LEONARDO DE AZEVEDO PEIXOTO

REDES NEURAIS ARTIFICIAIS NA PREDIÇÃO DO VALOR GENÉTICO

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

VIÇOSA
MINAS GERAIS - BRASIL
2013

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T
P378r
2013

Peixoto, Leonardo de Azevedo, 1990-
Redes neurais artificiais na predição do valor genético / Leonardo de Azevedo Peixoto. – Viçosa, MG, 2013.
xi, 97 f. : il. ; 29 cm.

Inclui anexo.
Orientador: Leonardo Lopes Bhering.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Redes neurais (Computação). 2. Métodos de simulação. 3. Seleção de plantas - Melhoramento genético. I. Universidade Federal de Viçosa. Departamento de Fitotecnia. Programa de Pós-Graduação em Genética e Melhoramento. II. Título.

CDD 22. ed. 006.32

LEONARDO DE AZEVEDO PEIXOTO

REDES NEURAIS ARTIFICIAIS NA PREDIÇÃO DO VALOR GENÉTICO

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

APROVADA: 20 de março de 2013

Moyses Nascimento

Cosme Damião Cruz
(Coorientador)

Leonardo Lopes Bhering
(Orientador)

A Deus,

Por sempre me iluminar nas escolhas
que devo fazer em minha vida

OFEREÇO

Aos meus pais Marise e Vanildo,
por toda dedicação e carinho
em todas as horas da minha vida

DEDICO

AGRADECIMENTOS

A Deus, por sempre iluminar meu caminho e me ajudar a escolher o que é certo na minha vida.

Aos meus pais, Marise Aparecida Azevedo Peixoto e Vanildo Peixoto Lacerda por toda dedicação durante estes anos e por tudo que fizeram por mim até hoje, sempre torcendo e rezando para que eu conquiste aquilo que almejo.

A minha afilhada Lara da Silva Azevedo, que sempre me proporciona momentos de muitas alegrias.

A Universidade Federal de Viçosa (UFV), pela oportunidade oferecida para realização dos trabalhos.

A Conselho Nacional do Desenvolvimento Cientifico e Tecnológico (Cnpq), pela concessão da bolsa de estudos.

Ao professor Leonardo Lopes Bhering, pela sua orientação, amizade, disponibilidade, dedicação e ensinamento.

Ao professor Cosme Damião Cruz pela sua co-orientação, disponibilidade e dedicação para execução deste trabalho.

Ao professor Pedro Crescêncio Souza Carneiro, pela sua co-orientação, dedicação e ensinamento.

Ao professor Moyses Nascimento, pela amizade e disponibilidade em participar da banca contribuindo para melhoria deste trabalho.

A todos os professores e colegas do Programa de Pós-Graduação em genética e Melhoramento, em especial aqueles que contribuíram de alguma forma para a execução deste trabalho e para minha formação acadêmica.

Aos amigos dos laboratórios de Bioinformática e Biometria pelos momentos compartilhados.

Aos amigos Edson, Marcos, Vinicius, Edineia, Solange, Romário, Adésio, Paula, Caio, Ana Maria, Wanderson, Glaucio, que me proporcionaram momentos muito felizes e sempre me apoiaram e torceram por mim em mais esta etapa da minha vida.

As minha primas Kelly, Eduarda e Roberta, pela amizade, dedicação, apoio e pelos bons momentos vividos juntos.

A todos que de alguma forma contribuíram com amizade, dedicação e apoio para que eu chegasse ao final de mais esta etapa da minha vida, deixo aqui meus sinceros e eternos agradecimentos.

"Tudo passa, tudo passará, e nossa história não estará pelo avesso assim sem um final feliz. Teremos coisas bonitas para contar. E até lá vamos viver, temos muito ainda por fazer, não olhe para traz, apenas começamos. O mundo começa agora, apenas começamos". (Renato Russo)

BIOGRAFIA

LEONARDO DE AZEVEDO PEIXOTO, filho de Marise Aparecida Azevedo Peixoto e Vanildo Peixoto Lacerda, nasceu em Alegre, Espírito Santo, no dia 18 de abril de 1990.

No Município de Alegre, cursou o ensino primário na Escola Unidocente Benjamim Barros, de 1996 a 1999. Na Escola Estadual de Ensino Fundamental e Médio Aristeu Aguiar cursou o ensino fundamental de 2000 a 2003.

Em 2003, iniciou o ensino médio e o curso de Técnico em Agropecuária pela Escola Agrotécnica Federal de Alegre (EAFA), onde estudou até 2006.

Em 2007, iniciou a graduação em Agronomia pelo Centro de Ciências Agrárias da Universidade Federal do Espírito Santo (CCAUFES), colando grau em fevereiro de 2012.

Em março de 2012, iniciou a pós graduação em Genética e Melhoramento pela Universidade Federal de Viçosa.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

RESUMO

PEIXOTO, Leonardo de Azevedo, M.Sc. Universidade Federal de Viçosa, março de 2013. **Redes neurais artificiais na predição do valor genético.** Orientador: Leonardo Lopes Bhering. Coorientadores: Cosme Damião Cruz e Pedro Crescêncio Souza Carneiro.

Os objetivos do presente trabalho foram a partir das populações simuladas, fazer a replicação ou a ampliação de conjuntos populacionais, com as mesmas características pontuais de média, herdabilidade e coeficiente de variação e de estruturação (matriz de covariância ou de correlações), por meio da utilização da t écnica de decomposição espectral, e avaliar a eficiência da utilização das redes neurais artificiais na predição do valor genético em experimentos em blocos casualizados. O delineamento utilizado para simulação dos experimentos foi blocos casualizados contendo seis repetições. Foram simuladas 80 cenários, que possuíam valores estabelecidos para a média, a herdabilidade e coeficiente de variação experimental. Os experimentos simulados foram utilizadas para treinamento (experimento com 5000 genótipos) e validação (100 experimentos com 150 ou 200 genótipos para cada cenário). Para medir a eficiência da RNA na predição do valor genético comparou-se a correlação do valor de rede com o valor genético e a correlação do valor fenotípico com a valor genético. A média, herdabilidade, CV e a matriz de variância e covariância foram mantidos constantes em todos os experimentos simulados. Isso foi possível através da utilização da técnica de decomposição espectral. As redes neurais foram eficientes para predição do valor genético com ganho de 0,64 a 10,3% em relação ao valor fenotípico independente do tamanho de população utilizada, da herdabilidade ou do coeficiente de variação simulado. Assim concluiu-se que a preservação da matriz de variâncias e covariâncias de dados experimentais foi eficientemente realizada por meio do uso da decomposição espectral da matriz observada. Portanto os conjuntos de dados simulados podem ser utilizados para treinamento das RNAs mantendo as características da população original. Assim verificou-se que as RNAs é uma técnica mais eficiente na predição do valor genético em experimentos balanceados, quando comparado ao valor fenotípico (média).

ABSTRACT

PEIXOTO, Leonardo de Azevedo, M.Sc. Universidade Federal de Viçosa, March, 2013. **Artificial neural networks in prediction of genetics value.** Advisor: Leonardo Lopes Bhering. Co-advisors: Cosme Damião Cruz and Pedro Crescêncio Souza Carneiro.

The aim this present work were from the simulated population, make replication or enlargement of populations sets, with the same characteristics off from average, herdability and variation coefficient and organization (covariance or correlation matrix), by means of spectral decomposition, and evaluable efficiency of use artificial neural network in the prediction genetics value in randomized blocks experiments. The design used to simulation of the experiments was randomized blocks with six repetitions. They were simulated 80 conformation, when get established values of average, herdability and experimental variation coefficient. Experiments simulated were used to training (experiment with 5000 genotypes) and validation (100 experiment with 150n and 200 genotypes by each conformation). Thus to measurement the efficiency of the ANN in prediction of genetic value compared its correlation between network value with genetic value and correlation between phenotypic value with genetic value. Average, herdability, variation coefficient and variance and covariance matrix were maintained constant in all experiments simulated. It was possible through spectral decomposition techniques. Neural network were efficient for prediction of genetic value with gain 0,64 until 10,3% in relation phenotypic value regardless of size from populations, herdability or variation coefficient simulated. Thus concluded that preservation variance and covariance matrix of experimental data was efficiently performed by using spectral decomposition of the matrix observed. Therefore datasets simulated can be used to training of the ANNs maintaining original populations characteristics. Thus it was found that ANNs is a technique more efficient by prediction of genetic value in balanced experiments, when compared with phenotype value (average).

1. INTRODUÇÃO GERAL

A identificação de genótipos superiores requer métodos de seleção capazes de explorar eficientemente o material genético disponível, maximizando o ganho genético em relação às características de interesse (Oda et al., 2007). Diversos métodos de seleção têm sido empregados nos programas de melhoramento, com destaque para a seleção entre e dentro de famílias (Paula et al., 2002), a seleção combinada (Martins et al., 2005) e a seleção por modelos mistos pelo método best linear unbiased prediction (BLUP) (Garcia & Nogueira, 2005).

Ganhos genéticos adicionais que possibilitam o aperfeiçoamento de linhagens, híbridos e variedades comerciais têm se tornado cada vez mais difíceis, no contexto de espécies submetidas a longos processos seletivos. Assim, recursos extras, além daqueles pertinentes à escolha de delineamentos genéticos, métodos de seleção e boa experimentação agrícola, fazem parte de uma tendência recente: o uso de procedimentos analíticos mais refinados, como o emprego de modelos lineares mistos (Hiraoka et al., 2011) e redes neurais artificiais (Mugnai et al., 2008), por exemplo, para o estudo mais detalhado dos componentes da média e da variância de um caráter, para tentar predizer a parte herdável da variância, ou seja, o valor genético.

De maneira geral, a grande questão envolvida no melhoramento genético é o conhecimento do valor genético do indivíduo, para que se possa praticar a seleção com o máximo de acurácia. Para melhor predizer o valor genético de uma característica, é possível recorrer à informação fenotípica deste próprio indivíduo ou de seus aparentados (descendentes ou ancestrais) ou de informações sobre outras características correlacionadas. A junção de todas estas informações tem sido objeto de estudo por vários biometristas.

O valor genético é baseado no modelo aditivo, e tem desempenhado um papel importante no ganho de seleção de características complexas em plantas e animais (Crossa et al., 2010). Além deste modelo aditivo, têm-se utilizado BLUP, interações bayesianas (Piepho et al., 2008) e seleção genômica em plantas (Jannink et al., 2010) e animais (González-Recio et al., 2008).

Rosado et al. (2009) compararam vários métodos de seleção em *Eucaliptus urofila* e concluíram que os métodos de seleção combinada e BLUP proporcionaram maiores ganhos em relação aos métodos de seleção entre e dentro de famílias para todas as características avaliadas. Rocha et al. (2009) trabalhando com árvores de

Dipteryx alata concluíram que a seleção combinada e BLUP proporcionaram maiores ganhos com relação aos métodos de seleção massal e seleção entre e dentro. Li e Lindgren (2006), usando dados simulados, compararam a seleção individual e a seleção combinada. Eles concluíram que a utilização de índices é vantajosa em situação de populações grandes e características com baixa ou média herdabilidade. David et al. (2003) compararam quatro métodos e 10 intensidades de seleção em mudas de *Pinus resinosa*, e concluíram que a seleção combinada foi o método que proporcionou maiores ganhos e maximizou a diversidade genética.

Um método mais recente para tornar mais eficiente a seleção de famílias são as redes neurais artificiais (RNAs). As RNAs tem sido utilizadas por inúmeros autores para classificação de imagens em sensoriamento remoto (Aitkenhead & Aalders, 2008), análise de diversidade genética (Barbosa et al., 2011), identificação de genótipos superiores (Mugnai et al., 2008) e predição do valor genético em animais (Ventura et al., 2012). No entanto não existe relato na literatura sobre utilização das RNAs para predição do valor genético em experimentos balanceados em plantas.

1.1. MÉTODOS DE SELEÇÃO NO MELHORAMENTO

Uma das grandes contribuições da genética quantitativa é a indicação de estratégias de melhoramento que proporcionem avanços na direção desejada, em relação às características de interesse (Cruz et al., 2012). Nesse sentido, procura-se, desenvolver métodos de seleção que sejam mais eficientes em aumentar a frequência de genes favoráveis, em comparação com métodos 'clássicos', como a seleção massal e a seleção entre e dentro (Arnhold et al., 2009).

A identificação de genótipos superiores requer métodos de seleção capazes de explorar eficientemente o material genético disponível, maximizando o ganho genético em relação às características de interesse (Borém, 2007). Diversos métodos de seleção tem sido utilizados no melhoramento genético vegetal tais como, seleção entre e dentro de famílias (Paula et al., 2002; Martins et al., 2003; Martins et al., 2005), seleção combinada (Martins et al., 2001; Martins, et al., 2005) e seleção por modelos mistos pelo método BLUP (best linear unbiased prediction) (Garcia & Nogueira, 2005; Rocha et al., 2006; Rocha et al., 2007; Lopes et al., 2012).

Na seleção massal a eficiência seletiva depende da quantidade de variabilidade existente na população-base a ser explorada, da herdabilidade do

caráter a ser melhorado e da extensão do ganho genético deste caráter selecionado (Hogarth, 1971). Para diminuir um pouco o efeito ambiental sobre a seleção massal, este método foi modificado dividindo a área em faixas com características ambientais homogêneas dentro destas faixas. Este método é conhecido como seleção massal estratificada (Gardner, 1961). Outro método clássico é a seleção entre e dentro de famílias, que tem sido muito utilizada e tem apresentado, em geral, bons resultados (Matta & Viana, 2003; Vilarinho et al., 2003; Daros et al., 2004; Santos et al., 2004; Carvalho & Souza, 2007; Santos et al., 2008).

Um método alternativo a esses métodos clássicos é a seleção combinada, o que é baseada em um índice que leva em consideração, simultaneamente, o comportamento dos indivíduos e de suas famílias (Vencovsky & Barriga, 1992). Inúmeros trabalhos mostraram que a seleção combinada supera os demais métodos citados (Costa et al. 2000; Martins, et al. 2005; Bhering et al. 2013). Arnhold et al. (2009) trabalhando com milho pipoca e Bhering et al. (2013) trabalhando com genótipos de *Jatropha curcas* L., visaram comparar a eficiência relativa da seleção massal, seleção entre e dentro e seleção combinada. Os autores concluíram que a seleção combinada proporcionou maiores ganhos e é mais adequada para estas espécies em comparação com os outros métodos.

Atualmente, para o estudo de famílias tem se adotado o método dos modelos mistos REML/BLUP (REML é a máxima verossimilhança restrita, e BLUP, a melhor predição linear não viciada), que permite estimar os parâmetros genéticos e predizer os valores genotípicos das famílias (Resende, 2002). O BLUP consiste basicamente na predição de valores genéticos dos efeitos aleatórios do modelo estatístico associado às observações fenotípicas, ajustando-se os dados aos efeitos fixos e ao número desigual de informações nas parcelas por meio da metodologia de modelos mistos (Resende, 2002).

Oliveira et al. (2011) objetivando comparar a seleção via procedimento BLUP individual simulado (BLUPIS) versus seleção massal em famílias de irmãos-completos de cana-de-açúcar, concluíram que a seleção clonal via procedimento BLUPIS indica maior número de clones promissores para caracteres quantitativos dentro de famílias com elevados efeitos genotípicos. Rosado et al. (2009) trabalharam com famílias de meio irmãos de *Eucalyptus urophylla* com o objetivo de comparar vários métodos de seleção como seleção entre e dentro, seleção combinada e seleção baseada em modelos mistos (REML/BLUP). Os autores concluíram que a seleção combinada e a seleção por modelos mistos (BLUP)

proporcionam estimativas de ganhos significativamente maiores às obtidas com a seleção entre e dentro, e maior eficiência na escolha dos melhores indivíduos dentro da população.

1.2. PREDIÇÃO DO VALOR GENÉTICO

O sucesso de um programa de melhoramento genético depende da acurácia da predição do valor genético a partir de valores fenotípicos. As estimativas de parâmetros genéticos podem orientar um melhorista durante as três fazes principais de um programa de melhoramento: i) criação da variabilidade genética; ii) seleção de indivíduos superiores na população segregante; iii) utilização dos indivíduos selecionados no programa. Assim, tem sido usual e indispensável a quantificação dos parâmetros genéticos especialmente a herdabilidade que é a medida do quadrado da correlação entre o valor genético e a média fenotípica.

A seleção de genótipos superiores tem sido uma tarefa de difícil execução, uma vez que os caracteres de importância agronômica, em sua maioria quantitativos, apresentam base genética complexa, além de serem altamente influenciados pelo ambiente (Vieira et al., 2005).

A redução da influência ambiental sobre o valor fenotípico tem sido obtida por meio da condução de experimentos de forma zelosa e a adoção de delineamentos experimentais apropriados observando sempre os princípios básicos da casualização, repetição e controle local. Esses princípios básicos estão presente no delineamento em blocos ao acaso.

Os blocos ao acaso constituem o tipo mais importante de delineamento e de ampla utilização nos programas de melhoramento genético onde o controle local é representado pelos blocos (Ramalho, 2005). Mesmo em situações mais simples, como avaliação de genótipos em ensaios balanceados em blocos ao acaso é comum defrontar com caracteres de importância econômica que foram avaliados em experimentos com baixa precisão (alto coeficiente de variação) e baixa herdabilidade. Nestes ensaios, toma-se a média fenotípica das repetições (ou blocos) como sendo a medida mais apropriada como indicador da superioridade genética e prediz-se o ganho genético ponderando-se o diferencial de seleção praticado pela estimativa da herdabilidade.

O sucesso de um programa de melhoramento genético depende da acurácia da predição do valor genético a partir de valores fenotípicos. Assim é possível

identificar a partir dos valores fenotípicos, os indivíduos de valores genotípicos desejáveis e a maior concentração de alelos favoráveis.

A metodologia de modelos mistos pressupõe que os componentes de variância sejam conhecidos, o que raramente ocorre na prática. Quando tais componentes são desconhecidos, utiliza-se como estratégia de análise a substituição dos valores paramétricos dos componentes de variância por estimativas de máxima verossimilhança restrita (Reml), recebendo esse método o nome de Blup Empírico (Eblup) (Carneiro Júnior et al., 2010).

Estimativas dos componentes de variância e coeficientes de herdabilidade têm-se mostrado heterogêneas de acordo com diferentes níveis de produção e diferentes classes de desvio-padrão genético ou ambiental. Quando a heterogeneidade não é considerada, diferenças de variâncias dentro das subclasses podem resultar na predição de valores genéticos viesados, redução no progresso genético e desproporcional número de indivíduos selecionados de ambientes com diferentes variâncias (Weigel & Gianola, 1992).

Os métodos, que buscam eliminar a heterogeneidade de variâncias, consistem na transformação de dados ou na aplicação de fatores de ajustamento de forma que os dados transformados ou ajustados apresentem homogeneidade de variâncias. Cardoso et al. (2001) relatam que as transformações de dados podem conduzir a pressuposições nem sempre realísticas e que a utilização de fatores multiplicativos, apesar da facilidade computacional, pode produzir avaliações genéticas viesadas.

Dentro do contexto relativo ao processo de seleção, o conhecimento dos componentes de variância é de fundamental importância para estimar a herdabilidade, predizer o ganho genético e avaliar as potencialidades de uma população e a eficiência relativa dos diferentes métodos de melhoramento. ALém disto pode auxiliar a identificar a estratégia de seleção mais adequada (Hallauger & Miranda Filho, 1981).

Outro conjunto de informações extremamente útil é aquele obtido através dos cálculos de correlações. Segundo Ferreira et al. (2003), a correlação fenotípica fornece uma estimativa da influência conjunta de causas genéticas e ambientais na expressão de uma dada característica. Por sua vez, os valores de correlação genotípica (que corresponde à porção genética da correlação fenotípica) têm sido empregados para orientar programas de melhoramento genético, uma vez que eles refletem a fração da expressão fenotípica que é de natureza herdável.

O conhecimento da associação entre caracteres agronômicos e morfológicos pode ser primordial quando existe a necessidade de fazer a seleção de várias características simultaneamente. Além disso, ao selecionar caracteres de alta herdabilidade e de fácil aferição, e que evidenciem alta correlação com o caráter desejado, o melhorista poderá obter progressos mais rápidos em relação ao uso de seleção direta (Santos & Vencovsky, 1986).

Quando os caracteres são de baixa herdabilidade a correlação fenotípica pode ter pouca aplicabilidade, podendo induzir o melhorista a erros. Assim, é importante distinguir as causas genéticas e de ambiente que, combinadas, resultam na correlação fenotípica (Almeida, 1988).

De acordo com Hallauger & Miranda Filho (1981), a correlação tem importância no melhoramento de plantas, porque mede o grau de associação genética ou não genética entre dois ou mais caracteres. Cruz. et al. (2012) ressaltaram a importância das correlações, afirmando que elas quantificam a possibilidade de ganhos indiretos por seleção e que caracteres de baixa herdabilidade têm a seleção mais eficiente quando realizada sobre caracteres correlacionados.

1.3. REDES NEURAIS ARTIFICIAIS (RNAs)

A bioinformática é o termo dado para aplicação de técnicas computacionais na análise genética, podendo ser utilizada em qualquer situação de cultivo na qual se quer prever alguma coisa, reconhecer algum padrão ou utilizar uma técnica de análise (Ruggiero et al., 2003)

O primeiro modelo artificial de um neurônio biológico data de 1943, resultado do trabalho do neuroanatomista e psiquiatra Warren McCulloch e do matemático Walter Pitts (McCulloch & Pitts, 1943). Alguns anos depois o modelo Perceptron Múltiplas Camadas (Multi Layer Perceptron - MLP) (Figura 1) e o algoritmo backpropagation tornaram as redes neurais artificiais uma metodologia amplamente utilizada em várias áreas da ciência (Braga et al., 2007).

Segundo Braga et al. (2007) as RNAs são modelos matemáticos que se assemelham às estruturas neurais biológicas (neurônios) e que têm capacidade computacional adquirida por meio de aprendizado e generalização.

O aprendizado de uma rede neural é um processo pelo qual os parâmetros são adaptados pelo processo de estimulação no ambiente onde a rede está inserida (Haykin, 2001). Esta etapa pode ser considerada como uma adaptação da RNA às

características intrínsecas de um problema, onde se procura cobrir um grande espectro de valores associados às variáveis pertinentes. Isso é feito para que a RNA adquira, através de uma melhora gradativa, uma boa capacidade de resposta para o maior número de situações possíveis. Espera-se que uma RNA treinada tenha uma boa capacidade de generalização, independentemente de ter sido utilizado o aprendizado supervisionado ou não supervisonado durante o treinamento.

As RNAs diferenciam-se pela sua arquitetura e pela forma como os pesos associados às conexões são ajustados durante o processo de aprendizado. A arquitetura de uma rede neural é definida pelo número de camadas (camada única ou múltiplas camadas), pelo número de nós em cada camada, pelo tipo de conexão entre os nós (feedforward ou feedback) e por sua topologia (Haykin, 2001).

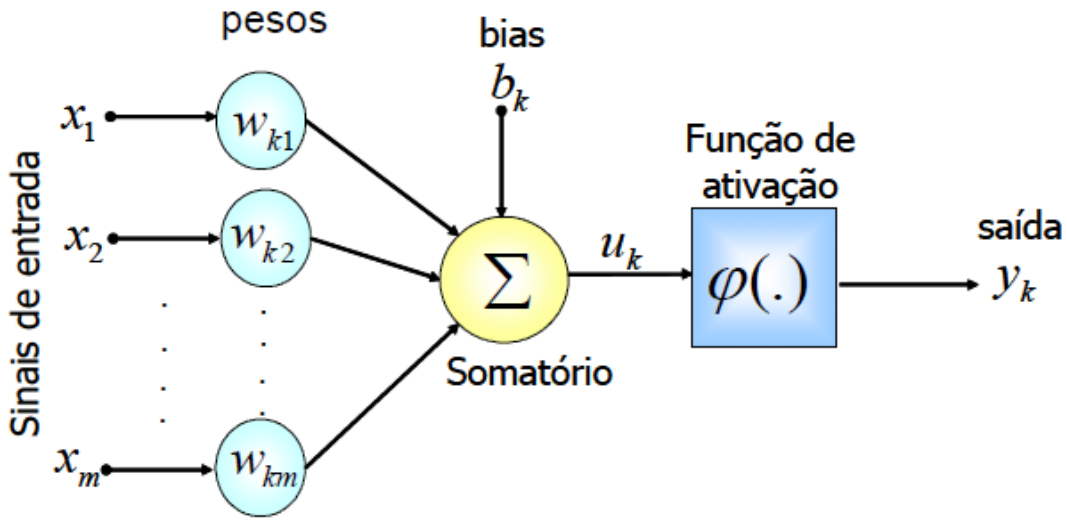

Figura 1. Modelo não linear de um neurônio artificial (Adaptado de Haykin, 2001). Onde x_1, x_2, ..., x_m são as entradas da rede; w_{k1}, w_{k2}, ..., w_{km} são os pesos, ou pesos sinápticos, associados a cada entrada; b_k é o termo bias; u_k é a combinação linear dos sinais de entrada; $\varphi(.)$ é a função de ativação e y_k é a saída do neurônio.

Pode-se dizer que é nos pesos que reside todo o conhecimento adquirido pela rede. Os pesos é que são os parâmetros ajustáveis que mudam e se adaptam à medida que o conjunto de treinamento é apresentado à rede. Assim, o processo de aprendizado supervisionado em uma RNA com pesos, resulta em sucessivos ajustes dos pesos sinápticos, de tal forma que a saída da rede seja a mais próxima possível da resposta desejada. Tipicamente, a ordem de amplitude normalizada da saída do neurônio está no intervalo [0, 1] ou alternativamente [−1, 1]. O modelo neural também inclui um termo chamado de “bias”, aplicado externamente, simbolizado por b_k. O b_k tem o efeito do acréscimo ou decréscimo da função de ativação na entrada da rede, dependendo se é positiva ou negativa, respectivamente.

As redes neurais artificiais (RNAs) têm a capacidade de aproximar funções não lineares, podendo, através de treinamento, mapear relações de entrada e saída. Essa habilidade permite, sem a necessidade de conhecer o parâmetro investigado, modelar um sistema conhecendo apenas os valores de entrada e saída.

As RNAs têm sido utilizadas de diferentes formas na agricultura como identificação e desenvolvimento de doenças em estágios iniciais e classificação de imagens de satélites (Chagas et al., 2009; França et al., 2010)

De acordo com Galvão et al. (1999) devido a sua estrutura não linear, as redes neurais podem trabalhar com características mais complexas, que nem sempre é possível com as técnicas tradicionais de estatística. Kuzmanovski and Aleksovska (2003) utilizaram RNAs para prever parâmetros celulares. Os autores utilizaram previsões em um mesmo arquivo de dados com características dependentes e independentes. As análises foram realizadas por redes neurais feedforward e cascade-forward e comparadas com o modelo de regressão linear múltipla pelo teste F. Os autores concluíram que as redes neurais possuem alta capacidade em relacionar características com distribuição não lineares, sendo superior a técnica de regressão linear múltipla. Delen et al. (2005) compararam redes neurais, árvore de decisão e regressão logística na predição de câncer de pulmão. Os autores concluíram que a árvore de decisão foi o mais acurado com 93,6%, seguido pela rede neural com 91,2% e pela regressão logística com 89,2% na predição do câncer de pulmão. Para Sudheer et al. (2003) a maior vantagem das redes neurais artificiais em relação aos métodos tradicionais é que eles não requerem informações detalhadas sobre os processos físicos do sistema a ser modelado.

A associação das redes neurais artificiais com métodos tradicionais é uma estratégia promissora, com grande vantagem de ser não-paramétrico e requererem pequenas amostras para treinamento (Kavzoglu & Mather, 2003) e tolerar a perda de dados (Bishop, 1995). Gallinari et al. (1991) apresentou resultados analíticos que estabeleciam um link entre análises discriminante e redes neurais multicamadas. Cheng e Titterington (1994) fizeram uma análise detalhada e compararam vários modelos de redes neurais e os métodos estatísticos convencionais. Ambos os trabalhos mostraram grande associação entre as redes Feedforward com as análises discriminantes e regressão. A rede neural artificial vem sendo usado em áreas de predição e de classificação, áreas onde modelos de regressão e outras

técnicas estatísticas tradicionais não conseguem obter bons resultados (Paliwal & Kumar, 2009)

Um grande número de estruturas diferentes pode ser gerado devido as diferentes possibilidades de conexões entre as camadas de neurônios. Para formação da rede é necessário que esta seja dividida em camadas. Usualmente as camadas são classificadas em três grupos: Camada de Entrada: onde os padrões são apresentados à rede; Camadas Intermediárias ou Ocultas: onde é feita a maior parte do processamento, através das conexões ponderadas, que podem ser consideradas como extratoras de características e Camada de Saída: onde o resultado final é concluído e apresentado (Figura 2).

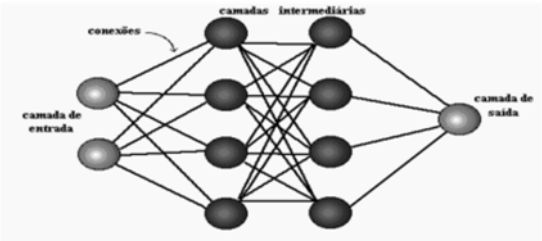

Figura 2. Representação das três camadas existentes em redes neurais.

1.3.1. Topologias das Redes Neurais

A priori, quanto mais camadas de neurônios, melhor seria o desempenho da rede neural, pois aumenta a capacidade de aprendizado, melhorando a precisão com que delimita regiões de decisão. Mas, na prática, aumentar o número de camadas intermediárias às vezes se torna inviável devido ao trabalho computacional.

Hechet-Nelsen (1989) afirma que com apenas uma camada intermediária na rede neural já é possível calcular uma função arbitrária qualquer a partir de dados fornecidos. De acordo com esses autores, a camada oculta deve ter por volta de (2i+1) neurônios, onde i é o número de variáveis de entrada.

Cybenko (1989), estudaram o número de camadas intermediárias necessárias à implementação de funções em RNAs. Seus resultados indicam que uma camada intermediária é suficiente para aproximar qualquer função contínua e que duas camadas intermediárias aproximam qualquer função matemática.

Lippmann (1987) afirma que, havendo uma segunda camada intermediária na rede neural, esta deve ter o dobro de neurônios da camada de saída; no caso de apenas uma camada oculta, deverá ter s(i+1) neurônios, onde s é o número de neurônios de saída e i, o número de neurônios na entrada.

1.3.2. Paradigmas de Aprendizagem

A propriedade que é de importância primordial para uma rede neural é a sua habilidade de aprender a partir de seu ambiente e de melhorar o desempenho através da aprendizagem. Uma rede neural aprende acerca do seu ambiente através de um processo interativo de ajustes aplicados a seus pesos sinápticos e níveis de bias.

Haykin (2001) define aprendizagem como um processo pelo qual os parâmetros livres de uma rede neural são adaptados através de um processo de estimulação pelo ambiente no qual a rede está inserida. Apesar de existirem diversos métodos (correção do erro, aprendizagem hebbiana, aprendizagem competitiva e aprendizagem boltzmann) para o treinamento de redes neurais, podem ser agrupados em dois grupos principais: os aprendizados supervisionado e não supervisionado. Existem ainda os aprendizados por reforço e por competição, mas o método de aprendizado supervisionado é o mais usado em RNAs, e possui este nome em função de que tanto entradas quanto saídas desejadas são fornecidas à rede durante o processo de treinamento Braga et al. (2007).

1.3.2.1. Aprendizado Supervisionado

Aprendizagem supervisionada, que é também denominada “aprendizagem com um professor” (Haykin, 2001), é um método desenvolvido pela disponibilização de um professor que verifique os desvios da rede a um determinado conjunto de dados de entrada.

Em termos conceituais, o professor é considerado como tendo conhecimento sobre o ambiente, sendo representado por um conjunto de exemplos de entrada-saída, (Haykin, 2001), como ilustrado na Figura 3. Entretanto, o ambiente é desconhecido pela rede neural de interesse.

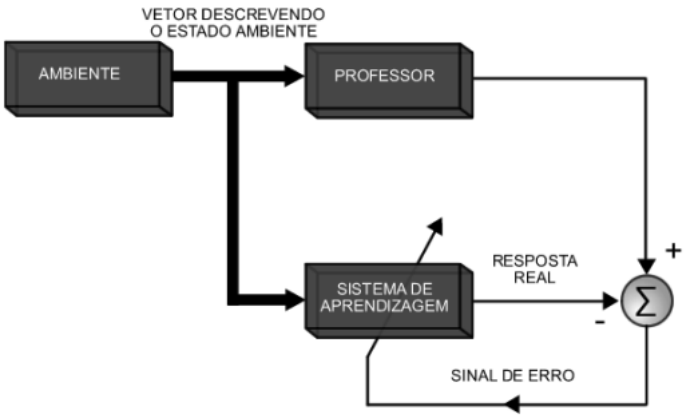

Figura 3. Diagrama em blocos de aprendizagem com o professor (Adaptado de Haykin, 2001).

Quando um vetor de treinamento é apresentado a uma rede neural que está usando o paradigma da aprendizagem supervisionada, em virtude do seu conhecimento prévio, o professor é capaz de fornecer à rede neural uma resposta desejada àquele vetor de treinamento. Na verdade, a resposta desejada representa a ação (saída) ótima a ser realizada pela rede neural. Assim os parâmetros (pesos sinápticos) da rede neural são ajustados através da influência combinada do vetor de treinamento e do sinal de erro indicado pelo professor. O sinal de erro é definido como a diferença entre a resposta desejada e a resposta real da rede. Este ajuste é realizado passo a passo, iterativamente, com o objetivo de fazer a rede neural utilizar o conhecimento do professor. Dessa forma, o conhecimento do ambiente disponível ao professor é passado para a rede através do treinamento. Ao alcançar esse objetivo, podemos, então, dispensar o professor e deixar a rede neural lidar sozinha com o ambiente.

Como medida de desempenho para o sistema, pode-se pensar em termos do erro médio quadrático ou da soma de erros médio quadráticos sobre a amostra de treinamento. O sinal de erro e(n) é dado por:

$$e(n) = d(n) – z(n)$$

onde d è a diferença entre a resposta desejada e z(n) é a saída da rede, em que o índice n varia no intervalo n = 1, 2, ..., N, sendo N o número total de exemplos de treinamento.

Para evitar que valores de erro com sinais positivos anulem valores com sinais negativos, toma-se o erro quadrático de cada amostra E(n), levando em consideração todas as saídas da rede dado por:

$$E(n) = \frac{1}{2} N_p \sum_{K=1}^{N_p} [e(n)]^2$$

Para uma análise geral do treinamento utiliza-se a média dos erros quadráticos das amostras de todo o conjunto de treinamento. Esta medida, denominada E(t), é dada por:

$$E(t) = \frac{1}{N} \sum_{n=1}^{N} E(n)$$

em que, t denota o número de épocas de treinamento. Uma época é contada a cada apresentação à rede neural de todos os exemplos do conjunto de treinamento.

1.3.3. Algoritmo da Retropropagação (Backpropagation)

O tipo mais utilizado de rede neural é o Perceptron múltiplas camadas (MLP) treinada com o algoritmo backpropagation (BP) ou retropropagação do erro (Patnaik & Mishra, 2000). A difusão deste algoritmo reporta da década de 1980, representando um marco na utilização das redes neurais e tendo o trabalho de Rumelhart et al. (1986) como uma das referências principais.

Para treinar a rede são utilizados vetores de entrada associados aos seus respectivos vetores de saída desejada, até que a rede aproxime uma determinada função e possa, a partir daí, oferecer saídas adequadas a vetores de entrada diferentes daqueles com os quais foi treinada. O backpropagation padrão é um algoritmo gradiente descendente, por meio do qual os pesos das conexões entre os neurônios são atualizados ao longo de um gradiente descendente de uma determinada função. O termo backpropagation refere-se à forma como o gradiente é calculado para redes de múltiplas camadas não lineares.

Basicamente, a aprendizagem por retropropagação de erro consiste em dois passos através das diferentes camadas da rede: um passo para frente, Feed-forward (a propagação), e um passo para trás, Feed-backward (retropropagação) (Haykin, 2001), como ilustrado na Figura 4.

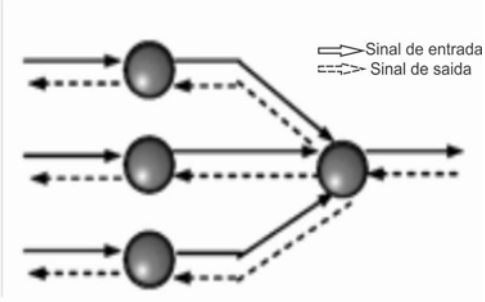

Figura 4. Representação Feed-Forward(Entrada) e Feed-backward(Erro) (Adaptado de Haykin, 2001).

O Feed-forward é um padrão de atividade (vetor de entrada) que se propaga pela rede, da camada de entrada até a camada de saída, e, finalmente, um conjunto de saídas é produzido como resposta real da rede. Durante esse passo os pesos sinápticos da rede são todos fixos;

O Feed-backward é a resposta real que é subtraída de uma resposta desejada (alvo) para produzir um sinal de erro. O erro se propaga na direção contrária ao fluxo de dados, indo da camada de saída até a primeira camada escondida, ajustando os pesos sinápticos das camadas.

O algoritmo básico pode ser desenvolvido tanto para o aprendizado incremental quanto para o aprendizado acumulativo. No modo de aprendizado incremental, os pesos são atualizados a cada apresentação de um novo padrão, mas tende a aprender melhor o último padrão apresentado. Já no aprendizado acumulativo, os pesos são ajustados apenas depois da apresentação de todos os padrões, ou seja, depois de um ciclo completo (após cada época) na apresentação dos padrões. Muitas das variações do algoritmo básico utilizam essa forma de aprendizado como tentativa de acelerar o processo de treinamento da rede.

1.3.3.1. O Perceptron Múltiplas Camadas (Multi Layer Perceptron - MLP)

Existem dezenas de diferentes modelos de RNA descritos na literatura, tais como MLP, Redes de função de Base Radial (RBF), Redes de Função Sample (SFNN), Redes de Fourier e Redes Wavelet. À primeira vista, o processo de seleção pode parecer uma tarefa difícil diante do número de possibilidades, porém é provável que poucos modelos forneçam uma solução excelente (Masters, 1994).

Segundo Sarle (1994), o modelo de uma RNA torna-se não linear quando é introduzido na rede uma camada intermediária, de forma que existam pesos estimados entre a camada de entrada e a camada oculta, e a função de ativação seja não linear (Logsig ou Tansig). O modelo resultante é chamado de Multi Layer Perceptron (Figura 5). De acordo com Haykin (2001), a MLP é uma extensão do perceptron simples, capaz de trabalhar com problemas não linearmente separáveis. Esse avanço foi possível pela utilização de, pelo menos, uma camada intermediária,

pois estas trabalham como um reconhecedor de características, que ficam armazenadas nos pesos sinápticos.

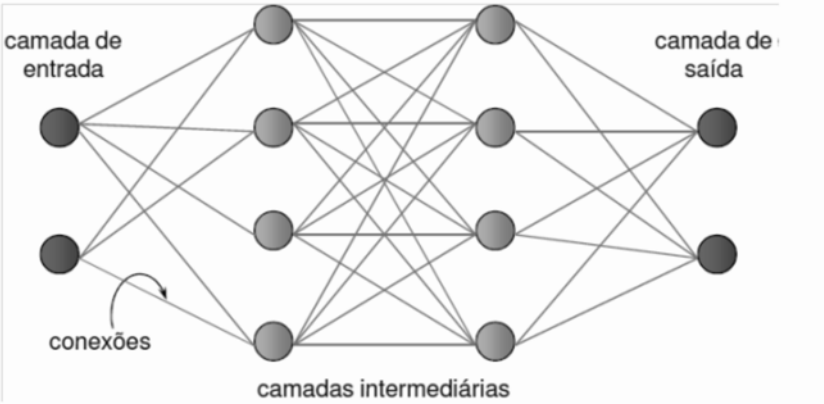

Figura 5. Representação de uma rede neural artificial tipo perceptron multi camadas (MLP).

Os perceptrons de múltiplas camadas têm sido aplicados com sucesso para resolver diversos problemas difíceis, através do seu treinamento de forma supervisionada com um algoritmo, conhecido como "algoritmo de retropropagação de erro" (error backpropagation) (Haykin, 2001). Um perceptron de múltiplas camadas tem três características distintivas:

1. O modelo de cada neurônio da rede inclui uma função de ativação não linear;

2. A rede contém uma ou mais camadas de neurônios ocultos, que não fazem parte da entrada ou da saída da rede;

3. A rede exibe um alto grau de conectividade, determinado pelas sinapses da rede.

É pela combinação dessas características, juntamente com a habilidade de aprender da experiência do treinamento, que a MLP deriva seu poder computacional. Essas mesmas características, entretanto, são também responsáveis pelas deficiências no estado atual de nosso conhecimento sobre o comportamento da rede por dois motivos principais: A presença de uma forma distribuída de não linearidade e a alta conectividade da rede tornam difícil a análise teórica de um perceptron de múltiplas camadas e a utilização de neurônios ocultos torna o processo de aprendizagem mais difícil de ser visualizado.

1.4. APLICAÇÃO DAS REDES NEURAIS ARTIFICIAIS NO MELHORAMENTO

Barbosa et al. (2011) trabalharam com 37 acessos de mamão (*Carica papaya* L.) e oito características quantitativas com o objetivo de avaliar uma estratégia alternativa para analises da diversidade genética. Eles comparam os resultados obtidos pelas redes neurais artificiais e pela análise discriminante de Anderson. A RNA classificou os acessos em 4 grupos. A análise discriminante de Anderson classificou 91,90% dos acessos nos grupos formados pela RNA. De acordo com a análise discriminante de Anderson, as redes neurais classificaram corretamente 94,44% no grupo 1, 100% no grupo 2, 88,89% no grupo 3 e 87,5% no grupo 4. Os autores concluíram que as redes neurais artificiais classificam de forma eficiente os acessos para estudo de diversidade genética.

Chen et al. (2010) visaram a classificação correta de variedades de milho através do uso de técnicas de processamento de imagens, análise discriminante, distância de Mahalanobis e redes neurais artificiais. Os autores avaliaram 750 grãos de milho (50 de cada variedade) e 58 características que foram extraídas através das técnicas de processamento de imagens. Essas características foram classificadas utilizando análise discriminante e os grãos foram classificados utilizando distância de Mahalanobis e redes neurais artificiais. As redes neurais artificiais conseguiram uma acurácia de 88% na classificação da variedade GAOYOU 115, 94% na variedade NONGDA 86 e 92% na variedade NONGDA 108. Os autores concluíram que a combinação da distância de Mahalanobis e redes neurais artificiais foram eficientes na classificação de variedades de milho. No entanto, novos estudos são necessários para estudar o desempenho do método para testar amostras de diferentes regiões e diferentes anos de produção da cultura.

Mugnai et al. (2008) utilizaram as redes neurais para avaliar as características morfológicas e identificar 25 acessos de *Camellia japonica* L. de uma série histórica. Os autores scanearam 50 folhas de cada acesso totalizando 1250 folhas. Como dados de entrada da rede foram utilizadas 14 características morfológicas. As análises mostraram que a rede neural artificial (Back propagation) pode ser eficiente para discriminar genótipos de *Camellia japonica* L.

Mancuso & Nicese (1999) utilizaram as redes neurais artificiais para diferenciar 10 cultivares de azeitona (*Olea europaea* L.), todas originárias do mediterrâneo. As análises foram baseadas em 17 características resultantes da análise de imagens. Esses parâmetros foram obtidos através da imagem scanneada de folhas colhidas em 10 plantas de cada cultivar. Foram utilizadas 40 folhas de

cada cultivar para realizar o treinamento. A melhor arquitetura de rede foi de 17 neurônios na camada de entrada, 20 na camada intermediária e 10 na camada de saída. As redes neurais conseguiram diferenciar todas as cultivares testadas. Os autores concluíram que as redes neurais artificiais podem ser utilizadas com grande eficiência para diferenciar acessos de azeitona através da utilização de parâmetros filométricos.

Neves (2007) estudou RNAs do tipo perceptron multicamadas para predição dos valores genéticos em características relacionadas à produção em gado de leite e comparou os resultados dessa metodologia com os obtidos pelo melhor preditor linear não viesado (BLUP), gerados por análises feitas por meio do programa Multiple Trait Derivative Free Restricted Maximum Likelihood (MTDFREML). Para os dados de entrada da rede, foram utilizados registros, coletados entre 1998 e 2002, referentes à produção de leite total, de gordura total, período de lactação e idade da vaca ao parto de 2.500 vacas da raça Pardo-Suíça. Por se tratar de uma rede com aprendizado supervisionado, foram utilizados como saída desejada os valores genéticos preditos pelo BLUP de cada uma dessas características. A comparação entre os resultados da RNA com os do BLUP apresentou uma correlação positiva moderada, indicando o potencial das RNAs na avaliação genética em bovinos de leite.

Meirelles (2005) estudou RNAs na avaliação e predição de valores genéticos para ganho de peso em bovinos de corte de raças zebuínas brasileiras. As características estudadas foram peso ao nascimento, peso ao desmame, peso ao sobreano e ganho de peso do desmame ao sobreano (GP345). Segundo a pesquisadora, os resultados apresentados indicaram o potencial das RNAs na avaliação genética, pois as RNAs apresentaram maior velocidade de obtenção dos dados e menor custo computacional necessário, fatos que justificam a continuidade dessa linha de pesquisa. Bertazzo (2006) pesquisou RNAs na predição do mérito genético de bovinos de corte, justificando seu uso em programas de melhoramento genético.

Ventura et al. (2012) avaliaram dados históricos de 19420 animais bovinos da raça Tabapuã provenientes de 152 fazendas localizadas em diversos estados brasileiros, nascidos entre 1976 e 1995. Esses dados foram utilizados para predição do valor genético do peso aos 205 dias de idade (VG_P205) por meio de redes neurais artificiais (RNAs) e usando o algoritmo LM (Levenberg Marquardt) para treinamento dos dados de entrada. Por se tratar de rede com aprendizado

supervisionado, foram utilizados, como saída desejada, os valores genéticos preditos pelo BLUP para a característica P205. Os valores genéticos do P205 obtidos pela RNA e os preditos pelo BLUP foram altamente correlacionados. A ordenação dos valores genéticos do P205 oriundos das RNAs e os valores preditos pelo BLUP (VG_P205_RNA) sugeriram que houve variação na classificação dos animais, indicando riscos no uso de RNAs para avaliação genética dessa característica. Inserções de novos animais necessitam de novo treinamento dos dados, sempre dependentes do BLUP.

1.5. REFERÊNCIAS BIBLIOGRÁFICAS

AITKENHEAD, M.; AALDERS, I. Classification of Landsat Thematic Mapper imagery for land cover using neural networks. **International Journal of Remote Sensing**, v. 29, n. 7, p. 2075-2084, 2008.

ALMEIDA, A. H. B. **Heterose e correlações em plantas branquícas e normais de jerimum caboclo (*Cucurbita maxima Duchesne*).** 1988. 96 f. Dissertação (Mestrado em Genética e Melhoramento) - Universidade Federal de Viçosa, Viçosa.

ARNHOLD, E.; VIANA, J. M. S.; SILVA, R. G.; MORA, F. Eficiências relativas de métodos de seleção de famílias endogâmicas em milho-pipoca. **Acta Scientiarum. Agronomy**, v. 31, n. 2, p. 203-207, 2009.

BARBOSA, C. D.; VIANA, A. P.; QUINTAL, S. S. R.; PEREIRA, M. G. Artificial neural network analysis of genetic diversity in *Carica papaya* L. **Crop Breeding and Applied Biotechnology**, v. 11, n. 3, p. 224-231, 2011.

BHERING, L. L.; BARRERA, C. F.; ORTEGA, D.; LAVIOLA, B. G.; ALVES, A. A.; ROSADO, T. B.; CRUZ, C. D. Differential response of *Jatropha* genotypes to different selection methods indicates that combined selection is more suited than other methods for rapid improvement of the species. **Industrial Crops and Products**, v. 41, p. 260-265, 2013.

BISHOP, C. M. **Neural networks for pattern recognition**. 1995.

BORÉM, A. **Biotecnologia florestal**. Universidade Federal de Viçosa, 2007.

BRAGA, A. P.; FERREIRA, A. C. P. L.; LUDERMIR, T. B. **Redes neurais artificiais: teoria e aplicações**. LTC Editora, 2007.

CARDOSO, F. F.; CARDELLINO, R. A.; CAMPOS, L. T. Fatores ambientais sobre escores de avaliação visual à desmama em bezerros Angus criados no Rio Grande do Sul. **Revista Brasileira de Zootecnia**, v. 30, n. 2, p. 318-325, 2001.

CARNEIRO JÚNIOR, J. M.; ASSIS, G. M. L. D.; EUCLYDES, R. F.; MARTINS, W. M. D. O.; WOLTER, P. F. Predição de valores genéticos utilizando inferência bayesiana e frequentista em dados simulados. **Acta Scientiarum Animal Sciences**, v. 32, n. 3, p. 337-344, 2010.

CARVALHO, H. W. L.; SOUZA, E. M. Ciclos de seleção de progênies de meios-irmãos do milho BR 5011 Sertanejo. **Pesq. agropec. bras.,** v. 42, n. 6, p. 803-809, 2007.

CHAGAS, C. S.; VIEIRA, C. A.; FERNANDES FILHO, E.; CARVALHO JUNIOR, W. Utilização de redes neurais artificiais na classificação de níveis de degradação em pastagens. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v. 13, n. 3, p. 319-327, 2009.

CHEN, X.; XUN, Y.; LI, W.; ZHANG, J. Combining discriminant analysis and neural networks for corn variety identification. **Computers and electronics in agriculture**, v. 71, p. S48-S53, 2010.

CHENG, B.; TITTERINGTON, D. M. Neural networks: A review from a statistical perspective. **Statistical Science**, v.9, n.1, p.2–30, 1994.

COSTA, R. B.; RESENDE, M. D. V.; ARAÚJO, A. J. Seleção combinada univariada e multivariada aplicada ao melhoramento genético da seringueira. **Pesq. agropec. bras.,** v. 35, n. 2, p. 381-388, 2000.

CROSSA, J.; DE LOS CAMPOS, G.; PEREZ, P.; GIANOLA, D.; BURGUENO, J.; ARAUS, J. L.; MAKUMBI, D.; SINGH, R. P.; DREISIGACKER, S.; YAN, J.; ARIEF, V.; BANZIGER, M.; BRAUN, H. J. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. **Genetics**, v. 186, n. 2, p. 713-724, 2010.

CRUZ, C. D.; REGAZZI, A. J.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético ed.5**. Viçosa, UFV**.** 480 p. 2012.

DAVID, A.; PIKE, C.; STINE, R. Comparison of selection methods for optimizing genetic gain and gene diversity in a red pine (*Pinus resinosa* Ait.) seedling seed orchard. **Theoretical and Applied Genetics**, v. 107, n. 5, p. 843-849, 2003.

DAROS, M.; AMARAL JR, A. T.; PEREIRA, M. G.; SANTOS, F. S.; GABRIEL, A. P. C.; SCAPIM, C. A.; FREITAS JR, S. P.; SILVÉRIO, L. Recurrent selection in inbred popcorn families. **Scientia Agricola**, v. 61, n. 6, p. 609-614, 2004.

DELEN, D.; WALKER, G.; KADAM, A. Predicting breast cancer survivability: A comparison of three data mining methods. **Artificial Intelligence in Medicine**, v.34, n.2, p.113–127, 2005.

FERREIRA, M.; QUEIROZ, M. A. D.; BRAZ, L. T.; VENCOVSKY, R. Correlações genotípicas, fenotípicas e de ambiente entre dez caracteres de melancia e suas implicações para o melhoramento genético. **Horticultura Brasileira**, v. 21, n. 3, p. 438-442, 2003.

FRANÇA, M. M.; FERNANDES FILHO, E. I.; DE LIMA, B. T. Análise do uso da terra no município de Viçosa-MG mediado por classificações supervisionadas com redes neurais artificiais e Maxver. **Revista Brasileira de Geografia Física**, v. 2, p. 92-101, 2010.

GALLINARI, P.; THIRIA, S.; BADRAN, F.; FOGELMAN-SOULIE, F. On the relations between discriminant analysis and multilayer perceptrons. **Neural Networks**, v.4, n.3, p.349–360, 1991.

GALVÃO, C. O.; VALENÇA, M. J. S.; VIEIRA, V. P. P. B.; DINIZ, L. S.; LACERDA, E. G. M.; CARVALHO, A. C. P. L. F.; LUDERMIR, T. B. **Sistemas inteligentes: Aplicações a recursos hídricos e ciências ambientais**. UFRGS: ABRH, 1999.

GARCIA, C.; NOGUEIRA, M. Utilização da metodologia REML/BLUP na seleção de clones de eucalipto. **Scientia Forestalis**, v. 68, p. 107-112, 2005.

GARDNER, C. An evaluation of effects of mass selection and seed irradiation with thermal neutrons on yield of corn. **Crop Science**, v. 1, p, 241-245, 1961.

GONZÁLEZ-RECIO, O.; GIANOLA, D.; LONG, N.; WEIGEL, K. A.; ROSA, G. J. M.; AVENDAÑO, S. Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. **Genetics**, v. 178, n. 4, p. 2305-2313, 2008.

HALLAUGER, A. R.; MIRANDA FILHO, J. B. **Quantitative genetics in maize Breeding**. Ames: Iowa State University Press, 1981. 468 p.

HAYKIN, S. **Redes Neurais: princípios e prática**. Porto Alegre: Bookman, 2001. 900 p.

HECHET-NELSEN, R. **Neurocomputing**. Boston: Addison-Wesley Longman, 1989.

HIRAOKA, Y.; KURAMOTO, N.; OHIRA, M.; OKAMURA, M.; TANIGUCHI, T.; FUJISAWA, Y. Estimation of genetic data and breeding values of traits related to wax production in *Rhus succedanea* L. clones using the REML/BLUP method. **Journal of Forest Research**, v. 16, n. 6, p. 509-517, 2011.

HOGARTH, J. **Sentencing as a human process**. University of Toronto Press Toronto, 1971.

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

KAVZOGLU, T.; MATHER, P. The use of backpropagating artificial neural networks in land cover classification. **International Journal of Remote Sensing**, v. 24, n. 23, p. 4907-4938, 2003.

KUZMANOVSKI, I.; ALEKSOVSKA, S.. Optimization of artificial neural networks for prediction of the unit cell parameters in orthorhombic perovskites: Comparison with multiple linear regression. **Chemometrics and Intelligent Laboratory Systems**, v.67, p.167–174, 2003.

LI, H.; LINDGREN, D. Comparison of phenotype and combined index selection at optimal breeding population size considering gain and gene diversity. **Silvae Genetica**, v. 55, n. 1, p. 13-18, 2006.

LIPPMANN, R. An introduction to computing with neural nets. **ASSP Magazine, IEEE**, v. 4, n. 2, p. 4-22, 1987.

LOPES, R.; CUNHA, R. N. V.; RESENDE, M. D. V. Produção de cachos e parâmetros genéticos de híbridos de caiaué com dendezeiro. **Pesquisa Agropecuária Brasileira**, v. 47, n. 10, p. 1496-1503, 2012.

MANCUSO, S.; NICESE, F. Identifying olive (*Olea europaea*) cultivars using artificial neural networks. **Journal of the American Society for Horticultural Science**, v. 124, n. 5, p. 527-531, 1999.

MARTINS, I. S.; CRUZ, C. D.; BARROS ROCHA, M. G.; REGAZZI, A. J.; PIRES, I. E. Comparação entre os processos de seleção entre e dentro e o de seleção combinada, em progênies de *Eucalytpus grandis*. **Cerne,** v. 11, n. 1, p. 16-24, 2005.

MARTINS, I. S.; CRUZ, C. D.; REGAZZI, A. J.; PIRES, I. E. Eficiência da seleção univariada direta e indireta e de índices de seleção em *Eucalyptus grandis*. **Revista Árvore**, v. 27, n. 3, p. 327-333, 2003.

MATTA, F. P; VIANA, J. M. S. Eficiências relativas dos processos de seleção entre e dentro de famílias de meios-irmãos em população de milho-pipoca. **Ciência e Agrotecnologia**, v. 27, n. 3, p. 548-556, 2003.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bull Math Biol**, v. 5, n. 4, p. 115-133, 1943.

MEIRELLES, F. D. P. **Modelo computacional de um rebanho bovino de corte virtual utilizando simulação Monte Carlo e redes neurais artificiais**. 2005, 104f. Tese (Doutorado em Zootecnia) – Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo, Pirassununga, SP.

MUGNAI, S.; PANDOLFI, C.; AZZARELLO, E.; MASI, E.; MANCUSO, S. *Camellia japonica* L. genotypes identified by an artificial neural network based on phyllometric and fractal parameters. **Plant systematics and evolution**, v. 270, n. 1, p. 95-108, 2008.

NEVES, A. R. M. **Aplicação de redes neurais artificiais na predição de valores genéticos em bovinos de leite da raça Pardo-Suiça**. 2007. (Dissertação Mestrado) - Universidade Federal do Pará, Belém , PA.

ODA, S.; MELLO, E. J.; SILVA, J. F.; SOUZA, I. C. G. Melhoramento florestal. In: BORÉM, A. (Ed.). **Biotecnologia Florestal**. Viçosa: UFV, 2007. p. 51-71.

OLIVEIRA, R. A.; DAROS, E.; DE RESENDE, M. D. V.; BESPALHOK-FILHO, J. C.; ZAMBON, J. L. C.; SOUZA, T. R.; LUCIUS, A. S. F. Procedimento Blupis e seleção massal em cana-de-açúcar. **Bragantia**, v. 70, n. 4, p. 796-800, 2011.

PALIWAL, M.; KUMAR, U. A. Neural networks and statistical techniques: A review of applications. **Expert Systems with Applications**, v.36, p. 2–17, 2009.

PATNAIK, A.; MISHRA, R. ANN techniques in microwave engineering. **Microwave Magazine, IEEE**, v. 1, n. 1, p. 55-60, 2000.

PAULA, R. C.; PIRES, I. E.; BORGES, R. C. G.; CRUZ, C. D. Predição de ganhos genéticos em melhoramento florestal. **Pesq. agropec. bras., Brasília**, v. 37, n. 2, p. 159-165, 2002.

PIEPHO, H.; MÖHRING, J.; MELCHINGER, A.; BÜCHSE, A. BLUP for phenotypic selection in plant breeding and variety testing. **Euphytica**, v. 161, n. 1, p. 209-228, 2008.

RAMALHO, M. A. P. **Experimentação em genética e melhoramento de plantas**. Ufla, 2005.

RESENDE, M. D. V. **Genética biométrica e estatística no melhoramento de plantas perenes**. Embrapa Informação Tecnológica, Colombo: Embrapa Florestas, 2002.

ROCHA, M. G. B.; PIRES, I. E.; ROCHA, R. B.; XAVIER, A.; CRUZ, C. D. Seleção de genitores de *Eucalyptus grandis* e de *Eucalyptus urophylla* para produção de híbridos interespecíficos utilizando REML/BLUP e informação de divergência genética. **Revista Árvore**, v. 31, 2007.

ROCHA, M. G. B.; PIRES, I. E.; XAVIER, A.; CRUZ, C. D.; ROCHA, R. B. Avaliação genética de progênies de meio-irmãos de *Eucalyptus urophylla* utilizando os procedimentos REML/BLUP e E (QM). **Ciência Florestal**, v. 16, n. 4, 2006.

ROCHA, R. B.; ROCHA, M. G. B.; SANTANA, R. C.; VIEIRA, A. H. Estimação de parâmetros genéticos e seleção de procedências e famílias de *Dipteryx alata* Vogel (baru) utilizando metodologia de REML/BLUP e E (QM). **Cerne**, n. 3, p. 331-338, 2009.

ROSADO, A. M.; ROSADO, T. B.; RESENDE JÚNIOR, M. F. R.; BHERING, L. L.; CRUZ, C. D. Ganhos genéticos preditos por diferentes métodos de seleção em progênies de *Eucalyptus urophylla*. **Pesqui. Agropecu. Bras**, v. 44, p. 1653-1659, 2009.

RUGGIERO, C.; J.F., D.; GOES, A.; NATALE, W.; BENASSI, A. C. Panorama da cultura do mamão no Brasil e no mundo: situação atual e tendências. In: MARTINS, D. S. (Ed.). **Papaya Brasil: qualidade do mamão para o mercado interno** ed. Vitória: INCAPER, 2003. p. 13-34.

RUMELHART, D. E.; HINTONT, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, v. 323, n. 6088, p. 533-536, 1986.

SANTOS, F. S.; AMARAL JÚNIOR, A. T.; JÚNIOR, F.; PAIVA, S.; RANGEL, R. M.; SCAPIM, C. A.; MORA, F. Genetic gain prediction of the third recurrent selection cycle in a popcorn population. **Acta Scientiarum. Agronomy**, v. 30, p. 651-655, 2008.

SANTOS, J.; VENCOVSKY, R. Correlação fenotípica e genética entre alguns caracteres agronômicos do feijoeiro (*Phaseolus vulgaris* L.). **Ciência e Prática**, v. 10, n. 3, p. 265-272, 1986.

SANTOS, J.; VIANA, J. M. S.; VILARINHO, A.; CÂMARA, T. Efficiency of S2 progeny selection strategies in popcorn. **Crop Breeding and Applied Biotechnology**, v. 4, n. 2, p. 183-191, 2004.

SARLE, W. S. **Neural networks and statistical models**. Proceedings of the Nineteenth Annual SAS Users Group International Conference, 1994. 13p.

SUDHEER, K.; GOSAIN, A.; RAMASASTRI, K. Estimating actual evapotranspiration from limited climatic data using neural computing technique. **Journal of Irrigation and Drainage Engineering**, v. 129, n. 3, p. 214-218, 2003.

VENCOVSKY, R.; BARRIGA, P. **Genética biométrica no fitomelhoramento**. Sociedade Brasileira de Genética Ribeirão Preto, 1992.

VENTURA, R.; SILVA, M.; MEDEIROS, T.; DIONELLO, N.; MADALENA, F.; FRIDRICH, A.; VALENTE, B.; SANTOS, G.; FREITAS, L.; WENCESLAU, R. Use of artificial neural networks in breeding values prediction for weight at 205 days in Tabapuã beef cattle. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 64, n. 2, p. 411-418, 2012.

VIEIRA, E. A.; COIMBRA, J. L. M.; FLOSS, I. P. V. E. L.; SILVA, I. B. G. O. Adaptabilidade e estabilidade em aveia em ambientes estratificados. **Ciência Rural**, v. 35, n. 2, 2005.

VILARINHO, A. A.; VIANA, J. M. S.; SANTOS, J. F.; CÂMARA, T. M. M. Eficiência da seleção de progênies S1 e S2 de milho-pipoca, visando à produção de linhagens. **Bragantia**, v. 62, n. 1, p. 9-17, 2003.

WEIGEL, K.; GIANOLA, D. Estimation of heterogeneous within-herd variance components using empirical Bayes methods: a simulation study. **J Dairy Sci**, v. 75, n. 10, p. 2824-2833, 1992.

2. OBJETIVO GERAL

Avaliar a eficiência na replicação e na ampliação de conjuntos populacionais a partir da população original, com as mesmas características pontuais de média, herdabilidade e coeficiente de variação e de estruturação (matriz de covariância ou de correlações), por meio da utilização da técnica de decomposição espectral, tendo em vista seu uso potencial em treinamento de redes neurais.

Avaliar a eficiência das redes neurais artificiais na predição do valor genético em experimentos em blocos casualizados, comparando a eficiência do valor de rede com a eficiência do valor fenotípico, e ainda definir uma estrutura de rede ideal para a predição do valor genético.

3. CAPITULO 1

SIMULAÇÃO E REPLICAÇÃO DE ESTRUTURAS DE DADOS EXPERIMENTAIS PARA FINS DE TREINAMENTO EM REDES NEURAIS ARTIFICIAIS

SIMULAÇÃO E REPLICAÇÃO DE ESTRUTURAS DE DADOS EXPERIMENTAIS PARA FINS DE TREINAMENTO EM REDES NEURAIS ARTIFICIAIS

3.1. RESUMO

O objetivo do presente trabalho foi simular populações a partir de características pré-estabelecidas (herdabilidade, coeficiente de variação e média). A partir das populações simuladas, fazer a replicação ou a ampliação de conjuntos populacionais, com as mesmas características pontuais de média, herdabilidade e coeficiente de variação e de estruturação (matriz de covariância ou de correlações), por meio da utilização da técnica de decomposição espectral, tendo em vista seu uso potencial em treinamento e validação de redes neurais. O delineamento utilizado para simulação dos experimentos foi blocos casualizados contendo seis repetições. Em cada experimento foram simulados 80 cenários. Para cada variável foram estabelecidos os valores da média (20, 40, 60, 80 e 100), da herdabilidade (10, 20, 30 e 40%) e do coeficiente de variação experimental (5 e 10%). A partir da decomposição espectral da matriz de variância e covariância foi possível simular outras populações como réplicas da primeira população simulada. E também simular populações com números de genótipos diferentes da original e manter as características genéticas. Para validação do processo de simulação, foram simulados experimentos com 5000 genótipos e seis blocos, e também foi simulado 100 réplicas com o mesmo número de genótipos da população inicial. Nas populações simuladas observou-se que a média, a herdabilidade, o CV e a matriz de variância e covariância foram mantidos constantes. Portanto a preservação da matriz de variâncias e covariâncias de dados experimentais pode ser eficientemente realizada por meio do uso da decomposição espectral da matriz original. Os Conjuntos de dados preservados ou ampliados podem ser apropriadamente gerados com potencial de uso, em redes neurais que demanda grande quantidade de informações para fins de treinamento e aprendizagem.

3.2. INTRODUÇÃO

O processo de simulação tem finalidade ampla e baseia-se na geração de dados experimentais para uso diverso. Na área da estatística experimental, a

simulação pode ser utilizada como a possibilidade de gerar ou replicar informações representativas das avaliações de um conjunto de genótipos em experimentos em delineamento apropriado, tais como o de blocos ao acaso em que são observados os princípios da casualização, aleatorização e controle local.

Cada característica simulada ou replicada deverá ter a propriedade de descrever uma variável com a mesma média, herdabilidade e precisão experimental da variável original ou de outra hipoteticamente referenciada (Cruz, 2006). Além das informações pontuais é indispensável que a simulação preserve a estrutura de variâncias e covariâncias dos dados originais ou hipotéticos para que fenômenos biológicos possam ser preservados e estudados com mais detalhes.

O processo de simulação de populações tem sido utilizado em inúmeras áreas da ciência, tais como análise de modelos não paramétricos (Gianola & De Los Campos, 2008), predição do valor genético em características quantitativas em animais (Gianola et al., 2010), mapeamento genético (Yu et al., 2008), detecção de QTL (Flint-Garcia et al., 2005) e redes neurais artificiais (Motsinger-Reif et al., 2008).

Yu et al. (2008) trabalharam com populações simuladas em milho para detecção de QTL. Para isto eles simularam oito populações com 625 indivíduos cada. Os autores concluíram que com 5000 indivíduos simulados de 30 a 79% dos QTLs foram encontrados. Motsinger-Reif et al. (2008) estudaram a eficiência das redes neurais artificiais para a detecção da interação gene-a-gene na epidemiologia de doenças. Eles comparam dados reais com dados simulados. Neste estudo para realizar a simulação foi mantido algumas características da população tais como a taxa de crossing over, taxa de mutação, tamanho da população, número máximo de gerações e tipo de seleção utilizada. O objetivo da simulação foi gerar um conjunto de dados com interação gene-a-gene para comparar o desempenho das metodologias utilizadas. Foram geradas populações com diferentes herdabilidades, número de polimorfismo na interação e o número de polimorfismo por indivíduo. Foram simulados indivíduos com 1000, 10000, 50000, 100000 e 500000 SNPs. Através destas populações simuladas, os autores concluíram que a metodologia GENN foi eficiente para a detecção da interação entre os genes.

Quando o objetivo da simulação inclui a necessidade de preservar a estrutura de covariância dos dados recomenda-se a análise de componentes principais que é uma técnica multivariada de modelagem da estrutura de covariância que foi introduzida por Pearson (1901). Componentes principais é um método estatístico que explora um conjunto de dados com grande número de mensurações,

reduzindo-os para poucos componentes principais para explicar determinado conjunto de dados (Reich et al., 2008). A análise de componentes principais tem sido muito utilizada para estudos de dados genéticos, principalmente em estudos de migrações humanas (Reich et al., 2008).

Os componentes principais de um grupo de k variáveis correlacionadas possibilitam estabelecer um grupo de k novas variáveis que são funções lineares das variáveis originais com as propriedades de serem não correlacionadas entre si e que conseguem explicar, o máximo de variação entre as k variáveis. O principal componente representa o maior eixo de variação entre as observações no espaço multidimensional; o segundo principal componente mostra a segunda maior variação entre as observações, e assim até o último componente principal que corresponde a menor variância (Baker et al., 1988). Assim, cada componente principal é uma combinação linear de um conjunto de covariáveis que permitem identificar as variáveis que mais contribuem na avaliação (Souza et al., 2010).

Para Chase et al. (2002), o componente principal tem sido usado para obter o posto das estimativas da matriz de covariância e obter também a decomposição dos seus vetores. Alternativamente componente principal é uma opção para se estimar os parâmetros fenotípicos e genéticos quando se estima novas variáveis. A melhor aproximação pode ser obtida estimando-os diretamente, e, ao mesmo tempo, restringindo-se aos mais importantes (Kirkpatrick & Meyer, 2004).

A técnica de componentes principais tem sido utilizada para resolver vários problemas na área cientifica como multicolineridade em regressão linear (Chao-Lung et al., 2010), estimação de fatores, modelagem da interação de fatores em experimentos sem repetição (AMML) (Kumar et al., 2012), estudos de divergência (Jombart et al., 2010) e agrupamento entre genótipos em estudos de genética e melhoramento de plantas e animais.

Khodadadi et al. (2011) avaliaram a diversidade genética de 36 cultivares de trigo proveniente de várias regiões do Ira através da técnica de agrupamentos. Para realizar os agrupamentos a matriz de distancia foi montada pelo quadrado da distancia euclidiana e pela análises de componentes principais (ACP). A ACP indicou que os cinco primeiros componentes explicavam mais de 97% da variância genética dos genótipos. Pela análise de agrupamento utilizando o quadrado da distancia euclidiana houve formação de sete grupos e pela ACP formou-se seis grupos. Os autores concluíram que a utilização dos componentes principais para formação de grupos em trigo é uma técnica com potencial.

Outra técnica matemática de grande importância é a decomposição espectral que consiste em obter diferentes estruturas para as matrizes de covariância. A decomposição espectral consiste em decompor a matriz de covariância nos autovalores e autovetores correspondentes (Soares & Mendonça, 2003).

Eeuwijk et al. (2010) visaram identificar a adequação da metodologia de modelos mistos atualmente executada no padrão PCs para mapeamento de QTLs em programas de seleção de híbridos e estudar a viabilidade do REML e dos modelos mistos através das técnicas Bayesianas para o mapeamento de QTL. Os autores utilizaram a decomposição espectral da matriz de variância Q para garantir que a matriz seja positiva definida. Assim a matriz Q foi decomposta em UU^t. Com a decomposição espectral foi possível descartar os auto vetores que possuíam autovalores negativos. Portanto a matriz Q foi recalculada $Q=U^*U^{*t}$, onde U* é uma matriz com apenas os autovetores positivos. Assim foi possível reproduzir a matriz Q com número inferior de linhas e colunas da matriz original.

As características dos dados de treinamento das redes neurais artificiais são importantes para o desempenho da rede. Os dados de treinamento das RNAs devem ser definido pelo analista como sendo dados representativos da sua população ou do experimento em que se pretende avaliar (Kavzoglu, 2009). A qualidade e o tamanho do arquivo de dados são de suma importância para o treinamento das RNAs. Para a predição do valor genético, dados com boa qualidade são dados que mantém as características da população que se pretende avaliar. Isto é possível através de séries históricas (banco de dados) (Ventura et al., 2012) e simulação de experimentos. Com relação ao tamanho do arquivo de treinamento Kavzoglu (2001) relata que arquivos de pequenas dimensões não são suficientes para que a rede possa reconhecer todas as classes possíveis, diminuindo o desempenho da rede.

A utilização de dados simulados podem otimizar a pesquisa tornando-a mais rápida nos testes iniciais, e depois que o método for eficiente para dados simulados, pode-se utiliza-los para dados reais. O uso de dados simulados em redes neurais é uma alternativa interessante, pois estes procedimentos requerem uma fase de treinamento em que o conhecimento prévio de dados possibilita o aprendizado eficaz das redes. Os dados para fins de treinamento são obtidos, normalmente, por bancos de dados históricos, mas em sua falta, o uso de dados simulados pode ser uma alternativa viável. Assim, o objetivo do presente trabalho foi simular populações

a partir de características pré-estabelecidas (herdabilidade, coeficiente de variação e média). A partir das populações simuladas, fazer a replicação ou a ampliação de conjuntos populacionais, com as mesmas características pontuais de média, herdabilidade e coeficiente de variação e de estruturação (matriz de covariância ou de correlações), por meio da utilização da técnica de decomposição espectral, tendo em vista seu uso potencial em treinamento de redes neurais.

3.3. MATERIAL E MÉTODOS

3.3.1. Simulação dos experimentos

O delineamento utilizado para simulação dos experimentos foi blocos casualizados contendo seis repetições. Foram simulados 80 cenários (experimentos). Em cada cenário foram estabelecidos os valores da média (20, 40, 60, 80 e 100), herdabilidade (10, 20, 30 e 40%), coeficiente de variação experimental (5 e 10%) e número de genótipos avaliados (150 e 200 genótipos por bloco). As interações entre essas características são mostradas na Tabela 1. Com o objetivo de validar as redes neurais artificiais (RNAs), foram simulados experimentos para um mesmo valor de média, herdabilidade, coeficiente de variação experimental e matriz de variância e covariância. O modelo estatístico utilizado no processo de simulação foi:

$$Y_{ij} = \mu + G_i + B_j + \varepsilon_{ij}$$

em que

Y_{ij}: observação simulada de uma dada característica;

μ: média geral da característica, cujos valores estão na Tabela 1;

G_i: efeito associado ao i-ésimo genótipo, sendo $G_i \sim N(0, \sigma^2_g)$;

B_j: efeito associado ao j-ésimo bloco;

ε_{ij}: erro aleatório, sendo $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Tabela 1. Estimativas dos parâmetros que foram estabelecidos para definir as populações simuladas.

Número de indivíduos por população					
150			200		
Média	h^2	CV (%)	Média	h^2	CV (%)
20	10	5	20	10	5
20	10	10	20	10	10
20	20	5	20	20	5
20	20	10	20	20	10

20	30	5	20	30	5
20	30	10	20	30	10
20	40	5	20	40	5
20	40	10	20	40	10
40	10	5	40	10	5
40	10	10	40	10	10
40	20	5	40	20	5
40	20	10	40	20	10
40	30	5	40	30	5
40	30	10	40	30	10
40	40	5	40	40	5
40	40	10	40	40	10
60	10	5	60	10	5
60	10	10	60	10	10
60	20	5	60	20	5
60	20	10	60	20	10
60	30	5	60	30	5
60	30	10	60	30	10
60	40	5	60	40	5
60	40	10	60	40	10
80	10	5	80	10	5
80	10	10	80	10	10
80	20	5	80	20	5
80	20	10	80	20	10
80	30	5	80	30	5
80	30	10	80	30	10
80	40	5	80	40	5
80	40	10	80	40	10
100	10	5	100	10	5
100	10	10	100	10	10
100	20	5	100	20	5
100	20	10	100	20	10
100	30	5	100	30	5
100	30	10	100	30	10
100	40	5	100	40	5
100	40	10	100	40	10

Média é a média da característica (20, 40, 60, 80 e 100), h^2 é a herdabilidade (10, 20, 30 e 40%), CV é o coeficiente de variação (5 e 10%). Foram simulados dois tamanhos de experimentos em blocos ao acaso com seis repetições (150 e 200 genótipos por população simulada).

Para simulação de dados, com distribuição normal, foram utilizadas as variáveis, proposto pelo teorema de Box-Muller:

$$x = \sqrt{-2\log_e(RND)}V\cos(2\pi RND)$$

e

$$y = \sqrt{-2\log_e(RND)V}\,Sen(2\pi RND)$$

sendo RND um número aleatório. Demonstra-se que os valores de x e y, assim obtidos, têm distribuição normal com média zero e variância V. Como o processo de simulação demandou-se a obtenção de n dados com média (μ) e variância (σ^2), foi utilizada a estratégia para gerar estes dados (z) por meio da seguinte expressão:

$$z = \mu + \frac{1}{2\theta}\sum_{i=1}^{\theta}(x_i + y_i)$$

sendo:

$$V = 2\theta\sigma^2$$

em que θ representa a repetibilidade de cada ponto simulado. Quanto maior o valor de θ estabelecido pelo usuário, mais precisa torna-se a simulação, porém mais lenta.

3.3.2. Simulação dos efeitos de blocos

Um conjunto de dados contendo n valores, em progressão aritmética, de razão r e média $\overline{X}$, em que o primeiro termo é denotado por X_1 e o último por X_n, a variância é dada por:

$$S^2 = \frac{n(n+1)}{3(n-1)^2}(X_n - \overline{X})^2$$

Assim, para estimar os efeitos de bloco foi admitida a existência de b efeitos fixos, cujos valores configuram uma progressão aritmética de razão r com a particularidade de que $B_1 = -B_b$ e $\overline{B} = 0$. Logo, o valor B_b é estimado por meio de:

$$B_b = \frac{(n-1)\sqrt{3\phi_b}}{\sqrt{n(n+1)}}$$

e os demais efeitos estabelecidos considerando a razão da progressão aritmética dada por:

$$r = \frac{B_b - B_1}{b-1}$$

3.3.3. Simulação dos efeitos de genótipos

Para estimar os efeitos de genótipos foi necessário conhecer o valor da variância genética, que foi obtida a partir das informações de herdabilidade (h^2) e o

coeficiente de variação experimental (CV_e). Assim, primeiro é obtido o valor da variância ambiental por meio de:

$$\sigma^2 = \left(\frac{\mu CV_e}{100}\right)^2$$

Onde

σ^2 é a variância ambiental.

A herdabilidade (h^2) é calculada por:

$$h^2 = \frac{100\sigma_G^2}{\sigma_G^2 + \frac{1}{b}\sigma^2}$$

Onde

σ_G^2 é a variância genotípica

logo,

$$\sigma_G^2 = \frac{\sigma^2 h^2}{b(100 - h^2)}$$

Neste caso foi considerado que $G_i \sim NID(0, \sigma_G^2)$. Como o valor de σ_G^2 era conhecido, estimou-se os efeitos usando a função randômica descrita anteriormente,

3.3.4. Simulação dos erros aleatórios

Foi considerado que $\varepsilon_i \sim NID(0, \sigma^2)$. Como foi fornecido o valor do coeficiente de variação experimental e da média da característica, o valor de σ^2 torna-se conhecido e, portanto, os erros aleatórios e independentes podem ser estimados usando a função randômica descrita anteriormente.

3.3.5. Estabelecimento dos valores fenotípicos e genotípicos

Conhecidos o valor da média da característica e dos efeitos envolvidos, os valores fenotípicos, de cada variável, foram estabelecidos por meio do modelo:

$$Y_{ij} = \mu + G_i + B_j + \varepsilon_{ij}$$

Os valores genotípicos foram obtidos por meio de:

$$Z_i = \mu + G_i$$

As simulações dos experimentos foram realizados utilizando a função simulação de ensaios em DBC (Integração - RNA - Predição do valor genético) no aplicativo computacional GENES (Cruz, 2006).

3.3.6. Obtenção dos componentes principais e decomposição espectral

Seja Y_{ij} a média padronizada no j-ésimo caráter (j = 1,2 3,..,n) avaliada no i-ésimo genótipo (i= 1,2,3,..,p) e R a matriz de covariância entre as médias fenotípicas dos blocos (ou a matriz de correlações fenotípicas entre as médias fenotípicas dos blocos, com base nos dados originais).

A utilização da técnica de componentes principais consistiu em transformar o conjunto das médias fenotípicas (Y_{ij}) dos genótipos simulados dos 6 blocos e do valor genético em um novo conjunto Y onde os Y's são funções lineares dos X's e independentes entre si (covariância nula).

Sendo Y_{i1} o primeiro componente principal, sua variância foi dada por:

$$V(Y_{i1})=V(Y_1)=\sum_{j}\sum_{j'} a_j a_{j'} r_{jj'}$$

onde $r_{jj'}$ é o elemento da j-ésima linha e da j'-ésima coluna da matriz R (matriz de variância e covariância).

Sob a forma matricial temos:

$$V(Y_1)=a'Ra$$

em que a' é um vetor 1 x n de elementos a_j (j=1,2,3..,n)

A obtenção dos componentes principais foi importante para garantir que os dados simulados pelo teorema de Box-Muller tivessem media zero e variância V.

Para simular outras populações como réplicas da primeira população simulada e simular populações com números de genótipos diferentes da original e manter as características genéticas de herdabilidade, coeficiente de variação e média foi utilizado a técnica multivariada da decomposição espectral.

Sendo X o conjunto representativo das informações de g (i=1,2...g) indivíduos em relação a b (b=1,2...b) informações com estrutura de covariância denotada por S, real e simétrica é possível, por meio da decomposição espectral encontrar F de modo que:

$$S^{-1} = FF'$$

atendendo a restrição:

$$F'SF = I$$

sendo I uma matriz identidade

É possível obter, pela decomposição espectral ou pela simulação, utilizando o teorema de Box-Muller, um novo conjunto de dados, de dimensões equivalentes (pelo menos em relação ao número de colunas, podendo ter o número de linhas

reduzido ou ampliado) a de X, denotado por W, que tenha matriz de covariâncias e variâncias igual a I (matriz identidade). Assim, a seguinte transformação pode ser estabelecida:

$$X^{*} = (F')^{-1}W$$

em que:

$$V(X^{*}) = (F')^{-1}V(W)(F)^{-1} = (FF')^{-1} = S$$

Assim, o conjunto de dados X^{*} representa uma réplica do conjunto original de observações com as mesmas propriedades de variância e covariância, porém com número diferentes de genótipos por bloco. As etapas do procedimento de simulação são mostradas na figura 6.

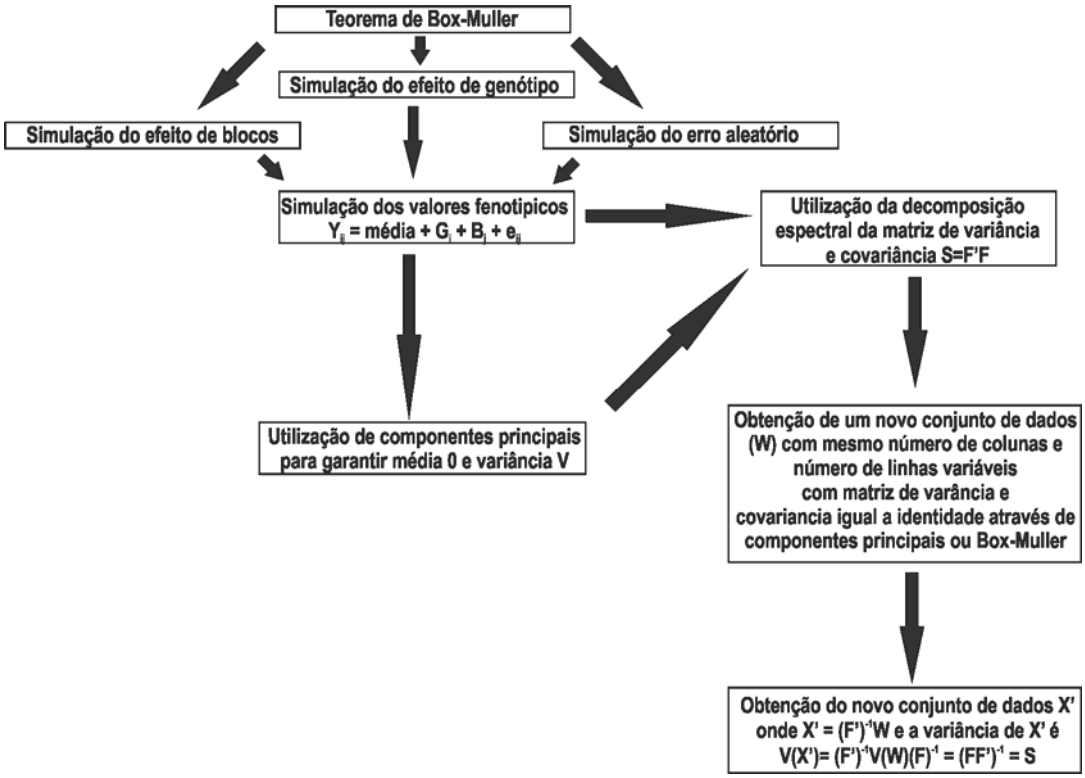

Figura 6. Esquema do procedimento das simulações.

3.3.7. Validação da simulação

Para validação da simulação de experimentos balanceados em blocos casualizados foram gerados dados simulados com as mesmas características dos dados simulados no primeiro processo de simulação de uma população, definidos pela media, herdabilidade, coeficiente de variação e matriz de variância e covariância.

Assim foram simulados experimentos com seis blocos e 5000 genótipos para verificar se a técnica de decomposição espectral conseguiria manter em uma

população aumentada a estrutura genética da população original. Dimensões maiores certamente trarão benefícios no processo de predição do valor genético correto. Assim, para cada variável, o número de entradas foi igual a 6 (seis blocos). Foi simulado uma população para cada cenário.

Foi utilizado um procedimento similar ao descrito acima, porém com número de genótipos mais próximo da situação real de teste. Assim, foram utilizadas informações de novos experimentos simulados, com a mesma caracterização da média, herdabilidade, coeficiente de variação experimental e matriz de variância e covariância, mas com apenas 150 e 200 genótipos. Foram simuladas 100 populações para cada cenário. Destaca-se o fato de que os ensaios para fins de validação, apesar de também ter origem da simulação, não são subamostras do conjunto de dados, mas novos ensaios com toda estrutura genética e ambiental preservada. Este fato é possível utilizando a técnica de decomposição espectral da matriz de covariância dos dados originais.

Visando verificar a homogeneidade da matriz de variância e covariância dos experimentos simulados foi realizado o teste de homogeneidade da matriz de covariância (Teste de Bartlett). Este pressupõe que a matriz de covariância dos n experimentos sejam iguais (hipótese H_0). A região critica obtida por meio do teste de razão de verossimilhança é dada por:

$$RC_1 = \left\{ Y \left| \lambda = n \ln\left|\widehat{\Sigma}\right| - \sum_{i=1}^{l} n_i \ln\left|\widehat{\sum}_i\right| > X^2_{\alpha,f} \right. \right\}$$

em que $\widehat{\sum}_i = \frac{n-1}{n} S_i$ e $\widehat{\Sigma} = \frac{\sum_{i=1}^{l}(n_i - 1)S_i}{n}$

onde $\widehat{\sum}_i$ é o estimador da matriz de covariância da i-ésima população e $\widehat{\Sigma}$ é o estimador da matriz de covariância comum entre as populações.

3.4. RESULTADO E DISCUSSÃO

A metodologia aplicada utilizando a técnica de decomposição espectral da matriz de covariâncias de uma população original hipotética mostrou-se eficiente para simular populações com as mesmas características das populações originais, tais como herdabilidade, coeficiente de variação, média e a matriz de variância e covariância. O estudo da estrutura populacional tem atraído pesquisadores de

várias áreas do conhecimento como populações biológicas, ecologia molecular e genética humana (Jombart et al., 2010).

Na simulação das características a partir de uma população com CV, h^2 e médias definidas, foi possível simular populações com estas características bem semelhantes (Tabela 2).

Tabela 2. Caracterização da população original (POP - tamanho da população, h^2 - herdabilidade, CV - coeficiente de variação esperado e X - média) e da população simulada para características (Ci) simuladas.

			C1			C2			C3			C4			C5		
Pop	h^2_{esp}	CV_{esp}	h^2_{sim}	CV	$\overline{X}$	h^2	CV	$\overline{X}$	h^2	CV	$\overline{X}$	h^2	CV	$\overline{X}$	h^2	CV	$\overline{X}$
150	10	5	11.02	5.01	20	10.78	4.99	40	8.50	5.03	60	9.55	5.00	80	8.95	5.04	100
150	10	10	9.21	10.01	20	11.84	9.96	40	8.90	9.95	60	8.74	9.99	80	10.08	9.98	100
150	20	5	21.67	4.97	20	21.13	4.93	40	18.24	5.03	60	19.7	5.02	80	18.06	5.02	100
150	20	10	20.28	10.01	20	19.29	9.99	40	18.41	10.1	60	20.06	9.98	80	20.50	10.01	100
150	30	5	30.82	5.03	20	29.19	5.01	40	29.13	5.02	60	30.16	4.94	80	31.10	5.03	100
150	30	10	30.18	10.00	20	31.42	9.89	40	28.71	10.03	60	29.57	10.06	80	29.72	9.98	100
150	40	5	39.05	5.01	20	38.18	5.04	40	40.69	4.99	60	41.59	5.02	80	41.56	5.00	100
150	40	10	40.86	9.93	20	38.03	10.05	40	40.94	9.90	60	40.76	9.94	80	38.25	9.98	100
200	10	5	8.65	5.04	20	10.77	4.99	40	11.03	5.00	60	10.51	5.02	80	9.97	4.98	100
200	10	10	11.19	10.01	20	11.84	9.96	40	11.06	9.95	60	11.36	9.97	80	9.14	9.95	100
200	20	5	21.12	4.99	20	18.98	5.01	40	18.60	5.02	60	18.26	5.03	80	21.02	5.01	100
200	20	10	19.30	10.02	20	18.51	10.00	40	21.00	9.95	60	21.22	9.98	80	21.93	9.96	100
200	30	5	31.04	5.00	20	30.84	4.95	40	28.21	5.00	60	31.57	4.99	80	28.33	4.97	100
200	30	10	28.73	10.01	20	30.38	10.00	40	31.63	9.85	60	31.39	10.00	80	29.92	9.82	100
200	40	5	40.00	4.98	20	40.10	4.97	40	40.94	4.96	60	38.50	5.04	80	39.06	5.01	100
200	40	10	39.38	9.91	20	39.95	9.98	40	41.20	9.90	60	41.30	10.00	80	38.26	10.04	100

(*) Médias esperadas para C1, C2, C3, C4 e C5 eram de 20, 40, 60, 80 e 100, respectivamente.

A média simulada foi idêntica a média esperada (média da população original) em todas as populações simuladas. Assim, o processo de simulação garante integralmente a manutenção da média esperada, que representa uma importante informação, em especial, em experimentos de natureza biológica. Deve estar atento que a média é uma propriedade inerente ao material genético avaliado e traduz a informação geral do potencial da cultura em relação às múltiplas características que são normalmente mensuradas num programa de melhoramento genético.

Na experimentação agrícola, é comum que pesquisadores procurem obter medidas que avaliem a precisão do experimento. A pesquisa é fundamental para a melhoria da produtividade, sendo que a avaliação da qualidade do experimento, usualmente, passa pela verificação do indicador da precisão, expressa pelos valores dos coeficientes de variação (CV) ou da diferença mínima significativa (dms) (SILVA et al., 2011). O coeficiente de variação simulado variou de 4,97 a 5,04 para C1, 4,93 a 5,04 para C2, 4,96 a 5,03 para C3, 4, 94 a 5,04 para C4 e 4, 97 a 5,04 para C5 para um coeficiente de variação esperado de 5. Para CV esperado de 10, o CV simulado variou de 9,91 a 10,02, 9,89 a 10,05, 9,85 a 10,03, 9,94 a 10,06 e 9,82 a 10,04 para C1, C2, C3, C4 e C5 respectivamente. Assim, a simulação gerou uma população onde o CV observado variou 1,2% e 1,8% para CV esperado de 5 e 10% respectivamente. Portanto a variação foi muito pequena. Dentre as medidas de dispersão, o CV é o único que permite a comparação da dispersão em diferentes trabalhos e diferentes variáveis, sendo considerada uma medida de dispersão relativa. Portanto, é indispensável manter constante o CV para manter a precisão experimental dos dados simulados, e assim garantir a acurácia da estimativa do valor genético pelas redes neurais.

As estimativas de parâmetros genéticos e fenotípicos são importantes em programas de melhoramento genético, pois possibilitam a escolha de métodos e caracteres utilizados nas etapas iniciais e avançadas dos programas de melhoramento, permitindo ainda, estudar mecanismos, valores genéticos e variabilidade para um caráter (Vasconcelos et al., 2012). Entre os parâmetros genéticos e fenotípicos que podem auxiliar o direcionamento da seleção de genótipos mais promissores, destacam-se as herdabilidades, as variâncias genéticas e fenotípicas e os progressos genéticos esperados (Ferrão et al., 2008). A estimativa de herdabilidade entre os diversos caracteres sob seleção é essencial

para permitir o estabelecimento de um conjunto de estratégias e métodos de melhoramento genéticos mais efetivos (Silva & Vieira, 2008).

A herdabilidade simulada variou de 8,5 a 11,84, 18,06 a 21,93, 28,21 a 31,63, 38,03 a 41,59 para herdabilidade esperada de 10, 20, 30 e 40 respectivamente. Assim, observamos uma variação no h^2 simulado de 18,4%, 9,7%, 6,9% e 4,93% para h^2 esperada de 10, 20, 30 e 40 respectivamente. A herdabilidade de uma característica combina tanto informações particulares dos genótipos avaliados quanto da natureza ambiental, incluindo componentes genéticos da variação e componentes aleatórios do erro.

A herdabilidade é um parâmetro essencial em programas de melhoramento genético, pois indica quanto das diferenças existentes no desempenho para uma característica são determinadas por fatores genéticos, tendo assim papel fundamental na predição dos valores genéticos (Custódio et al., 2012). Desta forma, pela sua importância, a herdabilidade deve ser conhecida para a condução de um programa de melhoramento, e muitas das decisões práticas são tomadas em função de sua magnitude (Ramalho et al., 2000). Portanto, torna-se extremamente necessário a manutenção constante da herdabilidade do experimento original nos experimentos simulados para treinamento e validação das redes neurais. Assim, a acurácia da predição do valor genético realizados pelas redes neurais tende a aumentar, pois a correlação do valor de rede com o valor genético aumenta.

A herdabilidade foi o parâmetro simulado em que foi observado maiores variações. Este fato ocorre devido a maior dificuldade em manter as informações sobre componentes quadráticos tais como a variação genética e ambiental. À medida que a h^2 esperada aumenta, a variação da h^2 simulada diminui. Portanto a metodologia apresentada, para fins de simulação de populações, tem mais eficiência para características qualitativas, ou seja, menos influenciada pelo ambiente. Assim é possível manter as variâncias esperadas e o valor da herdabilidade mais próximo do valor desejado. Com estudos de simulação Wang et al. (2007) mostrou que a utilização de componentes principais a partir das medidas de herdabilidade (PCH) tem um poder de ganho maior que a técnica de componentes principais utilizada sobre as médias das populações para estimar o valor genético dos indivíduos.

O tamanho das populações simuladas (150 ou 200 indivíduos) não influenciou no processo de simulação. A média não variou em nenhum dos tamanhos de população. O CV variou de 4,93 a 5,04% e 4,95 a 5,04% nas

populações com 150 e 200 genótipos respectivamente para o CV esperado de 5. Com o CV esperado de 10 a variação foi de 9,89 a 10,06%.

A partir das populações simuladas (Tabela 2) foram simuladas 100 réplicas mantendo a estrutura genética da população original (média, herdabilidade e coeficiente de variação). Através da decomposição espectral da matriz de covariâncias foi possível manter além desta estrutura, a matriz de variância e covariância próxima ao valor esperado. A eficiência do processo de simulação foi avaliada a partir da estrutura genética da população original (Tabelas 3, 4, 5, 6 e 7). Foi observado que as réplicas simuladas mantiveram constante a estrutura genética.

Para o treinamento das RNAs a manutenção da estrutura genética é importante, pois aumenta a acurácia da rede na validação e posteriormente no teste da RNAs. Isto ocorre porque a estrutura genética do arquivo de treinamento é igual aos arquivos de validação, facilitando o aprendizado da rede.

No entanto, deve ser destacado que o processo de simulação preservando apenas propriedades individuais, seja do conjunto de genótipos ou do ambiente da experimentação, não é suficiente para representar apropriadamente as informações geradas pela experimentação, em especial quando se tem por objetivo o uso destes dados para fins de treinamento de redes neurais. As redes tem seu aprendizado baseado no padrão de resposta manifestado em cada observação de forma a captar sinais, tendências e ruídos proporcionados pelo desempenho do material genético avaliado e o efeito indesejável do ambiente que prejudica a associação entre valores fenotípicos observáveis de valores genéticos verdadeiros.

Assim torna-se de suma importância, além da manutenção da média, herdabilidade e coeficiente de variação constantes, manter invariante a matriz de covariância, pois desta forma é possível manter constante a correlação entre o valor fenotípico dos blocos e o valor genotípico simulado. Assim a utilização da técnica de decomposição espectral sobre a matriz de covariância original é importante para simular populações para o treinamento e validação das redes neurais.

Uma vantagem da utilização da decomposição espectral é a versatilidade, pois esta técnica não dependem de um modelo genético e estão livres de pressuposições de equilíbrio de Hardy-Weinberg e desequilíbrio de ligação (Jombart et al., 2010). Portanto esta técnica podem ser utilizadas para qualquer tipo de população independente da ploidia e da recombinação gênica. Além disso os componentes principais podem ser aplicados a um conjunto de dados relativamente grande e obter resultados com um tempo computacional quase desprezível, quando

comparado aos métodos bayesianos do STRUCTURE (Falush et al., 2003) e BAPS (Tang et al., 2009), que são métodos que precisam de computadores altamente potentes para serem utilizados. Isto foi verificado neste trabalho, pois todas as populações simuladas não demoraram mais que 1 minuto para finalizar o processo de simulação, mostrando a eficiência do uso de técnicas multivariadas como decomposição espectral para simular experimentos.

Tabela 3. Caracterização da média das 100 populações simuladas a partir da primeira população simulada para a característica C1 (X = 20).

	Esperado			Observado						
Pop	**h^2_{esp}**	**CV_{esp}**	**$\bar{X}_{esp}$**	**h^2_{obs}**	**$\bar{X}_{obs}$**	**$h^2_{i\ obs}$**	**CV_{eobs}**	**CV_{gobs}**	**σ^2_{gobs}**	**σ^2_{obs}**
150	11.02	5.01	20	11.02	20	2.02	5.01	0.72	0.02	1.00
150	9.21	10.01	20	9.21	20	1.66	10.01	1.30	0.07	4.01
150	21.67	4.97	20	21.67	20	4.41	4.97	1.07	0.05	0.99
150	20.28	10.01	20	20.28	20	4.07	10.01	2.06	0.17	4.01
150	30.82	5.03	20	30.82	20	6.91	5.03	1.37	0.07	0.10
150	30.18	10.00	20	30.18	20	6.72	10.00	2.68	0.29	4.00
150	39.05	5.01	20	39.05	20	9.65	5.01	1.64	0.11	1.00
150	40.86	9.93	20	40.86	20	10.32	9.93	3.37	0.45	3.94
200	8.65	5.04	20	8.65	20	1.55	5.04	0.63	0.02	1.02
200	11.19	10.01	20	11.19	20	2.06	10.01	1.45	0.08	4.08
200	21.12	4.99	20	21.12	20	4.27	4.99	1.05	0.04	1.00
200	19.3	10.02	20	19.30	20	3.83	10.02	2.00	0.16	4.02
200	31.04	5.00	20	31.04	20	6.98	5.00	1.37	0.07	1.00
200	28.73	10.01	20	28.73	20	6.30	10.01	2.59	0.27	4.01
200	40.00	4.98	20	40.00	20	10.00	4.98	1.66	0.11	0.99
200	39.38	9.91	20	39.38	20	9.77	9.91	3.26	0.42	3.93

Pop - tamanho da população; h^2_{esp} - herdabilidade esperada; CV_{esp} - coeficiente de variação esperado; X_{esp} - média esperada; h^2_{obs} - herdabilidade média observada; X_{obs} - média observada; $h^2_{i\ obs}$ - herdabilidade individual média observada; $CV_{e\ obs}$ - coeficiente de variação experimental médio observado; $CV_{g\ obs}$ - coeficiente de variação genético médio observado; $\sigma^2_{g\ obs}$ - variância genética observada; $\sigma^2_{\ obs}$ - variância ambiental observada.

A média manteve-se constante nas 100 réplicas avaliadas, resultado este que já tinha sido obtido na primeira população simulada (Tabelas 3, 4, 5, 6 e 7). Assim, a média foi o único parâmetro avaliado que se manteve constante nos dois processos de simulação.

Jombart et al. (2010) simularam vários conjuntos de populações para diferentes ambientes e utilizaram componentes principais em conjunto com análise discriminante para separar as populações que pertenciam ao mesmo grupo e verificar a diversidade entre as populações simuladas. Eles concluíram que a

utilização da técnica de componentes principais foi tão acurada quanto aos resultados obtidos pelo software STRUCTURE na detecção da estrutura de aglomerados populacionais ocultos dentro de modelos populacionais simples. E ainda, relataram que os componentes principais mostraram-se mais acurado para detectar a estrutura populacional em modelos mais complexos. Assim, é possível verificar que a técnica de componentes principais é uma boa alternativa para detectar e manter a estrutura populacional, pois é possível manter constante a matriz de variância e covariância.

Tabela 4. Caracterização da média das 100 populações simuladas a partir da primeira população simulada para a característica C2 (X = 40).

	Esperado			Observado						
Pop	**h^2_{esp}**	**CV_{esp}**	**$\bar{X}_{esp}$**	**h^2_{obs}**	**$\bar{X}_{obs}$**	**$h^2_{i\ obs}$**	**CV_{eobs}**	**CV_{gobs}**	**σ^2_{gobs}**	**σ^2_{obs}**
150	10.78	4.99	40	10.78	40	1.97	4.99	0.71	0.08	3.99
150	11.84	9.96	40	11.84	40	2.19	9.96	1.49	0.36	15.88
150	21.13	4.93	40	21.13	40	4.27	4.93	1.04	0.17	3.89
150	19.29	9.99	40	19.29	40	3.83	9.99	1.49	0.64	15.98
150	29.19	5.01	40	29.19	40	6.14	5.01	1.28	0.26	4.01
150	31.42	9.89	40	31.42	40	7.09	9.89	2.73	1.19	15.65
150	38.18	5.04	40	38.18	40	9.33	5.04	1.62	0.42	4.06
150	38.03	10.05	40	38.03	40	9.28	10.05	3.21	1.65	16.17
200	10.77	4.99	40	10.77	40	1.97	4.99	0.71	0.08	3.99
200	11.84	9.96	40	11.84	40	2.19	9.96	1.49	0.36	15.88
200	18.98	5.01	40	18.98	40	3.76	5.01	1.00	0.16	4.02
200	18.51	10.00	40	18.51	40	3.65	10.00	1.95	0.61	16.00
200	30.84	4.95	40	30.84	40	6.92	4.95	1.35	0.29	3.92
200	30.38	10.00	40	30.38	40	6.78	10.00	2.70	1.16	15.99
200	40.10	4.97	40	40.10	40	10.04	4.97	1.66	0.44	3.96
200	39.95	9.98	40	39.95	40	9.98	9.98	3.32	1.77	15.39

Pop - tamanho da população; h^2_{esp} - herdabilidade esperada; CV_{esp} - coeficiente de variação esperado; X_{esp} - média esperada; h^2_{obs} - herdabilidade média observada; X_{obs} - média observada; $h^2_{i\ obs}$ - herdabilidade individual média observada; $CV_{e\ obs}$ - coeficiente de variação experimental médio observado; $CV_{g\ obs}$ - coeficiente de variação genético médio observado; $\sigma^2_{g\ obs}$ - variância genética observada; $\sigma^2_{\ obs}$ - variância ambiental observada.

O coeficiente de variação experimental manteve-se constante na simulação das 100 réplicas (Tabelas 3, 4, 5, 6 e 7). Portanto o efeito do ambiente sobre a expressão das características aqui simuladas foi mantido igual ao que foi estabelecido pelo pesquisador. Portanto foi verificado que as populações simuladas com a utilização da técnica da decomposição espectral mantém a estrutura da

população original. As técnicas multivariadas vêm sendo utilizadas por décadas para extrair vários tipos de informação de dados genéticos com grandes interesses para a produção (Price et al., 2006; Jombart, 2008). Em particular, a análise de componentes principais tem sido recentemente sugerida como alternativa a algumas análises bayesianas (Liu & Zhao, 2006; Patterson et al., 2006; Lee et al., 2009). A principal característica dos componentes principais é a habilidade para identificar estruturas genéticas em uma grande quantidade de dados dentro de um tempo computacional muito pequeno, e na ausência de qualquer hipótese sobre o modelo populacional (Jombart et al., 2010). Assim os componentes principais torna-se uma técnica atraente para serem utilizadas em programas de simulação para gerar réplicas de população de campo para utilização em processos de avaliação de características genéticas da população através de análises bayesianas ou redes neurais.

Wang et al. (2007) estimaram a matriz de variância e covariância para calcular os componentes principais. Eles observaram que em famílias de irmãos completos ou endogâmicas, as estimativas dos componentes de variância foram eficientes. No entanto quando a estrutura de famílias é mais complexa (polinização ao acaso) a estimativa dos componentes de variância não foi eficiente. Para melhorar esta eficiência em famílias com estruturas complexas foi estimado que a correlação entre indivíduos de duas famílias deveriam ser iguais. Assim, é necessário que o estudo de todos os parâmetros genéticos seja feito para garantir que o processo de simulação seja eficiente e que tenha confiabilidade para serem utilizados em estudos de estrutura genética.

A manutenção do coeficiente de variação garantiu que o efeito ambiental atuante sobre a característica manteve-se constante nas populações de treinamento e validação. Assim a eficiência das RNAs é garantida, pois o ruído causado pelo efeito ambiental na saída da rede é mantido no treinamento e na validação, tornando mais fácil o aprendizado da rede, e garantindo melhores resultados quando a rede for testada através de uma população de teste, onde o valor genotípico não é conhecido.

Tabela 5. Caracterização da média das 100 populações simuladas a partir da primeira população simulada para a característica C3 (X = 60).

	Esperado			Observado						
Pop	h^2_{esp}	CV_{esp}	$\bar{X}_{esp}$	h^2_{obs}	$\bar{X}_{obs}$	$h^2_{i\ obs}$	CV_{eobs}	CV_{gobs}	σ^2_{gobs}	σ^2_{obs}
150	8.50	5.03	60	8.50	60	1.53	5.03	0.63	0.14	9.10

150	8.90	9.95	60	8.90	60	1.60	9.95	1.27	0.58	35.61
150	18.24	5.03	60	18.24	60	3.58	5.03	0.97	0.34	9.11
150	18.41	10.10	60	18.41	60	3.62	10.10	1.96	1.38	36.72
150	29.13	5.02	60	19.13	60	3.79	5.02	1.00	0.36	9.08
150	28.71	10.03	60	28.71	60	6.29	10.03	2.60	2.43	36.24
150	40.69	4.99	60	40.69	60	10.26	4.99	1.69	1.02	8.95
150	40.94	9.90	60	40.94	60	10.36	9.90	3.37	4.08	35.29
200	11.03	5.00	60	11.03	60	2.02	5.00	0.72	0.19	9.01
200	11.06	9.95	60	11.06	60	2.03	9.95	1.43	0.74	35.63
200	18.60	5.02	60	18.60	60	3.67	5.02	0.98	0.34	9.07
200	21.00	9.95	60	21.00	60	4.24	9.95	2.09	1.58	35.63
200	28.21	5.00	60	28.21	60	6.15	5.00	1.28	0.59	9.00
200	31.63	9.85	60	31.63	60	6.62	9.85	3.40	4.15	34.94
200	40.94	4.96	60	40.94	60	10.36	4.96	1.68	1.02	8.84
200	41.20	9.90	60	41.20	60	10.46	9.90	3.38	4.12	35.32

Pop - tamanho da população; h^2_{esp} - herdabilidade esperada; CV_{esp} - coeficiente de variação esperado; X_{esp} - média esperada; h^2_{obs} - herdabilidade média observada; X_{obs} - média observada; $h^2_{i\ obs}$ - herdabilidade individual média observada; $CV_{e\ obs}$ - coeficiente de variação experimental médio observado; $CV_{g\ obs}$ - coeficiente de variação genético médio observado; $\sigma^2_{g\ obs}$ - variância genética observada; $\sigma^2_{\ obs}$ - variância ambiental observada.

A herdabilidade manteve-se constante em todas as réplicas e com valores iguais as populações originais (Tabelas 3, 4, 5, 6 e 7). Assim observou-se que a variância genética e a variância ambiental não se alteraram durante o processo de simulação. Portanto é possível através da utilização de redes neurais artificiais fazer predição do valor genético e calcular os componentes de variância da população original com eficiência.

Segundo Gianola et al. (2006) os processos de simulação podem ajudar na comparação e avaliação de métodos estatísticos paramétricos e não paramétricos, pois um número enorme de cenários simulados podem ser obtidos, abrangendo mais a área de estudo. Isto é possível através das combinações ilimitadas dos parâmetros utilizados no processo de simulação. Já os dados reais são normalmente de natureza local, ou seja, as conclusões são apenas dentro daquele local, daquele experimento, não podendo ser extrapolada.

A eficiência das RNAs esta correlacionada com o número de dados do arquivo de treinamento e do número de arquivos de validação. Assim, a simulação de experimentos com grande número de genótipos para treinamento das RNAs e a simulação de várias populações com o mesmo número de genótipos da população original para validação é importante para aumentar a eficiência das RNAs.

Em populações com a mesma estrutura genética da população original, a eficiência da predição do valor genético pelas RNAs será maior quanto maior for o número de genótipos no treinamento e o número de experimentos na validação. Portanto a técnica de simulação de dados utilizando decomposição espectral é uma alternativa promissora visto que a obtenção de dados históricos nem sempre é possível.

Segundo Novembre et al. (2008) a utilização dos componentes principais para estudos da estrutura genética de populações é eficiente e pode ajudar a verificar índices de migração em populações de humanos. Cavalli-Sforza e Minch (1997) utilizaram a técnica de componentes principais para separar os alelos de alta frequência em populações humanas dos alelos de baixa frequência, e conseguiram fazer um mapa de filogeografia com estes alelos utilizando a técnica de componentes principais. A técnica de componentes principais vem sendo utilizada largamente para estudos de estrutura genética principalmente em humanos. Esta técnica é interessante por refinar os dados obtidos pelo teorema de Box-Muller, e desta forma manter toda a estrutura genética da população.

Tabela 6. Caracterização da média das 100 populações simuladas a partir da primeira população simulada para a característica C4 (X = 80).

	Esperado			Observado						
Pop	**h^2_{esp}**	**CV_{esp}**	**$\bar{X}_{esp}$**	**h^2_{obs}**	**$\bar{X}_{obs}$**	**$h^2_{i\ obs}$**	**CV_{eobs}**	**CV_{gobs}**	**σ^2_{gobs}**	**σ^2_{obs}**
150	9.55	5.00	80	9.55	80	1.73	5.00	0.66	0.28	16.02
150	8.74	9.99	80	8.74	80	1.57	9.99	1.26	1.02	63.93
150	19.70	5.02	80	19.70	80	3.93	5.02	1.01	0.66	16.15
150	20.06	9.98	80	20.06	80	4.01	9.98	2.04	2.66	63.71
150	30.16	4.94	80	30.16	80	6.71	4.94	1.32	1.12	15.61
150	29.57	10.06	80	29.57	80	6.54	10.06	2.66	4.53	64.78
150	41.59	5.02	80	41.59	80	10.61	5.02	1.73	1.91	16.13
150	40.76	9.94	80	40.76	80	10.29	9.94	3.37	7.26	63.30
200	10.51	5.02	80	10.51	80	1.92	5.02	0.70	0.32	16.15
200	11.36	9.97	80	11.36	80	2.09	9.97	1.46	1.36	63.56
200	18.26	5.03	80	18.26	80	3.59	5.03	0.97	0.60	16.17
200	21.22	9.98	80	21.22	80	4.30	9.98	2.11	2.86	63.77
200	31.57	4.99	80	31.57	80	7.14	4.99	1.38	1.23	15.95
200	31.39	10.00	80	31.39	80	7.09	10.00	2.76	4.88	63.98
200	38.50	5.04	80	38.50	80	9.45	5.04	1.63	1.70	16.27
200	41.30	10.00	80	41.30	80	10.50	10.00	3.42	7.50	63.94

Pop - tamanho da população; h^2_{esp} - herdabilidade esperada; CV_{esp} - coeficiente de variação esperado; X_{esp} - média esperada; h^2_{obs} - herdabilidade média observada; X_{obs} - média observada; $h^2_{i\ obs}$ - herdabilidade individual média observada; $CV_{e\ obs}$ - coeficiente de variação experimental médio observado; $CV_{g\ obs}$ - coeficiente de

variação genético médio observado; $\sigma^2_{g\ obs}$ - variância genética observada; $\sigma^2_{\ obs}$ - variância ambiental observada.

A simulação de populações para estudos da estrutura genética é importante e altamente utilizada atualmente em vários ramos da ciência como em estudos da filogeografia humana (Novembre et al., 2008), análise discriminante em estruturas de populações (Jombart et al., 2010), estudos de genoma (Price et al., 2006), detecção das interações gene-a-gene e variância genética (Bhattacharya et al., 2010). Em todos estes estudos a técnica de componentes principais mostrou-se como uma metodologia promissora para estudar a estrutura genética das populações simuladas, sendo possível utiliza-la para simular populações com o objetivo de utilizadas para treinamento e validação das RNAs.

Tabela 7. Caracterização da média das 100 populações simuladas a partir da primeira população simulada para a característica C5 (X = 100).

	Esperado			Observado						
Pop	**h^2_{esp}**	**CV_{esp}**	**$\bar{X}_{esp}$**	**h^2_{obs}**	**$\bar{X}_{obs}$**	**$h^2_{i\ obs}$**	**CV_{eobs}**	**CV_{gobs}**	**σ^2_{gobs}**	**σ^2_{obs}**
150	8.95	5.04	100	8.95	100	1.61	5.04	0.65	0.42	25.39
150	10.08	9.98	100	10.08	100	1.83	9.98	1.36	1.86	99.62
150	18.06	5.02	100	18.06	100	3.54	5.02	0.96	0.93	25.26
150	20.50	10.01	100	20.50	100	4.12	10.01	2.07	4.30	100.17
150	31.10	5.03	100	21.10	100	4.27	5.03	1.06	1.13	25.30
150	29.72	9.98	100	29.72	100	6.58	9.98	2.65	7.02	99.61
150	41.56	5.00	100	41.56	100	10.60	5.00	1.72	2.96	24.99
150	38.25	9.98	100	38.25	100	9.36	9.98	3.21	10.30	99.71
200	9.97	4.98	100	9.97	100	1.81	4.98	0.68	0.46	24.77
200	9.14	9.95	100	9.14	100	1.65	9.95	1.29	1.66	98.93
200	21.02	5.01	100	21.02	100	4.25	5.01	1.05	1.11	25.11
200	21.93	9.96	100	21.93	100	4.47	9.96	2.16	4.65	99.31
200	28.33	4.97	100	28.33	100	6.18	4.97	1.28	1.63	24.69
200	29.92	9.82	100	29.92	100	6.64	9.82	2.62	6.87	96.52
200	39.06	5.01	100	39.06	100	9.65	5.01	1.64	2.69	25.14
200	38.26	10.04	100	38.26	100	9.36	10.04	3.23	10.41	100.77

Pop - tamanho da população; h^2_{esp} - herdabilidade esperada; CV_{esp} - coeficiente de variação esperado; X_{esp} - média esperada; h^2_{obs} - herdabilidade média observada; X_{obs} - média observada; $h^2_{i\ obs}$ - herdabilidade individual média observada; $CV_{e\ obs}$ - coeficiente de variação experimental médio observado; $CV_{g\ obs}$ - coeficiente de variação genético médio observado; $\sigma^2_{g\ obs}$ - variância genética observada; $\sigma^2_{\ obs}$ - variância ambiental observada.

A técnica de decomposição espectral também foi eficiente para manter as características das populações aumentadas (experimentos com 6 blocos e 5000 tratamentos por bloco). A herdabilidade, o coeficiente de variação e a média

mantiveram-se constante na simulação destas populações. Assim, é possível através de um experimento de campo gerar novas populações com valor fenotípico individual diferente, porém mantendo a estrutura genética da população. A manutenção da matriz de variância e covariância permite que seja possível manter a estrutura genética do experimento durante o processo de simulação.

Segundo Timm (2002) quando maximizamos a forma quadrática que representa a variância do componente principal, também maximizamos a média geométrica das variâncias dos componentes principais que equivale a maximizar a média geométrica dos autovalores da matriz de variância e covariância. Assim os componentes principais tem a capacidade de preservar as distâncias dos pontos coordenados originais para a origem invariante e maximizar as distâncias entre eles no novo sistema de eixos coordenados o que maximiza suas variâncias.

O processo de simulação foi eficiente para simular experimentos aumentados (maior número de genótipos), e através da técnica da decomposição espectral, foi possível manter a matriz de variância e covariância da população original e assim manter a estrutura genética (herdabilidade, coeficiente de variação e média) inalterada.

É possível através de um experimento de campo envolvendo seleção de genitores, onde a herdabilidade, o coeficiente de variação e a média são conhecidos, gerar populações com as mesmas características da população original para serem utilizadas em análises computacionais que requerem repetições como as redes neurais artificiais (RNAs). Nas RNAs é necessária uma população de treinamento com grande número de indivíduos e várias populações de validação do tamanho da população original, da forma que foi gerada no presente estudo, uma população com 5000 indivíduos para ser utilizadas para treinamento da rede e 100 réplicas com 150 e 200 indivíduos por bloco para serem utilizadas para validação da rede.

Através do teste de homogeneidade de variância verificou-se que em todas as populações simuladas, as matrizes de variâncias e covariâncias foram homogêneas. Esta homogeneidade foi perfeita, pois as populações simuladas possuem mesma matriz de variância e covariância das populações originais (Tabela 8).

Tabela 8. Covariância (COV) e correlação (COR) esperado e simulado para as populações de validação e treinamento das redes neurais artificiais utilizando decomposição espectral da matriz de covariância original.

Blocos			Esperado		Validação		Treinamento	
			COV	COR	COV	COR	COV	COR
1	X	2	0,160	0,150	0,160	0,150	0,160	0,150
1	X	3	0,130	0,120	0,130	0,120	0,130	0,120
1	X	4	-0,120	-0,120	-0,120	-0,120	-0,120	-0,120
1	X	5	0,080	0,080	0,080	0,080	0,080	0,080
1	X	6	-0,090	-0,080	-0,090	-0,080	-0,090	-0,080
2	X	3	-0,005	-0,005	-0,005	-0,005	-0,005	-0,005
2	X	4	-0,080	-0,090	-0,080	-0,090	-0,080	-0,090
2	X	5	0,050	0,050	0,050	0,050	0,050	0,050
2	X	6	0,090	0,080	0,090	0,080	0,090	0,080
3	X	4	0,090	0,080	0,090	0,080	0,090	0,080
3	X	5	-0,020	-0,030	-0,020	-0,030	-0,020	-0,030
3	X	6	0,070	0,060	0,070	0,060	0,070	0,060
4	X	5	-0,090	-0,100	-0,090	-0,100	-0,090	-0,100
4	X	6	0,090	0,090	0,090	0,090	0,090	0,090
6	X	6	-0,130	-0,130	-0,130	-0,130	-0,130	-0,130

3.5. CONCLUSÃO

A obtenção de dados experimentais preservando propriedades pontuais, tais como média, herdabilidade e coeficiente de variação, é eficiente e pode ser realizada usando princípios estocásticos de distribuição tais como enunciado no teorema de Box-Muller.

A decomposição espectral é uma técnica promissora para geração de experimentos simulados a partir de um experimento original, mantendo sua estrutura genética (média, herdabilidade, coeficiente de variação e matriz de variância e covariância)

Conjuntos de dados preservados ou ampliados podem ser apropriadamente gerados com potencial uso, em especial em redes neurais que demanda grande quantidade de informações para fins de treinamento e aprendizagem. Os dados apropriadamente simulados, por preservarem informações essenciais, podem agregar ou substituir dados históricos algumas vezes não tão facilmente disponíveis.

3.6. REFERÊNCIAS BIBLIOGRÁFICAS

BAKER, J. F.; STEWART, T. S.; LONG, C. R.; CARTWRIGHT, T. C. Multiple regression and principal components analysis of puberty and growth in cattle. **J Anim Sci**, v. 66, n. 9, p. 2147-2158, 1988.

BHATTACHARYA, S.; HEITMANN, K.; WHITE, M.; LUKIĆ, Z.; WAGNER, C.; HABIB, S. Mass Function Predictions Beyond LCDM. **arXiv preprint arXiv:1005.2239**, 2010.

CAVALLI-SFORZA, L.; MINCH, E. Paleolithic and Neolithic lineages in the European mitochondrial gene pool. **Am J Hum Genet**, v. 61, n. 1, p. 247, 1997.

CHAO-LUNG, Y.; YUEHWEM, Y.; YAN-FU, K.; GEORGE, C.; JAN, A. Improving Tone Prediction in Calibration of Electrophotographic Printers by Linear Regression: Using Principal Components to Account for Co-Linearity of Sensor Measurements. **Journal of Imaging Science and Technology**, v. 54, n. 5, p. 50302-50309, 2010.

CHASE, K.; CARRIER, D. R.; ADLER, F. R.; JARVIK, T.; OSTRANDER, E. A.; LORENTZEN, T. D.; LARK, K. G. Genetic basis for systems of skeletal quantitative traits: principal component analysis of the canid skeleton. **Proceedings of the National Academy of Sciences**, v. 99, p. 9930-9935, 2002.

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

CRUZ, C. D. **Programa Genes: Estatística Experimental e matrizes**. Viçosa: UFV, 2006.

CUSTÓDIO, T. N.; BALIZA, D. P.; CARVALHO, S. P.; REZENDE, T. T. Meta-análise para estimativas de herdabilidade de características do desenvolvimento e produção do *Coffea canephora* Pierre. **Semina: Ciências Agrárias**, v. 33, p. 2501-2510, 2012.

EEUWIJK, F. A. V.; BOER, M.; TOTIR, L. R.; BINK, M.; WRIGHT, D.; WINKLER, C. R.; PODLICH, D.; BOLDMAN, K.; BAUMGARTEN, A.; SMALLEY, M.; ARBELBIDE, M.; BRAAK, C. J. F. T.; COOPER, M. Mixed model approaches for the identification of QTLs within a maize hybrid breeding program. **Theor. Appl. Genet.**, v. 120, p.429–440, 2010.

FALUSH, D.; STEPHENS, M.; PRITCHARD, J. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. **Genetics**, v.164, p.1567-1587, 2003.

FERRÃO, R. G.; CRUZ, C. D.; FERREIRA, A.; CECON, P. R.; FERRÃO, M. A. G.; FONSECA, A. F. A. D.; CARNEIRO, P. C. D. S.; SILVA, M. F. D. Parâmetros genéticos em café Conilon. **Pesquisa Agropecuária Brasileira**, v. 43, n. 1, p. 61-69, 2008.

FLINT-GARCIA, S. A.; THUILLET, A. C.; YU, J.; PRESSOIR, G.; ROMERO, S. M.; MITCHELL, S. E.; DOEBLEY, J.; KRESOVICH, S.; GOODMAN, M. M.; BUCKLER, E. S. Maize association population: a high-resolution platform for quantitative trait locus dissection. **Plant J**, v. 44, n. 6, p. 1054-1064, 2005.

GIANOLA, D.; DE LOS CAMPOS, G. Inferring genetic values for quantitative traits non-parametrically. **Genet Res (Camb)**, v. 90, n. 6, p. 525-540, 2008.

GIANOLA, D.; FERNANDO, R. L.; STELLA, A. Genomic-assisted prediction of genetic value with semiparametric procedures. **Genetics**, v. 173, n. 3, p. 1761-1776, 2006.

GIANOLA, D.; WU, X. L.; MANFREDI, E.; SIMIANER, H. A non-parametric mixture model for genome-enabled prediction of genetic value for a quantitative trait. **Genetica**, v. 138, n. 9-10, p. 959-977, 2010.

JOMBART, T. adegenet: a R package for the multivariate analysis of genetic markers. **Bioinformatics**, v. 24, n. 11, p. 1403-1405, 2008.

JOMBART, T.; DEVILLARD, S.; BALLOUX, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. **BMC Genet**, v. 11, p. 94, 2010.

KAVZOGLU, T. **An investigation of the design and use of feed forward artificial neural networks in the classification of remotely sensed images**. University of Nottingham, 2001.

KAVZOGLU, T. Increasing the accuracy of neural network classification using refined training data. **Environmental Modelling & Software**, v. 24, n. 7, p. 850-858, 2009.

KIRKPATRICK, M.; MEYER, K. Direct Estimation of Genetic Principal Components Simplified Analysis of Complex Phenotypes. **Genetics**, v. 168, n. 4, p. 2295-2306, 2004.

KHODADADI, M.; FOTOKIAN, M. H.; MIRANSARI, M. Genetic diversity of wheat (*Triticum aestivum* L.) genotypes based on cluster and principal component analyses for breeding strategies. **Australian Journal of Crop Science**, v.5, n.1, p.17-24, 2011.

KUMAR, A.; VERULKAR, S. B.; MANDAL, N. P.; VARIAR, M.; SHUKLA, V. D.; DWIVEDI, J. L.; SINGH, B. N.; SINGH, O. N.; SWAIN, P.; MALL, A. K.; ROBIN, S.; CHANDRABABU, R.; JAIN, A.; HAEFELE, S. M.; PIEPHO, H. P.; RAMAN, A. High-yielding, drought-tolerant, stable rice genotypes for the shallow rainfed lowland drought-prone ecosystem. **Field Crops Research**, v. 133, p. 37-47, 2012.

LEE, C.; ABDOOL, A.; HUANG, C. H. PCA-based population structure inference with generic clustering algorithms. **BMC Bioinformatics**, v. 10, p. S1-S73, 2009.

LIU, N.; ZHAO, H. non-parametric approach to population structure inference using multilocus genotypes. **Hum Genomics**, v. 2, n. 6, p. 353-364, 2006.

MOTSINGER-REIF, A. A.; DUDEK, S. M.; HAHN, L. W.; RITCHIE, M. D. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. **Genet Epidemiol**, v. 32, n. 4, p. 325-340, 2008.

NOVEMBRE, J.; JOHNSON, T.; BRYC, K.; KUTALIK, Z.; BOYKO, A. R.; AUTON, A.; INDAP, A.; KING, K. S.; BERGMANN, S.; NELSON, M. R. Genes mirror geography within Europe. **Nature**, v. 456, n. 7218, p. 98-101, 2008.

PATTERSON, N.; PRICE, A. L.; REICH, D. Population structure and eigenanalysis. **PLoS Genet**, v. 2, n. 12, p. e190, 2006.

PEARSON, K. LIII. On lines and planes of closest fit to systems of points in space. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, v. 2, n. 11, p. 559-572, 1901.

PRICE, A. L.; PATTERSON, N. J.; PLENGE, R. M.; WEINBLATT, M. E.; SHADICK, N. A.; REICH, D. Principal components analysis corrects for stratification in genome-wide association studies. **Nat Genet**, v. 38, n. 8, p. 904-909, 2006.

RAMALHO, M. A. P.; FERREIRA, D. P.; OLIVEIRA, A. C. **Experimentação em genética e melhoramento de plantas**. Lavras: UFLA, 2000. 303 p.

REICH, D.; PRICE, A. L.; PATTERSON, N. Principal component analysis of genetic data. **Nat Genet**, v. 40, n. 5, p. 491-492, 2008.

SILVA, A. R.; CECON, P. R.; RÊGO, E. R.; NASCIMENTO, M. Avaliação do coeficiente de variação experimental para caracteres de frutos de pimenteiras. **Rev. Ceres**, Viçosa, v. 58, n.2, p. 168-171, 2011.

SILVA, G. O.; VIEIRA, J. V. Componentes genéticos e fenotípicos para caracteres de importância agronômica em população de cenoura sob seleção recorrente. **Hortic. bras**, v. 26, n. 4, 2008.

SOARES, T. M.; MENDONÇA, M. C. M. Construção de um modelo de regressão hierárquico para os dados do SIMAVE-2000. **Pesquisa Operacional**, v. 23, n. 3, p. 421-441, 2003.

SOUZA, J. C.; PEROTTO, D.; ABRAHÃO, J. J.; FREITAS, J. A.; FERRAZ FILHO, P. B.; WEABER, R. L.; LAMBERSON, W. R. Estimativa das distâncias genéticas e componentes principais em bovinos de corte no Brasil. **Archivos de Zootecnia**, v. 59, n. 228, p. 479-485, 2010.

TIMM, N. H. **Applied multivariate analysis**. Springer, 2002.

VASCONCELOS, E. S.; REIS, M. S.; SEDIYAMA, T.; CRUZ, C. D. Estimativas de parâmetros genéticos da qualidade fisiológica de sementes de genótipos de soja produzidas em diferentes regiões de Minas Gerais. **Semina: Ciências Agrárias**, v. 33, n. 1, p. 65-76, 2012.

VENTURA, R.; SILVA, M.; MEDEIROS, T.; DIONELLO, N.; MADALENA, F.; FRIDRICH, A.; VALENTE, B.; SANTOS, G.; FREITAS, L.; WENCESLAU, R. Use of artificial neural networks in breeding values prediction for weight at 205 days in Tabapuã beef cattle. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 64, n. 2, p. 411-418, 2012.

TANG, J.; HANAGE, W.P.; FRASER, C.; CORANDER, J. Identifying Currents in the Gene Pool for Bacterial Populations Using an Integrative Approach. **PLoS. Comput. Biol.,** v.5, n.8, p.e1000455, 2009.

WANG, Z.; CONG, P.; ZHOU, J.; ZHU, Z. Method for identification of external quality of wheat grain based on image processing and artificial neural network [J]. **Transactions of the Chinese Society of Agricultural Engineering**, v. 1, p. 029, 2007.

YU, J.; HOLLAND, J. B.; MCMULLEN, M. D.; BUCKLER, E. S. Genetic Design and Statistical Power of Nested Association Mapping in Maize. **Genetics**, v. 178, p. 539–551, 2008.

4. CAPITULO 2

EFICIÊNCIA DAS REDES NEURAIS ARTIFICIAIS NA PREDIÇÃO DO VALOR GENÉTICO EM EXPERIMENTOS BALANCEADOS

EFICIÊNCIA DAS REDES NEURAIS ARTIFICIAIS NA PREDIÇÃO DO VALOR GENÉTICO EM EXPERIMENTOS BALANCEADOS

4.1. RESUMO

O objetivo deste trabalho foi avaliar a eficiência da utilização das redes neurais artificiais na predição do valor genético em experimentos em blocos casualizados, comparando a eficiência do valor de rede com a eficiência do valor fenotípico, e ainda definir uma estrutura de rede ideal para a predição do valor genético. Foram simulados 80 cenários, onde variou a herdabilidade (10, 20, 30 e 40%), a média (20, 40, 60, 80 e 100), o coeficiente de variação (5 e 10%) e o número de genótipos por bloco (150 e 200 para validação e 5000 para treinamento da rede neural). Utilizou-se 100 populações de validação em cada cenário. A acurácia das RNAs foi avaliada comparando a correlação do valor de rede com o valor genético e o valor fenotípico com o valor genético. As redes neurais foram eficientes para predição do valor genético com ganho de 0,64 a 10,3% em relação ao valor fenotípico independente do tamanho de população utilizada, da herdabilidade ou do coeficiente de variação simulado. A estrutura de rede composta por três camadas intermediárias com 10, 20 e 8 neurônios mostrou-se eficiente para a predição do valor genético. Portanto a rede neural artificial é uma técnica promissora para a predição do valor genético em experimentos balanceados.

Palavras-chave: Backpropagation, herdabilidade, experimentação, programação, melhoramento

4.2. INTRODUÇÃO

A identificação de genótipos superiores requer métodos de seleção capazes de explorar eficientemente o material genético disponível, maximizando o ganho genético em relação às características de interesse (Oda et al., 2007). Diversos métodos de seleção têm sido empregados nos programas de melhoramento, com destaque para a seleção entre e dentro de famílias (Paula et al., 2002), a seleção combinada (Martins et al., 2005) e a seleção por modelos mistos pelo método best linear unbiased prediction (BLUP) (Garcia & Nogueira, 2005).

Na seleção massal, também denominada seleção individual, as plantas são selecionadas com base em seus valores fenotípicos. Assim, a avaliação é visual e com base em características indiretas de produção. Com isso, a eficiência seletiva depende da quantidade de variabilidade existente na população-base a ser explorada, da herdabilidade do caráter a ser melhorado e da extensão do ganho genético deste caráter selecionado (Bastos et al., 2003). Este método tem sido recomendado em situações em que há ampla variabilidade genética disponível e maiores facilidades de obtenção de ganhos.

Outra opção para obtenção de ganhos é a prática de seleção de famílias, ao invés de indivíduos, que consiste em selecionar aquelas com elevados valores genotípicos. Estudos mostrando o potencial de famílias com valores genotípicos superiores, quando comparados com famílias de valores inferiores, evidenciam que a seleção com base nas melhores famílias é efetiva para identificar quais possuem maior proporção de clones-elite. Neste caso, a seleção de famílias com base em caracteres quantitativos de produção, poderá possibilitar a identificação de clones promissores com maior probabilidade de serem mais produtivos (Kimbeng & Cox, 2003).

Com objetivo de aumentar o ganho populacional, além das opções de selecionar indivíduo, família ou progênie, adota-se estratégias de seleção que permite acumular, ciclo a ciclo, os ganhos obtidos nas gerações anteriores. Neste contexto, foi preconizada a seleção recorrente que é um método de melhoramento utilizado para aumentar gradativamente o desempenho da população por meio da obtenção de progênies, sua avaliação e posterior recombinação (Doná et al., 2012). Em cada ciclo da seleção recorrente a frequência dos alelos favoráveis na população aumenta. Este método é muito utilizado em plantas alógamas principalmente milho (Souza et al., 2010; Rovaris et al., 2011).

Ganhos genéticos adicionais que possibilitam o aperfeiçoamento de linhagens, híbridos e variedades comerciais têm se tornado cada vez mais difíceis, no contexto de espécies submetidas a longos processos seletivos. Assim, recursos extras, além daqueles pertinentes à escolha de delineamentos genéticos, métodos de seleção e boa experimentação agrícola, fazem parte de uma tendência recente: o uso de procedimentos analíticos mais refinados, como o emprego de modelos lineares mistos (Hiraoka et al., 2011) e redes neurais artificiais (Mugnai et al., 2008), por exemplo, para o estudo mais detalhado dos componentes da média e da

variância de um caráter, para tentar predizer a parte herdável da variância, ou seja, o valor genético.

De maneira geral, a grande questão envolvida no melhoramento genético é o conhecimento do valor genético do indivíduo, para que se possa praticar a seleção com o máximo de acurácia. Para melhor predizer o valor genético de uma característica, é possível recorrer à informação fenotípica deste próprio indivíduo ou de seus aparentados (descendentes ou ancestrais) ou de informações sobre outras características correlacionadas. A junção de todas estas informações tem sido objeto de estudo por vários biometristas.

O valor genético é baseado no modelo aditivo, e tem desempenhado um papel importante no ganho de seleção de características complexas em plantas e animais (Crossa et al., 2010). Além deste modelo aditivo, têm-se utilizado BLUP, interações bayesianas (Piepho et al., 2008) e seleção genômica em plantas (Jannink et al., 2010) e animais (González-Recio et al., 2008). Atualmente, para o estudo de famílias tem se adotado o método dos modelos mistos REML/BLUP (REML é a máxima verossimilhança restrita, e BLUP, a melhor predição linear não viciada), que permite estimar os parâmetros genéticos e predizer os valores genotípicos das famílias (Resende, 2002). A seleção via modelos mistos tem sido largamente utilizada em inúmeros programas de melhoramento como em cana de açúcar (Oliveira et al., 2011), eucalipto (Bush et al., 2011), café (Petek et al., 2008) e batata (Borges et al., 2010).

Rosado et al. (2009) compararam vários métodos de seleção em *Eucaliptus urofila* e concluíram que os métodos de seleção combinada e BLUP proporcionaram maiores ganhos em relação aos métodos de seleção entre e dentro de famílias para todas as características avaliadas. Rocha et al. (2009) trabalhando com árvores de *Dipteryx alata* concluíram que a seleção combinada e BLUP proporcionaram maiores ganhos com relação aos métodos de seleção massal e seleção entre e dentro. Li e Lindgren (2006), usando dados simulados, compararam a seleção individual e a seleção combinada. Eles concluíram que a utilização de índices é vantajosa em situação de populações grandes e características com baixa ou média herdabilidade. David et al. (2003) compararam quatro métodos e 10 intensidades de seleção em mudas de *Pinus resinosa*, e concluíram que a seleção combinada foi o método que proporcionou maiores ganhos e maximizou a diversidade genética.

Um método mais recente para tornar mais eficiente a seleção de famílias são as redes neurais artificiais (RNAs). As RNAs tem sido utilizadas por inúmeros

autores para classificação de imagens em sensoriamento remoto (Aitkenhead & Aalders, 2008), análise de diversidade genética (Barbosa et al., 2011), identificação de genótipos superiores (Mugnai et al., 2008) e predição do valor genético em animais (Ventura et al., 2012). No entanto não existe relato na literatura sobre utilização das RNAs para predição do valor genético em experimentos balanceados em plantas. Portanto, o objetivo deste trabalho foi avaliar a eficiência da utilização das redes neurais artificiais na predição do valor genético em experimentos em blocos casualizados, e ainda definir uma estrutura de rede ideal para a predição do valor genético.

4.3. MATERIAL E MÉTODOS

4.3.1. Simulação dos dados obtidos de experimentos em blocos ao acaso

O delineamento utilizado para simulação dos experimentos foi blocos casualizados contendo seis repetições. Foram simulados 80 cenários (experimentos). Para cada cenário foram estabelecidos os valores da média (20, 40, 60, 80 e 100), da herdabilidade (10, 20, 30 e 40%) e do coeficiente de variação (5 e 10%) experimental (Tabela 9). Com o objetivo de validar as redes neurais, foram simulados 100 populações (repetições) mantendo a estrutura genética (matriz de variância e covariância, média, herdabilidade e coeficiente de variação). O modelo estatístico utilizado no processo de simulação foi:

$$Y_{ij} = \mu + G_i + B_j + \varepsilon_{ij}$$

em que

Y_{ij}: observação simulada de uma dada característica;

μ: média geral da característica, cujos valores estão na Tabela 1;

G_i: efeito associado ao i-ésimo genótipo, sendo $G_i \sim N(0, \sigma^2_g)$;

B_j: efeito associado ao j-ésimo bloco;

ε_{ij}: erro aleatório, sendo $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Tabela 9. Estimativas dos parâmetros que foram estabelecidos para definir as populações simuladas. Foram simulados dois tamanhos de experimentos em blocos ao acaso com seis repetições (150 e 200 genótipos por bloco).

Número de indivíduos por população							
150				200			
Car	**Média**	**h^2**	**CV (%)**	**Car**	**Média**	**h^2**	**CV (%)**
1(m=20)	20	10	5	1(m=20)	20	10	5

1(m=20)	20	10	10	1(m=20)	20	10	10
1(m=20)	20	20	5	1(m=20)	20	20	5
1(m=20)	20	20	10	1(m=20)	20	20	10
1(m=20)	20	30	5	1(m=20)	20	30	5
1(m=20)	20	30	10	1(m=20)	20	30	10
1(m=20)	20	40	5	1(m=20)	20	40	5
1(m=20)	20	40	10	1(m=20)	20	40	10
2(m=40)	40	10	5	2(m=40)	40	10	5
2(m=40)	40	10	10	2(m=40)	40	10	10
2(m=40)	40	20	5	2(m=40)	40	20	5
2(m=40)	40	20	10	2(m=40)	40	20	10
2(m=40)	40	30	5	2(m=40)	40	30	5
2(m=40)	40	30	10	2(m=40)	40	30	10
2(m=40)	40	40	5	2(m=40)	40	40	5
2(m=40)	40	40	10	2(m=40)	40	40	10
3(m=60)	60	10	5	3(m=60)	60	10	5
3(m=60)	60	10	10	3(m=60)	60	10	10
3(m=60)	60	20	5	3(m=60)	60	20	5
3(m=60)	60	20	10	3(m=60)	60	20	10
3(m=60)	60	30	5	3(m=60)	60	30	5
3(m=60)	60	30	10	3(m=60)	60	30	10
3(m=60)	60	40	5	3(m=60)	60	40	5
3(m=60)	60	40	10	3(m=60)	60	40	10
4(m=80)	80	10	5	4(m=80)	80	10	5
4(m=80)	80	10	10	4(m=80)	80	10	10
4(m=80)	80	20	5	4(m=80)	80	20	5
4(m=80)	80	20	10	4(m=80)	80	20	10
4(m=80)	80	30	5	4(m=80)	80	30	5
4(m=80)	80	30	10	4(m=80)	80	30	10
4(m=80)	80	40	5	4(m=80)	80	40	5
4(m=80)	80	40	10	4(m=80)	80	40	10
5(m=100)	100	10	5	5(m=100)	100	10	5
5(m=100)	100	10	10	5(m=100)	100	10	10
5(m=100)	100	20	5	5(m=100)	100	20	5
5(m=100)	100	20	10	5(m=100)	100	20	10
5(m=100)	100	30	5	5(m=100)	100	30	5
5(m=100)	100	30	10	5(m=100)	100	30	10
5(m=100)	100	40	5	5(m=100)	100	40	5
5(m=100)	100	40	10	5(m=100)	100	40	10

Car é característica simulada, m é a média da característica (que variou de 20 a 100), h^2 é a herdabilidade, CV é o coeficiente de variação.

Para simulação de dados, com distribuição normal, foram utilizadas as variáveis, proposto pelo teorema de Box-Muller:

$$x = \sqrt{-2\log_e(RND)}V\cos(2\pi RND)$$

e

$$y = \sqrt{-2\log_e(RND)V}\,Sen(2\pi RND)$$

sendo RND um número aleatório. Demonstra-se que os valores de x e y, assim obtidos, têm distribuição normal com média zero e variância V. Como o processo de simulação demandou-se a obtenção de n dados com média (μ) e variância (σ^2), foi utilizado a estratégia para gerar estes dados (z) por meio da seguinte expressão:

$$z = \mu + \frac{1}{2\theta}\sum_{i=1}^{\theta}(x_i + y_i)$$

sendo:

$$V = 2\theta\sigma^2$$

em que θ representa a repetibilidade de cada ponto simulado. Quanto maior o valor de θ estabelecido pelo usuário, mais precisa torna-se a simulação, porém mais lenta.

4.3.1.1. Simulação dos efeitos de blocos

Um conjunto de dados contendo n valores, em progressão aritmética, de razão r e média $\overline{X}$, em que o primeiro termo é denotado por X_1 e o último por X_n, a variância é dada por:

$$S^2 = \frac{n(n+1)}{3(n-1)^2}(X_n - \overline{X})^2$$

Assim, para estimar os efeitos de bloco foi admitida a existência de b efeitos fixos, cujos valores configuram uma progressão aritmética de razão r com a particularidade de que B_1 = -B_b e $\overline{B} = 0$. Logo, o valor B_b é estimado por meio de:

$$B_b = \frac{(n-1)\sqrt{3\phi_b}}{\sqrt{n(n+1)}}$$

e os demais efeitos estabelecidos considerando a razão da progressão aritmética dada por:

$$r = \frac{B_b - B_1}{b-1}$$

4.3.1.2. Simulação dos efeitos de genótipos

Para estimar os efeitos de genótipos foi necessário conhecer o valor da variância genética, que foi obtida a partir das informações de herdabilidade (h^2) e o

coeficiente de variação experimental (CV_e). Assim, primeiro é obtido o valor da variância ambiental por meio de:

$$\sigma^2 = \left(\frac{\mu CV_e}{100}\right)^2$$

Onde

σ^2 é a variância ambiental.

A herdabilidade (h^2) é calculada por:

$$h^2 = \frac{100\sigma_G^2}{\sigma_G^2 + \frac{1}{b}\sigma^2}$$

Onde

σ_G^2 é a variância genotípica

logo,

$$\sigma_G^2 = \frac{\sigma^2 h^2}{b(100 - h^2)}$$

4.3.1.3. Efeito aleatório de genótipos

Neste caso foi considerado que $G_i \sim NID(0, \sigma_G^2)$. Como o valor de σ_G^2 era conhecido, estimou-se os efeitos usando a função randômica descrita anteriormente,

4.3.1.4. Simulação dos erros aleatórios

Foi considerado que $\varepsilon_i \sim NID(0, \sigma^2)$. Como foi fornecido o valor do coeficiente de variação experimental e da média da característica, o valor de σ^2 torna-se conhecido e, portanto, os erros aleatórios e independentes podem ser estimados usando a função randômica descrita anteriormente.

Todas as simulações foram realizadas utilizando o aplicativo computacional GENES (Cruz, 2006).

4.3.2. Estabelecimento dos valores fenotípicos e genotípicos

Conhecidos o valor da média da característica e dos efeitos envolvidos, os valores fenotípicos (Y_{ij}) e genético (Z_i) foram estabelecidos por meio do modelo:

$$Y_{ij} = \mu + G_i + B_j + \varepsilon_{ij}$$

O valor genético foi obtido a partir de

$$Z_i = \mu + G_i$$

4.3.3. Simulação dos dados obtidos de experimentos em blocos ao acaso para fins de treinamento da rede

Para fins de treinamento da rede foram gerados dados simulados com as mesmas características dos dados simulados para as populações de validação, definidos pela media, herdabilidade e coeficiente de variação.

Neste trabalho foram simulados, para fins de treinamento, experimentos com o mesmo número de blocos (seis) dos experimentos originais, mas com número ampliado de genótipos (5000). Dimensões maiores certamente trarão benefícios no processo de predição do valor correto.

Para esta simulação considerou que cada bloco representava um vetor de valores fenotípicos a ser reproduzido com a mesma média, variância e covariância com os valores dos demais blocos. O vetor de valores genotípicos também foi reproduzido considerando que a sua variância e covariância com os valores dos vetores fenotípicos dos blocos eram representadas pela estimativa da variância genética. Assim, o problema consistia em obter o conjunto de dados ampliados (Zij ~ N (μ,σ^2) onde i=1, 2,..., 5000, j=1, 2, ..., 7 e matriz de variância e covariancia de dimensão 7x7, sendo Z_{i7} o valor genotípico verdadeiro) a partir dos dados originais (Yij ~ N (μ,σ^2) onde i=1, 2, ..., 150 ou 200, j=1,2, ..., 6 e matriz de variância e covariância de dimensão 6x6)

A eficácia do processo de treinamento foi avaliada por meio da correlação entre os valores da rede e o valor genético (saída) dos genótipos avaliados.

4.3.4. Simulação dos dados obtidos de experimentos em blocos ao acaso para fins de Validação da rede

Foi utilizado um procedimento similar ao descrito na simulação para fins de treinamento, porém com número de genótipos mais próximo da situação real de teste. Assim, foram utilizadas informações de novos experimentos simulados, com a mesma caracterização da média, herdabilidade e coeficiente de variação experimental, mas com apenas 150 e 200 genótipos. Destaca-se o fato de que os ensaios para fins de validação, apesar de também ter origem da simulação, não são subamostras do conjunto de dados de treinamento, mas novos ensaios com toda estrutura genética e ambiental preservada.

4.3.5. Arquitetura da rede neural

Neste trabalho foi utilizado as redes neurais multicamadas. A rede neural artificial proposta possui 1 camada de entrada, 3 camadas intermediária e 1 camada de saída. A camada de entrada possui 6 entradas (valor fenotípico do individuo no bloco). A camada de entrada baseou-se em uma matriz nxm, onde n é o número de blocos que variou de 1 a 6 e m o número de indivíduos em cada bloco, que foi 5000 no experimento de treinamento e 150 ou 200 nos experimentos de validação. Na camada intermediária o número de neurônios por camada variou de 1 a 10 neurônios na primeira camada, de 1 a 20 na segunda camada e de 1 a 8 na terceira camada. A camada de saída foi composta por 1 neurônio e a saída foi o valor genético da população, onde esse valor era conhecido no treinamento e desconhecido na validação. A melhor arquitetura da rede foi estabelecida por aquela com acurácia média superior, considerando as 43200 possibilidades, calculada pela multiplicação do número de neurônios em cada camada e as funções de ativação possíveis (10X20X8X3X3X3). As funções de ativação utilizadas foram a linear (purelin), tangente hiperbólica (tansig) e logarítmica (Logsig) e para treinamento Trainbr. Foi utilizado 1000 épocas (iterações). Foi utilizado aplicativo computacional MATLAB. Os comandos utilizados nesta RNA estão no anexo A.

4.3.6. Eficácia da RNA em estudos genéticos

A eficácia da RNA foi considerada por meio da estimava da superioridade da correlação entre o valor da rede e os valores genéticos em relação à correlação entre as médias fenotípicas e os valores genéticos. O quadrado desta última correlação representa a herdabilidade da característica analisada.

4.4. RESULTADO E DISCUSSÃO

4.4.1. Obtenção dos dados experimentais

A metodologia aplicada utilizando a técnica de decomposição espectral para simular populações com a mesma estrutura das populações originais, tais como herdabilidade, coeficiente de variação, média e a matriz de variância e covariância foram eficientes no presente estudo. Na simulação das características a partir de uma população com CV, h^2 e médias definidas, foi possível simular populações com estas características bem semelhantes (Tabela 10).

Tabela 10. Caracterização da população original e da população simulada.

Pop	h^2_{esp}	CV_{esp}	X_{espC1}	h^2_{sim}	CV_{SIM}	X_{SIMC1}	X_{espC2}	h^2_{sim}	CV_{SIM}	X_{SIMC2}	X_{espC3}	h^2_{sim}	CV_{SIM}	X_{SIMC3}	X_{espC4}	h^2_{sim}	CV_{SIM}	X_{SIMC4}	X_{espC5}	h^2_{sim}	CV_{SIM}	X_{SIMC5}
150	10	5	20	11.02	5.01	20	40	10.78	4.99	40	60	8.5	5.03	60	80	9.55	5	80	100	8.95	5.04	100
150	10	10	20	9.21	10.01	20	40	11.84	9.96	40	60	8.9	9.95	60	80	8.74	9.99	80	100	10.08	9.98	100
150	20	5	20	21.67	4.97	20	40	21.13	4.93	40	60	18.24	5.03	60	80	19.7	5.02	80	100	18.06	5.02	100
150	20	10	20	20.28	10.01	20	40	19.29	9.99	40	60	18.41	10.1	60	80	20.06	9.98	80	100	20.5	10.01	100
150	30	5	20	30.82	5.03	20	40	29.19	5.01	40	60	29.13	5.02	60	80	30.16	4.94	80	100	31.1	5.03	100
150	30	10	20	30.18	10	20	40	31.42	9.89	40	60	28.71	10.03	60	80	29.57	10.06	80	100	29.72	9.98	100
150	40	5	20	39.05	5.01	20	40	38.18	5.04	40	60	40.69	4.99	60	80	41.59	5.02	80	100	41.56	5	100
150	40	10	20	40.86	9.93	20	40	38.03	10.05	40	60	40.94	9.9	60	80	40.76	9.94	80	100	38.25	9.98	100
200	10	5	20	8.65	5.04	20	40	10.77	4.99	40	60	11.03	5	60	80	10.51	5.02	80	100	9.97	4.98	100
200	10	10	20	11.19	10.01	20	40	11.84	9.96	40	60	11.06	9.95	60	80	11.36	9.97	80	100	9.14	9.95	100
200	20	5	20	21.12	4.99	20	40	18.98	5.01	40	60	18.6	5.02	60	80	18.26	5.03	80	100	21.02	5.01	100
200	20	10	20	19.3	10.02	20	40	18.51	10	40	60	21	9.95	60	80	21.22	9.98	80	100	21.93	9.96	100
200	30	5	20	31.04	5	20	40	30.84	4.95	40	60	28.21	5	60	80	31.57	4.99	80	100	28.33	4.97	100
200	30	10	20	28.73	10.01	20	40	30.38	10	40	60	31.63	9.85	60	80	31.39	10	80	100	29.92	9.82	100
200	40	5	20	40	4.98	20	40	40.1	4.97	40	60	40.94	4.96	60	80	38.5	5.04	80	100	39.06	5.01	100
200	40	10	20	39.38	9.91	20	40	39.95	9.98	40	60	41.2	9.9	60	80	41.3	10	80	100	38.26	10.04	100

POP - tamanho da população, h^2_{esp} - herdabilidade esperada, CV_{esp} - coeficiente de variação esperado, X_{esp} - média esperada, h^2_{sim} - herdabilidade simulada, CV_{SIM} - coeficiente de variação simulado, X_{SIMC1} - média simulada da característica 1, X_{SIMC2} - média simulada da característica 2, X_{SIMC3} - média simulada da característica 3, X_{SIMC4} - média simulada da característica 4, X_{SIMC5} - média simulada da característica 5.

A média simulada foi idêntica a média esperada em todas as populações simuladas. Assim, o processo de simulação garante em 100% dos casos a manutenção da média esperada.

O coeficiente de variação simulado variou de 4,97 a 5,04 para C1, 4,93 a 5,04 para C2, 4,96 a 5,03 para C3, 4, 94 a 5,04 para C4 e 4, 97 a 5,04 para C5 para um coeficiente de variação esperado de 5. Para CV esperado de 10 o CV simulado variou de 9,91 a 10,02, 9,89 a 10,05, 9,85 a 10,03, 9,94 a 10,06 e 9,82 a 10,04 para C1, C2, C3, C4 e C5 respectivamente. Assim, a simulação gerou uma população onde o CV variou de 1,2 e 1,8% para CV esperado de 5 e 10% respectivamente.

A herdabilidade simulada variou de 8,5 a 11,84, 18,06 a 21,93, 28,21 a 31,63, 38,03 a 41,59% para herdabilidade esperada de 10, 20, 30 e 40% respectivamente. Assim, observamos uma variação no h^2 simulado de 18,4%, 9,7%, 6,9% e 4,93% para h^2 esperada de 10, 20, 30 e 40% respectivamente. A herdabilidade foi o parâmetro estudado onde foi observado maiores variações. Este fato ocorre devido a maior dificuldade em manter constante a variância genética. A medida que a h^2 esperada aumenta, a variação da h^2 simulada diminui. Portanto a metodologia apresentada aqui para simulação de populações é mais eficiente para características qualitativas, ou seja, menos influenciada pelo ambiente, pois é possível manter as variâncias esperadas e assim, manter constante o valor da herdabilidade. Com estudos de simulação Wang et al. (2007) mostraram que a utilização de componentes principais a partir das medidas de herdabilidade (PCH) tem um poder de ganho muito maior que a técnica de componentes principais normalmente utilizada sobre as médias das populações.

O tamanho das populações simuladas (150 ou 200 indivíduos) não influenciou no processo de simulação. A média não variou em nenhum dos tamanhos de população. O CV variou de 4,93 a 5,04 e 4,95 a 5,04% nas populações com 150 e 200 genótipos respectivamente para o CV esperado de 5%. Com o CV esperado de 10% a variação foi de 9,89 a 10,06%.

4.4.2. Obtenção de dados ampliados para fins de treinamento da rede

A obtenção de dados de treinamento torna-se uma prática extremamente importante para garantir uma boa eficiência na validação das redes neurais. A rede neural toma os dados de treinamento como padrão (Mather & Koch, 2011). As características dos dados de treinamento selecionados para análise são de considerável importância para a performance das RNAs. Os dados de treinamento

devem ser definido para a analise como sendo dados representativos da sua população. A qualidade e o tamanho do conjunto de dados para treinamento são de extrema importância para a eficiência da rede neural (Kavzoglu, 2009). Portanto, é muito importante que os dados de treinamento utilizados para a predição do valor genético sejam dados representativos do experimento original, pois é necessário que a rede aprenda com estes dados para posteriormente estimar o valor genético nos demais experimentos.

O tamanho do conjunto de treinamento é crucial para o desempenho das redes. Existe uma relação direta entre o tamanho do conjunto de dados para treinamento e a confiabilidade das estimativas dos dados de validação da rede. O tamanho da amostra esta relacionada principalmente com as propriedades estatísticas utilizadas pelas redes neurais para o treinamento. Uma amostra com poucos indivíduos não é suficiente para uma rede neural reconhecer todas as classes possíveis. Uma amostra maior pode tornar a rede mais especifica e melhorar a confiabilidade dos resultados, porém exigem maior tempo computacional para a execução das tarefas de treinamento da rede (Kavzoglu, 2001).

Existe duas formas de obter um conjunto de treinamento eficiente para ser utilizado em uma RNA: a primeira é a utilização de séries históricas e a segunda é a simulação de dados. A utilização de séries históricas é reportada por Ventura et al. (2012) que trabalharam com dados de 19240 animais bovinos da raça Tabapuã, provenientes de 152 fazendas localizadas em diversos estados brasileiros, entre os anos de 1976 e 1995. Eles utilizaram para a predição do valor genético do peso aos 205 dias de idade e usaram o algoritmo LM (Levenberg Marquardt) para treinamento da rede. Os autores concluíram que os valores genéticos obtidos pela RNA era altamente correlacionado com os obtidos pelo BLUP.

No entanto quando não é possível obter uma série histórica para compor um conjunto de dados de treinamento, é possível obter estes dados através do processo de simulação, como foi utilizado no presente estudo.

Com o objetivo de conseguir uma população de treinamento que mantivesse a estrutura genética do experimento inicial, foi utilizado o processo de simulação de populações no programa GENES. Para a simulação a estrutura genética do experimento inicial foi mantido, ou seja a média, a herdabilidade e o coeficiente de variação são mantidos constantes. Além disso, através da utilização da técnica multivariada da decomposição espectral foi possível manter invariável a matriz de

variância e covariância. Com isto a população de treinamento foi simulada com a mesma estrutura genética da população inicial (Tabela 11).

Tabela 11. Estrutura genética da população de treinamento com 5000 genótipos por bloco.

Pop	h^2_{sim}	CV_{SIM}	X_{SIMC1}	h^2_{sim}	CV_{SIM}	X_{SIMC2}	h^2_{sim}	CV_{SIM}	X_{SIMC3}	h^2_{sim}	CV_{SIM}	X_{SIMC4}	h^2_{sim}	CV_{SIM}	X_{SIMC5}
150	11.02	5.01	20	10.78	4.99	40	8.50	5.03	60	9.55	5.00	80	8.95	5.04	100
150	9.21	10.01	20	11.84	9.96	40	8.90	9.95	60	8.74	9.99	80	10.08	9.98	100
150	21.67	4.97	20	21.13	4.93	40	18.24	5.03	60	19.70	5.02	80	18.06	5.02	100
150	20.28	10.01	20	19.29	9.99	40	18.41	10.10	60	20.06	9.98	80	20.50	10.01	100
150	30.82	5.03	20	29.19	5.01	40	29.13	5.02	60	30.16	4.94	80	31.10	5.03	100
150	30.18	10.00	20	31.42	9.89	40	28.71	10.03	60	29.57	10.06	80	29.72	9.98	100
150	39.05	5.01	20	38.18	5.04	40	40.69	4.99	60	41.59	5.02	80	41.56	5.00	100
150	40.86	9.93	20	38.03	10.05	40	40.94	9.90	60	40.76	9.94	80	38.25	9.98	100
200	8.65	5.04	20	10.77	4.99	40	11.03	5.00	60	10.51	5.02	80	9.97	4.98	100
200	11.19	10.01	20	11.84	9.96	40	11.06	9.95	60	11.36	9.97	80	9.14	9.95	100
200	21.12	4.99	20	18.98	5.01	40	18.60	5.02	60	18.26	5.03	80	21.02	5.01	100
200	19.30	10.02	20	18.51	10.00	40	21.00	9.95	60	21.22	9.98	80	21.93	9.96	100
200	31.04	5.00	20	30.84	4.95	40	28.21	5.00	60	31.57	4.99	80	28.33	4.97	100
200	28.73	10.01	20	30.38	10.00	40	31.63	9.85	60	31.39	10.00	80	29.92	9.82	100
200	40.00	4.98	20	40.10	4.97	40	40.94	4.96	60	38.50	5.04	80	39.06	5.01	100
200	39.38	9.91	20	39.95	9.98	40	41.20	9.90	60	41.30	10.00	80	38.26	10.04	100

POP - tamanho da população inicial, h^2_{sim} - herdabilidade simulada, CV_{SIM} - coeficiente de variação simulado, X_{SIMC1} - média simulada da característica 1, X_{SIMC2} - média simulada da característica 2, X_{SIMC3} - média simulada da característica 3, X_{SIMC4} - média simulada da característica 4, X_{SIMC5} - média simulada da característica 5

Podemos observar que as populações de treinamento simuladas com 5000 genótipos por bloco manteve toda a estrutura genética das populações simuladas inicialmente com 150 e 200 genótipos, ou seja, a herdabilidade, a média e o coeficiente de variação foram mantidos constantes, apesar do aumento do número de genótipos no experimento. Portanto é possível através de um experimento realizado em campo ou casa de vegetação, obter um experimento com maior número de genótipos que mantém a estrutura genética do experimento inicial.

Uma característica marcante do processo de simulação, que além da estrutura genética ser mantida, também é simulado o valor genético desta população de treinamento. Através do valor genético simulado a rede será capaz de aprender sobre aquele experimento e predizer o valor genético real do experimento inicial.

4.4.3. Desempenho das redes neurais para fins de predição do valor genético

Observou-se que, para a predição do valor genético as redes neurais foram superiores a utilização do valor fenotípico médio (Tabela 12). Em 70% das redes simuladas o ganho foi acima de 80%, do valor de rede em relação ao valor fenotípico. Portanto, a utilização das redes neurais para predição do valor genético é melhor que a utilização do valor fenotípico, pois a rede consegue diminuir o efeito ambiental (ruído) sobre um determinado experimento. Com isso, o valor de rede se aproxima mais do valor genético verdadeiro.

Chen et al. (2010) verificaram acurácia das RNAs variando de 88 a 94% para classificação de sementes de milho. A alta acurácia verificada com o uso das RNAs pode ser explicado pelas RNAs usarem modelos não-lineares para gerar uma saída (resposta) (Dai et al., 2011).

O uso das RNAs é muito importante na agricultura devido ao seu potencial em resolver problemas complexos para as técnicas computacionais e matemáticas convencionais (Huang et al., 2010), e principalmente na seleção de genótipos, pois o efeito ambiental pode mascarar o potencial de alguns genótipos, fazendo com que a seleção seja realizada de forma errônea. Como as RNAs conseguem diminuir o efeito ambiental (ruído) através da retropropagação do erro, ela é uma técnica com elevado potencial para a predição do valor genético.

Tabela 12. Porcentagem de ganho da rede (G) em relação a média nas 100 populações avaliadas para a característica: C1 (m=20), C2 (m=40), C3 (m=60), C4 (m=80), C5 (m=100).

POP	h^2	CV	G C1	G C2	G C3	G C4	G C5
150	10	5	82	99	67	100	86
150	10	10	100	89	91	96	99
150	20	5	61	63	92	100	71
150	20	10	90	98	93	99	81
150	30	5	66	96	97	75	98
150	30	10	93	79	89	86	86
150	40	5	99	99	80	90	65
150	40	10	96	90	59	92	97
200	10	5	83	100	78	65	61
200	10	10	100	79	99	77	94
200	20	5	100	85	74	84	68
200	20	10	94	97	94	80	67
200	30	5	100	79	75	53	88
200	30	10	87	91	93	91	95
200	40	5	71	90	91	86	73
200	40	10	70	73	83	89	82

POP - número de indivíduos em cada população; h^2 - herdabilidade; CV - coeficiente de variação.

Observou-se que não houve diferenças entre os experimentos simulados com diferentes herdabilidade (10, 20, 30 e 40%) e CV (5 e 10%). Portanto mesmo que no experimento o efeito ambiental seja alto, o valor de rede consegue se aproximar do valor genético verdadeiro. O estudo da predição do valor genético em diferentes níveis de herdabilidade tem sido realizado por inúmeros pesquisadores, principalmente para QTLs (Meuwissen et al., 2001; Bernardo & Yu, 2007).

As RNAs foram eficientes para a predição do valor genético para herdabilidade baixa (10 e 20%) e moderada (30 e 40%). Portanto, a utilização das redes em características quantitativas (baixa herdabilidade) pode ser realizada, sendo uma alternativa promissora para predição do valor genético.

Alguns autores também encontraram alta eficiência na predição do valor genético utilizando redes neurais (Cavero et al., 2008; Ventura et al., 2012) Em todos os trabalhos verificou-se que foram avaliadas características quantitativas e que as RNAs foram eficientes para a predição do valor genético diminuindo os gastos de tempo e menor custo computacional.

A correlação entre o valor genético e o valor de rede foi maior que a correlação entre o valor fenotípico e o valor de rede em todas as características. A

diferença entre estas correlações variou de 0,64 a 10,3% (Tabela 13, 14, 15, 16 e 17).

A herdabilidade pode ser definida como o quadrado da correlação. Assim a herdabilidade calculada pela rede foi maior que à calculada por meio da média fenotípica. Portanto, o ganho de seleção será maior se utilizarmos a correlação entre o valor de rede e o valor genotípico, pois o ganho de seleção esta correlacionada com a herdabilidade.

Tabela 13. Correlação entre o valor genético e o valor da rede (r VGxVR), valor genético e o valor fenotípico (r VGxVF) e o valor de rede e o valor fenotípico (r VRxVF) para a característica 1 (m=20).

POP	h^2	CV	r VGxVR	r VGxVF	r VRxVF
150	10	5	0.1536	0.1254	0.7758
150	10	10	0.1205	0.0984	0.8124
150	20	5	0.1972	0.1683	0.5962
150	20	10	0.2338	0.2111	0.9414
150	30	5	0.4069	0.3689	0.6699
150	30	10	0.3506	0.3313	0.9376
150	40	5	0.4283	0.3982	0.9334
150	40	10	0.4501	0.4075	0.8927
200	10	5	0.2242	0.1574	0.5496
200	10	10	0.1714	0.1173	0.6295
200	20	5	0.2142	0.1985	0.9184
200	20	10	0.2505	0.2302	0.9346
200	30	5	0.3385	0.3175	0.9414
200	30	10	0.3436	0.2989	0.8273
200	40	5	0.4278	0.4071	0.9179
200	40	10	0.3402	0.3294	0.9669

POP - número de indivíduos em cada população; h^2 - herdabilidade; CV - coeficiente de variação.

Não houve diferença entre os experimentos contendo 150 ou 200 genótipos por bloco (Tabela 13, 14, 15, 16 e 17). Portanto podemos utilizar as redes neurais artificiais na predição do valor genético para experimento com número reduzido de genótipos até experimentos com uma grande quantidade de genótipos, como experimentos de pré-melhoramento, onde são avaliados uma grande quantidade de famílias até experimentos em fase final de um programa de melhoramento, onde o número de genótipos é reduzido.

Crossa et al. (2010) verificaram correlações entre o BLUP e o valor fenotípico de 0,41 a 0,51. A correlação entre valor de rede e o valor fenotípico no

presente trabalho variou entre 0,5 e 0,98. Este alto valor de correlação é importante, pois verificou-se que o valor de rede esta bem próximo do valor real mensurado a campo. Heffner et al. (2009) concluíram que a correlação entre os valores genéticos verdadeiros e os valores genéticos estimados é suficiente para considerar a seleção dos melhores genótipos em um programa de melhoramento, principalmente utilizando seleção genômica, através de marcadores moleculares.

Crossa et al. (2010) concluíram que o maior ganho de seleção, obtido pela maior eficiência na predição do valor genético, pode ocorrer pelo maior número de características ou marcadores avaliados ou pela melhoria dos métodos utilizados para a predição do valor genético. No presente trabalho as RNAs foram eficientes por apresentarem uma alta correlação com o valor genético, principalmente em herdabilidades mais altas. Portanto, com a utilização das RNAs na predição do valor genético é possível avaliar com maior precisão cada genótipo sem o efeito ambiental sobre este. Assim, a seleção das melhores progênies em um programa de melhoramento será realizada com maior eficiência, sobre os melhores genótipos, e não sobre os melhores fenótipos.

Tabela 14. Correlação entre o valor genético e o valor da rede (r VGxVR), valor genético e o valor fenotípico (r VGxVF) e o valor de rede e o valor fenotípico (r VRxVF) para a característica 2 (m=40).

POP	h^2	CV	r VGxVR	r VGxVF	r VRxVF
150	10	5	0.1580	0.1156	0.7270
150	10	10	0.1175	0.0870	0.7347
150	20	5	0.1484	0.1361	0.8845
150	20	10	0.2280	0.1906	0.8145
150	30	5	0.3377	0.3012	0.8976
150	30	10	0.2603	0.2272	0.8745
150	40	5	0.4810	0.4321	0.8878
150	40	10	0.4413	0.4197	0.9495
200	10	5	0.1000	0.0692	0.6744
200	10	10	0.1538	0.1163	0.7090
200	20	5	0.2330	0.2164	0.9295
200	20	10	0.2092	0.1836	0.8975
200	30	5	0.2450	0.2252	0.9185
200	30	10	0.3267	0.3032	0.9195
200	40	5	0.4144	0.3916	0.9302
200	40	10	0.3836	0.3753	0.9807

POP - número de indivíduos em cada população; h^2 - herdabilidade; CV - coeficiente de variação.

Não foi observado diferença na predição do valor genético pelas RNAs quando variou o coeficiente de variação (5 e 10%) (Tabela 13, 14, 15, 16 e 17). Portanto pode-se estimar o valor genético através das redes neurais mesmo quando a variação dentro do bloco for alta. A eficiência da seleção em um programa de melhoramento depende da predição do valor genético a partir do valor fenotípico ou da eficiência de outro critério utilizado (Lorenzana e Bernardo, 2009). Assim as RNAs podem tornar a seleção de genótipos baseados no valor genético uma tarefa mais fácil e diminuir o erro causado pelo efeito do ambiente sobre o fenótipo.

Tabela 15. Correlação entre o valor genético e o valor da rede (r VGxVR), valor genético e o valor fenotípico (r VGxVF) e o valor de rede e o valor fenotípico (r VRxVF) para a característica 3 (m=60).

POP	h^2	CV	r VGxVR	r VGxVF	r VRxVF
150	10	5	0,1858	0,1603	0,7710
150	10	10	0,1231	0,0647	0,5032
150	20	5	0,2847	0,2193	0,6318
150	20	10	0,3473	0,2771	0,7328
150	30	5	0,3035	0,2340	0,7982
150	30	10	0,3785	0,3577	0,9351
150	40	5	0,4134	0,3914	0,9313
150	40	10	0,3649	0,3487	0,9224
200	10	5	0,1463	0,1349	0,9320
200	10	10	0,1014	0,0699	0,6791
200	20	5	0,2635	0,2361	0,8518
200	20	10	0,2034	0,1917	0,9338
200	30	5	0,3040	0,2942	0,9651
200	30	10	0,3434	0,3182	0,9277
200	40	5	0,9321	0,3715	0,9578
200	40	10	0,3610	0,3427	0,9614

POP - número de indivíduos em cada população; h^2 - herdabilidade; CV - coeficiente de variação.

Observou-se que em experimentos simulados com herdabilidade menor (10 e 20%) a diferença entre a correlação entre o valor de rede e o valor genético e a correlação entre o valor fenotípico e o valor genético foi maior (Tabela 13, 14, 15, 16 e 17). Este fato ocorre porque quando a herdabilidade é menor e o efeito ambiental sobre esta característica é maior, ou seja, o ruído é maior. Então, a rede consegue diminuir mais o ruído que ocorre em características de baixa herdabilidade. Quando a herdabilidade é mais alta (30 e 40%), o efeito ambiental é menor, e o ruído também. Assim, mesmo que a rede reduza o ruído, esta redução será pequena.

Ventura et al. (2012) trabalhando com características quantitativas em bovinos, ou seja, características de baixa herdabilidade e muito afetadas pelo ambiente, verificaram que as redes neurais foram eficazes na predição do valor genético e verificaram uma correlação variando de 0.68 a 0.74 entre o valor de rede e o BLUP para dados que já estejam inseridos no conjunto de dados de treinamento. Assim, as RNAs conseguem predizer o valor genético com a mesma acurácia do BLUP, que é a metodologia utilizada por inúmeros autores atualmente para a predição do valor genético (Piepho et al., 2008; Gianola et al., 2010; Arnhold et al., 2012).

Tabela 16. Correlação entre o valor genético e o valor da rede (r VGxVR), valor genético e o valor fenotípico (r VGxVF) e o valor de rede e o valor fenotípico (r VRxVF) para a característica 4 (m=80).

POP	h^2	CV	r VGxVR	r VGxVF	r VRxVF
150	10	5	0,1506	0,1110	0,7093
150	10	10	0,1612	0,0949	0,5719
150	20	5	0,3274	0,2374	0,7769
150	20	10	0,2878	0,1848	0,6247
150	30	5	0,2197	0,2034	0,9444
150	30	10	0,3759	0,3533	0,9385
150	40	5	0,4821	0,4592	0,9362
150	40	10	0,4264	0,3725	0,8509
200	10	5	0,1235	0,1102	0,8004
200	10	10	0,1407	0,1003	0,6721
200	20	5	0,2815	0,2480	0,8564
200	20	10	0,1963	0,1812	0,93
200	30	5	0,3449	0,3396	0,9482
200	30	10	0,3128	0,2495	0,8025
200	40	5	0,4416	0,4082	0,9242
200	40	10	0,3907	0,3606	0,9166

POP - número de indivíduos em cada população; h^2 - herdabilidade; CV - coeficiente de variação.

O modelo genético padrão, tendo como resultado o valor fenotípico é a soma do valor genético e valor ambiental (Crossa et al., 2010). Assim, como o valor obtido nos experimentos é o valor fenotípico, se o valor ambiental for muito alto, estaremos muito longe de obter o verdadeiro valor genético. Para um programa de melhoramento que visa a seleção dos melhores genótipos, a estimativa do valor genético é extremamente importante para a tomada de decisão, sem deixar que o ambiente influencie nesta decisão. Como as RNAs utilizam de equações não-

lineares para estimar o valor genético, é possível chegar mais próximo do verdadeiro valor genético.

Tabela 17. Correlação entre o valor genético e o valor da rede (r VGxVR), valor genético e o valor fenotípico (r VGxVF) e o valor de rede e o valor fenotípico (r VRxVF) para a característica 5 (m=100).

POP	h^2	CV	r VGxVR	r VGxVF	r VRxVF
150	10	5	0,1885	0,1798	0,9577
150	10	10	0,1332	0,0947	0,6902
150	20	5	0,2577	0,2245	0,6783
150	20	10	0,2572	0,2281	0,9314
150	30	5	0,3419	0,3106	0,9223
150	30	10	0,3016	0,2719	0,8941
150	40	5	0,4524	0,4288	0,8920
150	40	10	0,4195	0,3753	0,8914
200	10	5	0,0943	0,0802	0,7541
200	10	10	0,0972	0,0759	0,7287
200	20	5	0,2384	0,2142	0,8548
200	20	10	0,2164	0,1854	0,8069
200	30	5	0,2681	0,2531	0,9313
200	30	10	0,2346	0,1848	0,7407
200	40	5	0,4291	0,4227	0,9867
200	40	10	0,4248	0,3981	0,9168

POP - número de indivíduos em cada população; h^2 - herdabilidade; CV - coeficiente de variação.

A fase crucial para garantir a eficiência de uma RNA é a sua estruturação. A especificação do número e do tamanho das camadas ocultas na rede é um ponto crítico para garantir a capacidade da rede aprender as características dos conjuntos de dados de treinamento e depois reconhecer novos dados que são inseridos durante o processo de validação e teste. O número de nós nas camadas intermediárias definem a complexidade do modelo de rede neural para descrever as relações e a estrutura inerente aos dados de treinamento (Kavzoglu & Mather, 2003).

No presente estudo utilizou-se três camadas ocultas, pois o tempo computacional gasto em uma rede com mais camadas ocultas é muito longo. Este fato ocorre devido ao número de interações entre o número de camadas ocultas, o número de neurônios em cada camada e o número de funções de ativação. Mas e Flores (2008) trabalhando com redes neurais em sensoriamento remoto concluíram que uma camada intermediária já é o suficiente para conseguir boa acurácia na

utilização das RNAs. Ardö et al. (1997) compararam a acurácia das RNAs com uma, duas e três camadas intermediárias, e concluíram que não a relação entre o número de camadas intermediárias e acurácia da rede neural.

No presente estudo utilizou-se seis entradas que foram o valor fenotípico de cada bloco previamente simulado. A escolha correta dos dados de entrada é de suma importância para maior eficiência da rede, podendo a inclusão ou exclusão de qualquer variável influenciar positiva e negativamente na saída (Dai et al., 2011).

Para encontrar uma rede ideal na predição do valor genético, o número de neurônios de cada camada intermediária e a função de ativação utilizada foram testadas em diferentes configurações. Foi observado diferenças entre o número de neurônios nas camadas e as funções de ativação (Tabela 18). O número de neurônios na camada 1 variou de 2 a 10, na camada 2 de 2 a 10 e na camada 3 de 2 a 8.

Tabela 18. Número de neurônios ideais para as camadas intermediárias. O número de neurônios na camada intermediária variou de 2 a 10 na camada 1 (primeiro número), 2 a 20 na camada 2 (segundo número) e de 2 a 8 na camada 3 (terceiro núemro).

POP	h^2	CV	C1(m=20)	C2(m=40)	C3(m=60)	C4(m=80)	C5(m=100)
150	10	5	6-8-6	9-4-6	7-6-4	7-6-8	8-10-8
150	10	10	6-2-8	4-10-8	5-4-8	5-4-6	10-2-8
150	20	5	4-8-6	10-6-6	7-8-8	5-2-8	6-4-8
150	20	10	7-6-4	10-2-8	4-4-4	5-4-4	7-2-8
150	30	5	5-6-4	10-6-4	7-2-8	10-8-8	7-6-6
150	30	10	6-2-4	10-10-8	7-8-8	8-6-8	9-10-8
150	40	5	8-2-4	5-4-8	4-6-6	6-8-8	4-10-4
150	40	10	6-4-8	4-8-8	7-10-6	4-4-4	4-6-8
200	10	5	9-6-4	10-2-8	9-10-6	8-2-6	10-6-4
200	10	10	10-6-8	7-8-4	7-6-4	5-2-8	5-2-6
200	20	5	6-2-4	7-6-8	8-6-4	5-10-8	4-8-8
200	20	10	5-2-4	6-6-6	10-6-8	10-4-8	4-6-6
200	30	5	7-2-8	10-2-8	7-8-8	8-10-6	7-10-4
200	30	10	9-2-8	7-10-6	9-10-4	9-6-4	4-6-6
200	40	5	8-6-4	8-6-8	6-10-8	6-8-8	6-8-8
200	40	10	7-4-4	10-4-6	10-2-8	5-4-6	5-8-6

POP - número de indivíduos em cada população; h^2 - herdabilidade; CV - coeficiente de variação, C - característica

Na camada 2 foi utilizado 20 neurônios no processo de treinamento e validação. Portanto é possível diminuir o tempo gasto para validação da rede, diminuindo para 10 neurônios na segunda camada. O estudo do número de neurônios por camada é importante para diminuir o tempo gasto para validação da

rede e diminuir o custo computacional (Dai et al., 2011). Neste experimento foram gastos cerca de 30 horas para validação de cada rede, com 43200 iterações. Se o número de neurônios na camada 2 for reduzida para 10 neurônios, o tempo gasto reduzirá a metade, pois o número de iterações será reduzido para 21600. Pelos resultados da tabela 18 observamos que é possível reduzir o número de neurônios da camada intermediária sem perder eficiência, pois nenhuma das redes ideais tiveram mais que 10 neurônios na camada intermediária.

Kumar et al. (2002) trabalhando com estimação da evapotrasnpiração em grama, verificaram que uma rede com três camadas intermediárias com 7, 6 e 1 neurônios na primeira, segunda e terceira camada intermediária respectivamente, eram suficientes para obter erros mínimos quando comparado com os métodos convencionais como Penman-Monteith e lisímetro. Cavero et al. (2008) verificaram que três camadas intermediárias com 5, 10 e 1 neurônios na primeira, segunda e terceira camada intermediária respectivamente, são suficientes para a predição de mastite em animais bovinos.

As funções de ativação utilizadas em cada cenário estão dispostas na tabela 19. Podemos observar que não houve uma tendência na utilização das funções de ativação. Assim não foi possível determinar qual das três funções logsig, tansig ou linear foi melhor.

Tabela 19. Função de ativação da rede ideal. Os números correspondem a função de ativação utilizada na rede ideal nas camadas intermediárias 1, 2 e 3 respectivamente.

POP	h^2	CV	C1(m=20)	C2(m=40)	C3(m=60)	C4(m=80)	C5(m=100)
150	10	5	2-1-1	2-1-3	1-3-2	3-1-1	3-1-1
150	10	10	3-3-2	1-3-2	1-3-1	1-3-3	3-1-3
150	20	5	2-3-3	2-1-2	3-2-2	2-3-2	2-1-1
150	20	10	3-3-2	3-1-1	3-1-1	3-2-1	3-1-2
150	30	5	2-1-2	3-1-1	2-3-2	2-1-1	2-1-2
150	30	10	2-1-1	3-1-1	3-1-3	3-1-1	1-3-1
150	40	5	2-2-3	3-1-2	2-1-2	1-3-2	1-1-3
150	40	10	2-3-1	3-1-1	2-2-3	2-3-1	2-3-1
200	10	5	3-2-1	3-1-1	3-1-2	1-3-2	3-3-2
200	10	10	3-3-1	2-3-1	3-3-2	3-1-1	2-1-3
200	20	5	3-2-2	3-1-1	1-3-2	3-2-3	3-1-3
200	20	10	3-3-2	1-3-1	3-3-1	2-1-2	3-1-1
200	30	5	1-1-1	1-3-3	2-1-1	3-3-2	3-2-1
200	30	10	2-1-1	1-3-2	1-3-1	3-2-1	1-3-2
200	40	5	1-3-1	2-1-2	2-3-1	2-3-1	2-1-3
200	40	10	2-1-1	3-1-1	2-1-1	2-3-2	3-3-1

POP - número de indivíduos em cada população; h^2 - herdabilidade; CV - coeficiente de variação, C – característica. O número 1 corresponde a função tansig, o número 2 a função logsig e o número 3 a função purelin (linear).

4.5. CONCLUSÃO

As redes neurais artificiais foram eficientes para predizer o valor genético em experimentos balanceados em blocos ao acaso para características quantitativas (muito influenciadas pelo ambiente).

A estrutura da rede com três camadas intermediárias com 10 neurônios na primeira camada, 20 na segunda camada e 8 neurônios na terceira camada foram eficientes na predição do valor genético.

4.6. REFERÊNCIAS BIBLIOGRÁFICAS

AITKENHEAD, M.; AALDERS, I. Classification of Landsat Thematic Mapper imagery for land cover using neural networks. **International Journal of Remote Sensing**, v. 29, n. 7, p. 2075-2084, 2008.

ARDÖ, J.; PILESJÖ, P.; SKIDMORE, A. Neural networks, multitemporal Landsat Thematic Mapper data and topographic data to classify forest damages in the Czech Republic. **Canadian Journal of Remote Sensing**, v. 23, n. 3, p. 217-229, 1997.

ARNHOLD, E.; MORA, F.; PACHECO, C. A. P.; CARVALHO, H. W. L. Prediction of genotypic values of maize for the agricultural frontier region in northeastern Maranhão. **Crop Breeding and Applied Biotechnology**, v. 12, 2012.

BARBOSA, C. D.; VIANA, A. P.; QUINTAL, S. S. R.; PEREIRA, M. G. Artificial neural network analysis of genetic diversity in Carica papaya L. **Crop Breeding and Applied Biotechnology**, v. 11, n. 3, p. 224-231, 2011.

BASTOS, I. T.; BARBOSA, M. H. P.; CRUZ, C. D.; BURNQUIST, W. L.; BRESSIANI, J. A.; SILVA, F. Análise dialélica em clones de cana-de-açúcar. **Bragantia**, v. 62, n. 2, p. 199-206, 2003.

BERNARDO, R.; YU, J. Prospects for genomewide selection for quantitative traits in maize. **Crop Science**, v. 47, n. 3, p. 1082-1090, 2007.

BORGES, V.; FERREIRA, P. V.; SOARES, L.; SANTOS, G. M.; SANTOS, A. M. M. Sweet potato clone selection by REML/BLUP procedure. **Acta Scientiarum. Agronomy**, v. 32, n. 4, p. 643-649, 2010.

BUSH, D.; KAIN, D.; MATHESON, C.; KANOWSKI, P. Marker-based adjustment of the additive relationship matrix for estimation of genetic parameters - an example using *Eucalyptus cladocalyx*. **Tree Genetics & Genomes**, v. 7, n. 1, p. 23-35, 2011.

CAVERO, D.; TÖLLE, K. H.; HENZE, C.; BUXADÉ, C.; KRIETER, J. Mastitis detection in dairy cows by application of neural networks. **Livestock Science**, v. 114, n. 2, p. 280-286, 2008.

CHEN, X.; XUN, Y.; LI, W.; ZHANG, J. Combining discriminant analysis and neural networks for corn variety identification. **Computers and electronics in agriculture**, v. 71, p. S48-S53, 2010.

CROSSA, J.; DE LOS CAMPOS, G.; PEREZ, P.; GIANOLA, D.; BURGUENO, J.; ARAUS, J. L.; MAKUMBI, D.; SINGH, R. P.; DREISIGACKER, S.; YAN, J.; ARIEF, V.; BANZIGER, M.; BRAUN, H. J. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. **Genetics**, v. 186, n. 2, p. 713-724, 2010.

DAI, X.; HUO, Z.; WANG, H. Simulation for response of crop yield to soil moisture and salinity with artificial neural network. **Field Crops Research**, v. 121, n. 3, p. 441-449, 2011.

DAVID, A.; PIKE, C.; STINE, R. Comparison of selection methods for optimizing genetic gain and gene diversity in a red pine (*Pinus resinosa* Ait.) seedling seed orchard. **Theoretical and Applied Genetics**, v. 107, n. 5, p. 843-849, 2003.

DONÁ, A. A.; MIRANDA, G. V.; DELIMA, R. O.; CHAVES, L. G.; GAMA, E. E. G. Genetic parameters and predictive genetic gain in maize with modified recurrent selection method. **Chilean Journal of Agricultural Research**, v. 72, p. 1, 2012.

GARCIA, C.; NOGUEIRA, M. Utilização da metodologia REML/BLUP na seleção de clones de eucalipto. **Scientia Forestalis**, v. 68, p. 107-112, 2005.

GIANOLA, D.; WU, X. L.; MANFREDI, E.; SIMIANER, H. A non-parametric mixture model for genome-enabled prediction of genetic value for a quantitative trait. **Genetica**, v. 138, n. 9-10, p. 959-977, 2010.

GONZÁLEZ-RECIO, O.; GIANOLA, D.; LONG, N.; WEIGEL, K. A.; ROSA, G. J. M.; AVENDAÑO, S. Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. **Genetics**, v. 178, n. 4, p. 2305-2313, 2008.

HEFFNER, E. L.; SORRELLS, M. E.; JANNINK, J. L. Genomic selection for crop improvement. **Crop Science**, v. 49, n. 1, p. 1-12, 2009.

HIRAOKA, Y.; KURAMOTO, N.; OHIRA, M.; OKAMURA, M.; TANIGUCHI, T.; FUJISAWA, Y. Estimation of genetic data and breeding values of traits related to wax production in *Rhus succedanea* L. clones using the REML/BLUP method. **Journal of Forest Research**, v. 16, n. 6, p. 509-517, 2011.

HUANG, Y.; LAN, Y.; THOMSON, S. J.; FANG, A.; HOFFMANN, W. C.; LACEY, R. E. Development of soft computing and applications in agricultural and biological engineering. **Computers and Electronics in Agriculture**, v. 71, n. 2, p. 107-127, 2010.

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

JOMBART, T.; DEVILLARD, S.; BALLOUX, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. **BMC Genet**, v. 11, p. 94, 2010.

KAVZOGLU, T. **An investigation of the design and use of feed forward artificial neural networks in the classification of remotely sensed images**. University of Nottingham, 2001.

KAVZOGLU, T. Increasing the accuracy of neural network classification using refined training data. **Environmental Modelling & Software**, v. 24, n. 7, p. 850-858, 2009.

KAVZOGLU, T.; MATHER, P. The use of backpropagating artificial neural networks in land cover classification. **International Journal of Remote Sensing**, v. 24, n. 23, p. 4907-4938, 2003.

KIMBENG, C. A.; COX, M. C. Early generation selection of sugarcane families and clones in Australia: a review. **Journal American Society of Sugarcane Technolo gists**, v. 23, 2003.

KUMAR, M.; RAGHUWANSHI, N.; SINGH, R.; WALLENDER, W.; PRUITT, W. Estimating evapotranspiration using artificial neural network. **Journal of Irrigation and Drainage Engineering**, v. 128, n. 4, p. 224-233, 2002.

LI, H.; LINDGREN, D. Comparison of phenotype and combined index selection at optimal breeding population size considering gain and gene diversity. **Silvae Genetica**, v. 55, n. 1, p. 13-18, 2006.

MARTINS, I. S.; CRUZ, C. D.; ROCHA, M. D. G. B.; REGAZZI, A. J.; PIRES, I. E. Comparação entre os processos de seleção entre e dentro eo de seleção combinada, em progćnies de *Eucalytpus grandis*. **Cerne,** Lavras, v. 11, n. 1, p. 16-24, 2005.

MAS, J.; FLORES, J. The application of artificial neural networks to the analysis of remotely sensed data. **International Journal of Remote Sensing**, v. 29, n. 3, p. 617-663, 2008.

MATHER, P.; KOCH, M. **Computer processing of remotely-sensed images: an introduction**. Wiley, 2011.

MEUWISSEN, R.; LINN, S. C.; VALK, M. V. D.; MOOI, W. J.; BERNS, A. Mouse model for lung tumorigenesis through Cre/lox controlled sporadic activation of the K-Ras oncogene. **Oncogene**, v. 20, n. 45, p. 6551, 2001.

MUGNAI, S.; PANDOLFI, C.; AZZARELLO, E.; MASI, E.; MANCUSO, S. *Camellia japonica* L. genotypes identified by an artificial neural network based on phyllometric and fractal parameters. **Plant systematics and evolution**, v. 270, n. 1, p. 95-108, 2008.

ODA, S.; MELLO, E. J.; SILVA, J. F.; SOUZA, I. C. G. Melhoramento florestal. In: BORÉM, A. (Ed.). **Biotecnologia Florestal**. Viçosa: UFV, 2007. p. 51-71.

OLIVEIRA, R. A.; DAROS, E.; RESENDE, M. D. V.; BESPALHOK-FILHO, J. C.; ZAMBON, J. L. C.; SOUZA, T. R.; LUCIUS, A. S. F. Procedimento Blupis e seleção massal em cana-de-açúcar. **Bragantia**, v. 70, n. 4, p. 796-800, 2011.

PAULA, R. C.; PIRES, I. E.; BORGES, R. D. C. G.; CRUZ, C. D. Predição de ganhos genéticos em melhoramento florestal. **Pesq. agropec. bras., Brasília**, v. 37, n. 2, p. 159-165, 2002.

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**, v. 67, n. 1, p. 133-140, 2008.

PIEPHO, H.; MÖHRING, J.; MELCHINGER, A.; BÜCHSE, A. BLUP for phenotypic selection in plant breeding and variety testing. **Euphytica**, v. 161, n. 1, p. 209-228, 2008.

RESENDE, M. D. V. **Genética biométrica e estatística no melhoramento de plantas perenes**. Embrapa Informação Tecnológica, Colombo: Embrapa Florestas, 2002.

ROCHA, R. B.; ROCHA, M. G. B.; SANTANA, R. C.; VIEIRA, A. H. Estimação de parâmetros genéticos e seleção de procedências e famílias de *Dipteryx alata* Vogel (baru) utilizando metodologia de REML/BLUP e E (QM). **Cerne**, n. 3, p. 331-338, 2009.

ROSADO, A. M.; ROSADO, T. B.; RESENDE JÚNIOR, M. F. R.; BHERING, L. L.; CRUZ, C. D. Ganhos genéticos preditos por diferentes métodos de seleção em progênies de *Eucalyptus urophylla*. **Pesqui. Agropecu. Bras**, v. 44, p. 1653-1659, 2009.

ROVARIS, S. R. S.; ARAÚJO, P. M.; GARBUGLIO, D. D.; PRETE, C. E. C.; ZAGO, V. S.; SILVA, L. J. F. Estimates of genetic parameter in maize commercial variety IPR 114 at Paraná State, Brazil. **Acta Scientiarum. Agronomy**, v. 33, n. 4, p. 621-625, 2011.

SOUZA JR, C. L.; BARRIOS, S. C. L.; MORO, G. V. Performance of maize single-crosses developed from populations improved by a modified reciprocal recurrent selection. **Scientia Agricola**, v. 67, n. 2, p. 198-205, 2010.

VENTURA, R.; SILVA, M.; MEDEIROS, T.; DIONELLO, N.; MADALENA, F.; FRIDRICH, A.; VALENTE, B.; SANTOS, G.; FREITAS, L.; WENCESLAU, R. Use of artificial neural networks in breeding values prediction for weight at 205 days in Tabapuã beef cattle. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 64, n. 2, p. 411-418, 2012.

WANG, Z.; CONG, P.; ZHOU, J.; ZHU, Z. Method for identification of external quality of wheat grain based on image processing and artificial neural network [J]. **Transactions of the Chinese Society of Agricultural Engineering**, v. 1, p. 029, 2007.

5. CONCLUSÕES GERAIS

A obtenção de dados experimentais preservando propriedades pontuais, tais como média, herdabilidade e coeficiente de variação, foi eficiente e pode ser realizada usando princípios estocásticos de distribuição tais como enunciado no teorema de Box-Muller.

A preservação da matriz de variâncias e covariâncias de dados experimentais pode ser eficientemente realizada por meio do uso da decomposição espectral da matriz de variância e covariância original.

Conjuntos de dados preservados ou ampliados podem ser apropriadamente gerados com potencial de uso diverso, em especial em redes neurais que demanda grande quantidade de informações para fins de treinamento e aprendizagem. Os dados apropriadamente simulados, por preservarem informações essenciais, podem agregar ou substituir dados históricos algumas vezes não tão facilmente disponíveis.

As redes neurais artificiais foram eficientes para predizer o valor genético em experimento balanceados em blocos ao acaso para características quantitativas (muito influenciadas pelo ambiente).

A estrutura da rede com três camadas intermediárias com 10 neurônios na primeira camada, 20 na segunda camada e 8 neurônios na terceira camada foram eficientes na predição do valor genético.

ANEXO A. Script utilizado na Rede Neural Artificial

```
%% Inicializando
% ******************************************************************
%          PREDIÇÃO DE VALOR GENÉTICO - DBC
% ******************************************************************
clear,  close all, clc;
warning off;
%% Definindo constantes
numEntfinalTr =6;
numSaidas = 1;
minimosmaximos = [ zeros(numEntfinalTr,1) ones(numEntfinalTr,1) ];
epocas=500;
%% Listando arquivos
aqTreinamento = 'C:\dissertação\treincar2pop150h10cv10';
arquivo{1}='C:\dissertação\valcar2pop150h10cv10 1.dat';
arquivo{2}='C:\dissertação\valcar2pop150h10cv10 2.dat';
arquivo{3}='C:\dissertação\valcar2pop150h10cv10 3.dat';
arquivo{4}='C:\dissertação\valcar2pop150h10cv10 4.dat';
arquivo{5}='C:\dissertação\valcar2pop150h10cv10 5.dat';
arquivo{6}='C:\dissertação\valcar2pop150h10cv10 6.dat';
arquivo{7}='C:\dissertação\valcar2pop150h10cv10 7.dat';
arquivo{8}='C:\dissertação\valcar2pop150h10cv10 8.dat';
arquivo{9}='C:\dissertação\valcar2pop150h10cv10 9.dat';
arquivo{10}='C:\dissertação\valcar2pop150h10cv10 10.dat';
arquivo{11}='C:\dissertação\valcar2pop150h10cv10 11.dat';
arquivo{12}='C:\dissertação\valcar2pop150h10cv10 12.dat';
arquivo{13}='C:\dissertação\valcar2pop150h10cv10 13.dat';
arquivo{14}='C:\dissertação\valcar2pop150h10cv10 14.dat';
arquivo{15}='C:\dissertação\valcar2pop150h10cv10 15.dat';
arquivo{16}='C:\dissertação\valcar2pop150h10cv10 16.dat';
arquivo{17}='C:\dissertação\valcar2pop150h10cv10 17.dat';
arquivo{18}='C:\dissertação\valcar2pop150h10cv10 18.dat';
arquivo{19}='C:\dissertação\valcar2pop150h10cv10 19.dat';
arquivo{20}='C:\dissertação\valcar2pop150h10cv10 20.dat';
arquivo{21}='C:\dissertação\valcar2pop150h10cv10 21.dat';
arquivo{22}='C:\dissertação\valcar2pop150h10cv10 22.dat';
arquivo{23}='C:\dissertação\valcar2pop150h10cv10 23.dat';
arquivo{24}='C:\dissertação\valcar2pop150h10cv10 24.dat';
arquivo{25}='C:\dissertação\valcar2pop150h10cv10 25.dat';
arquivo{26}='C:\dissertação\valcar2pop150h10cv10 26.dat';
arquivo{27}='C:\dissertação\valcar2pop150h10cv10 27.dat';
arquivo{28}='C:\dissertação\valcar2pop150h10cv10 28.dat';
arquivo{29}='C:\dissertação\valcar2pop150h10cv10 29.dat';
arquivo{30}='C:\dissertação\valcar2pop150h10cv10 30.dat';
arquivo{31}='C:\dissertação\valcar2pop150h10cv10 31.dat';
arquivo{32}='C:\dissertação\valcar2pop150h10cv10 32.dat';
arquivo{33}='C:\dissertação\valcar2pop150h10cv10 33.dat';
arquivo{34}='C:\dissertação\valcar2pop150h10cv10 34.dat';
arquivo{35}='C:\dissertação\valcar2pop150h10cv10 35.dat';
arquivo{36}='C:\dissertação\valcar2pop150h10cv10 36.dat';
arquivo{37}='C:\dissertação\valcar2pop150h10cv10 37.dat';
arquivo{38}='C:\dissertação\valcar2pop150h10cv10 38.dat';
arquivo{39}='C:\dissertação\valcar2pop150h10cv10 39.dat';
arquivo{40}='C:\dissertação\valcar2pop150h10cv10 40.dat';
arquivo{41}='C:\dissertação\valcar2pop150h10cv10 41.dat';
arquivo{42}='C:\dissertação\valcar2pop150h10cv10 42.dat';
arquivo{43}='C:\dissertação\valcar2pop150h10cv10 43.dat';
arquivo{44}='C:\dissertação\valcar2pop150h10cv10 44.dat';
arquivo{45}='C:\dissertação\valcar2pop150h10cv10 45.dat';
```

```
arquivo{46}='C:\dissertação\valcar2pop150h10cv10 46.dat';
arquivo{47}='C:\dissertação\valcar2pop150h10cv10 47.dat';
arquivo{48}='C:\dissertação\valcar2pop150h10cv10 48.dat';
arquivo{49}='C:\dissertação\valcar2pop150h10cv10 49.dat';
arquivo{50}='C:\dissertação\valcar2pop150h10cv10 50.dat';
arquivo{51}='C:\dissertação\valcar2pop150h10cv10 51.dat';
arquivo{52}='C:\dissertação\valcar2pop150h10cv10 52.dat';
arquivo{53}='C:\dissertação\valcar2pop150h10cv10 53.dat';
arquivo{54}='C:\dissertação\valcar2pop150h10cv10 54.dat';
arquivo{55}='C:\dissertação\valcar2pop150h10cv10 55.dat';
arquivo{56}='C:\dissertação\valcar2pop150h10cv10 56.dat';
arquivo{57}='C:\dissertação\valcar2pop150h10cv10 57.dat';
arquivo{58}='C:\dissertação\valcar2pop150h10cv10 58.dat';
arquivo{59}='C:\dissertação\valcar2pop150h10cv10 59.dat';
arquivo{60}='C:\dissertação\valcar2pop150h10cv10 60.dat';
arquivo{61}='C:\dissertação\valcar2pop150h10cv10 61.dat';
arquivo{62}='C:\dissertação\valcar2pop150h10cv10 62.dat';
arquivo{63}='C:\dissertação\valcar2pop150h10cv10 63.dat';
arquivo{64}='C:\dissertação\valcar2pop150h10cv10 64.dat';
arquivo{65}='C:\dissertação\valcar2pop150h10cv10 65.dat';
arquivo{66}='C:\dissertação\valcar2pop150h10cv10 66.dat';
arquivo{67}='C:\dissertação\valcar2pop150h10cv10 67.dat';
arquivo{68}='C:\dissertação\valcar2pop150h10cv10 68.dat';
arquivo{69}='C:\dissertação\valcar2pop150h10cv10 69.dat';
arquivo{70}='C:\dissertação\valcar2pop150h10cv10 70.dat';
arquivo{71}='C:\dissertação\valcar2pop150h10cv10 71.dat';
arquivo{72}='C:\dissertação\valcar2pop150h10cv10 72.dat';
arquivo{73}='C:\dissertação\valcar2pop150h10cv10 73.dat';
arquivo{74}='C:\dissertação\valcar2pop150h10cv10 74.dat';
arquivo{75}='C:\dissertação\valcar2pop150h10cv10 75.dat';
arquivo{76}='C:\dissertação\valcar2pop150h10cv10 76.dat';
arquivo{77}='C:\dissertação\valcar2pop150h10cv10 77.dat';
arquivo{78}='C:\dissertação\valcar2pop150h10cv10 78.dat';
arquivo{79}='C:\dissertação\valcar2pop150h10cv10 79.dat';
arquivo{80}='C:\dissertação\valcar2pop150h10cv10 80.dat';
arquivo{81}='C:\dissertação\valcar2pop150h10cv10 81.dat';
arquivo{82}='C:\dissertação\valcar2pop150h10cv10 82.dat';
arquivo{83}='C:\dissertação\valcar2pop150h10cv10 83.dat';
arquivo{84}='C:\dissertação\valcar2pop150h10cv10 84.dat';
arquivo{85}='C:\dissertação\valcar2pop150h10cv10 85.dat';
arquivo{86}='C:\dissertação\valcar2pop150h10cv10 86.dat';
arquivo{87}='C:\dissertação\valcar2pop150h10cv10 87.dat';
arquivo{88}='C:\dissertação\valcar2pop150h10cv10 88.dat';
arquivo{89}='C:\dissertação\valcar2pop150h10cv10 89.dat';
arquivo{90}='C:\dissertação\valcar2pop150h10cv10 90.dat';
arquivo{91}='C:\dissertação\valcar2pop150h10cv10 91.dat';
arquivo{92}='C:\dissertação\valcar2pop150h10cv10 92.dat';
arquivo{93}='C:\dissertação\valcar2pop150h10cv10 93.dat';
arquivo{94}='C:\dissertação\valcar2pop150h10cv10 94.dat';
arquivo{95}='C:\dissertação\valcar2pop150h10cv10 95.dat';
arquivo{96}='C:\dissertação\valcar2pop150h10cv10 96.dat';
arquivo{97}='C:\dissertação\valcar2pop150h10cv10 97.dat';
arquivo{98}='C:\dissertação\valcar2pop150h10cv10 98.dat';
arquivo{99}='C:\dissertação\valcar2pop150h10cv10 99.dat';
arquivo{100}='C:\dissertação\valcar2pop150h10cv10 100.dat';
 %arquivo{11}='c:\dados\rede1.dat';

%% Preparação dos dados de treinamento e validação
dadosTreinamento = load(aqTreinamento);
[numEntTr numTr] = size(dadosTreinamento);
entradasTreinamento  = dadosTreinamento(1:numEntTr-1, :);
saidasTreinamento  = dadosTreinamento(numEntTr,:);
```

```
%entradasTreinamentoNormalizado= Normatiza(entradasTreinamento);
%saidasTreinamentoNormalizado= Normatiza(saidasTreinamento);
entradasTreinamentoNormalizado= entradasTreinamento;
saidasTreinamentoNormalizado= saidasTreinamento;


%% Laço principal para buscar a rede.
funcoes = ['tansig ' ; 'logsig '; 'purelin'];
contador = 1;
a0=8; a1=-2; a2=1;
b0=20; b1=-2;b2=1;
c0=8; c1=-2; c2=1;
d1=3;e1=3;f1=3;
steps = a1*b1*c1*d1*e1*f1;
acmmax =0;
acertomax=0;
rtreinamento =0;

for a = a0:a1:a2 % Neuronios Camada 1
   for b =b0:b1:b2 % Neuronios Camada 2
       for c =c0:c1:c2 % Neuronios Camada 2
         for d = 1:d1 % Função de Ativação Camada 1
            for e = 1:e1 % Função de Ativação Camada 2
             for f = 1:f1 % Função de Ativação Camada 2
                   %  traingdm, traingda, traingdx, trainlm,trainrp, traincgf, traincgb, trainscg,
traincgp,trainbfg.


                  net = newff(minimosmaximos, [ a b  c 1 ], { char(cellstr(funcoes(d,:)))
char(cellstr(funcoes(e,:))) char(cellstr(funcoes(f,:))) 'purelin' }, 'trainbr');
                  net = init(net);
                  % CONFIGURANDO OS PARAMETROS DA REDE:
                  net.trainParam.epochs = epocas; % NUMERO MAXIMO DE EPOCAS
                  net.trainParam.goal = 0.0001; % CONDIÇAO DE PARADA POR ERRO
                  net.trainParam.show = NaN; % INTERVALO PARA EXIBIÇAO NA TELA

                  % ********************** aleatorização do conjunto de dados ************************
                  embaralha = randperm(length(dadosTreinamento));

                  for k=1:length(dadosTreinamento)
                     entradaEmbaralhada(:,k) = entradasTreinamentoNormalizado(:,embaralha(k));
                     saidaEmbaralhada(k) = saidasTreinamento(embaralha(k));
                  end

                  %* *********** rede sem early stopping ***********
                  net= train(net,entradaEmbaralhada,saidaEmbaralhada);
                  % ************************************************

                     [numLin numTr] = size(saidasTreinamentoNormalizado);
                     eixoX = 1:numTr;
                     eixoYorigt = saidasTreinamentoNormalizado;% valor genético
                     simulacaoTreinamento = sim(net,entradasTreinamentoNormalizado);% valor da rede
                     eixoYnett = simulacaoTreinamento;
                     eixoYmediat = entradasTreinamentoNormalizado;
                     matriz = [eixoYorigt'  eixoYnett'  eixoYmediat'];
                     [rt,p] = corrcoef(matriz);
figure(1);
subplot(3,1,1,'replace')
hold on
xlabel('Indivíduos')
ylabel('Valor Genotípico')
plot(eixoX,eixoYorigt,'g.');
```

```
hold off

subplot(3,1,2,'replace')
hold on
xlabel('Indivíduos')
ylabel('Valor de Rede')
plot(eixoX,eixoYnett,'r.');
hold off

subplot(3,1,3,'replace')
hold on
xlabel('Indivíduos')
ylabel('Valor Fenotípico')
plot(eixoX,eixoYmediat,'b.');
hold off
                    if rt(1,2) > rtreinamento
                       rtreinamento =rt(1,2);
                    end


                % verificando desempenho da rede
                for kkk=1:length(arquivo)
                    aqValidacao= arquivo(1,kkk);
                    aqValidacao = cat(2,aqValidacao{:});



                    dadosValidacao = load(aqValidacao);
                    [numEntV numVal] = size(dadosValidacao);

                    entradasValidacao  = [dadosValidacao(1:numEntV-1, :)];
                    saidasValidacao  = dadosValidacao(numEntV,:);
                    %entradasValidacaoNormalizado  = Normatiza(entradasValidacao);
                    %saidasValidacaoNormalizado = Normatiza(saidasValidacao);
                    entradasValidacaoNormalizado  = entradasValidacao;
                    saidasValidacaoNormalizado = saidasValidacao;

                    eixoX = 1:numVal;
                    eixoYorig = saidasValidacaoNormalizado;% valor genético
                    simulacaoValidacao = sim(net,entradasValidacaoNormalizado);
                    eixoYnet = simulacaoValidacao;%valor da rede
                    eixoYmedia =  mean(dadosValidacao(1:numEntV-1, :));% valor fenotípico

                    matriz = [eixoYorig'  eixoYnet'  eixoYmedia'];% [VV  VR  VM]
                    [r,p] = corrcoef(matriz);
                    acuracia(kkk,1) =r(1,2)^2;% correlação do valor genético e valor de rede
                    acuracia(kkk,2) =r(1,3)^2;% correlação do valor genético e valor fenotípico
                    acuracia(kkk,3) =r(2,3)^2;% correlação do valor rede e valor fenotípico

                    %perfrede = mse(eixoYorig, eixoYnet)
                    %perfmedia = mse(eixoYorig, eixoYmedia)


 %figure(kkk+1);
%subplot(3,1,1,'replace')
%hold on
%xlabel('Indivíduos')
%ylabel('Valor Genotípico')
%plot(eixoX,eixoYorig,'g.');
%hold off
```

```
%subplot(3,1,2,'replace')
%hold on
%xlabel('Indivíduos')
%ylabel('Valor de Rede')
%plot(eixoX,eixoYnet,'r.');
%hold off

%subplot(3,1,3,'replace')
%hold on
%xlabel('Indivíduos')
%ylabel('Valor Fenotípico')
%plot(eixoX,eixoYmedia,'b.');
%hold off
```

```
end
acm = mean(acuracia);
delta = hardlim(acuracia(:,1)- acuracia(:,2));
acerto = sum(delta)/10;

if acm(1,1) > acmmax
%if acerto > acertomax

    acmmax = acm(1,1);
    acertomax= acerto

   [numLin numTr] = size(saidasTreinamentoNormalizado);
   eixoX = 1:numTr;
   eixoYorig = saidasTreinamentoNormalizado;
   simulacaoTreinamento = sim(net,entradasTreinamentoNormalizado);
   eixoYnet = simulacaoTreinamento;
   eixoYmedia = entradasTreinamentoNormalizado;

   matriz = [eixoYorig'  eixoYnet'  eixoYmedia'];
   [r,p] = corrcoef(matriz);

   fprintf('----------------------------  \n')
   fprintf('Treinamento:  %s\n',aqTreinamento)
   fprintf('----------------------------  \n')
   fprintf('Correlação VGenético x VRede:       %12.8f\n',r(1,2))
   fprintf('Correlação VGenético x VFenotípico:  %12.8f\n',r(1,3))
   fprintf('R2 VGenético x VRede:       %12.8f\n',r(1,2)^2)
   fprintf('R2 VGenético x VFenotípico:  %12.8f\n',r(1,3)^2)

   fprintf('Correlação VGenético x VRede máxima:       %12.8f\n',rtreinamento)
   fprintf('R2 VGenético x VRede máxima:       %12.8f\n',rtreinamento^2)

   texto = [ 'Neurônios 1 2 3 - Função de ativação 1 2 3 '];
   disp(texto);

    resp= 100000*a + 10000*b+ 1000*c+100*d+10*e+f;
    fprintf('Rede Ideal: %6.0f\n',resp)
    beep;
    disp('GxRede    GxFen    RedexFen     Ganho'  )
    disp('-----------------------------------')
   resposta = [acuracia delta]
   disp('-----------------------------------')
    mean(resposta)
   %acm
```

```
                    disp('-----------------------------------')



                    figure(2);

                    bar(acuracia(:,1:2));
                    texto = [ num2str((contador/steps)*100) '%' ];
                    title(texto)
                    legend('RNA','Média')

                end
                 contador = contador+1;

            end
        end
        end
    end
  end
end
```