UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

REDES NEURAIS ARTIFICIAIS APLICADAS NA PREDIÇÃO DE VALORES GENÉTICOS E CLASSIFICAÇÃO DE GENÓTIPOS DE SOJA DE DIFERENTES GRUPOS DE MATURIDADE RELATIVA

Lígia de Oliveira Amaral

Engenheira Agrônoma

2021

UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

REDES NEURAIS ARTIFICIAIS APLICADAS NA PREDIÇÃO DE VALORES GENÉTICOS E CLASSIFICAÇÃO DE GENÓTIPOS DE SOJA DE DIFERENTES GRUPOS DE MATURIDADE RELATIVA

Discente: Lígia de Oliveira Amaral

Orientadora: Profa. Dra. Sandra Helena Unêda-Trevisoli

Tese apresentada à Faculdade de Ciências Agrárias e Veterinárias – Unesp, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Doutora em Agronomia (Genética e Melhoramento de Plantas).

2021

A485r

Amaral, Lígia de Oliveira
Redes neurais artificiais aplicadas na predição de valores genéticos e classificação de genótipos de soja de diferentes grupos de maturidade relativa / Lígia de Oliveira Amaral. -- Jaboticabal, 2021
74 f. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal
Orientadora: Sandra Helena Unêda-Trevisoli

1. Genética vegetal. 2. Soja. 3. Redes neurais (Computação). 4. Análise multivariada. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: REDES NEURAIS ARTIFICIAIS APLICADAS A PREDIÇÃO DE VALORES GENÉTICOS E CLASSIFICAÇÃO DE GENÓTIPOS DE SOJA DE DIFERENTES GRUPOS DE MATURIDADE RELATIVA

AUTORA: LÍGIA DE OLIVEIRA AMARAL
ORIENTADORA: SANDRA HELENA UNÊDA TREVISOLI

Aprovada como parte das exigências para obtenção do Título de Doutora em AGRONOMIA (GENÉTICA E MELHORAMENTO DE PLANTAS), pela Comissão Examinadora:

Prof. Dr. SANDRA HELENA UNÊDA TREVISOLI (Participaçao Virtual)
Departamento de Ciências da Produção Agrícola / FCAV / UNESP - Jaboticabal

Prof. Dr. GLAUCO VIEIRA MIRANDA (Participaçao Virtual)
Universidade Tecnológica Federal do Paraná - Campus Santa Helena/UTFPR / Santa Helena/PR

Prof. Dr. RINALDO CESAR DE PAULA (Participaçao Virtual)
Departamento de Ciências da Produção Agrícola (Produção Vegetal) / FCAV / UNESP - Jaboticabal

Pesquisadora Dra. IVANA MARINO BÁRBARO TORNELI (Participaçao Virtual)
Pólo Regional Alta Mogiana-APTA / Colina/SP

Prof. Dr. ALAN RODRIGO PANOSSO (Participaçao Virtual)
Departamento de Engenharia e Ciências Exatas (DECEx) / FCAV / UNESP - Jaboticabal

Jaboticabal, 14 de junho de 2021

DADOS CURRICULARES DA AUTORA

LÍGIA DE OLIVEIRA AMARAL – nascida em 17 de novembro de 1989, na cidade de Lavras – MG, é Engenheira Agrônoma formada pela Universidade Federal de Lavras em 2014. Durante a graduação foi bolsista de iniciação científica CNPq (2011/2012) no Departamento de Engenharia – Agroecologia e Permacultura. Em março de 2015 ingressou no curso de Mestrado em Agronomia/Fitotecnia pela mesma instituição, área de concentração em Produção Vegetal com ênfase em Produção e Melhoramento Genético de soja, como bolsista CAPES sob a orientação do Prof. Dr. Adriano Teodoro Bruzi, obtendo o título de mestre em fevereiro de 2017. Ingressou no curso de Doutorado em Agronomia (Genética e Melhoramento de Plantas) em agosto de 2017 na Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências Agrárias e Veterinárias, Câmpus de Jaboticabal, sob a orientação da Profa. Dra. Sandra Helena Unêda-Trevisoli, como bolsista CAPES.

“Tenho a impressão de ter sido uma criança brincando a beira-mar, divertindo-me em descobrir uma pedrinha mais lisa ou uma concha mais bonita que as outras, enquanto o imenso oceano da verdade continua misterioso diante de meus olhos”.

(Isaac Newton)

Dedico,

Aos meus pais, Maciel (*in memoriam*) e Máguida, pelo apoio em todos os momentos, pelos ensinamentos e amor incondicional.

Aos meus irmãos Luiz Henrique, Lucio (*in memoriam*), Leandro e Lidiana pelo carinho e companheirismo.

Aos meus sobrinhos Yan e Maria Cecília pela alegria e esperança.

A toda minha família que é porto seguro e fonte de amor.

AGRADECIMENTOS

A Deus pela vida, pelos dons e por dar sentido a tudo que faço.

A minha família por me fortalecer e tornar possível todas as etapas.

À Universidade Estadual Paulista “Júlio de Mesquita Filho”, Faculdade de Ciências Agrárias e Veterinárias Câmpus de Jaboticabal e ao Programa de Pós-Graduação em Agronomia/Genética e Melhoramento de Plantas pelo aprendizado, experiências e toda estrutura e recursos disponibilizados para minha formação e crescimento.

À Profa. Dra. Sandra Helena Unêda-Trevisoli pela orientação, ensinamentos, direcionamentos e suporte no desenvolvimento e conclusão deste curso.

Ao Prof. Dr. Glauco Miranda por todo suporte e auxílio.

Ao Gustavo Fieno pelo amor, cuidado, carinho e presença em tudo e sempre.

Aos queridos Ari, Benícia, Gabriela e Lorenzo Fieno por terem me acolhido e sido minha família em Jaboticabal.

Aos amigos Daniela Konrad, Júlia Alexandrino, Maria Heloisa Julião, Antonio Pizolato Neto, Lucas Munaro, Sophia Mangussi, Gabriela Crivelenti, Maria Eugênia Ferraz, Alice Silva, Gabriel Salgado, Alyce Moitinho e Ana Paula Lira pela companhia, carinho, amizade e diversão.

Aos colegas do LBMP Paloma Libório, Guilherme Bevilacqua, Eduardo Bizari, Bruno Val, Hortência Kardec, Dardânia Cristeli, Saulo Dantas e Thayná Garcia pelo trabalho em equipe.

Às colegas do GEMP Edicleide Macedo, Naiara Zancanari, Amanda Baldassi e Samanta Carvalho pela confiança, sintonia, trabalho em equipe e aprendizado.

Aos membros da banca do exame geral de qualificação, Prof. Dr. Gustavo Môro e Prof. Dr. João Andrade, pela atenção e colaboração neste trabalho.

Aos funcionários da Fazenda de Ensino e Pesquisa da UNESP/FCAV, de maneira especial ao Marcelo (*in memoriam*) e Sr. João pela paciência e auxílio nos experimentos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

A todos que contribuíram de alguma forma para a conclusão deste curso.

De todo meu coração, MUITO OBRIGADA!

SUMÁRIO

Página

REDES NEURAIS ARTIFICIAIS APLICADAS NA PREDIÇÃO DE VALORES GENÉTICOS E CLASSIFICAÇÃO DE GENÓTIPOS DE SOJA DE DIFERENTES GRUPOS DE MATURIDADE RELATIVA

RESUMO – A soja [*Glycine max* (L.) Merrill] é uma espécie muito influenciada pelo fotoperíodo. Cultivares têm seu potencial produtivo maximizado quando cultivadas na faixa ótima de fotoperíodo. O contrário acontece quando os genótipos são submetidos a condições ambientais diferentes das ideais. A variabilidade genética gerada em uma população oriunda de cruzamentos divergentes deve ser estudada para que os genótipos sejam avaliados e selecionados em seus grupos de maturidade relativa corretos. Além da classificação e direcionamento dos genótipos, as metodologias eficientes de predição de valores genéticos podem promover o sucesso dos programas de melhoramento por auxiliarem no avanço de genótipos promissores para as características de interesse. Desta forma, os objetivos deste trabalho foram i) determinar a eficiência das redes neurais artificiais (RNA) em discriminar e classificar genótipos de soja pertencentes a diferentes grupos de maturidade relativa, ii) determinar a eficiência das RNAs em predizer valores genéticos relacionados à produtividade e à maturidade relativa de genótipos de soja pertencentes a populações de cruzamento amplo e restrito para grupo de maturidade relativa (GMR). Foram utilizados dados de três populações de soja coletados nos anos agrícolas 2017/2018, 2018/2019 e 2019/2020. Os dados foram submetidos às análises discriminantes linear de Fisher e quadrática de Anderson e tiveram seus resultados comparados aos obtidos por um modelo de RNA – MLP (*Multi Layer Perceptron*). Os dados de três populações, uma oriunda de cruzamento amplo e duas de cruzamentos restritos para GMR foram submetidos também à análise por modelos mistos para estimativa dos componentes de variância e valores genéticos. Posteriormente foi realizada a ordenação dos genótipos de acordo com o desempenho e, esta ordenação foi comparada àquela obtida pelas RNAs. Foram obtidas as porcentagens de coincidência na classificação dos melhores genótipos e o ganho esperado com a seleção (GS) para cada metodologia. As RNAs mostraram-se eficientes em discriminar e classificar corretamente os genótipos em suas populações quando comparadas às análises discriminantes que apresentaram taxa de erro aparente (TEA) acima de 50%. A mesma metodologia se mostrou eficiente na predição de valores genéticos apresentando ganhos com a seleção de até 11.91% para produção de grãos e -5.42% para ciclo total da cultura.

Palavras-chave: aprendizado de máquinas, *Glycine max*, maturidade relativa, mineração de dados, modelos mistos, REML/BLUP

ARTIFICIAL NEURAL NETWORKS APPLIED TO PREDICTION OF GENETICS VALUES AND CLASSIFICATION OF SOYBEAN GENOTYPES FROM DIFFERENT GROUPS OF RELATIVE MATURITY

ABSTRACT – The soybean [*Glycine max* (L.) Merril] is a species highly influenced by the photoperiod. Cultivars have their productive potential maximized when grown in the optimal photoperiod range. The opposite happens when genotypes are considered different from ideal environmental conditions. In addition to the classification and targeting of genotypes, efficient methodologies for predicting genetic values can promote the success of breeding programs to assist in the advancement of promising genotypes for the traits of interest. Thus, the main objectives were i) to determine the efficiency of artificial neural networks (ANN) in discriminating and classifying soybean genotypes belonging to different groups of relative maturity, ii) to determine the efficiency of ANNs in predicting genetic values related to grain yield and to relative maturity of genotypes belonging to populations of broad and narrow crossing for relative maturity group (RMG). Data from three soybean collected in the agricultural years 2017/2018, 2018/2019 and 2019/2020 were used. The data were submitted to multivariate analysis of principal components (APC), Fisher 's linear discriminant and Anderson' s quadratic and their results were compared to those obtained by an ANN - MLP (Multi Layer Perceptron) model. Data from three populations, one from broad cross and two from narrow crossings for RMG were also subjected to analysis by mixed models to estimate the components of variance and genetic values. Subsequently, the genotypes were ordered according to performance, and this order was compared to that obtained by the ANNs. The percentages of coincidence were obtained in the classification of the best genotypes and the expected gain with the selection (GS) for each methodology. The ANNs proved to be efficient in discriminating and correctly classifying the genotypes in their populations when compared to multivariate analyzes that showed an apparent error rate (AER) above 50%. The same methodology proved to be efficient in predicting genetic values, showing gains with the selection of up to 11.91% for grain yield and -5.42% for the total crop cycle.

Keywords: *Glycine max*, machine learning, maturity relative, mixed models, REML/BLUP

CAPÍTULO 1 – Considerações gerais

1 INTRODUÇÃO

A soja [*Glycine max* (L.) Merril], se destaca como uma das culturas de maior importância econômica mundial. O grão é utilizado como matéria prima de diversos produtos, sendo o óleo e aqueles destinados à alimentação animal, a exemplo do farelo, seus principais subprodutos. Na safra 2019/2020, o Brasil produziu cerca de 124 milhões de toneladas de soja tornando-se o maior produtor mundial do grão (CONAB, 2021). De acordo com o levantamento de maio de 2021 da CONAB, a área plantada nesta safra foi de cerca de 38,5 milhões de hectares e a produtividade foi de 3,517 kg ha^{-1}. O Mato Grosso se manteve como maior produtor brasileiro, responsável pela produção de 76,1 milhões de toneladas. A exportação da soja em grãos somada ao farelo e ao óleo movimentou um total de 32,6 bilhões de dólares em 2019 (Agrostat, 2020).

Quando foi introduzida oficialmente no Brasil, esta espécie era cultivada apenas na região Sul devido à semelhança climática com a região de origem das cultivares introduzidas, o sul dos Estados Unidos (Sediyama, 2009). A partir dos anos 1970, foi possível expandir a produção de soja para outras regiões brasileiras graças aos avanços nas áreas de produção vegetal e melhoramento genético. A obtenção de cultivares adaptadas a diferentes faixas de fotoperíodo (grupos de maturidade relativa), resistência a pragas e doenças limitantes à produção de grãos, e o incremento constante em produtividade foram os principais responsáveis pelo estabelecimento da espécie no Brasil, além de garantir a disponibilidade de sementes de qualidade no mercado (Bacaxixi et al., 2011; Sediyama et al., 2015).

Sendo a cultura da soja tão influenciada pelo fotoperíodo, é imprescindível que o desenvolvimento de cultivares tenha como objetivo maximizar a capacidade produtiva destas a partir da classificação e direcionamento dos genótipos para sua faixa ideal de cultivo. As populações oriundas de cruzamentos contrastantes são compostas por genótipos com diferentes combinações de genes relacionados ao florescimento e maturação, o que pode acarretar em uma seleção tendenciosa, já que as seleções são baseadas em avaliações dos genótipos direcionadas para o

local onde se encontram. Os estudos de discriminação e classificação de populações têm se baseado em metodologias de estatística multivariada como as análises discriminantes de Fisher e Anderson, que consistem em classificar os genótipos ou grupos em populações de acordo com um conjunto de informações obtidas destes indivíduos (Mingoti, 2005; Cruz et al., 2011). No âmbito da predição de valores genéticos, a acurácia determina o sucesso ou não dos programas de melhoramento visto que, estas estimativas orientam os melhoristas na seleção e condução dos genótipos superiores (Peixoto, 2013). O advento da abordagem de modelos mistos trouxe para as análises estatísticas de experimentos agrícolas maior liberdade exploratória das informações contidas nos dados. A metodologia REML/BLUP reúne a estimativa dos componentes de variância REML (Restricted Maximum Likelihood – Máxima Verossimilhança Restrita) à estimativa dos valores genéticos BLUP (Best Linear Unbiased Prediction – Melhor Predição Linear não Viesada) e tem sido largamente utilizada, principalmente na existência de dados desbalanceados (Carneiro Júnior et al., 2010; Resende et al., 2018).

Entretanto, existem situações nas quais informações como médias e variâncias não são suficientes, exigindo a percepção de características mais complexas das populações para respostas mais assertivas a respeito das classificações e predições. Metodologias mais flexíveis, que permitem generalizações e que sejam capazes de extrair informações adicionais de dados incompletos por meio de aprendizado têm sido procuradas e têm conquistado cada vez mais espaço na ciência de dados (Cruz et al., 2011; Sant'anna, 2014). Dada a importância da discriminação e classificação de genótipos em seus grupos de maturidade ideais e da predição para os programas de melhoramento, as Redes Neurais Artificias (RNA) têm se destacado como uma metodologia eficiente a favor da pesquisa agrícola. As RNAs têm capacidade de aprendizado e generalização, são aplicáveis a problemas não linearmente separáveis e não exigem nenhum tipo de pressuposição dos dados (Cruz e Nascimento, 2018; Sant'anna, 2018). A criação do modelo *MultiLayer Perceptron – MLP* e do algoritmo *backpropagation* possibilitou sua aplicação nos problemas enfrentados frequentemente no melhoramento vegetal e desde então, esta metodologia tem sido aplicada às mais diversas áreas e espécies, contribuindo de forma expressiva em todas as etapas de desenvolvimento

de cultivares produtivas (Braga et al., 2007). Portanto, os objetivos deste trabalho consistiram em determinar a eficiência das redes neurais artificiais i) na discriminação e classificação de genótipos de soja pertencentes a diferentes grupos de maturidade relativa (GRM) e ii) na predição de valores genéticos relacionados à produtividade e à maturidade relativa de genótipos pertencentes a populações de cruzamento amplo e restrito para grupo de maturidade relativa (GMR).

2 REVISÃO BIBLIOGRÁFICA

2.1 Fotoperíodo e grupos de maturidade relativa em soja

Garner e Allard (1920) estudaram a influência do comprimento do dia nas plantas de soja e demonstraram como esta afeta o florescimento e maturidade da espécie, sendo a soja considerada uma planta de dias curtos. A soja é cultivada no mundo inteiro, em diferentes faixas de latitude. Embora cada cultivar seja restrita a uma faixa relativamente estreita de latitude, hoje se sabe que esta capacidade de se adaptar às diversas regiões produtoras se deve a variabilidade genética nos principais loci de genes e locos de características quantitativas (QTLs) envolvidos no controle da floração e maturidade. Até o momento foram identificados os principais loci genéticos E1 a E11 e J, e vários QTLs, como Tof11 / Gp11, Tof12 / Gp1 / qFT12-1 e qDTF-J. De modo geral, com exceção dos genes *E6*, *E9*, *E11* e *J*, o alelo dominante dos genes *E* confere florescimento e maturidade tardios, ao passo que o aumento no número de alelos recessivos acarreta em precocidade da variedade (Samanfar et al., 2017; Wang et al., 2020; Lin et al., 2021).

De acordo com esta influência do fotoperíodo nas plantas de soja, as cultivares são distribuídas entre 13 grupos de maturidade relativa (GMR) classificados geograficamente com base no crescimento e desenvolvimento das plantas. O Brasil compreende os GMRs de 5 a 9, respectivamente do sul (latitude 30°) ao norte do país (latitude 0°) (Alliprandini et al., 2009). A sensibilidade ao fotoperíodo é importante para a adaptação local e o cultivo de variedades adequadas possibilita o pleno uso da estação de crescimento na região alvo. Cultivares em seus locais ideais de produção tem seu potencial de rendimento maximizado. A estimativa correta dos estágios fenológicos da planta possibilita

maior flexibilidade para modificar o seu desenvolvimento como um todo e em trabalhar com características controladas por outros genes, pois além de influenciar o tempo para floração e maturação, a maioria dos genes de maturidade e QTLs afetam várias características agronômicas que são dependentes do desenvolvimento reprodutivo, como rendimento de grãos e qualidade da semente (Miladinović et al., 2018; Lin et al., 2021). Já o cultivo de uma variedade em um GMR diferente daquele do seu adequado pode acarretar em alongamento ou redução indesejados do ciclo, desenvolvimento vegetativo insuficiente ou exagerado, suscetibilidade a pragas e doenças específicas de determinada época do ano e baixa produtividade (Miladinović & Đorđević, 2011).

2.2 Análises discriminantes

Na estatística multivariada, o conjunto de dados coletados é analisado e interpretado levando-se em consideração todas as informações das variáveis e as correlações entre elas, procurando evidenciar suas ligações, semelhanças ou diferenças com menor perda de informações possível (Hair et al., 2009). A grande vantagem, portanto, é analisar simultaneamente todas as variáveis resposta considerando a correlação existente para extrair informações pertinentes dos dados (Sartorio, 2008).

A análise discriminante é uma técnica de análise multivariada que permite alocar um determinado indivíduo a uma população previamente conhecida por meio de uma combinação linear de características mensuráveis, com poder discriminatório entre populações. Foi inicialmente descrita por Fisher em 1936, em seu estudo sobre a discriminação e classificação de grupos de cultivares de Iris. Fisher propôs funções matemáticas capazes de classificar um indivíduo “x” em uma entre várias populações, com base em medidas de um número “p” de características, buscando minimizar a probabilidade de classificação errônea (Regazzi, 2000).

No estudo da diversidade genética, o propósito das análises discriminantes é alocar um conjunto de genótipos em suas respectivas populações previamente definidas (Souza, 2017). Os métodos baseados em funções discriminantes de Fisher e de Anderson são aqueles que têm sido usados comumente com caracteres quantitativos e requerem pressuposições e probabilidade específica de distribuição,

sendo a multinormalidade a mais comum para o estudo das populações (Cruz et al., 2011), portanto, são métodos paramétricos.

A função discriminante linear consiste, basicamente, em separar duas classes de objetos ou fixar um novo objeto em uma das duas classes (Johnson & Wichern, 2002). A Função Discriminante Linear de Fisher trata-se de uma combinação linear das características observadas que apresenta melhor poder de discriminação entre todas as combinações lineares das variáveis envolvidas. Esta função tem a propriedade de minimizar a probabilidade de má classificação, quando as populações apresentam média e variância conhecidas. Contudo, tal situação pode não ocorrer na prática, necessitando-se, portanto de estimativas e métodos de estimação dessas probabilidades ótimas (Cruz et al., 2012). A análise baseada na função discriminante linear de Fisher será tanto mais eficiente quanto maior for a porcentagem da variância total a ela atribuída.

As análises discriminantes quadráticas são aplicadas quando o pressuposto da igualdade das matrizes de covariância (∑) não é atendido, ou seja, as “p” matrizes de variância são heterogêneas. De acordo com Anderson (1958), quando se dispõe de várias populações e se deseja alocar um novo indivíduo a cada uma delas, são necessárias algumas condições para se obter as funções discriminantes: as populações devem apresentar algum tipo de distribuição; que exista uma probabilidade a priori para cada população; e um custo de classificação inadequada. Estas funções permitem também a classificação de novos genótipos, de comportamento desconhecido, nas populações já conhecidas. A eficácia das variáveis utilizadas em promover a discriminação também é avaliada, permitindo conhecer a adequação da função estimada (Varella, 2008).

2.3 Modelos mistos na predição de valores genéticos

A utilização dos modelos mistos têm três objetivos principais: a estimação e teste dos efeitos fixos, a estimação e teste dos efeitos aleatórios, e a estimação dos componentes de variância devido aos fatores aleatórios (López & Iemma, 1998). Nessa classe de modelos, os níveis de um fator aleatório estão relacionados entre si por uma população de referência, resultando em covariância entre as observações (Duarte, 2000). Dentre os métodos de estimativa dos componentes de variância,

aqueles baseados em máxima verossimilhança, com destaque para REML (*Restricted Maximum Likelihood*), têm sido considerados os melhores devido às boas propriedades estatísticas dos seus estimadores como: consistência, suficiência, eficiência, não negatividade, variâncias amostrais menores, etc (Verneque, 1994). O método REML foi adotado para modelos mistos desbalanceados por Patterson & Thompson (1971) e desdobra a verossimilhança em duas funções, uma destas é livre dos efeitos fixos.

Com o advento da abordagem por modelos mistos, os componentes de variância passaram a ter importância fundamental na predição de valores genotípicos. Predição esta aplicada às etapas de seleção dos programas de melhoramento e, consequentemente, no desenvolvimento de cultivares com alto desempenho produtivo (Resende, 2002; Piepho et al., 2008). O BLUP (*Best Linear Unbiased Predictor*) são os valores realizados dos efeitos aleatórios (Gonçalves e Fritsche Neto, 2012) e uma combinação linear das observações já ajustadas para os efeitos ambientais. Entre as propriedades dos BLUPs é importante ressaltar que é o preditor de máxima correlação com o verdadeiro vetor de valores genotípicos e, considerando normalidade do caráter e dos valores genotípicos, é o critério ótimo de seleção, pois ordena os genótipos de maneira a maximizar a média populacional na geração seguinte (Jiménez & Villa, 1995). Estudos sobre a eficiência dos BLUP's apontaram menor erro padrão e maior correlação dos valores preditos com o real desempenho dos genótipos, facilitando a identificação e seleção de linhagens superiores (Panter & Allen, 1995).

2.4 Redes Neurais Artificiais (RNA)

2.4.1 Conceitos e definições

As redes neurais artificiais (RNA) são técnicas computacionais inspiradas nos neurônios biológicos e que, a exemplo destes, adquirem conhecimento por meio da experiência. As principais vantagens das RNAs são sua estrutura não linear, o que lhes permitem captar características mais complexas do conjunto de dados; baixa suscetibilidade a ruídos e *outliers*; não exigem informações detalhadas dos processos físicos do sistema a ser modelado; são altamente eficientes em

classificar, prever, reconhecer padrões e estabelecer agrupamentos (Kavzoglu e Mather, 2003; Sudheer et al., 2003; Haykin, 2008; Haykin, 2009).

O neurônio biológico (Figura 1) é constituído por corpo celular, axônio e dendritos. Os dendritos conduzem os sinais das extremidades para o corpo da célula. O corpo celular combina os sinais recebidos formando uma resposta excitante ou inibitória. Esta mensagem é transmitida pelos axônios os quais estão conectados aos dendritos de outros neurônios por meio das sinapses, formando uma rede (Guyton, 1988). De forma análoga, os neurônios artificiais (Figura 2) conduzem as informações ao sistema por meio das entradas, estes sinais são processados pelas camadas intermediárias e a resposta é apresentada pela camada de saída. Os pesos entre as conexões dos neurônios são os parâmetros ajustáveis que variam à medida que o conjunto de dados destinado ao treinamento é apresentado à rede. Os pesos são responsáveis pelo conhecimento adquirido, semelhantes às sinapses dos neurônios biológicos (McCulloch & Pitts, 1943).

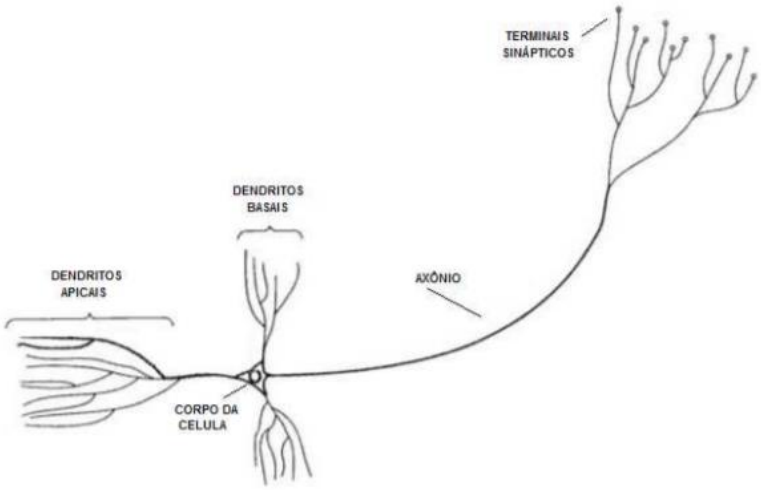

Figura 1. Esquema representativo de um neurônio biológico (Adaptado de Haykin, 2001).

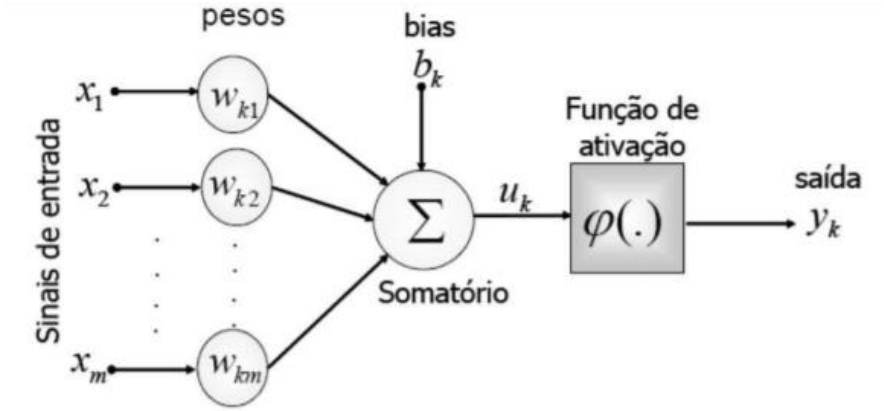

Figura 2. Esquema representativo de um neurônio artificial (Adaptado de Haykin, 2001).

A arquitetura de uma rede neural é definida pelo número de camadas e as conexões entre elas, pelo número de neurônios em cada camada e pelo tipo de conexão entre eles, e pelo algoritmo de aprendizado (Haykin, 2001). Sua estrutura é composta por: uma camada de entrada responsável por fornecer à rede as informações relacionadas às variáveis contidas nos dados; uma ou mais camadas intermediárias ou ocultas que possuem neurônios capazes de extrair e processar as características do sistema; e uma camada de saída, também composta por neurônios, que obtém e apresenta os resultados processados pelas camadas anteriores. O número de camadas ocultas é variável e dependente da complexidade da situação envolvida. Redes com apenas uma camada intermediária criam uma interação global entre os neurônios desta camada e prejudicam a sua capacidade de generalizar em algumas situações. As redes múltiplas camadas (Figura 3) exibem um alto grau de conectividade entre os neurônios das camadas existentes e cada neurônio possui uma função de ativação não linear, o que permite que a rede seja aplicada a problemas mais complexos (Haykin, 2001; Karsoliya, 2012).

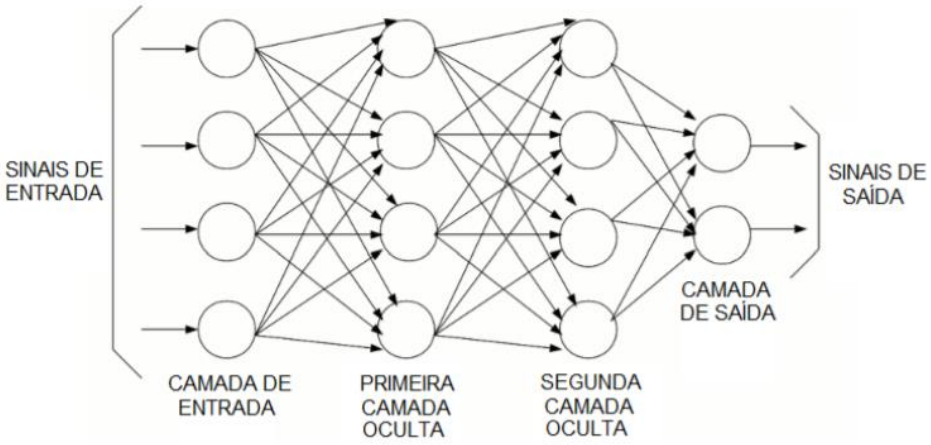

Figura 3. Esquema representativo de uma rede neural artificial de múltiplas camadas (Extraído de Silva, 2019).

O número de neurônios nas camadas ocultas é determinado por alguns métodos ou de forma empírica, a depender do conhecimento prático do pesquisador. Uma decisão errônea para um número de neurônios maior que o necessário pode levar ao *overfitting*, que é a memorização dos dados utilizados no treinamento ao invés do aprendizado que leva à generalização (Karsoliya, 2012). Por sua vez, um número de neurônios inferior ao demandado pelo problema pode acarretar em ineficiência da rede em consequência de um treinamento insuficiente, o chamado *underfitting* (Panchal e Panchal, 2014).

As RNAs podem ser classificadas de acordo com seu tipo de aprendizado em supervisionado e não supervisionado. Ao contrário do aprendizado não supervisionado, no qual não existe indicação de resposta desejada por um agente externo, no supervisionado, um agente externo assinala os erros e acertos de acordo com o padrão de entrada e a resposta desejada (Braga et al., 2000). Há necessidade de disponibilizar amostras, com entradas e saídas desejadas, representativas do processo a ser desenhado. Durante este aprendizado, a rede realiza o ajustamento dos pesos sinápticos e os limiares mediante ações comparativas entre os elementos de processamento. Isto é realizado até que os erros na camada de saída sejam mínimos possíveis. As redes *Perceptrons* são exemplos deste tipo de aprendizado (Quintão, 2015).

Durante a etapa de treinamento da rede é necessária a utilização de um conjunto de regras que orientam este treinamento, um algoritmo de aprendizado. O algoritmo mais utilizado nas redes tipo MLP (*MultiLayer Perceptron*) é o

backpropagation (retropropagação do erro) que se baseia no treinamento supervisionado e trabalha em duas etapas. Na primeira etapa, denominada passo para frente, as informações da camada de entrada são apresentadas à primeira camada oculta e propagadas em direção à camada de saída, de maneira que os sinais de saída de cada neurônio das diferentes camadas são calculados e os valores de saída da rede são comparados às saídas desejadas emitindo-se um sinal de erro. Na segunda etapa, o passo para trás, o sinal de erro emitido é propagado para as camadas anteriores ajustando-se os pesos sinápticos dos neurônios (Silva et al., 2010; Cruz e Nascimento, 2018).

A ativação do neurônio artificial, estímulos que excedam um limiar para que o estímulo seja transmitido, é obtida através de uma "função de ativação", que ativa ou não a saída, dependendo do valor das somas ponderadas de suas entradas (Braga et al., 2007). As funções de ativação fornecem o valor da saída de um neurônio a partir das somas ponderadas recebidas por ele e devem ser escolhida de acordo com o problema em estudo. Uma das funções de ativação mais utilizada é a sigmoidal que pode assumir todos os valores entre 0 e 1. Por definição, é monótona crescente com propriedades assintóticas e de suavidade e apresenta gráfico na forma de “s” (Sant’anna, 2014).

2.4.2 Aplicações no melhoramento genético vegetal

As RNAs têm sido utilizadas em diversas áreas da ciência. Nos últimos anos passaram a contribuir de forma expressiva na análise de dados oriundos de experimentos agrícolas. Devido à capacidade das redes em solucionar problemas não linearmente separáveis e captar características complexas de dados incompletos, são altamente recomendas ao melhoramento genético vegetal, acometido frequentemente por estas situações (Braga et al., 2007). Existem relados da sua utilização em estudos de diversidade genética (Barbosa et al. 2011), classificação e predição de valores genéticos (Peixoto, 2013; Nascimento et al., 2013), para várias espécies como soja (Abraham et al., 2019), milho (Kraisig et al., 2018), trigo (Trautmann, 2020), cana-de-açúcar (Ghazvinei et al., 2018), espécies florestais (Reis et al., 2018) e muitas outras.

O desempenho das RNAs na classificação de genótipos tem sido verificado a partir da comparação com análises discriminantes multivariadas convencionais. Nestes trabalhos são desenvolvidos modelos que reconhecem os padrões presentes nos dados fornecidos e respondem de acordo com as classes definidas. Os resultados têm mostrado maior capacidade das redes em discriminar genótipos e classificá-los corretamente nas populações. De Sá (2018) testou a eficiência das redes neurais em um estudo de dissimilaridade genética entre cultivares de soja e concluiu que, as redes foram eficientes em agrupar os genótipos de acordo com a divergência genética entre elas. Sant'anna (2014) comparou o desempenho das análises discriminantes de Fisher e Anderson com o das redes neurais quanto à capacidade de classificar corretamente genótipos sabidamente pertencentes a diferentes populações simuladas de retrocruzamento, com crescentes níveis de similaridade e viram que, as redes apresentaram um número menor de classificações errôneas sendo promissoras no que diz respeito à discriminação e classificação.

Na predição de valores genéticos, a vantagem das RNAs sobre outras metodologias é que as redes não exigem pressuposição sobre a distribuição dos dados (Barroso et al., 2013). Sant'anna (2018) avaliou a eficiência do método de seleção genômica RR-BLUP e das redes neurais artificiais dos tipos de base radial (RNA-RBF) e *Multilayer Perceptron* (RNA-MLP) na predição genômica na presença de interações epistáticas de uma população F_1 simulada com desequilíbrio gamético. Concluiu que, o uso das RNAs permite capturar interações epistáticas aumentando a confiabilidade na predição de valores genômicos. Guimarães (2020) avaliou modelos de *Machine Learning* voltados a predições para a cultura da soja utilizando dados de clima, solo e de produtividade. O autor desenvolveu os modelos de redes neurais *Multilayer Perceptron*, *Random Forest* e *Extreme Gradient Boosting*, comparou-os entre si e com o modelo de estimativa de produtividade adotado pela FAO. Os resultados apontaram para acurácias acima de 95% dos modelos de *machine learning* quando aplicados à predição de rendimento em soja. Silva (2019) testou o desempenho das RNAs do tipo múltiplas camadas e algoritmo de aprendizado de retropropagação do erro na predição de valores genéticos de clones comerciais de *Eucalyptus* sp., dentre outras aplicações. Observou que as RNAs do tipo múltiplas

camadas proporcionam bom desempenho na predição e seleção quanto à produtividade e que, neste caso, o modelo com três camadas ocultas foi superior.

REFERÊNCIAS

Abraham ER, Reis JGMD, Toloi RC, Souza AED, Colossetti AP. (2019). Estimativa da produção da soja brasileira utilizando redes neurais artificiais. **Agrarian**, 12(44), 261-271.

Alliprandini LF, Abatti C, Bertagnolli PF, Cavassim JE, Gabe HL, Kurek A, Steckling C (2009) Understanding soybean maturity groups in Brazil: environment, cultivar classification, and stability. **Crop Sci**. 49, 801-808.

Agrostat. **Estatística de comercio exterior do agronegócio brasileiro.** Indicadores Gerais Agrostat. 2020. Disponível em: <http://indicadores.agricultura.gov.br/agrostat/index.htm>. Acesso em: 16 Jan. 2021.

Anderson TW (1958) An Introduction to Multivariate Statistical Analysis. New York: John Wiley & Sons, 345 p.

Bacaxixi P, Rodrigues L, Brasil E, Bueno C, Ricardo H, Epiphanio P, Bosquê G (2011) A soja e seu desenvolvimento no melhoramento genético. **Revista Científica Eletrônica de Agronomia**, v. 10, n. 20.

Barbosa CD, Viana AP, Quintal SSR, Pereira MG (2011) Artificial neural network analysis of genetic diversity in Carica papaya L. **Crop Breeding and Applied Biotechnology**, v. 11, n. 3, p. 224-231.

Barroso LMA, Nascimento M, Nascimento ACC, Silva FF, Ferreira RP (2013) Uso do Método de Eberhart e Russell como informação a priori para aplicação de Redes Neurais Artificiais e Análise Discriminante visando a classificação de genótipos de alfafa quanto à adaptabilidade e estabilidade fenotípica. **Embrapa Pecuária Sudeste-Artigo em periódico indexado (ALICE)**, v.31, n.2, p.176-188. São Paulo.

Braga AP, Carvalho ACP, Ludemir TB (2000) Redes Neurais Artificiais - Teoria e Aplicações. Rio de Janeiro: LTC, 2a ed., p. 251.

Braga AP, Ferreira ACP, Ludermir TB (2007) Redes neurais artificiais: teoria e aplicações. LTC Editora. Rio de Janeiro.

Carneiro Júnior JM, Assis GMLD, Euclydes RF, Martins WMDO, Wolter PF (2010) Predição de valores genéticos utilizando inferência bayesiana e frequentista em dados simulados. **Acta Scientiarum Animal Sciences**, v.32, n. 3, p. 337-344.

CONAB. **Safra brasileira de grãos.** Brasília: Conab, 2021. Disponível em <https://www.conab.gov.br/info-agro/safras/graos>. Acesso em: 15 Mai. 2021.

Cruz CD, Ferreira FM, Pessoni LA (2011) Biometria aplicada ao estudo da diversidade genética. Suprema, Visconde do Rio Branco. 620p.

Cruz CD, Regazzi AJ, Carneiro PCS (2012) Modelos biométricos aplicados ao melhoramento genético ed.5. Viçosa, UFV. 480 p.

Cruz CD, Nascimento M (2018) Inteligência Computacional Aplicada ao Melhoramento Genético. Viçosa: UFV, 414p.

Duarte JB (2000) **Sobre o emprego e a análise estatística do delineamento em blocos aumentados no melhoramento genético vegetal.** 69 p. Tese (Doutorado em Genética e Melhoramento de Plantas) – ESALQ/USP, Piracicaba.

Garner, W.W., and Allard, H.A. (1920). Effect of the relative length of day and night and other factors of the environment on growth and reproduction in plants. *J. Agric. Res.* **48**: 553– 606.

Ghazvinei PT, Darvishi HH, Mosavi A, Yusof KBW, Alizamir M, Shamshirband S, Chau K (2018) Sugarcane growth prediction based on meteorological parameters using extreme learning machine and artificial neural network. **Engineering applications of computational fluid mechanics**. 12:738–749.

Gonçalves MC, Fritsche Neto R (2012) Tópicos especiais de biometria no melhoramento de plantas. 1ª Ed. Visconde do Rio Branco: Suprema. 282 p.

Guimarães EDS (2020) **Aprendizado de Máquina aplicado à predição da produtividade da cultura da soja utilizando dados de clima e solo.** 78 p. Tese (Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – USP, São Carlos.

Guyton AC (1988) Fisiologia Humana. 6ª ed., Rio de Janeiro: Interamericana.

Hair JF, Black W, Babin B, Anderson RE, Tatham RL (2009) Análise Multivariada de dados. Porto Alegre: Bookman, 6ª ed., 688 p.

Haykin S (2001) Redes Neurais: princípios e práticas. Porto Alegre: Bookman, 2 ed. 899p.

Haykin S (2008) Neural Networks and Learning Machines. 3rd ed. Pearson - Prentice Hall, Hamilton, p. 938.

Haykin S (2009) Neural networks and learning machines. 3 ed. Prentice Hall, 906p.

Jiménez RA, Villa FB (1995) Prediccion del valor genético: métodos. In: BUXADÉ, C. (Coord.) Zootecnia, bases de proccion animal. Tomo IV. Genética, patologia, higiene y resíduos animales. Madrid: Ediciones Mundi – Pensa. Cap. VI, p. 109 – 122.

Johnson RA, Wichern DW (2002) Applied multivariate statistical analysis. New Jersey: Prentice-Hall. 5 ed. 767 p.

Karsoliya S (2012) Approximating Number of Hidden layer neurons in Multiple Hidden Layer BPNN Architecture. **International Journal of Engineering Trends and Technology** 3:714-717.

Kavzoglu T, Mather PM (2003) The use of backpropagation artificial neural networks in land cover classification. **International Journal of Remote Sensing**, Germany, v.24 (13): 4907-4938.

Kraisig AR, Scremin OB, Mantai RD, Marolli A, Mamann ATW, Brezolin AP, Alessi O, Silva JAG (2018) Regressão por superfície de resposta pelo uso combinado de nitrogênio e hidrogel no sistema milho/aveia. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, v. 6, n. 2.

Lin X, Liu B, Weller JL, Abe J, Kong F (2021) Molecular mechanisms for the photoperiodic regulation of flowering in soybean. **Journal of Integrative Plant Biology**.

López LA, Iemma AF (1998) On the Estimation and Prediction in Mixed Linear Models. **Scientia Agricola**, v. 55, n. 2, p. 291-295.

Mcculloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, v. 5, n. 4, p. 115-133.

Miladinović J, Đorđević V (2011) Soybean morphology and stages of development. **U: Miladinović, J., Hrustić, Milica, Vidić, M.(ed): Soybean. Institute of Field and Vegetable Crops, Novi Sad and Sojaprotein,** Bečej, Grafika, Novi Sad, p. 45-71.

Miladinović J, Ćeran M, Đorđević V, Balešević-Tubić S, Petrović K, Đukić V, Miladinović D (2018) Allelic variation and distribution of the major maturity genes in different soybean collections. **Frontiers in plant science**, v. 9, p. 1286.

Mingoti AS (2005) Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Editora UFMG, 2005.

Nascimento M, Peternelli LA, Cruz CD, Nascimento ACC, Ferreira RP, Bhering LL, Salgado CC (2013) Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. **Crop Breeding and Applied Biotechnology** 13:152-156.

Panchal FS, Panchal M (2014) Review on Methods of Selecting Number of Hidden Nodes in Artificial Neural Network. **International Journal of Computer Science and Mobile Computing** 3:455 – 464.

Panter DM, Allen FL (1995) Using best linear unbiased predictions to enhance breeding for yield in soybean: I. Choosing parents. **Crop Science**, v.35, p.397-405.

Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. **Biometrika**, v. 58, p. 545 – 554.

Peixoto LA (2013) **Redes neurais artificiais na predição do valor genético.** 100 p. Dissertação (Mestrado em Genética e Melhoramento de Plantas) – UFV, Viçosa.

Piepho HP, Mohring J, Melchinger A, Buchse A (2008) BLUP for phenotypic selection in plant breeding and variety testing. **Euphytica**, v. 161, n. 1-2, p. 209-228.

Quintão VQ (2015) **Rede neural e lógica fuzzy aplicadas no melhoramento do feijoeiro.** 91 p. Dissertação (Mestrado em Genética e Melhoramento de Plantas) – UFV, Viçosa.

Regazzi AJ (2000) Análise multivariada, notas de aula INF 766, Departamento de Informática da Universidade Federal de Viçosa, v.2, 2000.

Reis LP, Souza AL, Reis PCM, Mazzei L, Binoti DHB, Leite HG (2018). Prognose da distribuição diamétrica na Amazônia utilizando redes neurais artificiais e autômatos celulares. **Floresta**, 48(1), 93-102.

Resende MDV (2002) Genética biométrica e estatística no melhoramento de plantas perenes. Brasília, Colombo: Embrapa Florestas, 975p.

Resende RT, Carneiro ACO, Ferreira RADC, Kuki KN, Teixeira RU, Zaidan UR, Santos RD, Leite HG, Resende, MDV (2018) Air-drying of eucalyptus logs: Genetic variations along time and stem profile. **Industrial Crops&Products**, 124:316–324.

Sá LG (2018) **Inteligência computacional aplicada ao estudo da divergência e fenotipagem em cultivares de soja.** 57 P. Dissertação (Mestrado em Produção Vegetal) – UFMG, Montes Claros.

Samanfar B, Molnar SJ, Charette M, Schoenrock A, Dehne F, Golshani A, Cober ER (2017) Mapping and identification of a potential candidate gene for a novel maturity locus, E10, in soybean. **Theoretical and Applied Genetics**, v. 130, n. 2, p. 377-390.

Sant'anna IC (2014) **Redes neurais artificiais na discriminação de populações de retrocruzamento com diferentes graus de similaridade.** 114 p. Dissertação (Mestrado em Genética e Melhoramento) – UFV, Viçosa.

Sant'anna IC (2018) **Redes neurais artificiais para predição genômica na presença de interações epistáticas.** 106 p. Tese (Doutorado em Genética e Melhoramento) - UFV, Viçosa.

Sartorio SD (2008) **Aplicações de técnicas de análise multivariada em experimentos agropecuários usando o software R.** 130 p. Dissertação (Mestrado em Genética e Melhoramento de Plantas) – ESALQ/USP, Piracicaba.

Sediyama T (2009) Tecnologias de produção e usos da soja. Londrina: Macenas, 314 p.

Sediyama T, Teixeira RC, Barros HB (2015) Origem, Evolução e importância econômica. In: Sediyama T. Tecnologias de produção e usos da soja. Londrina Mecenas, p.1-5.

Silva IN, Spatti HD, Flauzino RA (2010) Redes Neurais Artificiais: para engenharia e ciências aplicadas. São Paulo: Artliber, 399p.

Silva WDM (2019) **Redes neurais artificiais como ferramenta para prognose de crescimento e melhoramento genético florestal.** 87 p. Tese (Doutorado em Genética e Melhoramento de Plantas) – UNESP, Jaboticabal.

Souza MS (2017) **Análises discriminantes não paramétricas aplicadas ao estudo da diversidade genética baseado em dados fenotípicos quantitativos**. 78 p. Dissertação (Mestrado em Agronomia Tropical) – UFAM, Manaus.

Sudheer KP, Gosain AK, Ramasastri KS (2003) Estimating actual evapotranspiration from limited climatic data using neural computing technique. **Journal of Irrigation and Drainage Engineering**, Califórnia, v.129, p. 214-218.

Trautmann APB (2020) **Modelagem matemática e computacional da produtividade do trigo e otimização do uso do nitrogênio nas condições fenológicas e ambientais.** 324 p. Tese (Doutorado em Modelagem em Matemática) – UNIJUÍ, Ijuí.

Varella CAA (2008) **Análise multivariada aplicada as ciencias agrárias.** (Pós-Graduação em Agronomia Ciência do Solo) – UFR, Seropédica.

Verneque RS (1994) **Procedimentos numéricos e estimação de componentes de variância em análise multivariada pelo método da máxima verossimilhança restrita – modelos mistos aplicados ao melhoramento animal.** 157 p. Tese (Doutorado em Agronomia/Estatística e Experimentação Agronômica) – ESALQ/USP, Piracicaba.

Wang L, Sun S, Wu T, Liu L, Sun X, Cai Y, Han T (2020). Natural variation and CRISPR/Cas9-mediated mutation in GmPRR37 affect photoperiodic flowering and contribute to regional adaptation of soybean. **Plant biotechnology journal**, v. 18, n. 9, p. 1869-1881.

CAPÍTULO 2 – Redes Neurais Artificiais na discriminação e classificação de genótipos de soja pertencentes a diferentes grupos de maturidade relativa

RESUMO - As plantas de soja são afetadas pelo fotoperíodo e, o cultivo de uma determinada variedade é realizado na faixa de latitude que apresenta condições ideais para o seu desenvolvimento, de acordo com os grupos de maturidade relativa. Este trabalho teve o objetivo de classificar genótipos de soja de uma população com ampla variabilidade genética e duas populações com variabilidade estreita para o caráter maturidade relativa utilizando rede neural artificial. As três populações de soja foram obtidas a partir de cruzamentos biparentais entre genitores de grupos de maturidade GMR 5 (Sub-tropical 23^{o} LS) X GMR 9 (Tropical 0^{o} LS), GMR 7 (Tropical 20^{o} LS) X GMR 9 e GMR 5 X GMR 7. A critério de comparação com a arquitetura de rede neural artificial desenvolvida, as metodologias paramétricas discriminantes linear de Fisher e quadrática de Anderson foram aplicadas aos dados para discriminação e classificação dos genótipos. As redes neurais artificiais apresentaram taxa de erro aparente abaixo de 8,16% além de baixa influência de fatores ambientais classificando corretamente os genótipos em suas populações, inclusive nos casos de reduzida variabilidade genética como na população GMR 5 X GMR 6. Por sua vez, as funções discriminantes mostraram-se ineficientes em classificar corretamente os genótipos nas populações na ocorrência de similaridade genealógica (GMR 5 X GMR 6) e ampla variabilidade genética, apresentando taxa de erro acima de 50%. Com base nos resultados deste estudo, RNAs podem ser aplicadas na discriminação de genótipos em gerações iniciais de seleção em programas de melhoramento para o desenvolvimento de cultivares de alto desempenho para ampla e reduzida amplitudes de fotoperíodo em local único de seleção de maneira eficiente com redução de tempo e recursos. A RNA classifica corretamente populações de base estreita e linhagem pura.

Palavras-chave: aprendizado de máquina, fotoperíodo, *Glycine max*, maturidade relativa, mineração de dados , taxa de erro aparente

1 INTRODUÇÃO

A soja é altamente sensível ao fotoperíodo e considerada uma espécie de dias curtos. O comprimento do dia no local de cultivo afeta diretamente o crescimento das plantas. Além disto, o fotoperíodo influencia na mudança do estágio vegetativo para o reprodutivo e consequentemente no florescimento, ciclo total e produção de grãos (Garner & Allard, 1930). De acordo com esta influência, as condições ideais de fotoperíodo para as cultivares ficam restritas a uma determinada faixa de latitude. As cultivares são distribuídas entre 13 grupos de maturidade relativa (GMR) classificados geograficamente com base no crescimento e desenvolvimento das plantas. Devido a grande extensão territorial e variação de latitude, o Brasil compreende os GMRs de 5 a 9, respectivamente do sul (latitude 30°) ao norte do país (latitude 0°) (Alliprandini et al., 2009).

O cultivo de uma variedade em um GMR diferente daquele do seu adequado pode acarretar em alongamento ou redução indesejados do ciclo, desenvolvimento vegetativo insuficiente ou exagerado, suscetibilidade a pragas e doenças específicas de determinada época do ano e baixa produtividade (Miladinović & Đorđević, 2011). Geneticamente, o tempo para atingir o florescimento e a maturação é controlado pelos genes *E*. Até o momento, foram identificados na soja 11 *loci* principais (*E1*-*E10* e *J*) envolvidos no controle destas características (Samanfar et al., 2017). De modo geral, com exceção dos genes *E6*, *E9* e *J*, o alelo dominante dos genes *E* confere florescimento e maturidade tardios, ao passo que o aumento no número de alelos recessivos acarreta em precocidade da variedade.

Cultivares em seus locais adequados de produção tem seu potencial de rendimento maximizado. A estimativa correta dos estágios fenológicos da planta possibilita maior flexibilidade para modificar o seu desenvolvimento como um todo e em trabalhar com características controladas por outros genes que são afetadas pela duração dos estágios vegetativo e reprodutivo. Número de nós e vagens, hábito de crescimento, e ainda aquelas características relacionadas à ocorrência de temperaturas mais altas em determinado estágio da planta, como teores de óleo e

proteína e concentração de N e P nos grãos, podem ser mais bem exploradas, a depender dos objetivos (Miladinović et al., 2018).

O resultado do fenótipo é a soma dos efeitos genéticos, ambientais e da interação entre eles. A evolução da tecnologia e aperfeiçoamento das metodologias nos programas de melhoramento genético busca isolar o máximo possível os efeitos ambientais para aumentar a eficiência da seleção de genótipos com base no efeito genético. As redes neurais artificiais (RNAs) possuem alta capacidade em prever, reconhecer padrões, discriminar e classificar e, diferentemente das abordagens de estatística paramétrica, captam características complexas do conjunto dos dados, além de serem pouco suscetíveis a ruídos e outliers e se adequarem aos problemas não linearmente separáveis comuns à experimentação agrícola (Kavzoglu & Mather, 2003; Sudheer et al., 2003; Haykin, 2008). Atualmente, em nível experimental, modelos de RNA tem sido utilizados na predição de valores genéticos (Soares et al., 2015), adaptabilidade e estabilidade (Oda et al., 2019), fenotipagem (Sá, 2018), estimativas de produtividade (Alves, 2016), diversidade genética (Rahimi et al., 2019; Taratuhin et al., 2020) entre outros, e demonstrando que é possível aumentar a eficiência nas etapas de melhoramento, o que poderá reduzir o tempo e o custo de obtenção de cultivares de alto desempenho.

Este estudo testou a eficiência da RNA-MLP em discriminar e classificar genótipos de soja tropicais pertencentes a populações com variabilidade genética ampla e estreita para maturidade relativa e sua possível aplicação na obtenção de cultivares de alto desempenho para ampla faixa de fotoperíodo das regiões de cultivo de soja no Brasil.

2 MATERIAL E MÉTODOS

2.1 Material genético

Foram obtidas e avaliadas três populações de soja com diferentes amplitudes de variabilidade genética pelo Programa de Melhoramento de Soja da Universidade Estadual “Júlio de Mesquita Filho” em Jaboticabal, São Paulo. O cruzamento entre as cultivares BRS 278 RR (GMR 9.4) e BMX Veloz (GMR 5.0) originou a população

Brasil, caracterizada por seu caráter abrangente. As populações Norte e Sul, caracterizadas por seu caráter mais restrito, foram estabelecidas a partir dos cruzamentos entre as cultivares BMX Energia (GMR 5.3) e BMX Potência (GMR 6.7) e entre as cultivares BRS 245 RR (GMR 7.3) e BRS 278 RR (GMR 9.4), respectivamente. As testemunhas de cada população foram seus respectivos genitores, além das cultivares TMG 7262 RR (GMR 6.2), TMG 1174 RR (GMR 7.4) e TMG 1179 RR (GMR 7.9). Os dados utilizados foram obtidos nos anos agrícolas 2017/2018, 2018/2019 e 2019/2020, que correspondem às gerações filiais F_4 a F_6 para a população 1 e F_5 a F_7 para as populações 2 e 3.

2.2 Área experimental

As populações foram conduzidas na Fazenda de Ensino, Pesquisa e Extensão (FEPE) localizada na UNESP/FCAV – Campus de Jaboticabal – SP. Jaboticabal localiza-se no norte do Estado de São Paulo, a uma latitude 21º15’19’’ Sul e longitude 48º19’21’’ Oeste, altitude de 615 metros e possui uma área de 706,6 Km^2. Apresenta condições ideais de fotoperíodo para genótipos de GMR de 6 a 8 devido ao longo período de chuvas na região, de novembro (primavera) a abril (outono), permitindo cultivares de soja com ciclo de até 150 dias.

2.3 Condução das populações e delineamento experimental

As três populações foram conduzidas nos três anos agrícolas com um número variável de progênies, além de quatro cultivares comerciais como testemunhas. A população Brasil foi constituída por 220 progênies no ano 2017/2018, 252 no ano 2018/2019 e 252 no ano 2019/2020. A população Sul foi constituída por 120 progênies no ano 2017/2018, 168 no ano 2018/2019 e 168 no ano 2019/2020. A população Norte foi constituída por 60 no ano 2017/2018, 60 no ano 2018/2019 e 104 no ano 2019/2020. Os dados foram coletados de cinco plantas individuais selecionadas visualmente dentro de cada parcela.

O delineamento experimental foi o de blocos aumentados de Federer (1956), onde as populações foram dispostas em parcelas de uma linha de cinco metros de comprimento e espaçamento de 0,5 metros entre linhas. As testemunhas, os

respectivos genitores de cada população e outras duas cultivares comerciais, foram alocadas de forma aleatória dentro de cada bloco. A densidade de plantio foi de 15 sementes por metro, e todos os tratos culturais seguiram as recomendações técnicas para a cultura da soja (EMBRAPA, 2013).

2.4 Caracteres avaliados

Foram utilizadas informações dos caracteres avaliados:

- Número de dias para o florescimento (NDF): número de dias contados a partir da data de germinação até o florescimento total de pelo menos 50% da parcela;

- Número de dias para a maturidade (NDM): número de dias contados a partir da data de germinação até a maturidade fisiológica de pelo menos 50% da parcela;

- Ciclo total da cultura (CICLO): número de dias contados a partir da data de germinação até a colheita da parcela;

- Altura de inserção da primeira vagem (AIV): altura em centímetros medida com régua graduada do colo da planta à primeira vagem inserida no caule;

- Altura de planta na maturidade (APM): altura em centímetros medida com régua graduada do colo da planta até a inserção da vagem mais distal;

- Acamamento (Ac): avaliado com uma escala de notas visuais variando de 1 (todas as plantas eretas) a 5 (todas as plantas acamadas);

- Valor agronômico (VA): avaliado com uma escala de notas visuais a qual varia de 1 (plantas com características agronômicas ruins) a 5 (plantas com ótimas características agronômicas);

- Produção de grãos (PG): obtida pelo peso em gramas dos grãos de cada uma das cinco plantas selecionadas em cada parcela, após a colheita e beneficiamento das mesmas.

2.5 Análises discriminantes

Para a realização das análises, os genitores e as cultivares (testemunhas) foram considerados populações de indivíduos distintas, totalizando 11 populações (Tabela 1).

Tabela 1. Genealogia e grupo de maturidade relativa (GMR) de 11 populações de soja utilizadas no estudo.

População	Genealogia	GMR
Brasil	BRS 278 RR x 5953 RSF RR	9.4/5.0
Sul	BMX Potência RR x BMX Energia RR	6.7/5.3
Norte	BRS 245 RR x BRS 278 RR	7.3/9.4
GBN1	BRS 278 RR	9.4
GB2	5953 RSF RR	5.0
GS1	BMX Potência RR	6.7
GS2	BMX Energia RR	5.3
GN2	BRS 245 RR	7.3
TGM7	TMG 1174 RR	7.4
TGM6	TMG 7262 RR	6.2
TGM8	TMG 1179 RR	7.9

As funções discriminantes linear de Fisher e quadrática de Anderson, bem como a partição de dados para seleção de melhor conjunto de treinamento (80% dos dados) e validação (20% dos dados) foram realizadas no software estatístico Genes (Cruz, 2008).

2.5.1 Partição dos dados e validação cruzada

Foram realizadas várias partições dos dados em conjuntos de treinamento (80% dos dados) e conjuntos de validação (20% dos dados). As funções discriminantes de Fisher e Anderson foram dadas a partir das informações do conjunto de treinamento e o conjunto de 20% dos dados foi utilizado para validar as funções. Os vetores de médias e variâncias para cada variável foram obtidos e as médias e variâncias de cada par dos conjuntos gerados foram comparadas, sendo escolhido o par no qual as estimativas eram as mais próximas possíveis. Esta

abordagem é recomendada para que o conjunto teste, que é uma amostra do conjunto de treinamento, seja representativo e a validação seja eficiente.

2.5.2 Análise discriminante de Fisher

Esta análise foi baseada em uma combinação linear das variáveis que apresentam a melhor discriminação entre as populações em estudo. Considerando duas populações (π_i e π_i ') com vetor de médias *v-variado* μ_i e μ_i' e matriz de variâncias e covariâncias comuns $\sum$, de ordem v, define-se a função discriminante linear de Fisher pela expressão:

$$D_{ii'}(\tilde{X}) = \alpha'\tilde{X} = (\mu_i - \mu_{i'})'\Sigma^{-1}\tilde{X}$$

Assim, a função discriminante $D_{ii'}(\tilde{X})$ é uma combinação linear do conjunto de caracteres que possibilita alocar um determinado indivíduo, com vetor de observações $\tilde{X}$, em uma população π_i, ou π_i ', com máxima probabilidade de acerto. Define-se também o ponto médio entre duas populações π_i e π_i' pelo valor m, expresso por uma das equações:

$$m_{ii'} = \frac{1}{2}(\mu_i - \mu_{i'})'\Sigma^{-1}(\mu_i + \mu_{i'}) = \alpha'u = \frac{1}{2}(\alpha'\mu_1 + \alpha'\mu_2)$$

ou

$$m_{ii'} = \frac{1}{2}[D(\mu_1) + D(\mu_2)]$$

Com a função discriminante estimada, adota-se a regra de classificação conforme as expressões:

- Aloca-se x em π_i se:

$$D_{ii'}(\tilde{X}) = \alpha'\tilde{X} = (\mu_1 - \mu_2)'\Sigma^{-1}\tilde{X} \geq m_{ii'}$$

- Aloca-se X em π_i, se:

$$D_{ii'}(\tilde{X}) = \alpha'\tilde{X} = (\mu_1 - \mu_2)'\Sigma^{-1}\tilde{X} < m_{ii'}$$

A ideia básica da Análise Discriminante de Fisher foi transformar observações multivariadas X em observações univariadas Y, derivadas das populações π_1 e π_2 em que estas apresentassem o maior grau de separação possível. Fisher sugere tomar combinações lineares de X para criar as combinações Y's, pois tais combinações podem ser facilmente manipuladas.

2.5.3 Análise discriminante de Anderson

Com as informações sobre a probabilidade "a priori" de um indivíduo pertencer a uma determinada população, são geradas funções que são combinações entre as características avaliadas, obtendo a melhor discriminação entre os indivíduos e alocando-os nas populações corretas. As funções permitem posteriormente a classificação de novos indivíduos de comportamento desconhecido.

Para o estabelecimento da função discriminante de Anderson, considera-se que, para uma população π_j (*j* = 1, 2, ..., *g*), o vetor da variável aleatória *x* tem distribuição $N_v(\mu_j, \Sigma)$, com a seguinte função densidade de probabilidade:

$$f_j(\tilde{X}) = |2\pi\Sigma|^{-1/2}\, e^{-\frac{1}{2}[(\tilde{X}-\mu_j)'\Sigma^{-1}(\tilde{X}-\mu_j)]}$$

Também é admitido que a probabilidade de uma observação pertencer a uma determinada população é p_j ($\sum_{j=1}^{g} p_j = 1$, conhecida a priori. Assim, pode-se estabelecer a função discriminante, dada pela probabilidade de x pertencer a π_j, por meio do logaritmo da função densidade de probabilidade de x e da probabilidade a priori, de forma que se tenha:

$$D_j(\tilde{X}) = -\frac{1}{2}\left[\ln(2\pi) + ln|\Sigma_j|\right] - \frac{1}{2}\left[(\tilde{X} - \mu_j)'\Sigma_j^{-1}(\tilde{X} - \mu_j)\right] + \ln(p_j)$$

Supondo a não homogeneidade das matrizes de variância e covariância ($\Sigma_1 \neq \Sigma_2 \neq \ldots \Sigma_g \neq \Sigma$) temos a função quadrática de Anderson.

Com base nas médias de cada população e na matriz de variância e covariância entre as médias das populações, obtiveram-se as respectivas funções discriminantes. Cada função é uma combinação quadrática das v características avaliadas, existindo tantas funções quanto for o número de populações avaliadas. A partir das funções discriminantes, estima-se, para cada genótipo, o valor discriminante, permitindo, classificar o *i-ésimo* indivíduo, com vetor de média $\tilde{X}_i$, na população π_j se e somente se $D_j(\tilde{X}_i)$ for o maior entre os elementos do conjunto $\{D_1(\tilde{X}_i), D_2(\tilde{X}_i), \ldots, D_g(\tilde{X}_i)\}$.

2.6 Análises por redes neurais artificiais

A rede neural desenvolvida para o presente estudo consistiu no tipo *Multilayer Perceptron* (MLP). No total existem quatro camadas: camada de entrada, duas camadas ocultas e uma de saída. O número de unidades da camada de entrada não corresponde ao número de variáveis do problema, dado que foi necessário converter variáveis categóricas para a representação *one-hot*. Assim, a camada de entrada possui 12 neurônios, camadas ocultas com 64 e 128 neurônios, respectivamente, e camada de saída, com 11 neurônios.

As 12 variáveis de entrada foram as populações (POP), os caracteres avaliados - número de dias para o florescimento (NDF), número de dias para a maturidade (NDM), ciclo total da cultura (CICLO), altura de inserção da primeira vagem (AIV), altura de planta na maturidade (APM), acamamento (AC), valor agronômico (VA), produção de grãos (PG) e os três anos agrícolas (2017/2018, 2018/2019 e 2019/2020). Os três anos agrícolas foram considerados variáveis categóricas transformadas pelo processo *one hot*. O conjunto de dados possui 7288 exemplos.

A Rede MLP foi construída em Python 3.6 utilizando Keras como *Frontend* e TensorFlow 2.3.0 como *Backend* e Scikit-learn 0.22.2. Além disso, o pacote Scikit-learn foi utilizado para gerar as matrizes de confusão e validar o modelo.

A avaliação do modelo é baseada nas métricas de avaliação para classificadores, que em sua maioria são derivados da matriz de confusão. A matriz é gerada a partir dos dados que são separados para o teste. A matriz de confusão foi utilizada para analisar a qualidade das predições feitas pelos modelos. As demais métricas utilizadas foram: acurácia, precisão, *recall* e *f1-score*. A acurácia do classificador é a taxa de acertos (tanto para exemplos positivos e negativos) que o modelo fez sobre os dados de teste. A precisão diz respeito à taxa de acerto de exemplos positivos, enquanto o *recall* é cobertura de exemplos positivos corretos. O *f1-score* é o balanceamento entre as métricas de precisão e *recall*. O *f1-score* é usado quando a proporção de exemplos por classe não é equivalente.

Para a validação do modelo foram executados dois procedimentos. O primeiro divide o conjunto de dados em duas bases diferentes, uma para o conjunto de treinamento e outro para teste (Shalev-Shwartz & Ben-David, 2014). Este procedimento é referido como *hold-out*. No conjunto de treinamento, retirou-se 80% dos dados (tomados aleatoriamente) e os 20% restantes para o conjunto de teste. O segundo procedimento, divide o conjunto de dados em k partições, onde k-1 são os dados para o treinamento e k é o conjunto utilizado para o teste do modelo. Assim, cria-se k modelos, onde os dados para o treinamento e o teste são alterados para cada iteração (Shalev-Shwartz & Ben-David, 2014). A avaliação final do modelo é a média das métricas dos *k* modelos. Esse procedimento é chamado de *k-fold cross-validation*. É frequentemente utilizado para validar modelos cujos conjuntos de dados é relativamente pequeno. O k escolhido foi 10.

Foram testados duas funções de ativação diferentes para as camadas ocultas. A função de ativação escolhida foi a sigmóide-logística (Equação 1) devido ao seu desempenho superior. Na camada de saída foi utilizada a função *softmax* (Equação 2).

$$(x) = 11 + e - x \qquad \text{(Equação 1)}$$

$$\sigma(z)_j = \frac{e^{zj}}{\sum_{k=1}^{K} e^{zk}} \qquad \text{(Equação 2)}$$

para j = 1, ..., K.

Para o treinamento da rede, o algoritmo para atualização dos pesos foi o *backpropagation* que é o padrão para esse tipo de rede. O *backpropagation* é uma forma eficiente para calcular as derivadas parciais de cada camada e os pesos são atualizados utilizando a descida de gradiente que procura minimizar o erro produzido pela rede. De forma geral, o algoritmo aplica os dados e os repassa para as próximas camadas, denominado “passo para frente”. Então, calcula-se o erro da camada de saída e propaga-se o erro para trás, denominado “passo para trás”. Esses passos são repetidos até que o erro seja o menor possível. Um método eficiente para o cálculo da descida do gradiente é a sua versão estocástica (*Stochastic Gradient Descent* - SGD). Na prática, os otimizadores utilizados são variações do SGD. Neste trabalho foi utilizado otimizador ADAM.

O número de ciclos de treinamento foi fixado em 600 épocas. Teve-se o cuidado de limitar o número de iterações, para que esse não se tornasse excessivo, o que poderia levar à perda do poder de generalização.

A melhor arquitetura da rede foi estabelecida por aquela com acurácia média superior, considerando as possibilidades avaliadas, calculada pela multiplicação do número de neurônios em cada camada e as funções de ativação possíveis. Assim, foi escolhida a rede mais eficiente, para cada uma das estratégias adotando como critério a menor taxa de erro aparente.

3 RESULTADOS

As porcentagens de classificações errôneas obtidas pelas análises discriminantes de Fisher e Anderson foram de 58,60% e 50,59%, respectivamente, ambas superiores a 50%. As validações da RNA apresentaram TEA inferior para *k-fold* (5,62%) em relação a *hold-out* (8,16%) em mais de 2% das classificações (Tabela 2).

Tabela 2. Classificação de genótipos de soja em 11 populações de diferentes grupos de maturidade relativa e estimativa da Taxa de Erro Aparente (TEA) de acordo com as análises discriminantes de Fisher e Anderson e abordagens de RNA *hold-out* e *k-fold*.

Abordagem	Total de classificações	Classificações erradas	TEA (%)
Fisher	1517	889	58,60
Anderson	1517	769	50,59
Hold-out	1458	119	8,16
K-fold	729	41	5,62

A superioridade de *k-fold* sob *hold-out* pode ainda ser observada na Tabela 3. A avaliação da qualidade do modelo da rede neural apontou maiores métricas de acurácia, precisão, *recall* e *f1-score* para *k-fold* e maior perda (34,10%) para *hold-out*.

Tabela 3. Métricas de avaliação da qualidade de predição do modelo para as abordagens *hold-out* e *k-fold* na classificação de genótipos de soja em 11 populações de diferentes grupos de maturidade relativa.

Abordagem	Perda (%)	Acurácia (%)	Precisão (%)	Recall (%)	f1-score (%)
hold-out	34,10	91,84	92,14	91,70	91,94
k-fold	26,39	93,36	93,49	93,23	93,36

Considerando a melhor abordagem paramétrica e não paramétrica, as matrizes de confusão, geradas a partir do conjunto de dados destinado à validação, segundo a classificação por Anderson e *k-fold* estão apresentadas nas Tabelas 4 e 5, respectivamente. Para interpretar uma classificação dentro das matrizes de confusão considera-se a população da coluna como aquela a qual o genótipo pertence, e a população da linha aquela onde ele foi alocado pelo modelo em questão. Portanto, as classificações corretas encontram-se na diagonal em destaque, e as incorretas fora da mesma.

Tabela 4. Classificação de genótipos de soja em 11 populações de diferentes grupos de maturidade relativa segundo a Análise Discriminante de Anderson.

POP	Brasil	Sul	Norte	GBN1	GB2	GS1	GS2	GN2	TGM7	TGM6	TGM8
Brasil	**270**	96	84	2	10	70	7	15	27	28	25
Sul	22	**160**	0	0	19	197	3	0	0	40	12
Norte	0	0	**174**	3	0	0	0	15	1	0	0
GBN1	0	0	0	**27**	0	0	0	0	0	0	0
GB2	0	2	0	0	**19**	0	12	0	0	2	0
GS1	0	1	0	0	0	**16**	0	0	0	2	1
GS2	0	1	0	0	1	0	**17**	0	0	0	0
GN2	0	0	1	0	0	0	0	**10**	0	0	1
TGM7	0	0	1	0	0	0	0	3	**20**	0	7
TGM6	1	23	0	0	1	18	1	0	0	**14**	1
TGM8	0	0	0	0	0	1	0	0	8	4	**21**

A Tabela 4 aponta a alocação de genótipos pertencentes à população GS1 na população Sul como a de maior erro. A população Brasil foi a única que recebeu classificações incorretas vindas de todas as outras dez populações. Das 23 plantas pertencentes à população Brasil classificadas incorretamente, 22 foram para a população Sul. 34,5% das plantas da população Sul e 32,6% da população Norte foram classificadas como sendo da população Brasil. Entre as populações Sul e Norte não ocorreram erros de classificação.

Tabela 5. Classificação de genótipos de soja em 11 populações de diferentes grupos de maturidade relativa pela abordagem k-fold.

POP	Brasil	Sul	Norte	GBN1	GB2	GS1	GS2	GN2	TGM7	TGM6	TGM8
Brasil	**314**	11	0	0	0	0	0	0	0	3	0
Sul	4	**205**	0	0	1	4	0	0	0	3	0
Norte	1	0	**72**	0	0	0	0	1	0	0	0
GBN1	0	0	0	**12**	0	0	0	0	0	0	0
GB2	0	2	0	0	**18**	0	0	0	0	0	0
GS1	0	2	0	0	0	**11**	0	0	0	1	0
GS2	0	0	0	0	0	0	**9**	0	0	0	0
GN2	0	0	0	0	0	0	0	**2**	0	0	0
TGM7	1	0	0	0	0	0	0	0	**13**	0	0
TGM6	3	4	0	0	0	0	0	0	0	**25**	0
TGM8	0	0	0	0	0	0	0	0	0	0	**7**

Pela Tabela 5 é possível observar uma grande redução no número de classificações errôneas. As classificações incorretas ocorridas dos genótipos

pertencentes à população Sul na população Brasil foram as responsáveis pela maior parte dos 5,62% de erros da abordagem *k-fold*. A situação recíproca, genótipos pertencentes à população Brasil alocados na população Sul, também foi destacada pela presença de erros. Além destes pares de populações, ocorreram erros de classificação entre Brasil e Sul, GS1 e Sul, Sul e TGM6 e o recíproco, TGM6 e Brasil e o recíproco. A população GBN1 não recebeu alocações errôneas de nenhuma outra população assim como seus genótipos não foram classificados como pertencentes a outras.

Diferenças nas classificações realizadas pelos dois procedimentos podem ser observadas, além da redução dos erros de Anderson para *k-fold*. Nesta última, além da população GBN1, as populações GS2, GN2 e TGM8 não receberam nenhum genótipo errado, e as populações Norte, GS2, TGM7 e TGM8 não tiveram seus genótipos classificados de forma incorreta.

A identificação do ano de avaliação dos genótipos classificados incorretamente aponta um aumento desses erros do ano de 2018 para 2019 e de 2019 para 2020 (Figura 1).

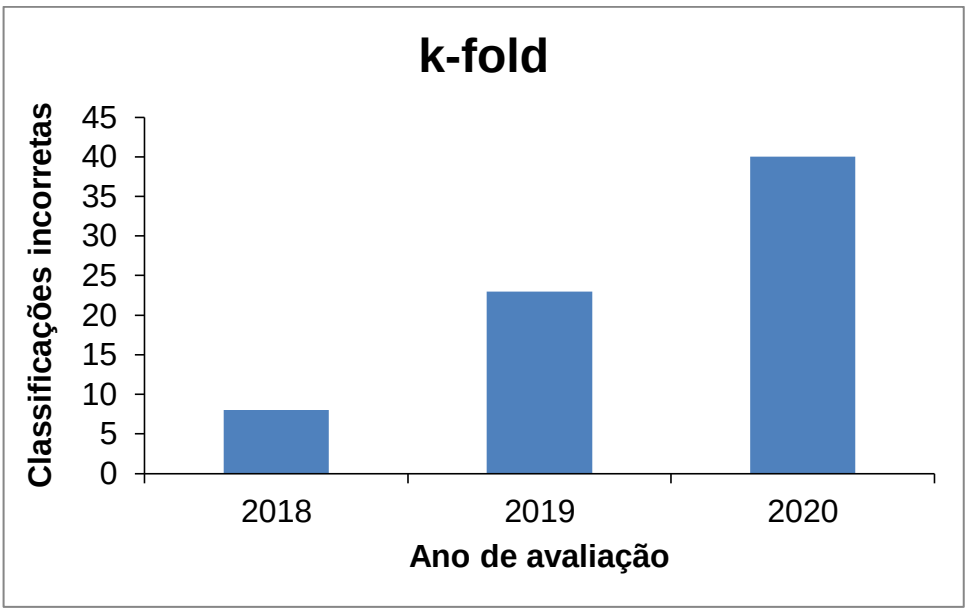

Figura 1. Classificações incorretas de genótipos de soja de 11 populações pela metodologia *k-fold* nos anos de 2017/2018, 2018/2019 e 2019/2020.

4 DISCUSSÃO

O insucesso em discriminar os genótipos e classificá-los em suas populações corretamente pode ser atribuído a quatro razões principais. As populações podem ser de fato, muito próximas em sua origem e genealogia. O número de variáveis avaliadas pode ser insuficiente, bem como baixa qualidade discriminatória das mesmas. Uma quarta razão seria a utilização de uma abordagem estatística inadequada (Cruz et al., 2014). Nas funções discriminantes quadráticas, como a de Anderson, com o aumento de heterogeneidade das matrizes de variâncias e covariâncias, tem-se maior não linearidade dos limiares de classificação, o que leva a um melhor desempenho na modelagem da estrutura da função (Carvalho, 2019). Concordando com esta afirmação, Anderson foi 8,01% mais precisa que Fisher. No entanto, ambas as funções discriminantes levam em consideração parâmetros e pressuposições que muitas vezes não são suficientes para explicar o conjunto de dados. A ocorrência de TEA acima de 50% dos exemplos aponta a limitação dessas metodologias em discriminar os genótipos deste trabalho, principalmente ao considerar a estrutura genética populacional e completamente ineficiente em classificar genótipos. As RNAs, por outro lado, aprendem com a experiência explorando de modo mais detalhado as características contidas nos dados capazes de aumentar a precisão das informações obtidas. *K-fold* é considerada uma abordagem precisa e adequada a pequenos conjuntos de dados (Shalev-Shwartz e Bem-David, 2014), também comprovado pelos resultados deste trabaho. A eficiência de *k-fold* em classificar corretamente mais de 94% dos exemplos mostra sua superioridade em relação a *hold-out*, assim como seus maiores valores de acurácia (93,36%), precisão (93,49%), *recall* (93,23%) e *f1-score* (93,36%). Em termos percentuais, a abordagem menos eficiente das RNAs classificou corretamente mais de 40% dos genótipos quando comparada à função discriminante mais eficiente.

A proximidade devido à origem e genealogia dos genótipos pertencentes a populações diferentes contribui em grande parte para ineficiência de discriminação entre elas como também constatado por Sant'anna (2014) ao trabalhar com RNA e populações de retrocruzamento com diferentes graus de similaridade. A função discriminante de Anderson mostrou ineficiência em discriminar genitores de seus descendentes visto que, a maioria das classificações incorretas ocorreu porque os

genótipos de GS1, representados pela cultivar BMX Potência RR, foram alocados na população Sul, sua descendente.

A população Brasil apresenta ampla variabilidade genética para maturidade relativa, acarretando na presença de muitas combinações alélicas dos genes que controlam o tempo para florescimento e maturidade. Este fator também foi limitante para o bom desempenho da metodologia paramétrica que classificou como pertencentes à população Brasil genótipos de todas as outras populações, 24% do total de erros, capitalizando toda a variação presente nas populações de base genética estreita e até de cultivares de linhagens puras. Ainda na população Brasil, apenas 23 plantas, pertencentes a esta, foram classificadas erroneamente, ou seja, 7,8% sendo que 22 destas foram alocadas na população Sul, mostrando que é amplamente representativa fenotipicamente de todas as outras populações e com pequena semelhança com a população Sul, caracterizando a ampla variabilidade genética presente na população Brasil. Por outro lado, 34,5% das plantas da população Sul foram classificadas como sendo da população Brasil que não apresentam relação de parentesco e explicação biológica conhecida e assim, provavelmente, a causa das classificações incorretas pode ser a técnica estatística aplicada. Por sua vez, na população Norte, 32,6% dos indivíduos foi classificado na população Brasil. Neste caso, as duas populações apresentam um genitor em comum, podendo ser também uma causa biológica para explicar os erros de classificação ocorridos. Entre as populações Norte e Sul não ocorreram classificações errôneas e não existe qualquer parentesco entre os genitores.

Ao considerar os genitores das populações e as outras três testemunhas, nota-se que na população Norte foi onde menos classificações incorretas ocorreram, porém os erros de classificação, no geral, foram sempre grandes, tanto em valores absolutos quanto em porcentagem, com uma tendência de não diferenciação do genitor com a população derivada, como no caso GS1 e populações Brasil e Sul e Sul e GN2 e populações Brasil e Norte. As testemunhas, mesmo sendo linhas puras, foram erroneamente classificadas como pertencentes às diferentes populações, principalmente plantas da população Sul na testemunha TGM6. Assim, pode-se observar que a técnica estatística mostra-se ineficiente em classificar genótipos de

linhagem pura ou ainda de base genética estreita, e parcialmente eficaz mesmo em populações de base genética ampla, como a Brasil.

O único caso em que não ocorreu erro por parte da metodologia de Anderson foi para com a população GBN1, correspondente à cultivar BRS 278 RR com GMR 9.4, que não recebeu genótipos incorretos de nenhuma outra população. O GMR desta população, considerado tardio para a região de avaliação, induz ao alongamento do estágio vegetativo resultando em plantas extremamente altas com baixa produção de grãos, contribuindo para um fenótipo atípico desta cultivar em Jaboticabal, cujos GMRs ideiais são aqueles entre 6 e 8.

Na matriz de confusão obtida por *k-fold*, a ampla variabilidade da população Brasil também levou a classificações incorretas de genótipos da população Sul, embora o número de erros tenha sido de 11 em comparação aos 96 cometidos em Anderson para este mesmo par de populações. As demais classificações incorretas de *k-fold* contidas nos 5,62% ocorreram entre Brasil e Sul, GS1 e Sul, Sul e TGM6 e o recíproco, TGM6 e Brasil e o recíproco, no entanto o número de erros é sempre igual ou menor que 4. A não ocorrência de erros de classificação dos genótipos da população GBN1, GS2, GN2 e TGM8 bem como a ausência de alocações incorretas nas populações Norte, GBN1, GS2, TGM7 e TGM8 demonstram que, mesmo com o parentesco entre o genitor e sua população derivada, a RNA é muito eficiente para populações de base ampla. Além disto, a RNA classifica corretamente populações de base estreita e linhas puras, aprendendo a diferenciar os genótipos. Mesmo a RNA teve sua rotina de classificação influenciada pelo genótipo, embora em menor intensidade do que aquela ocorrida em Anderson.

As avaliações a partir de caracteres tradicionais, nos três anos em um mesmo local, avaliados de forma mecânica não prejudicaram o desempenho da RNA, ao contrário, mostra a baixa influência de ruídos e perda de dados na qualidade das informações obtidas. Além disto, a identificação correta dos genitores e testemunhas mostra que, a proximidade genética entre os genótipos bem como a variabilidade genética existente, seja ela ampla ou restrita, não apresentam limitações para a precisão dos resultados demonstrando a utilidade da técnica em programas de melhoramento mesmo nas gerações F_6 e F_7.

A RNA criada identificou os genótipos que foram classificados incorretamente pelo seu número correspondente na tabela de dados. A partir desta informação, foi possível identificar em qual geração de autofecundação cada um destes genótipos se encontrava pelo ano de avaliação. À medida que as gerações avançam em homozigose e são realizadas seleções direcionadas dentro das populações, principalmente com base na produtividade, estas vão se tornando uniformes e parecidas entre si por serem, geralmente, destinadas ao cultivo no lugar onde se encontram como também constatado por Sant'anna (2014) em dados simulados. Genótipos cultivados fora das suas condições ideais de fotoperíodo sofrem redução do potencial produtivo e são eliminados no processo de seleção. O aumento em cerca de 40% das classificações incorretas do ano de 2018 para 2020 aponta para a importância em classificar e destinar os genótipos às suas regiões adequadas de avaliação. Nesta fase do programa de melhoramento local, seria ideal que as linhagens avançadas em F_5 ou F_6 fossem avaliadas na região de destino de acordo com o fotoperíodo. O conjunto de dados também leva a concluir que o desenvolvimento de linhagens em região intermediária (GMR7) mostra-se adequado para avançar linhagens para fotoperíodos extremos, como aqui para os GMRs 5 e 9.

A alta eficiência das RNAs em discriminar populações de variabilidade genética ampla e estreita possibilita a aplicação destas na obtenção de genótipos a serem testados em todo Brasil, a partir de uma mesma população base de ampla diversidade genética, para o desenvolvimento de novas cultivares com redução de tempo e recursos dos programas de melhoramento genético.

5 CONCLUSÕES

A RNA desenvolvida neste estudo mostrou-se altamente eficiente em discriminar e classificar corretamente genótipos de soja pertencentes a populações com variabilidade genética ampla e estreita para o caráter maturidade relativa.

As RNAs podem ser aplicadas à obtenção de grupos divergentes geneticamente avaliados em diferentes anos com baixa influência da interação genótipos x anos.

Populações de soja tropical com ampla variabilidade genética de diferentes grupos de maturidade relativa podem ser avançados em um único local para fornecimento de linhagens para regiões com diferentes fotoperíodos no Brasil.

REFERÊNCIAS

Alliprandini LF, Abatti C, Bertagnolli PF, Cavassim JE, Gabe HL, Kurek A, Steckling C (2009) Understanding soybean maturity groups in Brazil: environment, cultivar classification, and stability. **Crop Science**. 49, 801-808.

Alves GR (2016) **Estimativa da produtividade da soja com redes neurais artificiais.** 76 p. Dissertação (Mestrado em Engenharia Agrícola) – UEG, Anápolis.

Carvalho VP (2019) **Aprendizado de máquina e estatístico na discriminação de populações na presença de matrizes de covariâncias heterogêneas e vetores aleatórios não normais multivariados.** 59 p. Tese (Doutorado em Estatística Aplicada e Biometria) – UFV, Viçosa.

Cruz CD (2008) Programa genes: diversidade genética. Viçosa: Editora UFV.

Cruz CD, Regazzi AJ, Carneiro PCS (2014) Modelos biométricos aplicados ao melhoramento genético. 3 ed. Viçosa: Universidade Federal de Viçosa.

EMBRAPA - Empresa Brasileira de Pesquisa Agropecuária. Tecnologias de produção de soja – Região Central do Brasil 2014. Londrina: Embrapa Soja, 2013. p. 265.

Federer WT (1956) Augmented (or hoonuiaku) designs. **Hawaiian Planter's Records**, v. 55, p. 191 – 208.

Garner WW, Allard HA (1930) Photoperiodic response of soybeans in relation to temperature and other environmental factors. **J. Agric. Res**. 41, 719–735.

Haykin, S (2008) Neural Networks and Learning Machines. 3st edn. Pearson - Prentice Hall, Hamilton.

Kavzoglu T, Mather PM (2003) The use of backpropagation artificial neural networks in land cover classification. **International Journal of Remote Sensing**, Germany, v.24 (13): 4907-4938.

Miladinović J, Đorđević V (2011) Soybean morphology and stages of development. **U: Miladinović, J., Hrustić, Milica, Vidić, M.(ed): Soybean. Institute of Field and Vegetable Crops, Novi Sad and Sojaprotein,** Bečej, Grafika, Novi Sad, p. 45-71.

Miladinović J, Ćeran M, Đorđević V, Balešević-Tubić S, Petrović K, Đukić V, Miladinović D (2018) Allelic variation and distribution of the major maturity genes in different soybean collections. **Frontiers in plant science**, v. 9, p. 1286.

Oda MC, Sediyama T, Matsuo E, Nascimento M, Cruz CD (2019) Estabilidade e adaptabilidade de produção de grãos de soja por meio de metodologias tradicionais e redes neurais artificiais. **Scientia Agraria Paranaensis**, v. 18, n. 2, p. 117-124.

Rahimi Y, Bihamta MR, Taleei A, Alipour H, Ingvarsson PK (2019) Applying an artificial neural network approach for drought tolerance screening among Iranian wheat landraces and cultivars grown under well-watered and rain-fed conditions. **Acta Physiologiae Plantarum**, v. 41, n. 9, p. 1-17.

Sa LG (2018) **Inteligência computacional aplicada ao estudo da divergência e fenotipagem em cultivares de soja.** 57 P. Dissertação (Mestrado em Produção Vegetal) – UFMG, Montes Claros.

Samanfar B, Molnar SJ, Charette M, Schoenrock A, Dehne F, Golshani A, Cober ER (2017) Mapping and identification of a potential candidate gene for a novel maturity locus, E10, in soybean. **Theoretical and Applied Genetics**, v. 130, n. 2, p. 377-390.

Sant'anna IC (2014) **Redes neurais artificiais na discriminação de populações de retrocruzamento com diferentes graus de similaridade.** 114 p. Dissertação (Mestrado em Genética e Melhoramento) – UFV, Viçosa.

Shalev-Shwartz S, Ben-David S (2014) Understanding Machine Learning: From Theory to Algorithms. Cambridge: Cambridge University Press.

Soares FC, Robaina AD, Peiter MX, Russi JL (2015) Predição da produtividade da cultura do milho utilizando rede neural artificial. **Ciência Rural**, v. 45, n. 11, p. 1987-1993.

Sudheer KP, Gosain AK, Ramasastri KS (2003) Estimating actual evapotranspiration from limited climatic data using neural computing technique. **Journal of Irrigation and Drainage Engineering**, Califórnia, v.129, p. 214-218.

Taratuhin OD, Novikova LY, Seferova IV, Gerasimova TV, Nuzhdin SV, Samsonova MG, Kozlov KN (2020) An Artificial Neural Network Model to Predict the Phenology of Early-Maturing Soybean Varieties from Climatic Factors. **Biophysics**, v. 65, p. 106-117.

CAPÍTULO 3 – Redes Neurais Artificiais aplicadas à predição de valores genéticos de genótipos de soja oriundos de cruzamentos amplo e estreito

RESUMO – O objetivo principal dos programas de melhoramento é a obtenção de cultivares com alto desempenho em produtividade além de tolerância a doenças e pragas. A predição de valores genéticos é de suma importância para a etapa de seleção e avanço dos genótipos promissores. A busca por metodologias que explorem com eficiências as informações contidas nos dados e auxilie os melhoristas na tomada de decisão tem aberto espaço para a inteligência artificial. O objetivo deste trabalho consistiu em determinar a eficiência das Redes Neurais Artificiais (RNA) na predição de valores genéticos de genótipos de soja pertencentes a populações formadas a partir do cruzamento entre genitores de diferentes grupos de maturidade relativa (GMR). Foram utilizados dados de populações de soja de diferentes GMR, avaliadas nos anos agrícolas de 2017/18, 2018/19 e 2019/20. As populações foram conduzidas na Fazenda de Ensino, Pesquisa e Extensão (FEPE), na Unesp/FCAV Campus de Jaboticabal. O delineamento utilizado foi o de Blocos Aumentados, sendo utilizada a média de cinco plantas avaliadas em cada parcela e os caracteres avaliados foram ciclo total (CICLO) e produção de grãos (PG). Foram realizadas as análises individuais de cada ano para cada população e as análises conjuntas de cada população nos três anos agrícolas. Os componentes de variância foram obtidos por meio da metodologia de máxima verossimilhança restrita (REML) e os valores genéticos estimados pela melhor predição linear não viesada (BLUP). Posteriormente, a ordenação dos genótipos foi comparada àquela realizada pela RNA *Multilayer Perceptron* (MLP). A porcentagem de coincidência no ordenamento dos genótipos e o ganho esperado com a seleção (GS) destes genótipos para cada metodologia também foram comparados. Os valores de R (0,999) e RMSE (0,241) indicaram bom desempenho do modelo preditivo mostrando a eficiência das redes neurais em predizer valores genéticos para os genótipos pertencentes a populações de cruzamentos amplo e restrito. Os genótipos considerados como superiores pelas duas metodologias tiveram porcentagens consideráveis de coincidência embora tenham apresentado diferenças na ordenação dos mesmos. O ganho com a seleção estimado a partir dos genótipos indicados como melhores pelas redes neurais foi superior em mais de 47% ao estimado pelo REML/BLUP para produtividade e em 25% para ciclo.

Palavras-chave: componentes de variância, *Glycine max*, grupo de maturidade, modelos mistos, REML/BLUP

1 INTRODUÇÃO

A predição de valores genéticos é imprescindível aos programas de melhoramento genético. Durante as etapas de seleção, os melhoristas precisam identificar o real potencial genético dos indivíduos e decidir quais genótipos avançar e testar em experimentos mais criteriosos. Para isto, estimativas genéticas e ambientais devem ser altamente precisas. Para identificar estes genótipos superiores, faz necessária a aplicação de metodologias de seleção capazes de explorar, com eficiência, o material genético disponível, maximizando o ganho genético em relação às características de interesse (Oda et al., 2007).

A obtenção de informações fiéis a respeito do valor genético dos indivíduos é um ponto crítico nos programas de melhoramento (Peixoto, 2013). Mesmo com a grande quantidade de métodos de predição descritos na literatura, é comum utilizar a média fenotípica, correlação entre caracteres e/ou informações de genealogia para a predição de valores genéticos. A parametrização de efeitos genéticos e ambientais pode contribuir para a obtenção de informações mais precisas e, em caso de situações mais complexas, torna-se necessário um número maior de parâmetros. No entanto, o aumento da complexidade aumenta também a dificuldade na análise e decisão no momento da seleção (Cruz e Carneiro, 2006).

A abordagem por modelos mistos, proposta por Henderson (1977), trouxe grande evolução nas análises de dados. A estimativa dos componentes de variância pelo método da máxima verossimilhança restrita (REML) juntamente com a obtenção da melhor predição linear não viesada (BLUP) dos valores genéticos, têm sido preferida por muitos pesquisadores por maximizarem a acurácia seletiva (Resende, 2007).

A utilização de metodologias estatísticas paramétricas exigem pressuposições quanto à distribuição de probabilidades das variáveis, além de frequentemente assumirem natureza linear do fenômeno estudado. Isto pode acarretar em ineficiência das análises, visto que, estas condições ideais nem sempre são as que realmente ocorrem. Diferentemente dessas análises, as redes neurais artificiais (RNAs) apresentam vantagens que podem torná-las mais adequadas a

determinadas situações e auxiliar nas etapas de seleção e desenvolvimento de cultivares (Sá, 2018).

As RNAs são modelos computacionais semelhantes ao cérebro humano que baseiam-se no aprendizado por experiência. Por serem não paramétricas, não exigem pressuposição dos dados e captam características complexas oferecendo uma função aproximada, linear ou não (Gianola et al., 2011). Além disso, possuem outras vantagens como não necessitar de informações detalhadas do experimento a ser modelado, são tolerantes a ruídos, *outliers* e dados perdidos (Sudheer et al., 2003; Silva et al., 2014).

As RNAs têm sido utilizadas em diversas áreas da agricultura como previsão de produtividade das culturas (Galvão et al., 2018; Abraham et al., 2019; Silva e Schimiguel, 2020), atributos do solo (Liu et al., 2005; Daia et al., 2011; Bittar et al., 2018), interpretação de imagens (Albanez, 2017; Pertille et al., 2018; Magalhães Junior et al., 2019), entre outras. Na predição de valores genéticos, têm mostrado alta eficiência com relação a outras metodologias em diversos estudos (Coutinho et al., 2018; Sant'anna, 2018; Silva, 2019).

Assim, o objetivo deste trabalho foi verificar a eficiência das redes neurais artificiais (RNAs) na predição de valores genéticos de genótipos de soja pertencentes a populações oriundas de cruzamentos amplo e restrito para grupo de maturidade relativa (GMR).

2 MATERIAL E MÉTODOS

2.1 Material Genético

Foram obtidas e avaliadas três populações de soja com diferentes amplitudes de variabilidade genética pelo Programa de Melhoramento de Soja da Universidade Estadual "Júlio de Mesquita Filho" em Jaboticabal, São Paulo. O cruzamento entre as cultivares BRS 278 RR (GMR 9.4) e BMX Veloz (GMR 5.0) originou a população Brasil, caracterizada por seu caráter abrangente. As populações Norte e Sul, caracterizadas por seu caráter mais restrito, foram estabelecidas a partir dos cruzamentos entre as cultivares BMX Energia (GMR 5.3) e BMX Potência (GMR 6.7)

e entre as cultivares BRS 245 RR (GMR 7.3) e BRS 278 RR (GMR 9.4), respectivamente. As testemunhas de cada população foram seus respectivos genitores, além das cultivares TMG 7262 RR (GMR 6.2), TMG 1174 RR (GMR 7.4) e TMG 1179 RR (GMR 7.9). Os dados utilizados foram obtidos nos anos agrícolas 2017/2018, 2018/2019 e 2019/2020, que correspondem às gerações filiais F_4 a F_6 para a população 1 e F_5 a F_7 para as populações 2 e 3.

2.2 Área experimental

As populações foram conduzidas na Fazenda de Ensino, Pesquisa e Extensão (FEPE) localizada na UNESP/FCAV – Campus de Jaboticabal – SP. Jaboticabal localiza-se no norte do Estado de São Paulo, a uma latitude 21º15’19’’ Sul e longitude 48º19’21’’ Oeste, altitude de 615 metros e possui uma área de 706,6 Km^2. Apresenta condições ideais de fotoperíodo para genótipos de GMR de 6 a 8 devido ao longo período de chuvas na região, de novembro (primavera) a abril (outono), permitindo cultivares de soja com ciclo de até 150 dias.

2.3 Condução das populações e delineamento experimental

As três populações foram conduzidas nos três anos agrícolas com um número variável de progênies, além de quatro cultivares comerciais como testemunhas. A população Brasil foi constituída por 220 progênies no ano 2017/2018, 252 no ano 2018/2019 e 252 no ano 2019/2020. A população Sul foi constituída por 120 progênies no ano 2017/2018, 168 no ano 2018/2019 e 168 no ano 2019/2020. A população Norte foi constituída por 60 no ano 2017/2018, 60 no ano 2018/2019 e 104 no ano 2019/2020. Os dados foram coletados de cinco plantas individuais selecionadas visualmente dentro de cada parcela.

O delineamento experimental foi o de blocos aumentados de Federer (1956), onde as populações foram dispostas em parcelas de uma linha de cinco metros de comprimento e espaçamento de 0,5 metros entre linhas. As testemunhas, os respectivos genitores de cada população e outras duas cultivares comerciais, foram alocadas de forma aleatória dentro de cada bloco. A densidade de plantio foi de 15 sementes por metro, e todos os tratos culturais seguiram as recomendações técnicas para a cultura da soja (EMBRAPA, 2013).

2.4 Caracteres avaliados

Foram utilizadas informações dos caracteres avaliados:

- Ciclo total da cultura (CICLO): número de dias contados a partir da data de germinação até a colheita da parcela;

- Produção de grãos (PG): obtida pelo peso em gramas dos grãos das cinco plantas selecionadas em cada parcela, após a colheita e beneficiamento das mesmas.

2.5 Análises estatísticas

Os dados de ciclo (CICLO) e produção de grãos (PG) foram analisados com o auxílio do ambiente de programação R Development Core Team (2020), via abordagem de modelo misto proposto por Scott & Milliken (1993).

Foi realizada análise individual para cada uma das três populações em cada ano agrícola, seguindo o modelo:

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

em que:

Y_{ij} : é a resposta observada do *i-ésimo* tratamento no *j-ésimo* bloco;

μ : é a constante comum às observações;

τ_i : efeito de tratamentos decomposto em p_i = efeito aleatório da progênie *i* e t_i = efeito fixo da testemunha *i*;

β_j : é o efeito do *j-ésimo* bloco (j = 1, 2,..., b);

ε_{ij} : erro experimental aleatório associado à parcela com o *i-ésimo* tratamento, no *j-ésimo* bloco distribuído normal e independentemente, com média zero e variância σ_e^2 $(R = I\sigma_e^2)$.

Os componentes de variância foram estimados a partir da metodologia da Máxima Verossimilhança Restrita (REML). Os efeitos fixos foram verificados pelo teste de significância com o fator F e a significância das variâncias associadas aos

efeitos aleatórios foi verificada pelo teste da razão de verossimilhança “Likelihood Ratio Test” (BERNARDO, 2010).

A herdabilidade foi calculada a partir do estimador:

$$h^2 = \frac{\sigma_g^2}{\sigma_f^2}$$

em que,

σ_g^2: variância genética (progênies)

σ_f^2: variância fenotípica

Para avaliar a precisão experimental, foram determinados o coeficiente de variação experimental (CV) e a acurácia seletiva (rgg’) pelos seguintes estimadores:

$$CV = \frac{\sqrt{\sigma_e^2}}{\bar{x}}$$

Sendo,

σ_e^2: variância do erro

$\bar{x}$: média

$$rgg' = \sqrt{h^2}$$

Sendo,

h^2: herdabilidade a nível de progênie

2.6 Análises por redes neurais

A rede neural desenvolvida para o presente estudo consistiu no tipo *Multilayer Perceptron* (MLP). No total existem quatro camadas: camada de entrada, duas camadas ocultas e uma de saída. O número de unidades da camada de entrada não corresponde ao número de variáveis do problema, dado que foi necessário converter

variáveis categóricas para a representação *one-hot*. Assim, a camada de entrada possui 9 neurônios, camadas ocultas com 64 e 128 neurônios, respectivamente, e camada de saída, com um neurônio correspondendo ao ciclo total da cultura (CICLO), para predição de CICLO, e correspondendo à produção de grãos (PG), para predição de PG.

As 9 variáveis de entrada foram as populações (POP), os caracteres avaliados - número de dias para o florescimento (NDF), ciclo total da cultura (CICLO), altura de inserção da primeira vagem (AIV), altura de planta na maturidade (APM) e produção de grãos (PG), e os três anos agrícolas (2017/2018, 2018/2019 e 2019/2020). Os três anos agrícolas foram considerados variáveis categóricas e transformadas pelo processo *one hot*. O conjunto de dados possui 6158 exemplos.

A Rede MLP foi construída em Python 3.6 utilizando Keras como *Frontend* e TensorFlow 2.3.0 como *Backend* e Scikit-learn 0.22.2.

O conjunto de dados foi dividido em k partições, onde k-1 são os dados para o treinamento e k é o conjunto utilizado para o teste do modelo (*k-fold cross validation*). Assim, cria-se k modelos, onde os dados para o treinamento e o teste são alterados para cada iteração (Shalev-Shwartz & Ben-David, 2014). O k escolhido foi 10. A avaliação final do modelo foi feita a partir da correlação entre os valores observados e os preditos pela rede (R) e pelo parâmetro RMSE (*root mean squared error*) ou raiz do erro médio quadrático obtido pela equação:

$$RMSE = \sqrt{\frac{\sum(Q_{obs} - Q_{cal})^2}{N}}$$

onde,

Q_{obs}: é o valor observado;

Q_{cal}: é o valor previsto pela rede;

N: número de observações utilizadas pela rede.

A função de ativação escolhida foi a sigmóide-logística:

$$(x) = 11 + e - x$$

Na camada de saída foi utilizada a função softmax:

$$\sigma(z)_j = \frac{e^{zj}}{\sum_{k=1}^{K} e^{zk}}$$

para j = 1, ..., K.

Para o treinamento da rede, o algoritmo para atualização dos pesos foi o *backpropagation* que é o padrão para esse tipo de rede. O *backpropagation* é uma forma eficiente para calcular as derivadas parciais de cada camada e os pesos são atualizados utilizando a descida de gradiente que procura minimizar o erro produzido pela rede. De forma geral, o algoritmo aplica os dados e os repassa para as próximas camadas, denominado “passo para frente”. Então, calcula-se o erro da camada de saída e propaga-se o erro para trás, denominado “passo para trás”. Esses passos são repetidos até que o erro seja o menor possível (Bezerra, 2016). Um método eficiente para o cálculo da descida do gradiente é a sua versão estocástica (*Stochastic Gradient Descent* - SGD). Na prática, os otimizadores utilizados são variações do SGD. Neste trabalho foi utilizado o otimizador ADAM.

O número de ciclos de treinamento foi fixado em 600 épocas. Teve-se o cuidado de limitar o número de iterações, para que esse não se tornasse excessivo, o que poderia levar à perda do poder de generalização.

2.7 Comparação entre as metodologias de predição

A comparação da eficiência na predição de valores genéticos entre as metodologias de modelos mistos REML/BLUP e redes neurais artificiais foi realizada por meio do índice de coincidência (%) dos 10 e 20% melhores genótipos para cada caráter de acordo com cada metodologia e o ganho com a seleção considerando uma intensidade de seleção de 20%. Para o caráter ciclo total foram selecionados os genótipos com menores valores genéticos aditivos, visando à seleção de genótipos mais precoces, e para produção de grãos, com o intuito de aumentar estas estimativas, foram selecionados genótipos com maiores valores aditivos.

3 RESULTADOS

As estimativas dos parâmetros genéticos e fenotípicos do ciclo total da cultura (CICLO) e produção de grãos (PG) para as populações Brasil, Sul e Norte, respectivamente, nos três anos de avaliação estão apresentadas na Tabela 1.

A existência de variabilidade entre as progênies pode ser observada pela estimativa do componente de variância genética. Detectou-se diferença estatística a 1 e 5% de probabilidade em 12 das 18 estimativas. A população Sul não apresentou variância genética significativa para PG no ano 2018/2019. Este mesmo componente, na população Norte, foi não significativo para CICLO nos anos 2017/2018 e 2018/2019 e para PG nos anos 2018/2019 e 2019/2020. A herdabilidade variou de 0,03 (produção de grãos, população Norte 2017/2018) até 0,82 (ciclo, população Sul 2017/2018) indicando a proporção da variação observada devida aos componentes genéticos.

A precisão experimental, verificada a partir dos estimadores acurácia e coeficiente de variação ambiental (CV), variou de acordo com as características, as populações e os anos. As estimativas de acurácia variaram entre 17,32% (produção de grãos, população Norte 2017/2018) até 90,55% (ciclo, população Brasil 2017/2018). O coeficiente de variação ambiental também foi diferente a depender do caráter, população e ano agrícola. As estimativas estiveram entre 1,74% (ciclo, população Norte 2017/2018) e 29,12% (produção de grãos, população Sul 2018/2019).

Tabela 1. Estimativa de parâmetros genéticos e fenotípicos para ciclo total da cultura (CICLO) e produção de grãos (PG) de genótipos de soja pertencentes às populações Brasil, Sul e Norte respectivamente, em três anos agrícolas em Jaboticabal, SP.

População Brasil						
	2017/2018		**2018/2019**		**2019/2020**	
Parâmetros	**CICLO**	**PG**	**CICLO**	**PG**	**CICLO**	**PG**
σ^2_g	17,96*	22,10**	26,22**	141,08**	28,38**	97,74**
σ^2_e	80,11	10,95	19,64	28,27	28,87	66,05
h^2	0,82	0,33	0,43	0,17	0,5	0,4
Acurácia (%)	90,55	57,44	65,57	41,23	70,71	63,24
CV (%)	6,9	19,3	3,32	16,19	4,44	20,98
Média	129,79	17,14	133,58	32,83	120,89	38,74
População Sul						
	2017/2018		**2018/2019**		**2019/2020**	
Parâmetros	**CICLO**	**PG**	**CICLO**	**PG**	**CICLO**	**PG**
σ^2_g	82,81**	15,84**	27,81**	$36,57^{ns}$	15,41**	51,07**
σ^2_e	73,66	9,66	8,64	79,22	6,91	60,14
h^2	0,47	0,38	0,24	0,68	0,31	0,54
Acurácia (%)	68,56	61,64	49,00	82,46	55,68	73,48
CV (%)	8,25	20,08	2,46	29,12	2,29	19,73
Média	104,07	15,48	119,44	30,56	114,78	39,3
População Norte						
	2017/2018		**2018/2019**		**2019/2020**	
Parâmetros	**CICLO**	**PG**	**CICLO**	**PG**	**CICLO**	**PG**
σ^2_g	$19,19^{ns}$	130,60**	0^{ns}	69,13ns	33,15**	$15,09^{ns}$
σ^2_e	5,24	3,44	25,43	157,16	15,3	45,8
h^2	0,21	0,03	1	0,69	0,32	0,75
Acurácia (%)	45,82	17,32	1	83,07	56,57	80,6
CV (%)	1,74	4,72	3,83	23	3,75	19,58
Média	131,43	39,3	131,69	54,49	104,48	34,56

**/*/^{ns} Significativo a 1, a 5% de probabilidade e não significativo, respectivamente, pelo teste de razão de máxima verossimilhança; σ^2_g: variância genética; σ^2_e: variância ambiental; h^2: herdabilidade; CV: coeficiente de variação ambiental.

Os valores de R (correlação entre valores observados e preditos pela rede) e RMSE utilizados para avaliar o desempenho do modelo preditivo estão apresentados na Tabela 2. Pode-se observar que para R as estimativas foram superiores a 0,998 indicando uma alta correlação entre os valores observados, utilizados no treinamento, e os valores preditos pela RNA. Por sua vez, para as

estimativas de RMSE, os valores ficaram entre 0,241 e 0,391. O modelo MP 4 reuniu o menor valor de RMSE (0,241) e o maior valor de R (0,999).

Tabela 2. Desempenho das redes neurais artificiais na fase de treinamento.

Modelo	RMSE	R
MP 1	0,267	0,999
MP 2	0,264	0,999
MP 3	0,365	0,998
MP 4	0,241	0,999
MP 5	0,281	0,999
MP 6	0,294	0,999
MP 7	0,253	0,999
MP 8	0,270	0,999
MP 9	0,260	0,999
MP 10	0,391	0,998

RMSE: raiz quadrada do erro quadrático médio; R: correlação entre os valores observados e os preditos pela rede neural.

As Figuras 1 e 2 apresentam o ordenamento das 20% melhores progênies, para CICLO e PG respectivamente, de acordo com a média BLUP obtida a partir das análises por modelos mistos e RNA, para as três populações de soja no ano agrícola 2019/2020. O ordenamento dos genótipos nos anos agrícolas 2017/2018 e 2018/2019 encontram-se nas Figuras 1A, 2A, 3A e 4A em APÊNDICE A. É possível observar uma semelhança entre os genótipos indicados como os melhores pelas duas metodologias, muito embora ocorra grande divergência nos postos ocupados por eles. Para o caráter CICLO na população Norte, ano 2018/2019 não houve ordenamento dos melhores genótipos devido a ocorrência de variância genética zero pela análise por modelos mistos. De acordo com a análise por RNA, houve um ordenamento dos genótipos mesmo com a pequena diferença observada entre eles.

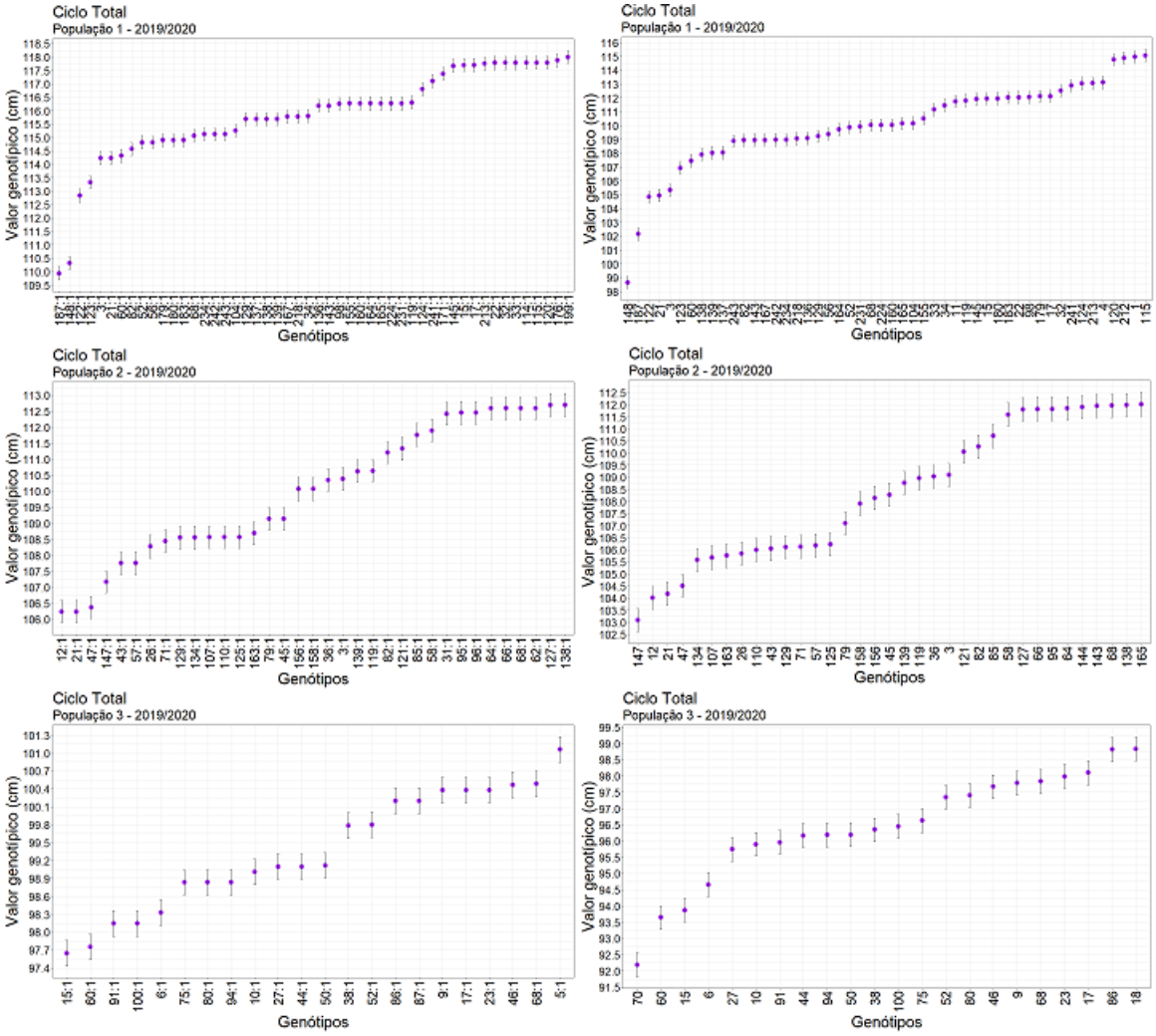

Figura 1. Valores genotípicos e erros associados das 20% melhores progênies para o caráter ciclo total (cm), obtidos pela metodologia REML/BLUP (coluna da esquerda) e RNA (coluna da direita), para as três populações de soja no ano agrícola 2019/2020. Jaboticabal, SP.

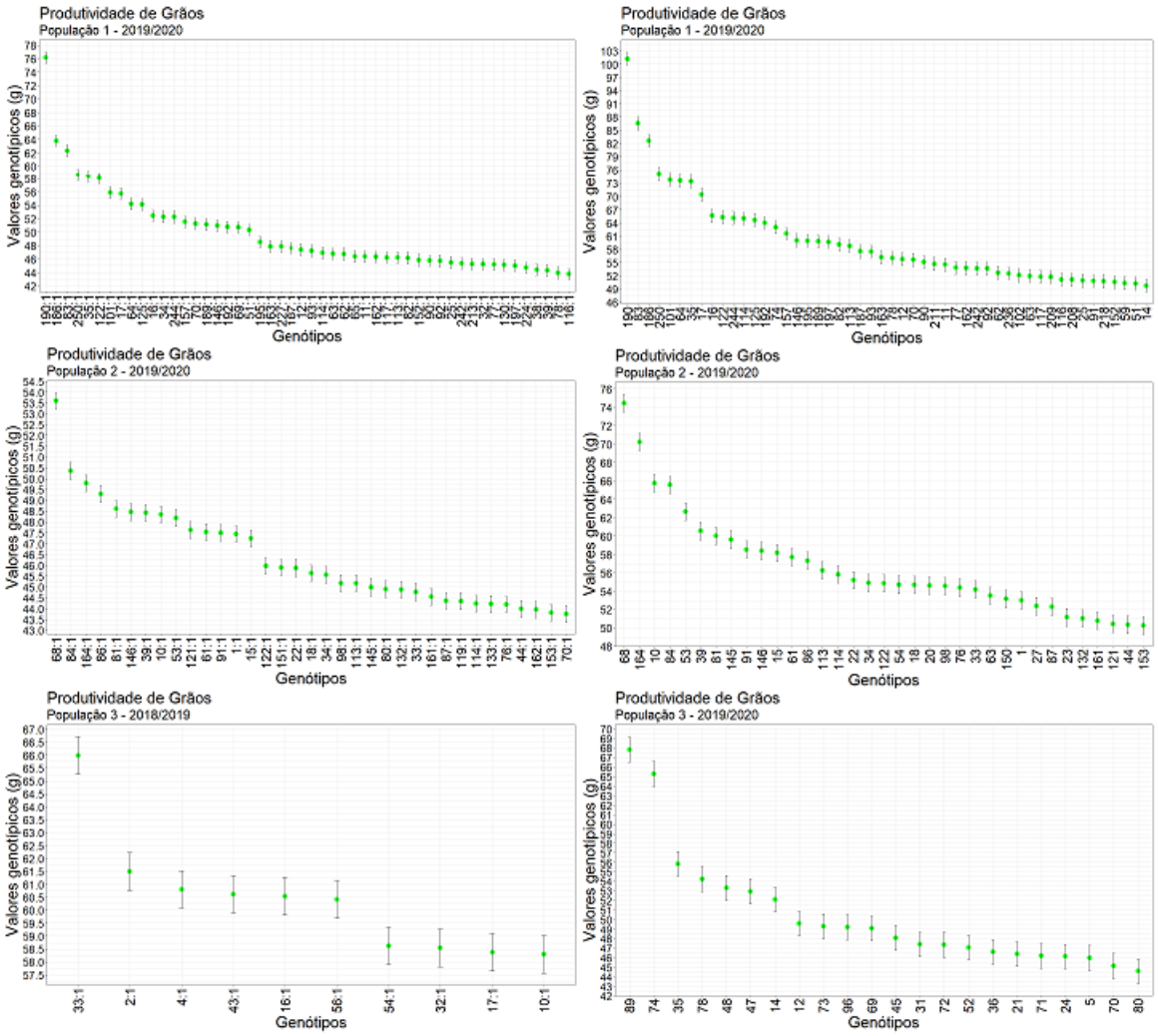

Figura 2. Valores genotípicos e erros associados das 20% melhores progênies para o caráter produção de grãos (g), obtidos pela metodologia REML/BLUP (coluna da esquerda) e RNA (coluna da direita), para as três populações de soja no ano agrícola 2019/2020. Jaboticabal, SP.

A semelhança entre os melhores genótipos indicados pelas duas metodologias também pode ser observada pelo índice de coincidência na Tabela 3.

Para a variável CICLO as porcentagens ficaram entre 30,77% (população Sul, 2017/2018) e 100% (população Sul, 2019/2020), considerando uma intensidade de seleção de 10% e entre 63,16% (população Brasil, 2017/2018) e 92% (população Brasil, 2019/2020), considerando uma intensidade de seleção de 20%.

Para PG, a menor coincidência para a intensidade de 10% ocorreu para a população Brasil no ano 2018/2019 (68,18%) e a maior para a população Brasil no ano 2017/2018 (89,47%). Considerando uma intensidade de seleção de 20%, as

porcentagens de coincidência foram de 68,18% (população Norte, 2019/2020) e 87,50% (população Sul, 2017/2018).

Tabela 3. Porcentagem de coincidência entre os 10 e 20% melhores genótipos para CICLO e PG de acordo com BLUPs e RNA para três populações de soja nos anos agrícolas 2017/2018, 2018/2019 e 2019/2020. Jaboticabal, SP.

	10%					
	2017/2018		2018/2019		2019/2020	
POPULAÇÃO	CICLO	PG	CICLO	PG	CICLO	PG
1	84,21	89,47	81,82	68,18	80,00	80,00
2	30,77	76,92	76,47	70,59	100,00	77,78
3	40,00	80,00	-	80,00	72,73	81,82
	20%					
	2017/2018		2018/2019		2019/2020	
POPULAÇÃO	CICLO	PG	CICLO	PG	CICLO	PG
1	63,16	84,21	79,54	84,09	92,00	80,00
2	68,00	87,50	91,18	85,29	91,43	82,86
3	70,00	70,00	-	70,00	90,91	68,18

Os ganhos esperados com a seleção, considerando uma intensidade de 20%, estão apresentados na Tabela 4. Para PG, os maiores ganhos foram obtidos pelas progênies ordenadas segundo a predição por redes neurais, chegando a 11,91 para a população Norte, enquanto que para a predição BLUP o maior ganho foi de 4,43 para a população Brasil.

Para CICLO, os ganhos foram negativos, pois o objetivo é reduzir o ciclo total da cultura. Os ganhos variaram entre as metodologias, sendo que o maior ganho foi de -5,42 (RNA, população Brasil) e o menor -1,49 (BLUP, população Sul).

Tabela 4. Ganhos esperados com a seleção das 20% melhores progênies para CICLO (cm) e PG (g) no ano agrícola 2019/2020.

	BLUPs		RNA	
POPULAÇÃO	CICLO	PG	CICLO	PG
1	-2,53	4,43	-5,42	8,47
2	-1,49	3,81	-1,99	9,35
3	-1,64	1,98	-2,57	11,91

4 DISCUSSÃO

A obtenção de cultivares cada vez mais produtivas e resistentes depende diretamente da existência de variabilidade genética nas populações de seleção (Bernardo, 2010; Ramalho et al., 2012). Os resultados deste trabalho apontaram a existência de variabilidade entre as progênies na maioria dos casos avaliados. No entanto, baixas estimativas de variância genética podem ser explicadas pela base genética estreita da soja, apontada por diversos autores (Hiromoto & Vello, 1986; Contreras-Soto et al., 2017; Gwinner et al., 2017). Isto pode implicar em baixa variabilidade dentro das populações, além da existência de parentesco entre os genitores, principalmente em caso de populações oriundas de cruzamentos biparentais, como é o caso deste trabalho. A ocorrência de variâncias genéticas não significativas para CICLO na população Norte pode ser explicada pela proximidade entre os grupos de maturidade relativa dos genitores, caracterizando um cruzamento de variabilidade estreita para o caráter, além da variabilidade comprometida pelo tipo de cruzamento biparental e a base genética da espécie.

Os parâmetros variância genética, herdabilidade e acurácia podem ser sub ou superestimados em análises individuais por não considerarem a interação genótipos x ambientes, principalmente quando se trata de caracteres quantitativos. Para o delineamento em blocos aumentados, a ausência de repetições dos genótipos no ano e entre os anos pode explicar a variação das estimativas destes parâmetros bem como a baixa magnitude das mesmas (Duarte, 2000; Rocha e Vello, 1999).

A precisão experimental nos programas de melhoramento é de extrema importância para a obtenção de estimativas confiáveis a respeito dos genótipos que estão sendo testados e, consequentemente, seleção daqueles que forem realmente superiores para as características de interesse do melhorista. Para testar esta precisão, os estimadores mais utilizados são o coeficiente de variação ambiental (CV), classificado como baixo quando menor que 10%, médio de 10 a 20%, alto de 20 a 30% e muito alto quando superior a 30% (Pimentel Gomes, 2009); e a acurácia, classificada como de alta precisão quando se apresenta acima de 70%, de média precisão entre 30 e 70% e de baixa precisão quando menor que 30% (Resende e Duarte, 2007).

O coeficiente de variação ambiental apontou menor precisão experimental para produção de grãos (PG), o que pode ser explicado pelo fato de que, ao se levar em conta esta estimativa, esperam-se maiores valores para caracteres com menores médias (Soares et al., 2015). Pela estimativa da acurácia, a qual não depende da média das progênies para determinada característica, pode-se observar que a maioria dos casos apresentou média ou alta precisão experimental. Existem ainda alguns fatores que contribuem para a elevação da variância ambiental e baixas estimativas de precisão experimental como aqueles típicos do delineamento em blocos aumentados. A avaliação de um grande número de progênies com pouco material disponível leva à utilização de blocos grandes e heterogêneos, parcelas pequenas e sem repetições (Duarte, 2000).

A eficiência do modelo preditivo criado pela rede neural foi verificada de acordo com o parâmetro R (correlação entre os valores observados e os preditos), podendo este variar entre 0 e 1, indicando maior correlação quanto mais próximo de 1; e com o parâmetro RMSE (raiz do erro médio quadrático) podendo este variar entre 0 e 1 e indicando menor erro e maior eficiência quanto mais próximo de 0. As altas estimativas positivas de R (acima de 0,998) e as baixas magnitudes de RMSE (abaixo de 0,391) indicam uma boa precisão dos modelos, baixas magnitudes dos erros e não indicam uma tendência em sub ou superestimar os valores preditos.

A semelhança na indicação dos melhores genótipos para CICLO e PG entre as duas metodologias de predição indica uma elevada eficiência da RNA visto que, a predição por modelos mistos é baseada em pressupostos e leva em consideração vários parâmetros genéticos e ambientais (Silva, 2019). A diferença nos postos ocupados pelos genótipos na ordenação pode ser explicada pela capacidade das redes em captar características complexas dos dados e se basear na experiência para predizer valores genéticos. Esta mesma característica das RNAs pode explicar a ordenação dos genótipos da população Norte no ano 2018/2019 (Figura 3 em ANEXOS) mesmo com a ausência de variância genética apontada pela análise REML/BLUP.

A ocorrência de menores porcentagens de coincidência nas populações no primeiro ano quando comparada àquelas apresentadas no terceiro ano para o caráter CICLO pode estar relacionada a uma maior variabilidade existente nas

populações em gerações onde pode ainda existir segregação, levando a uma divergência na ordenação realizada por cada metodologia.

A partir das estimativas do ganho esperado com a seleção podemos prever o sucesso em selecionar determinados genótipos. Foi possível verificar neste trabalho, por meio de ambas as metodologias de predição, que pode haver sucesso com a seleção baseado nas estimativas de ganho. Para PG, ganhos obtidos pelas progênies ordenadas segundo a predição por redes neurais foram maiores que os ganhos obtidos pela ordenação BLUP em 83,4% para a população Norte, 59,3% para a população Sul e 47,7% para a população Brasil. Ao reduzir a variabilidade espera-se um maior ganho com a seleção, o que pode explicar a ocorrência de maiores estimativas de GS nas populações oriundas de cruzamentos mais restritos.

Para CICLO, os ganhos foram negativos, pois o objetivo é reduzir o ciclo total da cultura e estes também foram superiores quando considerada a ordenação realizada pela rede neural nas proporções de 53,3% para a população Brasil, 36,2% para a população Norte e 25,1% para a população Sul. Embora a porcentagem de ganho tenha sido maior para a população de cruzamento mais amplo para o caráter, pode-se considerar que aí tenha ocorrido maior redução de variabilidade pela seleção direcionada ao local de avaliação de GMR ideal entre 6 e 8.

Muito embora possa ocorrer uma superestimação da RNA na predição de valores genéticos, as estimativas obtidas tanto para o caráter CICLO quanto para PG concordam com aquelas obtidas por Amaral et al., (2019) para a cultura da soja aplicando a mesma intensidade de seleção.

5 CONCLUSÃO

As redes neurais artificiais (RNAs) apresentaram eficiência na predição de valores genéticos quando aplicadas a populações de soja pertencentes a populações de cruzamentos amplo e restrito, apresentando estimativas de ganho genético com a seleção (GS) das progênies, sendo superior ao obtido pela metodologia REML/BLUP.

REFERÊNCIAS

Abraham ER, Reis JGM, Toloi RC, Souza AE, Colossetti AP (2019). Estimativa da produção da soja brasileira utilizando redes neurais artificiais. **Agrarian**, v. 12, n. 44, p. 261-271.

Albanez DDO (2017). **Redes neurais artificiais aplicadas à segmentação de imagens.** 156 f. Dissertação (Mestrado em Modelagem e Otimização (RC)) – UFG, Catalão.

Amaral LO, Bruzi AT, Resende PMD, Silva KB (2019) Pure line selection in a heterogeneous soybean cultivar. **Crop Breeding and Applied Biotechnology**, v. 19, n. 3, p. 277-284.

Bernardo R (2010) **Breeding for quantitative traits in Plants.** Woodbury: Stemma Press, Ed. 2, 400p.

Bittar RD, Alves SMF, Melo FR (2018). Estimativa de atributos de físicos e químicos de solo por meio de redes neurais artificiais. **Revista Caatinga**, *v. 31, n. 3, p.* 704-712.

Contreras-Soto RI, Oliveira MB, Silva DC, Scapim CA, Schuster I (2017) Population structure, genetic relatedness and linkage disequilibrium blocks in cultivars of tropical soybean (Glycine max). **Euphytica**, v. 213, n. 8, p. 173.

Coutinho AE, Neder DG, Silva MCD, Arcelino EC, Brito SGD, Carvalho Filho JLSD (2018) Predição de valores fenotípicos e genotípicos via RR-BLUP/GWS e redes neurais. **Revista Caatinga**, v. 31, n. 3, p. 532-540.

Cruz CD, Carneiro PCS (2006) Modelos biométricos aplicados ao melhoramento genético. 2 ed. v. 2, Viçosa: UFV, 585p.

Daia X, Huo Z, Wang H (2011) Simulation for response of crop yield to soil moisture and salinity with artificial neural network. **Field Crops Research**, v. 121, p. 441-449.

Duarte JB (2000) **Sobre o emprego e a análise estatística do delineamento em blocos aumentados no melhoramento genético vegetal.** 69 p. Tese (Doutorado em Genética e Melhoramento de Plantas) – ESALQ/USP, Piracicaba.

EMBRAPA - Empresa Brasileira de Pesquisa Agropecuária. Tecnologias de produção de soja – Região Central do Brasil 2014. Londrina: Embrapa Soja, 2013. p. 265.

Federer WT (1956) Augmented (or hoonuiaku) designs. **Hawaiian Planter's Records**, v. 55, p. 191 – 208.

Galvão GFP, Carvalho W, Rocha W, Costa JCS (2018). Visão computacional para detecção de doenças fúngicas na agricultura. **ÚNICA Cadernos Acadêmicos**, v. 2, n. 1.

Gianola D, Okut H, Weigel KA, Rosa GJM (2011) Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. **BMC Genetics**, p.12-87.

Gwinner R, Setotaw TA, Pasqual M, Santos JBD, Zuffo AM, Zambiazzi EV, Bruzi AT (2017) Genetic diversity in Brazilian soybean germplasm. **Crop Breeding and Applied Biotechnology**, v. 17, n. 4, p. 373-381.

Henderson CR (1977) Prediction of future records. International conference on quantitative genetics. Proceedings. Ames: IOWA Stte University, p.615-638.

Hiromoto DM, Vello NA (1986) The genetic base of Brasilian Soybean (Glycine max (L.) Merrill) cultivars. **Revista Brasileira de Genética**, v. 9, p. 295-306.

Liu G, Yang X, Li M (2005) An artificial neural network model for crop yield responding to soil parameters. **Lecture Notes in Computer Science**, p. 1017-1021.

Magalhães Junior AM, Santos PR, Sáfadi T (2019) Utilização de Redes Neurais Artificiais na classificação de danos em sementes de girassol. **Sigmae**, v. 8, n. 2, p. 564-575, 2019.

Oda S, Mello EJ, Silva JF, Souza ICG (2007) Melhoramento florestal. In: Borém A. (Ed.). **Biotecnologia Florestal.** Viçosa: UFV, p. 51-71.

Peixoto LA (2013) **Redes Neurais Artificiais na predição do valor genético.** 97 p. Dissertação (Mestrado em Genética e Melhoramento) – UFV, Viçosa.

Pertille CT, Silva GO, Souza CF, Nicoletti MF (2018). Estudo da Eficiência de Classificações Supervisionadas Aplicadas em Imagem de Média Resolução Espacial. **BIOFIX Scientific Journal**, v. 3, n. 2, p. 289-296.

Pimentel Gomes F (2009) **Curso de estatística experimental**. Piracicaba: FEALQ, 15.ed., 451p.

R version 4.0.2 (2020) "Taking Off Again". The R Foundation for Statistical Computing. Platform: x86_64-w64-mingw32/x64 (64-bit).

Ramalho MAP, Abreu ADF, Santos JD, Nunes JAR (2012) **Aplicações da Genética Quantitativa no Melhoramento de Plantas Autógamas.** Lavras: Ed. UFLA, 1 a edição. 522 p.

Resende MDV (2007) Matemática e Estatística na Análise de Experimentos e no Melhoramento Genético. Colombo: Embrapa Florestas, p.362.

Resende MDV, Duarte JB (2007) Precisão e controle de qualidade em experimentos de avaliação de cultivares. **Pesquisa Agropecuária Tropical**, v. 37, n. 3, p. 182-194.

Rocha MDM, Vello NA (1999) Interação genótipos e locais para rendimento de grãos de linhagens de soja com diferentes ciclos de maturação. **Bragantia**, v. 58, n. 1, p. 69–81.

Sá LG (2018) **Inteligência computacional aplicada ao estudo da divergência e fenotipagem em cultivares de soja.** 57 P. Dissertação (Mestrado em Produção Vegetal) – UFMG, Montes Claros.

Sant'anna IC (2018) **Redes neurais artificiais para predição genômica na presença de interações epistáticas.** 106 p. Tese (Doutorado em Genética e Melhoramento) - UFV, Viçosa.

Shalev-Shwartz S, Ben-David S (2014) Understanding Machine Learning: From Theory to Algorithms. Cambridge: Cambridge University Press.

Silva GN, Tomaz RS, Sant'anna IC, Nascimento M, Bhering LL, Cruz CD (2014) Neural networks for predicting breeding values and genetic gains. **Scientia Agricola**, v. 71, n. 6, p. 494- 498.

Silva MJ, Schimiguel J (2020) Identificação de Doenças em Plantas por meio de Processamento de Imagens: Redes Neurais Convolucionais como Auxílio à Agricultura. **Revista de Ubiquidade**, v. 3, n. 1, p. 91-111.

Silva WDM (2019) **Redes neurais artificiais como ferramenta para prognose de crescimento e melhoramento genético florestal.** 87 p. Tese (Doutorado em Genética e Melhoramento de Plantas) – UNESP, Jaboticabal.

Soares IO, Rezende PM, Bruzi AT, Zuffo AM, Zambiazzi EV, Fronza V, Teixeira CM (2015) Interaction between soybean cultivars and seed density. **American Journal of Plant Sciences**, v. 6, n. 09, p. 1425.

Sudheer KP, Gosain AK, Ramasastri KS (2003) Estimating actual evapotranspiration from limited climatic data using neural computing technique. **Journal of Irrigation and Drainage Engineering**, v. 129, n. 3, p. 214-218.

CAPÍTULO 4 – Considerações finais

Pode-se considerar, tomando-se por base os dados no presente trabalho, que as redes neurais artificiais (RNA's) são ferramentas que podem contribuir de maneira expressiva para estudos de classificação e discriminação de genótipos e de predição de valores genéticos para a cultura da soja, ambas fundamentais para o sucesso dos programas de melhoramento da espécie. Embora exista uma série de metodologias relatadas na literatura, as RNAs apresentam vantagens em sua estrutura e funcionamento para lidar com os problemas recorrentes na experimentação agrícola de campo. Análise de dados complexos, experimentos desbalanceados, problemas não linearmente separáveis, dentre outras situações comuns aos pesquisadores, tornam-se grandes desafios na utilização de métodos paramétricos de análise.

A redução do desempenho dos genótipos quando cultivados em condições ambientais diferentes daquelas consideradas ótimas deixa evidente a necessidade de se caracterizar as populações, classificar seus genótipos e direcioná-los para que as avaliações e seleções possam ocorrer no grupo de maturidade os quais maximizam a capacidade produtiva dos mesmos. Ao discriminar e classificar de forma assertiva os genótipos pertencentes a determinadas populações, a rede é capaz de formar grupos distintos quanto à maturidade relativa e classificar novos genótipos de origem desconhecida. Estas vantagens podem contribuir com o desenvolvimento de cultivares de alto desempenho em curtos prazos.

Em programas de melhoramento, a predição de valores genéticos por meio das RNAs permite explorar de forma mais livre e profunda um conjunto de dados de qualquer complexidade, captando informações e características complexas na obtenção de estimativas precisas.

Estas duas aplicações estão diretamente relacionadas ao desenvolvimento futuro de cultivares com características desejáveis às diferentes finalidades e com incremento crescente em produtividade de grãos, que corresponde à principal característica selecionada nos programas de melhoramento genético.

APÊNDICE

APÊNDICE A – Gráficos dos valores genotípicos e erros associados das melhores progênies pelas metodologias REML/BLUP e RNA.

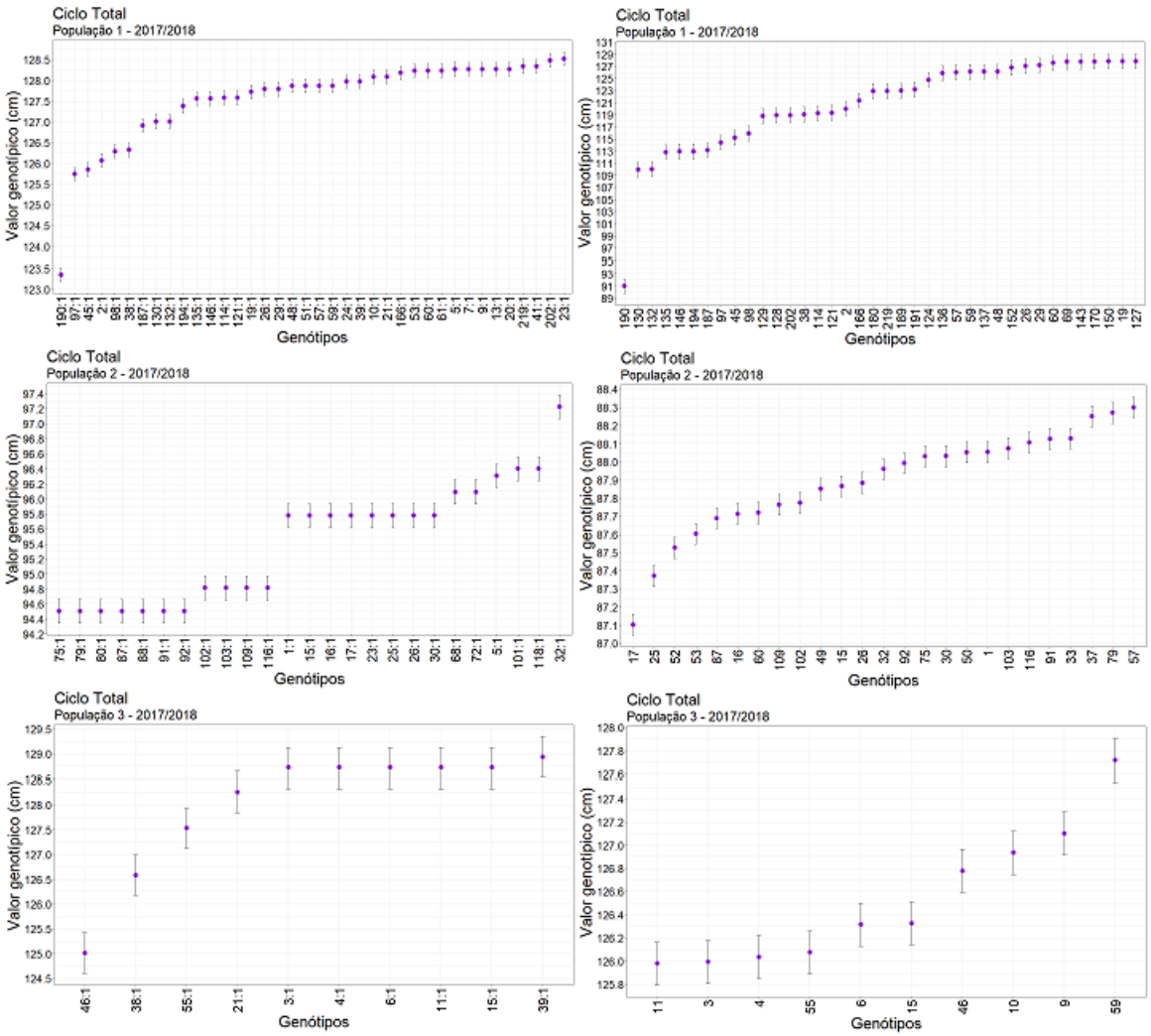

Figura 1A. Valores genotípicos e erros associados das 20% melhores progênies para o caráter ciclo total (cm), obtidos pela metodologia REML/BLUP (coluna da esquerda) e RNA (coluna da direita), para as três populações de soja no ano agrícola 2017/2018. Jaboticabal, SP.

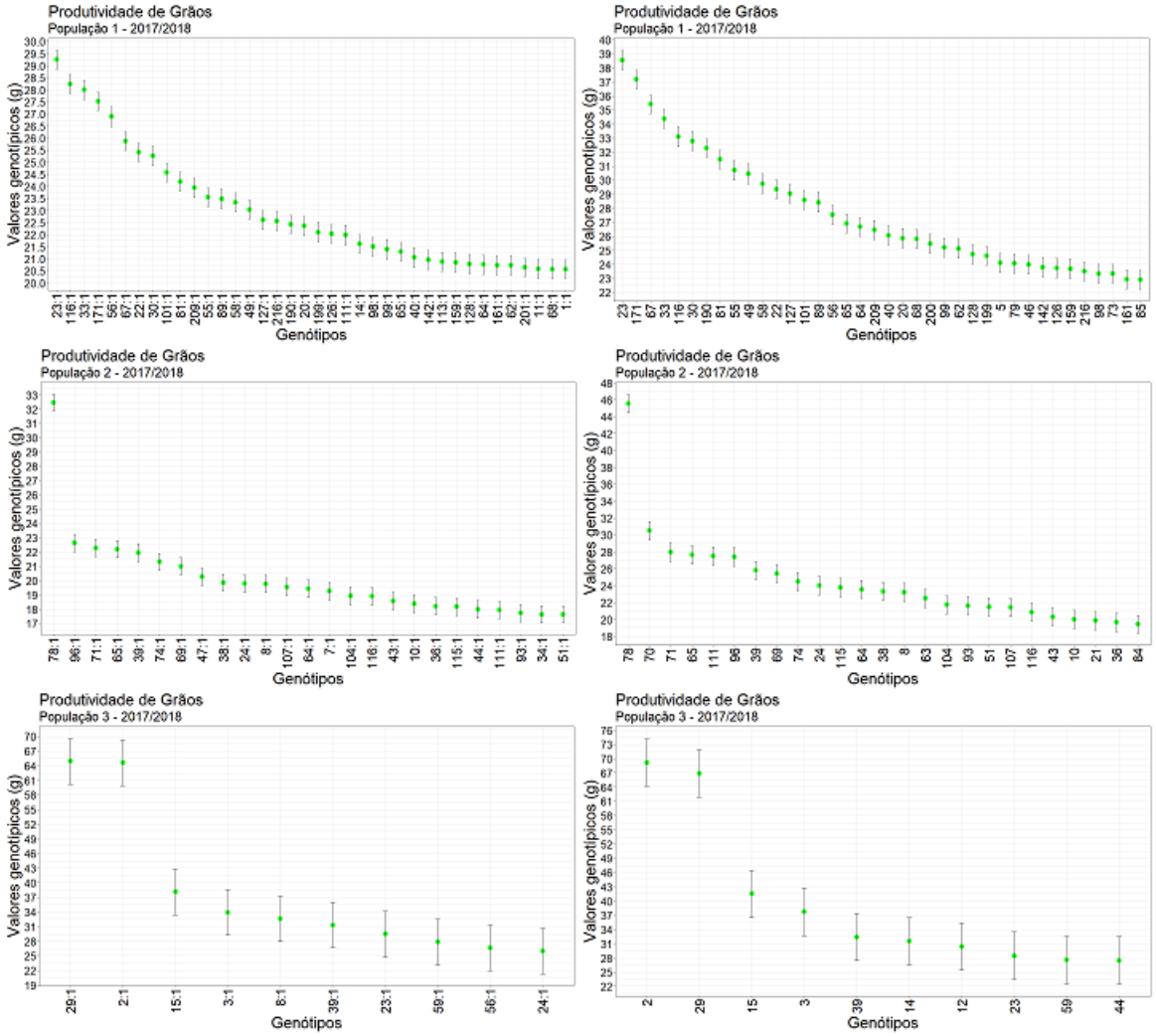

Figura 2A. Valores genotípicos e erros associados das 20% melhores progênies para o caráter produção de grãos (g), obtidos pela metodologia REML/BLUP (coluna da esquerda) e RNA (coluna da direita), para as três populações de soja no ano agrícola 2017/2018. Jaboticabal, SP.

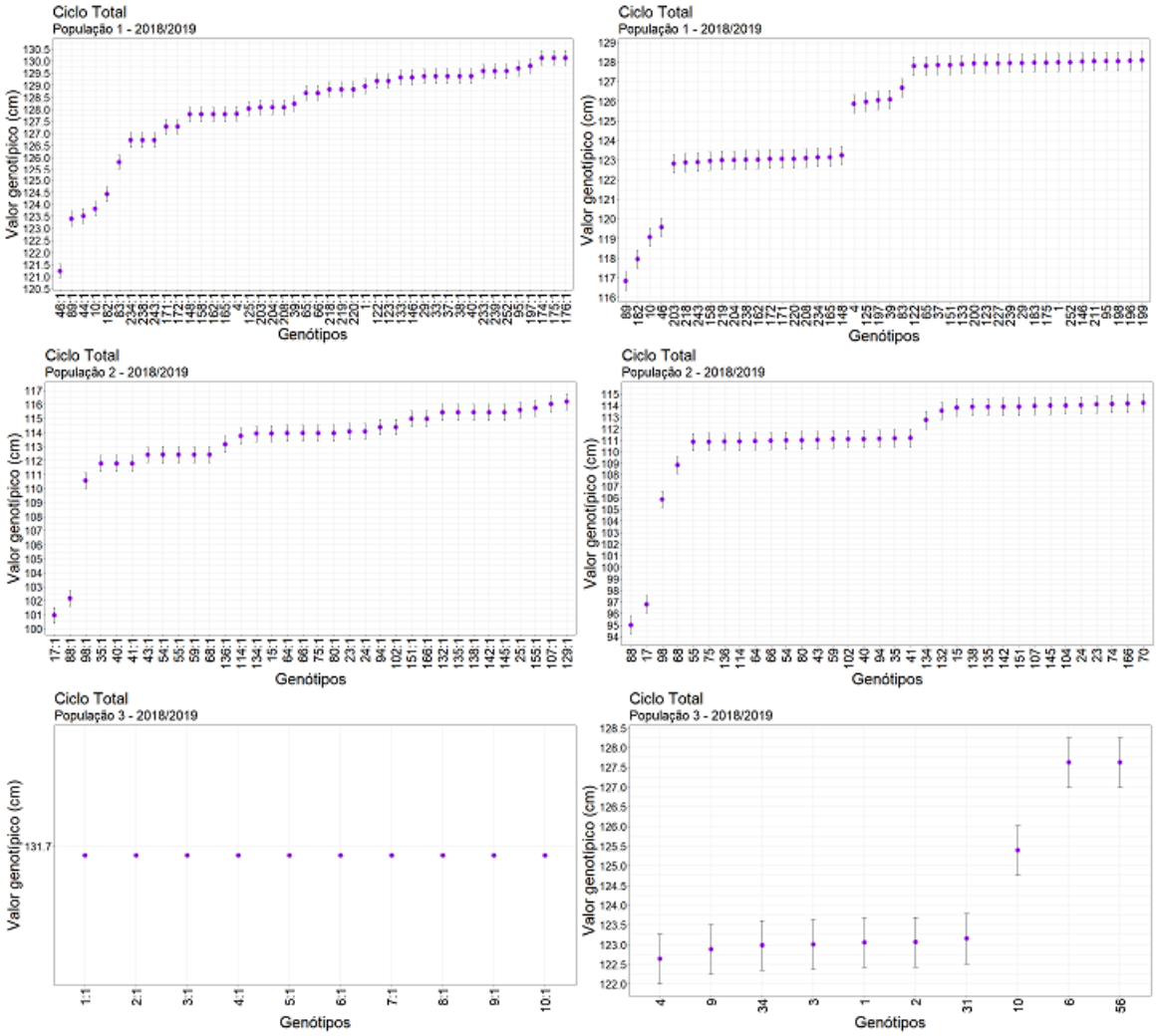

Figura 3A. Valores genotípicos e erros associados das 20% melhores progênies para o caráter ciclo total (cm), obtidos pela metodologia REML/BLUP (coluna da esquerda) e RNA (coluna da direita), para as três populações de soja no ano agrícola 2018/2019. Jaboticabal, SP.

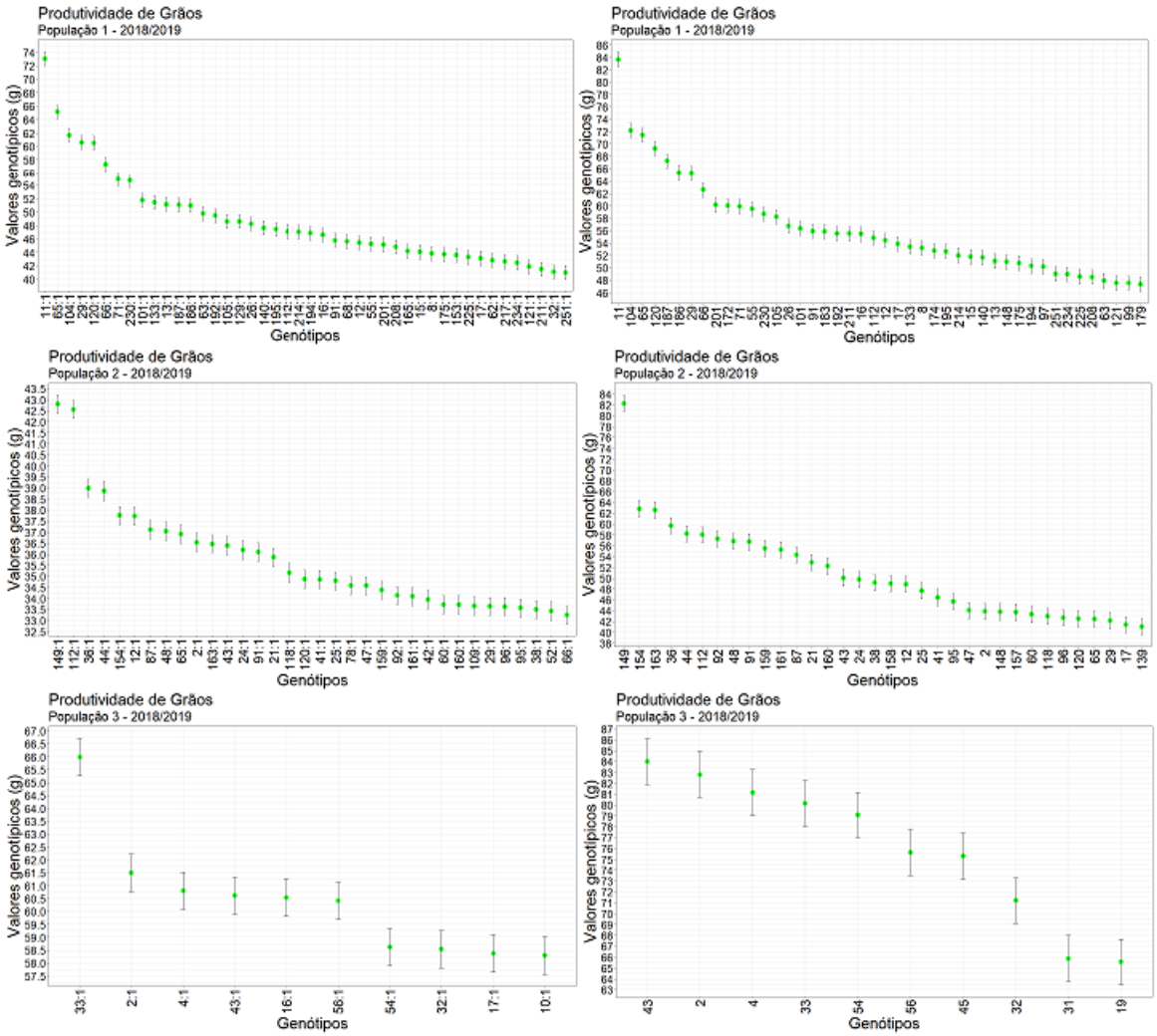

Figura 4A. Valores genotípicos e erros associados das 20% melhores progênies para o caráter produção de grãos (g), obtidos pela metodologia REML/BLUP (coluna da esquerda) e RNA (coluna da direita), para as três populações de soja no ano agrícola 2018/2019. Jaboticabal, SP.