REGULAR ARTICLE

Artificial intelligence applied to estimate soybean yield

Wesley Prado L. dos Santos[1*], Mariana Bonini Silva[2], Alfredo Bonini Neto[1], Carolina S. B. Bonini[2], Adônis Moreira[3]

[1] São Paulo State University (UNESP), School of Sciences and Engineering, Tupã, São Paulo State, Brazil.
[2] São Paulo State University (UNESP), College of Agricultural and Technological Sciences, Dracena, São Paulo State, Brazil.
[3] Department of Soil Science, Embrapa Soja, Londrina, Paraná State, Brazil.

Regular Section
Academic Editor: Celso Antonio Goulart

Statements and Declarations

Data availability
All data will be shared if requested.

Institutional Review Board Statement
Not applicable.

Conflicts of interest
The authors declare no conflict of interest.

Funding
This research did not receive external funding.

Autor contribution
WPLS: Experimental data collection, Data custody, Literature review, Writing the manuscript; Conceptualization, Supervision; MBS: Experimental data collection; ABN: Experimental data collection, Data analysis, Conceptualization, Supervision, Manuscript Review; CSBB: Conceptualization, Supervision, Manuscript Review; AM: Conceptualization, Data analysis, Writing the manuscript, Manuscript Review, Supervision.

Abstract

The application of mathematical models using biotic and abiotic factors for the efficient use of fertilizers to obtain maximum economic productivity can be an important tool to minimize the cost of soybean (Glycine max (L.) Merr.) grain yield. In this sense, using Artificial Neural Networks (ANN) is an important tool in studies involving optimization. This study aimed to estimate soybean yield in Luiziana, Paraná state, Brazil, by considering two growing seasons and an Artificial Neural Network (ANN) as a function of the morphological and nutritional parameters of the plants. Results reveal a well-trained network, with a margin of error of approximately 10^{-5}, thus acting as a tool to estimate soybean data. For the phases, model validation and network test, i.e., data that were not part of the training (validation), the errors averaged 10^{-3}. These results indicate that our approach is adequate for optimizing soybean yield estimates in the area studied.

Keywords

Artificial Neural Network; Forecast; Intelligent systems; Soy; Mathematical modelling.

Introduction

Soybeans (*Glycine max (L.) Merrill*) are one of the most important products cultivated and consumed in the world due to their chemical composition, nutritional quality, and productive potential (Kamali *et al.*, 2017). Despite this importance, only a small portion of farmers manage to fully exploit the productive potential of the crop due to problems in management, such as the inadequate use of fertilizers and correctives, as well as climatic factors, or combinations of these, which cause plants in more favourable conditions to have high performances (Hoeft, 2003).

In this regard, mathematical equations and Artificial Neural Networks (ANNs) for estimating crop development are important tools, whose simulation models are used to: i) verify theories and test hypotheses; ii) improve knowledge about certain processes by feeding databases with the information obtained; iii) make estimates of grain yield or plant biomass as a function of biotic and abiotic factors, and iv) quantify the dynamics of some element within an ecosystem (Boote *et al.*, 1996).

Methods to identify the variables involved in production with computational tools are increasingly being used for studies involving plant production mechanisms (Moreira *et al.*, 2023; Bonini et. al., 2023; Souza *et al.*, 2019; Putti *et al.*, 2017; Silva *et al.*, 2014), such as artificial neural networks (ANN) that mimic biological neurons as simple processing units that have the natural propensity to store experimental knowledge and make it available for use (Eliasmith & Anderson, 2003; Kovacs, 2006).

These systems resemble the human brain, in the sense that knowledge of both is acquired from the environment in a learning process given by the strength of connection (synaptic weights) between neurons and networks in the systems (Haykin, 2001). Since the publication of the work by Rumelhart and McClelland (1986), neural networks have been used in several areas in agriculture (Kovacs 2006; Braga *et al.* 2007) and studies have been developed using ANN with applications in chemical and physical attributes of the soil, as well as mapping, nutrient absorption models, and production estimates (Bonini Neto *et al.*, 2022; Bonini Neto *et al.*, 2021; Beuchera *et al.*, 2015; Silveira *et al.*, 2013; Mouazen *et al.*, 2010; Anagu *et al.*, 2009).

The training of a neural network can be supervised or unsupervised. While unsupervised training does not require a desired output (i.e., the network performs self-organizing training considering only the input data), supervised training considers what the network learns from input data and its respective desired outputs (Braga *et al.*, 2007). In other words, supervised training consists of knowing a target to be hit so that the network can adapt its weights, in a way that, later in the operation or diagnosis process (also known as the network test phase) one can classify or estimate data that was not part of the training process. Therefore, it can be said that the

* Corresponding author
E-mail address: wesley.prado@unesp.br (W.P.L Santos).

https://doi.org/10.18011/bioeng.2024.v18.1211
Received: 02 November 2023 / Accepted: 11 January 2024 / Available online: 14 March 2024

learning of an ANN takes place by adjusting its weights (Wi) during training and depending on input data whose outputs are known.

Within these precepts, the aim of this work was to use data on production as well as morphological and nutritional components of 16 soybean cultivars obtained during the experiment in two harvests, to estimate productivity using artificial neural networks (ANN).

Materials and methods

The experiment was conducted under rainfed conditions during two harvests (2018-2019 and 2019-2020) in the city of Luiziana, Paraná state, Brazil (23°23'30" SL, 51°11'05" WL, and altitude of 720 m above sea level), as shown in Figure 1(a). The area was managed under the direct planting system (NT) for 17 years. According to the Köppen classification (Alcarde *et al.* 2013), the climate is mesothermal, Cfb and Cfa, humid subtropical with minimum temperatures around 15ºC in July and average/maximum temperatures around 23ºC in February. Annual precipitation is approximately 1,600 mm, with rainy months between December and February and the driest months from June to August. Over the last 10 years, the city of Luiziana produced, on average, 150,934 tons of soybeans per year (IBGE, 2023).

Sixteen cultivars were used in the experiment: 1- NA 5909 RR, 2- TMG 7262 INOX, 3- BRS 1010 IPRO, 4- FPS Solar IPRO, 5- BMX 6663 RR, 6- M 5947 IPRO, 7- M 5917 IPRO, 8- DM 5958 IPRO, 9- M 6410 IPRO, 10- Agroeste 3610 IPRO, 11- V Top RR (NK 1059), 12- ND 6006 IPRO, 13- ND 6390 IPRO, 14- ND 6535 IPRO, 15 -BRS 1001 IPRO e 16- BRS 1003 IPRO (Embrapa 2023).

The RNA used was the Multilayer Perceptron (MLP), composed of n=18 neurons in the input layer – whose morphological and nutritional parameters were: cultivars, number of nodes, number of branches, height (cm), number of pods, number of grains, SPAD (Soil Plant Analysis Development), nitrogen, phosphorus, potassium, calcium, magnesium, sulphur, boron, copper, iron, manganese and zinc – m=30 neurons in the intermediate layer (best network performance), and i=1 neuron in the output layer that represents soybean production in the years 2018-2019 and 2019-2020 (Figure 1(b)), with backpropagation training algorithm. The activation function for the output of both layers was the hyperbolic tangent given by equation (1)

$$f(u) = (1 - e^{-\lambda u})/(1 + e^{-\lambda u}) \qquad (1)$$

where λ is an arbitrary constant and corresponds to the slope of the curve.

The platform used for the computational implementation of the ANN and obtention of results was MATLAB® (Mathworks, 2022).

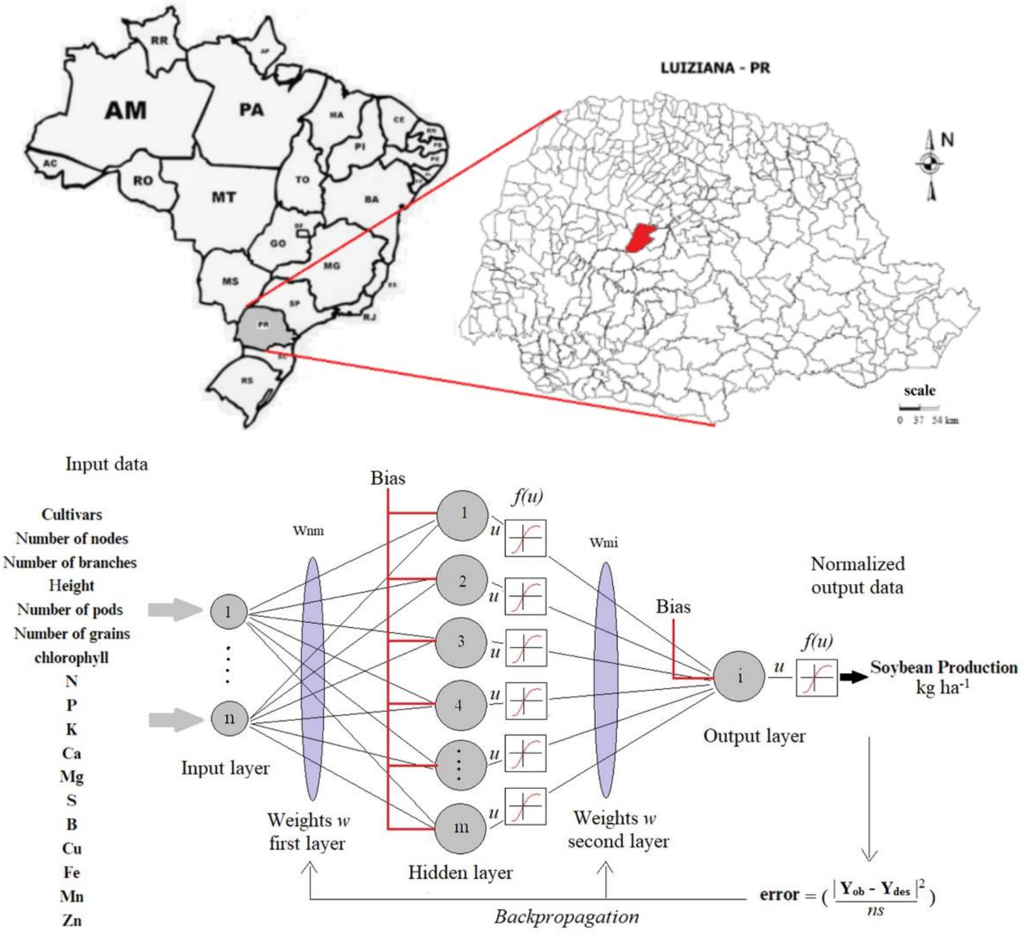


Figure 1. (a) Location of the city of Luiziana, Paraná state, Brazil, (b) ANN model used.

Results and discussion

The results found in the training, validation, and testing phase are shown in Figure 2. For network training, 76 samples (80%) were used as shown in Figure 2(a). The processing time was 2 seconds with an error of 0.000015. For validation and testing purposes, 10 samples were used in each set, which were not included in the training set. The errors obtained were 0.00255 and 0.00222, respectively, as shown in Figures 2(b) and 2(c). Figure 2(d) shows the network performance for 100% of the samples.

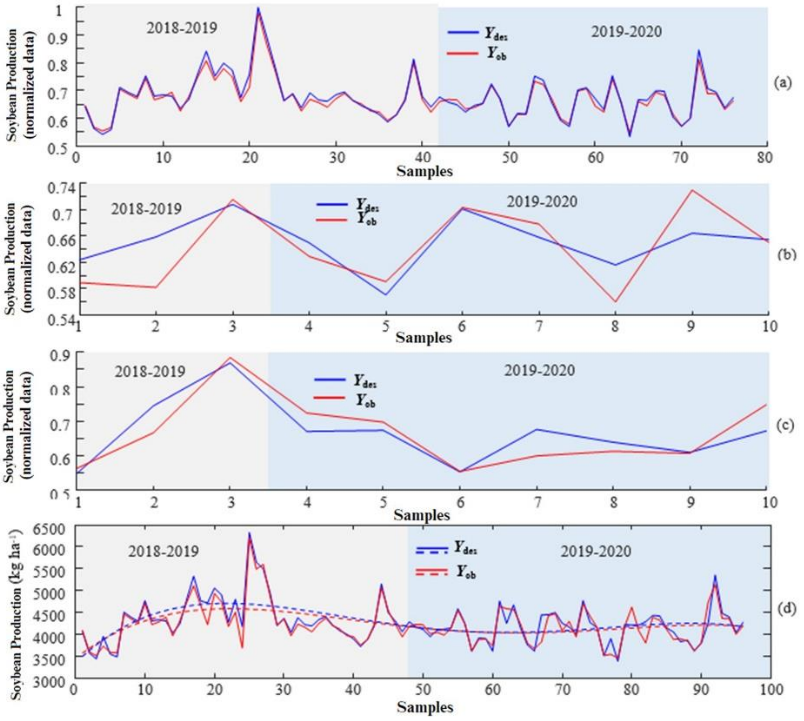
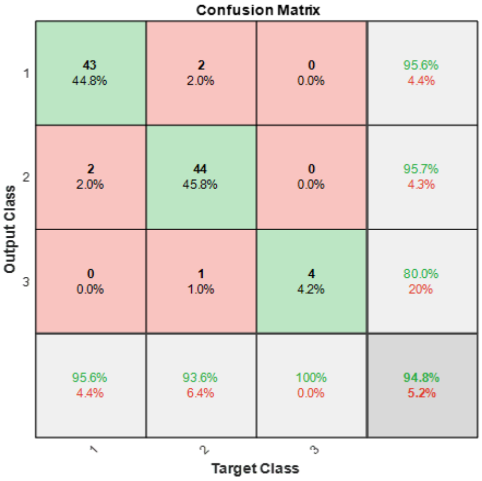


Figure 2. Analysis between variables: desired output (Y_{des}) and obtained output (Y_{ob}) with 30 neurons in the hidden layer, (a) training with 80% of the samples, (b) validation with 10% of the samples, (c) testing with 10% of the samples, and (d) all samples (100%).

Furthermore, from this data, it was possible to create the confusion matrix used to classify the categorical data. To do this, the desired output (Y_{des}) was transformed into categorical data, where Y_{des} was divided into three binary classes. Class 1 was defined as (0, 0, 1), indicating soybean production between 3300 and 4200.99 kg per hectare; class 2 was represented by (1, 0, 0), reflecting soybean production between 4201 and 5200.99 kg per hectare; and class 3 was designated as (1, 1, 1), corresponding to a soybean production between 5301 and 6400 kg per hectare. As a result of this process, the confusion matrix represented in Figure 3 was obtained.



	1	2	3	
1	43 44.8%	2 2.0%	0 0.0%	95.6% 4.4%
2	2 2.0%	44 45.8%	0 0.0%	95.7% 4.3%
3	0 0.0%	1 1.0%	4 4.2%	80.0% 20%
	95.6% 4.4%	93.6% 6.4%	100% 0.0%	94.8% 5.2%

Figure 3. Confusion matrix for 96 samples and 3 classes.

In the confusion matrix graph, the rows correspond to the predicted class (Output Class) and the columns correspond to the true class (Target Class). In the same sense, diagonal cells correspond to correctly classified observations while off-diagonal cells correspond to incorrectly classified observations (Mathworks, 2022). It is worth noting that both the number of observations and the percentage of the total number of observations are shown in each cell. As an example, we can analyse the value 2 in the first row and second column of the confusion matrix. This column represents class 2 (1, 0, 0) for the desired output (Y_{des}). It can be seen that two samples were incorrectly classified for the obtained output (Y_{ob}), i.e., these samples were mistakenly identified as belonging to class 1 instead of class 2.

The column on the far right of the graph shows the percentages of all examples predicted to belong to each class that are correctly and incorrectly classified. These metrics are often called ‘accuracy’ (or ‘positive predictive value’) and ‘false discovery’ rate, respectively. The line at the bottom of the graph shows the percentages of all examples belonging to each class that are correctly and incorrectly classified. These metrics are often called ‘recall’ (or ‘true positive rate’) and ‘false negative’ rate, respectively. The cell in the lower right corner of the graph shows overall accuracy (Mathworks, 2021).

In Figure 3, the first three diagonal cells show the number and percentage of correct classifications after training, validating, and testing the network. For example, 43 samples are correctly classified as (0, 0, 1). This corresponds to 44.8% of all 96 samples. Similarly, 44 cases are correctly classified as (1, 0, 0), which corresponds to 45.8% of all samples. Four samples were classified as (1, 1, 1), corresponding to 4.2% of all samples. Thus, out of 45 sample predictions (line 1 - 0, 0, 1), 95.6% are correct and 4.4% are wrong. Out of 46 (row 2 - 1, 0, 0) predictions, 95.7% are correct and 4.3% are wrong. Out of the 45 cases (column 1), 95.6% are correctly predicted as (0, 0, 1) and 4.4% are predicted as (1, 0, 0). Out of the 47 (column 2) cases, 93.6% are correctly classified as (1, 0, 0) and 6.4% are wrong. Same for the third row and column.

In summary, out of the 96 samples, 91 were correctly classified and 5 were incorrectly classified, representing a hit rate of 94.8% and an error rate of 5.2%, which can be seen in Figure 4 (where the desired outputs (Y_{des}) and outputs obtained via ANN (Y_{ob}) for categorical data are represented). Therefore, errors are also seen to be present in the classification of samples (five errors), which occurred only in the validation and testing phases.

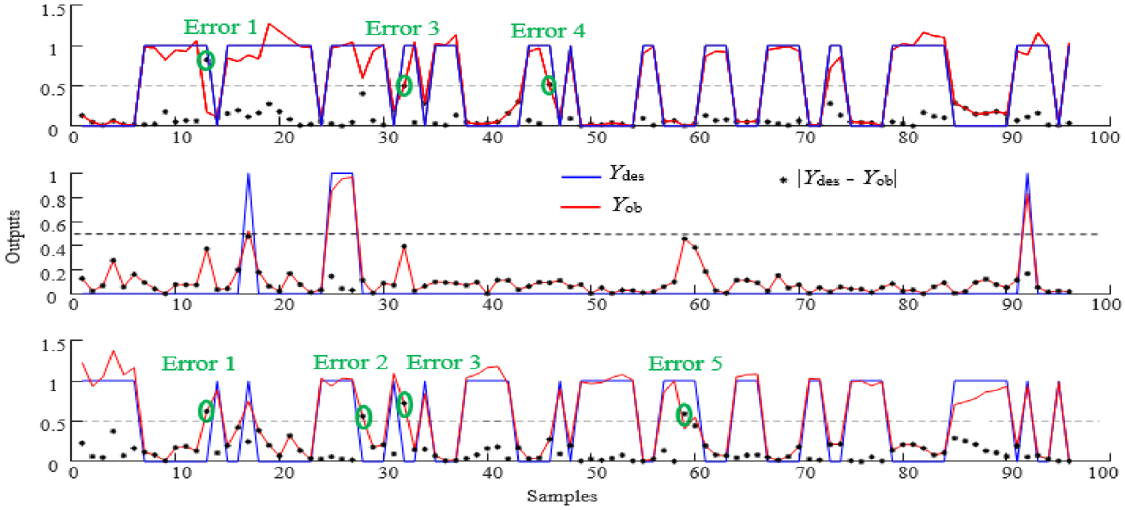


Figure 4. Desired outputs (Y_{des}) and outputs obtained via ANN (Y_{ob}) for categorical data.

The ANNs were properly trained, with a mean squared error of approximately 10^{-5} between the obtained (via ANN) and desired (via experimental field) outputs, equivalent to an average error of 70.07 kg ha^{-1} (1.58%). Under such a context, the best validation and testing performances were observed after the sixth iteration, with MSE values of 0.0025553 (average error of 180 kg ha^{-1}) and 0.0022251 (average error of 135 kg ha^{-1}), respectively.

In the validation and testing phases, the network predicted soybean productivity based on different morphological and nutritional data of the plant; values close to the one specified for the MSE (10^{-4}).

The ANN evaluated in this study demonstrated adequate performance, with adjusted weights and an average error of around 70.07 kg ha^{-1}. In the validation and testing phases, the network (Y_{ob}) showed mean squared errors (MSEs) of approximately 0.0025553 and 0.0022251, along with R values of 0.690 and 0.872, respectively, compared to the desired values (Y_{des}), which had an average of 157.5 kg ha^{-1} (5%).

In this sense, it is possible to emphasize the production of soybeans obtained via ANN in Figure 5, where (a) represents a surface graph of soybean production in kg ha^{-1} as a function of cultivars and years (two harvests), and (b) presents the soybean production level curves as a function of 16 cultivars in the years 2018-2019 and 2019-2020 (two harvests). It is noted that the lowest productions (3500 kg ha^{-1}) occurred for cultivars 2 and 14 (2018-2019 harvest), as well as 4 and 10 (2019-2020 harvest). The largest production, around 6000 kg ha^{-1}, was for the 2018-2019 harvest and cultivar 9- M 6410 IPRO.

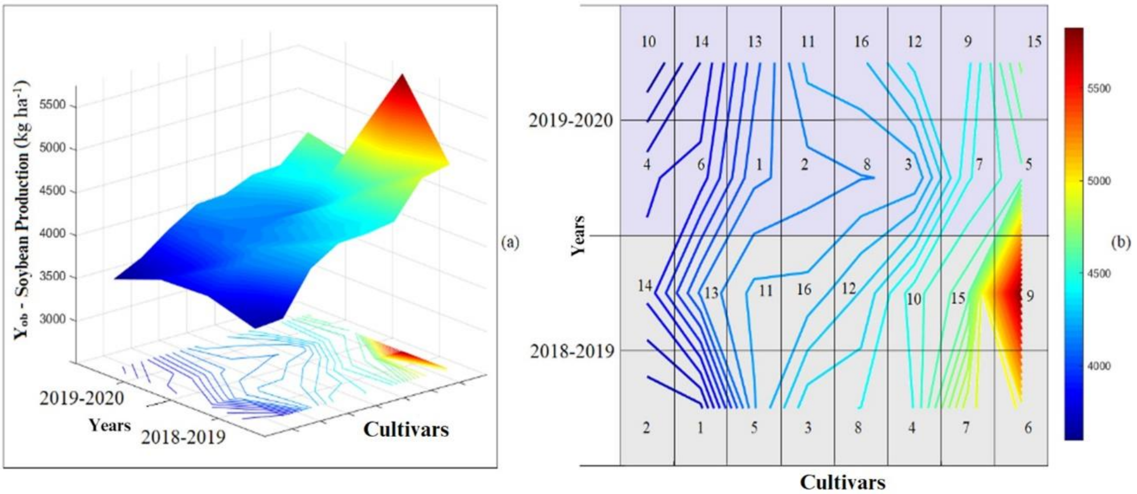


Figure 5. (a) Surface graph of soybean production as a function of crops and cultivars, (b) Soybean production level curves, represented for each cultivar and crop.

Conclusions

The study examined the use of an artificial neural network (ANN) in estimating soybean productivity, based on data on production components, morphological characteristics, and nutritional information of 16 soybean varieties, collected over two harvests. The results point to an effective performance of the ANN, with adequate weight adjustment and an average error of approximately 70.07 kg per hectare. It is important to highlight that the 2018-2019 harvest, especially in for the cultivar 9- M 6410 IPRO, recorded the highest production, reaching around 6,000 kg per hectare.

During the training, validation, and testing phases, the ANN showed mean squared errors (MSEs) of about 0.0025553 and 0.0022251, with R values of 0.690 and 0.872, respectively, compared to the desired values (Y_{des}), which had an average of 157.5 kg per hectare (5%). The confusion matrix used to evaluate the classification of categorical data revealed that, out of the 96 samples, 91 were classified correctly, while 5 were classified incorrectly, resulting in a hit rate of 94.8% and an error rate of 5.2%.

These results demonstrate the effectiveness of ANN in predicting soybean production based on plant morphological and nutritional information, allowing a better understanding of the factors that affect crop yield. The ANN approach also provides valuable information for decision-making in crop management, contributing to the productive potential of soybeans in the context of agriculture.

Acknowledgments

The authors would like to thank UNESP for the opportunity to contribute to the scientific community.

References

Alcarde, A. C.; Stape, J. L.; Sentelhas, P. C.; Gonçalves, J. L. M.; Sparovek, G. Köppen's climate classification map for Brazil. Meteorologische Zeitschrift. v. 22, n. 6, p. 711 - 728. 2013. 10.1127/0941-2948/2013/0507.

Anagu, I.; Ingwersen, J.; Utermann, J.; Streck, T. Estimation of heavy metal sorption in German soils using artificial neural networks. Geoderma, v. 152, Issues 1–2,15, p. 104-112. 2009. 10.1016/j.geoderma.2009.06.004.

Beuchera, A.; Siemssen, R.; Fröjdö, S.; Österholm, P.; Martinkauppi, A.; Edén, P. Artificial neural network for mapping and characterization of acid sulfate soils: Application to Sirppujoki River catchment, southwestern Finland. Geoderma. v. 247–248, p. 38–50. 2015. 10.1016/j.geoderma.2014.11.031.

Bonini Neto, A.; Fávaro, V. F. S.; Santos, W. P. L.; Mello, J. M.; Angela, A. V. Radial base neural network for the detection of banana maturation stages: perceptron multilayer network comparison. Brazilian Journal of Biosystems Engineering (UNESP), v. 16, p. 1-7, 2022. 10.18011/bioeng.2022.v16.1175.

Bonini Neto, A.; Moreira, A.; Bonini, C. S. B.; Campos, M.; Andrighetto, C. Fuzzy Logic and Artificial Neural Network Perceptron Multi-Layer and Radial Basis in Estimating Marandu Grass Yield in Integrated Systems. Communications in Soil Science and Plant Analysis, v. -, p. 1-12, 2023. 10.1080/00103624.2023.2252839.

Bonini Neto, A.; Criscimani, A. L.; Bonini, C. S. B.; Souza, J. F. D.; Oliverio, G. L.; Baretto, V. C. M.; Andrighetto, C. Artificial neural networks applied to the marandu grass production estimate in integrated systems. Brazilian Journal of Biosystems Engineering (UNESP), v. 15, p. 318-341, 2021. 10.18011/bioeng2021v15n2p318-341.

Boote, Kenneth J.; Jones, James W.; Pickering, Nigel B. Potential uses and limitations of crop models. Agronomy jornal. v. 88, n. 5, p. 704-716, 1996. 10.2134/agronj1996.00021962008800050005x.

Braga, A. P.; Carvalho, A. P. L. F.; Ludermir, T. B. Redes neurais artificiais: teoria e aplicações. 2. ed. Rio de Janeiro: LTC Editora, 2007. ISBN 8521615647

Eliasmith, C.; Anderson, C. H. Neural engineering: Computation, representation, and dynamics in neurobiological systems. MIT Press, Cambridge, MA, 2003. ISBN 9780262550604.

Embrapa - Cultivares de soja da Embrapa. Available at https://www.embrapa.br/cultivar/soja. Access: October 2023.

Haykin, S. Neural networks: a comprehensive foundation. 2. ed. Tsinghua University Press. 2001. ISBN 0132733501.

Hoeft, R.G. Desafios para a obtenção de altas produtividades de milho e de soja nos EUA. Piracicaba: Potafos, 2003. p.1-4. (Informações Agronômicas, 104).

IBGE - Instituto Brasileiro de Geografia e Estatística. Available at https://www.ibge.gov.br/. Access: October 2023.

Kamali, M, Hewage, K. Development of performance criteria for sustainability evaluation of modular versus conventional construction methods. J Clean Prod, v. 142, p. 3592-360620 2017. 10.1016/j.jclepro.2016.10.108.

Kovacs, Z. L. Redes Neurais Artificiais: Fundamentos e Aplicações: Um texto básico. 4ª ed. Editora Livraria da Física. 177 p., 2006. ISBN 8588325144.

Mathworks. Available at https://www.mathworks.com. Access: March 2022.

Moreira, A., Bonini Neto, A., Bonini, C. S. B., Moraes, L. A. C., Heinrichs, R. Prediction of soybean yield cultivated under subtropical conditions using artificial neural networks. Agronomy Journal, v. 115, p. 1981-1991. 2023. 10.1002/agj2.21360

Mouazen, A. M.; Kuang, B.; De Baerdemaeker, J. And Ramon, H. Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. Geoderma, v. 158, p. 23-31, 2010. 10.1016/j.geoderma.2010.03.001.

Putti, F. F.; Gabriel Filho, L. R. A.; Gabriel, C. P. C.; Bonini Neto, A.; Bonini, C. S. B.; Reis, A. R. A Fuzzy mathematical model to estimate the effects of global warming on the vitality of Laelia purpurata orchids. Mathematical Biosciences, v. 288, p. 124-129, 2017. 10.1016/j.mbs.2017.03.005.

Rummelhart, D. E.; Mcclelland, J. L. PDP Research Group. Parallel Distributed Processing - Explorations in the Microstructure of Cognition. v. 1: Foundations. A Bradford Book - The MIT Press. 1986. 10.7551/mitpress/5236.001.0001.

Silveira, C. T.; Oka-Fiori, C.; Santos, L. J. C.; Sirtoli, A. E.; Silva, C. R.; Botelho, M. F. Soil prediction using artificial neural networks and topographic attributes. Geoderma, v. 195–196, p. 165-172. 2013. 10.1016/j.geoderma.2012.11.016.

Souza, A. V.; Bonini Neto, A.; Piazentin, J. C.; Dainese Junior, B. J.; Gomes, E. P.; Bonini, C. S. B.; Putti, F. F. Artificial neural network modelling in the prediction of banana's harvest. Scientia Horticulturae, v. 257, p. 108724, 2019. 10.1016/j.scienta.2019.108724.