 10.4238/2015. August.19.24.

SILVA, G.N.; TOMAZ, R.S.; SANT'ANNA, I.C.; CARNEIRO, V.Q.; CRUZ, C.D.; NASCIMENTO, M. Evaluation of the efficiency of artificial neural networks for genetic value prediction. **Genetic Molecular Research**, v.15, p.1-11, 2016. DOI: 10.4238/ gmr.15017676, 2016.

SILVA, G.N.; TOMAZ, R.S.; SANT'ANNA, I. de C.; NASCIMENTO, M.; BHERING, L.L.; CRUZ, C.D. Neural networks for predicting breeding values and genetic gains. **Scientia Agricola**, v.71, p.494-498. DOI: 10.1590/0103- 9016-2014-0057. 2014.

SILVA, G.N. **Redes neurais artificiais: novo paradigma para a predição de valores genéticos**. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, 105p, 2014.

TORO, M.A.; VARONA, L. A note on mate allocation for dominance handling in genomic selection. **Genet Sel Evol**. 42: 33, 2010.

ZENG, J.; TOOSI, A.; FERNANDO, R.L.; DEKKERS, J.C.M.; GARRICK, D.J. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. **Genet Sel Evol**. 45: 11. 2013.

CAPÍTULO 2

REDUÇÃO DE DIMENSIONALIDADE COMO ESTRATÉGIA NA SELEÇÃO GENÔMICA PARA FINS DE MELHORIA DA EFICIÊNCIA PREDITIVA

Resumo
O objetivo deste trabalho foi propor duas metodologias de redução de dimensionalidade a fim de solucionar problemas de dimensionalidade e multicolinearidade enfrentados em Seleção Genômica Ampla. Os métodos de Stepwise e da Sonda foram utilizados para reduzir a matriz de inicial de marcadores da população simulada F1 em estudo. Com as novas matrizes de marcas, procedeu-se a análises genômicas por meio do método RR-BLUP na predição de valores genéticos para as doze características quantitativas avaliadas, que contemplavam diferentes estruturações quanto à modelo, graus médios de dominância e herdabilidade. Os resultados indicaram que o uso de métodos de redução de dimensionalidade das matrizes de marcas em detrimento ao uso da matriz de marcas original aumenta a eficiência do método RR-BLUP. No entanto, os resultados evidenciam que mesmo com a redução, o modelo do RR-BLUP ainda sofre decréscimos consideráveis nas acurácias preditivas e seletivas de características complexas que incluam efeitos ambientais e de interações intra e interalélicas bem pronunciados.
Termos para indexação: Seleção genômica, RR-BLUP, redução de dimensionalidade.

Abstract
The objective of this work was to propose two methodologies of dimensionality reduction in order to solve dimensionality and multicollinearity issues in Wide Genomic Selection. Stepwise and Sonda methods were used to reduce the initial order of markers matrix of the F1 simulated population under study. With the new matrices of markers, genomic analyzes were performed using the RR-BLUP method in the prediction of genetic values for the twelve quantitative traits evaluated, which included different model structures, mean degrees of dominance and heritability. The results indicated that the use of dimensionality reduction methods of markers matrices in detriment to the use of the original markers matrix increases the efficiency of the RR-BLUP method. However, the results show that even with the reduction, the RR-BLUP model still undergoes considerable decreases in predictive and selective accuracy of complex features that include environmental effects and well-pronounced inter- and intra-allelic interactions.
Index terms: Genome Selection, Ridge Regression, Dimensionality reduction.

INTRODUÇÃO

A seleção genômica ampla (*Genome Wide Selection – GWS*), proposta por Meuwissen (2001), revolucionou os programas de melhoramento genético pois permitiu a incorporação direta da informação de grande número de marcadores moleculares para a predição de valores genômicos (CRUZ et al., 2013). A partir daí diversas metodologias de GWS foram implementadas, tais como G-BLUP, RR-BLUP, Bayes A e B (MEUWISSEN et al., 2001), LASSO bayesiano (CAMPOS et al., 2009), dentre outras, diferindo umas das outras acerca das pressuposições sobre os efeitos dos marcadores, estabelecimento das *prioris* ou pelos parâmetros de penalização utilizados (RESENDE et al., 2014). No entanto, o elevado número de marcadores moleculares, além de prejudicar a capacidade preditiva dessas metodologias, acarreta em problemas de dimensionalidade e multicolinearidade. Os problemas de dimensão surgem porque a possibilidade de

mapear todo o genoma do indivíduo ou população de interesse faz com que o número de marcadores moleculares seja quase sempre muito maior que o número de indivíduos avaliados, o que dificulta a modelagem estatística do estudo (CRUZ, 2013). Os problemas de multicolinearidade se dão quando há existência de alta correlação entre as marcas avaliadas (AZEVEDO et al., 2014).

Para solucionar tais problemas, diversos autores têm proposto o uso de métodos estatísticos de redução de dimensionalidade (RESENDE et al., 2012; AZEVEDO, 2012; JAMES et al., 2013; AZEVEDO et al., 2014). James et al., 2013 propuseram três linhas metodológicas de redução, que podem ser baseadas na seleção de subamostras de variáveis, em métodos de penalização e em métodos de redução propriamente ditos via combinações lineares não correlacionadas, e destacaram que altas dimensionalidades podem acarretar em problemas de *overfitting*, multicolinearidade e alto viés da variância. Azevedo et al., 2014 utilizaram o método de redução de componentes independentes para seleção genômica em suínos e verificaram aumentos consideráveis na acurácia preditiva. Pantalião et al. (2016) concluíram que utilizar o Método *Stepwise* para selecionar SNPs sem efeitos de sobreposição em um estudo de associação genômica ampla para produtividade em arroz auxilia no processo de predição.

Além das metodologias estatísticas de redução listadas acima, uma técnica bastante utilizada, principalmente nas subáreas da ciência da computação, é o algoritmo genético. O algoritmo genético (AG), proposto por Holland (1975), consiste numa metodologia que baseia nos conceitos de seleção natural e de processos de evolução (JANG, 1997). Jang (1997) justifica a grande utilização desse algoritmo em alguns fatores: O AG não depende de modelos funcionais (é não paramétrico); pode ser aplicado tanto para o estudo de variáveis discretas quanto contínuas, e ainda, a flexibilidade do algoritmo facilita a estruturação e identificação de parâmetros em metodologias complexas, como Redes Neurais Artificiais e Sistemas *Neuro Fuzzy*.

Inspirado nesse algoritmo, propõe-se nesse estudo uma metodologia baseada na seleção de variáveis, a qual denominaremos Método da Sonda para fins de redução de dimensionalidade, como alternativa para os métodos convencionais de redução ou seleção de variáveis. Avaliaremos também a eficiência do método de seleção de variáveis *Stepwise*, dada sua consolidação e eficiência já verificada em outros estudos (JAMES et al., 2013; PANTALIÃO et al., 2016).

Diante do exposto, o objetivo desse trabalho foi avaliar os métodos de redução de dimensionalidade baseados em seleção de variáveis, a saber, *Stepwise* e Sonda, como

alternativas para melhorar as acurácias seletivas e preditivas do método de RR-BLUP para seleção genômica ampla.

MATERIAL E MÉTODOS

1. População avaliada

A população avaliada nesse capítulo foi exatamente a mesma população considerada no capítulo anterior. Portanto, a população simulada F1 oriunda do cruzamento de duas populações contrastantes, também obtidas via simulação, cujas informações genotípicas e fenotípicas foram descritas no capitulo anterior, foi objeto desse estudo. Foram avaliadas doze características quantitativas, com diferentes estruturações quanto à modelo, graus médios de dominância e herdabilidade para a população F1, constituída por 500 indivíduos e genotipada com 1000 marcadores moleculares. A Tabela 1 constitui a mesma tabela apresentada no capítulo 1, e segue apresentada novamente abaixo, a fim de estabelecer a mesma notação utilizada anteriormente.

Tabela 1. Características avaliadas no estudo com seus respectivos valores de herdabilidade, modelo adotado e grau médio de dominância (gmd).

Característica	Herdabilidade (%)	Modelo	gmd
V1 - D0H35_Ad	35	aditivo	0
V2 - D0H35_Ep	35	epistático	0
V3 - D0H70_Ad	70	aditivo	0
V4 - D0H70_Ep	70	epistático	0
V5 - D60H35_Ad	35	aditivo	0,6
V6 - D60H35_Ep	35	epistático	0,6
V7 - D60H70_Ad	70	aditivo	0,6
V8 - D60H70_Ep	70	epistático	0,6
V9 - D120H35_Ad	35	aditivo	1,2
V10 - D120H35_Ep	35	epistático	1,2
V11 - D120H70_Ad	70	aditivo	1,2
V12 - D120H70_Ep	70	epistático	1,2

2. Métodos de redução de dimensionalidade avaliados

Como visto no capítulo anterior, a dimensão da matriz de marcadores utilizada para genotipar a população F1 em estudo era superior que o numero de indivíduos, uma

vez que a análise será realizada com informações de 1000 marcadores genotipados em 500 indivíduos avaliados. Assim, com a finalidade de solucionar problemas dessa magnitude e/ou existência de multicolinearidade entre marcas, a literatura tem proposto o uso de técnicas de redução de dimensionalidade, que podem ser baseadas em estratégias de penalização (por regressão em cumeeira, por exemplo), de redução da dimensionalidade (por componentes principais, por exemplo) ou em metodologias que envolvam a seleção de subamostras de variáveis (métodos de exclusão) (JANG, 1997; AZEVEDO et al., 2014; PANTALIÃO, et al., 2016, CAO et al., 2003).

Neste capítulo, daremos enfoque a duas metodologias fundamentadas em subamostras (seleção de variáveis), a saber: o método já consolidado de Regressão *Stepwise* e um método proposto e implementado nesse estudo que denominamos Método da Sonda. Estes métodos foram utilizados em duas etapas, sendo a primeira utilizada para estabelecer o tamanho adequado da subamostra e a segunda para identificar quais marcadores seriam os mais importantes para compor a subamostra final a ser utilizada na predição pelas abordagens utilizadas neste estudo.

2.1 Estabelecimento do número de marcadores selecionados

Para estabelecer o número ótimo de marcadores que seriam inseridos nas matrizes reduzidas de ambas as metodologias, utilizou-se a característica mais complexa (V10) avaliada no estudo. A partir dela foram avaliados os valores de R^2, REQM, já definidos anteriormente, e o valor NC – número de condição – obtidos considerado a inclusão de 1 a 1000 marcas no modelo.

O NC consiste em uma técnica para identificar a existência de multicolinearidade na matriz de variáveis avaliadas, por meio da análise dos autovalores da matriz X^TX, em que X representa as variáveis independentes do modelo. Assim, o valor de NC é dado pela equação (1) abaixo:

$$NC = \frac{\max(\lambda_1,\lambda_2,\ldots,\lambda_p)}{\min(\lambda_1,\lambda_2,\ldots,\lambda_p)} \quad (1)$$

Em que:

λ_i: autovalor da i-ésima variável, i=1, ..., p

A decisão é tomada avaliando-se o valor obtido para NC, tal como sugerido por Montgomery & Peck (1981), de modo que:

- Se $NC < 100$ então não existem problemas de multicolinearidade;
- Se $100 < NC < 1000$ conclui-se que a multicolinearidade é moderada a forte;
- Se $NC > 1000$ tem-se acentuada multicolinearidade.

2.2 Método da Regressão *Stepwise*

Em problemas de regressão, em detrimento ao modelo completo – constituído pelas p variáveis independentes –, o pesquisador pode lançar mão de um modelo reduzido, uma vez que bom modelo não precisa necessariamente incluir todas as variáveis disponíveis. O importante, nesse caso, é estabelecer um modelo que inclua as variáveis que mais contribuem para a eficiência do modelo (SCHUSTER & CRUZ, 2013).

O Método da *Stepwise* consiste numa metodologia linear de seleção de variáveis, muito utilizada nas ciências da computação, e que apresenta grande potencial de aplicação em estudos de seleção genômica, uma vez que o elevado número de marcadores moleculares utilizados nos modelos preditivos de GWS tornam inviável a avaliação de todos os modelos possíveis (SCHUSTER & CRUZ, 2013). Além de tentar solucionar problemas computacionais e estatísticos de análises que envolvam um número muito elevado de variáveis, esse método visa selecionar as variáveis que mais influenciam o conjunto de saída, de modo a reduzir o modelo de regressão (JAMES, et al., 2013).

O *Stepwise* combina dois procedimentos: *Forward* – com a inclusão de variáveis – e *Backward* – com a exclusão de variáveis do modelo. A cada passo *forward*, avalia-se a possibilidade de inclusão de uma variável no modelo considerando, inicialmente, a maior correlação simples e, posteriormente, a maior magnitude da correlação parcial entre a variável pretendente e a variável resposta, removendo os efeitos das demais já presentes no modelo. A inclusão desta variável é ratificada por meio de uma estatítica F parcial e um nível de significância de probabilidade de entrada associado a um p-valor menor que a probabilidade de entrada e/ou saída fixada em 10%.

O passo *Backward,* que ocorre após uma variável ter entrado no modelo, consiste em certificar sobre a manutenção ou exclusão de qualquer outra variável pertencente ao modelo, utilizando também uma estatística F parcial e outro nível de significância referente a probabilidade de saída ou exclusão. Desse modo, a variável adicionada no passo anterior pode, ou não, ser importante para o modelo, dada sua correlação parcial com as demais variáveis e, assim, dependendo da sua estatística (F parcial), ela pode ser removida do modelo (*backward*).

O ajustamento do modelo pode ser mensurado por meio de estatísticas tais como: capacidade preditiva (*CP*), o coeficiente de determinação associado ao modelo (R^2), índice de Akaike (*AIC*) ou pelo Critério de Informação Bayesiano (*BIC*) (RESENDE et al., 2012; JAMES et al., 2013).

Como descrito acima, para esse estudo adotamos as correlações simples e parciais para determinar quais variáveis entrariam ou não no modelo final. O coeficiente de determinação simples e parcial (R^2 e $R^2_{parcial}$, respectivamente) são definidos pelas equações (2) e (3):

$$R^2 = \frac{SQR}{SQtot}.100 \quad (2)$$

e

$$R^2_{parcial} = \frac{SQR}{SQDr}.100 \quad (3)$$

Em que:

SQR: Soma de quadrados da regressão do modelo completo;

$SQtot$: Soma de quadrados total da análise de regressão.

SQDr: Soma de quadrado do desvio do modelo reduzido que não inclui a variável que está sendo considerada a inclusão no modelo.

2.3 Método da Sonda

O método da Sonda desenvolvido e proposto neste trabalho consiste num método de seleção de variáveis, inspirado no Algoritmo Genético (HOLLAND, 1975).

Como dito anteriormente, para muitos estudos é inviável testar todos os possíveis modelos. Para esse estudo, por exemplo, a população F1 era genotipada com 1000 indivíduos e a matriz de marcas foi reduzida para 150 marcadores moleculares com base em procedimento descrito no item 2.1 anterior. Se fôssemos testar todas as possíveis combinações, teríamos C^{150}_{1000} possibilidades resultando em demanda de tempo e recurso computacional muito grande. Assim, optamos por utilizar o método da sonda que envolve duas etapas, referente ao estabelecimento de um conjunto finito de soluções seguido da recombinação das melhores soluções para obtenção de uma solução única.

Consideramos na primeira etapa que, ao invés de testar todas as possíveis regressões, um número n de subamostras ou de "sondas" poderia ser suficiente para explorar as possibilidades de solução sendo que cada sonda, que representa uma amostra de variáveis explicativas, é tomada aleatoriamente no conjunto de variáveis original.

Nesse estudo foram consideradas arbitrariamente 20.000 subamostras (sondas) envolvendo n variáveis que foram utilizadas para gerar as regressões de Y cujos efeitos foram estimados e o ajuste do modelo foi quantificado. Deve ser ressaltado que o número *n* de variáveis que constituirão cada sonda dever ser pré-estabelecido arbitrariamente ou por ajuste considerando algum critério de otimização, conforme descrito em 2.1.

Assim, ao final da primeira etapa calcula-se um índice que indica a importância relativa (IR_{xi}) de cada variável conforme expressão (4) abaixo:

$$IR_{x_i} = \frac{1}{k}\sum_{i=1}^{k} \beta_i R_i^2, \text{ para i = 1,2, ..., m} \quad (4)$$

Em que:

k: número de vezes que uma determinada variável xi participou das 20000 sondagens efetuadas;

β_i: coeficiente da regressão associado à variável x_i incluída no modelo de regressão obtido numa determinada sonda S_i que tenha sido incluída;

R_i^2: coeficiente de determinação obtido pela regressão ajustada na sonda S_i.

m: número total de marcas estudadas (no estudo, igual a 1000)

A segunda etapa no procedimento de sonda proposto consiste na recombinação dos resultados, elegendo, para uma nova regressão, aquelas de melhor desempenho. Isto é feito a partir do ranqueamento fornecido pelo índices IR_{x_i}, que permite selecionar as n melhores variáveis para o calculo da nova regressão múltipla e do novo coeficiente R_{IR}^2 ou, que serão utilizadas em modelos de seleção genômica ampla para fins de avaliação de sua eficácia em processo preditivos.

3. Método estatístico empregado para fins de predição por seleção genômica ampla

Para tornar os resultados obtidos após a redução de dimensionalidade comparáveis com os resultados obtidos ao considerar a matriz original de 1000 marcadores moleculares, nesse capítulo procedeu-se a uma análise de seleção genômica ampla análoga à apresentada no capítulo 1. Portanto, adotou-se a metodologia de RR-BLUP (*Ridge Regression-Best Linear Unbiased Prediction*), tal como descrito no capítulo 1, considerando-se uma validação cruzada (*k-fold*) com k = 5 partições, ou seja, a população, agora com a matriz reduzida de marcas, foi particionada em cinco subconjuntos

mutuamente exclusivos e a cada rodada quatro desses subconjuntos constituíram a população de treinamento (totalizando 80% dos indivíduos) e o subconjunto restante constituiu a população de validação (20% da população total).

4. Medidas de eficiência do procedimento biométrico

Para avaliar as acurácias preditivas e seletivas da metodologia proposta, os mesmos parâmetros abordados no capítulo 1 foram utilizados. Assim, o trabalho contemplou a avaliação da eficiência do RR-BLUP/GWS por meio das estatísticas: raiz do erro quadrático médio ($REQM$); (r_t^2) e (r_v^2), correspondentes ao quadrado da correlação para as fases de treinamento e de validação, respectivamente e a capacidade preditiva do modelo (CP), cujas definições foram apresentadas no capítulo anterior. Adicionalmente avaliou-se o critério de informação de Akaike (*AIC*) definido a seguir.

O *AIC*, proposto por Akaike (1974), consiste num critério de seleção de modelos, que utiliza a Informação de Kullback-Leibler (K-L) para testar se um dado modelo é adequado. O critério adotado para a obtenção do *AIC* tenta escolher o modelo que minimize a divergência de Kullback-Leibler (K-L), de modo que menores valores do AIC indicam um melhor ajuste global (AKAIKE, 1974). O AIC é definido como:

$$AIC = -2\log\left(L(\hat{\theta})\right) + 2p \qquad (5)$$

Em que:

$L(\hat{\theta})$: máximo da função de verossimilhança do modelo considerado;

p: número de parâmetros a serem estimados no modelo.

5. Recursos Computacionais

O procedimento de simulação da população avaliada foi realizado no software GENES (CRUZ, 2016), tal como descrito no capítulo 1.

O procedimento para determinar o tamanho ideal da matriz reduzida de marcadores, e a metodologia de Sonda proposta foi desenvolvida no software MATLAB (MATLAB, 2011). Os métodos de redução de Regressão *Stepwise* e Sonda foram implementados no software GENES (CRUZ, 2016), no módulo integração com o software MATLAB (MATLAB, 2011).

O Método do RR-BLUP foi implementado no GENES, no módulo integração com o software R (R Core Team, 2018).

RESULTADOS E DISCUSSÃO

a) Determinação do número de marcadores selecionados

Antes de proceder às analises com as matrizes reduzidas, foi necessário determinar o número de marcadores moleculares a serem considerados nas novas matrizes. Desse modo, para determinar um número ideal de marcadores que satisfizesse ambas as metodologias propostas, avaliaram-se os valores de r^2, *REQM* e *NC* obtidos na avaliação da característica mais complexa do estudo (V10) após a inclusão de cada uma das 1000 marcas no modelo. V10 constitui a característica mais complexa pois é afetada fortemente pelos três fatores perturbadores no estudo, uma vez que foi obtida de um modelo epistático e, além disso, apresenta alto grau de dominância e baixa herdabilidade (CRUZ, 2010).

Os resultados obtidos para o Método *Stepwise* estão apresentados nas Figuras 1 e 2.

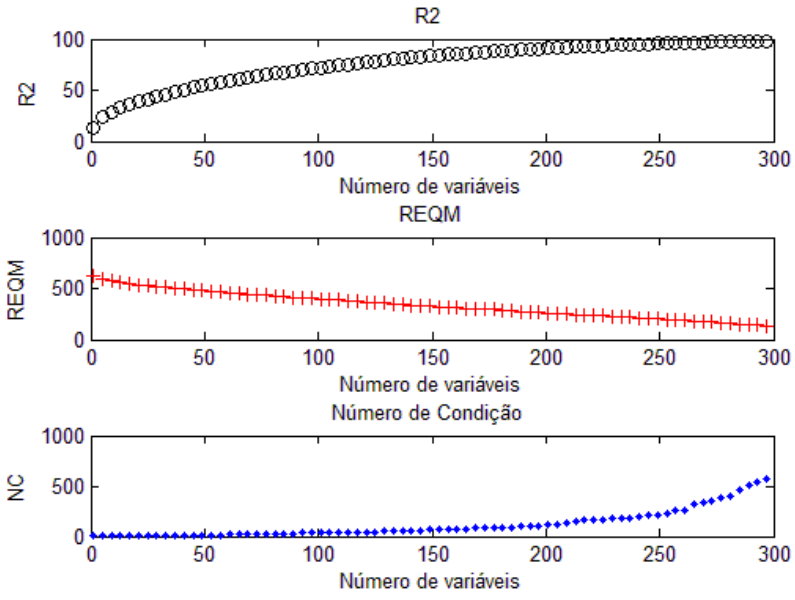

Figura 1. Representação gráfica dos valores de r^2, *REQM e NC* – gráficos preto, vermelho e azul, respectivamente – obtidos pelo Método de Stepwise. O eixo x representa o número de variáveis (marcadores moleculares) incluídos no modelo de regressão múltipla.

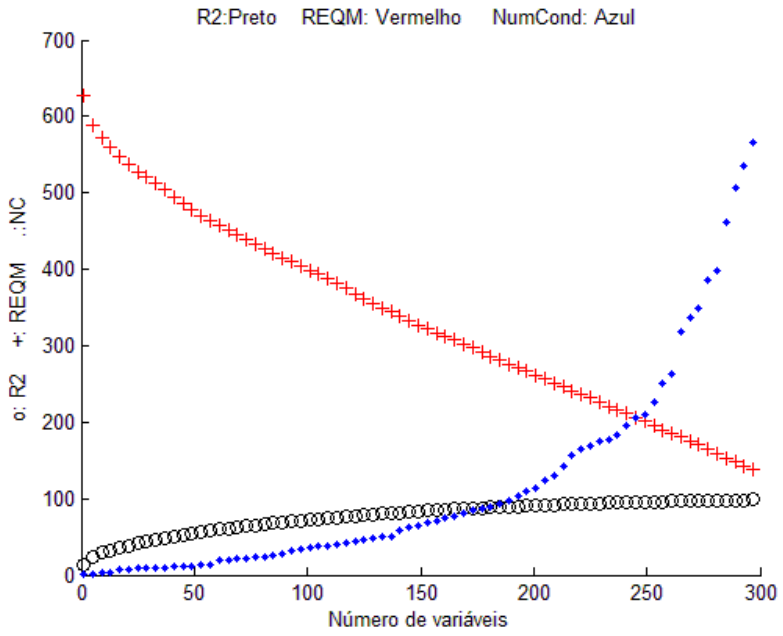

Figura 2. Representação gráfica conjunta dos valores de r^2, *REQM e NC* obtidos pelo Método de Stepwise. O eixo x representa o número de variáveis (marcadores moleculares) incluídos no modelo de regressão múltipla.

Ao observar os gráficos, nota-se que para essa metodologia, considerar a inclusão de 150 a 200 marcadores no modelo seria razoável, uma vez que nesse intervalo tem-se $NC < 100$, os valores do R^2 tendem a estabilizar e os valores de REQM caem substancialmente.

Avaliando agora os resultados obtidos para o Método da Sonda (Figuras 3 e 4), nota-se que o indicado também seria considerar a inclusão de um número de marcadores entre 150 e 200, uma vez que, para esse método, também foram obtidos $NC < 100$, valores estabilizados de R^2 e quedas substanciais no valor do REQM no intervalo sugerido. Desse modo, optou-se por aplicar os métodos de regressão *Stepwise* e Sonda visando a seleção de 150 marcadores moleculares.

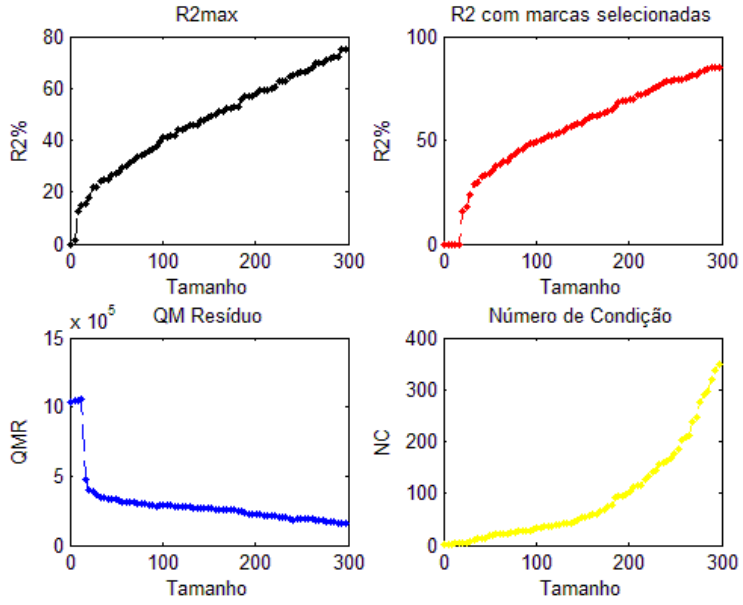

Figura 3. Representação gráfica dos valores de r^2max, r^2, *REQM e NC* – gráficos preto, vermelho, azul e amarelo, respectivamente – obtidos pelo Método da Sonda. O eixo x representa o número de variáveis (marcadores moleculares) incluídos no modelo de regressão múltipla.

b) Influência da dimensionalidade na eficiência da GWS

Como apresentado no decorrer desse capítulo, o elevado número de marcadores moleculares considerados em estudos de seleção genômica ampla tem levado a problemas computacionais, estatísticos e também de predição (SCHUSTER & CRUZ, 2013). Visando contornar tais transtornos, foi proposto nesse trabalho o uso de duas metodologias de redução de dimensionalidade, ambas baseadas na seleção de variáveis, almejando que a redução, além de aumentar a eficiência do método RR-BLUP avaliado, fosse capaz de eliminar eventuais problemas de multicolinearidade (AZEVEDO et al., 2014).

Para avaliar se a redução de dimensionalidade foi capaz de melhorar a eficiência do RR-BLUP/GWS, foi feito um histograma que compara os resultados de r^2 e $REQM$ obtidos para o RR-BLUP utilizando 1000 marcadores em comparação ao uso do método após os dois procedimentos de redução serem utilizados (Figuras 4 e 5).

Como pode ser observado, a eficiência do RR-BLUP apresentou aumentos bastante significativos. A acurácia seletiva dos modelos reduzidos (com 150 marcas) foi muito superior que a acurácia seletiva do modelo completo (com 1000 marcadores), uma

vez que os valores de r^2 se tornaram até cinco vezes maiores para algumas características após a redução (Figura 4). Para a característica V12, por exemplo, o quadrado da correlação era de 0.14 no modelo com 1000 marcadores moleculares, e passou para 0.7506 no modelo reduzido via Método *Stepwise*.

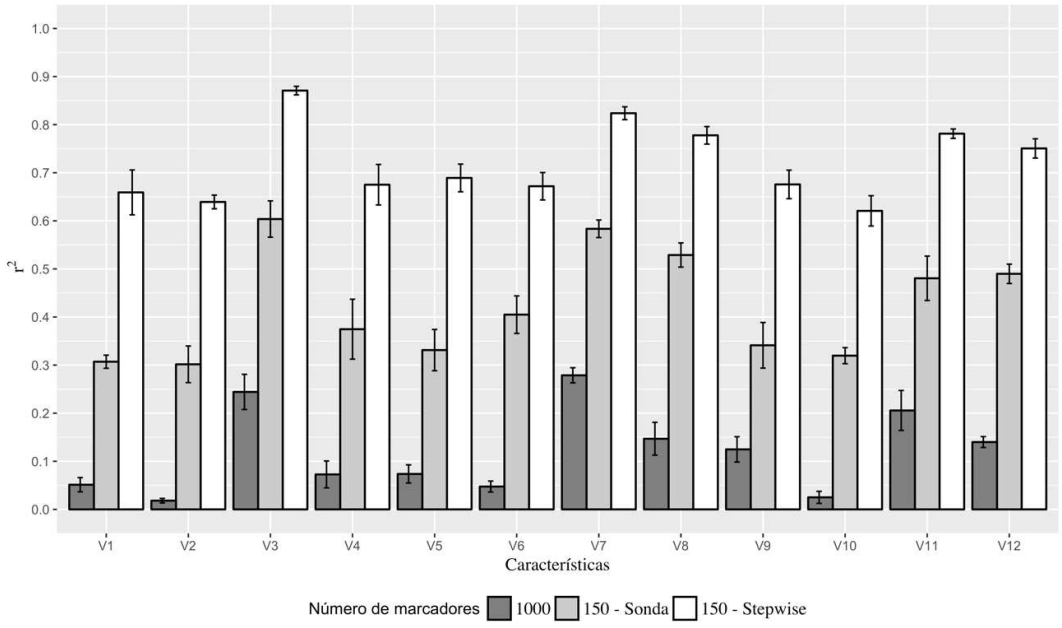

Figura 4. Histograma dos resultados obtidos para a estatística r^2 , obtida em conjunto de dados de validação, pela metodologia RR-BLUP ao utilizar três estratégias: análise com a matriz original de 1000 marcadores; análise com a matriz de reduzida de 150 marcadores selecionados por meio do Método da Sonda (150 - Sonda) e utilizando a matriz de reduzida de 150 marcadores selecionados por meio do Método da regressão Stepwise (150 - *Stepwise*).

Com relação à acurácia preditiva, pôde-se observar que a redução de dimensionalidade proporcionou quedas bastantes significativas nos valores do erro quadrático médio (Figura 5). Ao considerar a mesma variável V12 citada anteriormente, verifica-se que a raiz do erro quadrático médio foi reduzida de 792,08 para 553,06 ao utilizar o Método de Regressão *Stepwise*, e reduzida para 468,19 ao utilizar o Método da Sonda.

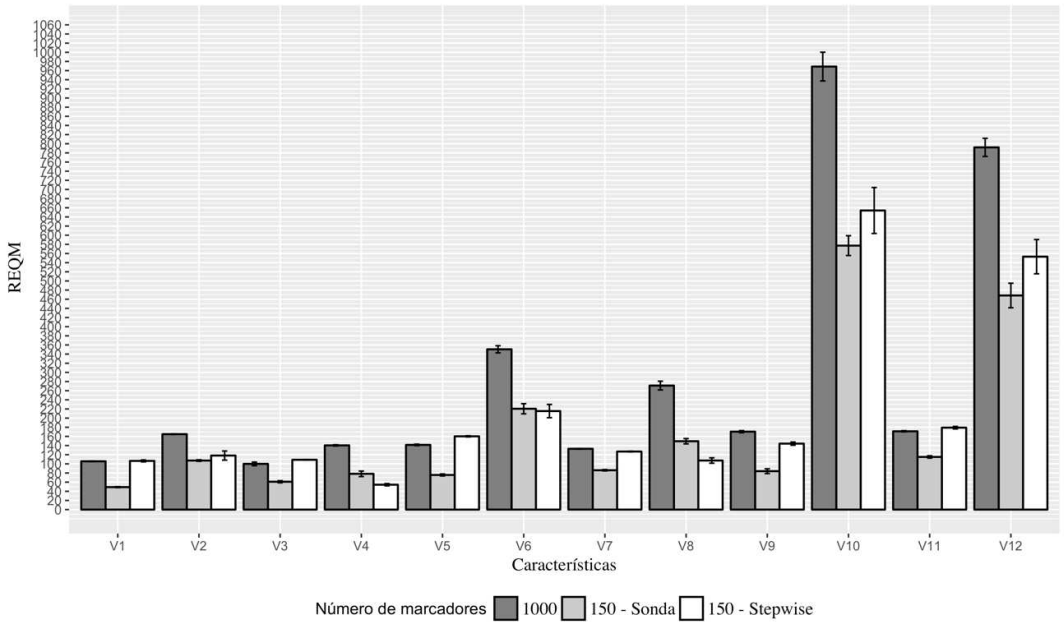

Figura 5. Histograma dos resultados obtidos para o parâmetro $REQM$ de validação pela metodologia RR-BLUP ao utilizar três estratégias: utilizando a matriz original de 1000 marcadores; utilizando a matriz de reduzida de 150 marcadores selecionados por meio do Método da Sonda (150 - Sonda) e utilizando a matriz de reduzida de 150 marcadores selecionados por meio do Método da Stepwise (150 - Stepwise).

Verifica-se, portanto, que ambas as técnicas de redução de dimensionalidade foram eficientes, tomando como critério as medidas estatísticas avaliadas (r^2 e $REQM$), e que o Método de Regressão *Stepwise* foi consistentemente superior quando se avaliam somente os valores de r^2, que expressam a acurácia seletiva do RR-BLUP (Figura 4). Entretanto, quando avaliamos somente os valores de $REQM$ – que expressam a acurácia preditiva, o Método da Sonda é mais promissor (Figura 5). Assim, a escolha do melhor método, neste contexto estatístico, está condicionada ao critério estabelecido pelo pesquisador. Sugere-se, então, que uma análise mais aprofundada sobre a capacidade das técnicas de redução em captar as dificuldades da aplicação da abordagem da GWS sob a presença de efeitos pertubadores tais como a dominância, a epistasia e o efeito ambiental seja realizada.

c) Influência da dominância sob a eficiência da GWS

A dominância é um efeito importante na expressão genotípica, entretanto nem todos os modelos genéticos contemplam, em suas causas conhecidas de variação, informações que permitam a estimação de tais efeito. Assim, retomando o modelo de

GWS adotado pelo método RR-BLUP e já apresentado no capítulo anterior, temos que, tal como proposto por Resende et al. (2008):

$$y = Wb + Xm + e \quad (5)$$

Em que:
y: vetor de observações fenotípicas;
b: vetor de efeitos fixos (média geral) com matriz de incidência W;
m: vetor dos efeitos aleatórios dos marcadores com matriz de incidência X;
X: Matriz de incidência composta pelos valores -1, 0 e 1 para o número de alelos do marcador dos genótipos mm, Mm e MM, respectivamente;
e: refere-se ao vetor de resíduos aleatórios.

Como visto, o modelo RR-BLUP/GWS adotado neste trabalho negligencia tanto os efeitos de dominância (interações intra-alélicas) quanto os de epistasia (interações interalélicas) e isso, certamente, afeta a eficiência do método no processo de predição de valores genéticos. No caso da dominância há o atenuante de que informações de heterozigotos contribuem, mesmo no modelo aditivo, para melhor ajuste de modelo e maior impacto sobre o efeito médio de substituição gênica que é fator determinante do ganho por seleção.

Analisando os resultados apresentados na Tabela 2, mas enfatizando a acurácia seletiva obtida ao utilizar o método da Sonda para redução, nota-se que, tal como verificado no estudo com as 1000 marcas, a inclusão da dominância geralmente gera ligeiros acréscimos no valor de r_v^2 (considerando os conjuntos (V1,V5, V9), (V2,V6, V10) ou (V4, V8,V12)). Há queda da acurácia seletiva após o acréscimo de dominância para os casos com alta herdabilidade e estabelecidos sob um modelo aditivo (V3, V7, V11). Nesse caso os valores de r_v^2 decaíram de 0,6037 para 0,4898.

Tabela 2. Estimativas obtidas pelo RR-BLUP com cinco validações cruzadas após reduzir a matriz de marcas utilizando o método da Sonda.

Características		r_t^2	r_v^2	$REQM_t$	$REQM_v$	AIC	CP	Viés
V1 - D0H35Ad	Md	**0,5755**	**0,3070**	**47,7098**	**49,092**	**3834,95**	**0,5536**	**0,4394**
	dv	0,0053	0,0305	1,72867	2,2939	8,76479	0,0279	0,0082
V2 - D0H35Ep	Md	**0,5945**	**0,3017**	**84,534**	**107,264**	**5178,48**	**0,545**	**0,4602**
	dv	0,0242	0,0852	2,871	3,60153	6,78286	0,0764	0,0284
V3 - D0H70Ad	Md	**0,7853**	**0,6037**	**60,5336**	**60,9465**	**3413,18**	**0,7755**	**0,6992**
	dv	0,0134	0,0844	5,48637	5,06006	15,0088	0,0542	0,018
V4 - D0H70Ep	Md	**0,6080**	**0,3747**	**66,1947**	**78,2181**	**4888,31**	**0,6027**	**0,4837**
	dv	0,0274	0,1393	6,43357	13,3804	24,5256	0,1199	0,0345
V5 - D60H35Ad	Md	**0,5732**	**0,3312**	**74,6042**	**75,7321**	**3997,29**	**0,5709**	**0,4415**
	dv	0,0258	0,0960	4,7012	4,8717	3,65932	0,0812	0,0313
V6 - D60H35Ep	Md	**0,6651**	**0,4050**	**173,887**	**220,485**	**5789,78**	**0,6336**	**0,5406**
	dv	0,0224	0,0872	6,88002	24,8458	22,7409	0,0664	0,0277
V7 - D60H70Ad	Md	**0,7587**	**0,5835**	**85,7335**	**86,1159**	**3555,78**	**0,7635**	**0,6677**
	dv	0,0107	0,0408	1,90702	3,24936	12,5052	0,0269	0,0134
V8 - D60H70Ep	Md	**0,7411**	**0,5289**	**119,027**	**149,463**	**5456,55**	**0,7264**	**0,6437**
	dv	0,0117	0,0564	11,437	12,9544	12,1048	0,0394	0,0154
V9 - D120H35Ad	Md	**0,5811**	**0,3412**	**82,5349**	**83,9852**	**4214,7**	**0,5790**	**0,4514**
	dv	0,0390	0,1061	11,3439	11,6866	12,2309	0,0866	0,0440
V10 - D120H35Ep	Md	**0,6029**	**0,3197**	**465,666**	**577,249**	**6515,48**	**0,5646**	**0,4644**
	dv	0,0103	0,0373	11,1547	48,3748	12,1552	0,0328	0,0131
V11 - D120H70Ad	Md	**0,6962**	**0,4806**	**114,492**	**115,167**	**3864,3**	**0,6899**	**0,5885**
	dv	0,0222	0,1032	6,22453	5,72198	18,9143	0,0757	0,0289
V12 - D120H70Ep	Md	**0,7088**	**0,4898**	**414,607**	**468,195**	**6182,21**	**0,6993**	**0,6032**
	dv	0,0067	0,0450	76,202	59,9894	7,46703	0,0320	0,0089

r_t^2 e r_v^2: referem-se aos coeficientes de determinação obtidos no treinamento e na validação, respectivamente; CP: capacidade preditiva; $REQM_t$ e $REQM_v$: raiz do erro quadrático médio para as fases da treinamento e validação, respectivamente; AIC: Índice de Akaíke.

No entanto, analisando a acurácia preditiva do RR-BLUP após a redução – para os mesmos conjuntos considerados acima – nota-se que a dominância exerce grande influência no sentido de inflacionar os valores de $REQM_v$ e *AIC*. A inclusão da dominância elevou o valor do $REQM_v$ de 49,092 para 75,7321, e depois para 83,9852 (para as variáveis V1, V5 e V9, respectivamente) para as variáveis com herdabilidade 35%. Para herdabilidade 70%, o $REQM_v$ aumentou de 60,9465 para 86,1159, e depois para 115,167 (para as variáveis V3, V7 e V11, respectivamente). A variável com maior grau médio de dominância e maior complexidade (V10), dado o modelo epistático adotado e a baixa herdabilidade considerada, foi a que apresentou maior erro quadrático médio, que nesse caso foi 577,249.

As mesmas conclusões acerca da acurácia seletiva podem ser tomadas ao analisar os resultados obtidos para o Método de *Stepwise* (Tabela 3). Os valores de r_v^2 também tendem a aumentar com a inclusão da dominância, salvo para os cenários com epistasia e alta herdabilidade (V3, V7, V11). Nesse caso o r_v^2 decai de 0,8704 para 0,7813.

Tabela 3. Estimativas obtidas pelo RR-BLUP com cinco validações cruzadas após reduzir a matriz de marcas utilizando o método de regressão *Stepwise*.

Características		r_t^2	r_v^2	$REQM_t$	$REQM_v$	AIC	CP	Viés
V1 - D0H35Ad	Md	**0,8624**	**0,6591**	**106,08**	**106,47**	**3762,07**	**0,8097**	**0,7789**
	dv	0,0141	0,1043	4,48	4,67	16,4859	0,0661	0,023
V2 - D0H35Ep	Md	**0,8553**	**0,6393**	**103,07**	**118,14**	**5128,95**	**0,7993**	**0,7667**
	dv	0,0071	0,0317	23,11	22,3052	11,3255	0,0199	0,0116
V3 - D0H70Ad	Md	**0,9474**	**0,8708**	**108,79**	**108,86**	**3213,47**	**0,9331**	**0,9166**
	dv	0,0017	0,0198	1,3214	1,098	5,84742	0,0107	0,0027
V4 - D0H70Ep	Md	**0,8610**	**0,6751**	**39,15**	**54,39**	**4825,17**	**0,8199**	**0,7785**
	dv	0,0161	0,0942	2,196	5,439	16,4933	0,0592	0,0262
V5 - D60H35Ad	Md	**0,8714**	**0,6892**	**159,75**	**160,16**	**3904,25**	**0,8295**	**0,7940**
	dv	0,0111	0,0643	2,999	2,779	17,834	0,0388	0,0181
V6 - D60H35Ep	Md	**0,8560**	**0,6719**	**180,26**	**215,40**	**5738,78**	**0,8190**	**0,7717**
	dv	0,0054	0,0637	39,28	32,396	19,8113	0,0393	0,0090
V7 - D60H70Ad	Md	**0,9301**	**0,8238**	**126,81**	**126,94**	**3390,14**	**0,9075**	**0,8890**
	dv	0,0036	0,0299	3,174	1,804	14,339	0,0165	0,0055
V8 - D60H70Ep	Md	**0,9102**	**0,7778**	**77,38**	**107,29**	**5362,01**	**0,8817**	**0,8577**
	dv	0,009	0,0408	12,73	13,173	9,13637	0,0229	0,0141
V9 - D120H35Ad	Md	**0,8611**	**0,6757**	**143,34**	**144,23**	**4120,12**	**0,8213**	**0,7797**
	dv	0,0097	0,0663	7,471	7,789	10,3152	0,0395	0,0161
V10 - D120H35Ep	Md	**0,8392**	**0,6207**	**568,26**	**654,03**	**6469,88**	**0,7868**	**0,7419**
	dv	0,0156	0,0706	113,25	112,29	11,5767	0,0446	0,0253
V11 - D120H70Ad	Md	**0,9104**	**0,7813**	**179,22**	**179,11**	**3726,57**	**0,8838**	**0,8582**
	dv	0,0050	0,0219	5,8910	6,082	9,33633	0,0123	0,0075
V12 - D120H70Ep	Md	**0,8974**	**0,7506**	**525,96**	**553,06**	**6098,36**	**0,8660**	**0,8374**
	dv	0,0077	0,0444	84,94	83,77	12,4904	0,0259	0,012

r_t^2 e r_v^2: referem-se aos coeficientes de determinação obtidos no treinamento e na validação, respectivamente; CP: capacidade preditiva; $REQM_t$ e $REQM_v$: raiz do erro quadrático médio para as fases da treinamento e validação, respectivamente; AIC: Índice de Akaíke.

A acurácia preditiva, por sua vez, também é inflacionada à medida que maiores graus médios de dominância são considerados. Para as variáveis V1, V5 e V1, por exemplo, os valores de $REQM_v$ para o método de *Stepwise* são 106,471, 160,158 e 179,11, respectivamente.

Os resultados confirmam a forte influência exercida por efeitos de dominância sobre a acurácia preditiva do RR-BLUP/GWS tal como observado por diversos autores

(RESENDE et al., 2008; ALMEIDA FILHO et al., 2016), e indicam que as abordagens de redução da dimensionalidade são eficazes no contexto estatístico, pelas estimativas favoráveis de r^2 e *REQM*, e no contexto genético por ainda traduzirem que esses fatores perturbadores continuam atuando negativamente sobre a eficiência da técnica adotada.

d) Influência da epistasia sob a eficiência da GWS

Tem-se o entendimento de que grande parte dos efeitos epistáticos são determinantes da fração não herdável da variância genotípica de uma população, cujos efeitos são provenientes da interação de dois ou mais alelos diferentes (CRUZ & CARNEIRO, 2003), mas precisam ser isolados no modelo de forma permitir estimar acuradamente os valores genéticos a partir de componentes aditivos e de interação aditiva x aditiva. Já foi relatado no capítulo anterior que os efeitos da epistasia, por serem importantes geneticamente embora não sejam contemplados no modelo, exercem impacto negativo sobre as predições genéticas.

Neste trabalho foi feita a redução da dimensionalidade a partir de técnicas –Sonda ou Regressão *Stepwise* – que levam em consideração apenas a melhor combinação linear na determinação da variável resposta podendo, a princípio, ignorar uma variável importante cuja ação era de ordem não-linear. Assim, é importante verificar se os efeitos perturbadores da epistasia, já detectados com a análise de todas as marcas moleculares, também seriam percebidos após redução da dimensionalidade.

Analisando os resultados obtidos (Tabelas 2 e 3) sob o enfoque somente de efeitos epistáticos, independente de dominância e herdabilidade – os pares de variáveis (V1 e V2), (V3 e V4), (V5 e V6), (V7 e V8), (V9 e V10) e (V11 e V12) – evidenciam que a redução de dimensionalidade amenizou, mas não mascarou os efeitos da epistasia sobre as acurácias seletivas e preditivas do método, uma vez que o r_v^2 caiu e o $REQM_v$ aumentou ao considerarmos a avalição de características que continham esse efeito como determinantes da expressão genotípica (V2, V4, V6, V8, V10 e V12).

Para V1 e V2, pelo o método da Sonda, o r_v^2 do RR-BLUP foi praticamente o mesmo na presença e ausência de epistasia mas o $REQM_v$ aumentou bastante, passando de 49,092 para 107,264 o que leva a concluir que este método de redução da dimensionalidade preserva a informação dos efeitos genéticos existentes no controle da característica. Ao usar o método *Stepwise* de redução, r_v^2 do RR-BLUP passou de 0,65913 para 0,65913 e o $REQM_v$ aumentou de 106,471 para 118,137 com a presença da epistasia.

Para características com dominância e baixa herdabilidade (V9 e V10), o $REQM_v$ aumentou ainda mais, passando de 144,227 para 654,034 após a redução via *Stepwis*e, corroborando com os resultados encontrados por outros autores (HAYES et al., 2009).

e) Influência da herdabilidade na eficiência da GWS

Já é sabido que a herdabilidade (h^2) de uma dada característica de interesse reflete a influência do meio sobre o genótipo, uma vez que expressa a fração do genótipo refletida pelo fenótipo (BUENO et al., 2001; CRUZ, 2005).

Para avaliar os efeitos do ambiente sobre a eficiência do método RR-BLUP após a redução via Sonda e Regressão *Stepwise*, avaliamos as mesmas características consideradas no capítulo anterior ((V1 e V3), (V2 e V4), (V5 e V7), (V6 e V8), (V9 e V11) e (V10 e V12)). Os resultados obtidos (Tabelas 2 e 3) evidenciam que a redução de dimensionalidade por meio do método *Stepwise* contribui bastante para a extrapolação dos resultados para a população de validação, uma vez que, de modo geral, foi possível se recuperar a herdabilidade das características avaliadas. Mesmo para o cenário mais complexo (V10), a redução por *Stepwise* propiciou ao RR-BLUP um r_v^2 de 0,75056, que superava a herdabilidade de 70% da característica considerada.

O método da sonda, no entanto, propiciou quadrados da correlação de validação inferiores à herdabilidade da característica para praticamente todos os cenários, superando o valor de h^2 somente para a variável V6, em que a herdabilidade era 35% e o r_v^2 obtido foi de 0,4050, indicativo de que o investimento em genotipagem para fins de aumento da acurácia seletiva, nesse caso, não seria vantajoso, pois a extrapolação para o conjunto de validação não foi muito eficiente. Autores como RESENDE et al. (2008) e CRUZ et al. (2013) já destacaram a importância de predizer valores genéticos recuperando a herdabilidade da característica econômica de interesse.

f) Comparação entre o Método da Sonda e o Método de Regressão Stepwise

Para comparar o desempenho das duas metodologias de redução de dimensionalidade abordadas nesse estudo, tomemos os cenários mais adversos avaliados, tal como feito no capitulo 1, considerando as médias dos parâmetros avaliados segundo o nível de dominância (ausência, parcial ou sobredominância); epistasia (presença ou ausência) e níveis de herdabilidades (alta ou baixa) (Tabela 4).

Tabela 4. Comparação entre os valores médios obtidos via RR-BLUP após redução pelos métodos Sonda e *Stepwise* ao agrupar as características segundo o nível de dominância; presença ou ausência de epistasia e segundo às duas herdabilidades consideradas, h^2=35% (baixo) e h^2=70% (alto).

Método da Sonda							
Cenários	r_t^2	r_v^2	$REQM_t$	$REQM_v$	AIC	CP	Viés
Sem Dominância	0,6408	0,3968	64,74	73,88	4328,73	0,6192	0,5206
Dominância parcial	0,6845	0,4621	113,31	132,95	4699,85	0,6736	0,5734
Sobredominância	0,6472	0,4078	269,32	311,15	5194,17	0,6332	0,5269
Sem Epistasia	0,6617	0,4412	77,60	78,51	3813,37	0,6554	0,5479
Com Epistasia	0,6534	0,4033	220,65	266,81	5668,47	0,6286	0,5326
h2 Baixo	0,5987	0,3343	154,82	185,63	4921,78	0,5744	0,4662
h2 Alto	0,7163	0,5102	143,43	159,68	4560,06	0,7095	0,6143
Método de Regressão *Stepwise*							
Sem Dominância	0,8815	0,7111	89,27	96,96	4232,41	0,8405	0,8102
Dominância parcial	0,8919	0,7407	136,05	152,45	4598,79	0,8594	0,8283
Sobredominância	0,8770	0,7071	354,19	382,61	5103,73	0,8395	0,8043
Sem Epistasia	0,8971	0,7500	137,33	137,63	3686,10	0,8642	0,8362
Com Epistasia	0,8699	0,6892	249,01	283,72	5603,86	0,8288	0,7923
h2 Baixo	0,8576	0,6593	210,13	233,07	4854,01	0,8109	0,7722
h2 Alto	0,9094	0,7799	176,22	188,27	4435,95	0,8820	0,8564

r_t^2: quadrado da correlação de treinamento; r_v^2: quadrado da correlação de validação; $REQM_t$: erro quadrático médio de treinamento; $REQM_v$: erro quadrático médio de validação; CP: capacidade preditiva

Ao comparar o Método de *Stepwise* com o Método da Sonda, verifica-se que, independente do efeito perturbador considerado – dominância, epistasia ou ambiental –, a acurácia seletiva (r_v^2 e *CP*) do Método Stepwise foi superior que o Método da Sonda na etapa de validação. Os valores de r_v^2 variaram entre 0,6593 e 0,7799 para o *Stepwise*, e entre 0,3343 e 0,5102 para a Sonda (Tabela 4). No geral, o efeito ambiental refletido no quadrado da correlação foi o agente que mais afetou a acurácia seletiva do RR-BLUP, independente do método de redução adotado. As médias de r_v^2, para o RR-BLUP considerando todas as características que possuíam h^2 = 35% foram de 0,3343 e 0,6593, ao utilizar a Sonda e o *Stepwise*, respectivamente.

Avaliando agora acurácia preditiva do RR-BLUP após a redução, percebe-se que a presença da dominância foi o fator mais perturbador, tal como foi antes da redução. Mesmo havendo reduzido consideravelmente quando comparados com o modelo considerando 1000 marcadores, os valores do $REQM_v$ e do *AIC* continuaram inflacionados. Na presença de sobredominância, o método da Sonda proporcionou ao RR-BLUP um $REQM_v$ = 311,15 e *AIC* = 5194,17 (Tabela 4). O *Stepwise*, por sua vez, resultou em $REQM_v$ = 382,61 e *AIC* = 5103,73 (Tabela 4).

Os valores de erro quadrático médio de validação foram superiores para o método *Stepwise* em praticamente todos os cenários variando entre 96,96 e 382,61, enquanto ao utilizar as Sondas esses valores variaram entre mas 73,88 e 311,15. O *AIC* no entanto, foi inferior para esse método – apresentando valores entre 3686,10 e 5103,73 para o *Stepwise* e entre 3813,37 e 5668,47 para o Método da Sonda.

Como pôde ser constatado, o problema de dimensionalidade em GWS já relatado por diversos autores (RESENDE et al., 2012; AZEVEDO et al., 2014) foi controlado por ambas as metodologias de redução, uma vez que as novas matrizes de marcas avaliadas no RR-BLUP, após a redução, possuíam um número de marcadores moleculares inferior ao número de indivíduos da população F1 avaliada (150 < 500). Os eventuais problemas de multicolinearidade também foram contornados, tal como ilustrado nas figuras 1, 2, 3 e 4 apresentadas anteriormente.

Avaliando as acurácias seletiva e preditiva, de modo geral, o Método de *Stepwise* foi mais eficiente e auxiliou melhor o RR-BLUP/GWS, no entanto, para as características mais complexas a acurácia preditiva e a capacidade de extrapolação do método RR-BLUP ainda apresentaram deficiências, dados os elevados valores de erro quadrático médio e índices de Akaíke encontrados. Esses valores inflacionados indicam que o modelo RR-BLUP adotado deve ser repensado, de modo a considerar uma melhor forma de avaliação dos efeitos pertubadores avaliados nesse estudo.

Nas situações mais corriqueiras dos estudos de predição realizados pelos programas de melhoramento, esses efeitos são simplesmente negligenciados e contabilizados no resíduo dos modelos utilizados (RESENDE, et al., 2012; CRUZ, 2012). Alguns autores, no entanto, tem proposto o estudo de características complexas (LOPES et al., 2014; TORO & VARONA, et al., 2010; ZENG et al., 2013), tentando superar as dificuldades enfrentadas pelo RR-BLUP por meio da inclusão da matriz de parentesco em modelos de GWS (JUNIOR et al., 2013; RESENDE, 2007) ou então por meio do uso de outras modelagens de GWS, como os modelos Bayesianos (ALMEIDA FILHO, et al., 2016; MUÑOZ et al., 2014), por exemplo.

Estudos mais recentes, por sua vez, foram além do simples ajuste das métodos de GWS e têm proposto o uso de metodologias baseadas em inteligência computacional, como as Redes Neurais Artificiais, para solucionar tais dificuldades. As Redes Neurais Artificiais consistem em uma metodologia já consolidada nas áreas de ciências da computação, engenharia e medicina (BRAGA, et al., 2011; SILVA, et al., 2010; LAIA & CRUVINEL, 2008) e tem demonstrado enorme potencial de aplicação nas diversas áreas do melhoramento animal e de plantas (CARNEIRO et al., 2017; SILVA et al., 2017;

SILVA et al., 2016; SANT'ANNA et al., 2015; SILVA et al., 2014; NASCIMENTO et al., 2013; GIANOLA et al., 2011).

CONCLUSÕES

A redução de dimensionalidade para fins de estudos de Seleção Genômica Ampla, além de propiciar menor demanda computacional, uma vez que as análises puderam ser realizadas de forma mais rápida, resultaram em aumentos na acurácia seletiva do RR-BLUP.

O método de *Stepwise,* de modo geral, propiciou resultados mais satisfatórios que o método da Sonda para o método RR-BLUP.

Mesmo com a redução, foi evidenciado que o efeito perturbador da dominância, epistasia e do ambiente reduz a eficiência da GWS. A forte influência de tais efeitos sobre a eficiência do modelo adotado sugere que, em estudos subsequentes, outras metodologias, tais como as preconizadas em inteligência computacional, sejam consideradas.

REFERÊNCIAS

AKAIKE, H. A new look at the statistical model identification **IEEE Transaction on Automatic Control**. v.19, p.716-723, 1974.

AZEVEDO, C.F.; SILVA, F.F.; RESENDE, M.D.V.; LOPES, M.S.; DUIJVESTEIJN, N.; GUIMARAES, S.E.F.; LOPES, P.S.; KELLY, M.J.; VIANA, J.M.S.; KNOL, E.F. Supervised independent component analysis as an alternative method for genomic selection in pigs. **J. Anim. Breed. Genet**. 131, 452–461, 2014.

BUENO, L.C.S.; MENDES, A.N.G.; CARVALHO, S.P. **Melhoramento genético de plantas: princípios e procedimentos**. Lavras: UFLA, 2001. 282p.

BRAGA, A.P.; CARVALHO, A.P.L.F.; LUDERMIR, T.B. **Redes Neurais Artificiais - Teoria e aplicações**. 2ed. Rio de Janeiro: LTV, 226p, 2011.

CAO, L.J.; CHUA, K.S.; CHONG, W.K.; LEE, H.P.; GU, Q.M. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. **Neurocomputing**, v.55, p.321 – 336, 2003.

CRUZ, C.D. GENES - A software package for analysis in experimental statistics and quantitative genetics. **Acta Sci.**35: 271-276, 2013.

CRUZ, C.D. **Princípios de genética quantitativa**. Viçosa: Ed. da UFV, 2012. 394p.

CRUZ, C.D.; REGAZZI, A.J.; CARNEIRO, P.C.S. (2012). **Modelos biométricos aplicados ao melhoramento genético**. UFV, Viçosa.

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

GIANOLA, D.; OKUT, H.; KENT A WEIGEL, K.A.; ROSA, G J.M. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. **BMC Genetics.** 12:87, 2011.

GODDARD, M. Genomic selection: prediction of accuracy and maximization of long term response. **Genetica**, v. 136, n. 2, p. 245–57, 2009.

HAYES, B. J.; BOWMAN, P. J.; CHAMBERLAIN, A. J.; GODDARD, M. E. Genomic selection in dairy cattle: progress and challenges. **Journal Dairy Science**, v. 92, n. 1, p. 433–443, 2009.

JAMES, G.; WITTEN, D.; HASTIE; T.; TIBSHIRANI, R. **An Introduction to Statistical Learning with Applications in R**. 2013.

JANG, J.S.R; SUN, C.T; MIZUTANI, E. **Neuro fuzzy and soft computing**. Upper Saddle River, NJ: Prentice Hall, 1997.

LAIA, M.A.M.; CRUVINEL, P.E. Filtragem de Projeções Tomográficas da Ciência do Solo Utilizando Kalman Discreto e Redes Neurais. **IEEE Latin America Transactions**, v.6, n1, 2008.

LONG, N.; GIANOLA, D.; ROSA, G.J.; WEIGEL, K.A. Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. **J Anim. Breed. Genet.** Aug; 128 (4): 247-57, 2011.

LOPES M.S.; BASTIAANSEN J.W.M.; HARLIZIUS B.; KNOL E.F.; BOVENHUIS H. A genome-wide association study reveals dominance effects on number of teats in pigs. **PLoS One** 9: e105867, 2014

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

MONTGOMERY, D.C.; PECK, E.A. **Introduction to linear regression analysis**. New York, J. Wiley, 1981. 504p.

MUÑOZ PR, RESENDE MFR, GEZAN SA, RESENDE MDV, DE LOS CAMPOS G, KIRST M et al. (2014). Unraveling additive from non-additive effects using genomic relationship matrices. **Genetics** 198: 1759–1768.

R CORE TEAM. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2017. Available at: <https://www.R-project.org/>.

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

RESENDE, M.D.V.; LOPES, P.S.; SILVA, R.L.; PIRES, I.E. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, Colombo, n.56, p.63-77, 2008.

SANT'ANNA, I.C.; TOMAZ, R.S.; SILVA, G.N.; NASCIMENTO, M.; BHERING, L.L.; CRUZ, C.D. Superiority of artificial neural networks for a genetic classification procedure. **Genetics and Molecular Research**, v.14, p.9898-9906, 2015. DOI: 10.4238/2015. August.19.24.

SCHUSTER, I.; CRUZ, C.D. Estatística genômica aplicada a populações derivadas de cruzamentos controlados. Viçosa: UFV, 568p. 2013.

SILVA, G.N.; TOMAZ, R.S.; SANT'ANNA, I.C.; CARNEIRO, V.Q.; CRUZ, C.D.; NASCIMENTO, M. Evaluation of the efficiency of artificial neural networks for genetic value prediction. **Genetic Molecular Research**, v.15, p.1-11, 2016. DOI: 10.4238/ gmr.15017676.

SILVA, G.N.; TOMAZ, R.S.; SANT'ANNA, I. de C.; NASCIMENTO, M.; BHERING, L.L.; CRUZ, C.D. Neural networks for predicting breeding values and genetic gains. **Scientia Agricola**, v.71, p.494-498, 2014. DOI: 10.1590/0103- 9016-2014-0057, 2014.

SILVA, I. N.; SPATTI, H. D.; FLAUZINO, R. A. **Redes Neurais Artificiais: para engenharia e ciências aplicadas**. São Paulo: Artliber, 399p. 2010.

TORO, M.A.; VARONA, L. A note on mate allocation for dominance handling in genomic selection. **Genet Sel. Evol**. 42: 33, 2010.

ZENG, J.; TOOSI, A.; FERNANDO, R.L.; DEKKERS, J.C.M.; GARRICK, D.J. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. **Genet Sel. Evol**. 45: 11. 2013.

CAPÍTULO 3

INTELIGÊNCIA COMPUTACIONAL COMO ALTERNATIVA PARA AUMENTAR A EFICIÊNCIA PREDITIVA DE VALORES GENÉTICOS

Resumo
O objetivo deste trabalho foi avaliar o uso das redes neurais, perceptron multicamadas e redes de função de base radial, na predição de valores genéticos, em detrimento à Seleção Genômica Ampla para doze características quantitativas, que contemplavam diferentes estruturações quanto à modelo, graus médios de dominância e herdabilidade. Foram utilizados 500 indivíduos provenientes de uma população simulada F1, oriunda do cruzamento de duas populações contrastantes, também obtidas via simulação. Os 500 indivíduos foram genotipados com 1000 marcadores. Procedeu-se então à redução de dimensionalidade das matrizes de marcadores via os métodos da Sondas e Regressão *Stepwise*. Para as análises foram utilizados os arquivos reduzidos, com 500 indivíduos e 150 marcas. Os resultados obtidos foram comparados com os da metodologia de RR-BLUP. A presença de epistasia e de dominância não foram fatores limitantes para as metodologias propostas, tendo em vista os altos valores de r^2 e capacidade preditiva alcançados para a característica de maior complexidade (r^2 = 0,6628 e *CP* = 0,8133 para RBF/*Stepwise* e r^2 = 0,6493 e *CP* = 0,8050 para RBF/*Stepwise*). Os resultados demonstraram que os métodos de Redes Perceptron Multicamadas e de Redes de Base Radial propiciaram predições mais acuradas que a metodologia de RR-BLUP, reduzindo o *REQM* de 654,03 para valores próximos de 20 após redução via Regressão *Stepwise* ao adotar RBF e MLP. Demonstram ainda que o uso da inteligência computacional no melhoramento consiste em uma promissora ferramenta para fins de predição.
Termos para indexação: Redes neurais artificiais; dominância; epistasia; predição.

Abstract
The objective of this work was to evaluate the use of multilayer perceptron neural network and radial basis function network for predicting genetic values, in detriment to the Genome Wide Selection for twelve quantitative traits, which contemplated different structures regarding the model, mean degrees of dominance and heritability. We used 500 individuals from a simulated F1 population, coming from the crossing of two contrasting populations, also obtained through simulation. The 500 individuals were genotyped with 1000 markers. The dimensionality reduction of the marker matrices was performed using Sonda and Stepwise Regression methodologies. For the analysis the reduced files were used, with 500 individuals and 150 markers. The results obtained were compared with those of the RR-BLUP methodology. The presence of epistasis and dominance were not limiting factors for the proposed methodologies, considering the high values of r^2 and predictive capacity achieved for the characteristic of greater complexity (r^2 = 0.6628 and *CP* = 0.8133 for RBF / Stepwise and r^2 = 0.6493 and *CP* = 0.8050 for RBF / Stepwise). The results showed that the Multilayer Perceptron Networks and Radial Base Networks methods provided more accurate predictions than RR-BLUP, reducing the *REQM* of 654.03 to values close to 20 after reduction by Stepwise Regression when adopting RBF and MLP. They also demonstrate that the use of computational intelligence in breeding programs is a promising tool for prediction purposes.
Index terms: Artificial neural network; dominance; epistasis; prediction.

INTRODUÇÃO

Nas diversas áreas do melhoramento genético, o desafio principal é basicamente o mesmo: recomendar genótipos superiores de modo a aumentar a produtividade e a qualidade do produto. Para tanto, uma gama de metodologias está disponível na literatura,

incluindo métodos de experimentação e análise estatística (RESENDE, 2007); genética biométrica aliada a conceitos de genética quantitativa; seleção assistida de marcadores, e finalmente a seleção genômica ampla, com o uso de diversos modelos estatísticos paramétricos (MEUWISSEN et al., 2001) ou semi-paramétricos (GIANOLA et al., 2006; GIANOLA & VAN KAAM, 2008) de predição. Para todas estas estratégias, uma informação de grande relevância para a avaliação desses genótipos é a acurácia seletiva (RESENDE & DUARTE, 2007).

No geral, tais metodologias divergem umas das outras quanto a modelos, pressuposições e aplicabilidade. No entanto, é praticamente unânime em todas elas o uso de informações fenotípicas fidedignas para inferir sobre as características genotípicas (CRUZ, 2012). Além disso, outro fator predominante é a instalação de experimentos nos quais os valores da média fenotípica são utilizados como indicadores de superioridade genética (PETEK et al., 2008). Mesmo para as metodologias que se destinam à prática da seleção com base em informações genotípicas, como a seleção genômica ampla, uma problemática enfrentada são as diversas suposições sobre distribuição dos dados, acerca de número e os efeitos de QTLs e sobre o modelo genético associado ao caráter quantitativo, diferindo umas das outras dependendo do modelo estatístico (ODILON JUNIOR, 2013). Além disto, em geral, os modelos genéticos adotados são modelos simplificados, tendo em vista a complexidade das características, pois se baseiam em modelos aditivos de predição e negligenciam os efeitos de dominância e epistasia no controle gênico do caráter em estudo.

Tendo em vista as dificuldades ao implementar metodologias tradicionais para fins de predição e classificação, alguns autores têm proposto o uso de técnicas baseadas em inteligência computacional, de Redes Neurais Artificiais, como as Redes Perceptron Multicamadas (MLP) (ROSENBLATT, 1959) e as Redes de Base Radial (RBF) (MOODY & DARKEN, 1989), que diferentemente das modelagens estocásticas, não possuem pressuposições quanto ao modelo, uma vez que seus resultados dependem do aprendizado e não da distribuição das variáveis em si (SILVA, et al., 2017). Esta é uma área de pesquisa que vem despertando crescente interesse nas mais diferentes linhas de pesquisa, pois permite a simulação das capacidades cognitivas de um ser humano (HAYKIN, 2001). Basicamente, as redes MLP e RBF consistem em modelos de processamento de dados que emulam uma rede de neurônios biológicos, capazes de recuperar rapidamente uma grande quantidade de dados e reconhecer padrões baseados na experiência, ou seja, tentam reproduzir as funções das redes biológicas, buscando implementar seu comportamento funcional e sua dinâmica (HAYKIN, 2001).

As redes MLP e RFB diferem entre si quanto a número de camadas ocultas intermediárias, funções de ativação e estratégia de treinamento adotados. Nas redes MLP podem ser utilizadas quantas camadas ocultas sejam necessárias e associadas a diferentes funções de ativação que levam a tomadas de decisão por intermédio de uma combinação de hiperplanos (SILVA et al., 2010). Além disso, o treinamento é totalmente realizado de forma supervisionada. As redes RBF, por sua vez, são compostas por uma única camada oculta, além das camadas de entrada e saída, e funções de ativação do tipo gaussianas, que conduzem a fronteiras delimitadoras definidas por campos hiperesféricos (SILVA et al., 2010). O treinamento nas redes RBF é realizado em duas etapas: etapa I não supervisionada e etapa II, supervisionada tal como no MLP. No modelo neural artificial de ambas, o desempenho está diretamente ligado às conexões entre os elementos que o compõe, calibrados e validados nas etapas de treinamento e validação, respectivamente (HAYKIN, 2001).

Além de não exigirem pressuposições, outra vantagem das metodologias de inteligência computacional baseadas em redes neurais artificiais, em detrimento aos modelos paramétricos usados na GWS, é que as redes não apresentam problemas de multicolinearidade e dimensionalidade, podendo ser aplicadas, sem nenhum problema, aos estudos envolvendo matrizes de marcadores moleculares de alta dimensão. No entanto, à medida que inserimos mais variáveis (marcadores moleculares no caso de estudos de predição) na camada de entrada, a demanda computacional e de tempo requeridas, pelas redes neurais, também aumentam (HAYKIN, 2001). Nesse aspecto, a redução da dimensionalidade das matrizes de entrada, pela seleção orientada de marcadores, viabiliza o emprego de arquiteturas de redes com maior número de camadas ocultas e/ou de neurônios por camada ocultas, que seriam mais eficazes em captar as relações não lineares dos fatores determinantes da característica analisada.

Com base no exposto, o objetivo deste capítulo foi avaliar a emprego da abordagem de inteligência computacional, por meio das metodologias de Redes Neurais Artificiais de Perceptron Multicamadas e de Base Radial, aplicadas em conjunto reduzido de entradas, como alternativa de predição em população simulada considerando diferentes cenários que representam situações de dificuldade, para o melhoramento fundamentado em reprodução sexuada, tais como a dominância, epistasia e efeitos ambientais.

MATERIAL E MÉTODOS

1. População avaliada

Com o intuito de tornar os resultados obtidos pelas metodologias de inteligência computacional propostas nesse capítulo comparáveis com os resultados obtidos no Capítulo 2, para o qual se utilizou a metodologia RR-BLUP/GWS para predição, a mesma população simulada F1 – incluindo informações fenotípicas, genotípicas e características avaliadas – foi utilizada para as análises. A Tabela 1 descreve as doze características avaliadas.

Tabela 1. Características avaliadas no estudo com seus respectivos valores de herdabilidade, modelo genético adotado e grau médio de dominância (gmd).

Característica	Herdabilidade (%)	Modelo	gmd
V1 - D0H35_Ad	35	aditivo	0
V2 - D0H35_Ep	35	epistático	0
V3 - D0H70_Ad	70	aditivo	0
V4 - D0H70_Ep	70	epistático	0
V5 - D60H35_Ad	35	aditivo-dominante	0,6
V6 - D60H35_Ep	35	epistático	0,6
V7 - D60H70_Ad	70	aditivo-dominante	0,6
V8 - D60H70_Ep	70	epistático	0,6
V9 - D120H35_Ad	35	aditivo-dominante	1,2
V10 - D120H35_Ep	35	epistático	1,2
V11 - D120H70_Ad	70	aditivo-dominante	1,2
V12 - D120H70_Ep	70	epistático	1,2

Neste estudo utilizamos a estratégia de redução da dimensionalidade uma vez que, no Capítulo 2 desse trabalho, constatou-se que o uso das metodologias de Sonda e Regressão *Stepwise*, para redução de dimensionalidade, propiciou aumentos consideráveis na acurácia do método de RR-BLUP. A redução solucionou o problema de dimensionalidade – uma vez que o estudo envolvia 500 indivíduos e 1000 marcadores moleculares – e possíveis problemas de multicolinearidade. Além disso, a redução otimiza a demanda computacional das análises. Por esse motivo, as análises de inteligência computacional foram realizadas utilizando os arquivos reduzidos com as 150 marcas pré-selecionadas para avaliar o RR-BLUP – 150 informações genotípicas selecionadas após a redução pelo método de Sonda e 150 informações genotípicas selecionadas após a redução pelo método Regressão *Stepwise*.

2. Métodos de inteligência computacional empregados para fins de predição

Buscando (avaliar) a aplicação de metodologias de inteligência computacional a estudos que envolvam predição e seleção genômica, neste trabalho abordaremos a metodologia de Redes Neurais Artificiais com enfoque nas Redes Perceptron Multicamadas (MPL) e nas Redes de funções de base radial (RBF).

2.1 Redes Perceptron Multicamadas (MLP)

Para o presente estudo, a arquitetura de MLP utilizada foi a *backpropagation,* com três camadas ocultas e considerando de um a quatro neurônios em cada camada. Como informação de entrada, considerou-se a matriz de marcadores moleculares selecionados [X_1 X_2 ... X_{150}], de modo que a saída desejada era o valor genotípico verdadeiro – valor conhecido uma vez que as populações foram geradas via simulação. Na camada de saída, a RNA retornou o valor predito de cada indivíduo (Y_{Rede}), tal como ilustrado na Figura 1.

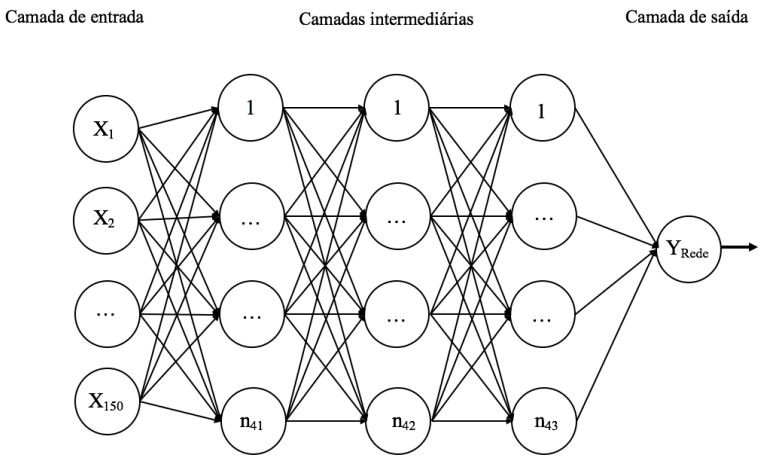

Figura 1. Esquema de uma rede MLP *Backpropagation*. Entradas X_1 a X_{150} na camada de entrada se referem aos 150 marcadores considerados nas análises. Três camadas ocultas intermediárias (n_{i1}, n_{i2} e n_{i3}) constituída de i neurônios (i = 1, ... ,4). A RNA retorna o vetor de valores preditos (Y_{Rede}).

As funções de ativação utilizadas foram a sigmoide logística (*logsig**) e tangente hiperbólica (*tansig**), definidas pelas Equações (1) e (2) abaixo.

$$\varphi(x) = \frac{1}{1+e^{-ax}} \quad (1)$$

$$\varphi(x) = \tanh\left({}^{x}/_{2}\right) = \frac{1+e^{-x}}{1+e^{x}} \quad (2)$$

O processo de treinamento da rede MLP foi realizado por meio do algoritmo de retropropagação de erro (*backpropagation)* (SILVA, et al., 2010). Nesse algoritmo, a rede é alimentada para frente (*forward*) e para trás (*backward*).

2.2 Redes de funções de base radial (RBF)

A arquitetura de RBF utilizada foi a *feedforward*, com uma camada oculta intermediária considerando de 1 a 400 neurônios e um raio de tamanho r, variando de 1 a 80. Tal como para rede neural MLP, a matriz de marcadores moleculares [X_1 X_2 ... X_{150}] é considerada como informação de entrada, de modo que na camada de saída a RBF retorna o vetor de valores preditos (Y_{Rede}).

A função de ativação utilizada na camada oculta foi a gaussiana (Equação 4).

$$g(u) = e^{\frac{-(u-c)^2}{2\sigma^2}} \quad (4)$$

Em que:
c: centro da função gaussiana
σ^2: variância da função gaussiana
u: potencial de ativação

Na camada de saída, utilizou-se uma função de ativação do tipo linear (Equação 5).

$$y_{ri} = g\left(x_0 w_0 + \sum_{j=1}^{q} f_{x_j}(x_i) w_j\right) = (z_0 w_0 + \sum_{j=1}^{q} z_j w_j) \quad (5)$$

**Notação adotada pelo MATLAB*

Em que:

x_i: i-ésima entrada;

w_j: j-ésimo peso sináptico;

f_{x_j}: função de ativação da camada oculta associada à entrada x_i ponderada por seu respectivo peso;

A Figura 2 ilustra a configuração de uma rede RBF.

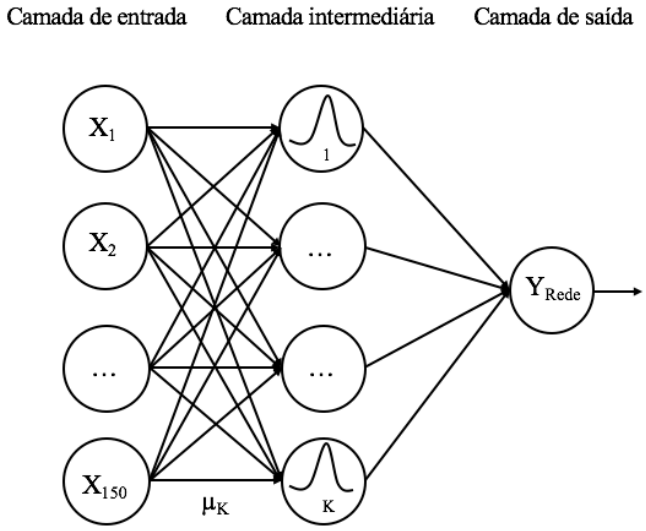

Figura 2. Esquema de uma Rede de Base Radial. Entradas X_1 a X_{150}na camada de entrada se referem aos 150 marcadores considerados nas análises. Uma camada oculta considerando raios de tamanho r (r variando de 1 a 80) e constituída de k neurônios (k = 1, ..., 400). A RBF retorna o vetor de valores preditos (Y_{Rede}).

3. Populações de treinamento e validação

A mesma população F1 foi utilizada para a realização das fases de treinamento e validação para ambas as metodologias – MLP e RBF –, considerando-se uma validação cruzada com k = 5 partições, tal como feito para avaliar o RR-BLUP nos capítulos precedentes a esse. Desse modo, a população foi particionada em cinco subconjuntos mutuamente exclusivos e a cada rodada, quatro desses subconjuntos constituíram a população de treinamento (totalizando 80% dos indivíduos) e o subconjunto restante constituiu a população de validação (20% da população total).

4. Medidas de eficiência da MLP e da RBF

Para avaliar as acurácias preditivas e seletivas das metodologias propostas, os mesmos parâmetros abordados nos capítulos 1 e 2 foram utilizados. Assim, o trabalho contemplou a avaliação da eficiência da MLP e da RBF por meio do ($REQM$) – raiz do erro quadrático médio (Equação 9) – (r_t^2) e (r_v^2) – que correspondem ao quadrado da correlação (Equação 10) para as fases de treinamento e de validação, respectivamente; a capacidade preditiva (CP) e finalmente o viés (β) do modelo.

$$REQM = \sqrt{\frac{\sum(\hat{Y}-Y)^2}{n}} \tag{9}$$

$$r^2 = (cor(GEBV, Y))^2 = (cor(\hat{Y}, Y))^2 \tag{10}$$

$$CP = cor(GEBV, Y) = cor(\hat{Y}, Y) \tag{11}$$

$$\beta = \frac{cov(\hat{Y},Y)}{var(Y)} \tag{12}$$

5. Recursos Computacionais

A simulação da população avaliada foi realizada no software PORTAL GENES (CRUZ, 2016), tal como descrito no capítulo 1. As metodologias de inteligência computacional de MLP e RBF propostas também foram implementados no PORTAL GENES (CRUZ, 2016), no módulo integração com o software MATLAB (MATLAB, 2011).

RESULTADOS E DISCUSSÃO

Dadas a expressiva melhora na acurácia do RR-BLUP após o uso dos métodos de redução de dimensionalidade de Sonda e Regressão *Stepwise* e a necessidade de tornar os resultados obtidos pelas metodologias de inteligência computacional propostas no presente capítulo comparáveis com os resultados do RR-BLUP apresentados no Capítulo 2, as análises de predição por meio de inteligência computacional foram realizadas utilizando-se as matrizes, de marcadores moleculares, reduzidas. Diversos autores já têm demonstrado que o uso de métodos de redução de dimensionalidade e seleção de variáveis

contribui para o aumento da eficiência dos métodos de seleção e predição (CAO et al., 2003; AZEVEDO et al., 2014).

Desse modo, exatamente as mesmas matrizes reduzidas de marcadores utilizadas no capítulo 2 – obtidas pelo Método de Regressão Stepwise e pelo Método da Sonda – foram utilizadas nesse capítulo.

a) Influência da dimensionalidade na eficiência de metodologias de inteligência computacional

Como observado nos Capítulos 1 e 2 deste estudo, a alta dimensão da matriz de marcadores moleculares, além de levar a problemas de multicolinearidade – dado o elevado número de condição ($NC > 100$) obtido ao considerar a inclusão de todos os marcadores moleculares na análise –, pode também prejudicar a acurácia dos métodos de GWS, como observado para o RR-BLUP avaliado. Ao utilizar metodologias de inteligência computacional, por sua vez, a dimensionalidade não representa empecilho no sentido de "prejudicar" a modelagem ou ajuste do problema em estudo como no caso do RR-BLUP – para o qual o número de marcas superior ao número de indivíduos acarreta em problemas no ajuste do modelo linear adotado – uma vez que tais metodologias constituem modelos não estocásticos que se baseiam apenas no processo de aprendizado avaliado (SILVA et al., 2017).

No entanto, se o intuito for avaliar a influência da dimensionalidade no tempo de execução das análises e na demanda computacional requerida, conclui-se que, para metodologias baseadas em inteligência computacional, há grande exigência com relação a esses quesitos, uma vez que em problemas envolvendo redes neurais artificiais, as próprias variáveis de entrada, amplificadas em camadas ocultas, constituem os parâmetros que serão utilizados para treinar e validar os sistemas de MLP e RBF, o que faz com que as análises levem muitas horas e até semanas rodando. Além disso, o algoritmo de aprendizado de *backpropagation* utilizado nesse estudo – que foi adotado devido à complexidade do estudo avaliado – constitui um algoritmo que requer demanda computacional ainda maior para executar as análises (HAYKIN, 2001). Tais dificuldades relacionadas à demanda computacional e tempo, aliadas aos bons resultados obtidos no Capítulo 2 ao agregar técnicas de redução de dimensionalidade ao estudo de GWS serviram de motivação para a condução dos estudos com MLP e RBF utilizando as mesmas matrizes reduzidas de marcas consideradas no capítulo 2 – após a redução de dimensionalidade via Sonda e Regressão *Stepwise*.

Técnicas de redução de dimensionalidade já vêm sendo utilizadas com sucesso dentro de programas de melhoramento genético. Long et al. (2011) utilizaram técnicas de redução de dimensionalidade para predizer a produção de leite em Holsteins e os resultados obtidos demonstraram o grande potencial existente ao combinar redução de dimensão e seleção de variáveis para uma previsão acurada e econômica do valor genético genômico. Hahn et al. (2003) desenvolveram um método de redução da dimensionalidade multifatorial (MDR) para colapsar dados genéticos de alta dimensão em uma única dimensão e concluíram que a redução proposta constituiu abordagem promissora para superar algumas das limitações da regressão logística para a detecção e caracterização das interações gene-gene e gene-ambiente. Para estudos de inteligência computacional, técnicas de redução foram por abordadas. Silva & Schmidt (2016) realizaram um estudo de redução de variáveis de entrada via componentes principais para a modelagem de oxigênio dissolvido por meio de redes neurais artificiais e concluíram que a redução exigiu menor número de neurônios na camada de entrada, poupando esforço computacional e tempo de treinamento.

b) Comparação global entre GWS (RR-BLUP) e RNA (RBF e MLP)

Uma comparação global entre as metodologias de GWS e de RNA utilizadas foi realizada por meio dos histogramas apresentados a seguir (Figuras 4 e 5), nos quais são apresentados ao leitor os resultados médios de r^2 e $REQM$ de validação obtidos pelas metodologias de MLP, RBF e RR-BLUP/GWS pós redução do número de entradas (marcadores) em cada abordagem estatística.

Ao avaliar os resultados descritos nos histogramas (Figura 5), dois aspectos se destacam. O primeiro é que a redução da dimensionalidade pelo método fundamentado em regressão *Stepwise* conduz a resultados mais satisfatórios em promover maiores estimativas de r^2. O segundo, refere-se a similaridade entre as acurácias seletivas (r^2) das metodologias avaliadas para as 12 características estudadas. Mais especificamente, para a redução realizada por meio do método da Sonda, os valores de r^2 de validação variaram entre 0,3017 e 0,6037 para o método RR-BLUP (Tabela 5 (anexo)); 0,2843 e 0,5944 para o RBF (Tabela 1 (anexo)) e entre 0,2016 e 0,5755 para a metodologia de MLP (Tabela 3 (anexo)).

A redução via Regressão *Stepwise* propiciou r^2 de validação variando entre 0,6207 e 0,8708 para a metodologia de RR-BLUP (Tabela 6 (anexo)); entre 0,5562 e 0,8216 para o RBF (Tabela 2 (anexo)) e entre 0,6493 e 0,8731 para a metodologia de

MLP (Tabela 4 (anexo)). Apesar de termos detectado pequenas diferenças entre as acurácias dos métodos avaliados, observou-se que para variável V10 por exemplo, que constitui característica de alta complexidade dado o alto nível de dominância, baixa herdabilidade e a presença de epistasia, os valores de r^2 foram de 0,6207, 0,6493 e 0,6628 para o RR-BLUP, MLP e RBF, respectivamente, evidenciando a superioridade, ainda que tênue, das metodologias de MLP e RBF na avaliação de cenários de maior complexidade.

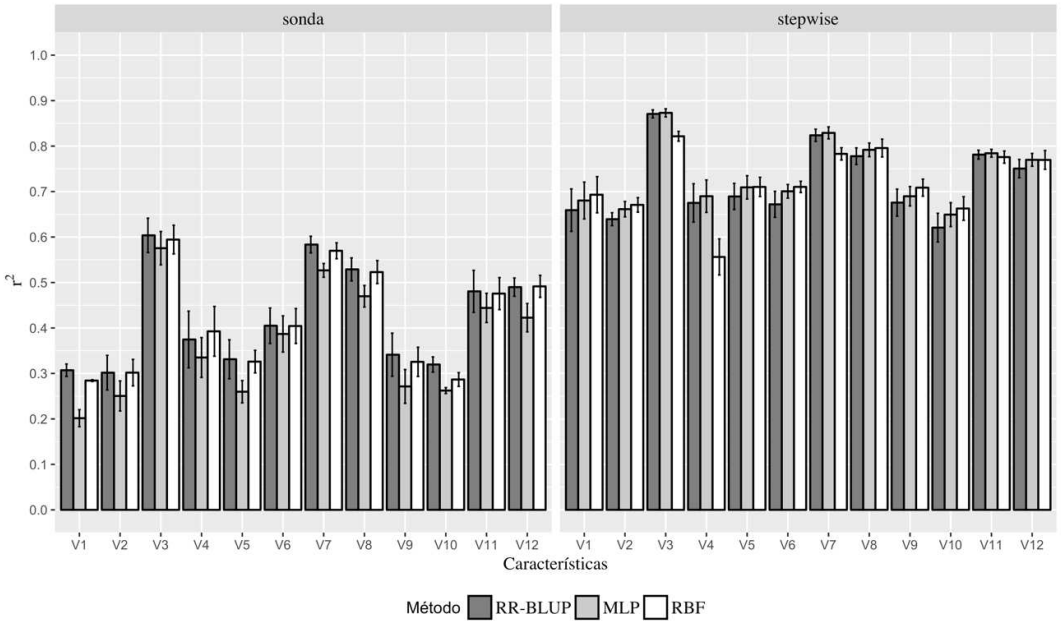

Figura 4. Histograma dos resultados obtidos para a estatística r^2, obtida em conjunto de dados de validação, pelas metodologias de RR-BLUP, MLP e RBF, após utilizar as duas estratégias de redução pelo Método da Sonda e Método de Regressão *Stepwise*.

Uma comparação geral entre as acurácias preditivas das metodologias propostas também foi realizada. A acurácia preditiva de metodologias de seleção e/ou predição pode ser avaliada por meio do erro quadrático médio ou pela raiz desse valor ($REQM$). Basicamente, o $REQM$ expressa a diferença entre valores esperados e preditos (RESENDE et al., 2014). Os histogramas obtidos (Figura 5) demonstraram que, para esta medida, a redução da dimensionalidade por meio da técnica de sondas é tão apropriada quanto a obtida por meio da regressão *Stepwise*. Adicionalmente, é revelado expressiva superioridade da acurácia preditiva das metodologias de inteligência computacional propostas quando estas são comparadas à seleção genômica ampla, dados os valores inferiores de $REQM$ obtidos pela MLP e pelo RBF.

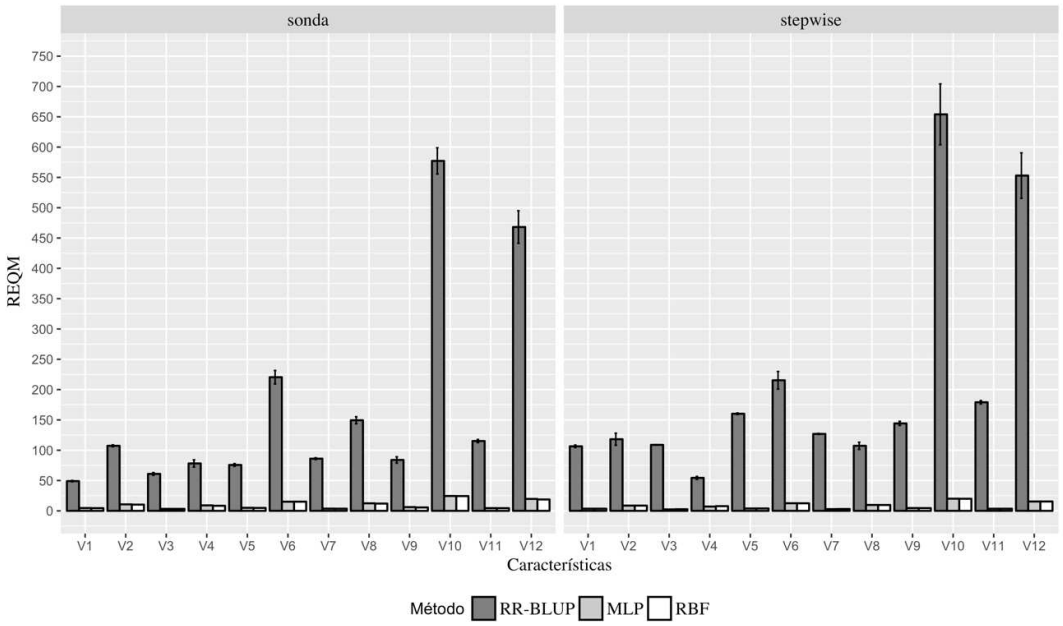

Figura 5. Histograma dos resultados obtidos para a estatística $REQM$, obtida em conjunto de dados de validação, pelas metodologias de RR-BLUP, MLP e RBF, após utilizar as duas estratégias de redução pelo Método da Sonda e Método de Regressão *Stepwise*.

O erro quadrático de validação reduziu em até 10 vezes – como no caso da variável V10, que decaiu de 654,03 para aproximadamente 20 após redução via Regressão *Stepwise* – ao adotar as metodologias de inteligência computacional propostas.

c) Influência da dominância sob a eficiência da RBF e da MLP

A identificação e posterior recomendação de indivíduos superiores, de maneira geral, sofre influência de três fatores perturbadores que dificultam o reconhecimento da superioridade genética destes dentro nas populações ou famílias avaliadas: a dominância, a epistasia e os efeitos aleatórios devido ao ambiente (CRUZ, 2012). Por esse motivo, a possibilidade de utilizar uma metodologia que seja capaz de detectar tais efeitos, geralmente negligenciados pelas metodologias convencionalmente adotadas para seleção, representa um grande avanço para toda a comunidade científica. Como já apresentado nos capítulos anteriores desse estudo, a influência da dominância pode ser mensurada por meio do grau médio de dominância (gmd), que expressa a posição relativa do heterozigoto em relação à média dos homozigotos (CRUZ, 2012). A influência da dominância sobre o RR-BLUP, RBF e MLP, independentemente do modelo adotado (aditivo ou epistático) e da herdabilidade (h^2=35% ou h^2=70%), encontra-se explicitada nas Figuras 6 e 7.

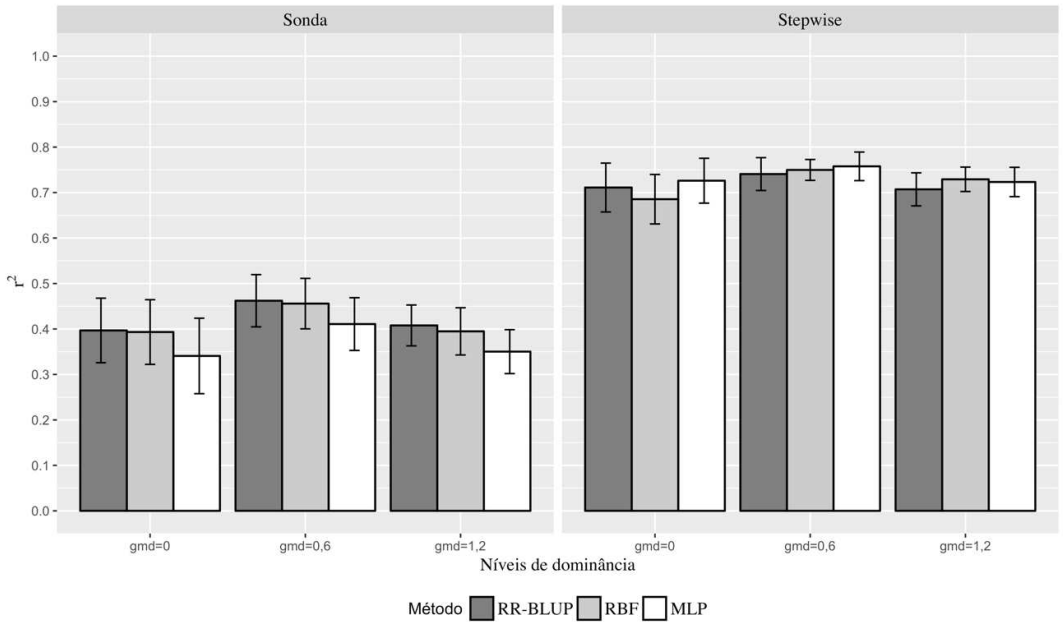

Figura 6. Resultados médios obtidos para a estatística r^2, em conjunto de dados de validação, pelas metodologias de RR-BLUP, MLP e RBF, após utilizar as duas estratégias de redução pelo Método da Sonda e Método de Regressão *Stepwise* considerando ausência de dominância (gmd = 0), dominância parcial (gmd = 0,6) e sobredominância (gmd = 1,2).

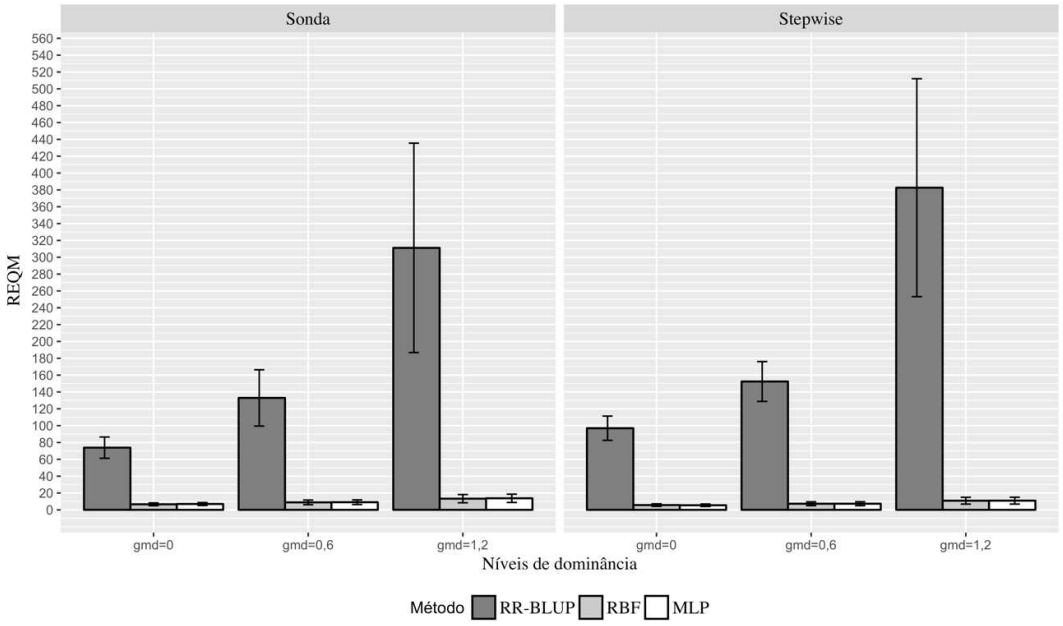

Figura 7. Resultados médios obtidos para a estatística $REQM$, em conjunto de dados de validação, pelas metodologias de RR-BLUP, MLP e RBF, após utilizar as duas estratégias de redução pelo Método da Sonda e Método de Regressão Stepwise considerando ausência de dominância (gmd = 0), dominância parcial (gmd = 0,6) e sobredominância (gmd = 1,2).

Avaliando a acurácia seletiva (r^2) dos métodos de RBF e MLP propostos obtida, verificou-se que, tal como para o RR-BLUP, a inclusão da dominância como fator determinante da característica, de modo geral, propiciou ligeiros acréscimos no valor de r_v^2 (Figura 6), reafirmando um resultado já abordado no Capítulo 1, o qual nos permitiu concluir que a inclusão de efeitos de dominância na população F1 avaliada não afetou a acurácia seletiva do RR-BLUP mas impactou negativamente a capacidade de predição dessa metodologia, ou seja, a aproximação do valor predito com o valor real, dados os elevados valores de REQM de validação obtidos para o RR-BLUP (Figura 7).

Comparando as metodologias de inteligência computacional, por sua vez, percebeu-se que a eficiência seletiva e preditiva do RBF e do MLP não foram comprometidas pela inclusão da dominância (gmd = 0,6 e gmd = 1,2), resultados evidenciados pela semelhança entre as acurácias seletivas (r^2) das três metodologias abordadas (Figura 6) e ressaltado pela superioridade da acurácia preditiva (*REQM*) dos métodos RBF e MLP, independentemente da técnica de redução aplicada (Figura 7). Os valores de *REQM* de validação via Sonda do RR-BLUP passaram, em média, de 73,8802 para 311,1491 na análise de características que incluíam dominância em seu controle gênico; ao usar inteligência computacional esses valores passaram de 6,6084 para 13,2927 para o RBF e de 6,9464 para 13,7216 para o MLP (Tabela 6). Para a estatística dada pelo *CP* os valores também se mantiveram semelhantes (Tabelas 2 e 3).

Tabela 2. Comparação média entre a eficiência do RR-BLUP, RBF e MLP após redução pelo Método da Sonda ao agrupar as características segundo o nível de dominância (gmd = 0, gmd = 0,6 ou gmd = 1,2).

Método	Dominância	$r_t^{2\ *}$	r_v^2	$REQM_t$	$REQM_v$	CP	Viés
RR-BLUP	gmd = 0	0,6408	0,3968	64,7430	73,8802	0,6192	0,5206
	gmd = 0,6	0,6845	0,4621	113,3129	132,9490	0,6736	0,5734
	gmd = 1,2	0,6472	0,4078	269,3250	311,1491	0,6332	0,5269
RBF	gmd = 0	0,8761	0,3933	4,4079	6,6084	0,6168	0,6598
	gmd = 0,6	0,9078	0,4558	5,6208	8,9008	0,6694	0,7027
	gmd = 1,2	0,9070	0,3948	8,1691	13,2927	0,6223	0,6531
MLP	gmd = 0	0,6909	0,3407	5,3574	6,9464	0,5675	0,7360
	gmd = 0,6	0,7196	0,4108	7,3832	9,0744	0,6336	0,7938
	gmd = 1,2	0,7197	0,3502	10,7927	13,7216	0,5850	0,7413

r_t^2: quadrado da correlação de treinamento; r_v^2: quadrado da correlação de validação; $REQM_t$: erro quadrático médio de treinamento; $REQM_v$: erro quadrático médio de validação; CP: capacidade preditiva

Tabela 3. Comparação média entre a eficiência do RR-BLUP, RBF e MLP após redução pelo Método de Regressão *Stepwise* ao agrupar as características segundo o nível de dominância (gmd = 0, gmd = 0,6 ou gmd = 1,2).

Método	Dominância	r_t^2	r_v^2	$REQM_t$	$REQM_v$	CP	Viés
RR-BLUP	gmd = 0	0,8815	0,7111	89,2724	96,9634	0,8405	0,8102
	gmd = 0,6	0,8919	0,7407	136,0487	152,4468	0,8594	0,8283
	gmd = 1,2	0,8770	0,7071	354,1946	382,6071	0,8395	0,8043
RBF	gmd = 0	0,8762	0,6854	4,4078	5,6641	0,8247	0,8808
	gmd = 0,6	0,9078	0,7498	5,6217	7,2808	0,8653	0,9082
	gmd = 1,2	0,9069	0,7292	8,1714	10,8688	0,8531	0,8958
MLP	gmd = 0	0,8879	0,7261	4,3442	5,4552	0,8498	0,9220
	gmd = 0,6	0,8999	0,7577	5,7972	7,2814	0,8696	0,9347
	gmd = 1,2	0,8849	0,7233	8,7286	10,9231	0,8494	0,9236

r_t^2: quadrado da correlação de treinamento; r_v^2: quadrado da correlação de validação; $REQM_t$: erro quadrático médio de treinamento; $REQM_v$: erro quadrático médio de validação; CP: capacidade preditiva

Uma análise comparativa de resultados para as características que não incluíram efeito da dominância em seu controle gênico deveria, a princípio, indicar a superioridade da técnica RR-BLUP sobre as demais. No modelo RR-BLUP considera-se apenas os efeitos de dose alélica de cada marcador de forma que, ao ser aplicado para estudo de característica sem efeito de dominância, conseguiria captar toda a contribuição aditiva dos locos e a redução na eficácia estaria apenas comprometida pelos efeitos perturbadores do ambiente. Nesta condição, seria esperado que as técnicas de inteligência computacional poderiam proporcionar resultados apenas equivalentes na medida que os neurônios utilizados teriam o papel de detectar unicamente as influências lineares. A menor eficiência das técnicas de inteligência computacional, nestas condições, seria atribuída ao processo iterativo, aos critérios de parada, à convergência pelo uso algoritmo de treinamento que poderia estacionar a solução em pontos locais ao invés de globais.

Quando a característica é determinada por efeitos da dominância pode-se questionar se técnicas como o RR-BLUP teria sua eficácia preservada mesmo quando se utilizam modelos em que os regressores representam apenas o efeito de dose alélica, não contemplando interações intra e nem inter-alélica no modelo estatístico. Entretanto, deve-se considerar que na regressão RR-BLUP o efeito estimado representa o efeito de substituição gênica que, segundo teoria de genética quantitativa, é determinado tanto pelo valor genotípico do homozigoto (*a*) quando do heterozigoto (*d*) e, assim, a presença da dominância no controle da característica e a ausência de uma matriz de incidência da dominância no modelo estatístico não representaria grandes problemas.

Assim, feita a comparação global entre os métodos RR-BLUP, RBF e MLP, uma avaliação mais criteriosa das metodologias de inteligência computacional propostas pode ser realizada. Após a redução por Sonda, verificou-se acréscimos, ainda que tênues, na acurácia seletiva (r^2 e *CP*) do método RBF para as características de baixa herdabilidade (h^2=35%), no entanto esse valor não foi recuperado de modo satisfatório por essa metodologia (considerando os conjuntos (V1,V5, V9), (V2,V6, V10) ou (V4, V8,V12)). Por exemplo, os valores de r_v^2 foram 0,2843, 0,3260 e 0,3255 para V1, V5 e V9, respectivamente (Tabela 1 (anexo)). As mesmas conclusões puderam ser tiradas ao adotar a metodologia de inteligência computacional de MLP após redução via Sonda, dado que os valores de r_v^2 nesse caso foram de 0,2016, 0,2598 e 0,2714 para V1, V5 e V9, respectivamente (Tabela 3 (anexo)).

Para as características de alta herdabilidade (h^2=70%) estabelecidas segundo um modelo aditivo (considerando o conjunto (V3, V7, V11)), observou-se decaimento nos valores de r^2 de validação à medida que os graus de dominância aumentavam, independente da técnica de redução e da metodologia de inteligência artificial adotada, observando-se r_v^2 iguais a (V3 = 0,5944, V7 = 0,5698, V11 = 0,4756) ao adotar RBF/Sonda (Tabela 1 (anexo)); (V3 = 0,5755, V7 = 0,5267, V11 = 0,4441) ao adotar MLP/Sonda (Tabela 3 (anexo)).

A redução via regressão *Stepwise*, por sua vez, proporcionou acurácias seletivas bastante satisfatórias de modo que as herdabilidades das características puderam ser recuperadas. Obtiveram-se valores de r^2 de validação de 0,6931, 0,7102 e 0,7086 ao adotar a metodologia de RBF (Tabela 2 (anexo)) e valores de 0,6803, 0,7093 e 0,6897 ao adotar a metodologia de MLP (Tabela 4 (anexo)) para as variáveis V1, V5 e V9 respectivamente.

Tomando por base os resultados obtidos para a acurácia preditiva, expressos pelo *REQM*, confirma-se o impacto sofrido por tal estatística com a inclusão de efeitos de dominância no controle da característica, sendo um pouco mais acentuados para características de baixa herdabilidade. Para o MLP/Sonda, por exemplo, o *REQM* passou de 4,6959 para 5,0393 e depois para 6,1647 para gmd = 0, gmd = 0,6 e gmd = 1,2, respectivamente, considerando características com herdabilidade de 35% e sem efeitos de epistasia (Tabela 3 (anexo)). Para características de herdabilidade alta (70%), por sua vez, obtiveram-se $REQM_v$ de 3,3275, 3,7288 e 4,5116, para gmd = 0, gmd = 0,6 e gmd = 1,2, respectivamente (Tabela 3 (anexo)).

O aumento da acurácia seletiva de metodologias aplicadas ao melhoramento genético com a inclusão de efeitos de dominância nos modelos de predição constitui uma

importante discussão já abordada por outros autores na literatura. Almeida Filho et al. (2016) realizaram estudos preliminares de modo a detectar a contribuição da dominância na predição fenotípica no melhoramento de *pinus* e também em populações simuladas. Os resultados obtidos indicaram que o uso de modelos que contemplem a inclusão de efeitos de dominância pode melhorar a acurácia (r) na predição fenotípica. Nishio & Satoh (2014) incluíram efeitos de dominância no modelo GBLUP por meio da estimação da variância e predição do mérito genético em suínos por meio de simulação e concluíram que a inclusão da dominância contribuiu para o aumento da acurácia dos efeitos genéticos globais estimados e que o GBLUP-D (GBLUP com dominância) é uma abordagem viável para melhorar o desempenho genético em populações mestiças com grande variação genética de dominância e identificar sistemas de acoplamento com boa capacidade de combinação.

Note que para o uso dessas abordagens de GWS, os efeitos de dominância devem ser incluídos nos modelos adotados – com a inclusão de matrizes de parentesco, por exemplo – ao passo que a abordagem de redes neurais permite que a inclusão de efeitos de dominância no controle da característica seja captada por meio das conexões entre neurônios e funções de ativação, sem a necessidade de uma modelagem estocástica já que a tomada de decisão é realizada com base somente no processo de aprendizado (SILVA et al., 2017).

d) Influência da epistasia sob a eficiência da RBF e da MLP

Assim como a dominância, as interações epistáticas aliadas às dificuldades de realizar o controle gênico de características quantitativas de interesse, uma vez que o caráter quantitativo é controlado por muitos genes, também dificultam a prática do melhoramento e seleção, pois fenótipos superiores podem ser atribuídos a populações que possuem elevado número de indivíduos geneticamente distintos (VENCOSVSKY, 1973).

Comparando de forma mais geral a influência dos efeitos epistáticos sobre as metodologias de inteligência computacional de RBF e MLP com a influência de tais efeitos sobre a preconizada metodologia de seleção genômica ampla de RR-BLUP (Figuras 8 e 9), cujo modelo incluía apenas efeito aditivos de dose alélica dos marcadores, foi possível reafirmar as dificuldades enfrentadas pelos programas de melhoramento genético em contabilizar os efeitos de epistasia, que constituem toda e qualquer interação entre diferentes genes (CRUZ & CARNEIRO, 2003).

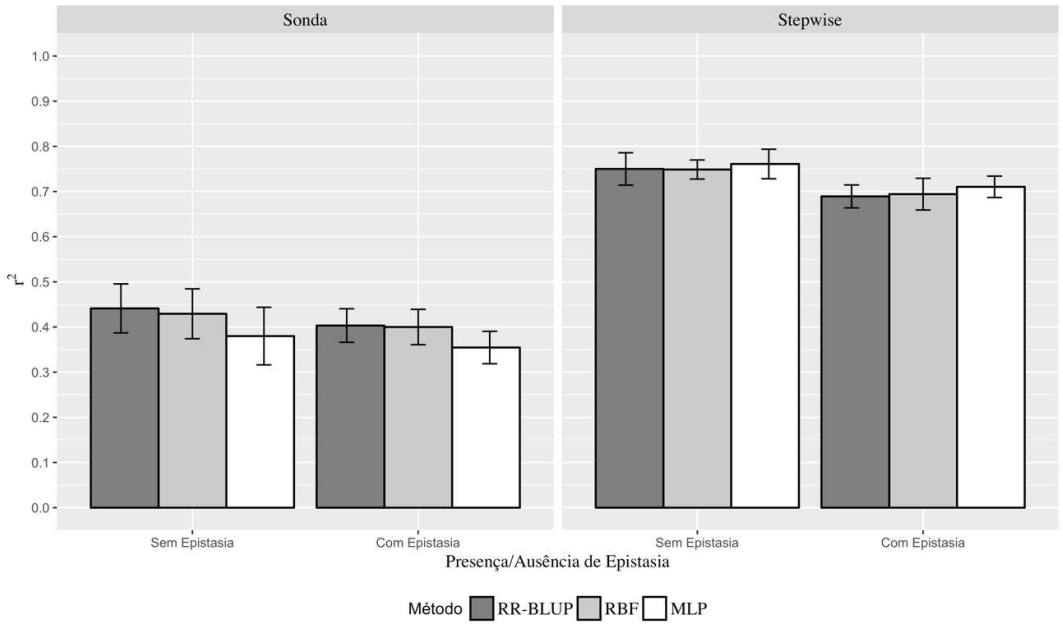

Figura 8. Resultados médios obtidos para a estatística r^2, em conjunto de dados de validação, pelas metodologias de RR-BLUP, MLP e RBF, após utilizar as duas estratégias de redução pelo Método da Sonda e Método de Regressão *Stepwise* considerando a utilização de um modelo aditivo (Sem Epistasia) e a utilização de um modelo epistático (Com Epistasia).

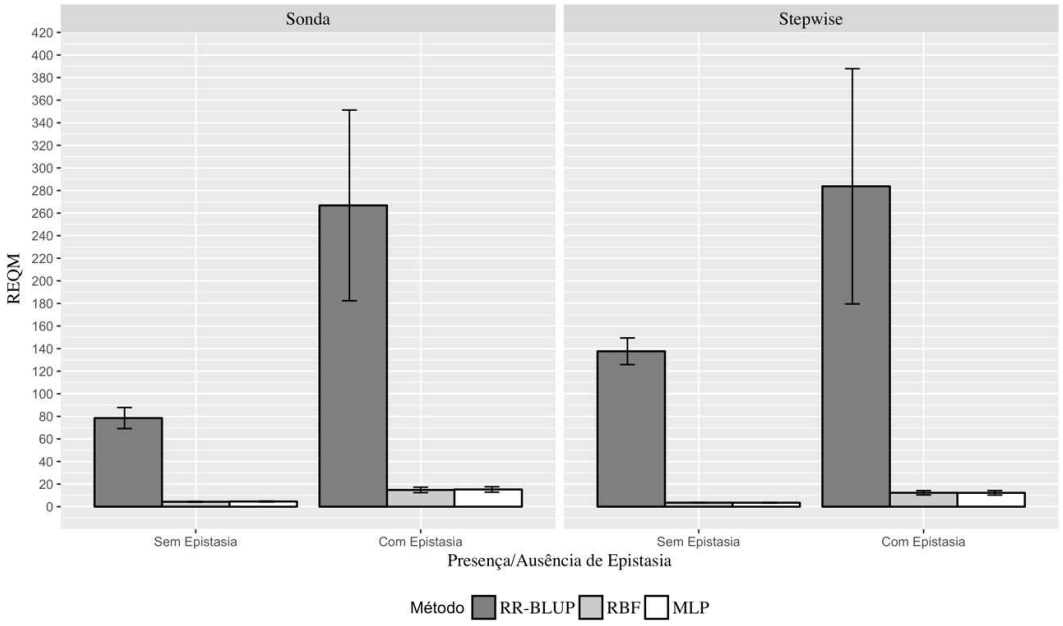

Figura 9. Resultados médios obtidos para a estatística $REQM$, em conjunto de dados de validação, pelas metodologias de RR-BLUP, MLP e RBF, após utilizar as duas estratégias de redução pelo Método da Sonda e Método de Regressão *Stepwise* considerando a utilização de um modelo aditivo (Sem Epistasia) e a utilização de um modelo epistático (Com Epistasia).

Os resultados evidenciaram que a inclusão da epistasia propiciou decréscimos nas acurácias para todos os métodos abordados (Figuras 8 e 9).

A acurácia seletiva do RR-BLUP e dos métodos RBF e MLP foi ligeiramente superior ao utilizarmos o método da Sonda para redução das matrizes de marcas, apresentando r_v^2 de 0,4033, 0,40 e 0,3546 para o RR-BLUP, RBF e MLP, respectivamente, ao considerar características com epistasia (Tabela 4).

Tabela 4. Comparação média entre a eficiência do RR-BLUP, RBF e MLP após redução pelo Método da Sonda ao agrupar as características segundo presença ou ausência de epistasia.

Método	Modelo	r_t^2	r_v^2	$REQM_t$	$REQM_v$	CP	Viés
RR-BLUP	Sem epistasia	0,6617	0,4412	77,6013	78,5065	0,6554	0,5479
	Com epistasia	0,6534	0,4033	220,6526	266,8124	0,6286	0,5326
RBF	Sem epistasia	0,9055	0,4293	2,7695	4,3770	0,6468	0,6797
	Com epistasia	0,8884	0,4000	9,3623	14,8242	0,6255	0,6640
MLP	Sem epistasia	0,7166	0,3799	3,5322	4,5830	0,6026	0,7604
	Com epistasia	0,7036	0,3546	12,1566	15,2453	0,5881	0,7536

r_t^2: quadrado da correlação de treinamento; r_v^2: quadrado da correlação de validação; $REQM_t$: erro quadrático médio de treinamento; $REQM_v$: erro quadrático médio de validação; CP: capacidade preditiva

Ao adotar a técnica de Regressão *Stepwise* para redução, os métodos propostos RBF e MLP apresentaram resultados mais evidentes para as características com epistasia, obtendo-se neste caso r_v^2 de 0,6882, 0,6942 e 0,7104 para o RR-BLUP, RBF e MLP, respectivamente (Tabela 5).

Tabela 5. Comparação média entre a eficiência do RR-BLUP, RBF e MLP após redução pelo Método de Regressão *Stepwise* ao agrupar as características segundo presença ou ausência de epistasia.

Método	Modelo	r_t^2	r_v^2	$REQM_t$	$REQM_v$	CP	Viés
RR-BLUP	Sem epistasia	0,8971	0,7500	137,3320	137,6266	0,8642	0,8362
	Com epistasia	0,8699	0,6892	249,0118	283,7183	0,8288	0,7923
RBF	Sem epistasia	0,9054	0,7487	2,7703	3,5606	0,8643	0,9084
	Com epistasia	0,8885	0,6942	9,3636	12,3152	0,8311	0,8815
MLP	Sem epistasia	0,9024	0,7610	2,7923	3,5119	0,8708	0,9342
	Com epistasia	0,8794	0,7104	9,7877	12,2611	0,8418	0,9194

r_t^2: quadrado da correlação de treinamento; r_v^2: quadrado da correlação de validação; $REQM_t$: erro quadrático médio de treinamento; $REQM_v$: erro quadrático médio de validação; CP: capacidade preditiva

Apesar do impacto negativo do fator genético epistasia ter sido evidenciado tanto na acurácia preditiva ($REQM$) quanto na acurácia seletiva (r_v^2) do RR-BLUP e também para as metodologias de inteligência computacional propostas nesse capítulo, se

retomarmos à figura 9 na qual tem-se uma visão geral da acurácia preditiva do RR-BLUP, RBF e MLP dada em termos do erro quadrático médio (*REQM*), percebe-se que mesmo sofrendo a influência desse fator perturbador, os métodos de RBF e MLP propiciaram uma diminuição expressiva nos valores do *REQM*, que constitui estatística importante em estudos de predição de valores genéticos pois mede a aproximação do valor predito com o valor real, de modo que minimizar tal erro significa maximizar a acurácia (HENDERSON, 1984).

Abordando essa estatística, nota-se que os valores de *REQM* de validação passaram de 266,8124 para 14,8242 e 15,2453 ao optar pelas metodologias de RBF/Sonda e MLP/Sonda, respectivamente, em detrimento ao RR-BLUP/Sonda (Tabela 4); e passaram de 283,7183 para 12,3152 e 12,2611 ao utilizar redução via Regressão *Stepwise* (Tabela 5).

Para avaliar de forma mais criteriosa os impactos da epistasia sobre a acurácia das metodologias de RBF e MLP propostas, tal como feito no Capítulo 1, concentraremos nossa interpretação nos pares de variáveis (V1 e V2), (V3 e V4), (V5 e V6), (V7 e V8), (V9 e V10) e (V11 e V12) apresentados nas tabelas 2, 3, 4 e 5. Esta análise nos permitiu concluir que as características de baixa herdabilidade ($h^2 = 35\%$) sofreram maior impacto da epistasia, uma vez que os valores de *REQM* de validação destas foram inflacionados ainda mais ao considerar um modelo epistático. Para V9 e V10, por exemplo, que constituía um cenário de alta complexidade pois além da epistasia considerou-se nesse caso nível de dominância gmd = 1,2, o REQM subiu de 5,6134 para 24,3591 (Tabela 1 (anexo)) e de 4,5473 para 19,93 (Tabela 2 (anexo)) ao usar os métodos RBF/Sonda e RBF/*Stepwise*, respectivamente; e foi inflacionado de 6,1947 para 24,4833 (Tabela 3 (anexo)) e de 4,6329 para 20,06761 (Tabela 4 (anexo)) ao usar os métodos MLP/Sonda e MLP/*Stepwise*, respectivamente.

Para as características de herdabilidade $h^2 = 70\%$ o maior impacto epistasia na acurácia também foi maior para as características afetadas também por um maior nível de dominância gmd = 1,2 (V11 e V12). Os valores de *REQM* de validação para esses cenários passaram de 4,4306 para 18,7675 (Tabela 1 (anexo)) e de 3,5524 para 15,4453 (Tabela 2 (anexo)) ao usar o método RBF/Sonda e RBF/*Stepwise*, respectivamente; e passaram de 4,5116 para 19,69683 (Tabela 3 (anexo)) e de 3,5421 para 15,4497 (Tabela 4 (anexo)) ao usar o método MLP/Sonda e MLP/*Stepwise*, respectivamente.

Com base nos resultados, evidenciou-se as metodologias de inteligência computacional RBF e MLP foram capazes de captar os efeitos de epistasia existentes nas características de interesse desse estudo por meio de suas redes de neurônios, ao contrário

do RR-BLUP, que negligencia tais efeitos, uma vez que seu modelo contempla somente o termo Xm, que em sua equação representa o espaço de incidência dos efeitos de dose dos marcadores estudados e não a interação entre os mesmos (RESENDE et al., 2014). Nas técnicas de inteligência computacional, as entradas são também representadas pelas informações de Xm marcadores, porém as camadas ocultadas empregadas tornam-se indispensáveis para captar efeitos além daqueles relacionado a ação aditiva das informações. Estudos abordando metodologias baseadas em redes neurais artificiais para predição começaram a ser realizados por alguns autores. Gonzàlez-Camacho et al. (2012) utilizaram as redes RBF, modelos de reprodução dos espaços de Hilbert (RKHS) e LASSO bayesiano para predição com marcadores moleculares densos em dados simulados e linhagens de milho e os resultados obtidos indicaram que os modelos apresentaram uma precisão de previsão geral semelhante e que o RKHS e RBFNN capturaram efeitos epistáticos.

e) Influência da herdabilidade sob a eficiência da RBF e da MLP

Como abordado nos capítulos anteriores, o estudo do herdabilidade (h^2) é de suma importância para o sucesso dos programas de melhoramento pois auxilia na determinação da variação genética dos indivíduos e sua influência nas gerações segregantes, de modo que se possa identificar, a partir dos valores fenotípicos, os indivíduos que apresentam os valores genotípicos desejáveis e a maior concentração de alelos favoráveis (CRUZ, 2012). Além disso, por meio da herdabilidade o pesquisador pode inferir a respeito da influência ambiental sobre as características de interesse, de modo que maiores variações ambientais resultam numa diminuição na herdabilidade da característica (BUENO et al., 2001). Por exemplo, se atribuirmos herdabilidade zero para determinada população, concluiremos que a variação atribuída à causa genética é zero e que a variação entre os indivíduos nesse caso foi unicamente de natureza ambiental (CRUZ, 2012).

O interesse em verificar se as metodologias de redes neurais de RBF e MLP propostas foram capazes de captar, com maior eficiência, a influência (ou ruído) dos efeitos ambientais sobre a população estudada, levou-nos a fazer um paralelo entre essas e o RR-BLUP, tal como realizado nos itens anteriores. Para tanto, comparou-se as acurácias médias obtidas para os métodos de RBF, MLP e RR-BLUP para a herdabilidade baixa (h^2 = 35%) e para a herdabilidade alta (h^2 = 70%), desconsiderando os efeitos dos demais fatores perturbadores também abordados nesse estudo (dominância e epistasia). As Figuras 10 e 11 expressam os resultados obtidos.

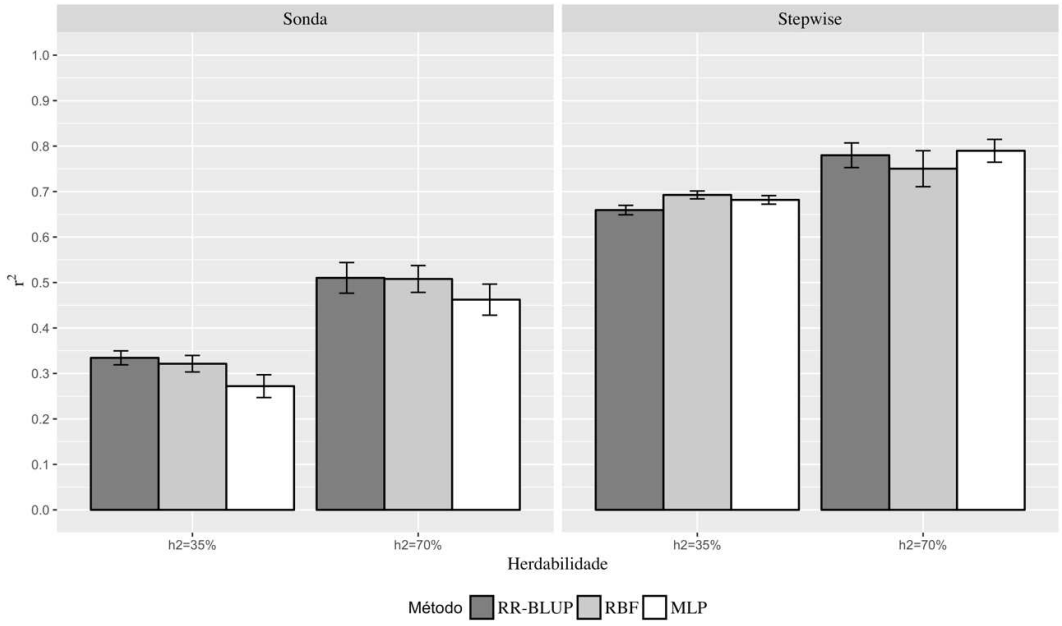

Figura 10. Resultados médios obtidos para a estatística r^2, em conjunto de dados de validação, pelas metodologias de RR-BLUP, MLP e RBF, após utilizar as duas estratégias de redução pelo Método da Sonda e Método de Regressão *Stepwise* considerando baixa herdabilidade (h^2 = 35%) e alta herdabilidade (h^2 = 70%).

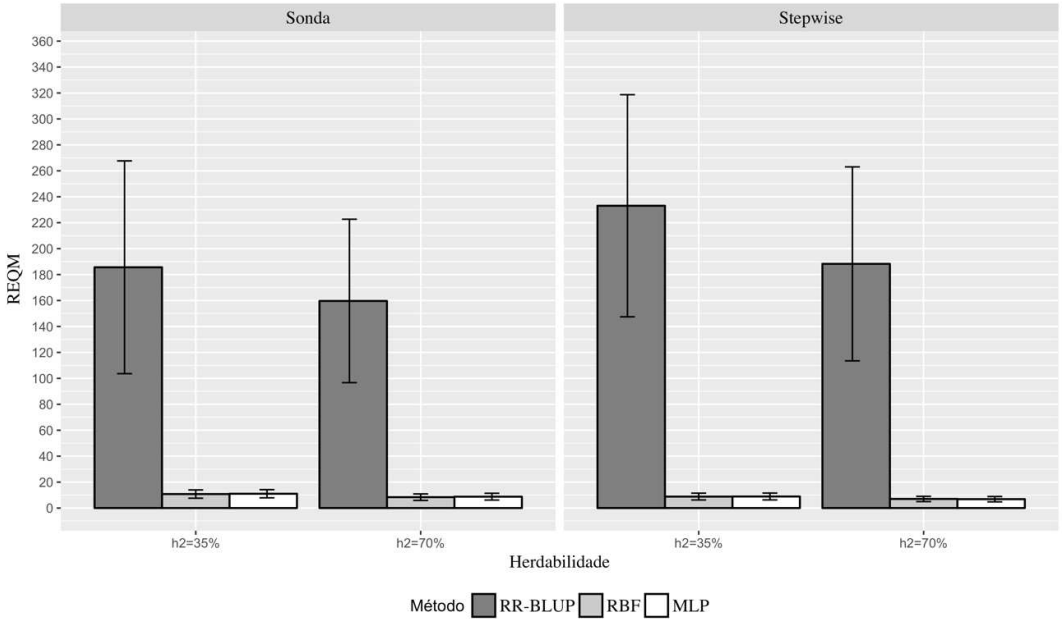

Figura 11. Resultados médios obtidos para a estatística $REQM$, em conjunto de dados de validação, pelas metodologias de RR-BLUP, MLP e RBF, após utilizar as duas estratégias de redução pelo Método da Sonda e Método de Regressão *Stepwise* considerando baixa herdabilidade (h^2 = 35%) e alta herdabilidade (h^2 = 70%).

Os resultados demonstraram que baixas herdabilidade dificultam o processo de predição, dada a diminuição dos valores de r^2 à medida que a herdabilidade diminui (Figura 10). As dificuldades em predizer caracteres de baixa herdabilidade já foram relatadas por diversos autores (CRUZ, 2012; HAYES ET AL., 2009; GODDARD, 2009).

A acurácia seletiva de ambas metodologias abordadas foi semelhante, no entanto, ao utilizar os arquivos reduzidos pela técnica da Sonda, percebe-se que as herdabilidades das características não puderam ser recuperadas na fase de validação dos métodos. Nesta etapa, os valores de r^2 médios obtidos para o RR-BLUP foram de 0,3343 e 0,5102 para h^2 de 35% e 70%, respectivamente; para o RBF foram de 0,3214 e 0,5078 para h^2 de 35% e 70%, respectivamente e para o método de MLP foram de 0,2721 e 0,4623 para h^2 de 35% e 70%, respectivamente (Tabela 6). A técnica de Regressão *Stepwise*, por sua vez, contribuiu para que as herdabilidades das características pudessem ser recuperadas (Tabela 7).

Tabela 6. Comparação média entre a eficiência do RR-BLUP, RBF e MLP após redução pelo Método da Sonda ao agrupar as características segundo a herdabilidade: h^2 = 35% ou h^2 = 70%.

Método	Modelo	r^2_t	r^2_v	$REQM_t$	$REQM_v$	CP	Viés
RR-BLUP	h^2 = 35%	0,5987	0,3343	154,8227	185,6346	0,5744	0,4662
	h^2 = 70%	0,7163	0,5102	143,4313	159,6843	0,7095	0,6143
RBF	h^2 = 35%	0,9024	0,3214	6,6042	10,7901	0,5637	0,5935
	h^2 = 70%	0,8916	0,5078	5,5276	8,4111	0,7086	0,7503
MLP	h^2 = 35%	0,6617	0,2721	8,8199	11,0394	0,5156	0,6940
	h^2 = 70%	0,7584	0,4623	6,8690	8,7889	0,6751	0,8201

r_t^2: quadrado da correlação de treinamento; r_v^2: quadrado da correlação de validação; $REQM_t$: erro quadrático médio de treinamento; $REQM_v$: erro quadrático médio de validação; CP: capacidade preditiva

Tabela 7. Comparação média entre a eficiência do RR-BLUP, RBF e MLP após redução pelo Método de Regressão *Stepwise* ao agrupar as características segundo a herdabilidade: h^2 = 35% ou h^2 = 70%.

Método	Modelo	r^2_t	r^2_v	$REQM_t$	$REQM_v$	CP	Viés
RR-BLUP	h^2 = 35%	0,8576	0,6593	210,1260	233,0717	0,8109	0,7722
	h^2 = 70%	0,9094	0,7799	176,2178	188,2732	0,8820	0,8564
RBF	h^2 = 35%	0,9022	0,6926	6,6056	8,8585	0,8315	0,8754
	h^2 = 70%	0,8917	0,7503	5,5283	7,0174	0,8639	0,9144
MLP	h^2 = 35%	0,8674	0,6817	7,1375	8,9214	0,8248	0,9095
	h^2 = 70%	0,9144	0,7897	5,4425	6,8516	0,8877	0,9441

r_t^2: quadrado da correlação de treinamento; r_v^2: quadrado da correlação de validação; $REQM_t$: erro quadrático médio de treinamento; $REQM_v$: erro quadrático médio de validação; CP: capacidade preditiva

Avaliando agora a acurácia preditiva, expressa pelo REQM, percebe-se que enquanto o RR-BLUP sofreu impactos negativos nesses valores para os cenários com herdabilidade baixa (h^2 = 35%), o impacto de tal fator perturbador sobre a estatística de *REQM* foi praticamente irrisório para os métodos de inteligência computacional de RBF e MLP propostos. O RR-BLUP propiciou *REQMv* de 185.6346 e 233.0717 para h^2 = 35% por meio do RR-BLUP/Sonda e RR-BLUP/*Stepwise*, respectivamente (Tabelas 6 e 7). As metodologias de RBF e o MLP reduziram esses valores para 10,7901 (RBF/Sonda), 8,8585 (RBF/*Stepwise*), 11,0394 (MLP/Sonda) e 8,9214 (MLP/*Stepwise*) (Tabelas 6 e 7).

Concentrando a avaliação da influência da herdabilidade sobre a acurácia das redes neurais, consideraram-se os pares de variáveis (V1 e V3), (V2 e V4), (V5 e V7), (V6 e V8), (V9 e V11) e (V10 e V12) percebeu-se que menores herdabilidades resultaram em maiores valores *REQMv* de modo que a inclusão de dominância e epistasia inflacionaram ainda mais o erro de predição. Para as características sem efeito de dominância e epistasia V1 e V3 os valores de *REQMv* foram de 4,4586 e 3,2824, respectivamente. A presença de sobredominância (gmd = 1,2) e o uso do modelo epistático na entanto, resultaram em valores de *REQMv* de 24,3591 e 18,7675 para V10 e V12 respectivamente para o método RBF/Sonda (Tabela 1 (anexo)). Resultados semelhantes foram obtidos ao usar a estratégia de Regressão *Stepwise* para redução (Tabela 2 (anexo)) e também para o método MLP (Tabelas 3 e 4 (anexo)).

Nesse capítulo propusemos o uso de duas metodologias de inteligência computacional baseadas em Redes Neurais Artificiais – RBF e MLP – para estudos de predição de características complexas. Essa abordagem demonstrou ser poderosa em um amplo espectro que comtemplou diferentes graus médios de dominância, ausência e/ou presença de epistasia e influência ambiental expressa pela herdabilidade, de modo que os resultados obtidos para o RBF e para o MLP evidenciaram a superioridade dessas metodologias em comparação com o método RR-BLUP/GWS.

Acredita-se que metodologias de redes neurais artificiais – por meio de suas redes de neurônios artificiais, funções de ativação e algoritmos de aprendizado – sejam capazes de capturar efeitos de fatores perturbadores negligenciados por outras metodologias e assim proporcionar resultados mais eficientes para estudos de adaptabilidade e estabilidade (TEODORO et al., 2015; BARROSO, et al., 2013; NASCIMENTO et al., 2013); estudos de análise discriminante (SANT`ANNA et al., 2015) e estudos de predição (SILVA et al., 2017; GIANOLA et al., 2016; SILVA et al., 2016; BHERING et al., 2015; SILVA et al., 2014). Glória et al. (2016) estudaram efeitos de marcadores e estimativas de herdabilidade a partir da predição do genoma por meio das redes neurais regularizadas

Bayesianas e concluíram que as redes neurais são ferramentas quantitativas promissoras para estudos de predição genômica. Beam et al. (2014) utilizaram Redes Neurais Bayesianas para detectar epistasia em estudos de associação genética para diversos cenários simulados e concluíram que as redes neurais constituem técnica poderosa para estudos de associação, sendo capazes de captar efeitos epistáticos; González-Camacho, et al. (2012) avaliaram as redes RFB e concluíram que essa metodologia é superior ao método de Lasso Bayesiano para estudos de predição de valores genéticos em dados simulados.

CONCLUSÕES

As metodologias RR-BLUP e de inteligência computacional (RBF e MLP) mostraram ser igualmente eficientes na predição de valores genéticos em características com controle gênico aditivo.

As metodologias de inteligência computacional baseadas em redes neurais artificiais de RBF e MLP propostas apresentam-se vantajosas para estudos de características complexas, com controle gênico envolvendo efeitos aditivos, dominantes e epistáticos, de interesse dos programas de melhoramento genético.

As metodologias baseadas em inteligência computacional apresentam acurácia preditiva, expressa por erro quadrático médio, superior à apresentada por RR-BLUP.

A redução do número de entradas (marcadores) demonstrou ser apropriada para fins de predição de valores genéticos.

REFERÊNCIAS

AZEVEDO, C.F.; SILVA, F.F.; RESENDE, M.D.V.; LOPES, M.S.; DUIJVESTEIJN, N.; GUIMARAES, S.E.F.; LOPES, P.S.; KELLY, M.J.; VIANA, J.M.S.; KNOL, E.F. Supervised independent component analysis as an alternative method for genomic selection in pigs. **J. Anim. Breed. Genet**. 131, 452–461, 2014.

BARROSO, L.M.A.; NASCIMENTO, M.; NASCIMENTO, A.C.C.; SILVA, F.F.; FERREIRA, R.P. Uso do método de Eberhart e Russell como informação a priori para aplicação de redes neurais artificiais e análise discriminante visando a classificação de

genótipos de alfafa quanto à adaptabilidade e estabilidade fenotípica. **Rev. Bras. Biom.** São Paulo, v.31, n.2, p.176-188, 2013.

BEAM, A.L., MOTSINGER-REIF, A., DOYLE, J., 2014. Bayesian neural networks for detecting epistasis in genetic association studies. **BMC Bioinform**. 15 (1), 368.

BHERING, L.L.; CRUZ, C.D.; PEIXOTO, L.A.; ROSADO, A.M.; LAVIOLA, B.G.; NASCIMENTO, M. Application of neural networks to predict volume in eucalyptus. **Crop Breeding and Applied Biotechnology**, v15, p.125-131, 2015.

BUENO, L.C.S.; MENDES, A.N.G.; CARVALHO, S.P. **Melhoramento genético de plantas: princípios e procedimentos**. Lavras: UFLA, 2001. 282p.

CAO, L.J.; CHUA, K.S.; CHONG, W.K.; LEE, H.P.; GU, Q.M. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. **Neurocomputing**, v.55, p.321 – 336, 2003.

CRUZ, C.D. Genes Software – extended and integrated with the R, Matlab and Selegen. **Acta Scientiarum. Agronomy**. Maringá, v. 38, n. 4, p. 547-552, Oct.-Dec., 2016.

CRUZ, C.D. **Princípios de genética quantitativa**. Viçosa: Ed. da UFV, 2012. 394p.

GIANOLA, D.; OKUT, H.; KENT A WEIGEL, K.A.; ROSA, G J.M. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. **BMC Genetics.** 12: 87, 2011.

GLÓRIA, L.S.; CRUZ, C.D.; VIEIRA, R.A.M.; RESENDE, M.D.V.; LOPES, P.S.; SIQUEIRA, O.H.G.B.D.; SILVA, F.F. Accessing marker effects and heritability estimates from genome prediction by Bayesian regularized neural networks. **Livestock Science**, 191, p:91–96, 2016.

GONZÀLEZ-CAMACHO, J.M.; DE LOS CAMPOS, G.; Pérez, P.; GIANOLA, D.; CAIRNS, J.E.; MAHUKU G., BABU, R.; CROSSA, J. Genome-enabled prediction of genetic values using radial basis function neural networks. **Theor. Appl. Genet.** 125:759–771, 2012.

HAHN, L.W.; RITCHIE, M.D.; MOORE, J.H. Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. **Bioinformatics**. v19, n3, pages 376–382, 2003.

HAYKIN, S. **Redes neurais: princípios e prática**. 2ed. Porto Alegre: Bookman, 2001.

HENDERSON, C.R. 1984. **Applications of linear models in animal breeding**. University of Guelph, Guelph. 462 p.

JANG, J.S.R; SUN, C.T; MIZUTANI, E. **Neuro fuzzy and soft computing**. Upper Saddle River, NJ: Prentice Hall, 1997.

LONG, N.; GIANOLA, D.; ROSA, G.J.; WEIGEL, K.A. Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. **Journal Anim. Breed Genet**. Aug;128(4):247-57, 2011.

MATLAB (2010). Matlab Version 7.10.0. Natick, Massachusetts: The Math Works Inc.

NASCIMENTO, M.; PETERNELLI, L. A.; CRUZ, C. D.; NASCIMENTO, A. C. C.; FERREIRA, R. de P.; BHERING, L. L.; SALGADO, C. C. Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. **Crop Breeding and Applied Biotechnology**. v.13, p.152-156, 2013.

NISHIO, M.; SATOH, M. Including Dominance Effects in the Genomic BLUP Method for Genomic Evaluation. **Plos One**, v.9, e85792, p.1-6, 2014.

PETEK, M.R.; SERA, T.; FONSECA, I.C.B. Prediction of genetic additive values for development of a coffee cultivar with increased rust resistance. **Bragantia,** 2008, v67 p:133-140.

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

RESENDE, M.D.V. **Matemática e estatística na análise de experimentos e no melhoramento genético**. Colombo: Embrapa Florestas, 2007. 561p.

RESENDE, M.D.V.; DUARTE, J.B. Precisão e controle de qualidade em experimentos de avaliação de cultivares. **Pesq Agropec Trop**. 37(3), p.182-194, 2007.

SANT'ANNA, I.C.; TOMAZ, R.S.; SILVA, G.N.; NASCIMENTO, M.; BHERING, L.L.; CRUZ, C.D. Superiority of artificial neural networks for a genetic classification procedure. **Genetics and Molecular Research**, v.14, p.9898-9906, 2015.

SILVA, G.N.; NASCIMENTO, M.; SANT'ANNA, I.C.; CRUZ, C.D.; CAIXETA, E.T.; CARNEIRO, P.C.S.; ROSADO, R.D.S.; PESTANA, K.N.; ALMEIDA, D.P.A.; OLIVEIRA, M.S. Artificial neural networks compared with Bayesian generalized linear regression for leaf rust resistance prediction in *Arabica coffee*. **Pesq. agropec. bras.**, Brasília, v.52, n.3, p.186-193, 2017.

SILVA, G.N.; TOMAZ, R.S.; SANT'ANNA, I.C.; CARNEIRO, V.Q.; CRUZ, C.D.; NASCIMENTO, M. Evaluation of the efficiency of artificial neural networks for genetic value prediction. **Genetic Molecular Research**, v.15, p.1-11, 2016. DOI: 10.4238/ gmr.15017676, 2016.

SILVA, G.N.; TOMAZ, R.S.; SANT'ANNA, I. de C.; NASCIMENTO, M.; BHERING, L.L.; CRUZ, C.D. Neural networks for predicting breeding values and genetic gains. **Scientia Agricola**, v.71, p.494-498.

SILVA, G.N. **Redes neurais artificiais: novo paradigma para a predição de valores genéticos**. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, 105p, 2014.

SILVA, I. N.; SPATTI, H. D.; FLAUZINO, R. A. **Redes Neurais Artificiais: para engenharia e ciências aplicadas**. São Paulo: Artliber, 2010. 399p.

TEODORO, P.E.; BARROSO, L.M.A.; NASCIMENTO, M.; TORRES, F.E., SAGRILO, E., dos SANTOS, A.; RIBEIRO, L.P. Redes neurais artificiais para identificar genótipos

de feijão-caupi semiprostrado com alta adaptabilidade e estabilidade fenotípicas. **Pesq. agropec. bras.** Brasília, v.50, n.11, p.1054-1060, 2015.

VENCOVSKY, R. **Princípios de genética quantitativa**. Piracicaba: Esalq, 1973.

CONCLUSÕES GERAIS

Os resultados obtidos pelo método RR-BLUP de Seleção Genômica Ampla destacam a necessidade do uso de modelos melhor parametrizados para estudos de predição de valores genéticos que envolvam características de controle gênico complexo e contemplam a inclusão de efeitos de dominância, epistasia e ambiental.

Além do efeito perturbador da dominância, epistasia e do ambiente sobre a acurácia do RR-BLUP, o número de marcadores moleculares, superior ao número de indivíduos, também afetou a eficiência do método, de modo que a redução de dimensionalidade para fins de estudos de Seleção Genômica Ampla, por meio dos Métodos de Regressão Stepwise e da Sonda propostos, propiciou menor demanda computacional e aumento na acurácia seletiva do RR-BLUP. No entanto, mesmo com a redução, foi evidenciado que o efeito perturbador da dominância, epistasia e do ambiente reduz a eficiência da GWS.

As metodologias de inteligência computacional de Redes Neurais Perceptron Multicamadas e Redes de Base Radial revelaram grandes potencialidades para a predição de valores genéticos que envolvam o estudo de características de controle gênico complexo, uma vez que apresentaram acurácia preditiva, expressa por erro quadrático médio, superior à apresentada pelo RR-BLUP.

ANEXO

Tabela 1. Estimativas obtidas pela Rede de Bas Radial (RBF) com cinco validações cruzadas após reduzir a matriz de marcas utilizando o método da Sonda.

Características		r_t^2	r_v^2	$REQM_t$	$REQM_v$	CP	Viés
V1	Md	**0,8993**	**0,2843**	**2,71**	**4,4586**	**0,5332**	**0,5622**
	dv	0,0026	0,0043	0,005	0,1219	0,0040	0,0046
V2	Md	**0,9011**	**0,3018**	**6,30**	**10,3088**	**0,5461**	**0,5754**
	dv	0,0036	0,0650	0,006	0,1687	0,0601	0,0644
V3	Md	**0,9122**	**0,5944**	**2,23**	**3,2824**	**0,7696**	**0,8059**
	dv	0,0015	0,0706	0,0017	0,1261	0,0456	0,0480
V4	Md	**0,7918**	**0,3926**	**6,39**	**8,3838**	**0,6182**	**0,6957**
	dv	0,0164	0,1224	0,0058	0,4111	0,1021	0,1199
V5	Md	**0,9110**	**0,3260**	**2,92**	**4,8653**	**0,5690**	**0,5962**
	dv	0,0027	0,0555	0,0012	0,0629	0,0474	0,0505
V6	Md	**0,9035**	**0,4043**	**9,3790**	**15,1352**	**0,6322**	**0,6652**
	dv	0,0026	0,0859	0,0118	0,3536	0,0674	0,0708
V7	Md	**0,9004**	**0,5698**	**2,4815**	**3,6120**	**0,7544**	**0,7950**
	dv	0,0017	0,0392	0,0046	0,1137	0,0262	0,0276
V8	Md	**0,9163**	**0,5229**	**7,7003**	**11,9907**	**0,7220**	**0,7544**
	dv	0,0038	0,0566	0,0060	0,4514	0,0398	0,0426
V9	Md	**0,9105**	**0,3255**	**3,36**	**5,6134**	**0,5673**	**0,5947**
	dv	0,0058	0,0721	0,0029	0,1715	0,0609	0,0658
V10	Md	**0,8890**	**0,2867**	**14,95**	**24,3591**	**0,5345**	**0,5670**
	dv	0,0041	0,0338	0,0200	0,3228	0,0317	0,0345
V11	Md	**0,8997**	**0,4756**	**2,91**	**4,4306**	**0,6872**	**0,7245**
	dv	0,0010	0,0786	0,0021	0,1709	0,0580	0,0614
V12	Md	**0,9290**	**0,4915**	**11,45**	**18,7675**	**0,7000**	**0,7263**
	dv	0,0019	0,0542	0,0099	0,3701	0,0379	0,0400

r_t^2 e r_v^2: referem-se aos coeficientes de determinação obtidos no treinamento e na validação, respectivamente; CP: capacidade preditiva; $REQM_t$ e $REQM_v$: raiz do erro quadrático médio para as fases da treinamento e validação, respectivamente.

Tabela 2. Estimativas obtidas pela Rede de Base Radial (RBF) com cinco validações cruzadas após reduzir a matriz de marcas utilizando o método de regressão *Stepwise*.

Características		r^2t	r^2v	$REQM_t$	$REQM_v$	CP	Viés
V1	Md	**0,8985**	**0,6931**	**2,7178**	**3,6023**	**0,8307**	**0,8765**
	dv	0,0024	0,0889	0,0027	0,1766	0,0546	0,0586
V2	Md	**0,9008**	**0,6707**	**6,3024**	**8,6498**	**0,8186**	**0,8626**
	dv	0,0035	0,0356	0,0038	0,1743	0,0220	0,0243
V3	Md	**0,9125**	**0,8216**	**2,2265**	**2,6631**	**0,9063**	**0,9488**
	dv	0,0019	0,0239	0,0036	0,0655	0,0132	0,0147
V4	Md	**0,7929**	**0,5562**	**6,3844**	**7,7413**	**0,7433**	**0,8353**
	dv	0,0171	0,0885	0,0095	0,3549	0,0613	0,0747
V5	Md	**0,9112**	**0,7102**	**2,9207**	**3,9613**	**0,8423**	**0,8824**
	dv	0,0023	0,0471	0,0038	0,1070	0,0280	0,0301
V6	Md	**0,9034**	**0,7103**	**9,3798**	**12,46**	**0,8426**	**0,8866**
	dv	0,0031	0,0277	0,0113	0,3413	0,0165	0,0179
V7	Md	**0,9004**	**0,7829**	**2,4815**	**3,0372**	**0,8846**	**0,9323**
	dv	0,0015	0,0303	0,0012	0,1218	0,0170	0,0179
V8	Md	**0,9161**	**0,7957**	**7,7049**	**9,6648**	**0,8917**	**0,9317**
	dv	0,0037	0,0438	0,0044	0,2969	0,0243	0,0270
V9	Md	**0,9102**	**0,7086**	**3,3662**	**4,5473**	**0,8414**	**0,8821**
	dv	0,0054	0,0418	0,0025	0,0908	0,0244	0,0280
V10	Md	**0,8891**	**0,6628**	**14,9466**	**19,930**	**0,8133**	**0,8627**
	dv	0,0038	0,0579	0,0264	0,7669	0,0359	0,0398
V11	Md	**0,8998**	**0,7758**	**2,9088**	**3,5524**	**0,8807**	**0,9284**
	dv	0,0010	0,0299	0,0016	0,1186	0,0168	0,0179
V12	Md	**0,9287**	**0,7696**	**11,4638**	**15,4453**	**0,87688**	**0,9100**
	dv	0,0021	0,0460	0,0084	0,5330	0,0261	0,028

r_t^2 e r_v^2: referem-se aos coeficientes de determinação obtidos no treinamento e na validação, respectivamente; CP: capacidade preditiva; $REQM_t$ e $REQM_v$: raiz do erro quadrático médio para as fases da treinamento e validação, respectivamente.

Tabela 3. Estimativas obtidas pela Rede Perceptron Multicamadas (MLP) com cinco validações cruzadas após reduzir a matriz de marcas utilizando o método da Sonda.

Características		$r^2{}_t$	$r^2{}_v$	$REQM_t$	$REQM_v$	CP	Viés
V1	Md	**0,6163**	**0,2016**	**3,8034**	**4,6959**	**0,4464**	**0,6285**
	dv	0,0108	0,0417	0,0189	0,0444	0,0485	0,0748
V2	Md	**0,6350**	**0,2505**	**8,7541**	**10,7184**	**0,4946**	**0,6818**
	dv	0,0221	0,0739	0,0879	0,1862	0,0766	0,1189
V3	Md	**0,8046**	**0,5755**	**2,7254**	**3,3275**	**0,7567**	**0,8787**
	dv	0,0104	0,0820	0,0333	0,1447	0,0544	0,0712
V4	Md	**0,7078**	**0,3351**	**6,1468**	**9,0436**	**0,5721**	**0,7550**
	dv	0,1467	0,0978	2,4259	0,7998	0,0879	0,1796
V5	Md	**0,6121**	**0,2598**	**4,2372**	**5,0393**	**0,5070**	**0,7158**
	dv	0,0129	0,0547	0,0127	0,0684	0,0528	0,0869
V6	Md	**0,7080**	**0,3868**	**12,3971**	**15,1046**	**0,6177**	**0,7866**
	dv	0,0178	0,0892	0,2118	0,5944	0,0723	0,1037
V7	Md	**0,7836**	**0,5267**	**3,0173**	**3,7288**	**0,7254**	**0,8583**
	dv	0,0108	0,0337	0,0316	0,0627	0,0232	0,0378
V8	Md	**0,7746**	**0,4698**	**9,8812**	**12,4249**	**0,6842**	**0,8146**
	dv	0,0094	0,0532	0,0765	0,2815	0,0400	0,0521
V9	Md	**0,7538**	**0,2714**	**3,6740**	**6,1947**	**0,5159**	**0,6567**
	dv	0,1750	0,0831	1,6406	0,4727	0,0730	0,1663
V10	Md	**0,6451**	**0,2624**	**20,0534**	**24,4833**	**0,5120**	**0,6945**
	dv	0,0092	0,0146	0,2129	0,7404	0,0144	0,0249
V11	Md	**0,7290**	**0,4441**	**3,7362**	**4,5116**	**0,6642**	**0,8245**
	dv	0,0160	0,0718	0,0544	0,1339	0,0549	0,0797
V12	Md	**0,75078**	**0,42267**	**15,70707**	**19,69683**	**0,6479**	**0,78926**
	dv	0,00779	0,06976	0,1188	0,35392	0,05384	0,06976

r_t^2 e r_v^2: referem-se aos coeficientes de determinação obtidos no treinamento e na validação, respectivamente; CP: capacidade preditiva; $REQM_t$ e $REQM_v$: raiz do erro quadrático médio para as fases da treinamento e validação, respectivamente.

Tabela 4. Estimativas obtidas pela Rede Perceptron Multicamadas (MLP) com cinco validações cruzadas após reduzir a matriz de marcas utilizando o método de Regressão *Stepwise*.

Características		r^2t	r^2v	$REQM_t$	$REQM_v$	CP	Viés
V1	Md	**0,8714**	**0,6803**	**2,885**	**3,6312**	**0,8229**	**0,9051**
	dv	0,0104	0,0906	0,045	0,18081	0,0559	0,0681
V2	Md	**0,8640**	**0,6613**	**6,826**	**8,66093**	**0,8129**	**0,8981**
	dv	0,0054	0,0378	0,055	0,20562	0,0233	0,0295
V3	Md	**0,9473**	**0,8731**	**1,962**	**2,44854**	**0,9343**	**0,9701**
	dv	0,0025	0,0199	0,022	0,06565	0,0107	0,0130
V4	Md	**0,8688**	**0,6898**	**5,703**	**7,07997**	**0,8291**	**0,9148**
	dv	0,0112	0,0799	0,087	0,36167	0,0491	0,0615
V5	Md	**0,8796**	**0,7093**	**3,153**	**3,95008**	**0,8415**	**0,9187**
	dv	0,0107	0,0573	0,061	0,16386	0,0339	0,0448
V6	Md	**0,8693**	**0,7006**	**10,128**	**12,58596**	**0,8368**	**0,9215**
	dv	0,0046	0,0334	0,100	0,28851	0,0201	0,0262
V7	Md	**0,9325**	**0,8291**	**2,252**	**2,86676**	**0,9104**	**0,9558**
	dv	0,0049	0,0296	0,042	0,13038	0,0162	0,0223
V8	Md	**0,9182**	**0,7918**	**7,656**	**9,7227**	**0,8896**	**0,9428**
	dv	0,0064	0,0335	0,083	0,25331	0,0188	0,0243
V9	Md	**0,8704**	**0,6897**	**3,694**	**4,6329**	**0,83**	**0,9126**
	dv	0,0067	0,0476	0,031	0,10632	0,0283	0,0358
V10	Md	**0,8497**	**0,6493**	**16,139**	**20,06761**	**0,8050**	**0,9009**
	dv	0,0114	0,0587	0,195	0,67094	0,0362	0,0492
V11	Md	**0,9131**	**0,7843**	**2,808**	**3,54205**	**0,8856**	**0,9427**
	dv	0,0043	0,0192	0,041	0,11671	0,0108	0,0138
V12	Md	**0,9066**	**0,7698**	**12,273**	**15,44969**	**0,8772**	**0,9382**
	dv	0,0046	0,0316	0,109	0,37834	0,0180	0,0209

r_t^2 e r_v^2: referem-se aos coeficientes de determinação obtidos no treinamento e na validação, respectivamente; CP: capacidade preditiva; $REQM_t$ e $REQM_v$: raiz do erro quadrático médio para as fases da treinamento e validação, respectivamente.

Tabela 5. Estimativas obtidas pelo RR-BLUP com cinco validações cruzadas após reduzir a matriz de marcas utilizando o método da Sonda.

Características		r_t^2	r_v^2	$REQM_t$	$REQM_v$	AIC	CP	Viés
V1	Md	**0,5755**	**0,3070**	**47,7098**	**49,092**	**3834,95**	**0,5536**	**0,4394**
	dv	0,0053	0,0305	1,72867	2,2939	8,76479	0,0279	0,0082
V2	Md	**0,5945**	**0,3017**	**84,534**	**107,264**	**5178,48**	**0,545**	**0,4602**
	dv	0,0242	0,0852	2,871	3,60153	6,78286	0,0764	0,0284
V3	Md	**0,7853**	**0,6037**	**60,5336**	**60,9465**	**3413,18**	**0,7755**	**0,6992**
	dv	0,0134	0,0844	5,48637	5,06006	15,0088	0,0542	0,018
V4	Md	**0,6080**	**0,3747**	**66,1947**	**78,2181**	**4888,31**	**0,6027**	**0,4837**
	dv	0,0274	0,1393	6,43357	13,3804	24,5256	0,1199	0,0345
V5	Md	**0,5732**	**0,3312**	**74,6042**	**75,7321**	**3997,29**	**0,5709**	**0,4415**
	dv	0,0258	0,0960	4,7012	4,8717	3,65932	0,0812	0,0313
V6	Md	**0,6651**	**0,4050**	**173,887**	**220,485**	**5789,78**	**0,6336**	**0,5406**
	dv	0,0224	0,0872	6,88002	24,8458	22,7409	0,0664	0,0277
V7	Md	**0,7587**	**0,5835**	**85,7335**	**86,1159**	**3555,78**	**0,7635**	**0,6677**
	dv	0,0107	0,0408	1,90702	3,24936	12,5052	0,0269	0,0134
V8	Md	**0,7411**	**0,5289**	**119,027**	**149,463**	**5456,55**	**0,7264**	**0,6437**
	dv	0,0117	0,0564	11,437	12,9544	12,1048	0,0394	0,0154
V9	Md	**0,5811**	**0,3412**	**82,5349**	**83,9852**	**4214,7**	**0,5790**	**0,4514**
	dv	0,0390	0,1061	11,3439	11,6866	12,2309	0,0866	0,0440
V10	Md	**0,6029**	**0,3197**	**465,666**	**577,249**	**6515,48**	**0,5646**	**0,4644**
	dv	0,0103	0,0373	11,1547	48,3748	12,1552	0,0328	0,0131
V11	Md	**0,6962**	**0,4806**	**114,492**	**115,167**	**3864,3**	**0,6899**	**0,5885**
	dv	0,0222	0,1032	6,22453	5,72198	18,9143	0,0757	0,0289
V12	Md	**0,7088**	**0,4898**	**414,607**	**468,195**	**6182,21**	**0,6993**	**0,6032**
	dv	0,0067	0,0450	76,202	59,9894	7,46703	0,0320	0,0089

r_t^2 e r_v^2: referem-se aos coeficientes de determinação obtidos no treinamento e na validação, respectivamente; CP: capacidade preditiva; $REQM_t$ e $REQM_v$: raiz do erro quadrático médio para as fases da treinamento e validação, respectivamente; AIC: Índice de Akaíke.

Tabela 6. Estimativas obtidas pelo RR-BLUP com cinco validações cruzadas após reduzir a matriz de marcas utilizando o método de regressão *Stepwise*.

Características		r_t^2	r_v^2	$REQM_t$	$REQM_v$	AIC	CP	Viés
V1	Md	**0,8624**	**0,6591**	**106,08**	**106,47**	**3762,07**	**0,8097**	**0,7789**
	dv	0,0141	0,1043	4,48	4,67	16,4859	0,0661	0,023
V2	Md	**0,8553**	**0,6393**	**103,07**	**118,14**	**5128,95**	**0,7993**	**0,7667**
	dv	0,0071	0,0317	23,11	22,3052	11,3255	0,0199	0,0116
V3	Md	**0,9474**	**0,8708**	**108,79**	**108,86**	**3213,47**	**0,9331**	**0,9166**
	dv	0,0017	0,0198	1,3214	1,098	5,84742	0,0107	0,0027
V4	Md	**0,8610**	**0,6751**	**39,15**	**54,39**	**4825,17**	**0,8199**	**0,7785**
	dv	0,0161	0,0942	2,196	5,439	16,4933	0,0592	0,0262
V5	Md	**0,8714**	**0,6892**	**159,75**	**160,16**	**3904,25**	**0,8295**	**0,7940**
	dv	0,0111	0,0643	2,999	2,779	17,834	0,0388	0,0181
V6	Md	**0,8560**	**0,6719**	**180,26**	**215,40**	**5738,78**	**0,8190**	**0,7717**
	dv	0,0054	0,0637	39,28	32,396	19,8113	0,0393	0,0090
V7	Md	**0,9301**	**0,8238**	**126,81**	**126,94**	**3390,14**	**0,9075**	**0,8890**
	dv	0,0036	0,0299	3,174	1,804	14,339	0,0165	0,0055
V8	Md	**0,9102**	**0,7778**	**77,38**	**107,29**	**5362,01**	**0,8817**	**0,8577**
	dv	0,009	0,0408	12,73	13,173	9,13637	0,0229	0,0141
V9	Md	**0,8611**	**0,6757**	**143,34**	**144,23**	**4120,12**	**0,8213**	**0,7797**
	dv	0,0097	0,0663	7,471	7,789	10,3152	0,0395	0,0161
V10	Md	**0,8392**	**0,6207**	**568,26**	**654,03**	**6469,88**	**0,7868**	**0,7419**
	dv	0,0156	0,0706	113,25	112,29	11,5767	0,0446	0,0253
V11	Md	**0,9104**	**0,7813**	**179,22**	**179,11**	**3726,57**	**0,8838**	**0,8582**
	dv	0,0050	0,0219	5,8910	6,082	9,33633	0,0123	0,0075
V12	Md	**0,8974**	**0,7506**	**525,96**	**553,06**	**6098,36**	**0,8660**	**0,8374**
	dv	0,0077	0,0444	84,94	83,77	12,4904	0,0259	0,012

r_t^2 e r_v^2: referem-se aos coeficientes de determinação obtidos no treinamento e na validação, respectivamente; CP: capacidade preditiva; $REQM_t$ e $REQM_v$: raiz do erro quadrático médio para as fases da treinamento e validação, respectivamente; AIC: Índice de Akaíke.