Estimativa da produção de arroz irrigado por meio de Regressão Múltipla e Redes Neurais Artificiais

João Guilherme Simões[1], Gabi Nunes Silva[2], Laís M. Azevedo Barroso[3]

[1]Fundação Universidade Federal de Rondônia - simoesj.guilherme@gmail.com
[2]Fundação Universidade Federal de Rondônia - gabi.silva@unir.br
[3]Fundação Universidade Federal de Rondônia - lais.barroso@unir.br

Introdução

O arroz (*Oryza sativa L.*), é um cereal de grande importância econômica em todo o mundo e constitui a base alimentar de cerca de 2,4 bilhões de pessoas. Dentre os caracteres agronômicos dessa cultura, a produção de grãos constitui a variável de maior importância econômica e dos programas de melhoramento genético de arroz irrigado, que visam tornar a orizicultura irrigada mais atrativa, diminuindo o custo de produção e obtendo cultivares com qualidade nutricional melhor em relação às já existentes (JÚNIOR, 2017). No entanto, embora haja elevado interesse no aumento e qualificação da produção de arroz, o conhecimento sobre quais fatores ou caracteres influenciam diretamente na produção ainda é raso.

Em geral, a predição da produção é realizada por meio da metodologia de Regressão Linear Múltipla, entretanto essa metodologia enfrenta algumas limitações, tais como: pressupostos do modelo, dimensionalidade, ajusta modelos lineares (RESENDE, et al., 2007). Nesse sentido, estudos mais recentes têm indicado o uso das Redes Neurais Artificiais (RNAs) (HAYKIN, 2007), uma vez que tal metodologia não tem restrições quanto a pressupostos, dimensionalidade do modelo, e ainda permite a captação de efeitos não lineares entre as variáveis (CASTRO et al., 2017; SILVA et al., 2017).

Diante do exposto, o presente trabalho tem por objetivo mensurar o grau de influência de variáveis agronômicas de arroz irrigado sobre a variável de interesse produção de grãos por meio da técnica de Regressão Linear Múltipla e da metodologia de Redes Neurais Artificiais.

Metodologia

Foram avaliadas oito caracteres de arroz irrigado: produção de grãos (Prod), altura (Alt), comprimento de panícula (Cpan), número de grãos cheios/panícula (Ng/Pan), porcentagem de grãos cheios (%g), largura (Larg), espessura (Esp), e peso de 100 grãos (Pes). As variáveis foram denotadas Y, X_1, X_2, X_3, X_4, X_5, X_6 e X_7, respectivamente, provenientes de 25 genótipos de arroz irrigado do programa de melhoramento de arroz irrigado de Minas Gerais. O experimento foi conduzido sob o Delineamento de Blocos Casualizados com três repetições, na fazenda experimental de Leopoldina na safra de 2012/2013.

O modelo de Regressão Linear Múltipla (RLM) foi estimado pelo Método dos Mínimos Quadrados (DEMÉTRIO e ZOCCHI, 2006) segundo o modelo:$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_i X_i + \varepsilon$, i=1,2,3,...,n, em que Y é a produção de grãos de arroz irrigado (em g/parcela); B_i são os coeficientes de regressão a serem estimados; X_i são as variáveis independentes (caracteres agronômicos) e ε é o erro aleatório do modelo.

Com o modelo ajustado, procedeu-se ao teste T de Student para os betas da Regressão afim de avaliar a significância das variáveis explicativas e também utilizou-se o método de seleção automática de variáveis denominado *Backward* (MINGOTI, 2007). As medidas utilizadas para obtenção do melhor modelo foram o Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano (BIC). O ajuste dos modelos foi avaliado por meio do coeficiente de determinação (R^2).

A Rede Neural Artificial (RNA) implementada neste trabalho foi uma rede Perceptron de Camada Única (*Single Hidden Layer Neural Network*), com algoritmo de treinamento *back-propagation*, testando-se de 1 a 20 neurônios na camada oculta e um máximo de 8000 iterações. O modelo neural utilizado foi estabelecido tal como descrito por Nascimento et al. (2013). A topologia de rede adotada segue representada de forma funcional na figura 1.

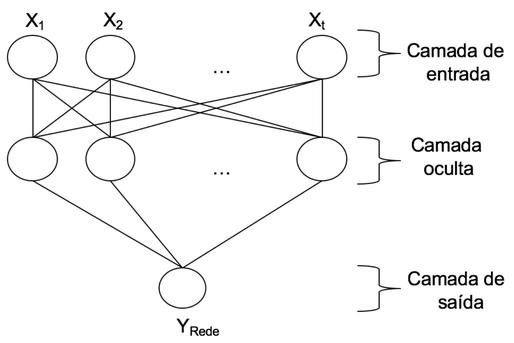

Figura 1: Esquema de uma Rede Neural Artificial *feed-forward* multicamadas. X_1, X_2, ...,X_t , t=1,...,7 são as variáveis preditoras (na camada de entrada) e Y_{Rede} representa a variável resposta (na camada de saída), nas camadas ocultas estão representados os neurônios
Fonte: próprios autores.

As funções de ativação adotadas foram as funções tangente hiperbólica $\gamma\ (x) = (1 + e^{-x})/(1 + e^{x})$ e sigmoide logística $\gamma\ (v) = 1/(1 + e^{-v})$.

Para o treinamento da rede (obtenção dos pesos) o conjunto de observações foi dividido em dois conjuntos: o primeiro, denotado por conjunto de treinamento contendo 100 observações (100 genótipos) foi simulado mantendo os mesmos parâmetros estatísticos de média e variância dos dados originais e o segundo conjunto constituído pelas demais observações, para fins de validação da rede, assim como Nascimento et al. (2013) sugerem. Tal como para o método de RLM, o ajuste do modelo via RNA foi avaliado por meio do coeficiente de determinação (R^2).

As análises estatísticas foram realizadas mediante a utilização do software livre R (R Core Team, 2019) e do programa GENES (CRUZ, 2008) no módulo integração ao software MATLAB (MOLER, 2011).

Resultados e Discussão

Análises de variância individuais e conjuntas foram realizadas e indicaram a significância das variáveis consideradas nesse estudo. Procedeu-se então ao método de RLM. O modelo ajustado com as 7 variáveis explicativas obtido foi:

$$Y = 1572,65 - 25,89X_1 + 97,13X_2 + 42,85X_3 - 3098,74X_4 - 1860,9X_5 + 2320,97X_6 + 304X_7$$

Pelo método *Backward*, somente as variáveis explicativas comprimento de panícula (Cpan) e número de grãos cheios/panícula (Ng/Pan), representadas por X_2 e X_3, respectivamente, deveriam ser consideradas no modelo, dada a significância dos betas e os menores valores de AIC=382,7 e BIC=387,6. O modelo obtido foi:

$$Y = -2548,8 + 117,6X_2 + 34,5X_3.$$

Os coeficiente de determinação (R^2) obtidos para o modelo com todas as 7 variáveis (modelo completo) e considerando somente as variáveis X_2 e X_3 selecionadas pelo método de seleção de *Backward* (modelo reduzido) encontram-se na tabela 1. O R^2 indica quão explicativa é a reta ajustada.

Tabela 1- Comparação dos coeficientes de determinação obtidos para o método RLM. Modelo completo: com as 7 variáveis explicativas. Modelo reduzido: com as variáveis X_2 e X_3 selecionadas via *Backward*

Método	Método adotado	R^2 (%)
RLM	Modelo completo	66,37
	Modelo reduzido	57

Fonte: próprios autores.

A RNA apresentou resultados semelhantes ao modelo de RLM. A RNA com melhor configuração foi a rede com 3 neurônios na camada oculta. Na tabela 2 pode-se observar os R^2 para os conjuntos de treinamento e teste da Rede.

Tabela 2- Comparação dos coeficientes de determinação obtidos para o método RNA. Modelo completo: com as 7 variáveis explicativas. Modelo reduzido: com as variáveis X_2 e X_3 selecionadas via *Backward*

Método	Método adotado	R^2_t	R^2_v
RNA	Modelo completo	78,34	61,52
	Modelo reduzido	59,09	55,15

Fonte: próprios autores.

Ao utilizar a RNA, a importância dos caracteres avaliados pode ser feita mensurando-se o impacto sofrido no modelo ajustado ao retirar o efeito de determinada característica. Observou-se que, tal como para o modelo RLM, as RNAs também identificaram as variáveis X_2 e X_3 – comprimento de panícula (Cpan) e número de grãos cheios/panícula (Ng/Pan), respectivamente, como sendo as mais importantes para estimar a produção de grãos, evidenciando concordância entre os dois métodos – RLM e RNA – avaliados neste estudo.

Os baixos valores de R^2 podem ser justificados pelas baixas correlações identificadas preliminarmente entre as variáveis explicativas e a variável dependente produção de grãos, prejudicando diretamente a análise pelo seu pouco poder explicativo (HAIR et al., 2009). A matriz de correlação das variáveis incluídas nesse estudo encontra-se na figura 2.

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Y	1							
X_1	0,1957	1						
X_2	0,5271	0,4745	1					
X_3	0,6637	0,3718	0,2754	1				
X_4	0,0278	0,2565	0,1422	0,1794	1			
X_5	-0,4697	-0,2451	-0,7270	-0,2227	-0,3511	1		
X_6	-0,2500	-0,1149	-0,3389	-0,2937	-0,2762	0,6300	1	
X_7	-0,3492	-0,2620	-0,4297	-0,4643	-0,2986	0,4131	0,3995	1

Figura 2: Matriz de correlação dos caracteres agronômicos de arroz irrigado
Fonte: próprios autores.

Note que, apesar de o método *Backward* ter indicado a inclusão somente das variáveis X_2 e X_3, os modelos que incluíam todas as variáveis apresentaram maiores valores de R^2. A literatura afirma que modelos com maior número de parâmetros tendem a apresentar maiores coeficientes, mas alertam a necessidade de se avaliar as correlações, verificar presença ou não de colinearidade entre as variáveis, além de utilizar métodos de seleção de variáveis para evitar problemas de superestimação, multicolinearidade, dentre outros (RESENDE, et al., 2007).

Vale ressaltar ainda que, para as RNAs, o baixo número de entradas (25 genótipos de arroz irrigado) também é um fator complicador, uma vez que para um bom desempenho e maior eficácia no processo de treinamento e aprendizado de uma RNA, quanto maior o número de observações, melhor ela atua no processo de captação de informações para sua sinapses (SILVA; SPATTI e FLAUZINO, 2010).

Comparando as metodologias utilizadas, pode-se destacar duas vantagens da RNA sobre o modelo de RLM. A primeira é que, como afirmam Barroso et al. (2013), na técnica de RNA é feita nenhuma pressuposição acerca da distribuição dos dados e da estrutura de covariâncias, diferentemente do modelo de RLM, em que os pressupostos sobre os resíduos precisam ser atendidos para que as predições sejam válidas. Outra vantagem é a dificuldade da interpretação do modelo de Regressão na presença de muitas variáveis em estudo, o que não acontece na técnica de RNA.

Conclusões

Para o problema em questão, assim como para estudos de adaptabilidade e estabilidade na área de melhoramento genético vegetal já realizados por alguns autores, as RNAs apresentaram elevada concordância com a metodologia tradicional de Regressão Linear, destacando-se também por suas consideráveis vantagens mencionadas frente a eventuais problemas do modelo de RLM.

De modo geral, pode-se afirmar que as variáveis comprimento de panícula e número de grãos cheios/panícula exercem maior influência sobre a variável de maior interesse econômico, a produção de grãos.

Vale salientar que este trabalho consistiu em um estudo de caso para fornecer aos leitores a possibilidade do uso de outras metodologias para a solução de problemas que envolvam ajuste de modelos e para apresentar as potencialidades das RNAs.

Bibliografia

BARROSO, Laís Mayara Azevedo et al. 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)**, 2013.

CASTRO, Carla Aparecida de O. et al. High-performance prediction of macauba fruit biomass for agricultural and industrial purposes using Artificial Neural Networks. **Industrial Crops and Products**, v. 108, p. 806-813, 2017.

DEMÉTRIO, Clarice Garcia Borges; ZOCCHI, Sılvio Sandoval. Modelos de regressão. **Piracicaba: ESALQ**, 2006.

HAIR, Joseph F. et al. **Análise multivariada de dados**. Bookman Editora, 2009.

HAYKIN, Simon. **Redes neurais: princípios e prática**. Bookman Editora, 2007.

JÚNIOR, A. C. D. S. **Progresso genético do programa de melhoramento de arroz irrigado em minas gerais no período 1993/94 a 2015/2016**. 2017. Tese de Doutorado. Universidade Federal de Viçosa.

MINGOTI, Sueli Aparecida. Análise de dados através de métodos estatística multivariada: uma abordagem aplicada. In: **Análise de dados através de métodos estatística multivariada: uma abordagem aplicada**. 2007. p. 295-295.

MOLER, Cleve B. **Experiments with MATLAB**. Society for Industrial and Applied Mathematics, 2011.

NASCIMENTO, Moysés et al. Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. **Crop Breeding and Applied Biotechnology**, v. 13, n. 2, p. 152-156, 2013.

RESENDE, M. D. V. Matemática e estatística na análise de experimentos e no melhoramento genético, Embrapa Florestas, Colombo. **Forestry Embrapa, Colombo, PR, Brazil**, 2007.

SILVA, Gabi Nunes et al. Artificial neural networks compared with Bayesian generalized linear regression for leaf rust resistance prediction in Arabica coffee. **Pesquisa Agropecuária Brasileira**, v. 52, n. 3, p. 186-193, 2017.

SILVA, IN da; SPATTI, Danilo Hernane; FLAUZINO, Rogério Andrade. Redes neurais artificiais para engenharia e ciências aplicadas. **São Paulo: Artliber**, v. 23, n. 5, p. 33-111, 2010.

Agradecimentos