ÉDIMO FERNANDO ALVES MOREIRA

REDES NEURAIS ARTIFICIAIS E ANÁLISE DISCRIMINANTE LINEAR COMO ALTERNATIVAS PARA SELEÇÃO ENTRE FAMÍLIAS DE CANA-DE-AÇÚCAR

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

VIÇOSA
MINAS GERAIS – BRASIL
2014

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

T

M838r
2014

Moreira, Édimo Fernando Alves, 1988-
Redes neurais artificiais e análise discriminante linear como alternativas para seleção entre famílias de cana-de- açucar / Édimo Fernando Alves Moreira. – Viçosa, MG, 2014.
vii, 45f. : il. (algumas color.) ; 29 cm.

Inclui apêndice.
Orientador: Luiz Alexandre Peternelli.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Cana-de-açucar - Melhoramento genético. 2. Redes neurais. 3. Inteligência artificial. 4. Modelos lineares (Estatística). 5. Plantas - Melhoramento genético.
I. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 632.612

ÉDIMO FERNANDO ALVES MOREIRA

REDES NEURAIS ARTIFICIAIS E ANÁLISE DISCRIMINANTE LINEAR COMO ALTERNATIVAS PARA SELEÇÃO ENTRE FAMÍLIAS DE CANA-DE-AÇÚCAR

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

APROVADA: 25 de fevereiro de 2014.

Moysés Nascimento

Felipe Lopes da Silva

Luiz Alexandre Peternelli
(Orientador)

Agradeço a Deus pela saúde e proteção.
Aos meus Pais José Alves Moreira e Maria Clélia Moreira, pelo amor, carinho, dedicação e força dedicados a mim.
À minha irmã e segunda mãe, Clenilda, pelo esforço e carinho dedicado a mim.
Aos professores Décio e Walderez, porta de entrada para a minha graduação.
Aos companheiros de estudo, Eduardo, Denise, Dora, Laís e Gabi pela parceria.
À minha namorada, Mayana, pela companhia e paciência.
Ao meu orientador, Luiz Alexandre Peternelli, pela orientação e confiança depositados desde a graduação, pelas oportunidades oferecidas, pela paciência e ensinamentos.
Aos professores do departamento de estatística, Carlos Henrique, Fabiano, Policarpo, José Ivo, Nerilson e Moysés que tanto contribuíram para a minha formação.
À secretária da pós-graduação em Estatística, Carla, pela disposição em ajudar com essa burocracia brasileira.
Ao CNPq, FAPEMIG e CAPES pelo incentivo financeiro à pesquisa.

BIOGRAFIA

Édimo Fernando Alves Moreira, filho de Maria Clélia Moreira e José Alves Moreira, nasceu na Vila Santo Antônio, Minas Gerais, em 19 de fevereiro de 1988.

Em janeiro de 2012, graduou-se em Agronomia pela Universidade Federal de Viçosa, Viçosa-MG.

Em fevereiro de 2012 ingressou no curso de Mestrado em Estatística Aplicada e Biometria pela Universidade Federal de Viçosa, Viçosa-MG.

SUMÁRIO

RESUMO

MOREIRA, Édimo Fernando Alves, M. Sc., Universidade Federal de Viçosa, fevereiro de 2014. **Redes neurais artificiais e análise discriminante linear como alternativas para seleção entre famílias de cana-de-açúcar.** Orientador: Luiz Alexandre Peternelli. Coorientadores: Márcio Henrique Pereira Barbosa e Cosme Damião Cruz

Um dos grandes desafios nos programas de melhoramento genético de cana-de-açúcar é a seleção eficiente de genótipos nas fases iniciais. Esse desafio advém da grande quantidade de genótipos avaliados e da dificuldade operacional da pesagem das parcelas do experimento, necessária nos principais métodos de seleção. O objetivo deste trabalho é comparar a modelagem por redes neurais, a análise discriminante linear de Fisher e a seleção de famílias usando a variável tonelada de cana por hectare estimada (TCHe) como alternativas para seleção de famílias promissoras em cana-de-açúcar com base nos caracteres indiretos número de colmos (NC), diâmetro de colmos (DC) e altura de colmos (AC). Incialmente foi feita a modelagem via redes neurais em 4 diferentes cenários: com simulação e com padronização das variáveis; com simulação e sem padronização das variáveis ; sem simulação e com padronização das variáveis; e sem simulação e sem padronização das variáveis. Os piores resultados ocorreram no cenário 4, sem padronização e sem simulação e os melhores ocorreram no cenário 1, onde as variáveis foram padronizadas e foram simulados valores de DC, NC, AC e TCHR para 1000 famílias. Posteriormente, foi feita a modelagem via análise discriminante no melhor cenário, ou seja, naquele onde houve simulação e padronização das variáveis de entrada. Para avaliação dos métodos – redes neurais, análise discriminante e seleção via TCHe - foi utilizada a taxa de erro aparente (TEA) e a taxa de erro 1 (TE1) obtidas a partir da matriz de confusão. A simulação e a padronização melhoram o desempenho das redes neurais. A modelagem via redes neurais artificiais e a análise discriminante linear de Fisher fornecem melhores resultados quando comparadas a estratégia usualmente utilizada, que é baseada na estimação da variável tonelada de cana por hectare. Comparando os modelos de redes neurais com a análise discriminante, a rede neural fornece melhores resultados.

ABSTRACT

MOREIRA, Édimo Fernando Alves, M.Sc., Universidade Federal de Viçosa, February, 2014. **Artificial neural networks and linear discriminant analysis as alternatives for selecting among families of cane sugar.** Adviser: Luiz Alexandre Peternelli. Co-advisers: Márcio Henrique Pereira Barbosa e Cosme Damião Cruz.

One of the challenges in breeding of cane sugar programs is the efficient selection of genotypes in the early stages. This challenge arises from the large number of genotypes evaluated and the operational difficulty of weighing all the main plots, required in the main selection methods. The objective of this study is to compare the modeling by artificial neural networks, Fisher's linear discriminant analysis and the selection of families above the overall average for the variable ton of cane per hectare estimated (TCHe) as alternatives for selecting promising families in cane sugar based on indirect character number of stalk (NC), stalk diameter (DC) and stalk height (AC). First was done the analysis for modelling for neural networks in 4 different scenarios (with simulation and standardization of variables; with simulation and without standardization of variables; without simulation and with standardization of variables; without simulation and without standardization of variables). The worst results occur in scenario 4, without standardization and without simulation and the best results occurred in scenario 1, where the variables were standardized and were simulated values of DC, NC, AC and TCHr to 1,000 families. Subsequently, the modeling was done through discriminant analysis in the best scenario, ie, that where there was simulation, and standardization of input variables. To evaluate the methods, neural networks, linear discriminant analysis and selection for TCHe, will be used the apparent error rate (TEA) and one error rate (TE1) obtained for matrix confusion. The simulation and standardization improve the performance of neural network models. The modeling via neural networks and discriminant analysis provide best results when compared to commonly used strategy, which is based on the estimation of the variable ton of cane per hectare. Comparing the models of neural networks with discriminant analysis, neural network gives better results.

INTRODUÇÃO GERAL

Introduzida no período colonial, a cana-de-açúcar (*Saccharum spp*) se tornou uma das mais importantes culturas para a economia brasileira. O Brasil é o maior produtor mundial de açúcar e etanol do mundo, e conquista cada vez mais o mercado externo com o uso do biocombustível como alternativa energética (BARBOSA e SILVEIRA, 2010).

Para a safra 2013/14, a previsão é que o Brasil tenha um acréscimo na área cultivada de 314 mil hectares, equivalendo a 3,7% em relação à safra 2012/13. O acréscimo é reflexo do aumento de área plantada na Região Centro-Sul. São Paulo, Minas Gerais, Goiás e Mato Grosso do Sul deverão ser os estados com maiores acréscimos de área (CONAB, 2013). Esses resultados demonstram a expansão da lavoura de cana-de-açúcar nos principais estados produtores do Brasil.

Segundo Waclawovsky et al. (2010) a média de produtividade mundial é de 80 t/ha e a produtividade máxima teórica da cana-de-açúcar é de 380 t/ha. A produtividade média brasileira na safra 2012/2013 foi de 69.4 t/ha. Portanto, nossos cultivares estão bastante abaixo do potencial genético da cultura, o que reflete a necessidade do melhoramento genético da cultura para que se tenha uma maior produção sem que novas áreas sejam incorporadas.

O programa de melhoramento genético de cana-de-açúcar da Rede Interuniversitária para o Desenvolvimento do Setor Sucroalcooleiro (RIDESA) envolve cinco fases, três fases de teste (T1, T2 e T3), uma fase experimental e uma fase de multiplicação (BARBOSA e SILVEIRA, 2010). A etapa mais importante do programa é a fase inicial, chamada de T1, uma vez que é nesta que são selecionados os primeiros materiais. Nesta fase é produzida uma grande quantidade de genótipos, de maneira que a avaliação destes é um trabalho bastante moroso.

No que se refere aos métodos de estatística e genética utilizados para a seleção de materiais superiores em cana-de-açúcar na fase T1 destacam-se o BLUP (*Best Linear Unbiased Predictor*) individual (BLUPI) (RESENDE, 2002) e o BLUP individual simulado (BLUPIS) (RESENDE e BARBOSA, 2006). No entanto estes procedimentos têm sido usados com restrição nos programas de melhoramento de cana-de-açúcar uma vez que necessitam a pesagem de todas as parcelas do experimento para a obtenção da variável de interesse

tonelada de colmos por hectare real (TCHr), o que é um grande problema operacional.

Uma alternativa que tem sido utilizada para contornar o problema de pesagem no campo é estimar a variável tonelada de colmos por hectare (TCHe), sendo selecionados aqueles com TCHe acima da média geral. Esta estimativa é obtida através dos caracteres indiretos número de colmos (NC), diâmetro de colmos (DC) e altura de colmos (AC), comumente usados para avaliar o rendimento de cana-de-açúcar (CHANG e MILLIGAN, 1992).

Um novo método que tem sido utilizado, ainda de forma tênue, em programas de melhoramento vegetal, é o uso de redes neurais artificiais. Essa abordagem é baseada em princípios de aprendizado por experiência e inteligência artificial.

O principio central da teoria de redes neurais está no fato de que fornecendo exemplos do relacionamento entre variáveis de entrada x e um alvo t, a rede neural irá capturar a relação entre as variáveis, podendo generalizar essas informações para novos casos (MACKAY, 1994).

Neste contexto, as variáveis NC, DC e AC podem ser usadas como variáveis de entrada em modelos de redes neurais para predição das probabilidades de cada família ser selecionada.

Assim, as redes neurais surgem como uma poderosa estrutura de modelagem não linear, capaz de predizer se determinada família será selecionada ou não, baseado em características indiretas, otimizando o processo de seleção de famílias promissoras em cana-de-açúcar.

O objetivo deste trabalho é utilizar modelos de redes neurais para seleção de famílias promissoras em cana-de-açúcar com base nos componentes de rendimento altura de colmos, diâmetro de colmos e número de colmos e compará-los com o método utilizando a variável tonelada de cana por hectare estimada (TCHe) e o método da análise discriminante.

Bibliografia

BARBOSA, M.H.P.; SILVEIRA, L.C.I. (2010) Melhoramento Genético e Recomendação de Cultivares. In: Santos, F.; Borém, A. e Caldas, C. Editores. Cana-de-açúcar: Bioenergia, Açúcar e Álcool - Tecnologias e Perspectivas. Viçosa, MG – Suprema, 578p.

BRASIL. Companhia Nacional de Abastecimento. Acompanhamento da Safra Brasileira. Safra 2013/2014. 2° Levantamento da Cana-de-Açúcar Ago/2013 Brasília: Conab, 2013. 19p. Disponível em:<http://www.conab.gov.br/OlalaCMS/uploads/arquivos/13_08_08_09_39_29_boletim_cana_portugues_-_abril_2013_1o_lev.pdf >> Acesso em: 31 de outubro de 2013.

CHANG, Y.S.; MILLIGAN S.B. (1992) Estimating the potential of sugarcane families to procude elite genotypes using univariate cross prediction methods.Theoretical and Applied Genetics, Berlin, v. 84, p. 662-671.

FALCONER D.S., MCKAY T.F.C.: (1996) Introduction to Quantitative Genetics. Malaysia: Longmans Green, 463p.

WACLAWOVSKY, A. J.; SATO, P.M.; LEMBKE, C.G.; MOORE, P.H.; SOUZA, G.M. (2010). Sugarcane for bioenergy production: an assessment of yield and regulation of sucrose content. Plant Biotechnology Journal, 8(3): 263-276.

CAPÍTULO 1

Revisão de literatura

1. Seleção em cana-de-açúcar

O programa de melhoramento genético de cana-de-açúcar envolve cinco fases, três fases de teste (T1, T2 e T3), uma fase experimental e uma fase de multiplicação (BARBOSA e SILVEIRA, 2010).

Uma das etapas mais importantes do programa é a fase inicial, chamada de T1, uma vez que é nesta que são selecionados os primeiros materiais (OLIVEIRA et al., 2011). Nas outras fases são usados os materiais selecionados em T1, de maneira que uma seleção eficiente nesta fase é de suma importância para os objetivos do programa.

O que vem sendo empregado nos programas de melhoramento de cana-de-açúcar é a seleção massal. Este método se resume na seleção visual individual de clones através de características correlacionadas com o rendimento e sanidade da cana-de-açúcar, sendo realizada por profissionais com mais de 25 anos de experiência no programa de melhoramento de cana-de-açúcar.

Kimbeng e Cox (2003) e Stringer et al. (2011) sugerem que ao invés da seleção individual de clones seja feita uma seleção de famílias seguido da seleção dos melhores genótipos dentro das melhores famílias, tendo em vista que a herdabilidade dos caracteres relacionados ao rendimento baseado em famílias é maior que em plantas individuais. Assim, é preferível que seja dado maior importância na seleção de famílias promissoras seguida da seleção individual de clones dentro das melhores famílias.

Em cana-de-açúcar, a estratégia de seleção ótima seria através da predição de valores genotípicos dos indivíduos usando BLUP (*Best Linear Unbiased Predictor*) individual (BLUPI). Este método usa simultaneamente a informação de família e de indivíduos para a seleção (RESENDE, 2002). No entanto, esse procedimento tem sido usado com restrição nos programas de melhoramento devido a dificuldades operacionais em se obter dados a nível de indivíduo.

Para contornar este problema, Resende e Barbosa (2006) propuseram a seleção das famílias com valores genotípicos acima da média geral, seguido da simulação do número de indivíduos a serem selecionados de cada família de acordo com a relação entre os seus valores genotípicos e do número de indivíduos que se deseja selecionar na melhor família. Este procedimento foi denominado BLUP individual simulado (BLUPIS).

O método BLUPIS elimina automaticamente as famílias com efeito genotípico negativo, ou seja, aquelas abaixo da média geral do experimento. Isto é razoável quando se considera a baixíssima probabilidade de encontrar clones superiores nestas famílias (RESENDE e BARBOSA, 2006).

Em especial, para o caso de dados balanceados, oriundos do delineamento em blocos completos casualizados, a seleção das melhores famílias pode ser simplificada. Nesse caso pode-se provar que as melhores famílias podem ser obtidas selecionando-se aquelas com média fenotípica acima da média geral do experimento para a variável de interesse.

2. Redes Neurais

Redes Neurais são modelos matemáticos inspirados na estrutura neural de organismos inteligentes e que adquirem conhecimento através de experiência para resolver problemas de predição, reconhecimento de padrões e classificação (MACKAY, 1994).

O principio central da teoria de redes neurais está no fato de que fornecendo exemplos do relacionamento entre variáveis de entrada x e um alvo t, a rede neural irá capturar a relação entre as variáveis, podendo generalizar essas informações para novos casos (MACKAY, 1994).

2.1. Inspiração biológica

Como dito anteriormente, os modelos de redes neurais são baseadas no funcionamento do cérebro humano. Este por sua vez tem como unidade básica os neurônios, células especializadas no armazenamento e transmissão das informações.

Um neurônio típico apresenta três partes distintas: os dendritos, o corpo celular e o axônio.

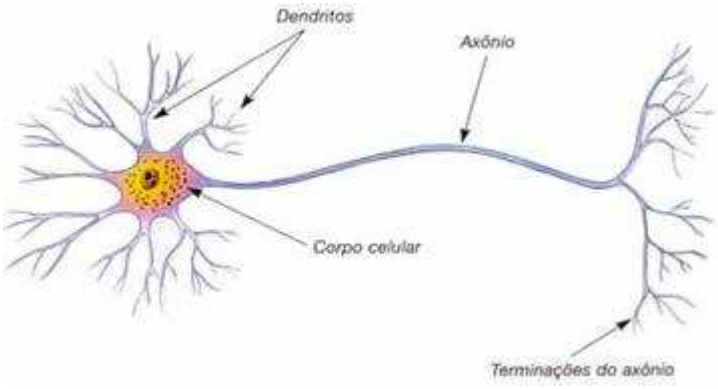


Figura 1: Modelo de neurônio biológico

Fonte: (http://www.sogab.com.br/anatomia/neuronio.jpg)

Os dendritos são os responsáveis por receber os estímulos e transmiti-lós até o corpo celular. No corpo celular, os estímulos serão processados e enviados até os axônios, que são os responsáveis por transmitir os impulsos ou estímulos para a outra célula.

O ponto de contato entre os dendritos e as terminações do axônio é chamado de sinapse, que é então a responsável por regular o fluxo dos estímulos e impulsos, de maneira que se estes impulsos atingem um certo limiar, o estímulo é transmitido, caso contrário não.

A atuação em conjunto dos neurônios é o responsável por regular as nossas atividades, desde as mais simples até as mais complexas. A essas células (neurônios) interligadas é que damos o nome de redes neurais (BRAGA et al., 2000).

2.2. Modelo matemático dos neurônios biológicos

Diversos pesquisadores propuseram modelos para simular em computadores o funcionamento do cérebro humano, sendo o mais aceito o modelo proposto por McCulloch e Pitts (1943), conhecido como *perceptron*.

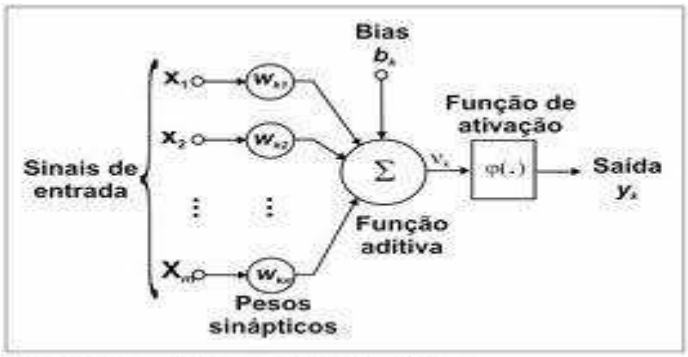


Figura 2: Modelo de neurônio artificial de McCulloch e Pitts (1943)
Fonte: (HAYKIN, 2001)

Neste modelo, tem-se de maneira simplificada os componentes e o funcionamento de um neurônio biológico. Os sinais de entrada correspondem aos impulsos ou informações que chegam aos dendritos, os pesos representam o grau de excitação do impulso, a função aditiva ou somador, corresponde ao corpo celular, ou seja, é onde a informação é processada. O *bias* ou viés corresponde a uma tendência sistemática de o estimador fornecer uma resposta errada. A função de ativação representa a sinapse, pois é ela que irá regular a saída da informação. Finalmente os axônios seriam a saída.

2.3. Funções de ativação

Matematicamente uma rede neural pode ser representada por uma função $y(x;w)$, onde a variável saída da rede y é uma função das variáveis de entrada x, parametrizada pelos pesos w (MACKAY, 1994). Estas funções são chamadas funções de ativação.

Para a modelagem em redes neurais podem ser utilizadas diferentes funções de ativação. Em teoria, qualquer função contínua e monotômica crescente onde $x \in R$ e $y(x;w) \in [-1,1]$ pode ser utilizada como função de ativação.

As mais utilizadas são:

a) A função sigmoidal $y(x;w) = \frac{1}{1+e^{-wx}}$

b) A função tangente hiperbólica $y(x;w) = \frac{1-e^{-wx}}{1+e^{-wx}}$

Estas funções são amplamente utilizadas, pois são funções não lineares com um comportamento levemente linear, o que permite inserir a não linearidade presente em problemas mais complexos. Além disso, são funções facilmente diferenciáveis permitindo a obtenção dos estimadores de maneira mais simples (HAYKIN, 2001).

2.4. Arquitetura da rede neural artificial

A arquitetura em redes neurais se refere à disposição dos neurônios na rede, ou seja, como a rede está estruturada. As arquiteturas mais comuns são a *Feed-forward networks* ou rede alimentada a diante e *Back-forward networks* ou redes recorrentes (HAYKIN, 2001).

Em problemas de predição e classificação, problema do presente trabalho, a rede neural mais utilizada é a *Feed-forward networks* (HAYKIN, 2001).

As *Feed-forward networks* são redes estruturadas em camadas, onde os neurônios são dispostos em conjuntos ordenados e em sequência. Estas redes podem apresentar uma ou mais camadas, sendo chamadas de rede de camada única e rede de múltiplas camadas respectivamente (HAYKIN, 2001).

A rede de camada única é capaz de realizar a classificação com apenas duas classes, no entanto aumentando o número de neurônios, podemos realizar classificação com mais classes desde que essas sejam linearmente separáveis (HAYKIN, 2001).

Para problemas mais complexos, nos quais não obtemos uma solução satisfatória com a rede de camada única, propõe-se o uso das redes *perceptron* de multicamadas – *Multilayer Perceptron* (MLP) (HAYKIN, 2001).

2.5. Redes perceptron de multicamadas

O modelo de rede neural mais utilizado em problemas envolvendo classificação e predição de dados é o perceptron multicamadas (MLP).

Redes MLP são redes *Feed-forward* de múltiplas camadas que possuem uma ou mais camadas de neurônios entre as camadas de entrada e saída, chamada de camada oculta (LIPPMANN, 1987). Nestas redes, o fluxo

de informação se propaga pra frente, camada por camada, desde a camada de entrada até a camada de saída como mostra a figura 3.

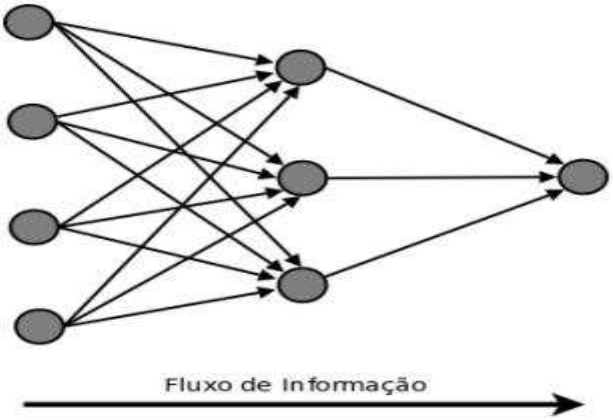


Figura 3: Fluxo de informação em uma *Feed-forward networks* multicamadas
Fonte: (HAYKIN, 2001)

Este modelo de rede tem treinamento baseado no algoritmo de retro propagação (*back-propagation*), um algoritmo de aprendizado supervisionado.

O algoritmo de treinamento é o responsável pelo ajuste dos pesos. No *back-propagation* o ajuste é feito através de um processo de correção do erro. Este é dado pela diferença entre a resposta da camada de saída da rede e a resposta desejada. A correção do erro é feita reajustando os pesos a cada entrada que é apresentada a rede (ZEIDENBERG, 1990).

2.6. Aplicações das redes neurais na área agronômica

Aplicações agronômicas de modelos de redes neurais incluem predição de produção em grãos (YANG, S., 2007), seleção em populações de *seedling* em cana-de-açúcar (ZHOU et al., 2011), previsão de características complexas em vacas da raça Jersey e trigo (GIANOLA et al., 2011), modelagem da resistência à penetração no solo (SANTOS, F. L., 2012), predição para valores genéticos de peso para bovinos da raça tabapuã (VENTURA et al., 2012), avaliação da adaptabilidade e estabilidade em genótipo de alfafa (NASCIMENTO et al., 2013) entre outros.

Yang (2007) mostra que, apesar de mais demorado para desenvolver comparado com modelos de regressão múltipla, os modelos de redes neurais são mais precisos para prever a produção de arroz nas condições climáticas específicas de Fujian na China.

Zhou et al. (2011) relata que modelos de redes neurais são superiores quando comparados ao método de seleção visual em identificar genótipos de cana-de-açúcar com alto potencial produtivo.

Em seu artigo sobre previsão de características complexas em vacas da raça Jersey e trigo, Gianola et al. (2011) mostra que modelos de redes neurais não lineares superaram modelos lineares na capacidade preditiva em ambos os conjuntos de dados, mas de forma mais clara no trigo.

Santos et al. (2012) em seu artigo sobre modelagem da resistência a penetração no solo mostra que a análise de regressão apresentou um coeficiente de determinação de 0,92, e a modelagem de redes neurais apresentou um coeficiente de determinação de 0,98. Os resultados evidenciam que a modelagem com redes neurais apresentou melhores resultados do que o modelo matemático obtido a partir da análise de regressão.

Ventura et al. (2012) relata que os valores genéticos de peso preditos para bovinos da raça tabapuã obtidos pela rede neural e os preditos pelo BLUP foram altamente correlacionados.

De acordo com Nascimento et al. (2013) a rede neural artificial foi capaz de classificar satisfatoriamente os genótipos de alfafa. Além disso, a análise apresentou alto índice de concordância com os resultados da análise de adaptabilidade e estabilidade segundo o método de Eberhart e Russell (1966).

3. Análise Discriminante

A análise discriminante consiste em obter funções que permitem classificar uma observação x, com base em medidas de um número p de características desta observação, em uma de várias populações $\prod_i$, com $i = 1,2,...,n$, distintas (KHATTREE e NAIK, 2000).

Esta classificação deve ser feita de maneira que a probabilidade de má classificação seja mínima, ou seja, a probabilidade de classificar erroneamente um indivíduo em uma população $\prod_i$, quando ele na verdade pertence à população $\prod_j$, com $i \neq j$, deve ser mínima (CRUZ e CARNEIRO, 2006).

Assim, por exemplo, no campo agronômico, a fim de diagnosticar se uma determinada família de plantas será selecionada para etapas posteriores de um programa de melhoramento genético, um pesquisador pode dispor de

uma série de variáveis tomadas numa família qualquer e com base nesses resultados procurar classificá-la como selecionada ou não selecionada.

O problema de discriminação entre duas ou mais populações passa pela obtenção de uma combinação linear das características observadas que apresenta maior poder de discriminação. A essa combinação dá-se o nome de Função Discriminante Linear de Fisher (RENCHER, 1995).

3.1. Classificação em uma de duas populações

3.1.1. Função Discriminante Linear de Fisher

Uma das soluções para a classificação de uma observação em uma de duas populações consiste em utilizar uma função que produza a separação máxima entre as duas populações (RENCHER, 1995).

Se considerarmos que o vetor de médias $\boldsymbol{\mu}_i$ multivariado e a matriz de covariância $\boldsymbol{\Sigma}$ comum das populações são conhecidos, demonstra-se que a função linear que produz separação máxima entre as populações é denominada Função Discriminante Linear de Fisher (FERREIRA, 2011) sendo escrita como:

$$D(\mathbf{X}) = [\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2]^t \boldsymbol{\Sigma}^{-1}\mathbf{X}.$$

Seja uma nova observação $\mathbf{x}_0$, o valor da função discriminante é

$$D(\mathbf{x}_0) = [\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2]^t \boldsymbol{\Sigma}^{-1}\mathbf{x}_0$$

e seja

$m = \frac{1}{2}[D(\boldsymbol{\mu}_1) + D(\boldsymbol{\mu}_2)]$ o ponto médio entre as duas médias populacionais univariadas $D(\boldsymbol{\mu}_1)$ e $D(\boldsymbol{\mu}_2)$.

A regra de classificação é:

$$\mathbf{x}_0 \in \prod_1 \text{ se } D(\mathbf{x}_0) = [\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2]^t \boldsymbol{\Sigma}^{-1}\mathbf{x}_0 \geq m$$

$$\mathbf{x}_0 \in \prod_2 \text{ se } D(\mathbf{x}_0) = [\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2]^t \boldsymbol{\Sigma}^{-1}\mathbf{x}_0 < m.$$

Na prática, é muito difícil conhecermos os vetores de médias e as matrizes de covariâncias. Assim, na maior parte dos casos é necessário obter os estimadores de $\boldsymbol{\mu}_i$ e $\boldsymbol{\Sigma}$. Felizmente, as regras de discriminação continuam

as mesmas descritas anteriormente, utilizando-se os estimadores ao invés dos parâmetros (FERREIRA, 2011).

Então, os parâmetros $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ e $\boldsymbol{\Sigma}$ serão substituídos pelos respectivos estimadores $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ e $\mathbf{S}_c$. Assim teremos a Função Discriminante Linear Amostral de Fisher escrita como:

$$D(\mathbf{x}) = \left[\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right]^t \mathbf{S}_c^{-1}\mathbf{x}$$

O ponto médio, m, entre as duas médias amostrais univariadas é dado por:

$$m = \frac{1}{2}\left(D(\bar{\mathbf{x}}_1) + D(\bar{\mathbf{x}}_2)\right).$$

A regra de classificação baseada nas amostras é apresentada por (FERREIRA, 2011):

$$\mathbf{x}_0 \in \prod_1 \text{ se } D(\mathbf{x}_0) = \left[\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right]^t \mathbf{S}_c^{-1}\mathbf{x}_0 \geq m$$

$$\mathbf{x}_0 \in \prod_2 \text{ se } D(\mathbf{x}_0) = \left[\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right]^t \mathbf{S}_c^{-1}\mathbf{x}_0 < m$$

4. Bibliografia

BARBOSA, M.H.P.; SILVEIRA, L.C.I. (2010) Melhoramento Genético e Recomendação de Cultivares. In: Santos, F.; Borém, A. e Caldas, C. Editores. Cana-de-açúcar: Bioenergia, Açúcar e Álcool - Tecnologias e Perspectivas. Viçosa, MG – Suprema, 578p.

BRAGA, A. P.; CARVALHO, A. C. P. L. F.; LUDERMIR, T. B. (2000) *Redes neurais artificiais: teoria e aplicações*. Rio de Janeiro, LTC – Livros Técnicos e Científicos, 262p.

CRUZ, C. D.; CARNEIRO, P. C. S. (2006) *Modelos Biométricos aplicados ao melhoramento genético*. Vol. 2, Viçosa, MG: UFV. 585p.

FERREIRA, D. F. (2011) *Estatística Multivariada*. Lavras, MG, Ed. UFLA. 675p.

GIANOLA, D.; OKUT, H.; KENT A. W.; ROSA, J. M. R. (2011) Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genetics*, 45:34.

GUPTA, M. M.; JIN, L.; HOMMA, N. (2003). *Static and Dynamic Neural Networks: From Fundamentals to Advanced Theory*. Wiley-IEEE Press. 752p.

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

JI, B.; SUN, Y.;YANG, S.; WAN, J. (2007) Artificial neural networks for rice yield prediction in mountainous regions. *Journal of Agricultural Science*, 145, 249–261.

KHATTREE, R.; NAIK, D. N. (2000) *Multivariate Data Reduction and Discrimination with SAS Software,* Cary, NC: SAS Institute Inc. 558p.

KIMBENG, C.A.; COX, M.C. (2003) Early generation selection of sugarcane families and clones in Australia: a review.*Journal American Society of Sugarcane Technologists*. 23:20-39.

LIPPMANN, R. (1987) An introduction to computing with neural nets. *ASSP Magazine*, v.4, n.2, p. 422.

MACKAY, D. J. C. (1994) Bayesian non-linear modelling for the prediction competition. In: *ASHRAE Transactions,* ASHRAE, Atlanta Georgia. Vol. 100, pp. 1053–1062.

MCCULLOCH, W. S.; PITTS, W. H. (1943) A logical calculus of the ideas immanent in nervous activity.*Bulletin of Mathematical Biophysics*, v. 5, p. 115-133.

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

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

RENTCHER, A. C. (1995) *Methods of multivariate analysis*. John Wiley e Sons, INC. 738p.

RESENDE, M.D.V. de. (2002) *Genética biométrica e estatística no melhoramento de plantas perenes*. Brasília: Embrapa Informação Tecnológica, 975p.

RESENDE, M.D.V.; BARBOSA, M.H.P. (2006) Selection via simulated Blup base on family genotypic effects in sugarcane. *Pesquisa Agropecuária Brasileira*, 41(3):421-429.

SANTOS, F. L.; JESUS, V. A. M., VALENTE, D. S. M. (2011) Modeling of soil penetration resistance using statistical analyses and artificial neural networks. *Acta Scientiarium*, 34(2):219-224.

STRINGER, J.K.; COX, M.C.; ATKIN, F.C.; WEI, X.; HOGARTH. (2011) Family Selection Improves the Efficiency and Effectiveness of Selecting Original Seedlings and Parents. *Sugar Tech*, 13(1):36-41.

VENTURA, R. V.; SILVA, A. M.; MEDEIROS, T. H.; DIONELLO, N. L; MADALENA, F.E.; FRIDRICH, A.B.; VALENTE, B.D.; SANTOS, G.G.; FREITAS, L.S.; WENCESLAU, R.R.; FELIPE, V. P. S.; CORRÊA, G.S.S. (2012) Uso de redes neurais artificais na predição de valores genéticos para peso aos 205 dias em bovinos da raça Tabapuã. *Arq. Bras. Med. Vet. Zootec.*, 64(2): 411-418.

ZEIDENBERG, M. (1990) *Neural Networks in Artificial Inteligence.* Ellis Horwood Series in Artificial Intelligence, 268p.

ZHOU, M. M.; TEW, L.T.; KIMBENG, A. C.; GRAVOIS, A. K.; PONTIF, M. J. (2011) Artificial Neural Network Models as a Decision Support Tool for Selection in Sugarcane: A Case Study Using Seedling Populations. *Crop Science*, 12:87.

CAPÍTULO 2

Seleção entre famílias de cana-de-açúcar via redes neurais artificiais

Resumo

A base para aumento da produtividade em cana-de-açúcar é o melhoramento genético. A etapa mais importante do programa é a fase inicial, chamada de T1, uma vez que é nesta que são selecionados os primeiros genótipos. Nesta fase a seleção de famílias pode ser preferida quando a seleção é praticada com base em caracteres indiretos. O objetivo deste trabalho é utilizar modelos de redes neurais para seleção de famílias promissoras em cana-de-açúcar com base nos caracteres indiretos altura de colmos (AC), diâmetro de colmos (DC) e número de colmos (NC) – variáveis de entrada da rede – e o resultado do processo de seleção via TCHr (Tonelada de Cana por Hectare real), utilizada como variável de saída, e compará-los com a seleção de famílias acima da média geral para a variável tonelada de cana por hectare estimada (TCHe). O conjunto de dados foi dividido em dois subconjuntos, um para treinamento e outro para teste da rede. A análise foi feita em 4 diferentes cenários: com simulação e com padronização das variáveis; com simulação e sem padronização das variáveis ; sem simulação e com padronização das variáveis; e sem simulação e sem padronização das variáveis. A rede usada neste trabalho é uma rede de múltiplas camadas (*Multilayer Perceptron* - MLP) com uma camada intermediária entre a camada de entrada e a camada de saída. Para avaliação dos métodos foi utilizada a taxa de erro aparente (TEA) e a taxa de erro 1 (TE1). Os piores resultados ocorrem no cenário 4, sem padronização e sem simulação e os melhores ocorreram no cenário 1, onde as variáveis foram padronizadas e foram simulados valores de DC, NC, AC e TCHR para 1000 famílias. Isso mostra que a simulação e a padronização melhoram o desempenho dos modelos de redes neurais. A modelagem via redes neurais, usando as ferramentas adequadas (simulação e padronização), fornece baixas taxas de erro aparente podendo ser utilizadas na seleção de materiais promissores em cana-de-açúcar.

1. Introdução

A base para aumento da produtividade em cana-de-açúcar é o melhoramento genético. O programa de melhoramento genético de cana-de-açúcar envolve cinco fases, três fases de teste (T1, T2 e T3), uma fase experimental e uma fase de multiplicação (BARBOSA e SILVEIRA, 2010). Uma das etapas mais importante do programa é a fase inicial, chamada de T1, uma vez que é nesta que são selecionados os primeiros genótipos (OLIVEIRA et al., 2011). Uma seleção ineficiente nessa fase pode comprometer o sucesso de todo o programa.

De maneira geral, nas fases iniciais, a seleção de famílias pode ser preferida quando a seleção é praticada com base em caracteres indiretos (FALCONER e MACKAY, 1996). Durante a seleção das melhores famílias, diâmetro dos colmos (DC), número de colmos (NC) e altura de colmos (AC) são comumente usados para avaliar o rendimento de cana-de-açúcar (CHANG e MILLIGAN, 1992).

Usualmente, o que vem sendo realizado nos programas de melhoramento de cana é a seleção massal, ou seja, uma seleção individual de clones (MATSUOKA et al., 2005). No entanto, esse método aplicado nos estágios iniciais dos programas tem se mostrado ineficiente.

No que se refere aos métodos de estatística e genética utilizados para a seleção de materiais superiores em cana-de-açúcar na fase T1 destacam-se o BLUP (*Best Linear Unbiased Predictor*) individual (BLUPI) (RESENDE, 2002) e o BLUP individual simulado (BLUPIS) (RESENDE e BARBOSA, 2006). Para o caso de dados balanceados, como neste trabalho, a seleção pode ser simplificada. Neste caso, basta selecionar aquelas com tonelada de cana por hectare real (TCHr) acima da média geral do experimento.

No entanto, para o uso destas metodologias, é necessário pesar de forma integral todas as parcelas dos experimentos, restringindo a seleção dentro das melhores famílias na fase de cana-soca e o número de famílias a serem avaliadas por vez.

A alternativa que tem sido utilizada é obter a variável tonelada de cana por hectare estimada (TCHe) em função das variáveis NC, DC e AC (CHANG e MILLIGAN, 1992), e assim selecionar aquelas com TCHe acima da média geral fenotípica.

Um novo método que tem sido utilizado, ainda de forma sútil, em programas de melhoramento vegetal, é o uso de redes neurais artificiais. Essa abordagem é baseada em princípios da inteligência artificial.

O principio básico de uma rede neural é que fornecendo exemplos do relacionamento entre variáveis de entrada e saída, pode-se fazer com que a rede neural “aprenda” a relacionar essas variáveis. Essa estrutura de modelagem flexível permite que sejam descritas relações mais complexas entre as variáveis do que as obtidas pelos modelos tradicionais (MACKAY, 1994).

Assim, podemos utilizar as variáveis DC, NC e AC como variáveis de entrada da rede e o resultado do processo de seleção via TCHr como variável de saída, a fim de selecionar as melhores famílias com base em características indiretas contornando assim o problema de pesagem no campo e otimizando o processo de seleção de famílias promissoras em cana-de-açúcar.

O objetivo deste trabalho é utilizar modelos de redes neurais para seleção de famílias promissoras em cana-de-açúcar com base nos componentes de rendimento NC, DC e AC e compará-los com a seleção de famílias acima da média geral para a variável TCHe.

2. Material e métodos

2.1. Material vegetal e conjunto de dados

Foram utilizados dados de cinco experimentos conduzidos no Centro de Pesquisa e Melhoramento de Cana-de-açúcar (CECA), da Universidade Federal de Viçosa, localizado no município de Oratórios, Minas Gerais, com latitude 20°25’S; longitude 42°48’W. Em cada experimento foram alocadas 22 famílias de cana-de-açúcar, sendo utilizado o delineamento em blocos casualizados com cinco repetições. A unidade experimental foi constituída por 20 plantas, distribuídas em dois sulcos de 5 m de comprimento, espaçados em 1,40 m.

Os seguintes caracteres foram avaliados: altura de colmos (AC) em metros, mensurando-se um colmo de cada touceira, desde a base até o primeiro *dewlap* visível; diâmetro de colmos (DC) em centímetros, medido com paquímetro digital no terceiro internódio, contado da base do colmo para o

ápice, tomando o mesmo colmo onde foi coletada a altura; número total de colmos por parcela (NC) e tonelada de cana por hectare real (TCHr) obtida com a pesagem de cada parcela.

Para seleção das melhores famílias foi obtida a variável de interesse tonelada de colmos por hectare real (TCHr) sendo selecionadas as famílias com médias superiores a média geral do experimento.

Tendo em vista que 110 famílias poderiam não ser suficientes para a obtenção de um modelo de rede neural com ampla capacidade de generalização, foram simulados valores de DC, NC, AC e TCHr para 1000 famílias. Os valores foram simulados usando a estrutura de média e de covariância de cada um dos cinco experimentos separadamente e a modelagem por redes neurais foi feita com e sem simulação.

A simulação foi feita utilizando a função mvrnorm do pacote MASS (VENABLES e RIPLEY, 2002) implementado no software R (R Core Team, 2013).

Desde que a escala das variáveis de entrada determinam a escala das estimativas dos pesos na camada de saída, isso pode acarretar num grande efeito na qualidade da solução final. A princípio, o melhor é padronizar todas as entradas para a média zero e desvio padrão um. Isso permite tratar da mesma forma todas as entradas no processo de regularização, e permite escolher um intervalo significativo para os pesos iniciais (HASTIE et al., 2009). Assim, para efeito de comparação, decidiu-se por realizar a modelagem via redes neurais com as variáveis de entrada (NC, AC e DC) na escala original e padronizada.
A padronização foi feita utilizando a seguinte expressão:

$$Z_i = \frac{V_i - \overline{V}}{Sd(V)}$$

onde:

Z_i é o i-ésimo valor da variável padronizada;

V_i é o i-ésimo valor da variável a ser padronizada;

$\overline{V}$ é a média geral da variável a ser padronizada;

$Sd(V)$ é o desvio-padrão da variável a ser padronizada.

O conjunto de dados compreendeu, então, valores de DC, AC, NC e o resultado do processo de seleção via TCHr em 4 diferentes cenários: com simulação e com padronização, com simulação e sem padronização, sem

simulação e com padronização, e sem simulação e sem padronização (Tabela 1).

2.2. Modelagem via redes neurais artificiais

Para modelagem usando rede neural, dois conjuntos são necessários: um conjunto para treinamento da rede e outro conjunto para o teste da rede. Estes conjuntos, no presente trabalho, variam de acordo com o cenário utilizado como mostra a Tabela 1.

Tabela 1: Conjunto de treinamento e teste para quatro diferentes cenários definidos em função da realização ou não de simulação e padronização.

Cenários	Padronização	Simulação	Exp. trein	Exp. teste
1	Sim	Sim	Exp. 1	Exp. 2,3,4,5
1	Sim	Sim	Exp. 2	Exp. 1,3,4,5
1	Sim	Sim	Exp. 3	Exp. 1,2,4,5
1	Sim	Sim	Exp. 4	Exp. 1,2,3,5
1	Sim	Sim	Exp. 5	Exp. 1,2,3,4
2	Sim	Não	Exp. 1	Exp. 2,3,4,5
2	Sim	Não	Exp. 2	Exp. 1,3,4,5
2	Sim	Não	Exp. 3	Exp. 1,2,4,5
2	Sim	Não	Exp. 4	Exp. 1,2,3,5
2	Sim	Não	Exp. 5	Exp. 1,2,3,4
3	Não	Sim	Exp. 1	Exp. 2,3,4,5
3	Não	Sim	Exp. 2	Exp. 1,3,4,5
3	Não	Sim	Exp. 3	Exp. 1,2,4,5
3	Não	Sim	Exp. 4	Exp. 1,2,3,5
3	Não	Sim	Exp. 5	Exp. 1,2,3,4
4	Não	Não	Exp. 1	Exp. 2,3,4,5
4	Não	Não	Exp. 2	Exp. 1,3,4,5
4	Não	Não	Exp. 3	Exp. 1,2,4,5
4	Não	Não	Exp. 4	Exp. 1,2,3,5
4	Não	Não	Exp. 5	Exp. 1,2,3,4

* Exp. trein = Experimento de treinamento; Exp.test= Experimento de teste

Assim, no cenário 1, o conjunto de treinamento corresponde às 22 famílias de cada um dos cinco experimentos acrescido de 1000 famílias simuladas com o vetor de médias e a matriz de covariâncias do experimento correspondente. O conjunto de teste corresponde às famílias dos quatro

experimentos restantes. A leitura para os demais cenários é feita da mesma forma.

A rede usada neste trabalho é uma rede de múltiplas camadas (*Multilayer Perceptron* - MLP) com uma camada intermediária entre a camada de entrada e a camada de saída (HASTIE et al., 2009) como mostra a Figura 4.

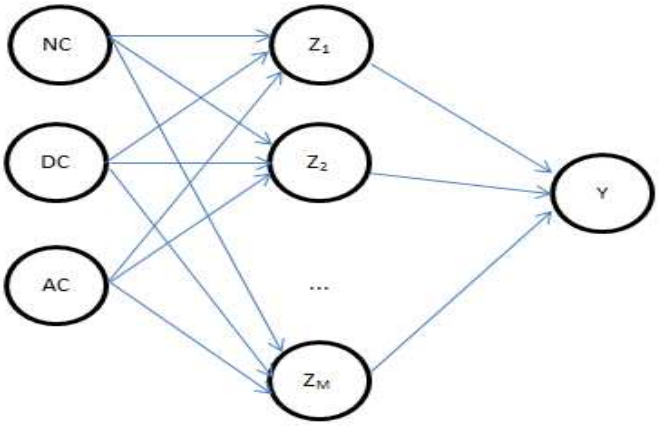


Figura 4. Esquema da rede MLP com uma camada intermediária.

Matematicamente, temos que NC, DC e AC são as variáveis de entrada. Z_m, onde $m = 1,2,...,M$, são funções responsáveis pela soma ponderada das entradas, esta ponderação é feita de acordo com os parâmetros da rede W_i, com $i = 1,2,...,I$. Y_k, com $k = 1,2,...,K$, é a saída da rede, ou seja, o resultado do processo de seleção via rede neural (HASTIE et al, 2009).

A função utilizada foi a função sigmoide, dada por:

$$y(x;w) = \frac{1}{1+e^{-wx}} \ .$$

Os parâmetros da rede, ou pesos, são estimados pela minimização da soma de quadrados dos erros, escrita como

$$R(\theta) = \sum_{k=1}^{K}\sum_{i=1}^{N}\left[y_{ik} - f_k(x_i)\right]^2 \ .$$

A minimização é feita pela aplicação de um algoritmo de gradiente decrescente conhecido como *back-propagation* (HASTIE et al., 2009).

O processo de treinamento da rede consiste em fornecer valores de entrada e saída para que a rede “aprenda” a relacioná-los e com isso estimar os parâmetros. No entanto, para iniciar o processo de treinamento é necessário fornecer valores iniciais para os parâmetros. Segundo Venables e Ripley

(2002) os valores iniciais podem ser escolhidos ao acaso em um intervalo que satisfaça a seguinte equação $LS * máx(|X|) \approx 1$, onde LS é o limite superior do intervalo e $máx(|X|)$ é o maior valor em módulo do conjunto de treinamento.

Após treinada a rede temos a fase de teste. Nesta fase a rede foi aplicada ao conjunto de dados de teste e o resultado foi confrontado com o resultado obtido pelo seleção das famílias com médias de TCHr acima da média geral.

A análise via redes neurais foi feita utilizando a função nnet do pacote nnet (VENABLES e RIPLEY, 2002) implementado no software R (R Core Team, 2013).

2.3. Seleção via tonelada de colmos por hectare estimada

Para os cenários onde foi feita a modelagem via redes neurais também foi feita a seleção das melhores famílias selecionando aquelas com tonelada de colmos por hectare estimada (TCHe) acima da média geral.

A variável TCHe pode ser obtida através da seguinte expressão (FERREIRA et al., 2007):

$$TCHe = d \times \pi \times NC \times AC \times \left(\frac{DC}{2}\right)^2 \times \frac{1}{100 \times tp}$$

onde:

d é a densidade específica do colmo. Chang e Milligan (1992) sugerem utilizar para d o valor de 1000 kg m^{-3};

tp é o tamanho da parcela em m^2.

2.4. Avaliação e comparação das técnicas

Para avaliação dos classificadores – rede neural artificial e seleção via TCHe – gerou-se matrizes de confusão para cada um dos cenários em ambos os métodos. A matriz de confusão oferece uma medida efetiva do classificador utilizado ao mostrar o número de classificações corretas usando a variável TCHr *versus* o número de classificações preditas pelo classificador considerado para cada classe sobre o conjunto de teste.

A seguir é apresentada o esquema da matriz de confusão para o classificador rede neural artificial tendo em vista que o raciocínio é o mesmo considerando a seleção via TCHe (Tabela 2).

Tabela 2: Matriz de confusão entre as estratégias de seleção via TCHr e redes neurais, acompanhada das taxas de erro de classificação.

Classificação	NS (Rede)	S (Rede)
NS (TCHr)	VN	FP
S (TCHr)	FN	VP

*S= Famílias selecionadas, NS= Famílias não selecionadas, VP= Verdadeiro Positivo, FN= Falso Negativo, FP= Falso Positivo, VN= Verdadeiro Negativo

É fácil perceber que VP indica o número de famílias selecionados por ambos os métodos, FN corresponde ao número de famílias selecionados via TCHr que não foram selecionadas pela rede neural, FP indica o número de famílias não selecionadas via TCHr que foram selecionadas pela rede neural e VN corresponde ao número de famílias não selecionadas por ambos.

Várias são as medidas que se originam da matriz de confusão para avaliação do classificador. Especificamente para seleção de famílias em cana-de-açúcar estamos interessados, principalmente, nas taxas de erros, que trataremos aqui de taxa de erro 1 (TE1) e taxa de erro aparente (TEA).

A TE1 é escrita como:

$$\mathrm{TE1} = \frac{\mathrm{FP}}{\mathrm{N}}$$

onde:

N é o número de famílias da população de validação.

A TEA é dada por:

$$\mathrm{TEA} = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{N}}$$

onde:

N é o número de famílias da população de validação

3. Resultados e discussão

A taxa de erro aparente (TEA) e a taxa de erro 1 (TE1) via modelagem por redes neurais e via tonelada de colmos por hectare estimada (TCHe), o conjunto de treinamento e o conjunto de teste para os quatro diferentes cenários analisados encontram-se na Tabela 3.

Tabela 3: Taxa de erro aparente via rede neural (TEA-RN), taxa de erro 1 via rede neural (TE1-RN), taxa de erro aparente via tonelada de colmos por hectare estimada (TEA-TCHe) e taxa de erro 1 via tonelada de colmos por hectare estimada (TE1-TCHe) para os quatro diferentes cenários avaliados.

Cenário	Padr.	Simul.	Exp. trein	Exp. teste	TEA-RN	TE1-RN	TEA-TCHe	TE1-TCHe
1	Sim	Sim	1	2,3,4,5	0.10227	0.07954	0.17045	0.09091
1	Sim	Sim	2	1,3,4,5	0.11363	0.06818	0.18182	0.09091
1	Sim	Sim	3	1,2,4,5	0.14773	0.09091	0.14773	0.06818
1	Sim	Sim	4	1,2,3,5	0.125	0.09091	0.15909	0.07954
1	Sim	Sim	5	1,2,3,4	0.09091	0.06818	0.15909	0.07954
2	Sim	Não	1	2,3,4,5	0.14772	0.09091	0.17045	0.09091
2	Sim	Não	2	1,3,4,5	0.13636	0.02273	0.18182	0.09091
2	Sim	Não	3	1,2,4,5	0.125	0.07954	0.14773	0.06818
2	Sim	Não	4	1,2,3,5	0.13636	0.07954	0.15909	0.07954
2	Sim	Não	5	1,2,3,4	0.125	0.06818	0.15909	0.07954
3	Não	Sim	1	2,3,4,5	0.13636	0.125	0.17045	0.09091
3	Não	Sim	2	1,3,4,5	0.125	0.10227	0.18182	0.09091
3	Não	Sim	3	1,2,4,5	0.15909	0.02273	0.14773	0.06818
3	Não	Sim	4	1,2,3,5	0.125	0.11363	0.15909	0.07954
3	Não	Sim	5	1,2,3,4	0.125	0.09091	0.15909	0.07954
4	Não	Não	1	2,3,4,5	0.18182	0.10227	0.17045	0.09091
4	Não	Não	2	1,3,4,5	0.18182	0.06818	0.18182	0.09091
4	Não	Não	3	1,2,4,5	0.19318	0.02273	0.14773	0.06818
4	Não	Não	4	1,2,3,5	0.21591	0.14773	0.15909	0.07954
4	Não	Não	5	1,2,3,4	0.21591	0.09091	0.15909	0.07954

* Padr.= Padronização; Simul.= Simulação; Exp. trein= Experimento de treinamento; Exp. test= Experimentos de teste.

Percebe-se que a taxa de erro aparente via redes neurais (TEA-RN) é baixa nos quatro cenários (Tabela 3). Isso mostra que há uma alta concordância entre a modelagem via redes neurais e o método de seleção via TCHr. É importante ressaltar ainda que grande parte da TEA-RN deve-se a

taxa de erro 1 via redes neurais (TE1-RN), que não representa um problema na prática. Selecionar famílias que não deveriam ter sido selecionadas não é um problema, pois esses materiais poderão ser eliminados em seleções posteriores.

A grande vantagem de utilizar modelos de redes neurais é que seria necessária a pesagem de apenas uma pequena parte do material. Neste estudo, com a pesagem de apenas um dos cinco experimentos, houve uma excelente generalização para os quatro experimentos restantes. Assim, é evidente que tal estratégia poderia reduzir consideravelmente o trabalho no campo, otimizando o processo de seleção.

As maiores taxas de erro aparente ocorrem no cenário 4, sem padronização e sem simulação, e as menores ocorreram no cenário 1, onde as variáveis foram padronizadas e foram simulados valores de DC, NC, AC e TCHR para 1000 famílias. Isso mostra que a simulação e a padronização melhoram o desempenho dos modelos de redes neurais.

A simulação é importante porque o conjunto de dados original para o treinamento é muito pequeno, e assim o aprendizado da rede fica comprometido. Nascimento (2013) em estudos de adaptabilidade e estabilidade em genótipos de alfafa também faz uso da simulação para compor o conjunto de treinamento, obtendo bons resultados.

A padronização das variáveis melhora o desempenho dos modelos de redes neurais, pois elimina o efeito da escala nos parâmetros da rede, aqui chamados de pesos, como também observado por Hastie (2009).

Vale ressaltar que as taxas de erro aparente em cada cenário são próximas, independente do experimento utilizado para simulação e treinamento.

Pode ser observado que os modelos de redes neurais tem desempenho superior à estratégia usando a variável TCHe, comumente usada nos programas de melhoramento de cana-de-açúcar devido à facilidade operacional, nos cenários 1, 2 e 3, onde foi feita simulação e/ou padronização (Tabela 3). No cenário 4, onde são utilizados os dados originais, a estratégia usando a variável TCHe é superior. O resultado acima pode ser melhor visualizado na Figura 5.

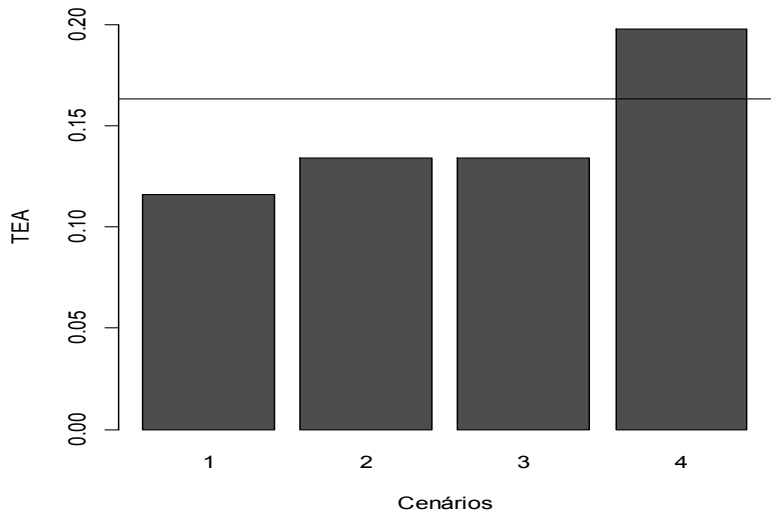


Figura 5: Valores médios para a taxa de erro aparente (TEA) via modelagem por redes neurais nos quatros cenários (1: com padronização e com simulação; 2: com padronização e sem simulação; 3: sem padronização e com simulação; 4: sem padronização e sem simulação) e via tonelada de colmos por hectare estimado (TCHe), indicada pela linha horizontal.

Comparativamente, a taxa média de erro aparente obtido no cenário 1, onde ambas as estratégias, simulação e padronização, foram utilizadas, foi de 0.1159, o que representa uma concordância de 88.41% com o método de seleção via TCHr. No cenário 4, onde nenhuma das estratégias foi utilizada a taxa média de erro aparente foi de 0.1977, que corresponde a uma concordância de 80.23% (Figura 5). Nos cenário 2 e 3, onde apenas uma das duas estratégias foi utilizada , a taxa média de erro aparente foi 0.1341, o que corresponde uma taxa média de 86.59% de concordância com o método ideal (Figura 5).

Isso mostra que a modelagem via redes neurais, usando as ferramentas de simulação e padronização, pode ser utilizada com sucesso na seleção das melhores famílias em cana-de-açúcar.

Se usarmos a diferença entre a taxa de erro aparente e a taxa de erro 1, medida correspondente ao erro que seria um problema no critério de seleção, para comparar a rede neural artificial com a seleção usando a variável TCHe é ainda mais evidente que a rede neural fornece melhores resultados , como mostra a figura 6.

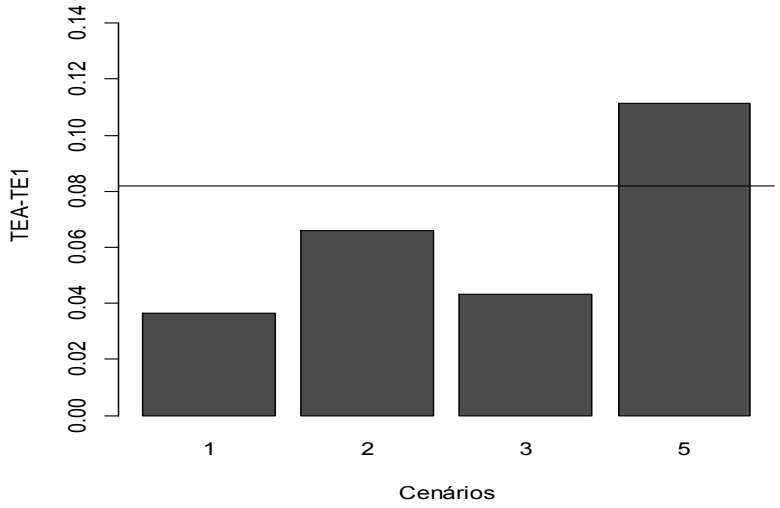


Figura 6: Valores médios para a diferença entre a taxa de erro aparente (TEA) e a taxa de erro 1 (TE1) via modelagem por redes neurais nos quatros cenários (1: com padronização e com simulação; 2: com padronização e sem simulação; 3: sem padronização e com simulação; 4: sem padronização e sem simulação) e via tonelada de colmos por hectare estimado (TCHe), indicada pela linha horizontal.

Adicionalmente, podemos perceber que os cenários onde foi feita simulação para compor o conjunto de treinamento são os que fornecem menor diferença média entre a TEA e a TE1.

Apesar dos bons resultados obtidos pela modelagem via redes neurais, outros estudos são necessários para avaliar a eficácia real da técnica. Estes estudos esclareceriam, por exemplo, se a modelagem via redes neurais é mais eficiente que outras técnicas que também podem ser usadas para classificação, como a análise discriminante e a regressão logística (HAIR et al., 2007).

Outra dúvida que surge é se os modelos de redes neurais continuariam tendo desempenho superior para taxas menores de seleção. Neste estudo a taxa de seleção é de aproximadamente 50 % considerando distribuição normal, uma vez que são selecionadas as famílias que apresentem valores de TCHr acima da média geral do experimento.

4. Conclusões

Os modelos de redes neurais apresentam alta concordância com a seleção via tonelada de cana por hectare real.

A modelagem via redes neurais, usando as ferramentas de simulação e/ou padronização, fornece melhores resultados quando comparada a estratégia usualmente utilizada baseada na seleção via tonelada de cana por hectare estimada.

5. Bibliografia

CHANG, Y.S.; MILLIGAN S.B. (1992) Estimating the potential of sugarcane families to procude elite genotypes using univariate cross prediction methods. Theoretical and Applied Genetics. Berlin, 84: 662-671.

FALCONER D.S.; MCKAY T.F.C. (1996) Introduction to Quantitative Genetics. Malaysia: Longmans Green, 463p.

FERREIRA, F. M.; BARROS, W. S.; SILVA, F. L.; BARBOSA, M. H. P.; CRUZ, C. D.; BASTOS, I. T. (2007) Relações fenotípicas e genotípicas entre componentes de produção em cana-de-açúcar. Bragantia, 66(4).

HAIR J. F.; BLACK W.; ANDERSON R. E.; TATHAM R. L. (2007) Análise Multivariada de Dados, 6ª edição, BOOKMAN, 593p.

HASTIE T.; TIBSHIRANI R.; FRIEDMAN J. (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York, 745p.

MACKAY, D. J. C. (1994) Bayesian non-linear modelling for the prediction competition. In: ASHRAE Transactions, ASHRAE, Atlanta Georgia. Vol. 100, p. 1053–1062.

MATSUOKA, S.; GARCIA, A.A.F.; ARIZONO, H. (2005) Melhoramento da cana-de-açúcar. In: BORÉM, A. (Ed.) Melhoramento de espécies cultivadas. Viçosa: Ed. da UFV. 969p.

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

R Development Core Team (2010) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (http://www.r-project.org).

RESENDE, M.D.V.; BARBOSA, M.H.P. (2006) Selection via simulated Blup base on family genotypic effects in sugarcane. Pesquisa Agropecuária Brasileira, 41(3):421-429.

VENABLES W.N.; RIPLEY B.D. (2002) Modern applied statistics with s. Springer, New York, 493p.

.

CAPÍTULO 3

Comparação entre Redes Neurais e Análise Discriminante como alternativas para seleção de famílias em cana-de-açúcar

Resumo

Um dos grandes desafios nos programas de melhoramento genético de cana-de-açúcar é a seleção eficiente de genótipos nas fases iniciais. Esse desafio advém da grande quantidade de genótipos avaliados e da dificuldade operacional da pesagem das parcelas do experimento, necessária nos principais métodos de seleção. O objetivo deste trabalho é comparar a modelagem por redes neurais, a análise discriminante linear de Fisher e a seleção de famílias usando a variável tonelada de cana por hectare estimada (TCHe) como alternativas para seleção de famílias promissoras em cana-de-açúcar com base nos caracteres indiretos número de colmos (NC), diâmetro de colmos (DC) e altura de colmos (AC). Para comparação e avaliação dos métodos empregados foi utilizada a taxa de erro aparente (TEA) e a taxa de erro 1 (TE1) obtidas a partir da matriz de confusão. A modelagem via redes neurais artificiais e a análise discriminante linear de Fisher fornecem melhores resultados quando comparadas a estratégia usualmente utilizada, que é baseada na estimação da variável tonelada de cana por hectare. Comparando os modelos de redes neurais com a análise discriminante, a rede neural fornece melhores resultados.

.

1. Introdução

O melhoramento genético é a base de todo o agronegócio da cana-de-açúcar. As cultivares melhoradas permitem o aumento da produtividade em cana-de-açúcar e a melhoria da matéria prima para a fabricação de açúcar e álcool (BARBOSA et al., 2006).

A etapa mais importante nos programas de melhoramento genético de cana-de-açúcar é a fase inicial, chamada de T1 (Oliveira et al., 2011). Nesta são selecionados os primeiros materiais, de maneira que, uma seleção mal feita pode comprometer o sucesso de todo o programa (BARBOSA e SIIVEIRA, 2010).

De maneira geral, nas fases iniciais, a seleção de famílias pode ser preferida quando a seleção é praticada com base em caracteres indiretos (FALCONER e MACKAY, 1996). O diâmetro dos colmos (DC), o número de colmos (NC) e a altura de colmos (AC) são os caracteres indiretos comumente usados para avaliar o rendimento em cana-de-açúcar e selecionar as melhores famílias (CHANG e MILLIGAN, 1992).

No que se refere aos métodos de estatística e genética utilizados para a seleção de materiais superiores em cana-de-açúcar na fase T1 destacam-se o BLUP (*Best Linear Unbiased Predictor*) individual (BLUPI) (RESENDE, 2002) e o BLUP individual simulado (BLUPIS) (RESENDE e BARBOSA, 2006). No entanto estes procedimentos têm sido usados com restrição nos programas de melhoramento de cana-de-açúcar uma vez que necessitam da pesagem de todas as parcelas do experimento para a obtenção da variável de interesse tonelada de colmos por hectare real (TCHr), o que é um grande problema operacional.

A alternativa que tem sido utilizada é obter a variável tonelada de cana por hectare estimada (TCHe) em função das variáveis NC, DC e AC e assim selecionar as de TCHe acima da média geral fenotípica.

Um novo método que tem sido utilizado em diferentes programas de melhoramento genético é o uso de redes neurais artificiais (GIANOLA, 2011; VENTURA, 2012; NASCIMENTO, 2013).

O principio básico de uma rede neural artificial é que fornecendo exemplos do relacionamento entre variáveis de entrada e saída, nós podemos fazer com que a rede neural “aprenda” a relacionar essas variáveis (MACKAY,

1994). Assim, podemos utilizar as variáveis DC, NC e AC como variáveis de entrada da rede e o resultado do processo de seleção via TCHr como variável de saída, a fim de selecionar as melhores famílias.

No entanto, para verificar a eficiência real desta técnica convém compará-la com outras metodologias como por exemplo a análise discriminante.

Semelhantemente, na análise discriminante linear, fornecemos exemplos de características de duas populações ou grupos conseguimos estabelecer uma função que produza a máxima separação entre as populações (CRUZ e CARNEIRO, 2006). Então, podemos fornecer as variáveis DC, AC e NC das famílias selecionadas e não selecionadas com base na seleção via TCHr e estabelecer uma função discriminante linear de Fisher para alocar novas observações em um desses grupos, ou seja, para classificar novas famílias em selecionadas ou não selecionadas.

O objetivo deste trabalho é comparar a modelagem por redes neurais e a análise discriminante linear de Fisher após simulação e padronização das variáveis de entrada como alternativas para seleção entre famílias em cana-de-açúcar.

2. Material e métodos

2.1. Material vegetal e conjunto de dados

Os dados são provenientes de cinco experimentos conduzidos no Centro de Pesquisa e Melhoramento de Cana-de-açúcar (CECA), da Universidade Federal de Viçosa, localizado no município de Oratórios, Minas Gerais, com latitude 20°25'S; longitude 42°48'W. Os Experimentos foram montados em blocos casualizados com 5 repetições e 22 famílias cada. A unidade experimental foi constituída por 20 plantas, distribuídas em dois sulcos de 5 m de comprimento, espaçados em 1,40 m.

Os seguintes caracteres foram avaliados: altura de colmos (AC) em metros, mensurando-se um colmo de cada touceira, desde a base até o primeiro *dewlap* visível; diâmetro de colmos (DC) em centímetros, medido com paquímetro digital no terceiro internódio, contado da base do colmo para o

ápice; número total de colmos por parcela (NC) e tonelada de colmos por hectare real (TCHR) medida pesando a parcela.

Para seleção das melhores famílias foram pesadas todas as famílias de cada um dos cinco experimentos sendo selecionadas aquelas que apresentavam TCHr acima da média geral do experimento.

Foram simulados valores de NC, AC, DC e TCHr para 1000 famílias tomando como base a estrutura de médias e covariâncias de cada um dos cinco experimentos separadamente. A simulação foi feita utilizando a função mvrnorm do pacote MASS (VENABLES e RIPLEY, 2002) implementado no software R (R Core Team, 2013).

As variáveis NC, AC e DC foram padronizadas para a média zero e variância um. A padronização foi feita utilizando a seguinte expressão:

$$Z_i = \frac{V_i - \overline{V}}{Sd(V)}$$

onde:

Z_i é o i-ésimo valor da variável padronizada;

V_i é o i-ésimo valor da variável a ser padronizada;

$\overline{V}$ é a média geral da variável a ser padronizada;

$Sd(V)$ é o desvio-padrão da variável a ser padronizada.

O conjunto de dados compreendeu, então, valores de DC, AC e NC padronizados e o resultado do processo de seleção via TCHr. Para o treinamento dos métodos foram usados dados simulados e o experimento correspondente utilizado como base para a simulação. Na fase de teste foram usados os dados originais.

2.2. Modelagem via redes neurais

Para modelagem usando rede neural, dois conjuntos são necessários: um conjunto para treinamento e outro conjunto para o teste da rede. O primeiro conjunto destina-se ao ajuste de pesos sinápticos, e o segundo à fase de avaliação e desempenho da rede.

De maneira geral, o conjunto de treinamento corresponde sempre ao experimento que foi utilizado para a simulação acrescido das 1000 famílias simuladas. O conjunto de teste é constituído pelos experimentos restantes.

A rede usada neste trabalho é uma rede de múltiplas camadas (*Multilayer Perceptron* - MLP) com uma camada intermediária entre a camada de entrada e a camada de saída (HASTIE et. al., 2009) como mostra a Figura 7.

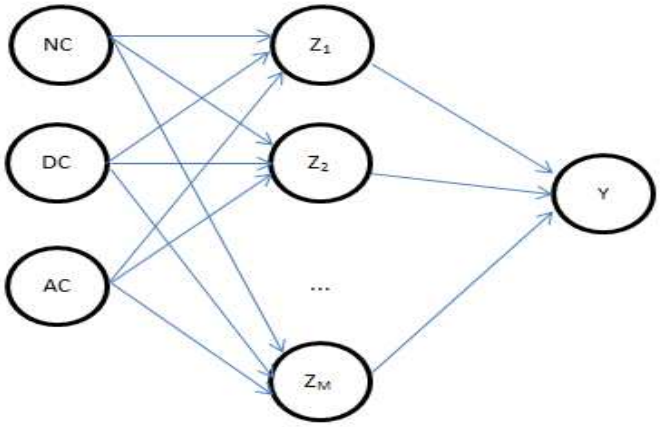


Figura 7. Esquema de uma rede MLP com uma camada intermediária.

Matematicamente, temos que NC, DC e AC são as variáveis de entrada. Z_m, onde $m = 1,2,...,M$, são funções responsáveis pela soma ponderada das entradas, esta ponderação é feita de acordo com os parâmetros da rede W_i, com $i = 1,2,...,I$. Y_k, com $k = 1,2,...,K$, é a saída da rede, ou seja, o resultado do processo de seleção via rede neural (HASTIE et al, 2009).

A função utilizada foi a função sigmóide, dada por:

$$y(x;w) = \frac{1}{1+e^{-wx}} \ .$$

Os parâmetros da rede, ou pesos, são estimados pela minimização da soma de quadrados dos erros, escrita como

$$R(\theta) = \sum_{k=1}^{K}\sum_{i=1}^{N}\left[y_{ik} - f_k(x_i)\right]^2 \ .$$

A minimização é feita pela aplicação de um algoritmo de gradiente decrescente conhecido como *back-propagation* (HASTIE et al., 2009).

O processo de treinamento da rede consiste em fornecer valores de entrada e saída para que a rede “aprenda” a relacioná-los e com isso estimar

os parâmetros. No entanto, para iniciar o processo de treinamento é necessário fornecer valores iniciais para os parâmetros. Segundo Venables e Ripley (2002) os valores iniciais podem ser escolhidos ao acaso em um intervalo que satisfaça a seguinte equação $LS * máx(|X|) \approx 1$, onde LS é o limite superior do intervalo e $máx(|X|)$ é o maior valor em módulo do conjunto de treinamento.

Após treinada a rede temos a fase de teste. Nesta fase a rede foi aplicada ao conjunto de dados de teste e o resultado foi confrontado com o resultado obtido pelo seleção das famílias com médias de TCHr acima da média geral.

A análise via redes neurais foi feita utilizando a função nnet do pacote nnet (VENABLES e RIPLEY, 2002) implementado no software R (R Core Team, 2013).

2.3. Seleção via tonelada de cana por hectare estimada

Para os cenários onde foi feita a modelagem via redes neurais também foi feita a seleção das melhores famílias selecionando aquelas com tonelada de cana por hectare estimada (TCHe) acima da média geral.

A variável TCHe pode ser obtida através da seguinte expressão (FERREIRA et al., 2007):

$$TCHe = d \times \pi \times NC \times AC \times \left(\frac{DC}{2}\right)^2 \times \frac{1}{100 \times tp}$$

onde:

d é a densidade específica do colmo. Chang e Milligan (1992) sugerem utilizar para d o valor de 1000 kg m^{-3};

tp é o tamanho da parcela em m^2.

2.4. Modelagem via análise discriminante

O treinamento e a estimação das funções discriminantes foram estabelecidos a partir do conjunto de dados de treinamento em cada um das cinco situações estudadas e para verificar a eficiência do método foi utilizado o respectivo conjunto de teste para a obtenção da taxa de erro aparente.

Foi usada a função discriminante de Fisher (1938) para discriminar as famílias dos experimentos de teste em famílias selecionadas ou não selecionadas. Se considerarmos que o vetor de médias $\boldsymbol{\mu}_i$ multivariado e a matriz de covariância $\boldsymbol{\Sigma}$ comum das populações são conhecidos, demonstra-se que a função linear que produz separação máxima entre as populações é denominada Função Discriminante Linear de Fisher (FERREIRA, 2011) sendo escrita como:

$$D(\mathbf{X}) = [\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2]^t \boldsymbol{\Sigma}^{-1}\mathbf{X}.$$

Seja uma nova observação $\mathbf{x}_0$, o valor da função discriminante é

$$D(\mathbf{x}_0) = [\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2]^t \boldsymbol{\Sigma}^{-1}\mathbf{x}_0$$

e seja

$m = \frac{1}{2}(D(\boldsymbol{\mu}_1) + D(\boldsymbol{\mu}_2))$ o ponto médio entre as duas médias populacionais univariadas $D(\boldsymbol{\mu}_1)$ e $D(\boldsymbol{\mu}_2)$.

A regra de classificação é:

$$\mathbf{x}_0 \in \prod_1 \text{ se } D(\mathbf{x}_0) = [\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2]^t \boldsymbol{\Sigma}^{-1}\mathbf{x}_0 \geq m$$

$$\mathbf{x}_0 \in \prod_2 \text{ se } D(\mathbf{x}_0) = [\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2]^t \boldsymbol{\Sigma}^{-1}\mathbf{x}_0 < m.$$

Na prática não conhecemos os vetores de médias e as matrizes de covariâncias. Assim é necessário supor normalidade da população e obter os estimadores de $\boldsymbol{\mu}_i$ e $\boldsymbol{\Sigma}$. Felizmente, as regras de discriminação continuam as mesmas descritas anteriormente, utilizando-se os estimadores ao invés dos parâmetros.

Então, os parâmetros $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ e $\boldsymbol{\Sigma}$ serão substituídos pelos respectivos estimadores $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ e $\mathbf{S}_c$. Assim teremos a Função Discriminante Linear Amostral de Fisher escrita como:

$$D(\mathbf{x}) = [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2]^t \mathbf{S}_c^{-1}\mathbf{x}$$

O ponto médio, m, entre as duas médias amostrais univariadas é dado por:

$$m = \frac{1}{2}(D(\bar{\mathbf{x}}_1) + D(\bar{\mathbf{x}}_2)).$$

A regra de classificação baseada nas amostras é apresentada por (FERREIRA, 2011):

$$\mathbf{x}_0 \in \prod_1 \text{ se } D(\mathbf{x}_0) = [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2]^t \mathbf{S}_c^{-1} \mathbf{x}_0 \geq m$$

$$\mathbf{x}_0 \in \prod_2 \text{ se } D(\mathbf{x}_0) = [\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2]^t \mathbf{S}_c^{-1} \mathbf{x}_0 < m$$

A análise discriminante foi feita utilizando a função lda do pacote MASS implementado no software R (R Development Core Team, 2013).

2.5. Avaliação e comparação das técnicas

Para avaliação dos classificadores – rede neural artificial, análise discriminante e seleção via TCHe – gerou-se matrizes de confusão para cada um dos cenários em ambos os métodos. A matriz de confusão oferece uma medida efetiva do classificador utilizado ao mostrar o número de classificações corretas usando a variável TCHr *versus* o número de classificações preditas pelo classificador considerado para cada classe sobre o conjunto de teste.

A seguir é apresentada o esquema da matriz de confusão para o classificador rede neural artificial tendo em vista que o raciocínio é o mesmo considerando a seleção via TCHe (Tabela 4).

Tabela 4: Matriz de confusão entre as estratégias de seleção via TCHr e redes neurais, acompanhada das taxas de erro de classificação.

Classificação	NS (Rede)	S (Rede)
NS (TCHr)	VN	FP
S (TCHr)	FN	VP

*S= Famílias selecionadas, NS= Famílias não selecionadas, VP= Verdadeiro Positivo, FN= Falso Negativo, FP= Falso Positivo, VN= Verdadeiro Negativo

É fácil perceber que VP indica o número de famílias selecionados por ambos os métodos, FN corresponde ao número de famílias selecionados via TCHr que não foram selecionadas pela rede neural, FP indica o número de famílias não selecionadas via TCHr que foram selecionadas pela rede neural e VN corresponde ao número de famílias não selecionadas por ambos.

Várias são as medidas que se originam da matriz de confusão para avaliação do classificador. Especificamente para seleção de famílias em cana-de-açúcar estamos interessados, principalmente, nas taxas de erros, que trataremos aqui de taxa de erro 1 (TE1) e taxa de erro aparente (TEA).

A TE1 é escrita como:

$$TE1 = \frac{FP}{N}$$

onde:

N é o número de famílias da população de validação.

A TEA é dada por:

$$TEA = \frac{FP + FN}{N}$$

onde:

N é o número de famílias da população de validação.

3. Resultados e discussão

As taxas de erro aparente via modelagem por redes neurais (TEA-RN), análise discriminante (TEA-AD), tonelada de cana por hectare estimada (TEA-TCHe) e as taxas de erro 1 via modelagem por redes neurais (TE1-RN), análise discriminante (TE1-AD), tonelada de colmos por hectare estimada (TE1-TCHe) encontram-se na Tabela 5.

Tabela 5. Taxa de erro aparente via rede neural artificial (TEA-RNA), análise discriminante (TEA-AD) e tonelada de cana por hectare estimada (TEA-TCHe) e taxa de erro 1 via rede neural (TE1-RNA), análise discriminante (TE1-AD) e tonelada de cana por hectare estimada (TE1-TCHe).

Exp. trein	Exp. teste	TEA -RNA	TE1-RNA	TEA-AD	TE1-AD	TEA-TCHe	TE1-TCHe
1	2,3,4,5	0,10227	0.07954	0,13636	0,09091	0,17045	0,07954
2	1,3,4,5	0,11363	0.06818	0,13636	0,07954	0,18182	0,09091
3	1,2,4,5	0,14773	0.09091	0,17045	0,07954	0,14773	0,07955
4	1,2,3,5	0,125	0.09091	0,14772	0,10227	0,15909	0,07955
5	1,2,3,4	0,09091	0.06818	0,125	0,07954	0,15909	0,07955

Exp. trein= Experimento de treinamento; Exp.teste= Experimentos de teste

Pode ser observado que a modelagem via redes neurais fornece menor taxa de erro aparente em todos os casos, ou seja, é o método que apresenta maior concordância com o método via TCHr (Tabela 5). A maior TEA-RN é 0,1477, que pode ser considerado um valor baixo, pois representa uma taxa de concordância de 85,23% com o método de seleção via TCHr. Quando o

experimento 5 é utilizado a taxa de concordância entre o modelo de rede neural e o método via TCHr foi superior a 90%.

A análise discriminante também apresenta bons resultados, a maior taxa de erro aparente foi obtida quando o experimento 3 é utilizado para treinamento com um valor de 0,1704, que também pode ser considerado um valor baixo pois corresponde uma taxa de concordância de 82,95 %. A modelagem via análise discriminante chega alcançar taxas de concordância de 87,50% quando o experimento 5 é utilizado para simulação e treinamento e de 86,36% quando são utilizados os experimentos 1 e 2 (Tabela 5).

O método que apresentou maiores TEA e, consequentemente, menores taxas de concordância com o método via TCHr foi o baseado na estimação da variável tonelada de cana por hectare. Apesar de inferior quando comparado aos demais métodos, os resultados não são ruins, tanto que o TCHe tem sido usado com grande frequência na prática (BARBOSA e SILVEIRA, 2010). As maiores taxas de erros são 0,1818 e 0,1704, quando são utilizados para simulação e treinamento os experimentos 2 e 1 respectivamente (Tabela 5).

Os resultados acima estão resumidos na figura 8. Como pode ser observado, a rede neural fornece menor taxa de erro aparente média seguida da análise discriminante e da seleção via TCHe.

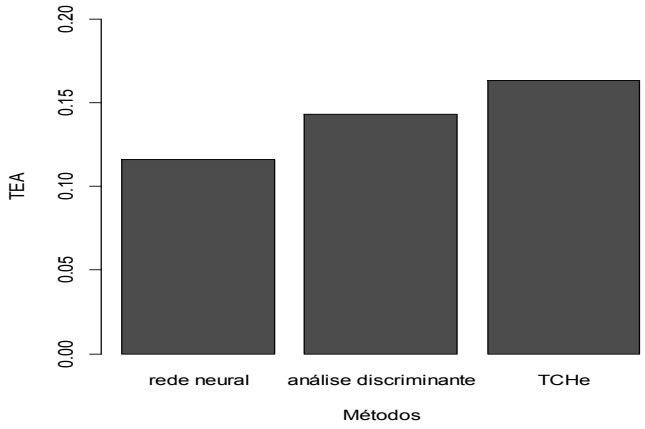


Figura 8: Taxa de erro aparente média via rede neural, análise discriminante e seleção via tonelada de cana por hectare estimada (TCHe).

Se tomarmos como uma medida de erro efetivo cometido pelo método a diferença média entre a TEA e a TE1, uma vez que a TE1 não representa um problema na prática, os melhores resultados são também provenientes da classificação via rede neural, como mostra a Figura 9.

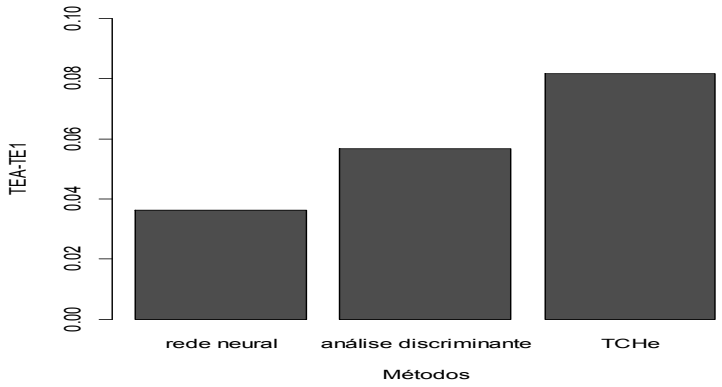


Figura 9: Diferença média entre a taxa de erro aparente (TEA) e a taxa de erro 1 (TE1) via rede neural, análise discriminante e tonelada de cana por hectare estimada.

Estes resultados indicam que outras alternativas, como o modelagem via redes neurais e a análise discriminante linear, podem ser utilizadas e com melhores resultados quando comparado com a simples obtenção do TCHe.

Os melhores resultados, tanto com relação a taxa de erro aparente quanto com relação a diferença entre a TEA e a TE1, foram obtidos com o uso das redes neurais, um reflexo da capacidade de generalização do modelo resultante desse procedimento.

Os modelos de redes neurais são mais flexíveis e com isso são capazes de descrever relações mais gerais nos dados que os modelos tradicionais (HAYKIN, 2001).

Os bons resultados fornecidos pela análise discriminante linear são um indício que os grupos, referentes às famílias selecionadas e não selecionadas, são linearmente separáveis (FERREIRA, 2011) de maneira que mesmo métodos mais simples podem fornecer resultados satisfatórios.

É importante ressaltar que, na prática, para a obtenção dos dados necessários para uso da rede neural ou da análise discriminante, seria necessária a pesagem das parcelas de apenas um dos cinco experimentos,

mostrando que estas são estratégias que podem ser utilizadas para a seleção de materiais promissores em cana-de-açúcar. Dessa forma pode ser reduzido significativamente o problema de pesagem no campo experimental presente nos métodos tidos como ideais. Obviamente, necessitaríamos da tomada de informações do NC, DC e AC em todos os experimentos, porém estamos assumindo que tal coleta de dados seria mais prática de ser realizada.

4. Conclusões

A modelagem via redes neurais e a análise discriminante fornecem menores erros quando comparadas a estratégia usualmente utilizada, que é baseada na estimação da variável tonelada de cana por hectare.

Comparando os modelos de redes neurais com a análise discriminante, a rede neural fornece melhores resultados.

5. Bibliografia

BARBOSA, M.H.P.; SILVEIRA, L.C.I. (2010) Melhoramento Genético e Recomendação de Cultivares. In: Santos, F.; Borém, A. e Caldas, C. Editores. Cana-de-açúcar: Bioenergia, Açúcar e Álcool - Tecnologias e Perspectivas. Viçosa, MG – Suprema, 578 p.

CHANG, Y.S.; MILLIGAN S.B. Estimating the potential of sugarcane families to procude elite genotypes using univariate cross prediction methods Theoretical and Applied Genetics, Berlin, v. 84, p. 662-671, 1992.

CRUZ, C. D.; CARNEIRO, P. C. S. (2006) Modelos Biométricos aplicados ao melhoramento genético. Vol. 2, Viçosa, MG: UFV. 585p.

FALCONER D.S.; MCKAY T.F.C. (1996) Introduction to Quantitative Genetics. Malaysia: Longmans Green, 463p.

FERREIRA, R.P.; BHERING, L. L.; SALGADO, C.C. (2013) Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. Crop Breeding and Applied Biotechnology 13: 152-156.

FISHER, R.A. (1938) The Statistical Utilization of Multiple Measurements, Annals of Eugenics, (8): 376-386.

GIANOLA, D.; OKUT, H.; KENT A. W.; ROSA, J. M. R. (2011) Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genetics*, 45:34.

HASTIE T.; TIBSHIRANI R.; FRIEDMAN J. (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York, 745p.

MACKAY, D. J. C. (1994) Bayesian non-linear modelling for the prediction competition. In: ASHRAE Transactions, ASHRAE, Atlanta Georgia. Vol. 100, pp. 1053–1062.

MATSUOKA, S.; GARCIA, A.A.F.; ARIZONO, H. (2005) Melhoramento da cana-de-açúcar. In: BORÉM, A. (Ed.) Melhoramento de espécies cultivadas. Viçosa: Ed. da UFV. 969p.

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

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

R Development Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (http://www.r-project.org).

RESENDE, M.D.V.; BARBOSA, M.H.P. (2006) Selection via simulated Blup base on family genotypic effects in sugarcane. Pesquisa Agropecuária Brasileira, 41(3):421-429.

SIMMONDS N.W. (1996) Family selection in plant breeding. Euphytica (90):201–208

STRINGER, J.K.; COX, M.C.; ATKIN, F.C.; WEI, X.; HOGARTH. (2011) Family Selection Improves the Efficiency and Effectiveness of Selecting Original Seedlings and Parents. Sugar Tech, 13(1):36–41.

VENABLES W.N.; RIPLEY B.D. (2002) Modern applied statistics with s. Springer, New York, 493p.

VENTURA, R. V.; SILVA, A. M.; MEDEIROS, T. H.; DIONELLO, N. L; MADALENA, F.E.; FRIDRICH, A.B.; VALENTE, B.D.; SANTOS, G.G.; FREITAS, L.S.; WENCESLAU, R.R.; FELIPE, V. P. S.; CORRÊA, G.S.S. (2012) Uso de redes neurais artificais na predição de valores genéticos para peso aos 205 dias em bovinos da raça Tabapuã. *Arq. Bras. Med. Vet. Zootec.*, 64(2): 411-418.

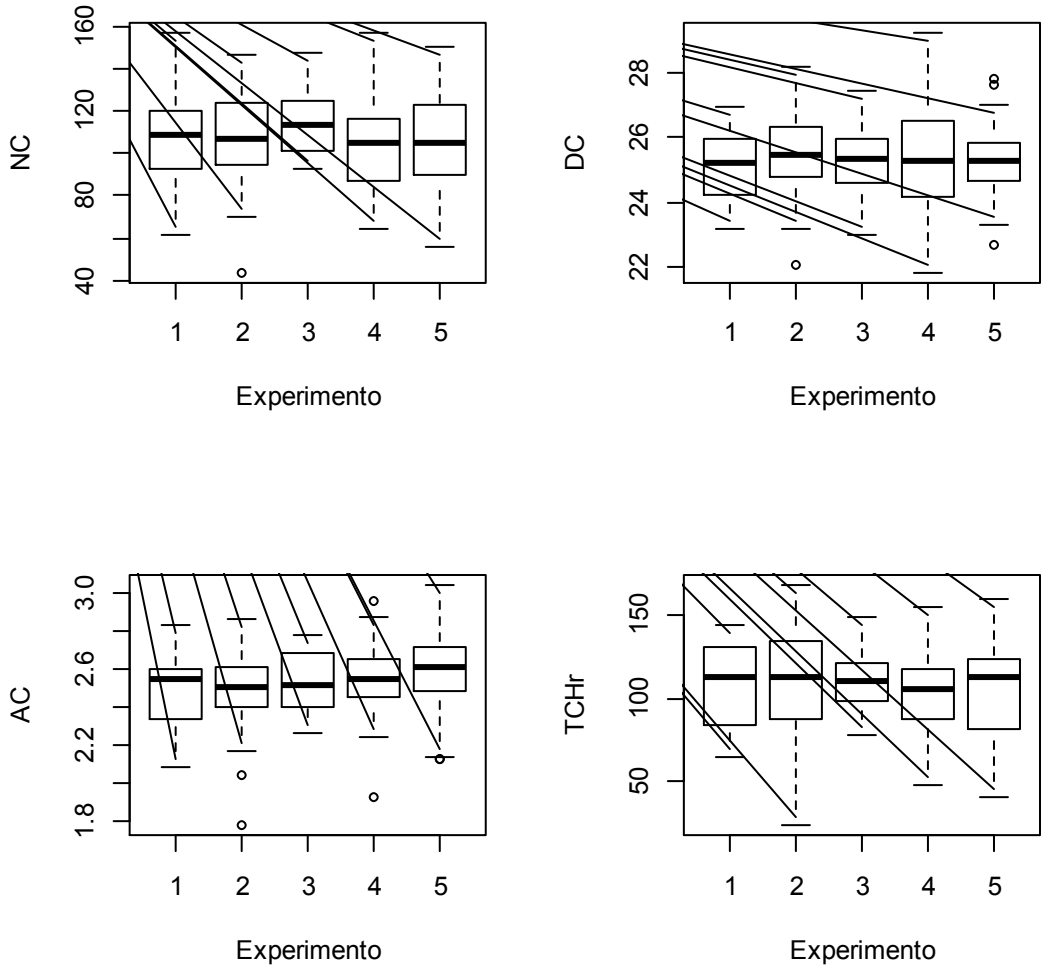
CONCLUSÕES GERAIS

A modelagem via redes neurais, usando as ferramentas de simulação e padronização, apresenta alta concordância com a seleção via tonelada de cana por hectare.

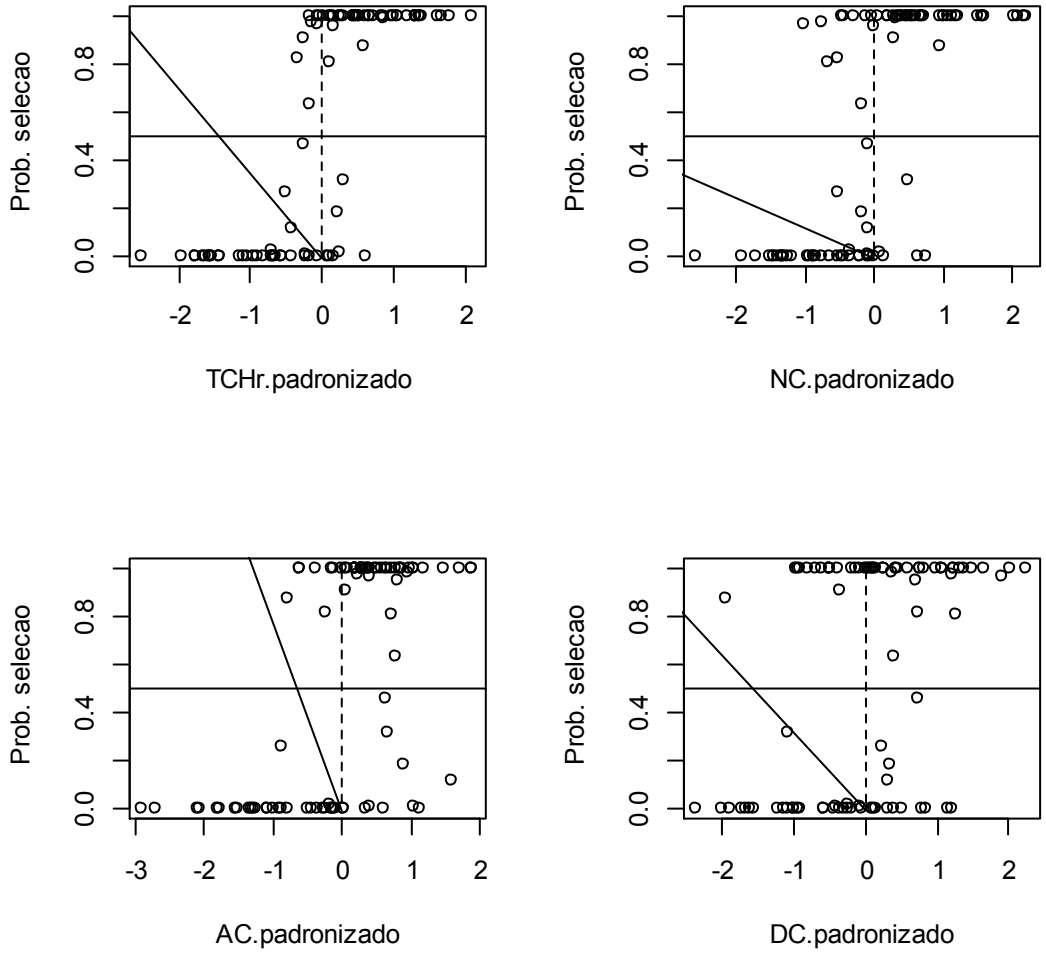
A modelagem via redes neurais e a análise discriminante fornecem menores erros quando comparadas a estratégia usualmente utilizada, que é baseada na estimação da variável tonelada de cana por hectare.

Comparando os modelos de redes neurais com a análise discriminante, a rede neural fornece melhores resultados.

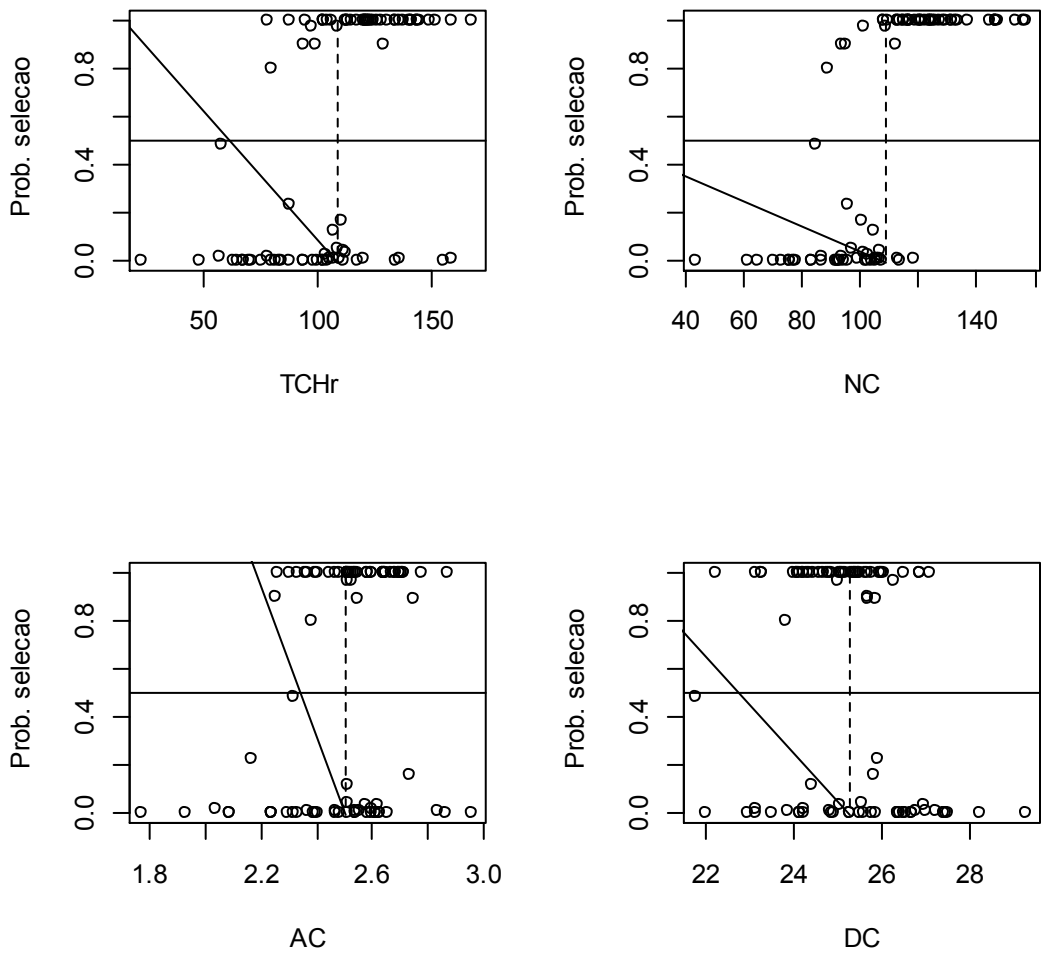
Apêndice



1) Boxplot para as variáveis DC, NC, AC e TCHR nos cinco experimentos.



2. Gráfico de probabilidade de seleção versus variável no melhor cenário (com simulação e com padronização) (Capitulo 2).



3. Gráfico de probabilidade de seleção versus variável no pior cenário (sem simulação e sem padronização) (Capítulo 2).