GMR Genetics and Molecular Research ISSN 16765680

Clustering of soybean genotypes via Ward-MLM and ANNs associated with mixed models

P.E. Teodoro[1], **F.E. Torres**[2] **and A.M. Corrêa**[2]

[1]Departamento de Biologia Geral, Universidade Federal de Viçosa, Viçosa, MG, Brasil
[2]Departamento de Fitotecnia, Universidade Estadual do Mato Grosso do Sul, Aquidauana, MS, Brasil

Corresponding author: P.E. Teodoro
E-mail: eduteodoro@hotmail.com

Genet. Mol. Res. 15 (3): gmr.15038384
Received January 4, 2016
Accepted March 11, 2016
Published August 5, 2016
DOI http://dx.doi.org/10.4238/gmr.15038384

ABSTRACT. The objectives of this study were to use mixed models to confirm the presence of genetic variability in 16 soybean genotypes, to compare clusters generated by artificial neural networks (ANNs) with those created by the Ward modified location model (MLM) technique, and to indicate parental combinations that hold promise for obtaining superior segregating populations of soybean. A field trial was conducted between November 2014 and February 2015 at Universidade Estadual de Mato Grosso do Sul, Aquidauana, MS. The experimental design consisted of four replications of randomized blocks, each containing 16 treatments. We assessed the following agronomic traits: plant height, first pod height, number of branches per plant, number of pods per plant, number of grains per pod, hundred-grain weight, and grain yield. Mixed models were used to estimate variance components and genetic parameters, and obtain genotypic values for each trait. After verifying the presence of genetic variability for all traits, genotypic values were submitted to both a Ward-MLM procedure and ANNs to estimate genetic divergence among genotypes. The number of groups formed by both methods was the same, but there were differences in group constitutions. ANN analysis improved soybean genotypes clustering patterns compared to Ward-MLM procedure. Based on these methods, divergent crosses may be made between genotype 97R73 with genotypes AS3797 and SYN9070, whereas convergent crosses may be made between genotypes AS3797 and SYN9070.

Key words: *Glycine max*; genetic divergence; clustering method

INTRODUCTION

Soybean [*Glycine max* (L.) Merrill] is the most widely grown oilseed crop in the world with a production of 260 million tons of grain in 2014/2015, and with production in Brazil accounting for 25% of this total, thus characterizing this country as the second largest producer globally (FAO, 2015). An average yield increase of 37 kg·ha^{-1}·year^{-1} from 1976 to 2015 (Conab, 2015) has been provided by advances in breeding and by improving cultivation techniques and a cultivated area that is constantly expanding.

In soybean breeding, in order to obtain segregating populations, a choice of parents to be crossed is required. Artificial hybridization in autogamous plants often involve two-parent crosses. Major limitations of breeding autogamous species include narrow genetic variation and low recombination rates that are due to the subsequent self-fertilization process. Thus, one method to obtain superior progeny is to gather information on agronomic superiority and genetic diversity in order to enable appropriate combinations among parents, and thus identify a broader gene set and the feasibility of crossing (Cruz et al., 2014).

The use of multivariate techniques has enabled studies on genetic divergence among genotypes in soybean (Niu et al., 2015; Torres et al., 2015). Multivariate analyses are based on algorithms, or distance measurements, which simultaneously consider several features and allow unification of much data from a feature set. Among the available techniques, agglomerative methods are most often used as they bring together genotypes into groups such that there is homogeneity within groups and heterogeneity between groups (Mohammadi and Priasanna, 2003).

Recently, artificial neural networks (ANNs) have emerged as a new method employed for plant breeding research. ANNs are models that emulate a biological neural network and are able to quickly process a large amount of data and recognize patterns using self-learning techniques. Barbosa et al. (2011) reported that the use of ANNs as a method of genotype clustering is promising because these would act as non-parametric classifiers, require small samples for training, and in addition tolerate the default and/or loss of data (Haykin, 2009).

Alternatively, the Ward-modified location model (Ward-MLM) procedure proposed by Franco et al. (1998) is a new approach for characterizing variability using quantitative and/or qualitative variables. This procedure, when combined with analysis of variance (ANOVA), has been proven to enable efficient differentiation between genotypes of maize (Gutiérrez et al., 2003; Franco et al., 2005; Ortiz et al., 2008), forage turnip (Padilla et al., 2005), tomato (Gonçalves et al., 2009), common bean (Cabral et al., 2010; da Costa Barbé et al., 2010), pepper (Sudré et al., 2010), banana (Pestana et al., 2011), and cassava (Oliveira et al., 2015). However, the high environmental influence suffered by quantitative traits, which include traits most commonly used in genotypic selection by soybean breeding programs, yields Ward-MLM results that may be less accurate than those obtained using other techniques (Duarte and Vencovsky, 2001).

Therefore, the use of mixed linear models to analyze genetic divergence can generate greater accuracy because these techniques have the advantage of using genotypic values rather than phenotypic values, thus promoting more accurate results than those obtained using conventional statistical methods (de Resende, 2004). This has been shown in studies with *Eucalyptus* (de Barros Rocha et al., 2007), castor bean (Oliveira et al., 2013) and sugar cane (Lopes et al., 2014).

According to de Resende (2007), ANOVA, since its creation, along with regression analysis were the basis of analysis and statistical modeling for many years. However, best linear unbiased prediction (BLUP) methods developed in 1940 (Bernardo, 1996) and restricted maximum likelihood (REML) methods developed in 1971 (de Resende, 2002) have replaced ANOVA methods due to their higher accuracy in a range of applications.

BLUP presumes knowledge of variance component values, and as this is not possible, these components are estimated via REML; both are then associated with a mixed linear model. In this model, blocks are considered to be fixed effects whereas other effects (genotypes and error) are considered to be random effects (de Resende, 2004). The consideration of treatment effects as random is essential for plant breeding because it is only under this assumption that genetic selection can be accomplished; otherwise, selection is purely phenotypic. This approach of treatment effects as random is recognized by several authors (Duarte and Vencovsky, 2001; Crossa and Franco, 2004; de Resende, 2007; Piepho et al., 2007).

However, despite the importance of using mixed models on the analysis of genetic divergence, no work thus far has combined this analysis with a Ward-MLM procedure and/or ANNs. As such, the aims of this study were i) to use mixed models to check the presence of genetic variability in 16 soybean genotypes, ii) to compare clusters generated by ANNs with those created by the Ward-MLM procedure, and iii) to indicate promising combinations for obtaining superior segregating populations of soybean.

MATERIAL AND METHODS

The field trial was conducted between November 2014 and February 2015 at Universidade Estadual de Mato Grosso do Sul - Unit of Aquidauana (UEMS/UUA), in the municipality of Aquidauana, MS (20°27'S and 55°40'W; 170 m average altitude). The soil of the area was classified as Ultisol sandy loam texture, with the following features at 0 to 0.20 m depth: pH (H_2O) = 6.2; Al exchangeable (cmolc/dm^3) = 0.0; Ca+Mg (cmolc/dm^3) = 4.31; P (mg/dm^3) = 41.3; K (cmolc/dm^3) = 0.2; organic matter (g/dm^3) = 19.74; V (%) = 45; m (%) = 0.0; sum of bases (cmolc/dm^3) = 2.3; CEC (cmolc/dm^3) = 5.1. The regional climate is classified as Aw (Savanna Tropical) according to the Köppen classification system. Accumulated rainfall was 464 mm and maximum and minimum averages temperatures were 37.7° and 16.9°C, respectively, during the field trial.

Experimental design consisted of a randomized block design with sixteen treatments and four replications. Each plot had a length of 5 m, spacing between rows of 0.45 m, and a density of 15 plants/m. Treatments consisted of 16 Roundup-Ready® soybean cultivars: 97R21, 97R71, 97R73, 97Y07, AS3610, AS3730, AS3797, B4184, B4377, CD238, MOSOY6410, P98Y11, POTÊNCIA, SYN1163, SYN13671, and SYN9070.

Before seeding, a commercial formulation of glyphosate was applied at 6 L/ha with a bar sprayer equipped with a cone-type nozzle. Seeds were treated with fungicide (pyraclostrobin + thiophanate-methyl) and insecticide (fipronil) at 200 mL commercial product per 100 kg seed in order to ensure protection against pests and soil fungi. Seed was then inoculated with *Bradyrhizobium* at 200 mL concentrated inoculum per 100 kg of seeds to encourage nitrogen fixation.

Soil tillage consisted of one heavy harrowing and two leveler harrowing passes, after which grooves were mechanically opened. Base fertilization was not conducted as soil fertility was high. Methyomil (600 mL/ha of commercial product) and thiamethoxam + lambda-cyhalothrin (200 mL/ha) were applied with a nozzle cone Coastal sprayer to control pests. Glyphosate (6 L/ha) and, subsequently, hand weeding were used to control weeds.

At maturation, the following agronomic traits were assessed: plant height (PH), first pod height (FPH), number of branches per plant (NB), number of pods per plant (NP), number of grains per pod (NGP), hundred-grain weight (HGW), and grain yield (YIE). Ten plants from each plot were selected at random and PH and FPH were measured in centimeters with a tape measure. NB, NP, and NGP were then quantified for each of these plants. The central rows of each plot were then harvested to determine YIE, whereby grain from these plants was weighed, mass measurement was corrected to 13% moisture, and yields were extrapolated to kg/ha values. Similarly, a sample was collected, weighed, and correct to 13% moisture in order to calculate HGW (g).

To estimate variance components, data were submitted to a REML procedure. The following statistical model was used: y = Xb + Zg + e, wherein y, b, g and e correspond to the data vector, block effects (fixed), genotype effects (random), and random errors, respectively, and where X and Z are incidence matrices for b and g, respectively. The assumed distributions and structures of means and variances were:

$$E\begin{bmatrix} y \\ g \\ e \end{bmatrix} = \begin{bmatrix} Xb \\ 0 \\ 0 \end{bmatrix} \qquad \text{(Equation 1)}$$

$$\mathrm{Var}\begin{bmatrix} g \\ e \end{bmatrix} = \begin{bmatrix} I\sigma_g^2 & 0 \\ 0 & I\sigma_e^2 \end{bmatrix} \qquad \text{(Equation 2)}$$

Model fit was obtained from the mixed model equations:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda_1 \end{bmatrix} \times \begin{bmatrix} \hat{b} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix} \qquad \text{(Equation 3)}$$

Where

$$\lambda_1 = \frac{\sigma_e^2}{\sigma_g^2} \qquad \text{(Equation 4)}$$

Iterative estimators of variance components by REML were calculated:

$$\hat{\sigma}_g^2 = \frac{\left[\hat{g}'\hat{g} + \hat{\sigma}_e^2\right]}{q} \qquad \text{(Equation 5)}$$

where $\hat{\sigma}_g^2$ is the genotypic variance and q is the number of genotypes;

$$\hat{\sigma}_e^2 = \frac{\lfloor y'y - \hat{b}'X'y - \hat{g}'Z'y \rfloor}{[N - r(x)]} \qquad \text{(Equation 6)}$$

where $\hat{\sigma}_e^2$ is the environmental variance, r(x) is the rank of the X matrix and N is the total number of data points;

$$\hat{\sigma}_f^2 = \hat{\sigma}_g^2 + \hat{\sigma}_e^2 \qquad \text{(Equation 7)}$$

where $\hat{\sigma}_f^2$ is the phenotypic variance;

$$h_g^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \qquad \text{(Equation 8)}$$

where h_g^2 is broad sense heritability;

$$h_{mg}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_e^2}{J}} \qquad \text{(Equation 9)}$$

where h_{mg}^2 is mean genotypic heritability and J is the number of blocks;

$$Ac = \sqrt{h_{mg}^2} \qquad \text{(Equation 10)}$$

where Ac is the accuracy of genotypic selection;

$$CV_g(\%) = \left(\sqrt{\frac{\hat{\sigma}_g^2}{\mu}}\right) \times 100 \qquad \text{(Equation 11)}$$

where CV_g is the coefficient of genotypic variation and μ is the overall mean;

$$CV_e(\%) = \left(\sqrt{\frac{\hat{\sigma}_e^2}{\mu}}\right) \times 100 \qquad \text{(Equation 12)}$$

where CV_e is the coefficient of experimental variation; and

$$b = \frac{CV_g(\%)}{CV_e(\%)} \qquad \text{(Equation 13)}$$

where b is named b-quotient.

Genotypic means for each trait were obtained by the BLUP method, given by $\hat{\mu} + \hat{g}$, wherein $\hat{\mu}$ is the overall trait mean and $\hat{g}$ is the predicted genotypic effect. Statistical analyses to obtain variance components, genetic parameters, and genotypic values were conducted using the Selegen-Reml/Blup software (de Resende, 2002).

On the methodology of mixed models, model effects should not be tested via an F test, as is done using analysis of variance (ANOVA) (de Resende, 2004). In this case, the recommended test for random effects is the likelihood ratio test by analysis of deviance. This analysis, suggested by de Resende (2002), represents a generalization of the ANOVA and indicates the quality of the model fit. In addition, this procedure allows consideration of correlated treatments (multicollinearity), a situation that is common in breeding but that is ignored by ANOVA methods, which assume independence of treatment effect errors (de Resende, 2007).

In order to verify genetic divergence among soybean genotypes, the predicted genotypic values for each trait were submitted simultaneously to the Ward-MLM procedure for the composition of the groups through the CLUSTER and IML procedures from SAS® software version 9.1.3 (SAS Institute, 2003). The first step in this process was to estimate the similarity matrix using the Gower algorithm (Gower, 1971). The definition of the optimal number of groups was completed according to the pseudo-F and pseudo-t^2 criteria combined with the likelihood profile associated with the likelihood ratio test. Subsequently, the genotypes were clustered in groups according to the Ward (1963) agglomerative hierarchical method. To verify dissimilarity among the groups, the distance proposed by Franco et al. (1998) was determined.

In order to study genetic divergence among the soybean genotypes via ANNs, a computational routine based on the Kohonen model was implemented with the assistance of the MatLab program (MathWorks, 2010). ANNs were supplied by an input matrix i x j, wherein i genotypes (i = 1, 2, ..., 16) and j traits (PH, FPH, NB, NGP, HGW, and YIE) were considered, and which together represent the input vector X. The intermediate layer was composed of n neurons and output of k neurons representing groups in which the genotypes could be clustered. After the data set was submitted to the ANNs, only one output neuron was activated, recording the group to which the genotype belonged. Groups were composed of similar genotypes so that there was homogeneity within groups and heterogeneity between groups. As such, a rating by similarity of values was made after 5000 simulations.

Considering that j = 7 traits, each input vector element corresponded to the trait set evaluated in genotype i. In the process of implementing the ANNs, the input data set was completely and repeatedly supplied. The synaptic weight vector of each neuron k from the output layer had the same dimension of the input space. Spatial location of a topological neuronal neighborhood was performed by genotyping individuals determined to be centroids. For each output group there was one centroid individual, which was the genotype with the most significant (or typical) features of that group. Thus, the centroids indicated the most typical location of any genotype from a particular group. Finally, to verify dissimilarity among the resulting groups, we determined the distance proposed by Franco et al. (1998).

RESULTS AND DISCUSSION

Analysis of deviance indicated a significant genotype effect (P < 0.01) for all assessed traits (Table 1). As such, the variance and heritability coefficients were significantly different from zero and the existence of genetic variability for these traits was indicated. Similar results were obtained in other studies with soybeans grown in this region (da Silva Jr et al., 2014; Torres et al., 2014; Teodoro et al., 2015a,b; Torres et al., 2015).

Table 1. Variance components, genetic parameters and analysis of deviance for plant height (PH, cm), first pod height (FPH, cm), number of branches (NB), number of pods (NP), number of grains per pod (NGP), hundred-grain weight (HGW, g) and grain yield (YIE, kg/ha) traits assessed in 16 soybean genotypes grown in the Cerrado-Pantanal ecotone.

Parameters	PH	FPH	NB	NP	NGP	HGW	YIE
$\hat{\sigma}^2_g$	103.66	5.29	1.25	115.66	0.11	1.73	317,032.45
$\hat{\sigma}^2_e$	82.41	2.23	0.18	93.85	0.02	0.71	230,506.22
$\hat{\sigma}^2_f$	186.07	7.52	1.43	209.51	0.13	2.44	547,538.67
h^2_g	0.56	0.70	0.87	0.55	0.82	0.71	0.58
h^2_{mg}	0.83	0.90	0.97	0.83	0.95	0.91	0.85
Ac	0.91	0.95	0.98	0.91	0.97	0.95	0.92
CV_g	14.61	19.02	22.23	17.34	13.75	8.20	14.97
CV_e	13.03	12.35	8.44	15.62	6.46	5.26	12.76
b	1.12	1.54	2.63	1.11	2.13	1.56	1.17
Genotype	362.75*	185.82*	91.29*	377.82*	51.79*	86.09*	851.54*
Mean	69.70	12.09	5.62	73.63	2.71	16.86	3.761.17

$\hat{\sigma}^2_g$ = genotypic variance; $\hat{\sigma}^2_e$ = environmental variance; $\hat{\sigma}^2_f$ = phenotypic variance; h^2_g = heritability in the broad sense; h^2_{mg} = mean of genotype heritability; Ac = accuracy in the selection of genotypes; CV_g = coefficient of genotypic variation; CV_e = coefficient of experimental variation; b = b-quotient; *significant at 1% probability by the chi-square test.

The variance of genotypic effects ($\hat{\sigma}^2_g$) was an order of magnitude higher than the variance of environmental effects ($\hat{\sigma}^2_e$) for all traits assessed, and revealing predominantly genetic control of phenotypic variance, and thus indicating the possibility of obtaining gains by selection (Cruz et al., 2014). Mean genotype heritability (h^2_{mg}) is estimated when using block means as evaluation and/or selection criteria (de Resende, 2007). When considering calculated heritability values (≥0.83), there is a reliable indication of the more divergent combinations based on the predicted genotypic values (de Resende, 2004). When estimating broad sense heritability ($\hat{h}^2_g$), we considered the total genetic dispersion, which is relevant as we seek to explore all $\hat{\sigma}^2_g$ among the soybean genotypes. According to de Resende (2002), estimates of $\hat{h}^2_g$ for FPH, NB, NGP, and HGW can be considered high magnitude (>60%) and accurate. Similar results were obtained by da Silva Jr et al. (2014) for these traits.

The accuracy in the selection of genotypes (Ac) reflects the quality of the information and procedures used in the prediction of genetic values. This measure is related to the accuracy of selection and refers to the correlation between predicted values and real genetic values for individuals (de Resende, 2007). As such, a higher Ac for a given trait indicates a greater confidence in the assessment and predicted genetic value for genotypes. In the present study, all traits showed significant values of Ac and h^2_{mg}, indicating high additive genetic variability and precision in its identification, and thus the possibility of success in the study of divergence (Cruz et al., 2014).

The coefficient of experimental variation (CV_e) ranged between 5.26% (HGW) and 15.62% (NP), results that resemble those obtained in other studies with soybean (da Silva Jr et al., 2014; Torres et al., 2014; Torres et al., 2015). According to Cruz et al. (2014), phenotypic traits with continuous distribution and CV_e values lower than 20% indicate excellent experimental accuracy. According to another interpretation, the coefficient of genetic variation (CV_g) quantifies the magnitude of genetic variation for selection and, therefore, higher values are desired. The b-quotient is the ratio between CV_g and CV_e, the calculated value (≥1.00) of which that was obtained for all traits in the present study indicates a situation conducive to selection (de Resende, 2007). Thus, according to the various genetic parameters evaluated, it is possible to infer that the data obtained are suitable for the study of genetic diversity, since variability among genotypes was indicated for all traits.

Results in Table 2 indicate that the Ward-MLM procedure and ANNs analysis formed the same number of genotypic groups (four each). Differences in constitution between groups were due to particular statistical procedures of each method. Gonçalves et al. (2009) and da Costa Barbé et al. (2010) report that determining groups using the Ward-MLM procedure is less subjective than when using ANNs methods because the Ward-MLM procedure is based on analysis of the likelihood function (pseudo-F and pseudo-t^2). Conversely, the ANNs created in this study clustered genotypes based on a group centroid, which is the genotype with the most significant features of the group. According to Barbosa et al. (2011), the use of ANNs as a clustering technique is promising because of its nonlinear structure, which allows capture of the more complex features of a data set.

Table 2. Clustering of 16 soybean genotypes grown in the Cerrado-Pantanal ecotone by Ward-MLM procedure and artificial neural networks (ANNs) and number of genotypes matching in each group generated.

Group	Genotypes clustered by Ward-MLM	Genotypes clustered by ANNs	Number of genotypes in common
I	**97R21,** B4377, **MOSOY64**, **SYN1163** and **SYN1367**	**97R21**, **MOSOY64**, **SYN1163** and **SYN1367**	4
II	AS3610 and **CD238**	B4377, **CD238** and B4184	1
III	**97R73**, AS3730, B4184 and P98Y11	**97R73**, AS3610, 97R71, 97Y07 and POTÊNCIA	1
IV	97R71, 97Y07, **AS3797**, POTÊNCIA and **SYN9070**	AS3730, P98Y11, **AS3797** and **SYN9070**	2

Genotypes in bold correspond to those that were allocated in the same groups in both used procedures.

The Ward-MLM procedure clustered five genotypes into group I, whereas ANNs clustered four genotypes into group I. Comparing the genotypic means of group I according to the clustering method (Table 3) verifies that the values are similar for all traits. This was due to the similar clustering pattern, which allocated genotypes 97R21, MOSOY64, SYN1163, and SYN1367 to that group. Group II contained two genotypes according to the Ward-MLM procedure and three according to ANNs analysis. The only common genotype in both approaches was CD238. Group III contained four and five genotypes according to the Ward-MLM procedure and ANNs analysis, respectively. Only genotype 97R73 was shared between both groups. The Ward-MLM procedure clustered five genotypes into group IV, whereas ANNs clustered four genotypes into group IV, and genotypes AS3797 and SYN9070 were included in the groups created by both methods. Regardless of the method used, group II was formed by soybean genotypes with intermediate agronomic performance, whereas groups III and IV were composed of genotypes with improved performance in the growing region.

Table 4 expresses the distance between groups formed by the Ward-MLM procedure and the ANNs analysis according to the criteria proposed by Franco et al. (1998). Regardless of the pair of groups, the distance suggested by the ANNs analysis was superior to that suggested by the Ward-MLM procedure. Initially, this shows that the clustering method created in this study is superior because it maximized the distance between the groups, which is a basic premise of plant breeding (Cruz et al., 2014).

Table 3. Predicted genotypic values according to the clustering by Ward-MLM procedure and artificial neural networks (ANNs) for various plant traits assessed in 16 soybean genotypes grown in the Cerrado-Pantanal ecotone.

Group	Procedure	PH	FPH	NB	NP	NGP	HGW	YIE
I	Ward-MLM	71.33	11.97	5.07	57.81	2.54	17.20	3288.02
	ANNs	69.17	11.40	5.02	55.72	2.50	17.27	3206.39
II	Ward-MLM	65.63	10.53	5.38	70.55	2.60	15.86	3830.56
	ANNs	76.09	12.77	5.23	63.35	2.63	15.44	3590.00
III	Ward-MLM	78.81	13.01	6.21	82.92	2.82	16.78	4000.92
	ANNs	68.00	11.99	5.64	76.72	2.76	15.88	3910.28
IV	Ward-MLM	67.85	12.10	5.81	82.94	2.85	17.00	4014.78
	ANNs	74.36	12.40	6.50	95.00	2.93	18.75	4257.95

PH = plant height (cm); FPH = first pod height (cm); NB = number of branches; NP = number of pods; NGP = number of grains per pod; HGW = hundred-grain weight (g); YIE = grain yield (kg/ha).

Table 4. Distance between the groups formed by the Ward-MLM procedure and artificial neural networks (ANNs) as criteria proposed by Franco et al. (1998).

Ward-MLM/ANNs*				
Group	I	II	III	IV
I	0	24.49 / 26.18	19.01 / 20.16	37.92 / 42.65
II		0	19.37 / 21.04	33.32 / 38.06
III			0	44.35 / 49.05
IV				0

*The first value refers to the distance of groups formed by Ward-MLM procedure, while the second refers to groups generated by ANNs.

Regardless of clustering method, genotypes in group III showed a high similarity with genotypes from groups I and II. In order to not restrict genetic variability in breeding programs, these groups are not recommended for hybridization use, in order to not prevent realization of possible gains with a given selection differential. This occurs because genetically related parents tend to share more genes or alleles, and when two of these parents are crossed, there is little variation from which to select, given the low level of allelic heterozygosity (Cruz et al., 2014).

According to the outcomes of both analysis methods, the greatest distance was observed between the groups III and IV. This high divergence, initially, allows a recommendation to cross between these pairs in order to maximize the appropriate selection differential in progeny and increase the possibility of segregants in advanced generations due to increased numbers of loci conferring dominance effects (Cruz et al., 2014). In addition, individuals from these groups have the highest genotypic means for YIE, an extremely important trait in the selection of superior genotypes in a breeding program. As such, it may be possible to generate genotypes with a high heterotic effect due to different numbers of loci on which dominance effects are evident.

Finally, based on group classifications by both methods and on superior agronomic performance, we recommend crosses between genotype 97R73 (group III) and genotypes AS3797 and SYN9070 (group IV) when the aim of the breeding program is to generate divergent genotypes and exploit a higher selection differential in the population. Conversely, when the aim is to increase the population mean, convergent crosses between genotypes AS3797 and SYN9070 (group IV) can be made.

Conflicts of interest

The authors declare no conflicts of interest.

ACKNOWLEDGMENTS

We thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for financial support.

REFERENCES

Barbosa CD, Viana AP, Quintal SSR and Pereira MG (2011). Artificial neural network analysis of genetic diversity in *Carica papaya* L. *Crop Breed. Appl. Biotechnol.* 11: 224-231.

Bernardo R (1996). Best linear unbiased prediction of maize-single cross performance given erroneous inbred relationships. *Crop Sci.* 36: 862-866. http://dx.doi.org/10.2135/cropsci1996.0011183X003600040007x

Cabral PDS, Soares TCB, Gonçalves LSA, Amaral Júnior AT, et al. (2010). Quantification of the diversity among common bean accessions using Ward-MLM strategy. *Pesquisa Agropecu. Bras.* 45: 1124-1132. http://dx.doi.org/10.1590/S0100-204X2010001000011

Conab (2015). Acompanhamento da safra Brasileira. Available at [http://www.conab.gov.br/OlalaCMS/uploads/arquivos/14_09_10_14_35_09_boletim_graos_setembro_2014.pdf]. Accessed August 23, 2015.

Crossa J and Franco J (2004). Statistical methods for classifying genotypes. *Euphytica* 137: 19-37. http://dx.doi.org/10.1023/B:EUPH.0000040500.86428.e8

Cruz CD, Carneiro PCS and Regazzi AJ (2014). Modelos biométricos aplicados ao melhoramento genético. 3rd edn. Editora UFV, Viçosa.

da Costa Barbé T, do Amaral Júnior AT, Gonçalves LSA, Rodrigues R, et al. (2010). Association between advanced generations and genealogy in inbred lines of snap bean by the Ward-Modified Location Model. *Euphytica* 173: 337-343. http://dx.doi.org/10.1007/s10681-009-0089-z

da Silva Jr CA, Teodoro PE, Silva GFC, Ribeiro LP, et al. (2014). Correlations and genetic parameters between morphological descriptors in soybean. *J. Agron.* 13: 117-121. http://dx.doi.org/10.3923/ja.2014.117.121

de Barros Rocha MG, Pires IE, Barros Rocha R, Xavier A, et al. (2007). Seleção de genitores de *Eucalyptus grandis* e de *Eucalyptus urophylla* para produção de híbridos interespecíficos utilizando REML/BLUP e informação de divergência genética. *Rev. Arv* 31: 977-987.

de Resende MDV (2002). O Software Selegen–Reml/Blup. Embrapa Florestas, Curitiba.

de Resende MDV (2004). Métodos estatísticos ótimos na análise de experimentos de campo. Embrapa Floresta, Colombo.

de Resende MDV (2007). Matemática e estatística na análise de experimentos e no melhoramento genético. Embrapa Florestas, Colombo.

Duarte JB and Vencovsky R (2001). Estimação e predição por modelo linear misto com ênfase na ordenação de médias de tratamentos genéticos. *Sci. Agric.* 58: 109-117. http://dx.doi.org/10.1590/S0103-90162001000100017

FAO (2015). FAOSTAT. Food and Agriculture Organization of the United Nations. Available at [http://faostat.fao.org]. Accessed August 23, 2015.

Franco J, Crossa J, Villaseñor J, Taba S, et al. (1998). Classifying genetic resources by categorical and continuous variables. *Crop Sci.* 38: 1688-1696. http://dx.doi.org/10.2135/cropsci1998.0011183X003800060045x

Franco J, Crossa J, Taba S and Shands H (2005). A sampling strategy for conserving genetic diversity when forming core subsets. *Crop Sci.* 45: 1035-1044. http://dx.doi.org/10.2135/cropsci2004.0292

Gonçalves LSA, Rodrigues R, do Amaral Júnior AT, Karasawa M, et al. (2009). Heirloom tomato gene bank: assessing genetic divergence based on morphological, agronomic and molecular data using a Ward-modified location model. *Genet. Mol. Res.* 8: 364-374. PubMed http://dx.doi.org/10.4238/vol8-1gmr549

Gower JC (1971). A general coefficient of similarity and some of its properties. *Biometrics* 27: 857-871. http://dx.doi.org/10.2307/2528823

Gutiérrez L, Franco J, Crossa J and Abadie T (2003). Comparing a preliminary racial classification with a numerical classification of the maize landraces of Uruguay. *Crop Sci.* 43: 718-727. http://dx.doi.org/10.2135/cropsci2003.0718

Haykin SO (2009). Neural networks and learning machines. 3rd edn. Prentice Hall, New York.

Lopes VR, Bespalhock Filho JC, Daros E, Oliveira RA, et al. (2014). Divergência genética entre clones de cana-de-açúcar usando análise multivariada associada a modelos mistos. *Semina: Cien. Agra* 35: 125-134.

MathWorks (2010). MATLAB 2010. Available at [http://www.mathworks.com].

Mohammadi SA and Priasanna BM (2003). Analysis of genetic diversity in crop plants-salient statistical tools and considerations. *Crop Sci.* 43: 1235-1248. http://dx.doi.org/10.2135/cropsci2003.1235

Niu YL, Guo WY, Bai LR and Zhao JC (2015). Genetic diversity and the conservation priority of *Glycine soja* populations from Northern China. *Genet. Mol. Res.* 14: 16608-16615. http://dx.doi.org/10.4238/2015.December.11.8

Oliveira EJ, Oliveira Filho OS and Santos VS (2015). Classification of cassava genotypes based on qualitative and quantitative data. *Genet. Mol. Res.* 14: 906-924. http://dx.doi.org/10.4238/2015.February.2.14

Oliveira RS, Silva AS, Brasileiro BP, Medeiros EP, et al. (2013). Genetic divergence on castor bean using the ward-mlm strategy. *Rev. Ciênc. Agron.* 44: 564-570.

Ortiz R, Crossa J, Franco J, Sevilla R, et al. (2008). Classification of Peruvian highland maize races using plant traits. *Genet. Resour. Crop Evol.* 55: 151-162. http://dx.doi.org/10.1007/s10722-007-9224-7

Padilla G, Cartea ME, Rodríguez VM and Ordás A (2005). Genetic diversity in a germplasm collection of *Brassica rapa* subsp. *rapa* L. from northwestern Spain. *Euphytica* 145: 171-180. http://dx.doi.org/10.1007/s10681-005-0895-x

Pestana RKN, Amorim EP, Ferreira CF, de Oliveira Amorim VB, et al. (2011). Agronomic and molecular characterization of gamma ray induced banana (*Musa* sp.) mutants using a multivariate statistical algorithm. *Euphytica* 178: 151-158. http://dx.doi.org/10.1007/s10681-010-0329-2

Piepho HP, Möhring J, Melchinger AE and Büchse A (2007). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161: 209-228. http://dx.doi.org/10.1007/s10681-007-9449-8

SAS Institute (2003). SAS®, Version 9.1.3. SAS Institute, Cary.

Sudré CP, Gonçalves LSA, Rodrigues R, do Amaral Júnior AT, et al. (2010). Genetic variability in domesticated *Capsicum spp* as assessed by morphological and agronomic data in mixed statistical analysis. *Genet. Mol. Res.* 9: 283-294. http://dx.doi.org/10.4238/vol9-1gmr698

Teodoro PE, Ribeiro LP, Corrêa CCG, da Luz RAA Jr. et al. (2015a). Path analysis in soybean genotypes as function of growth habit. *Biosci. J.* 31: 794-799. http://dx.doi.org/10.14393/BJ-v31n1a2015-26094

Teodoro PE, Rigon JPG, Torres FE, Ribeiro LP, et al. (2015b). Comparison of clustering methods for study of genetic dissimilarity in soybean genotypes. *Afr. J. Agric. Res.* 10: 1331-1337.

Torres FE, Silva EC and Teodoro PE (2014). Desempenho de genótipos de soja nas condições edafoclimáticas do ecótono Cerrado-Pantanal. *Interações* 15: 71-78.

Torres FE, David GV, Teodoro PE, Ribeiro LP, et al. (2015). Desempenho agronómico e dissimilaridade genética entre genótipos de soja. *Rev. Ciên. Agrár.* 38: 111-117.

Ward JH (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58: 236-244. http://dx.doi.org/10.1080/01621459.1963.10500845