THAIS DO PRADO HASHIMOTO

SEGMENTED MODEL AS PRIOR INFORMATION FOR THE APPLICATION OF ARTIFICIAL NEURAL NETWORKS TO CLASSIFTY SOYBEAN GENOTYPES IN TERMS OF PHENOTYPIC ADAPTABILITY AND STABILITY

Dissertation submitted to the Genetics and Breending Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Magister Scientiae*.

Adviser: Moysés Nascimento

Co-advisers: Ana Carolina C. Nascimento
Eder Matsuo

VIÇOSA - MINAS GERAIS
2023

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

H348s
2023

Hashimoto, Thais do Prado, 1998-

Segmented model as prior information for the application of artificial neural networks to classifty soybean genotypes in terms of fenotypic adaptability and stability / Thais do Prado Hashimoto. – Viçosa, MG, 2023.

1 dissertação eletrônica (30 f.): il.

Texto em inglês.

Orientador: Moysés Nascimento.

Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Agronomia, 2023.

Referências bibliográficas: f. 28-30.

DOI: https://doi.org/10.47328/ufvbbt.2023.577

Modo de acesso: World Wide Web.

1. Soja - Melhoramento genético. 2. Mapeamento cromossômico - Métodos estatísticos. 3. Redes neurais (Computação). I. Nascimento, Moysés, 1979-. II. Universidade Federal de Viçosa. Departamento de Agronomia. Mestrado em Genética e Melhoramento. III. Título.

CDD 22. ed. 631.52

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

THAIS DO PRADO HASHIMOTO

SEGMENTED MODEL AS PRIOR INFORMATION FOR THE APPLICATION OF ARTIFICIAL NEURAL NETWORKS TO CLASSIFTY SOYBEAN GENOTYPES IN TERMS OF PHENOTYPIC ADAPTABILITY AND STABILITY

Dissertation submitted to the Genetics and Breending Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Magister Scientiae*.

APPROVED: July 18, 2023.

Assent:

Documento assinado digitalmente
gov.br
THAIS DO PRADO HASHIMOTO
Data: 05/10/2023 19:38:30-0300
Verifique em https://validar.iti.gov.br

Thais do Prado Hashimoto
Author

Documento assinado digitalmente
gov.br
MOYSES NASCIMENTO
Data: 05/10/2023 13:48:10-0300
Verifique em https://validar.iti.gov.br

Mosyés Nascimento
Adviser

To God, my family, my friends and all the teachers I had during my academic life.

"The pleasure in learning, is an incentive for the pursuit of any knowledge. And this pleasure can take you so close to those who have the pleasure of teaching al-ways."

Eronildo Paulino.

ACKNOWLEDGEMENTS

To God, for having given me this realization of this dream.

"I am grateful to my spirit guides for helping me find my path and move forward with courage."

To everyone who participated, directly or indirectly, in the development of this research work, enriching my learning process.

To the people come I lived with throughout the years of this course, who encouraged me and certainly had an impact on my academic training. My heroes are humans who overcame the most diverse difficulties to raise me. I love you mom and dad. To my parents and pillars of my life, Gizele Maria do Prado Hashimoto and Ricardo Hashimoto.

My sister, Heloisa of the Hashimoto Meadow, you where the best gift God gave me, a wonderful sister and a true friend.

My family and friends for the incentive.

I dedicate to my friends that I made during my stay in Viçosa/ Arapongas, João Fernandes, Tales Henrique, Luis Bahia, Mariana Fajardo, Larissa Macedo, Leonardo Henriques, Thainá Costa, Victor Silva Signorini, João Souza, Emilene Pedrero, Nicoly Humai, Míria Grasiele, Ana Carolina Melo, David Lima Batista, Luisa Crauss, Fernando Teixeira, Nicolo Lenner, Hermes Strassacapa, Ana Konkol, Eleniz Dias, Gabriela França, Paloma Brás, Dr. Maicon Nardino, my Masters friends, Hirlanda Brito, Bruna de Paula, Cynthia Barreto, Pedro Medeiros, Carla Fernandes, Vitor Sagae, Matheus Suela, Aline Marçal, Wagner Barbosa, Vinicius Begnami, Sheila Faria, Diego Souza, Mauricio Celeri, Stephanie Locatelli, Andréa Bastos Andrade, Bárbara Antunes, Edson Amorim and Rafael Guimarães my friends Maranhenses, Teacher Lygia Barranco da Silveira thank you for helping me with my English classes, for your support and affection! And all the other friends I made during the master's, thank you for the companionship my stay in Viçosa was more fun with you!

My second Family my heart, Dona Shiley, Senhor Vitório, Adriana Vidal, and I extend to the Regassins brothers and Dona Risa and the whole family. To my advisor Dr. Moysés Nascimento, for friendship, patience, encouragement, concern and knowledge.

To my advisor of PIBIC/CNPq, Dr. Nelson da Silva Fonseca Júnior, for friendship, patience, encouragement and also for presenting me statistics, and show me how the science of variation made me find myself in the area of Agronomy.

To my undergraduate friends and to all the teachers I had during my academic life, for the shared knowledge prior to my master's degree.

To my friends of the Republic Intact, for the companionship, friendship and the welcome, my days were more fun with you!

To my friend at heart João Fernandes, for the companionship, and for always supporting me in my decisions, and being by my side.

My friend Neide Noguchi, for the friendship I made during graduation and always, propel me in my dreams.

To the professors and staff of the Genetics and Breeding department and the Statistics department for their teachings, patience, friendship and encouragement. To my friends at the Laboratory of Computational Intelligence and Statistical Learning (LICAE) for friendship, encouragement and teachings.

To my co-supervisors Dra. Ana Carolina Campana Nascimento, Dr. Eder Matsuo for his contribution to my learning and patience. The Fundação MS para a Pesquisa e Difusão de Tecnologias Agropecuárias for cooperation and for making the data available for the accomplishment of this work, in special do Dr. André Ricardo Gomes Bezerra.

To the members of the board, Professor Dr,. Moysés Nascimento, Professor Dr, Kaio Olimpo das Graças and Professor Eder Matsuo.

To the Federal University of Vicosa, for the opportunity and realization of my dream of being a Master.

Capes, CNPq, FAPEMIG.

To the Federal University of Viçosa, for the opportunity to complete the postgraduate course.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. To the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), to granting the scholarship.

I am grateful to my spirit guides for reminding me that there is a higher purpose in my life and that I can make a difference in this world."

I thank everyone who contributed directly or indirectly to the realization of this work, dream and professional and personal fulfillment.

Gratitude!!!

“Kashikoku jikan wo toushi shiyo”

Japanese wisdom

ABSTRACT

HASHIMOTO, Thais do Prado Hashimoto, M.Sc., Universidade Federal de Viçosa, july 2023. **Segmented model as prior information for the application of artificial neural networks to classify soybean genotypes in terms of phenotypic adaptability and stability**. Advisor: Moysés Nascimento. Co-advisers: Ana Carolina Campana Nascimento and Eder Matsuo.

Unlike models based on simple linear regressions, segmented models can better assess the adaptability and stability of genotypes, which can demonstrate a non-linear pattern of response to environmental variation. Therefore, this work aimed to transpose the concepts of adaptability and stability from the statistical analysis of a segmented model to the strong discriminatory potential of an artificial neural network (ANN) and use it to classify soybean genotypes *Glycine max*. A total of 9,000 simulated soybean genotypes were previously arranged into 18 different classes, which represented the combination of nine adaptability classes by the method of Verma and collaborators (VCM) and two stability classes by the method of Finlay & Wilkinson. There was 90% agreement between the ANN and VCM analyses regarding adaptability classification and 20% regarding stability. With the methods presented in this work, it was demonstrated that the potential of using ANNs to evaluate the adaptability of genotypes is strong. These auxiliary parameters were used in an algorithm programmed in the R software using the *nnet* function of the *nnet* package to find an ANN configuration whose maximum classification error in the testing phase was 1%. After choosing the ANN model with the smallest error, the set of real soybean genotypes was submitted to it for classification in terms of adaptability and stability. The R codes used in this manuscript are available at https://github.com/licaeufv. An ANN based on a segmented model as the VCM model were powerful to classify soybean genotypes regarding their adaptability and, possibly, can help breeders interpret data from the behavior of any cultivar in face of environmental variations considering adapted ANN models for each situation. In addition, since the stability was introduced in the ANN as a different concept from that used to classify the genotypes by the (VCM) statistical method, such classification needs to be reviewed and further improved.

Keywords: *Glycine Max*. Artificial Intelligence. Genotypes × Environments Interaction. Data Simulation. Bioinformatics. Artificial Neural Network.

RESUMO

HASHIMOTO, Thais do Prado Hashimoto, M.Sc., Universidade Federal de Viçosa, julho de 2023. **Modelo segmentado como informação prévia para aplicação de redes neurais artificias para classificar genótipos de soja quanto a adaptabilidade e estabilidade fenotípicas.** Orientador: Moysés Nascimento. Coorientadores: Ana Carolina Campana Nascimento e Eder Matsuo.

Diferentemente dos modelos baseados em regressões lineares simples, os modelos segmentados podem avaliar melhor a adaptabilidade e a estabilidade dos genótipos, que podem demonstrar um padrão não linear de resposta à variação ambiental. Portanto, este trabalho teve como objetivo transpor os conceitos de adaptabilidade e estabilidade da análise estatística de um modelo segmentado para o forte potencial discriminatório de uma rede neural artificial (ANN) e utilizá-lo para classificar genótipos de soja *Glycine max*. Um total de 9.000 genótipos de soja simulados foram previamente organizados em 18 classes diferentes, que representavam a combinação de nove classes de adaptabilidade pelo método de Verma e colaboradores (VCM) e duas classes de estabilidade pelo método de Finlay & Wilkinson. Houve 90% de concordância entre as análises ANN e VCM quanto à classificação de adaptabilidade e 20% quanto à estabilidade. Com os métodos apresentados neste trabalho, foi demonstrado que o potencial do uso de RNA para avaliar a adaptabilidade de genótipos é forte. Esses parâmetros auxiliares foram utilizados em um algoritmo programado no software R utilizando a função nnet do pacote nnet para encontrar uma configuração de RNA cujo erro máximo de classificação na fase de testes fosse de 1%. Os códigos R utilizados neste manuscrito estão disponíveis em https://github.com/licaeufv. Uma RNA baseada em um modelo segmentado como o modelo VCM foi poderosa para classificar genótipos de soja quanto à sua adaptabilidade e, possivelmente, pode auxiliar os melhoristas a interpretar dados do comportamento de qualquer cultivar frente às variações ambientais considerando modelos de RNA adaptados para cada situação. Além disso, como a estabilidade foi introduzida na RNA como um conceito diferente daquele utilizado para classificar os genótipos pelo método estatístico (VCM), essa classificação precisa ser revista e melhorada.

Palavras-chave: *Glycine Max*. Inteligência Artificial. Interação Genótipos × Ambientes. Simulação de Dados. Bioinformática.

SUMMARY

GENERAL INTRODUCTION

In plant breeding, when the objective is to select and or recommend genotypes for planting, the existence of genotypic interaction by environments (G x E) is one of the greatest difficulties (Cruz et al., 2013; Barroso et al, 2013). If it exists, it is possible that superior genotypes in one environment are not in another (Nascimento et al., 2021; Eeuwijk et al., 2016). For the environments, managed stress treatments can be treated as fixed when they refer to (hopefully) repeatable environmental conditions. Similarly, repeatable G x E for a small number of genotypes under various levels of a well-defined managed stress factor will be fixed (Nascimento et al., 2021; Eeuwijk et al., 2016). For locations and years, we prefer to take the main effects, which are a kind of intercept terms, fixed, where the year main effects may be taken random when it concerns many years. For the G x E interactions, the genotype ´ location interaction is fixed for repeatable locations and selected genotypes (Annicchiarico, 1997; Annicchiarico et al., 2005) (Nascimento et al., 2021; Eeuwijk et al., 2016)

The literature presents several adaptability and stability methodologies that allow the identification and recommendation of superior cultivars in different environments. These methodologies differ in relation to statistical principles, biometric procedure and consequently in the interpretation of the results obtained. Specifically, the methods can be derived from statistical techniques such as analysis of variance (Yates and Cochran,1938; Plaisted and Peterson,1959; Wricke,1965), simple regression analysis (Finlay and Wilkinson,1963; Eberhart and Russell,1966; Tai,1971), segmented (Verma, Chahal and Murty,1978; Cruz, Torres and Vencovsky, 1989), quantile (Barroso et al, 2015) and non-parametric (Nascimento et al., 2010). Bayesian methods (Couto et al., 2015; Nascimento et al., 2011; Nascimento et al., 2020) and multivariate analyses (Nascimento et al., 2010; Nascimento et al., 2015; Yan et al. (2000); Gauch, 2006). Besides these non-parametric methods were also used (Lin Binns, 1988; Huehn, 1990; Annicchiarico, 1992; Rocha, Muro-Abad, Araujo and Cruz, 2005).

The use of methods based on Computational Intelligence, for example, in artificial neural networks (ANN) is increasing. Briefly, RNAs are models that function as a network of biological neurons capable of processing a large amount of data through self-learning (Haykin, 2009). Compared to a statistical framework, RNAs have

the advantage of not requiring a priori assumptions about the model, which allows their adjustment to the most diverse problems (ROSADO et al., 2022). In the context of adaptability and stability studies, Nascimento et al. (2013) proposed a methodology for the classification of genotypes by artificial neural networks (ANN). In this methodology the classification was based on the learning of the network based on classes of pre-defined recommendations according to the methodology of Eberhart and Russell (1966). The authors justified the use of the methodology of Eberhart and Russell (1966) for its wide use in plant breeding due to its easy application and interpretation. This methodology has been successfully used to study adaptability and stability in several crops, such as cowpea (Teodoro et al, 2015), Papaya (Luz et al., 2018) and soybean (Oda et al, 2022).

Although interesting, a critique of the methods for the study of adaptability and stability based on simple linear regression models refers to the existence of a possible nonlinear pattern of genotype responses to environmental variation (Ferreira et al. 2006). In order to obtain a solution to this question, Verma et al. (1978) proposed the segmented regression model based on two linear regression coefficients which are used to separate the evaluation of the responses of cultivars to unfavorable and favorable environments. Therefore, segmented regression allows to find the "ideal" genotype, which has high productive performance, high stability and low sensitivity to adverse conditions. This proposal already used under the frequentist context (Verma et al., 1978; Cruz et al., 1989) and Bayesian (Nascimento et al., 2020) has not yet been elucidated under the Artificial Intelligence paradigm. Given the above, the objective of this study was: (1) to propose the development of a segmented model based on RNA to evaluate the adaptability and stability of genotypes; (2) the application of this model for the classification of soybean genotypes (*Glycine max* (L.) Merr.) using real data; (3) To compare the classification of soybean genotypes as to their adaptability and stability with the results from the segmented model proposed by Verma et al (1978).

Chapter 1 - Artificial neural networks based on segmented model for adaptability and stability evaluation of soybean genotypes

ABSTRACT - Unlike models based on simple linear regressions, segmented models can better assess the adaptability and stability of genotypes, which can demonstrate a nonlinear pattern of response over the environmental variation. However, these methods can be under statistical limitations such as the increase of the Type Error II and biased estimates. Therefore, this work aimed to transpose the concepts of adaptability and stability from the statistical analysis of a segmented model to the strong discriminatory potential of an artificial neural network (ANN) and use it to classify soybean (*Glycine max* (L.) Merr.) genotypes. An ANN training was carried out with the grain yield of 7,200 soybean genotypes simulated in 15 different environments; the ANN topology chosen was the one that had less than 1% of error in the testing phase with 1,800 simulated genotypes. A total of 9,000 simulated soybean genotypes were previously arranged in 18 different classes, which represented the combination of nine classes of adaptability by the Verma and collaborators (VCM) method and two classes of stability (invariability concept) by the Finlay & Wilkinson (FW) method. Finally, the grain production of ten real soybean genotypes was inputted into the ANN trained model, and the classification regarding adaptability and stability was obtained. There was 90% agreement between the ANN and VCM analyses regarding the adaptability classification and 20% regarding stability. With the methods presented at this work, it was demonstrated that the potential of using ANNs to assess the adaptability of genotypes is strong. In addition, since the and stability was introduced in the ANN as a different concept from that used to classify the genotypes by the statistical method, such classification needs to be reviewed and further improved.

Keywords: *Glycine max*; artificial intelligence, genotypes × environments interaction; data simulation, bioinformatics.

1. INTRODUCTION

The use of methods based on machine learning in agronomy, such as artificial neural networks (ANN), is increasing (SOUSA et al., 2022). Briefly, ANNs are models

that work as a network of biological neurons capable to process a large amount of data using self-learning (Haykin, 2009). Compared with a statistical framework, ANNs have the advantage of not requiring *priori* assumptions about the model, which allows their adjustment to the most diversified problems (ROSADO et al., 2022).

It is not surprising, therefore, that ANN models have been used in breeding programs to predict genetic values of animals and plants (ABDOLLAHI-ARPANAHI et al., 2020; ROSADO et al., 2020). Still, in the context of plant breeding, the use of ANN models has also become an interesting approach to deal with the interaction between genotypes and environments in multi-environmental trials (MET's) (ALVES et al., 2019; NASCIMENTO et al., 2013). For exemple, NASCIMENTO et al. (2013) proposed a methodology of adaptability and phenotypic stability based on the training of an ANN, considering the methodology of Eberhart & Russell (ER). The authors chose the ER method since it is widely used because of its simplicity and efficiency for analyzing MET's (Janick, 2003).

Although attractive, the ER is based on the fit of simple linear regression models (Cruz et al., 2012). Therefore, it does not allow to study the potential nonlinear pattern of genotype responses throughout the environmental variation (NASCIMENTO et al., 2020) and makes ER-based ANN models equally deficient in assessing the adaptability and stability of genotypes. On the other hand, segmented regression models like the one proposed by VERMA et al. (1978) (VCM) are able to distinctly evaluate the performance of genotypes in unfavorable and favorable environments. This allows such models finding the "ideal" genotype which should present low sensitivity to adverse conditions and increasing yield as the environment improves, besides a high stability.

As mentioned in previous works, the use of ANN models to assess adaptability and stability is preferred to avoid the statistical limitations of the simpler methods, such as biases in the estimates of regression coefficients and the increase of Type Error II (Nascimento et al., 2013; Teodoro et al., 2015). Thus, joining the improvement aspects of the VCM method with an ANN approach, a new method can be established to remove the limitations described above. In addition, as proposed by NASCIMENTO et al. (2013), the ANN approach can also include an adapted method based on the

FINLAY & WILKINSON (1963) method (FW) to assess stability, which is based in the invariant classify of a given genotype after data linearization.

Therefore, because of the issues raised above, this study (1) proposed the development of an ANN-based segmented model to evaluate the adaptability and stability of genotypes and (2) the application of such model for the classification of soybean (*Glycine max* (L.) Merr.) genotypes using real data. Finally, the classification of the soybean genotypes regarding to their adaptability and stability could be compared with that from regular statistical analysis, i.e., that upon which the neural network was based.

2. RESULTADS AND DISCUSSION

The analysis of variance (Table 2) indicated that the soybean genotypes presented distinguished performances in the face of different environmental conditions, which is attested by the significant interaction ($P \leq 0.01$) between genotypes, environments and GxE interaction.

Table 2. Variance analysis of the grain yield (kg ha^{-1}) of the real soybean genotypes (*G. max*).

FV	GL	SQ	QM	F
Blocks/Environments	30	4252146.55	141738.22	
Genotypes	9	8535633.04	948403.67	6.56**
Environments	14	90127512.44	6437679.46	45.42**
Genotypes × Environments	126	18211847.33	144538.47	3.13**
Error	270	12457753.65	46139.83	
TOTAL	449	133584893.00		

**Significant at 1% probability by F test. The coefficient of variation was 5.27 %.

Under such situation, studies of adaptability and stability become necessary to detail the behavior of each genotype within the different environments evaluated as pointed out by CRUZ et al. (2012).

In addition, as the methods implemented require that genotypes are analyzed in both unfavorable and favorable environments, the variance analyses were also performed considering these two conditions. In these separated analyses, the MSE for unfavorable and favorable environments were 50,141.65 (df = 108) and 43,471.95 (df = 162), which were used to implement the normally distributed random deviations into the grain production and test the angular coefficients of the estimated ER models of the simulated genotypes (these analyses are not shown since they are not required to interpret the results, but just to provide estimates for data simulation). In possession of the simulated data set, ANN models were trained and tested. A model with 15 neurons in the single hidden layer and 0.94% errors in the testing phase, which converged after 528 iterations, was selected, and subsequently used to classify the real soybean genotypes regarding adaptability and stability.

None of these genotypes was considered “ideal” among those that presented productivity above the overall mean of the environments (4,074.78 $kg\ ha^{-1}$) (Table 3). The best scenarios were found for the genotypes CZ 26B36 IPRO, DM 66I68 RSF IPRO, and ST 644 IPRO, which were classified as having general adaptability and high stability/invariability by the ANN output. In addition, the M 6210 IPRO genotype was considered exclusively responsive to favorable environments by the VCM method and the ANN (Table 3).

Table 3. Mean grain yield and classification regarding adaptability and stability of 10 soybean genotypes (*G. max*) evaluated in 15 environments in the State of Mato Grosso do Sul by the methods of Verma et al. (1978) and artificial neural networks (ANN).

	Verma, Chahal & Murty	ANN

Genotype	Grain yield (kg ha^{-1})[1]	Adaptabilty[2] (Unf. | Fav.)	Stability[3] (Both)	Adaptabilty[2] (Unf. | Fav.)	Stability[4] (Both)
AS 3680 IPRO	3993.02	(=1 | =1)	High	(=1 | =1)	High
BRASMAX GARRA IPRO	3942.01	(>1 | =1)	Low	(>1 | =1)	High
BMX POTÊNCIA RR	3919.56	(=1 | =1)	Low	(=1 | =1)	High
BS 2606 IPRO	3958.68	(<1 | =1)	Low	(=1 | =1)	Low
CZ 26B36 IPRO	**4075.17**	(=1 | =1)	Low	(=1 | =1)	High
DM 66I68 RSF IPRO	**4301.00**	(=1 | =1)	Low	(=1 | =1)	High
M 6210 IPRO	**4152.15**	(=1 | >1)	Low	(=1 | >1)	High
M 6410 IPRO	**4269.23**	(=1 | <1)	Low	(=1 | <1)	High
ST 644 IPRO	**4196.45**	(=1 | =1)	Low	(=1 | =1)	High
TEC 7022 IPRO	3940.54	(=1 | =1)	Low	(=1 | =1)	High
Overall mean	4074.78				

[1]Average of the genotype above the overall mean of the experiment is in bold. [2]Adaptability based on the method of Verma, Chahal & Murty (1978) in terms of β_{1i} values, respectively, for unfavorable (left) and favorable (right) environments. [3]Stability based on reclassification for both environments according to Table 1. [4]Stability based on invariability after data linearization (adapted from Finlay & Wilkinson method).

The ANN showed 90% agreement with the VCM method to discriminate the adaptability of soybean genotypes. However, only 20% agreement on stability (Table 3). The low agreement in terms of the stability parameter can be explained by the difference in the concept of this parameter used in each of these approaches, one based on invariance and the other based on regression deviations (NASCIMENTO et al., 2013). The VCM method mirrors the concept of stability from its auxiliary method, i.e., the ER method, which is applied separately in unfavorable and favorable environments. Its stability concept is based on the predictability of genotype behavior (simplified in Table 1). This stability concept differs from that used by ANN, which was based on the invariance of genotype behavior after linearization of the data, performing

an adaptation of the FW method. However, a comparison with the FW method is not feasible as it is not a bi-segmented regression.

As seen in other previous attempts, TEODORO et al. (2015) showed greater agreement between the ER and ANN methods with 100 and 70%, respectively, regarding the phenotypic adaptability and stability of cowpea (*Vigna unguiculata* (L.) Walp.) genotypes and NASCIMENTO et al. (2013) reported respective 93 and 85% for alfalfa (*Medicago sativa* L.) genotypes. These results showed that our findings are even better than those previously reported for the adaptability classification by the segmented model-based ANN. Finally, the interesting point is that using neural networks to assess phenotypic adaptability and stability allows simulating genotypes based on different methodologies. In this way, it is possible to create networks that classify genotypes based on different concepts according to the researcher's interest.

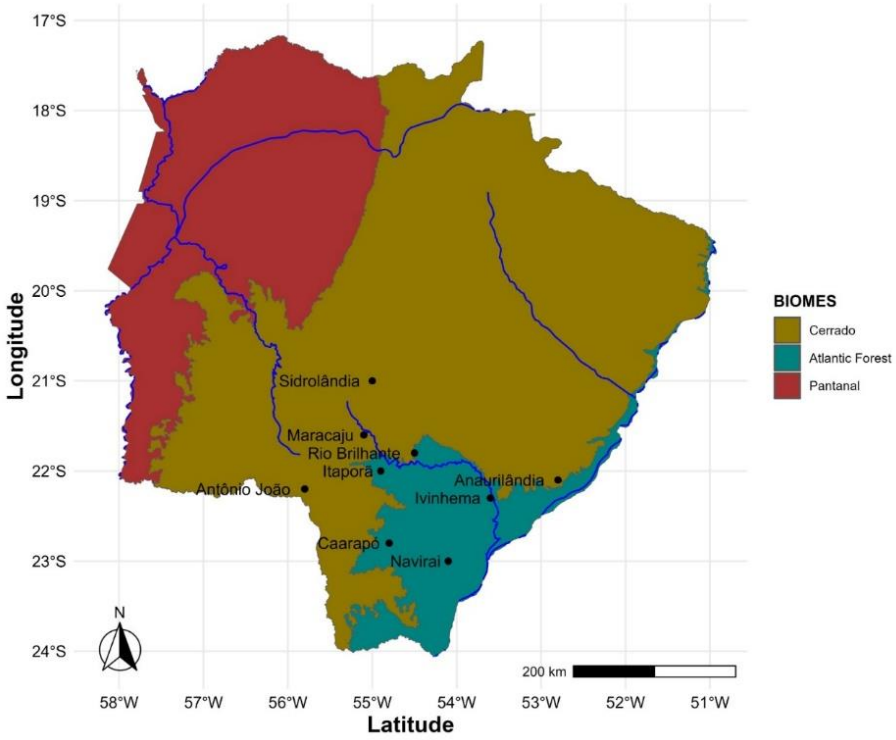
3. MATERIALS AND METHODS

Firstly, it worth to mention that to stablish the ANN based on VCM and FW methods as proposed in this work, a real data needed to be present followed by statistical analysis in which some estimated parameters were used to simulate the training and testing data used by the ANN models.

Experimental data

The data used here were obtained from experiments carried out by the Phytotechnics sector of the Fundação MS para Pesquisa e Difusão de Tecnologias Agropecuárias with ten soybean cultivars named as AS 3680 IPRO, BRASMAX GARRA IPRO (63i64RSF IPRO), BMX POTÊNCIA RR, BS 2606 IPRO, CZ 26B36 IPRO, DM 66I68 RSF IPRO, M 6210 IPRO, M 6410 IPRO, ST 644 IPRO and TEC 7022 IPRO. The cultivars were planted in the crop year 2020/2021 in 15 different experimental areas distributed among nine municipalities (Anaurilândia (22°08'S; 52°45'W; 370 msnm), Antônio João (22°10'S; 55°46'W; 630 msnm), Caarapó (22°45'S; 54°47'W; 390 msnm), Itaporã (22°03'S; 54°55'W; 400 msnm), Ivinhema (22°20'S; 53°39'W; 370 msnm), Maracaju (21°38'S; 55°06'W; 360 msnm), Navirai (22°59'S; 54°06'W; 370 msnm), Rio Brilhante (21°50'S; 54°32'W; 310 msnm) and Sidrolândia (21°00'S; 54°59'W; 450 msnm)) of the State of Mato Grosso do Sul, Brazil.

The experiments were carried out under randomized blocks (3 blocks) and the experimental unit consisted of four 5.0 m long rows spaced 0.5 m from each other. The useful area of each plot was 4.0 m^2, with the two central rows being harvested, discounting 0.50 m of border at the ends to obtain the production per plot. With this result it was obtained the yield productivity in kg ha^{-1}.



Font: Program R

Analyses of variance

The soybean yield ($kg\ ha^{-1}$) data were submitted to a joint analysis of variance in the software Genes (Cruz, 2013). The model adopted for the analysis was $Y_{ijk} = \mu + R/E_{k(j)} + G_i + E_j + GE_{ij} + \xi_{ijk}$, where Y_{ijk} is the phenotypic mean; μ is the general mean; $R/E_{k(j)}$ is the effect of the kth repetition (block) in the jth environment; G_i is the fixed effect of the ith genotype; E_j is the effect of the jth environment Normally

Independently Distributed (NID); GE_{ij} is the effect of the interaction of the ith genotype in the jth environment NID with mean equal to 0 and variance denoted by σ^2_{ge}; and ξ_{ijk} is the experimental error NID with mean equal to 0 and variance denoted by σ^2_e. In addition, variance analyses considering the same model as above-mentioned were carried out with the data, which were previously split into according to unfavorable and favorable environments (more details are given below).

Segmented model for adaptability evaluation

The VCM method characterizes the adaptability and stability of genotypes by the interpretation, respectively, of the angular coefficients and the regression deviation of two simple linear regressions, which are estimated for each genotype in unfavorable and favorable environments. Then, initially, it is necessary to recognize the unfavorable and favorable environments through the environmental index (I_j) given by $I_j = \frac{1}{g}\sum_i^g \quad Y_{ij} - \frac{1}{ga}\sum_i^g \quad \sum_j^a \quad Y_{ij}$.

Once the environments are identified, the response of each genotype to environmental variations can be analyzed within two distinct groups of environments (unfavorable and favorable) through a simple regression model, which, in this case, is based on the ER method. The statistical model considered by these authors is defined by $Y_{ij} = \beta_{0i} + \beta_{1i}I_j + \psi_{ij}$; where Y_{ij} is the mean of the ith genotype in the jth environment; β_{0i} is the regression coefficient that measures the response of the ith genotype throughout the environments; I_j is the environmental index; and ψ_{ij} is the random effect, which is decomposed as follows $\psi_{ij} = \delta_{ij} + \underline{\varepsilon_{ij}}$, where δ_{ij} is the regression deviation and $\underline{\varepsilon_{ij}}$ is the mean experimental error. The estimators of the adaptability and stability parameters are defined, respectively, by $\hat{\beta}_{1i} = \frac{\sum_j \quad Y_{ij}}{\sum_j \quad I_j^2}$

$and\ \hat{\sigma}^2_{d_i} = \frac{MSD_i - MSE}{r}$, where MSD_i is the mean square deviation of the ith genotype, MSE is the mean square of the error and, r is the number of repetitions. In this case, since the ER method is applied in two groups of environments, the j number of environments has size a_u and a_f, denoting, respectively, the number of unfavorable and favorable environments. In addition, thirty-six classes of genotypes are possible to be generated from the arrangement [i.e., $(3 \times 3) \times (2 \times 2)$] among parametric values

of β_{1i} and $\sigma^2_{d\,i}$ (Table 1), which are tested under the respective hypotheses: $H_{0(\beta_1)}$: $\beta_{1i} = 1$ versus $H_{1(\beta_1)}$: $\beta_{1i} \neq 1$ (Student's t-test) and $H_{0(\sigma^2_d)}$: $\sigma^2_{di} = 0$ versus $H_{1(\sigma^2_d)}$: $\sigma^2_{di} > 0$ (F-test), considering $\alpha = 0.05$.

Table 1. Adaptability and stability classes of genotypes based on, respectively, the arrangement of the β_{1i} and $\sigma^2_{d_i}$ parametric values according to the Verma et al. (1978) method.

Adaptability Classes	Parametric values of β_{1i}		Adaptability[1]
	Unfavorable	Favorable	
1	$\beta_{1i} = 1$	$\beta_{1i} = 1$	Overall

2	$\beta_{1i} < 1$	$\beta_{1i} < 1$	Specific for unfavorable environments
3	$\beta_{1i} > 1$	$\beta_{1i} > 1$	Specific for favorable environments
4	$\beta_{1i} = 1$	$\beta_{1i} < 1$	Not recommended
5	$\beta_{1i} < 1$	$\beta_{1i} > 1$	Ideal
6	$\beta_{1i} > 1$	$\beta_{1i} = 1$	Not recommended
7	$\beta_{1i} = 1$	$\beta_{1i} > 1$	Specific for favorable environments
8	$\beta_{1i} < 1$	$\beta_{1i} = 1$	Specific for unfavorable environments
9	$\beta_{1i} > 1$	$\beta_{1i} < 1$	Not recommended

Stability Classes	Parametric values of $\sigma^2_{d_i}$		Stability[1]
	Unfavorable	Favorable	
1	$\sigma^2_{d_i} = 0$	$\sigma^2_{d_i} = 0$	High
2	$\sigma^2_{d_i} = 0$	$\sigma^2_{d_i} > 0$	Low
3	$\sigma^2_{d_i} > 0$	$\sigma^2_{d_i} = 0$	Low
4	$\sigma^2_{d_i} > 0$	$\sigma^2_{d_i} > 0$	Low

[1]Simplified adaptability and stability classes of the genotypes throughout both unfavorable and favorable environments.

Artificial neural networks based on segmented model

Initially, aiming to expand the data set for the training and testing of the network, yield data were simulated based on the information from the experiment data under study. Therefore, the values of I_j were firstly estimated for each environment and then, from its sign (positive or negative), two sets of environments were defined, one containing six unfavorable environments (i.e., $a_u = 6$ with negative values) and the other containing nine favorable ones (i.e., $a_f = 9$ with positive values). Posteriorly, 500

vectors containing values of Y_{ij} were simulated by the application of the ER model considering each possible parametric values of β_{1i} (i.e., $\beta_{1i} < 1$, $\beta_{1i} = 1$ and $\beta_{1i} > 1$) for each set of the previously categorized environments; i.e., vectors representing grain yield in unfavorable environments had six values of Y_{ij} simulated from the three possible values of β_{1i} and, vectors representing grain yield in favorable environments had nine values of Y_{ij} simulated from the three possible values of β_{1i}. Finally, the groups of vectors were arranged in nine classes by concatenating the vectors (without changing the order) obtained from each value of β_{1i} for both sets of favorable and favorable environments. This procedure created a set of 4,500 simulated genotypes representing their production behavior throughout the total of 15 environments.

The parametric values used to simulate each value of Y_{ij} according to the model $Y_{ij} = \beta_{0i} + \beta_{1i} I_j + \psi_{ij}$ were $\beta_{0i} = \underline{X}$ $\beta_{0i} = \bar{X}$ (general average of the grain yield data of the real soybean genotypes); β_{1i} = a random value generated from a uniform distribution with the parameters a and b (i.e., $U[a; b]$), where $U[0.90; 1.10]$, $U[0.00; 0.89]$ and $U[1.11; 2.00]$ were used to the respective classes $\beta_{1i} = 1$, $\beta_{1i} < 1$ ou $\beta_{1i} > 1$ and, ψ_{ij} = a random value of a Normal distribution $N(0, \hat{\sigma}^2)$ for the ith genotype in the jth environment, where $\hat{\sigma}^2$ was $\hat{\sigma}_u^2$, whether it was the estimated MSE from the variance analysis carried out with only unfavorable environments and, $\hat{\sigma}^2$ was $\hat{\sigma}_f^2$, whether it was the estimated MSE from the variance analysis carried out with only favorable environments. In addition, simulated Y_{ij} values were included in the dataset of the 4,500 genotypes only if, a bilateral t test, under the hypotheses H_0: $\beta_{1i} = 1$ versus H_1: $\beta_{1i} \neq 1$, indicated the belonging of $\hat{\beta}_{1i}$ in the possible three classes of the parameter β_{1i}.

The ANN used these set of 4,500 simulated soybean genotypes to classify the real genotypes according to the nine classes of adaptability coming from Table 1. However, since simulated values of Y_{ij} identifying the four classes of stability in (Table 1) may overlap each other when considering $\sigma_{d_i}^2 = 0$ and $\sigma_{d_i}^2 > 0$, the phenotypic stability classification by the ANN was carried out via an adapted concept of stability, which is the invariance of results after data linearization. Thus, the simulated values of Y_{ij} of the 4,500 genotypes were submitted to a logarithmic transformation producing a linearization in which regression deviations are expected to be equal to zero

(NASCIMENTO et al., 2013). These linearized Y_{ij} values were added to the set of 4,500 previously simulated genotypes (totaling 9,000 genotypes) so that an adapted concept of invariance of genotype behavior could be inserted, as proposed by FINLAY & WILKINSON (1963). Therefore, the ANN could classify a genotype of high stability/invariability if, after linearization, the classification matched the origin class before linearization, whereas, if the classification was another, the genotype was considered of low stability/invariability (NASCIMENTO et al., 2013).

Since linearization was performed simultaneously in Y_{ij} values representing the total of the 15 environments for each simulated genotype, only two classes (high and low) of adapted stability/invariability could be generated (i.e., not for each set of unfavorable and favorable environments); therefore, the ANN classification output was compared to those two simplified stability classes of Table 1 and 18 classes of genotypes were then generated by data simulation. Finally, from the total of 9,000 simulated genotypes, 80% (i.e., 400 per class) were randomly chosen and used for the training of ANN models and the remaining 20% (i.e., 100 per class) were used for the testing phase.

The ANN models were built by single-layer backpropagation neural networks (Hastie et al., 2009), in which the variables Z_i are weighted functions of the input variables X_i. The outputs Y_k are modeled as functions of these combinations. The sigmoid activation function was used in the single-layer and, the “softmax” function was used as output function. The estimation of network parameters (weights) was performed by minimizing the sum of squares of the errors using the gradient descent algorithm.

For the training of the ANN models with the simulated dataset, auxiliary parameters were also defined as the learning rate ($L = 0.0005$), the maximum number of iterations ($Iter_{max.} = 5{,}000$), the initialization interval of weights [-0.0002; 0.0002] and, the number of hidden layer neurons that varied from 4 to 15 in each ANN convergence attempt. These auxiliary parameters were used in an algorithm programmed in the R software using the *nnet* function of the *nnet* package (Venables and Ripley, 2002) to find an ANN configuration whose maximum classification error in the testing phase was 1%. After choosing the ANN model with the smallest error, the

set of real soybean genotypes was submitted to it for classification in terms of adaptability and stability. The R codes used in this manuscript are available at https://github.com/licaeufv.

4. CONCLUSION

The ANN based on a segmented model as the VCM model were powerful to classify soybean genotypes regarding their adaptability and, possibly, can help breeders interpret data from the behavior of any cultivar in face of environmental variations considering adapted ANN models for each situation. In addition, since the stability was introduced in the ANN as a different concept from that used to classify the genotypes by the (VCM) statistical method, such classification needs to be reviewed and further improved.

ACKNOWLEDGMENTS

The authors are grateful to the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Code 001, Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) for granting scholarships and Fundação MS for carrying out the field experiments.

AUTHORS' CONTRIBUTIONS

All authors contributed equally for the manuscript, critically revised it and approved its final version.

REFERENCES

Abdollahi-Arpanahi R, Gianola D, Peñagaricano F (2020) Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genetics Selection Evolution. 52: 12. https://doi.org/10.1186/s12711-020-00531-z

Alves GF, Nogueira JPG, Machado R, Ferreira S da C, Nascimento M, Matsuo E (2019) Stability of the hypocotyl length of soybean cultivars using neural networks and

traditional methods. Ciência Rural 49(3), e20180300. https://doi.org/10.1590/0103-8478cr20180300

Carvalho LP de, Teodoro PE, Barroso LMA, Farias FJC, Morello C de L, Nascimento M (2018) Artificial neural networks classify cotton genotypes for fiber length. Crop Breeding and Applied Biotechnology 18: 200-204. https://doi.org/10.1590/1984-70332018v18n2n28

Cruz, CD (2013) Genes: a software package for analysis in experimental statistics and quantitative genetics. Acta Scientiarum. Agronomy 35, 271-276. DOI:10.4025/actasciagron.v35i3.21251

Cruz CD, Regazzi AJ, Carneiro PC (2012) Métodos biométricos aplicados ao melhoramento genético. Viçosa*:* UFV*,* Imprensa Universitária*.* 514p.

Finlay KW, Wilkinson GN (1963) The analysis of adaptation in a plant-breeding programme. Australian journal of agricultural research. 14:742-754. https://doi.org/10.1071/AR9630742

Gauch HG (2006) Statistical analysis of yield trials by AMMI and GGE. Crop Sci 46:1488–1500. https://doi.org/10.2135/cropsci2005.07-0193

González-Camacho JM, Crossa J, Pérez-Rodríguez P, Ornella L, Gianola D (2016) Genome-enabled prediction using probabilistic neural network classifiers. BMC Genomics. 17: 208. https://doi.org/10.1186/s12864-016-2553-1.

Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd ed. Springer. https://doi.org/10.1111/j.1751-5823.2009.00095_18.x

Haykin S (2009) Neural networks and learning machines*,* 3ed. Pearson Education India.

Janick J (2003) Plant Breeding Reviews, Volume 24, Part 1: Long-term Selection: Maize, 24th ed. John Wiley & Sons. https://doi.org/10.1002/9780470650240

Kujawa S, Niedbała G (2021) Artificial Neural Networks in Agriculture. Agriculture .11 : 497. https://doi.org/10.3390/agriculture11060497

Lin CS, Binns MR (1988) A superiority measure of cultivar performance for cultivar x location data. Can J Plant Sci 68:193–198. https://doi.org/10.4141/cjps88-018

Nascimento M. Nascimento ACC, Silva FF e, Teodoro PE, Azevedo CF, Oliveira TRA de, Amaral-Junior AT do, Cruz CD, Farias FJC, Carvalho LP de (2020) Bayesian segmented regression model for adaptability and stability evaluation of cotton genotypes. Euphytica. 216: 30. https://doi.org/10.1007/s10681-020-2564-5

Nascimento M, Peternelli LA, Cruz CD, Nascimento ACC, Ferreira R de P, Bhering LL, Salgado CC (2013) Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. Crop Breeding and Applied Biotechnology. 13: 152-156. DOI:10.1590/S1984-70332013000200008

Rosado RDS, Cruz CD, Barili LD, Carneiro JE de S, Carneiro PC, Carneiro VQ, Silva, JT da, Nascimento M (2020) Artificial Neural Networks in the Prediction of Genetic Merit to Flowering Traits in Bean Cultivars. Agriculture. https://doi.org/10.3390/agriculture10120638

Rosado RDS, Penso GA, Serafini GAD, Magalhães dos Santos CE, Picoli EA de T, Cruz CD, Barreto CAV, Nascimento M, Cecon PR (2022) Artificial neural network as an alternative for peach fruit mass prediction by non-destructive method. Scientia Horticulturae. 299: 111014. https://doi.org/10.1016/j.scienta.2022.111014

Sousa IC de, Nascimento M, Sant'anna I de C, Caixeta ET, Azevedo CF, Cruz, CD, Silva FL da, Alkimim ER, Nascimento, ACC, Serão NVL (2022) Marker effects and heritability estimates using additive-dominance genomic architectures via artificial neural networks in Coffea canephora. PLOS ONE. 17: e0262055. https://doi.org/10.1371/journal.pone.0262055

Sousa IC de, Nascimento M, Silva GN, Nascimento ACC, Cruz CD, Almeida DP de, Pestana, KN, Azevedo CF, Zambolim L, Caixeta ET (2020) Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. Scientia Agricola. 78: e20200021. https://doi.org/10.1590/1678-992X-2020-0021

Teodoro P E, Barroso LMA, Nascimento M, Torres FE, Sagrilo E, Santos AD, Ribeiro LP (2015). Redes neurais artificiais para identificar genótipos de feijão-caupi semiprostrado com alta adaptabilidade e estabilidade fenotípicas. Pesquisa Agropecuária Brasileira. 50: 1054-1060. https://doi.org/10.1590/S0100-204X2015001100008

Venables WN, Ripley BD (2002). Modern applied statistics with S 4 edition Springer. New York.

VAN EEUWIJK, Fred A.; BUSTOS-KORTS, Daniela V.; MALOSETTI, Marcos. What should students in plant breeding know about the statistical aspects of genotype https://doi.org/10.2135/cropsci2015.06.0375

Verma MM, Chahal GS, Murty BR (1978) Limitations of conventional regression analysis a proposed modification. Theoretical and Applied Genetics. 53: 89-91. https://doi.org/10.1007/BF00274335

Yan W, Hunt LA, Sheng Q, Szlavniscs Z (2000) Cultivar evaluation and mega-environment investigation based on the GGE Biplot. Crop Sci 40:597-605. https://doi.org/10.2135/cropsci2000.403597x