Cluster analysis combined with AMMI model and bootstrap resampling: application in soybean with insect control

Priscila Neves Faria
Universidade Federal de Uberlândia, Uberlândia, Brazil – priscila@famat.ufu.br

Carlos Tadeu dos Santos Dias
Escola Superior de Agricultura "Luiz de Queiroz", Piracicaba, Brazil – ctsdias@usp.br

Lúcio Borges Araújo
Universidade Federal de Uberlândia, Uberlândia, Brazil - araujolb@gmail.com

Marcelo Ângelo Cirillo
Universidade Federal de Lavras, Lavras, Brazil - macufla@gmail.com

Abstract

In this work the objective was identify genotypes that unite the characteristics of high yield and tolerance to chewing and sucking insects. To assess the accuracy of the scores of genotypes and environments, the 'bootstrap' resampling technique was used, enhancing the quality of inferences about phenotypic adaptabilities estimated by the AMMI model. In addition, the Euclidean distance between genotypes scores was used as similarity measure and subsequently for clustering. From the analysis, the genotypes 97-8011, 97-8029, 97-8050 (33), and the control IAS-5 (44) can be widely recommended, not only for being stable, but also in view of the excellent mean grain yield.

Keywords: Similarity, grouping, yield, interaction.

1. Introduction

The inclusion of breeding for resistance against insect pests as a goal of breeding programs is a new approach, in response to the yield losses caused by the occurrence of insects that attack, for example, soybean, and which had formerly been controlled mainly by chemical methods (PINHEIRO, 1998).

With regard to soybean, which is extensively grown in diverse environments, a differential response of the genotypes is observed. In this sense, the interaction between genotype and environment represents an important aspect in the context of breeding.

The AMMI analysis (Zobel et al. 1988) allows the estimation of genotypes with elimination of the noise effects from genotype responses, which is of crucial importance for recommending varieties. Moreover, it enables the analysis of genetic diversity by clustering methods and identifies genotypes with high productivity, stable and widely adapted to environments. The idea of refining the analysis of phenotypic stability with the AMMI bootstrap method has a great complementary potential to the AMMI analysis, with additional resources to ensure a more judicious selection of genotypes and environments for phenotypic stability.

The purpose of this study was to implement a more accurate and reliable method of predicting the phenotypic stability of genotypes and environments, analyzing at the same time, the genetic divergence in the evaluation of soybean lines by cluster analysis, identifying genotypes that unite the characteristics of high yield and tolerance to chewing and sucking insects, and grouping similar genotypes at the end of the analysis.

2. AMMI and Bootstrap resampling Methodologies

In the experiments that generated the data of this study involving soybean lines of the population with insect control, derived from the selection for grain yield, insecticide was applied during the entire cycle to control the chewing and sucking insects. The experiments were carried out at two locations in the municipality of Piracicaba, São Paulo (Estação Experimental Anhembi e Fazenda Areão), in two growing seasons (1999/00, 2000/01), using two management systems: intensive and ecological insect control - IIC and EIC, respectively.

For the analysis of interaction, four environments (E1, E2, E3, and E4) were defined, namely: Anhembi-IIC (E1), Anhembi-EIC (E2), Areão-IIC (E3), Areão-EIC (E4). This was repeated in each growing season.

The experiments were conducted in a randomized complete block design with replications stratified in experimental sets, all with the controls: IAC-100, OCEPAR-4, IAS-5 and Primavera. The population effects were considered fixed (lines and common controls) and the environmental effect was considered random.

The AMMI analysis was applied in two successive steps: The additive part of the main effects (overall mean, effect of genotype and environment) were adjusted by analysis of variance (ANOVA), resulting in a non-additive residue (GE interaction); the interaction (multiplicative part of the model) was analyzed by Principal Component Analysis (PCA). The general model was the one proposed by Duarte and Vencovsky (1999).

The interaction effect of the i^{-th} genotype with the j^{-th} environment $(ge)_{ij}$, was calculated in the mathematical model according to the equation $(ge)_{ij} = \sum_{k=1}^{n} \lambda_k \gamma_{ik} \alpha_{jk} + \rho_{ij}$, where λ_k is the singular value of the k^{-th} principal component of the interaction (PCI) retained in the AMMI model; γ_{ik} is the singular vector of the i^{-th} genotype in the k^{-th} PCI; α_{jk} is the singular vector of the j^{-th} environment in the k^{-th} PCI; ρ_{ij} is the residue of the GE interaction or AMMI residue (noise present in the data); k are the non-zero characteristic roots, i.e., the number of PCIs retained in the model, where k = (1,2,..., p), in which rank = min(g-1, e-1) is the rank of the GE matrix.

Thus, the matrix of GE interaction is modeled by the above equation for $(ge)_{ij}$, under the restrictions (marginal conditions). Therefore, the term GE (interaction in the traditional model) in the AMMI method is represented by the sum of p plots, each resulting from the multiplication of λ_k, expressed in the same unit as Y_{ij}, by a genotypic effect (γ_{ik}) and an environmental effect (α_{jk}), both dimension-less, i.e., $\sum_{k=1}^{n} \lambda_k \gamma_{ik} \alpha_{jk}$ (n = terms of the interaction). The term λ_k contains information on the k^{-th} plot of GE interaction; the effects γ_{ik} and α_{jk} represent the weights of genotype i and environment j, in that plot of the interaction.

Thus, the additive and multiplicative components of the AMMI model can be written by $Y_{ij} = \mu + g_i + e_j + \sum_{k=1}^{n} \lambda_k \gamma_{ik} \alpha_{jk} + \rho_{ij} + \bar{\varepsilon}_{ij}$, where the additive part is represented by $\mu + g_i + e_j$ and the multiplicative part $\sum_{k=1}^{n} \lambda_k \gamma_{ik} \alpha_{jk} + \rho_{ij}$ can be decomposed into what we call "standard" and "noise" $\rho_{ij} = \sum_{k=n+1}^{p} \lambda_k \gamma_{ik} \alpha_{jk}$, with n ≤ p.

Then the F_{Gollob} test proposed by Gollob (1968) was used to verify the presence of significant interaction, which is in principle the condition to continue this study. Thereafter, the association between "bootstrap" and AMMI (Lavoranti, 2003) was based on resampling of the residue matrix (noise-free), obtained from the values estimated by the AMMI method proposed by Gollob.

Bootstrap resampling was performed in the non-parametric version and, in agreement with the AMMI model, resampling was performed in the columns of the interaction-effect matrix, which was obtained from the estimated values.

All statistical analyses were performed using the R statistical software (The R Development Core Team, 2008) using the packages: fields (FURRER; NYCHKA, 2009) and agricolae (De MENDIBURU, 2009).

3. Results

In the study population, in the growing season of 2000, the individual analysis of variance (for each environment) was obtained for a statistical evaluation of the genetic variability among treatments (soybean lines) and the experimental precision. When differences between treatments were detected, the combined analysis of variance (Table 1) was carried out and the significance of the genotype - environment interaction was diagnosed by the F test.

Table 1 - Combined **a**nalysis of variance for grain yield data in kg/ha, including the decomposition of the interaction of 44 soybean genotypes evaluated in four environments.

Source of variation			**DF**	**SS**	**MS**	**F**	**Pr (>F)**
Environment			3	3827879	1.275.959	242.826	< 0.0001**
Blocks (Environment)			4	210185	52546	2.197	0.070626
Genotype			43	2392397	55637	2.326	< 0.0001**
G×E			129	4542867	35216	1.472	0.006804**
Residue			204	4879923	23921		
IPCA	**%**	**% Ac.**					
$IPCA_1$	49.1	49.1	45	2.162.507	48.055.71	2.01	0.0006
$IPCA_2$	34.8	83.9	43	1.534.622	35.688.88	1.49	0.036
$IPCA_3$	16.1	100	41	710.244.2	17.323.03	0.72	0.8945
$IPCA_4$	0	100	39	0	0	0	1

The analysis of GE interaction by principal components showed significance ($p < 0.05$) of the first two axes ($IPCA_1$ and $IPCA_2$), which explained 83.9 % of the contribution of SS_{GE}. Thus, the genotype scores were obtained according to the $AMMI_2$ model. The first singular axis of AMMI analysis captures the highest percentage of "standard" and, with subsequent accumulation of the dimensions of the axes, the "standard" percentage decreases and "noise" increases.

The biplot graph showed that the environment which contributed least to the interaction was environment 4, with a low score (vector size). Moreover, the environment 1, which was the environment that contributed most to the GE interaction (Figure 2).

The genotypes with low scores, close to zero, were those that contributed little or nearly nothing to the interaction, and were therefore considered stable. The genotypes that met this condition (Figure 2) were 97-8011 (no. 8), 97 -8029 (no. 24), 97-8050 (no. 33), and the control IAS-5 (no. 44). These genotypes that can be widely recommended are those that combine high mean grain yields with stability in relation to the study environments All means of these genotypes exceeded the overall mean GY.

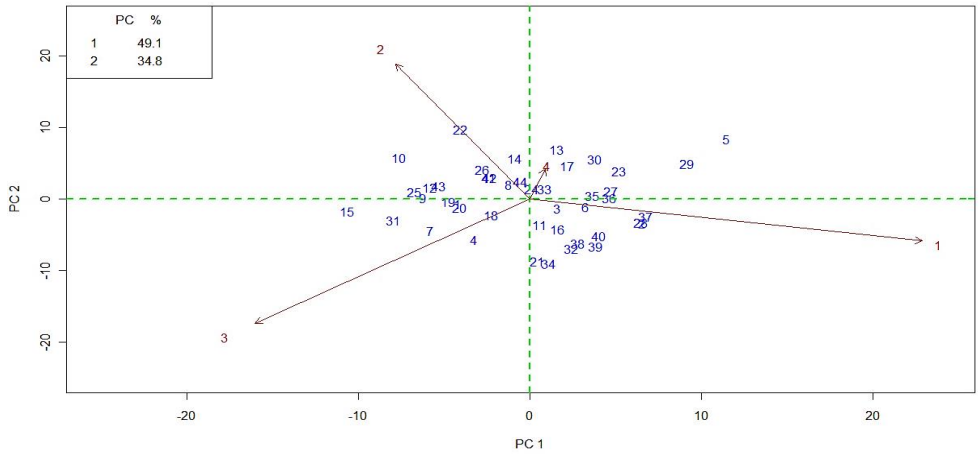

Figure 1 - $AMMI_2$ biplot for grain yield data in kg/ha.

The divergence between the genotypes was evaluated by Ward's hierarchical clustering method, based on the Euclidean distance. From the cluster analysis in the matrix, which is the mean of the resampled distance matrices, a representative dendrogram was constructed of the grouping between the studied genotypes (Figure 3).

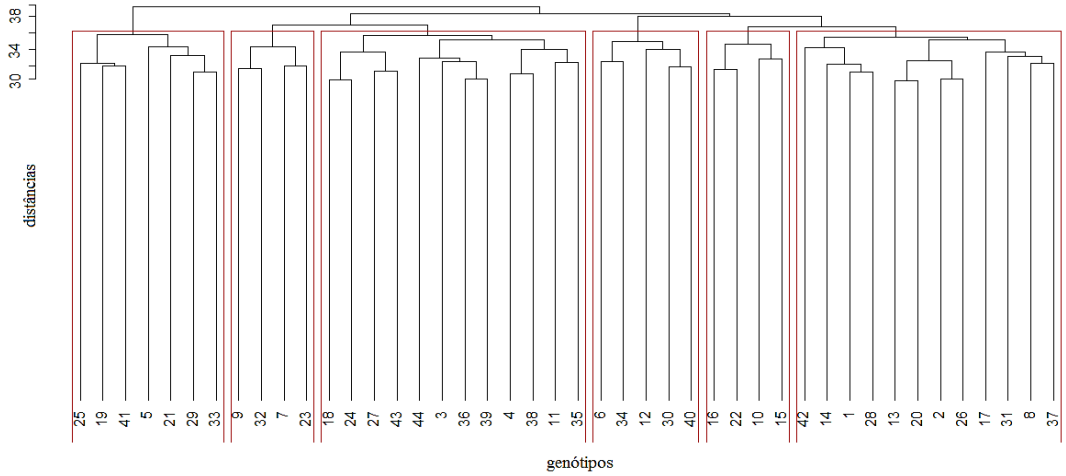

Figure 2 - Dendrogram of the Euclidean distances between the "bootstrap" scores of $AMMI_2$ genotype markers for grain yield data in kg/ha.

The dendrogram indicated the formation of six distinct groups, three of which include controls among the genotypes. This demonstrates that the characteristics of the controls differ from each other, resulting in the grouping in different clusters, based on their similarity. However, the clusters containing the controls stood out from the others in terms of yield and stability of their genotypes.

By the bootstrap intervals, it was possible to calculate confidence intervals of approximately 100 (1-α) % for the parameter of interest (trace of covariance matrix of the GE matrix). The analysis of the percentile interval based on bootstrap percentiles, in relation to the parameter estimates, showed that this estimate (equal to 2,203,686) is contained in the range defined by the 10thpercentile (equal to 1,881,002), within the lower limit, and the 90thpercentile (equal to 2,738,849), in the upper limit. Therefore, since the estimated value of the parameter of interest is included in the calculated interval, it is considered that there is no tendency. According to Monico et al. (2009), in cases in which there is no tendency, the accuracy is contained in the precision measure, which was obtained by bootstrap re-sampling.

4. Conclusions

The analyses were conclusive with regard to selecting the most stable genotypes with high mean for the character grain yield (GY). In the population with insect control in the 2000 growing season concluded that genotype 97-8056 would be a good recommendation for environment 1, with the highest grain yield score. The genotypes 97-8011, 97-8029, 97-8050 (33), and the control IAS-5 (44) can be widely recommended, not only for being stable, but also in view of the excellent mean GY. The mean for this trait was highest in environment 1 (1408.602 kg/ha).

References

De MENDIBURU, F. (2009) **agricolae: Statistical Procedures for Agricultural Research.** R package. Version 1.0-7. URL http://cran.r-project.org/web/packages=agricolae The R Foundation for Statistical Computing. Vienna, Austria. Accessed on May 3, 2012.

EFRON, B.; TIBSHIRANI, R.J. (1993) **An introduction to the bootstrap.** Chapman & Hall, London, 436p.

FURRER, R and NYCHKA, D. Sain S. (2009) **fields: Tools for Spatial Data.** R package. Version 6.01. URL http://CRAN.R-project.org/package=fields. Accessed on May 3, 2012.

GOLLOB, H.F. (1968) A statistical model which combines features of factor analytic and analysis of variance techniques. **Psychometrika 33, n.1:** 73-115.

LAVORANTI, O.J. (2003) **Estabilidade e adaptabilidade fenotípica através da reamostragem "Bootstrap" no modelo AMMI.** Tese (Doutorado) - Escola Superior de Agricultura "Luiz de Queiroz", Piracicaba, 166p.

MONICO, J. F. G.; DAL POZ, A. P.; GALO, M.; SANTOS, M. C. and CASTRO, L. de O. de (2009) Acurácia e Precisão: Revendo os Conceitos de Forma Acurada. **Boletim de Ciências Geodésicas 15:** 469-483.

PINHEIRO, J. B. (1998) **Seleção para caracteres agronômicos, em diferentes épocas de cultivo, de populações de soja com resistência a insetos.** Tese (Doutorado) - Escola Superior de Agricultura "Luiz de Queiroz", Piracicaba, 143p.

R version 2.7.1. (2008) **The R Foundation for Statistical Computing.** ISBN 3-900051-07-0.

ZOBEL, R.W.; WRIGHT, M.J. and GAUCH, H.G (1988) Statistical analysis of a yield trial. **Agronomy Journal 80:** 388-393.