



Evaluation of the efficiency of artificial neural networks for genetic value prediction

G.N. Silva^{1,4}, R.S. Tomaz², I.C. Sant'Anna^{3,4}, V.Q. Carneiro^{3,4}, C.D. Cruz^{3,4} and M. Nascimento^{1,4}

¹Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, MG, Brasil

²Departamento de Engenharia Agrônômica, Universidade Estadual Paulista "Júlio de Mesquita Filho", Dracena, SP, Brasil

³Departamento de Biologia Geral Universidade Federal de Viçosa Universidade Federal de Viçosa, Viçosa, MG, Brasil

⁴Laboratório de Bioinformática, Viçosa, MG, Brasil

Corresponding author: G.N. Silva
E-mail: gabi.nuessilva@yahoo.com.br

Genet. Mol. Res. 15 (1): gmr.15017676

Received August 18, 2015

Accepted November 17, 2015

Published March 28, 2016

DOI <http://dx.doi.org/10.4238/gmr.15017676>

ABSTRACT. Artificial neural networks have shown great potential when applied to breeding programs. In this study, we propose the use of artificial neural networks as a viable alternative to conventional prediction methods. We conduct a thorough evaluation of the efficiency of these networks with respect to the prediction of breeding values. Therefore, we considered eight simulated scenarios, and for the purpose of genetic value prediction, seven statistical parameters in addition to the phenotypic mean in a network designed as a multilayer perceptron. After an evaluation of different network configurations, the results demonstrated the superiority of neural networks compared to estimation procedures based on linear models, and indicated high predictive accuracy and network efficiency.

Key words: Artificial intelligence; Simulation; Accuracy

INTRODUCTION

Researchers conducting plant or animal genetic breeding research daily face the challenge of recommending superior genotypes with a view towards an increased yield of the population under study. Therefore, biometric genetics, together with concepts from quantitative genetics and population studies, has contributed considerably to genetic breeding through methods that assist a breeder in identifying superior genotypes, predicting genetic gains, and recommending varieties (Cruz, 2005). A highly relevant parameter for evaluating the quality of an experiment is selective accuracy (Resende and Duarte, 2007). To recommend or select individuals, a researcher turns to experimental methods and statistical analyses (Resende, 2007). Among the methods of statistical genetics for selection, we can highlight mass selection (Gardner, 1961; Méndez, 1971), progeny selection (Oliveira et al., 2008; Rosado et al., 2012), selection index (Silva, 1982; Cruz and Regazzi, 2012), and selection based on mixed models such as Best linear unbiased prediction/restricted maximum likelihood (BLUP/REML) (Patterson and Thompson, 1971).

In general, these are divergent methodologies; however, in all of them, it is practically unanimous that all information on the genotypic characteristics of individuals may be obtained through phenotype-based predictions (Cruz et al., 2012). Another predominant factor in studies based on these methodologies is setting up experiments with low reliability of predicting the true genotypic value and, consequently, according to Petek et al. (2008), the use of mean phenotypic values as an indicator of genetic superiority. Upon checking the literature, however, a breeder may find a series of other statistics and parameters, not currently used in selection procedures, providing a considerable amount of information regarding the genotype of individuals.

Unlike all the previously cited methodologies, artificial neural networks (ANNs) present the possibility of applying computational intelligence to genetic breeding for the purpose of predicting genetic values. The application of neural networks to an analysis of biological data fits within the area of research called bioinformatics and involves computational techniques and tools to resolve biological problems (Braga et al., 2011).

Various authors have been evaluating the use of ANNs in problems involving genetics (Grimaldi et al., 2011; Sant'Anna et al., 2015). Ventura et al. (2012) investigated the application of these networks to predicting genetic values in cattle and reported their considerable potential for genetic evaluation. Barroso et al. (2013) compared the results obtained by the Eberhart and Russel method with discriminant analysis and analysis through neural networks to analyze adaptability and stability in alfalfa genotypes. Nascimento et al. (2013) used these networks for a classification of alfalfa genotypes, and Silva et al. (2014) used neural networks to predict superior genotypes. Sant'Anna et al. (2015) concluded that the use of ANNs is a promising technique to solve classification problems, since the number of individuals classified incorrectly by an ANN was always lower than that by discriminant functions. Furthermore, according to Barroso et al. (2013), in comparison to other methodologies, these networks have the advantage of not requiring presuppositions with respect to the distribution of the data to be used.

In view of the growing potential of computational intelligence shown by ANNs, this study was conducted for evaluating this new approach for the prediction of genetic values and gains by using simulated data and considering information often disregarded by breeders and researchers as input for the network.

MATERIAL AND METHODS

The data simulation process was carried out through the GENES computational application, as described by Cruz (2013). Eight scenarios were simulated with the mean values, coefficient of variation, and heritability well established. A randomized block experimental design (RBD) was adopted for the simulation, with 100 genotypes and six blocks, assuming heritability values of 40, 45, 50, 55, 60, 65, 70, and 80%, a mean value of 100, and a coefficient of variation of 15%. The values of the mean, heritability, and coefficient of experimental variation were established for each variable, and the following statistical model was used:

$$Y_{ij} = \mu + G_i + B_j + \varepsilon_{ij} \quad (\text{Equation 1})$$

where Y_{ij} denotes the simulated observation of a given characteristic; μ , the overall mean of the characteristic whose value is specified by the researcher; G_i , the effect associated with the i -th genotype; B_j , the effect associated with the j^{th} block; and ε_{ij} , the random error, which is $\varepsilon_{ij} \sim N(0, \sigma^2)$.

To estimate the effects of the genotypes, it is necessary to know the value of genetic variance, which is obtained from the information on heritability (h^2) and the coefficient of experimental variation (CV_e). Thus, the value of environmental variance is obtained through $\sigma^2 = \left(\frac{\mu CV_e}{100}\right)^2$ knowing that

$$h^2 = \frac{100\sigma_G^2}{\sigma_G^2 + \frac{1}{b}\sigma^2}, \sigma_G^2 = \frac{\sigma^2 h^2}{b(100 - h^2)} \quad (\text{Equation 2})$$

On the basis of this information, the fixed or random effects of genotypes may be established. In the first case, the existence of g fixed effects is admitted, whose values constitute an arithmetic progression of ratio r with $G_1 = -G_g$ and $\bar{G} = 0$. Thus, the value G_g is estimated by $G_g = \frac{(g-1)\sqrt{3}\sigma_G^2}{\sqrt{g(g+1)}}$ and the other effects estimated considering the ratio of arithmetic progression given as follows:

$$r = \frac{G_g - G_1}{g - 1} \quad (\text{Equation 3})$$

To simulate the effect of random errors, it is considered that $\varepsilon_{ij} \sim \text{NID}(0, \sigma^2)$. As the researcher supplies the value of the coefficient of experimental variation and of the mean value of the characteristic, the value of σ^2 becomes known; therefore, random and independent errors may be estimated using the previously described random function.

Finally, once the mean value of the characteristic and of the effects involved are known, the phenotypic values of each variable, in each environment (when it is the case), are established through the model as follows:

$$Y_{ij} = \mu + G_i + B_j + \varepsilon_{ij} \quad (\text{Equation 4})$$

For the purpose of network training, we expanded the initial dataset by using the following process:

The simulated values were taken as a random variable $Y \sim N(\phi, \Sigma)$. The data were

transformed into a random variable $Z \sim N(\phi, 1)$ through the linear transformation $Z = F'Y$, in which F was obtained through the spectral decomposition process of Σ , such that $\Sigma^{-1} = FF'$. The expansion process consisted of the simulation of new values of Y , considering $Y \sim N(\phi, (F')^{-1}Z)$. An expanded data file was considered, consisting of 5000 genotypes for training the neural network. The data expansion process was also carried out by using the integration module in the GENES computational application (Cruz, 2013).

In dataset simulation procedures, or the replication of a known data structure, or even the expansion of a set based on the structure of another, some presuppositions must be made. The first of them is that the dataset must have a known distribution and, in the beginning, a mean value equal to zero and variance equal to V . To satisfy this requirement, a Box-Muller transform was used, which ensures that the variables x and y are normally distributed with a mean of zero and variance V as follows:

$$x = \sqrt{-2\log_e(\text{RND})V} \cos(2\pi\text{RND}) \quad (\text{Equation 5})$$

$$y = \sqrt{-2\log_e(\text{RND})V} \sin(2\pi\text{RND}) \quad (\text{Equation 6})$$

where RND denotes a random number.

To ensure that the covariance of the X set of data is null, the recommended methodology is the use of the principal components technique and is based on a simplification of the dataset to a reduced set of components, which have the properties of retaining the maximum of the variation originally available and being independent of each other (Cruz, 2006).

Consider the random variable $Y \sim N(\phi, \Sigma)$ that we wish to transform into the random variable $Z \sim N(\phi, 1)$. Through the spectral decomposition process, we have:

$$\begin{aligned} \Sigma^{-1} = FF'. \text{ Thus, } (\Sigma^{-1})^{-1} = (FF')^{-1} = (F')^{-1}F^{-1} = \Sigma. \text{ If } Z = F'Y, \text{ then} \\ E(Z) = E(F'Y) = F'E(Y) = F'\phi \text{ and } V(Z) = F'V(Y)F = F'\Sigma F = F'[(F')^{-1}F^{-1}]F = 1 \end{aligned} \quad (\text{Equations 7 and 8})$$

Let us consider that $Z \sim N(\phi, 1)$. If $Z = F'Y$, then $Y = (F')^{-1}Z$. Therefore, $V(Y) = V((F')^{-1}Z) = (F')^{-1}V(Z)[(F')^{-1}]' = (F')^{-1}IF^{-1} = \Sigma$

In the expanded dataset, files with 5000 genotypes were obtained for training the ANN. In breeding programs for the selection of superior genotypes, the conventionally adopted model was based on a decomposition of the phenotypic values of the individuals at the mean value of the characteristic and on the effects involved, just as in the equation $Y_{ij} = \mu + G_i + B_j + \varepsilon_{ij}$, already described.

As seen throughout the study, selection is usually practiced on the basis of the mean phenotypic values obtained. Thus, given two individuals i and i' , their mean values are given by $\bar{Y}_i = \hat{\mu} + \hat{G}_i = \hat{\mu}_i$ and $\bar{Y}_{i'} = \hat{\mu} + \hat{G}_{i'} = \hat{\mu}_{i'}$. Given the equations obtained, the individual recommended by the breeder will be the one that exhibits the greatest value for the estimated phenotypic mean ($\hat{\mu}_i$).

In the methodology proposed in this study, we believe that a consideration of other additional statistics in the decision-making process increases the selection accuracy; i.e., an expansion of the information regarding the genotypes evaluated adds value to the genotypes and assists the breeder in selection. Thus, for a given genotype G_i , the following model is proposed:

$$G_i = \sum_k w_k E_{ik}, \varphi(\cdot), \theta \quad (\text{Equation 9})$$

where w_k denotes the synaptic weight associated with the k^{th} statistic; E_{ik} , statistic k associated with genotype i ; $\varphi(\cdot)$, the activation function used; and θ , the activation (bias) threshold of the network.

Having established the model, given two genotypes G_i and G_j , we only need to compare the values obtained for each of them, just as described in the equation above. The individual that has the greatest value for G_i is considered superior.

A schematic representation of the ANN model proposed is shown in Figure 1.

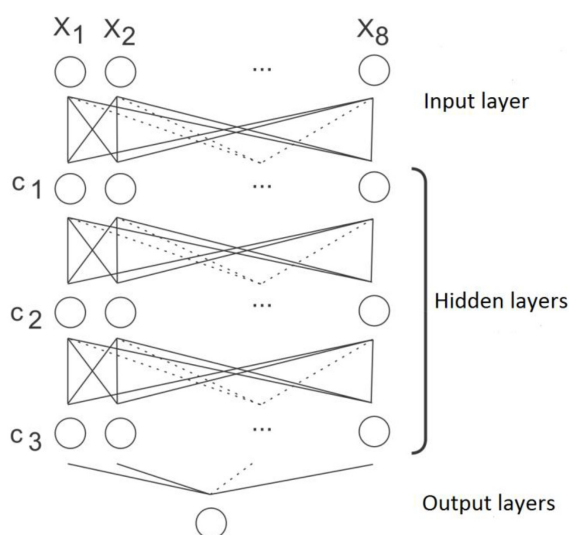


Figure 1. Schematic representation of an MLP network, with an input layer, three hidden layers, and an output layer.

The architecture of the network used was multilayer perceptron (MLP), which in this study was established by adopting three hidden layers of neurons. The MLP network was created using the Matlab software (Matlab, 2010) through the script from the integration module of the GENES computational application. To train the network, the *trainlm* training algorithm was adopted, and the activation functions investigated to establish the best architecture were the function *purellin*, the hyperbolic tangent function – *tansig*, and the sigmoidal function – *logsig*.

The number of neurons per layer ranged from one to seven, with the maximum number of iterations (or epochs) equal to 2000. Combining the variation in the number of neurons per layer and the possibilities of the activation function, a total of $7 \times 7 \times 7 \times 3 \times 3 \times 3 = 9261$ network architectures were tested.

As an alternative to the usual selection methods, which base their selection only on the mean phenotypic values of the genotypes, the network was trained and validated using the true

genotypic values as the desired output and, as the input, the information from seven other statistics (in addition to the mean) as follows: x_1 denotes the mean phenotypic value; x_2 , the maximum value of each genotype in b blocks; x_3 , the minimum value of each genotype in b blocks; x_4 , the value of the standard deviation of each genotype; x_5 , the value of the coefficient of variation for each genotype; x_6 , the value of the sum of rank (or positions) obtained in the b blocks for each genotype; x_7 , the value of the P_i statistic, adapted from Lin and Binns (1988) for studies on adaptability and stability in genotypes, given by $\sum_{j=1}^b \frac{(x_{ij} - M_j)^2}{2b}$, where X_{ij} represents the phenotypic value of the i^{th} genotype in the j^{th} block, M_j indicates the maximum response observed among all the genotypes in block j , and b refers to the number of blocks of the experiment; and x_8 , the discrete values, codified from 1 to 6, established from the transformation of the phenotypic mean considering its variation around the mean value in the lower and upper limits, taken as one, two, or three standard deviations.

To verify the efficiency of the ANN, validation files were generated and also obtained through a trial simulation process in an RBD, with the numbers of blocks and genotypes exactly equal to those of the original file. These files were considered replicas of the test file and maintained the same mean vector and the same variance and covariance structure among the values obtained by the genotypes in the different blocks. Thus, 120 validation files were used as replicas of the simulated experiments. The criterion adopted by the ANN to conclude the analyses using computational intelligence was based on the number of times in which the reliability of the genotypic prediction provided by the network proved to be superior to that of the phenotypic mean. Reliability was determined by using the square of the correlation between the ANN response and the true genotypic value, analogous to the square of the correlation between the phenotypic mean and the true phenotypic value, which expresses the heritability of the characteristic, applied to all the combinations of neurons, number of hidden layers, and activation functions used. Thus, a comparison was made between the measure of reliability of the network and the heritability of the characteristic, which is the criterion used conventionally for the prediction of gains, identification of selective units, and choice of selection strategies.

RESULTS

The methodology presented in this study involved the use of a set of expanded data, which have the particular feature of preserving the characteristics of the experiment from which they are derived. For the purpose of showing that the expansion of the data did not change the structure of the original dataset, as established by the researcher, an analysis of variance of the data from the different files was carried out, and the effectiveness of the simulation process in conserving characteristics was shown.

The correlation between the statistics used in the input layer and the real genetic values was analyzed to verify whether the simulation process, like the expansion process, affected the structure of the expanded experiment. The correlations (r) proved to be statistically equal, according to a t -test ($P < 0.01$) that evaluated the correlations for the simulated file and the expanded file. Furthermore, although the statistics used in the input layer proved to be sufficient for the prediction of genetic values, additional studies are necessary to quantify the influence of each of these in the prediction process, as well as to evaluate the removal or inclusion of statistics.

The same evaluation was carried out using the 120 replicated files used in the validation process. To facilitate the evaluation of the correlations between the additional statistics and the real genetic values for each heritability value, a boxplot diagram was drawn up for all the heritability

values. Figure 2 shows the box plot for the heritability of 80%. We found results similar to those obtained from the simulation and from the expansion of the original file. The reasoning underlying this step is based on the fact that the researcher may question the validity in the validation process since this process may sometimes be carried out with the training file or a part of it. The replicated files serve to emulate situations in which the ANN created is used in the prediction of the contemporary experiments of the genotypes evaluated.

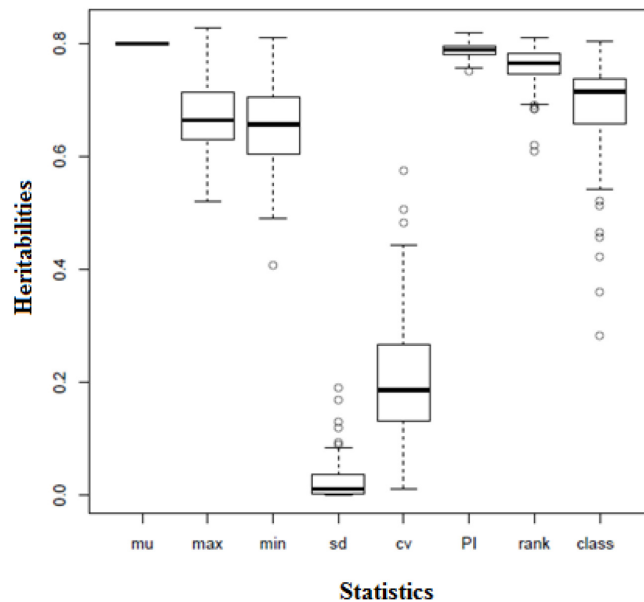


Figure 2. Boxplot for the heritability of 80%.

As the network stop criterion, a hit was considered to be the number of times in which the square of the correlation between the response of the neural network and the true genotypic value was greater than the square of the correlation between the phenotypic mean and the phenotypic value, considering all the combinations of neurons, number of hidden layers, and activation functions used, as proposed by Silva et al. (2014). Thus, a comparison was made between a measure of reliability of the network and the heritability of the characteristic, which is the criterion used conventionally for the prediction of gains and the definition of selection strategies.

Since an improvement of the selection efficiency is still a challenge shared by breeders and biometricians, the selection efficiency was evaluated using the capacity of the ANN for rejecting an inferior genotype and/or selecting a superior genotype. The results obtained through the ANNs were at least as good as those obtained through traditional methods, such as estimates obtained by the least squares method. In 75% of the scenarios, the ANN proved to be more effective than the least squares method in the attainment of the genetic value (Table 1). The estimates of the prediction accuracy through the neural network, considering the 120 validation experiments, were an average of 0.6916% superior considering the heritability values of 40, 45, 50, 55, 65, and 70%. For the heritability values of 60 and 80%, the network provided lower accuracies; nevertheless, they were very close to those obtained by the least squares method used in linear models (LMs). The percentage values of the superiority of the neural network to the LM are shown in Table 1.

Table 1. Mean values of the estimates of accuracies obtained using ANN versus those obtained using LM, as well as the mean value of superiority (ANNSup. (%)) of the network for each heritability value.

| Accuracy (%) | h ² (%) | | | | | | | |
|--------------|--------------------|--------|--------|--------|--------|--------|--------|--------|
| | 40 | 45 | 50 | 55 | 60 | 65 | 70 | 80 |
| LM | 0.6325 | 0.6708 | 0.7071 | 0.7416 | 0.7746 | 0.8062 | 0.8366 | 0.8944 |
| ANN | 0.6411 | 0.6711 | 0.716 | 0.7432 | 0.7738 | 0.8128 | 0.8403 | 0.8911 |
| ANNSup. (%) | 1.3697 | 0.0447 | 1.2587 | 0.2157 | - | 0.8187 | 0.4423 | - |

Moreover, it may be seen that upon fitting a regression for the accuracy values obtained as a function of the heritability values evaluated, the relationship is practically linear, showing a real relationship between these values.

After the prediction and ordering of the genetic values, variation was perceived in the classification of the genotypes. According to Ventura et al. (2012), this may lead to risks in the use of the networks for a genetic evaluation of the characteristic under study. Estimates of coincidence in the classification of the selected genotypes, intensities of 5 and 10%, and rejected genotypes, intensities of 5 and 10%, based on the real genetic values and on the genetic values obtained through the ANN and through the LM are shown in Table 2.

Table 2. Mean percentage of coincidence of genotypes selected (Selec.) and rejected (Rej.) by prediction using ANN and LM.

| | | Coincidence (%) for each h ² | | | | | | | |
|--------------|--------|---|-------|-------|-------|-------|-------|-------|-------|
| | | 40 | 45 | 50 | 55 | 60 | 65 | 70 | 80 |
| Selec. (5%) | LM | 30.17 | 37.33 | 40.83 | 42.67 | 48 | 52 | 50.67 | 59.17 |
| | ANN | 34.17 | 36.67 | 44 | 42.67 | 47.17 | 52.33 | 52.33 | 59 |
| | LM-ANN | 69.67 | 93.33 | 85.5 | 89.67 | 90.67 | 93 | 91.83 | 91.67 |
| Selec. (10%) | LM | 38.58 | 43 | 44.58 | 46.08 | 51.92 | 53.42 | 57.92 | 67.58 |
| | ANN | 39.83 | 42.75 | 47.33 | 45.92 | 52.25 | 53.67 | 58.17 | 65.67 |
| | LM-ANN | 80.92 | 95.25 | 89.33 | 88.33 | 93.5 | 95.58 | 94.92 | 93.33 |
| Rej. (20%) | LM | 51.29 | 52.17 | 56.29 | 59 | 61.54 | 63.63 | 67.58 | 71.29 |
| | ANN | 52.33 | 52.92 | 56.46 | 59.29 | 60.92 | 64.38 | 67.67 | 71.08 |
| | LM-ANN | 89.63 | 89.13 | 97.04 | 93.58 | 95.21 | 96.04 | 98.25 | 94.25 |
| Rej. (10%) | LM | 42.5 | 46.42 | 48.92 | 52.42 | 54.58 | 59 | 62.33 | 67 |
| | ANN | 46 | 47.75 | 49.5 | 51.92 | 53.75 | 60 | 62.92 | 67.33 |
| | LM-ANN | 81.87 | 83.75 | 94.17 | 92.5 | 93.5 | 95.17 | 97.17 | 93.08 |

DISCUSSION

In conventional statistical procedures used in the process of predicting the genetic gain, only the mean genotypic values are considered, assuming the existence of their correlation with the real genetic values of the genotypes denoted by the heritability of the characteristic. Admittedly, the mean is the primary piece of information for judging superiority among genotypes and, in cases of low heritability, the selective process has a relatively low reliability. Nevertheless, a more thorough analysis may be carried out, taking other relevant statistics or statistics of a highly practical nature into account, beyond the mean value of the genotypes; however, this makes decision making more complex for the data analyst. Thus, two genotypes may have similar mean values but very different behaviors in terms of variation, ranking, and minimum and maximum values, among other aspects. In the biometric approach, this additional information is disregarded; however, when a computational intelligence approach is used, such information is indispensable and constitutes a relevant input to be considered in the ANN in the training processes.

Upon considering the overall values of different statistics of each genotype obtained in a determined trial, it may be seen that more reliable criteria are obtained for predicting genotypic

superiority than basing oneself solely on the mean value obtained for the genotype in trials, in which there may be significant variation in genotypic variability and experimental precision.

The usual procedure for determining the criteria of the optimization of network architecture, also called early stopping rules of the iterative process of the ANN, is the mean squared prediction error. Once the mean squared error reaches an ideal level, training is interrupted and the values obtained until then are the best estimates of the network coefficients (Haykin, 2001; Braga et al., 2011).

These early stopping rules are normally adopted when the ANN validation process is carried out with a single set of validation data, and they have been used by several authors such as Sant'Anna et al. (2015), Silva et al. (2014), and Gianola et al. (2011). As 120 datasets were used in this study (representative of the replicas of a base experiment) for network validation, an alternative was to adopt the maximum number of hits obtained by the ANN as the early stopping rule, considering the 120 replicas.

From the results obtained in Table 1, it is clear that ANNs show considerable potential for use in studies involving prediction procedures. Nevertheless, the accuracy values found did not exhibit superiority greater than 1.5% in any of the eight scenarios evaluated, which shows that the results obtained are still quite close to the results found when the researcher makes use of procedures based on the least squares method.

In contrast, the paradigm in effect was constructed using methodologies based on the LM, and the results obtained in this study are nearly at par with the results of these methodologies. Thus, it is still difficult to evaluate the effectiveness of the network based on a model originally conceived by using the mean of the genotype as the best value of its superiority disturbed only by noises brought about by the environmental effect, over which biometrics attempts to establish the best prediction models considering the stochastic distribution.

In spite of the difficulty in evaluating the ANNs, studies based on them represent a considerable advance with respect to the use of trial simulation processes for the prediction of superior genotypes. In breeding programs, which involve experimental statistics, a simulation may be used as an alternative for generating or replicating information representative of the evaluations of a set of genotypes in experiments of suitable design, such as randomized blocks, in which the principles of replication, randomization, and local control are observed.

Simulation processes have been increasingly used in diverse areas of science. Coelho and Barbin (2006) used simulated data to compare different methods for the estimation of variance components. Júnior et al. (2010) used simulated data to compare the Empirical Best Linear Unbiased Predictor (EBLUP) and Bayesian methodologies on data with homogeneity of variances, heterogeneity of genetic variance, and heterogeneity of genetic and environmental variance.

The comparison in this study was performed under artificial conditions considering the use of simulated data. The use of this type of data was important to develop the study because these data speed up the process of comparing methodologies; moreover, the training process of the network may be refined when the researcher has a considerable deal of information available.

In this study, the paradigm of the selective process was based on the estimate of the true genotypic mean value, which is not available in real situations of experimentation. Therefore, once more the use of simulated data was of utmost importance since this procedure allows the researcher to obtain the true genetic values of the genotypes by using the information on the coefficient of variation, mean value, and heritability.

According to Table 2, it is noted that, in all cases, the ANN led to mean coincidences superior to those brought about by the selective process using LMs, indicating that the predictions allowed a better discrimination of the genotypes based on their genetic value. The low percentage

of coincidence of genotypes selected and rejected by prediction using ANN and LM (Table 2) reflect the difficulty in selecting genotypes in a quantitative characteristic. This effect being more pronounced the lower the heritability of the characteristic (40%).

In contrast, coincidences greater than 69% were observed when genotypes selected by the ANN and by the traditional method were compared, showing that, in most cases, the same genotypes were selected by both approaches.

The ANNs proved to be efficient in predicting values and genetic gains in the simulated trials under the randomized block design. However, it should be noted that as the entire study was developed using simulated data, subsequent studies in real breeding populations are necessary for confirming the superiority of predictions by the ANN.

The proposed strategy for training the ANNs constitutes a new approach for analysis, which may be used for increasing the efficiency of the selection process in genetic breeding programs.

The ANN shows superiority in relation to the traditional method of discriminating genotypes on the basis of their real genetic value; moreover, the model fitted to the ANN obliges the researcher to consider additional statistics of a highly practical nature in the process of predicting genetic values. In short, the ANN shows considerable potential for use as an alternative method for predicting genetic values and as a tool for genotypic selection.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

The authors thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Pesquisa do Estado de Minas Gerais (FAPEMIG), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for financial support and scholarships.

REFERENCES

- Barroso LMA, Nascimento M, Nascimento ACC, Silva FF, et al. (2013). Uso do Método de Eberhart e Russell como informação a priori para aplicação de Redes Neurais Artificiais e Análise Discriminante visando a classificação de genótipos de alfafa quanto à adaptabilidade e estabilidade fenotípica. *Rev. Bras. Biol.* 31: 176-188.
- Braga AP, Carvalho APLF and Ludermir TB (2011). Redes neurais artificiais - Teoria e aplicações. LTV, Rio de Janeiro.
- Coelho AM and Barbin D (2006). Simulação de dados visando à estimação de componentes de variância e coeficientes de herdabilidade. *Rev. Mat. Estat.* 24: 103-120.
- Cruz CD (2005). Princípios de genética quantitativa. UFV, Viçosa.
- Cruz CD (2006). Programas genes: Análise multivariada e simulação. 1st edn. UFV, Viçosa.
- Cruz CD (2013). GENES - A software package for analysis in experimental statistics and quantitative genetics. *Acta Sci.* 35: 271-276.
- Cruz CD, Regazzi AJ and Carneiro PCS (2012). Modelos biométricos aplicados ao melhoramento genético. UFV, Viçosa.
- Gardner CO (1961). An evaluation of effects of mass selection and seed irradiation with thermal neutrons on yield of corn. *Crop Sci.* 1: 124-245. <http://dx.doi.org/10.2135/cropsci1961.0011183X000100040004x>
- Gianola D, Okut H, Weigel KA and Rosa GJM (2011). Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genet.* 12: 87. <http://dx.doi.org/10.1186/1471-2156-12-87>
- Grimaldi M, Visintainer R and Jurman G (2011). Regn ANN: Reverse engineering gene networks using artificial neural networks. *PLoS One* 6: e28646. <http://dx.doi.org/10.1371/journal.pone.0028646>
- Haykin SS (2001). Redes Neurais: Princípios e práticas. Symon Haykin; trad Paulo Martins Engel. 2nd edn., Bookman, Porto Alegre.

- Júnior JMC, Assis GML, Euclides RF, Martins WMO, et al. (2010). Predição de valores genéticos utilizando inferência Bayesiana e frequentista em dados simulados. *Acta Sci.* 32: 337-344.
- Lin CS and Binns MR (1988). A method of analyzing cultivar x location x year experiments: a new stability parameter. *Theor. Appl. Genet.* 76: 425-430. <http://dx.doi.org/10.1007/BF00265344>
- Matlab (2010). Matlab Version 7.10.0. Natick, Massachusetts: The Math Works Inc.
- Méndez RI (1971). Refinamiento a la técnica de seleccionmasal moderna. *Agrociencia* 6: 87-91.
- Nascimento M, Peternelli LA, Cruz CD, Nascimento ACC, et al. (2013). Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. *Crop Breed. Appl. Biotechnol.* 12: 152-156. <http://dx.doi.org/10.1590/S1984-70332013000200008>
- Oliveira EJ, Santos VS, Lima DS, Machado MD, et al. (2008). Seleção em progênies de maracujazeiro-amarelo com base em índices multivariados. *Pesq. Agropec. Bras.* 43: 1543-1549. <http://dx.doi.org/10.1590/S0100-204X2008001100013>
- Patterson HD and Thompson R (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* 58: 545-554. <http://dx.doi.org/10.1093/biomet/58.3.545>
- Petek MR, Sera T and Fonseca ICB (2008). Prediction of genetic additive values for development of a coffee cultivar with increased rust resistance. *Bragantia* 67: 133-140. <http://dx.doi.org/10.1590/S0006-87052008000100016>
- Resende MDV (2007). Matemática e estatística na análise de experimentos e no melhoramento genético. Embrapa Florestas, Colombo.
- Resende MDV and Duarte JB (2007). Precisão e controle de qualidade em experimentos de avaliação de cultivares. *Pesq. Agropec. Trop.* 37: 182-194.
- Rosado LDS, Santos CEM, Bruckner CH, Nunes ES, et al. (2012). Simultaneous selection in progenies of yellow passion fruit using selection indices. *Rev. Ceres* 59: 95-101. <http://dx.doi.org/10.1590/S0034-737X2012000100014>
- Sant'Anna IC, Tomaz RS, Silva GN, Nascimento M, et al. (2015). Superiority of artificial neural networks for a genetic classification procedure. *Genet. Mol. Res.* 14: 9898-9906. <http://dx.doi.org/10.4238/2015.August.19.24>
- Silva GN, Tomaz RS, Sant'anna IC, Nascimento M, et al. (2014). Neural networks for predicting breeding values and genetic gains. *Sci. Agric.* 71: 494-498. <http://dx.doi.org/10.1590/0103-9016-2014-0057>
- Silva RG (1982). Métodos de genética quantitativa aplicados ao melhoramento animal. Sociedade Brasileira de Genética, Ribeirão Preto.
- Ventura RV, Silva MA, Medeiros TH, Dionello NL, et al. (2012). Use of artificial neural networks in breeding values prediction for weight at 205 days in Tabapuã beef cattle. *Arq. Bras. Med. Vet. Zootec.* 64: 411-418. <http://dx.doi.org/10.1590/S0102-09352012000200022>