# Factor analysis using mixed models of multi-environment trials with different levels of unbalancing

**J.J. Nuvunga[1], L.A. Oliveira[2], A.K.A. Pamplona[1], C.P. Silva[1], R.R. Lima[1] and M. Balestre[1]**

[1]Departamento de Ciências Exatas, Universidade Federal de Lavras, Lavras, MG, Brasil
[2]Faculdade de Ciências Exatas e Tecnologia, Universidade Federal da Grande Dourados, Dourados, MS, Brasil

Corresponding author: M. Balestre
E-mail: marciobalestre@dex.ufla.br

Genet. Mol. Res. 14 (4): 14262-14278 (2015)
Received March 16, 2015
Accepted August 25, 2015
Published November 13, 2015
DOI http://dx.doi.org/10.4238/2015.November.13.10

**ABSTRACT.** This study aimed to analyze the robustness of mixed models for the study of genotype-environment interactions (G x E). Simulated unbalancing of real data was used to determine if the method could predict missing genotypes and select stable genotypes. Data from multi-environment trials containing 55 maize hybrids, collected during the 2005-2006 harvest season, were used in this study. Analyses were performed in two steps: the variance components were estimated by restricted maximum likelihood, using the expectation-maximization (EM) algorithm, and factor analysis (FA) was used to calculate the factor scores and relative position of each genotype in the biplot. Random unbalancing of the data was performed by removing 10, 30, and 50% of the plots; the scores were then re-estimated using the FA model. It was observed that 10, 30, and 50% unbalancing exhibited mean correlation values of 0.7, 0.6, and 0.56, respectively. Overall, the genotypes classified as stable in the biplot had smaller prediction error sum of squares (PRESS) value and prediction

 

amplitude of ellipses. Therefore, our results revealed the applicability of the PRESS statistic to evaluate the performance of stable genotypes in the biplot. This result was confirmed by the sizes of the prediction ellipses, which were smaller for the stable genotypes. Therefore, mixed models can confidently be used to evaluate stability in plant breeding programs, even with highly unbalanced data.

**Key words:** G x E interaction; Unstructured variance; Adaptability; Stability; Factor analysis

## INTRODUCTION

Identifying genotypes with high production yields and stability, and which are adaptable to the widest range of environments, is one of the main objectives of breeding programs. However, this selection (of the best genotype) may be affected by genotype-environment interactions (G x E). Several methods have been applied to the evaluation of G x E interactions; however, the choice of the best method depends on the experimental design, number of environments available, required precision, and the type of desired information (Cruz et al., 2004).

One of the methods used to evaluate G x E interactions is based on multiplicative analysis, which explores the response of the genotypes to specific environments. Therefore, this analysis can rigorously describe the G x E interaction (Resende and Thompson, 2004). The advantage of multiplicative analyses lies in their ability to group similar environments and genotypes in biplots, which allows for the identification of genotypes with the greatest potential in each subgroup of environments. Multi-environment trial (MET) data are analyzed in two steps in this method; a joint analysis is initially conducted, followed by decomposition of the interaction using the principal components. Difficulties arise during the first step in case of heterogeneous variances or if there are missing values that can compromise the inferences made by analysis of variance (ANOVA). It could be difficult to address complete unbalancing in the environment when using principal components in the second step.

The use of mixed models with genotypes and environments as the major effects (at least one of which is random) and random G x E interactions (Patterson et al., 1977) would provide at least one solution to such restrictions. Smith et al. (2001b) mentioned a number of authors, including Patterson and Nabugoomu (1992), who recognized the possibility of heterogeneous variances; in this context, models that account for the variance heterogeneity in the G x E interaction and which are less strict regarding the assumption of independence may be required.

Therefore, the major criticism from statisticians working on breeding programs focuses on the lack of rigorous analyses of G x E interaction structure, which may affect the recommended cultivars. Traditionally, an analysis of this structure is superficial because it does not show the effects of the complexity of the interaction (Lavoranti et al., 2007).

Therefore, factor analytic multiplicative mixed (FAMM) models performed with a restricted maximum likelihood/best-linear unbiased predictor (REML/BLUP) procedure, is an especially important recent method that adequately explains the major effects (genotype and environment) and their interactions.

Piepho (1997, 1998), Smith et al. (2001b, 2005), and Kelly et al. (2007) showed that FA models exhibit a superior performance in the study of G x E interactions. However, these

studies were limited to comparisons between models and the structures of genetic variance, and covariance matrices with heterogeneous variances. Despite demonstrating the sufficiency of these models in the study of G x E interactions of unbalanced data (not all genotypes are grown in all of the sites), none of these studies demonstrated the robustness of the model for the analysis of highly unbalanced data. Burgueño et al. (2011) recently showed the robustness of FA models with unbalanced data; however, they did not test the different levels of missing data.

Despite the attractiveness of this technique for plant breeding, one of the difficulties encountered by researchers in adopting FA models is their computational implementation, as the available packages do not explore the regression models in which the FA model is grounded (Smith et al., 2001b). Consequently, the equations for the mixed models are relatively dense, which seriously reduces the computational speed of the analyses for datasets with a large number of environments, or when fitting the factor analytic variance models with several factors (Thompson et al., 2003). To improve computational stability, Thompson et al. (2003) suggested the use of sparse matrices in the FA structure; however, such an implementation is also computationally intensive, which was observed in the twelve steps proposed by the authors. The other practical problem with the FA model is the frequent occurrence of Heywood cases, wherein certain parameters in the FA structure become null or negative, potentially hindering the analysis (Smith et al., 2001b; Thompson et al., 2003; Costa e Silva and Dutkowski, 2006).

Piepho and Möhring (2006) and Resende (2007) reported that the best method to model the treatment structure in MET is a multivariate mixed model with unstructured (UN) variance and covariance matrices, as this method accounts for heterogeneous variance and covariance between sites, in addition to being the best method for addressing unbalanced data. However, this structure does not provide a method for direct evaluation of the genotypic adaptability and stability; in addition, it requires overparameterized models. Therefore, it is evident that the discovery of an appropriate (co)variance matrix structure depends on the available data; although FA structures have been widely used, such structures are intended to describe unstructured matrices, and depend on good estimates of the dispersion components. Balestre et al. (2012) indicated that models that are more parsimonious may be preferred when the parameters for the UN matrices are not well estimated.

One relevant characteristic of using mixed models in the MET approach is the ability to predict missing data. The behavior of a given genotype in a specific environment determines the accuracy of the prediction; therefore, if a genotype is stable or predictable in a specific environment, the marginal mean of this genotype can be directly assigned to the unevaluated environment. However, this value may be very different from the true value of a genotype with low homeostasis, which would make this technique impractical (Lin et al., 1986). Therefore, cross-validation (Lavoranti et al., 2003; Yang et al., 2009) using the prediction error sum of squares (PRESS) could be used to confirm the stability of a genotype in a biplot. As unbalancing does not destroy the structure of the interactions in this approach, as shown by the additive main effects multiplicative interaction (AMMI) or genotype and genotype-environment (GGE) bootstrap analyses (Lavoranti et al., 2007; Yang et al., 2009; Yan et al., 2010), the accuracy of the confidence ellipses obtained in the cross validation has a direct genetic interpretation in terms of stability; that is, if the performance of a genotype is not sensitive to its loss from an environment, the genotype can be said to possess greater stability.

Therefore, this study sought to evaluate the robustness of FA models under various levels of unbalancing using cross-validation, and to assess the capacity of these analyses in the selection of stable genotypes, using biplots validated with confidence ellipses.

## MATERIAL AND METHODS

The data used in this study has been previously described by Machado et al. (2008). Experiments were conducted in nine environments during the 2005/06 agricultural year at experimental stations and farms (Table 1). Fifty-five maize hybrids were analyzed (coded as G1, G2, G3, ..., G55) using a randomized block design with three replicates, where an experimental plot consisted of two 3 m long rows. A population density of 55,000 plants per hectare was obtained after thinning. The husked ear yield (t/ha) trait, corrected for 13% moisture content, was evaluated in this study.

**Table 1.** Characteristics of the experimental environments.

| Environment | Municipality | Latitude | Longitude | CV² (%) | Mean yield (t/ha) |
|---|---|---|---|---|---|
| Experimental area/DBI (E1) | Lavras, MG | 21°13'S | 44°58'W | 14.1 | 10.803 |
| Geneze experimental area (E2) | Guarda-Mor, MG | 17°34'S | 47°08'W | 13.5 | 6.212 |
| Bionacional experimental area (E3) | Barreiras, BA | 12°08'S | 45°00'W | 12.0 | 4.549 |
| Prezzotto experimental area (E4) | Jussara, GO | 23°35'S | 52°28'W | 12.5 | 5.152 |
| Vitorinha farm (E5) | Lavras, MG | 21°12'S | 44°58'W | 20.8 | 6.246 |
| Coopadao experimental area (E6) | São Gotardo, MG | 19°18'S | 46°03'W | 11.3 | 8.085 |
| Faepe farm (E7) | Ijaci, MG | 21°09'S | 44°56'W | 10.1 | 13.192 |
| Faepe farm (E8) | Ijaci, MG | 21°10'S | 44°56'W | 10.7 | 8.896 |
| Mato Dentro farm (E9) | Lavras, MG | 21°10'S | 45°03'W | 14.6 | 8.737 |

¹Source: Machado et al., 2007; ²coefficient of variation.

### Multivariate mixed model (MMM)

This dataset was subjected to a joint analysis using the following linear model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{e} \quad \text{(Equation 1)}$$

Here, *y* denotes the vector of the plot observations in each environment, *β* and *u* are the fixed (blocks) and random (genotype) effect vectors, respectively, *e* denotes the random error vector, and *X* and *Z* are the incidence matrices for the fixed and random effects. For this dataset, it was assumed that

$\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$,

And, $\mathbf{u} \sim N(\mathbf{0}, \mathbf{\Sigma})$.

In general, this equation can be expanded as follows:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} X_1 & O & \dots & \dots & \dots & O \\ O & X_2 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & O \\ O & \dots & \dots & \dots & O & X_n \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_n \end{bmatrix} + \begin{bmatrix} Z_1 & O & \dots & \dots & \dots & O \\ O & Z_2 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & O \\ O & \dots & \dots & \dots & O & Z_n \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ . \\ . \\ . \\ u_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix} \quad \text{(Equation 2)}$$

where each subscript corresponds to the subvectors and submatrices of the observations, and experimental design in each environment.

The following mixed model equation matrix (MMEM):

$$C = \begin{pmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + \Sigma^{-1} \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \quad \text{(Equation 3)}$$

Provided the solutions for $\beta$ and $u$, as follows:

$$\begin{pmatrix} \hat{\beta} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + \Sigma^{-1} \end{pmatrix}^{-1} \begin{pmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{pmatrix} \quad \text{(Equation 4)}$$

The expectation-maximization (EM) algorithm described by Dempster et al. (1979) was used to classify the vector $u$, which represents the random effects, as missing data. Therefore, the REML solutions of the elements of matrix $\Sigma$ are given as follows:

$$\hat{\sigma}_{u_{ij}} = \left[ u_i^T u_j + tr\left( C_{ij}^{-1} \right) \right] \Big/ t \quad \text{(Equation 5)}$$

where

$$\tilde{\sigma}_{u_{ij}} = \begin{cases} \sigma_{u_k}^2 & if\ i = j \\ \sigma_{u_{ij}} & if\ otherwise \end{cases} \quad \text{(Equation 6)}$$

the matrix $C_{\hat{y}}^{-1}$ corresponds to the submatrices $i\ j$ of $C_{22}$, which are contained in the inverse matrix $C^{-1}$. The residual (co)variance estimator contained in $R$ can be given as follows:

$$\tilde{\sigma}_{e_{ij}} = \left\{ e_i^T e_j + tr\left( \left[ KC^{-1}K^t \right]_{ij} \right) \right\} \Big/ n^* \quad \text{(Equation 7)}$$

$$\tilde{\sigma}_{e_{ij}} = \begin{cases} \sigma_{e_k}^2 & if\ i = j \\ \sigma_{e_{ij}} & if\ otherwise \end{cases} \quad \text{(Equation 8)}$$

where $K = \{X,Z\}$, the trace depends on the submatrices $i$ and $j$, and $n^*$ denotes the length of the vector $\{j,i\}$.

This approach estimates all dispersion parameters of an unstructured (co)variance matrix; that is, all variances {$\sigma^2_{ek}$, $\sigma^2_{uk}$} and covariances {$\sigma^2_{eij}$, $\sigma^2_{eij}$} are estimated simultaneously with the fixed empirical best linear unbiased estimator (EBLUEs) and random effects (EBLUPs). This procedure allows for the performance of a MET analysis with missing data and variance heterogeneity. However, the stability and adaptability of the genotypes cannot be obtained directly. Therefore, a restricted likelihood factor analysis must be performed in a subsequent step.

## Restricted maximum likelihood factor analysis

Restricted maximum likelihood factor analysis can be easily implemented with the covariance matrices $\Sigma$ and $R$ from step one. It was assumed that $\Sigma$ can be represented by an FA structure, such as ($LL' + \Psi$), and the BLUPs can be represented by common factors in the form ($u = Lf + \delta$). The application of this transformation to the mixed linear model (1) produces the following equation:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}[\mathbf{Lf} + \boldsymbol{\delta}] + \mathbf{e} \quad \text{(Equation 9)}$$

where $f$ is the vector of the factor scores for the missing data (BLUPs), $\delta$ is the specific variance, $L$ represents the matrix of the factor loadings, and $X$ and $Z$ represent the design matrices. In addition, it is assumed that

$f \sim N(\mathbf{0}, \mathbf{I})$,

$\delta \sim N(\mathbf{0}, \boldsymbol{\Psi})$, and

$e \sim N(\mathbf{0}, \mathbf{R})$.

Therefore, the reparametrized matrix solution to the mixed model equations ($W = ZL$), can be given as follows:

$$\begin{pmatrix} \hat{\beta} \\ \hat{f} \\ \hat{\delta} \end{pmatrix} = \begin{pmatrix} X'R^{-1}X & X'R^{-1}W & X'R^{-1}Z \\ W'R^{-1}X & W'R^{-1}W + I & W'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}W & Z'R^{-1}Z + \Psi^{-1} \otimes I \end{pmatrix}^{-1} \begin{pmatrix} X'R^{-1}y \\ W'R^{-1}y \\ ZR^{-1}\acute{y} \end{pmatrix} \quad \text{(Equation 10)}$$

Meyer (2009) reported the solutions for the fixed and random effects to be similar, that is:

$$\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}(y - \text{Wf-}Z\delta) \quad \text{(Equation 11)}$$

$$\hat{f} = (W'\Sigma^{-1}W + I)^{-1}W'\Sigma^{-1}(y - Xb - Z\delta) \quad \text{(Equation 12)}$$

$$\hat{\delta} = (\text{Z}'\Sigma^{-1}Z + \Psi^{-1} \otimes \text{I})^{-1}\text{Z}'\Sigma^{-1}(y - Xb - \text{Wf}) \quad \text{(Equation 13)}$$

This procedure guarantees that the model will converge within the parameter space, thereby avoiding the Heywood cases, which is a major advantage of this procedure.

To fit the number of loadings in the factor analysis, one, two, and three factor loadings (similar to the FA(1), FA(2), and FA(3) models, respectively) were tested to determine the number of factors required to explain the genetic variability and predict the genotypes. The goodness-of-fit for the number of factors in the FA model was assessed using a modified likelihood ratio test reported by Bartlett (1954), and defined by Johnson and Wichern (2007):

$$\lambda = n \ln \frac{(|\hat{\mathbf{L}}\hat{\mathbf{L}}'+\hat{\mathbf{\Psi}}|)}{|\mathbf{R}|} \sim \chi_n^2 \quad \text{(Equation 14)}$$

where $\mathbf{\Sigma} = \hat{\mathbf{L}}\hat{\mathbf{L}}'+\hat{\mathbf{\Psi}}$ is the "variance and covariance matrix" (defined as positive) of the fitted factor model and **R** is the standardized variable matrix.

## Rotation of loadings

The loading matrix is not unique when the number of loadings is greater than one (k > 1). This lack of singularity requires corrective measures when choosing FA-k models, as the variance model is not identifiable. These corrective measures include restricting the identification, or rationing of the factor loadings.

The uniqueness of the choice for *L* was guaranteed using the following restriction:

$$L^t \Psi^{-1} L = \Delta \quad \text{(Equation 15)}$$

where Δ is a diagonal matrix. The key for the estimation of *L* and *Ψ* is to obtain the matrix $\Psi^{-1/2}(A-\Psi\ )\Psi^{-1/2}$.

Therefore, the EM algorithm can be used to obtained $\hat{L} = \Psi^{1/2}\hat{\mathrm{P}}$; $\hat{\mathrm{P}}$ provides the spectral decomposition of $\Psi^{-1/2}(A-\Psi\ )\Psi^{-1/2}$ and $\bar{\Psi} = diag(A-\hat{L}\hat{L}^t)$.

Here, an iterative process occurs that continues until the *L* and *Ψ* matrices converge. Therefore, *L* can describe the loading factors for each environment. The next step consists of obtaining factor scores for the effects of *û*.

## Cross-validation

The data were unbalanced by the random separation of the training and validation populations. The original dataset contains 1,485 observations, which were randomly removed at different levels (10, 30, and 50%). The process was repeated 1000 times for each level of missing data, totaling 3000 different incomplete datasets. One hundred and forty five (10%) elements were removed from the data matrix of the first group, 446 were removed from the second group (30%) and 743 elements were removed to create each unbalanced dataset from the third group (50%).

The predictive ability of the models was measured by calculating the PRESS, and determining the correlation between the predicted and observed values. PRESS can be expressed as follows:

$$PRESS(n) = \sum_{i=1}^{g} (\bar{\hat{f}}_i^n - f_i)^2 \quad \text{(Equation 16)}$$

where $\overline{\hat{f}}_i^n$ is the mean value of the predictions for the nth factor score of the ith genotype, and $f_i$ is the value assumed to be parametric. Based on the empirical distribution of the jth predicted value ($\hat{f}_j^n$), the ellipses can be constructed and homeostasis of the genotype can be measured such that lesser the ellipse spreads out from $\overline{\hat{f}}_i^n$; larger buffering effects of the genotype and smaller individual PRESS given by $\sum_{j=1}^{\hat{}}(\hat{f}_j - f)$ indicate predictions of greater reliability.

## Confidence ellipses for the predictions

Ellipses were obtained using the R software (R Core Team, 2014).

All analyses performed in this study were performed using PROC IML in the SAS (Statistical Analysis System, 2013) and R (R Core Team, 2014) software platforms.

## RESULTS

Of the proposed FA models, the model with two factor loadings [similar to a 2nd order factor analysis (FA2)] best explained the total genotypic variation (85%). Table 2 summarizes the results of the factor loadings, rotated by the varimax method. The first factor explained almost 70% of the total variance, while the second explained 15%. The specific variances were low for all environments studied. The high values for the common variances revealed that two factors explained a large percentage of the variance in each environment, with the FA2 model showing the best fit for the dataset. The model could explain almost all of the environments, except for environments E5, E6, E7, and E8. Table 2 also shows that environment E1 is highly correlated with Factor 1, whereas environments E2, E4, E7, and E9 correlate to Factor 2. The genotypic variance-covariance matrix for the environments and residuals for the balanced data, which is fitted using the mixed model with a UN variance and covariance matrix, is shown in Table 3.

**Table 2.** Estimated loadings (correlation scale) for the balanced data fit using the FA2 model.

| Environment | Factor 1 | Factor2 | Common variance | Specific variance |
|---|---|---|---|---|
| E1 | 0.997 | -0.030 | 0.995 | 0.000 |
| E2 | 0.152 | 0.697 | 0.982 | 0.000 |
| E3 | 0.190 | 0.423 | 0.966 | 0.170 |
| E4 | 0.570 | 0.710 | 0.993 | 0.035 |
| E5 | 0.275 | 0.453 | 0.806 | 0.348 |
| E6 | 0.511 | 0.611 | 0.907 | 0.438 |
| E7 | 0.253 | 0.765 | 0.974 | 0.402 |
| E8 | 0.032 | 0.623 | 0.997 | 0.350 |
| E9 | 0.521 | 0.800 | 0.998 | 0.000 |
| Percentage of variance (%) | 70 | 15 | - | - |

Table 3 clearly shows the genetic and residual heterogeneity of the variances, which justifies the decision to relax this restriction in the model. The genetic covariances are also heterogeneous between sites (these covariances represent the genotypic variance, in addition to the variance of the interaction between pairs of sites). The residual variance was larger for environments one, five, and seven, suggesting the influence of the mean of these environments on the magnitude of this variance (Table 1).

Figure 1 contains a plot of the factor scores obtained from the FA2 model, fitted to the balanced data. This biplot has characteristics that are similar to those of the GGE biplot proposed by Yan

et al. (2000) for phenotypic means (Figure 2). The GGE analysis identified 64% of the total genotypic variance in the first two components, which is less than the 85% explained by the FA model (2).

**Table 3.** Genotypic and residual (in red) variance-covariance matrix for the datasets of maize grown at nine different sites.

| Environment | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 |
|---|---|---|---|---|---|---|---|---|---|
| E1 | 2.78(2.33) | 0.45 | -0.12 | -0.06 | 0.18 | -0.06 | -0.20 | 0.19 | -0.06 |
| E2 | 0.17 | 0.60(0.68) | -0.08 | -0.07 | -0.06 | -0.08 | 0.20 | -0.08 | 0.13 |
| E3 | 0.14 | 0.29 | 0.22(0.29) | -0.01 | 0.01 | 0.01 | -0.08 | -0.03 | 0.05 |
| E4 | 0.48 | 0.22 | 0.10 | 0.28(0.41) | 0.01 | 0.06 | -0.04 | 0.09 | 0.04 |
| E5 | 0.29 | 0.21 | 0.17 | 0.20 | 0.47(1.69) | -0.13 | -0.13 | -0.07 | -0.15 |
| E6 | 0.88 | 0.42 | 0.25 | 0.40 | 0.40 | 1.17(0.83) | -0.05 | 0.10 | -0.12 |
| E7 | 0.45 | 0.37 | 0.19 | 0.47 | 0.36 | 1.05 | 1.48(1.79) | 0.25 | -0.11 |
| E8 | 0.01 | 0.45 | 0.10 | 0.13 | 0.10 | 0.19 | 0.17 | 0.62(0.94) | 0.11 |
| E9 | 0.44 | 0.28 | 0.08 | 0.27 | 0.17 | 0.45 | 0.53 | 0.30 | 0.34(1.16) |

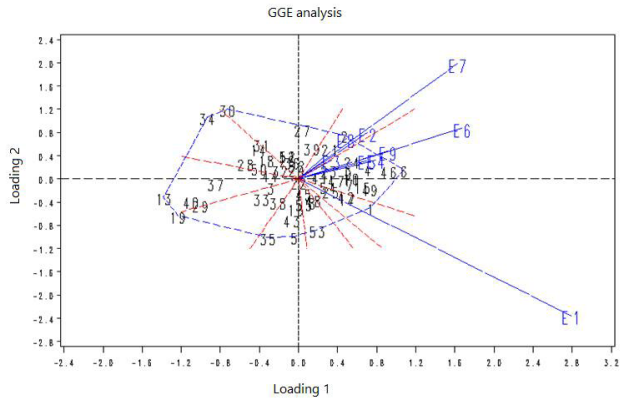


**Figure 1.** Biplot depicting the yield (t/ha) of 55 genotypes (G) of maize grain grown in 9 types of environments (E), using the FA model.

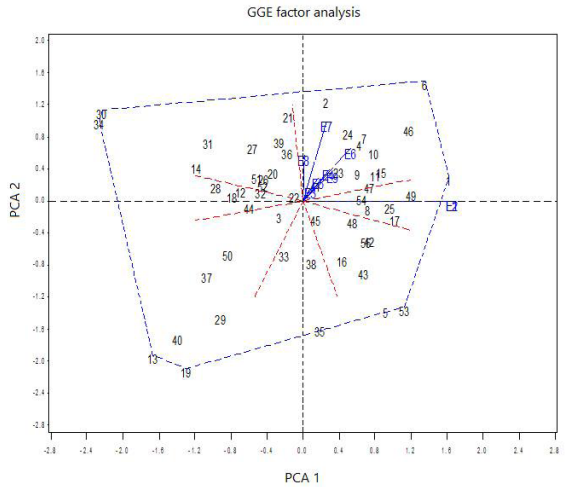


**Figure 2.** GGE-biplot depicting the yield (t/ha) of 55 genotypes (G) maize grain grown in 9 different type of environments.

A comparison of the two methods (represented by Figures 1 and 2) reveals only a small difference in the ranks of the genotypes and environments for the first axis. The adaptability and stability for most of the genotypes obtained from the GGE and FA biplots were similar for some environments.

Smith et al. (2002) demonstrated the relationship between the FA2 and AMMI2 models, while Burgueño et al. (2008) revealed the relationship between FA2 and SREG2. The FA model used in this study is very similar to the GGE2 model, wherein the effect of G is confounded with the G x E interaction. A visual comparison of the biplots generated by these models is clear because of the properties of the FA2 model; therefore, the biplots are similar to that of the GGE model biplot shown by Yan et al. (2009) (see also Crossa et al. (2010) and Stefanova and Buirchell (2010)). The adaptability of the genotypes in the GGE biplot is analyzed using approximate estimates given by the scores of principal component 1 (which is also related to the simple part of the G x E interaction). This relationship is contained in factor score one in the FA model, as negative factor loadings were not observed for the environments, and the factors scores were highly correlated with BLUPs (0.90). The stability of a genotype can be described by principal component 2 (which corresponds to the complex part of the G x E interaction) in the GGE analysis; similarly, the stability can also be described by the factor 2 score in the FA analysis. Therefore, productive and stable genotypes should have high Factor 1 scores, but values closer to zero for Factor 2, which suggests that these scores correspond to genotypes that are not specific to groups of environments.

Low scores indicate genotypes and/or environments that contribute little to the G x E interaction (or are not explained by the environment loadings), and are therefore considered as being stable. The genotypes classified as being stable were G12, G18, G22, G25, G32, G44, G45, G49, and G54. These genotypes can be highly recommended as long as they have high means; this was observed only for genotype G49. The environments that contributed the least to the interaction were E3, E5, E6, and E8.

The genotypes farthest from the origin are those that contribute the most to the interaction (that is, have a specific response to a group of environments). These were G5, G13, G6, G19, G29, G30, G34, G36, G37, G40, G46, and G53.

The results of cross-validation revealed that the FA model can be used to predict the performance of maize hybrids. Table 4 shows a moderate correlation (0.56-0.70) for all levels of unbalancing, which was inversely proportional to the level of missing data used. This table also revealed that the modes of the correlations ranged from 0.64 to 0.90, while the median ranged from 0.71 to 0.56, with a standard deviation of 0.13-0.21.

As the plots were removed at random, the percentage of unbalanced hybrids in the dataset also varied for each cycle. For example, with 10% missing plots, the total number of hybrids with missing data ranged from 25 to 38; however, a clear difference was not observed between the correlations of these results (Table 5).

**Table 4.** Descriptive statistics for the correlations between the observed and predicted values.

| Level of unbalancing (%) | Mean | Standard error | Median | Mode | Standard deviation | Minimum | Maximum |
|---|---|---|---|---|---|---|---|
| 10 | 0.70 | 0.01 | 0.71 | 0.90 | 0.21 | 0.12 | 0.96 |
| 30 | 0.60 | 0.05 | 0.58 | 0.79 | 0.14 | 0.23 | 0.91 |
| 50 | 0.56 | 0.04 | 0.56 | 0.64 | 0.13 | 0.03 | 0.86 |

**Table 5.** Descriptive statistics for the correlations between the observed and predicted values for an unbalanced dataset with 10% of the plots missing.

| Number of unbalanced hybrids | Mean | Minimum | Maximum | Variance |
|---|---|---|---|---|
| 25 | 0.82 | 0.75 | 0.91 | 0.05 |
| 26 | 0.69 | 0.53 | 0.89 | 0.03 |
| 27 | 0.66 | 0.53 | 0.89 | 0.03 |
| 28 | 0.70 | 0.34 | 0.94 | 0.04 |
| 29 | 0.64 | 0.23 | 0.97 | 0.05 |
| 30 | 0.71 | 0.05 | 0.95 | 0.05 |
| 31 | 0.71 | 0.12 | 0.97 | 0.06 |
| 32 | 0.68 | 0.02 | 0.96 | 0.05 |
| 33 | 0.71 | 0.21 | 0.96 | 0.05 |
| 34 | 0.71 | 0.25 | 0.97 | 0.04 |
| 35 | 0.65 | 0.33 | 0.96 | 0.04 |
| 36 | 0.70 | 0.39 | 0.97 | 0.03 |
| 37 | 0.58 | 0.10 | 0.88 | 0.13 |
| 38 | 0.82 | 0.39 | 0.76 | 0.06 |

The number of hybrids with missing data for the 30% unbalanced data ranged from 48 to 55 hybrids; almost all of the hybrids had missing data at a level of 50% of the missing plots. Regardless of the amount of unbalancing used, the correlation coefficients were above 0.5. Importantly, all hybrids were missing at least once from each site.

Another parameter used for validation was the PRESS statistic, and its variance. Table 6 summarizes the results for the PRESS statistic for the genotypes classified as stable or unstable (in Figure 2).

**Table 6.** Cross-validation of unbalanced data under different levels of missing plots in the various environments.

| Level of Unbalancing (%) | Genotypes | | | |
|---|---|---|---|---|
| | Stable | | Unstable | |
| | PRESS | Var (PRESS) | PRESS | Var (PRESS) |
| 10 | 0.19 | 0.06 | 0.50 | 0.56 |
| 30 | 0.31 | 0.06 | 0.77 | 0.82 |
| 50 | 0.32 | 0.07 | 0.78 | 0.98 |

PRESS statistic: prediction error sum of squares statistic.

As shown in Table 6, the PRESS was generally low, indicating that the model successfully predicted the BLUPs of the unbalanced genotypes. Moreover, the PRESS statistic differed for the stable and unstable groups of genotypes in the biplot analysis. The variances in PRESS values were used to improve the accuracy of the factor scores of the genotypes with greater stability than those of the unstable genotypes. This result suggested that the second factor score can be used to describe stable genotypes in the FA analysis with G + GE.

The PRESS statistic of the 10% unbalanced data was lower than that of the other levels of unbalancing for the two groups of genotypes, although the dispersion was practically the same. The predictive ability of the model for the 30 and 50% missing plot levels in the environments was similar within the two groups of genotypes. However, the dispersion of the PRESS statistic with 50% unbalancing was much higher in the unstable hybrids. Despite this, these results indicated that even with unbalancing levels of 30 to 50% of missing plots, the FA model effectively predicted the genotypic values of the missing hybrids.

The graphical representation (Figure 2) identifies the genotypes that are adaptable and stable in various environments, or the stable and unstable genotypes that contribute to the interaction. However, the graphical representation is only descriptive because it does not account for the uncertainty of the scores or factor loadings. Because of this limitation, this study proposes the use of empirical regions generated by factor score predictions for the measurement and validation of the stability of a given genotype. The goal was to incorporate this uncertainty, and facilitate the interpretation of the predicted scores of the missing genotypes, in a MET data analysis.

Figures 3, 4, and 5 display the confidence ellipses for genotypic score predictions, with 10, 30, and 50% unbalancing, respectively. These figures show that the stable genotypes (a) have ellipses that are concentrated greater around the mean, whereas unstable genotypes (b) have wider confidence regions.

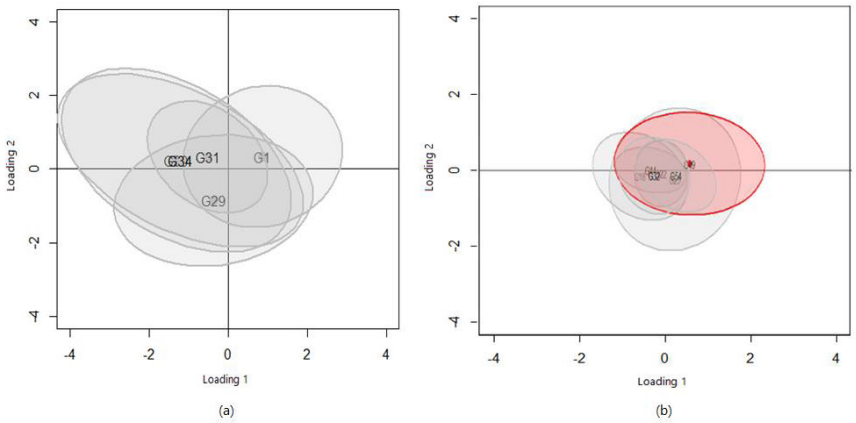


**Figure 3.** Empirical 95% confidence regions for the prediction of factor scores of (**a**) unstable and (**b**) stable genotypes for 10% unbalanced production data.

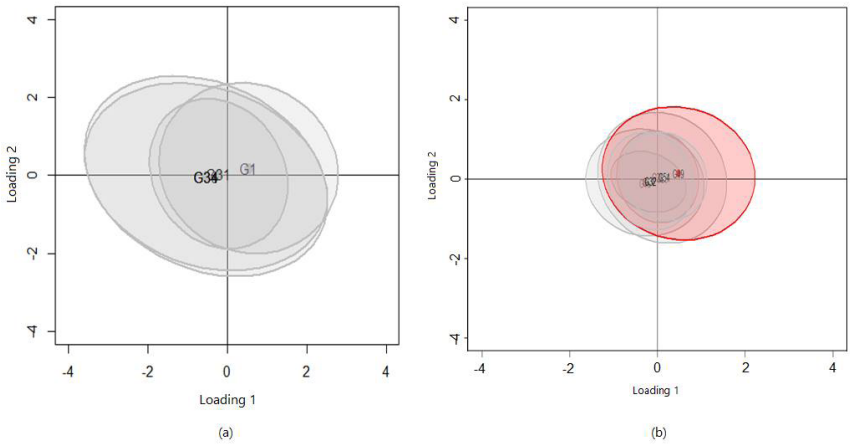


**Figure 4.** Empirical 95% confidence regions for the prediction of factor scores of (**a**) unstable and (**b**) stable genotypes for 30% unbalanced production data.

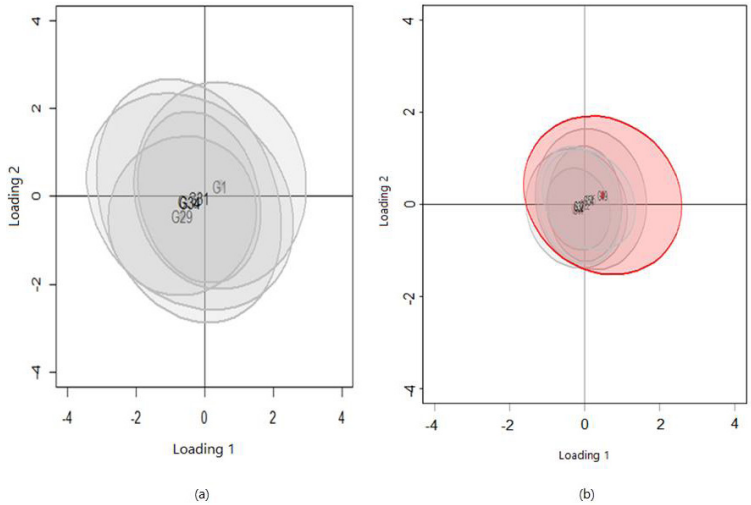


**Figure 5.** Empirical 95% confidence regions for the prediction of factor scores (BLUP) of (**a**) unstable and (**b**) stable genotypes for 50% unbalanced production data.

The exception to the rule was observed for the G49 (highlighted in red) genotype, which was classified as stable and productive in the biplot; however, its behavior was not particularly predictable in the cross-validation. Therefore, if G49 did not significantly contribute to the interaction, its prediction would have a small dispersion in the loading of the second factor. However, the identification of the stable genotypes in the factor score biplot corroborates its predictive ability in the cross-validation, suggesting that this technique can be used to measure the uncertainty of the stability captured in the biplot.

## DISCUSSION

The dynamics of a maize-breeding program requires the breeder to have a certain amount of flexibility when dealing with the introduction and removal of materials; Piepho (1998), Smith et al. (2001a,b, 2005), and Crossa et al. (2006) have attempted to calculate an approximation between multi-trait analysis and multi-environment analysis, using each environment as a variable. This approximation allowed the analysis of unbalanced and correlated data, in addition to heterogeneous variances.

Our results revealed that the use of mixed models for data from multiple environments can be applied to factors other than the study of genotypic-environment interactions. This model fits within the real dynamics of a breeding program, where the data is unbalanced by nature.

The predictive ability and fit of the FA2 analysis using the simulated unbalancing were deemed to be satisfactory; in addition, the components of the UN structure were minimally affected, even with 50% unbalancing. Although the FA model was not treated as an alternative to a UN model, instead treating it as a complementary technique for biplot analysis (or as a classic factor analysis in restricted likelihood), several results have demonstrated little difference in the predictions upon modeling of different structures (Kelly et al., 2007; Burgueño et al., 2007, 2008; Burgueño et al., 2011). These studies indicate that models with more than two factor loadings may (or may not)

increase the predictive capacity, while increasing the complexity of the model, and employing a larger information criterion. The results obtained by Burgueño et al. (2007, 2008) revealed that the FA models with more than two factors improved the variance and covariance estimates; however, this was not reflected in the EBLUPs of the genotypes (BLUPs). Crossa et al. (2006) reported that the use of an analytic factor G x E structure with two to nine factors only slightly affected the BLUPs, and did not alter the classification of the genotypes.

This study utilized a classic factor analysis as an alternative method to estimate the analytic factor models; here, the model was fit in two steps: through a multivariate mixed model, which had unstructured variance and covariance matrices, and an estimation of the FA model, which graphically visualized the genotype-environment relationship and studied the adaptability and stability.

Direct modeling of the UN matrix via FA is widely recommended in literature, although it is not always the best approach for a specific dataset. Often, this approach becomes a computational solution whose FA values have no direct meaning in terms of genetic variance, as it only captures a portion of the magnitude of the genetic variance, or reflects measurements outside of the parametric space [see the application of sparse matrices detailed by Thompson et al. (2003)].

The advantage of fitting the model in two steps via EM is justified because of the guaranteed convergence in the parametric space, thereby avoiding Heywood cases and leading to gains in the convergence and reduced computational demand for simpler datasets; in addition, models need not be selected directly with the estimation of parameters, which was proposed by Piepho (1998), Kelly (2007), and Resende (2007). Therefore, the UN structure may exhibit a large increase in predictive ability in cases with a large number of genotypes and good estimates of the (co)variance components (Balestre et al., 2012).

The major disadvantage of G x E variance structures of greater complexity is that a larger number of dispersion or covariance parameters must be estimated. In the presence of a large number of environments and few genotypes, the increased number of estimates for required variance components can result in problems of convergence, loss of efficiency, and increased computational demand (Welham et al., 2010). In addition, poor estimation of the (co)variance components can significantly reduce the predictive ability of the UN structures, compared to diagonal models (Balestre et al., 2012).

In traditional MET analyses, where GE is directly inserted into a mixed linear model, it is frequently assumed that the residual variance is similar for all observations. However, as observed in this study (and in others), heterogeneous residual effects must be included in the analysis (Smith et al., 2001b; Resende, 2007; Kelly et al., 2007; Rönnegard et al., 2010). Our results show that both the variances and residual covariances were heterogeneous. This may help the breeder group environments, as this value reflects the ranking of the genotypic interactions in the plots; environments with high residual correlations may be structurally more similar than those that only have similar genotype rankings.

According to Kelly et al. (2007), a common practice among plant breeders is to independently analyze the results of each trial. This method is equivalent to using the predictions from the diagonal variance model, which allows for heterogeneous variance but does not show correlations among the performances of the genotypes in all of the trials, or a potential residual covariance. The FA model adopted in this study is conceptually superior to this approach. This is because, in addition to the heterogeneity of the residual variance and its covariances, it captures a more complex covariance structure with regard to the genetic effect, which provides accurate predictions, both for individual trials and MET.

The values of the correlations measured in this study were moderate to high, depending on the position measurement used. The difference in these measurements revealed the asymmetric nature of the correlations; and the values with the highest density (mode) in the empirical distribution ranged from 0.9 to 0.64. These results showed the robustness of the G x E interaction prediction, modeled using mixed models in heterogeneous variance structures. For example, the correlation between predicted and observed values decreased by 4% for the 30 and 50% levels. This result suggested that the increase in the percentage of missing hybrids in the sites relative to the loss of plots did not interfere substantially with the prediction. Although these results are encouraging, we reiterate that the environments, genotypes, and levels of unbalancing were the limiting factors, compared to the large data set utilized in MET.

The uncertainty of ranking the genotypes in the biplots according to stability has been controversial (Yan et al., 2009; Yang et al., 2009; Crossa et al., 2010). Several analytical tools have been used, including bootstrap analysis, Bayesian analysis, and asymptotic confidence interval (Denis and Gower, 1996; Lavoranti et al., 2007; Crossa et al., 2011). In this study, we propose the application of the predictive ability of the FA models via cross-validation to validate the position of a genotype in the biplot. This approach does not destroy the GE structure, as posited by Yang et al. (2009), and allows for the prediction of missing data. Overall, the biplot results were consistent with those obtained by cross-validation, demonstrating the predictable behaviors of stable genotypes; this predictability was confirmed by the lower PRESS value of the stable genotypes compared to the unstable genotypes, and by the decreased prediction amplitude of the ellipses. Therefore, factor analysis with mixed models allows the unbalanced data and heterogeneous variances to be addressed, while also providing a clear graphical interpretation of the biplot similar to SREG2; in addition, this method provides the breeder with a method of validating stability via cross-validation, as this analysis tolerates high levels of unbalancing.

Therefore, our results suggest that PRESS can be used as an alternative approach for the evaluation of the performance of stable genotypes in the biplot. This result is confirmed by the amplitude of the prediction ellipses, which was smaller for these genotypes. Factor analysis using mixed models is robust under various levels of unbalanced data, with moderate to high correlations depending on the level of missing data. Therefore, this type of analysis could be used to evaluate the stability of plant breeding programs.

## Conflicts of interest

The authors declare no conflict of interest.

## ACKNOWLEDGMENTS

The authors thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico-CNPq, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-CAPES, Fundação de Amparo à Pesquisa do Estado de Minas Gerais-FAPEMIG, Fundção Colouste Gulbekian and Ministério da Ciência e Tecnologia de Moçambique-MCT, for financial support.

## REFERENCES

Balestre M, Pinho RGV, Sousa Junior CL and Bueno Filho JSS (2012). Bayesian mapping of multiple traits in maize: the importance of pleiotropic effects in studying the inheritance of quantitative traits. *Theor. Appl. Genet.* 125: 479-493.

Burgueño J, Crossa J, Cornelius PL, Trethowan R, et al. (2007). Modeling additive x environment and additive x additive x environment using genetic covariances of relatives of wheat genotypes. *Crop Sci.* 47: 311-320.

Burgueño J, Crossa J, Cornelius PL and Yang RC (2008). Using factor analytic models for joining environments and genotypes without crossover genotype x environment interaction. *Crop Sci.* 48: 1291-1305.

Burgueño J, Crossa J, Cotes JM, Vicente FS, et al. (2011). Prediction assessment of linear mixed models for multienvironment trials. *Crop Sci.* 51: 944-954.

Costa e Silva J, Potts BM and Dutkowski G (2006). Genotype by environment interaction for growth of Eucalyptus globulus in Australia. *Tree Genet. Genomes* 2: 61-75.

Crossa J, Yang RC and Cornelius PL (2004). Studying crossover genotype x environment interaction using linear-bilinear models and mixed models. *J. Agric. Biol. Environ. Stat.* 9: 362-380.

Crossa J, Burgueño J, Cornelius PL, McLaren G, et al. (2006). Modeling genotype x environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. *Crop Sci.* 46: 1722-1733.

Crossa J, de Los Campos G, Pérez P, Gianola D, et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186: 713-724.

Crossa J, Perez-Elizalde S, Jarquin D, Cotes JM, et al. (2011). Bayesian estimation of the additive main effects and multiplicative interaction model. *Crop Sci.* 51: 1458-1469.

Cruz CD, Regazzi AJ and Carneiro PCS (2004). Modelos biométricos aplicados ao melhoramento genético. 3rd edn. Editora UFV, Viçosa.

Dempster AP, Laird NM and Rubin DB (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. B* 39: 1-38.

Denis JB and Gower JC (1996). Asymptotic confidence regions for biadditive models: Interpreting genotype-environment interactions. *Appl. Statist.* 45: 479-493.

Johnson RA and Wichern DW (2007). Applied multivariate statistical analysis. 6th edn. Prentice-Hall Inc., New Jersey.

Kelly AM, Smith AB, Eccleston JA and Cullis BR (2007). The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. *Crop Sci.* 47: 1063-1070.

Lavoranti OJ, Dias CT dos S and Kraznowsk WJ (2007). Phenotypic stability via AMMI model with bootstrap re-sampling. *Bol. Pesq. Florestal* 2: 45-52.

Lin CS, Binns MR and Lefkovitch LP (1986). Stability analysis: Where do we stand? *Crop Sci.* 26: 894-900.

Machado JC, Souza JC, Ramalho MAP and Lima JL (2008). Estabilidade de produção de híbridos simples e duplos de milho oriundos de um mesmo conjunto gênico. *Bragantia* 67: 627-631.

Patterson HD and Nabugoomu F (1992). REML and the analysis of series of crop variety trials. Proceedings of the XVIth International Biometric Conference. Hamilton, 77-93.

Patterson HD, Silvey V, Talbot M and Weatherup STC (1977). Variability of yields of cereal varieties in U.K. trials. *J. Agric. Sci.* 89: 239-245.

Piepho HP (1997). Analysing genotype-environment data by mixed models with multiplicative terms. *Biometrics* 53: 761-767.

Piepho HP (1998). Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. *Theor. Appl. Genet.* 97: 195-201.

Piepho HP and Möhring J (2006). Selection in cultivar trials: Is it ignorable? *Crop Sci.* 46: 192-201.

R Core Team (2014). R: a language and environment for statistical computing. Available at [http://www.R-project.org]. Accessed May 10, 2014.

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

Resende MDV and Thompson R (2004). Factor analytic multiplicative mixed models in the analysis of multiple experiments. *Rev. Mat. Estat.* 22: 31-52.

Rönnegard L, Felleki M, Fikse F, Mulder HA, et al. (2010). Genetic heterogeneity of residual variance - estimation of variance components using double hierarchical generalized linear models. *Genet. Sel. Evol.* 42: 8.

Stefanova K and Buirchell B (2010). Multiplicative mixed models for genetic gain assessment in Lupin breeding. *Crop Sci*. 50: 880-891.

Statistical Analysis System Institute. SAS. Version 9.3. Cary, 2013. Available at [http://hostname:port/SASLogon/sas-environment.xml]. Accessed May 20, 2014.

Smith A, Cullis B and Gilmour A (2001a). Applications: the analysis of crop variety evaluation data in Australia. *Aust. N.Z. J. Stat.* 43: 129-145.

Smith A, Cullis B and Thompson R (2001b). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* 57: 1138-1147.

Smith A, Cullis B, Luckett D, Hollamby G, et al. (2002). Exploring variety-environment data using random effects AMMI models

with adjustments for spatial field trend. Part 2: Applications. In: Quantitative genetics, genomics and plant breeding (Kang MS, ed.). CABI Publishing, 337-351.

Smith AB, Cullis BR and Thompson R (2005). The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. *J. Agric. Sci.* 143: 449-462.

Thompson R, Cullis B, Smith A and Gilmour AR (2003). A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. *Aust. N.Z. J. Stat*. 45: 445-459.

Welham SJ, Gogel BJ, Smith AB, Thompson R, et al. (2010). A comparison of analysis methods for late-stage variety evaluation trials. *Aust. N.Z. J. Stat.* 52: 125-149.

Yan W, Glover KD and Kang MS (2010). Comment on "Biplot analysis of genotype x environment interaction: proceed with caution", by Yang R-C, Crossa J, Cornelius PL, and Burgueño J in 2009, 49: 1564-1576. *Crop Sci.* 50: 1121-1123.

Yan W, Hunt LA, Sheng Q and Szlavnics Z (2000). Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Sci.* 40: 597-605.

Yan W, Kang MS, Ma B, Woods S and Cornelius PL (2007). GGE biplot vs. AMMI analysis of genotype-by-environment data. *Crop Sci.* 47: 643-653.

Yang RC, Crossa J, Cornelius PL and Burgueño J (2009). Biplot analysis of genotype x environment interaction: proceed with caution. *Crop Sci.* 49: 1564-1576.