FACTOR ANALYTIC MULTIPLICATIVE MIXED MODELS IN THE ANALYSIS OF MULTIPLE EXPERIMENTS

Marcos Deon Vilela de RESENDE[1]
Robin THOMPSON[2]

- ABSTRACT: Analysis of groups of experiments or multi-environment trials (MET) has been traditionally based on simple models assuming error variance homogeneity between trials, independent error within trials, genotype x environment (*g* x *e*) effects as a set of independent random effects. The combined analysis of MET data through realistic models is a complex statistical problem which requires extensions to the standard linear mixed model. The relaxation of the assumption concerning the independence of *g* x *e* effects can be achieved with the use of multiplicative models. Such models have been popularised as additive main effects and multiplicative interaction effects (AMMI) and a number of applications have been found. However, AMMI analysis presents at least five great limitations: it considers the genotype and *g* x *e* effects as fixed; it is suitable only for balanced data sets; it does not consider spatial variation within trials; it does not consider the heterogeneity of variance between trials; it does not consider the different number of replications across sites. These features are not realistic in analysing field data. In a mixed model setting, Piepho (1998) presented a factor analytic multiplicative mixed (FAMM) model with random genotype and *g* x *e* effects which is conceptually and functionally better than AMMI. In the same context, Smith et al. (2001) presented a general class of FAMM models that encompass the approach of Piepho (1998) and include separate spatial errors for each environment (FAMMS). Such general class of models provides a full realistic approach for analysing MET data. The present paper deals with the application of FAMM and FAMMS models in two large unbalanced data sets (on eucalypt and tea plant) aiming at emphasising their advantages over AMMI models in terms of the assumptions of error variance homogeneity between trials and independent error within trials. Also, the ability of FAMM models in providing parsimonious models is also stressed. Parsimonious FAMM models were found for the two data sets. There were great advantages of heterogeneous variance FAMM models over homogeneous variance FAMM models. This reveals the superiority of FAMM models over AMMI models. It was noted that there was heterogeneity among the specific variances in individual environments; therefore, factor analytic models with common specific variances for all sites were not suitable. FAMM models provided estimates of the full correlation structure, facilitating practical decisions to be made. FAMM models with heterogeneous variance among traits and spatial errors within traits were advantageous over FAMM models with variance homogeneity and non-spatial error. This also shows the superiority of FAMM models over AMMI models, which do not allow for dependent or spatial errors. For analysing multi-environment data sets with longitudinal data, FAMMS models proved to be a very useful tool.

[1] Embrapa Florestas - EMBRAPA, Caixa Postal 319, CEP: 83411-000, Colombo, PR, Brasil. Email: deon@cnpf.embrapa.br
[2] Biomathematics Unit, Rothamsted Research, AL5 2JQ – Harpenden - Herts, England. E-mail: robin.thompson@bbsrc.ac.uk

▪ KEYWORDS: Factor analytic multiplicative mixed models; factor analytic multiplicative mixed spatial models; additive main effects and multiplicative interaction effects; restricted maximum likelihood; best linear unbiased prediction; multi-environment trials; stability analysis.

1 Introduction

Analysis of experiments repeated on several sites or environments are very common and important in agriculture. Such trials aim at providing inferences concerning for responses on both broad (in the average of all sites) and specific environments. To attain this, all the information should be analysed simultaneously. Traditional analysis of these multi-environment trials (MET) has been made through joint analysis of variance (ANOVA) and linear regression techniques. In general, stability and adaptability approaches (Finlay and Wilkinson, 1963; Eberhart and Russell, 1966) have been used to study treatment x environment interaction, mainly referred to as genotype x environment interaction or *g* x *e*. In spite of their generalised use, these regression-based methods present limitations that have been reported in the literature, such as inefficiency in the presence of non-linearity, generating simplified response models (Crossa, 1990; Duarte and Vencovsky, 1999). Some proposed models (Cruz et al., 1989; Toler and Burrows, 1998) correct this inefficiency, but the *g* x *e* component has been estimated but not decomposed into the pattern (tendency) and noise components.

A first attempt to circumvent these limitations was the proposed technique called AMMI (Additive Main Effects and Multiplicative Interaction Analysis). This technique was well described by Gauch (1988; 1992) and attributed to Fisher and Mackenzie (1923) and Gollob (1968). Another denomination of the method is PCA (Doubled Centred Principal Components Analysis). AMMI may be viewed as a procedure to separate pattern (the *g* x *e* interaction) from noise (mean error of treatment mean within trials). This is achieved by PCA, where the first axes (i.e. the axes with the largest eigenvalues) recover most of the pattern, whilst most of the noise ends up in later axes. The pattern can be viewed as the whole *g* x *e* effect weighed by an estimate of the pattern-to-noise ratio associated with the respective effect. This pattern-to-noise ratio is a variance component ratio analogue to a repeatability or heritability coefficient (Piepho, 1994). Multiplicative models AMMI have been popularised in a fixed model context and a number of applications have been found (Gauch, 1988; 1992; Crossa et al., 1990). AMMI analysis combines, in a model, additive components for main effects (treatments and environments) and multiplicative components for *g* x *e* effects. It combines a univariate technique (ANOVA) for the main effects and a multivariate technique (PCA-principal component analysis) for *g* x *e* effects. Crossa (1990) suggests that the use of multivariate techniques permits a better use of information than the traditional regression methods.

Although useful, AMMI models present at least five great limitations: they consider genotype and *g* x *e* effects as fixed; they are suitable only for balanced data sets; they do not consider spatial variation within trials; they do not consider heterogeneity of variance between trials; they do not consider different number of replications across sites. These features are not realistic in analysing field data, where the data are generally unbalanced and many of treatments (genotypes) do not support the assumption of fixed genotype effects (implicit heritability at mean level equal to 1). The AMMI model estimates phenotypic and non-genotypic values. If genotypes are considered as random, effects can be predicted by the best linear unbiased

prediction (BLUP). Hill and Rosenberger (1985) and Stroup and Mulitze (1991) showed that assuming random genotypes may be preferable in terms of predictive accuracy even when genotypes would be considered fixed by conventional standards. Assuming genotype as random effects, it is possible to obtain shrinkage predictions of the random interaction *g* x *e* terms and thus separate pattern and noise as do AMMI models. In this sense, BLUP and AMMI can be seen as two approaches to achieve the same goal, namely to separate pattern from noise. The BLUP procedure produces the generalised least square (GLS) estimates of interaction effects and then weighs them by an estimate of the correspondent pattern-to-noise ratios. However, the BLUP procedure has a number of advantages that circumvent all the limitations of AMMI. It has also been shown that BLUP can be predictively more accurate than AMMI models (Piepho, 1994).

The full multivariate BLUP model is the best approach for analysing data on multiple experiments. This model provides response on each environment through the use of all information and also considers variance heterogeneity. However, with a large number of experiments the mixed model analysis is unlikely to converge. The variance-covariance matrix in this case is completely unstructured, which means a large number of parameters to be estimated. So, the parsimonious model behind AMMI is an interesting feature. Van Eeuwijk et al. (1995) suggested obtaining a genotype by environment BLUP and then subject this table to AMMI analysis, using a single value decomposition procedure. A better approach was found by Piepho (1998). In a mixed model setting, he presented a multiplicative factor analytic model with random genotype and *g* x *e* effects which is conceptually and functionally better than AMMI. In the same context, Smith et al. (2001) presented a general class of factor analytic multiplicative mixed models that encompasses the approach of Piepho (1998) and includes separate spatial errors for each environment. Such general class of models provides a full realistic approach for analysing MET data (Thompson et al., 2003).

The multivariate technique of factor analysis (Lawley and Maxwell, 1971; Mardia et al. 1988) provides simplification of correlated multivariate data as do other multivariate methods such as principal components analysis and canonical transformation. These techniques consider the correlation between variables and generate a new set of independent (non-correlated) variables. The factor analysis technique can be considered as an extension of the principal component analysis. The factor analytic variance-covariance structure may be regarded as an approximation to the completely unstructured variance-covariance matrix and can provide parsimonious models.

Analysis of multi-environment trials (MET) has also been traditionally based on simple models assuming error variance homogeneity between trials, independent error within trials, genotype x environment (*g* x *e*) effects as a set of independent random effects. The combined analysis of MET data through realistic models is a complex statistical problem which requires extensions to the standard linear mixed model. Such extensions have been done recently. Cullis et al. (1998) presented a spatial mixed model analysis for MET data, which fits a separate error structure for each site, circumventing the assumptions of error variance homogeneity among trials and independent error within trials. The relaxation of the assumption concerning the independence of *g* x *e* effects can be achieved with the use of multiplicative models.

In a mixed model setting, multiplicative models for random *g* x *e* interaction terms induce correlations between the interactions. Mixed models with multiplicative terms are closely related to the so-called factor analytic variance-covariance structure advocated by Jennrich and Schluchter (1986). Piepho (1997) proposed multiplicative

mixed models for multi-environment analysis but assumed random environment rather than random genotype effects. The same author proposed the use of factor analytic multiplicative mixed (FAMM) models with random genotype effects (Piepho, 1998). Smith et al. (2001) presented a general class of FAMM models that encompass the approach of Piepho (1998) and provides: accounting of heterogeneity of *g* x *e* variance; accounting of correlation among *g* x *e* interactions; appropriate spatial error variance structures for individual trials. This factor analytic multiplicative mixed spatial (FAMMS) model provides parsimonious models for large multivariate data sets and a better conceptual approach for interaction effects based on the multiplicative model. The model can be regarded as a random effects analogue of AMMI. Smith et al. (2001) reported that the advantages of FAMMS models are numerous and include: (i) within trial spatial variation can be accommodated; (ii) between trial error variance heterogeneity can be accommodated; (iii) unbalanced data are easily handled; (iv) genotype effects and *g* x *e* interactions can be regarded as random, leading to better predictions; (v) the goodness of fit of the model, i.e., number of multiplicative terms needed, can be formally tested through residual maximum likelihood ratio tests (REMLRT). Through a unified mixed model approach stability and adaptability parameters are integrated into broad (selection for an average environment), specific (selection for specific environments) and new-environment (selection for a non-tested environment) inferences.

The present paper deals with the application of FAMM and FAMMS models in two large unbalanced data sets aiming at emphasising their advantages over AMMI models in terms of the assumptions of error variance homogeneity between trials and independent error within trials. Also, the ability of FAMM models in providing parsimonious models is stressed.

2 Material and methods

2.1 Factor analytic models

A model concerning the evaluation of several treatments or genotypes in several environments is given by:

$$Y_{ij} = \mu + g_i + e_j + ge_{ij} + \varepsilon_{ij},$$

where: μ, *g, e, ge* and ε are the fixed constant, genotype, environment, genotype x environment interaction and within environment error effects, respectively. The μ and *e* effects can be regarded as fixed and the others as random. A model referring to random genotype effects in each environment can be written as:

$$Y_{ij} = \mu + g_{ij} + e_j + \varepsilon_{ij}.$$

In the context of MET data, the factor analysis approach can be used to provide a class of structures for the variance-covariance matrix of g_{ij} (*G*). The model is postulated in terms of the unobservable genotype effects in different environments:

$$g_{ij} = \sum_{r=1}^{k} \lambda_{jr} f_{ir} + \delta_{ij} \ ,$$

where:

g_{ij} : effect of genotype *i* in environment *j*;

λ_{jr} : loading for factor *r* in environment *j*;

f_{ir} : score for genotype i in factor *r*;

δ_{ij} : error representing the lack of fit of the model.

The FAMM model is presented according to Smith et al. (2001). Applied to *g* genotype effects on s environments, the factor analytic model postulates dependence on a set of random hypothetical factors $f_r^{(g\,x1)},(r=1....k<s)$. In vector notation, the factor analytic model for these effects is

$$g_s=(\lambda_1\otimes I_g)f_1+....+(\lambda_k\otimes I_g)f_k+\delta\ ,$$

where:

$\lambda_r^{(s\,x1)}$: loadings or weights of the factors in environments;

$\delta^{(gs\,x1)}$: vector of residuals or lack of fit for the model (also called vector of specific factors).

In a compact way, the model is:

$$g_s=(\Lambda\otimes I_g)f\ +\delta\ ,$$

where:

$\Lambda^{(s\,x\,k)}=[\lambda_1.....\lambda_k]$;

$f^{(gk\,x1)}=(f_1^{'},f_2^{'}....f_k^{'})'$.

The joint distribution of f *and* δ is given by

$$\begin{pmatrix} f \\ \delta \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix},\begin{pmatrix} I_k\otimes I_g & 0 \\ 0 & \Psi\otimes I_g \end{pmatrix}\right],$$

where:

$\Psi=\text{diag}(\psi_1....\psi_p)$;

ψ_i : specific variance for the *i*th trial.

The variance matrix for genotype effects on environments is given by

$$\text{var}(g_s)=(\Lambda\otimes I_g)\,\text{var}(f)(\Lambda^{'}\otimes I_g)\ +\text{var}(\delta)=(\Lambda\Lambda^{'}+\Psi)\otimes I_g.$$

The model for genotype effects in each environment leads to a model for *G* in which:

$\sigma_{g_{jj'}}=\sum_{r=1}^{k}\lambda_{jr}^2+\psi_j$: genotype variance in environment *j*;

$\sigma_{g_{jj'}}=\sum_{r=1}^{k}\lambda_{jr}\lambda_{j'r}$: genotype covariance between environments *j* and *j*';

$\rho_{g_{jj'}}=\sum_{r=1}^{k}\lambda_{jr}\lambda_{j'r}/[(\sum_{r=1}^{k}\lambda_{jr}^2+\psi_j)(\sum_{r=1}^{k}\lambda_{j'r}^2+\psi_{j'})]^{1/2}$: genotype correlation between environments *j* and *j*'

The equation for g_s has the form of a (random) regression on *k* environmental covariates $\lambda_1 ....\lambda_k$ in which all regressions pass through the origin. It may be more appropriate to allow a separate (non-zero) intercept for each genotype. This is equivalent to the model with genotype main effects, g, and a k-factor analytic model for *g* x *e* interaction. Then, the expression for g_s turns to

$$g_s = (1_s \otimes I_g)g + ge = (1_s \otimes I_g)g + (\Lambda \otimes I_g)f + \delta.$$

Vector *g* has mean zero and variance $\sigma_g^2 I$ or $\sigma_g^2 A$, where A is a genetic relationship matrix. The model can be written as

$$g_s = (\sigma_g 1_s \otimes I_g)f_0 + (\Lambda \otimes I_g)f + \delta = (\Lambda_g \otimes I_g)f_g + \delta,$$

where:

$\Lambda_g^{s(k+1)} = [\sigma_g 1_s \ \Lambda]; \quad f_0 = g/\sigma_g; \quad f_g^{'} = (f_0^{'} f^{'}).$

Thus the model with genotype main effects and a k-factor analytic model for *g* x *e* interactions is a special case of a (*k*+1)-factor analytic genotype effects in each environment, in which the loadings in the first set are constrained to be equal.

The feature that distinguishes equations for *g*, from standard random multivariate regression problems is that both the covariates and the regression coefficients are unknown and therefore must be estimated from the data. The model is then a multiplicative model of environment and genotypes coefficients (known as loadings and factorial scores, respectively). Here lies the analogy with AMMI models. However, a key difference is that the multiplicative model in equation for g_s accommodates random effects, whereas AMMI is a fixed-effects model. FAMM models are also called random AMMI.

2.2 General linear mixed model and REML estimation of factor analytic, multivariate and spatial models

A general linear mixed model has the form (Henderson, 1984; Thompson et al., 2003):

$$y = X\beta + Z\tau + \varepsilon, \tag{1}$$

with the following distributions and structures of means and variances:

$\tau \sim N(0, G)$ $\qquad$ $E(y) = X\beta$

$\varepsilon \sim N(0, R)$ $\qquad$ $Var(y) = V = ZGZ' + R$

where:

y: known vector of observations.

β: parametric vector of fixed effects, with incidence matrix *X*.

τ: parametric vector of random effects, with incidence matrix Z.

ε: unknown vector of errors.

G: variance-covariance matrix of random effects.

R: variance-covariance matrix of errors.

0: null vector.

Assuming *G* and *R* as known, the simultaneous estimation of fixed effects and the prediction of the random effects can be obtained through the mixed model equations given by:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z+G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \tilde{\tau} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

The solution to this system of equations for $\hat{\beta}$ and $\tilde{\tau}$ leads to identical results as that obtained by:

$\hat{\beta} = (X'V^{-1}X)^{-}X'V^{-1}y$: generalised least square estimator (GLS) or best linear unbiased estimator (BLUE) of β;

$\tilde{\tau} = GZ'V^{-1}(y - X\hat{\beta}) = C'V^{-1}(y - X\hat{\beta})$: best linear unbiased predictor (BLUP) of τ; where *C' = GZ'*: covariance matrix between τ and *y*.

When *G* and *R* are not known, the variance components associated can be estimated efficiently through the REML procedure (Patterson & Thompson, 1971; Thompson and Welham, 2003). Except for a constant, the residual likelihood function (in terms of its log) to be maximised is given by:

$$L = -\frac{1}{2}\ (\log|X'V^{-1}X| + \log|V| + v\ \log\ \sigma_\varepsilon^2 + y'Py/\sigma_\varepsilon^2)$$

$$= -\frac{1}{2}\ (\log|C^*| + \log|R| + \log|G| + v\ \log\ \sigma_\varepsilon^2 + y'Py/\sigma_\varepsilon^2)$$

where:

$V = R + ZGZ'; \qquad P = V^{-1} - V^{-1}X\ (X'V^{-1}X)^{-1}X'V^{-1}.$

v = N-r(x): degrees of freedom, where *N* is the total number of data and *r(x)* is the rank of matrix X.

*C**: Coefficient matrix of the mixed model equations.

Being general, the model (1) encompasses several models inherent to different situations, such as:

Multivariate models

In the bivariate case:

$$Z = \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix}; \quad \tau = \begin{bmatrix} \tau_1 \\ \tau_2 \end{bmatrix};$$

$$G = A \otimes G_O; \quad R = I \otimes R_O;$$

$$G_O = \begin{bmatrix} \sigma_{\tau_1}^2 & \sigma_{\tau_{12}} \\ \sigma_{\tau_{21}} & \sigma_{\tau_2}^2 \end{bmatrix}; \quad R_O = \begin{bmatrix} \sigma_{\varepsilon_1}^2 & \sigma_{\varepsilon_{12}} \\ \sigma_{\varepsilon_{21}} & \sigma_{\varepsilon_2}^2 \end{bmatrix} \quad or \quad R_O = \begin{bmatrix} \sigma_{\varepsilon_1}^2 & 0 \\ 0 & \sigma_{\varepsilon_2}^2 \end{bmatrix}, \ where:$$

$\sigma_{\tau_{12}}$: random treatment effects covariance between variables 1 and 2.

$\sigma_{\varepsilon_{12}}$: residual covariance between variables 1 and 2.

Spatial models (time series or geostatistical)

$R = \Sigma$: non-diagonal matrix that considers the correlation between residuals through ARIMA models or covariance based on adjusted semivariance.

In the context of agricultural experiments, the general spatial model developed by Martin (1990) and Cullis and Gleeson (1991) has the following form:

$y = X\beta + Z\tau + \xi + \eta$, where:

y: known vector of data, ordered as columns and rows within columns;

τ: unknown vector of treatment effects;

β: unknown vector representing spatial variation at large scale or global tendency (block effects, polynomial tendency);

ξ: unknown vector representing spatial variation at small scale (within blocks) or local tendency, modelled as a random vector with zero mean and spatially dependent variance;

η: unknown vector of independent and identically distributed errors.

Through ARIMA models, the error is modelled as a function of a tendency effect (ξ) plus a non correlated random residual (η). So, the vector of errors is partitioned into $\varepsilon = \xi + \eta$, where ξ and η refer to spatially correlated and independent errors, respectively. Traditional models of analysis do not include the ξ component.

Considering an experiment with rectangular shape in a grid of c columns and r rows, the residuals can be arranged in a matrix in a way that they can be considered as correlated within columns and rows. Writing these residuals in a vector following the field order (by putting each column beneath another), the variance of residuals is given by $Var(\varepsilon) = Var(\xi + \eta) = R = \Sigma = \sigma_\xi^2[\sum_c(\Phi_c) \otimes \sum_r(\Phi_r)] + I\sigma_\eta^2$, where σ_ξ^2 is the variance due to local tendency and σ_η^2 is the variance of the independent residuals.

Matrices $\sum_c(\Phi_c)$ *and* $\sum_r(\Phi_r)$ refer to first order autoregressive correlation matrices with auto-correlation parameters Φ_c and Φ_r and order equal to the number of columns and rows, respectively. In this case, ξ is modelled as a separable first order auto-regressive process (AR1 x AR1) with covariance matrix

$$Var(\xi) = \sigma_\xi^2[\sum_c(\Phi_c) \otimes \sum_r(\Phi_r)] \cdot$$

The mixed model equations and variance structure for spatial factor analytic models can be given by

$$\begin{bmatrix} \hat{\beta} \\ \tilde{g}_s \\ \tilde{\kappa} \end{bmatrix} = \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z & X'R^{-1}W \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} & Z'R^{-1}W \\ W'R^{-1}X & W'R^{-1}Z & W'R^{-1}W + C^{-1} \end{bmatrix}^{-1} \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \\ W'R^{-1}y \end{bmatrix}$$

where:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_s \end{bmatrix}; \qquad \tilde{g}_s = \begin{bmatrix} \tilde{g}_1 \\ \vdots \\ \tilde{g}_s \end{bmatrix}; \qquad \tilde{\kappa} = \begin{bmatrix} \tilde{\kappa}_1 \\ \vdots \\ \tilde{\kappa}_s \end{bmatrix}$$

$$R^{-1} = R_O^{-1} \otimes H^{-1}; \qquad G^{-1} = G_O^{-1} \otimes A^{-1}; \qquad C^{-1} = C_O^{-1} \otimes I$$

$$R_O = \begin{bmatrix} \sigma^2_{\xi_1} & 0 \\ 0 & \sigma^2_{\xi_s} \end{bmatrix}; \qquad G_O = \begin{bmatrix} \sigma_{g11} & \sigma_{g1s} \\ \sigma_{g1s} & \sigma_{gss} \end{bmatrix}; \qquad C_O = \begin{bmatrix} \sigma^2_{\kappa_1} & 0 \\ 0 & \sigma^2_{\kappa_s} \end{bmatrix},$$

where:

$R^{-1} = \begin{bmatrix} H_1\sigma^2_{\xi_1} & 0 \\ 0 & H_s\sigma^2_{\xi_s} \end{bmatrix}^{-1}$;

β and κ: vectors of fixed effects and random plot effects, respectively.;

$H_1 = [\sum_{c_1}(\Phi_{c_1}) \otimes \sum_{\eta}(\Phi_{\eta})]$: spatial correlation matrix for environment 1;

$H_s = [\sum_{c_s}(\Phi_{c_s}) \otimes \sum_{r_s}(\Phi_{r_s})]$: spatial correlation matrix for environment *s*;

$H = \begin{bmatrix} H_1 & 0 \\ 0 & H_s \end{bmatrix}$.

In this case, the genotype main effects are fitted implicitly in $\tilde{g}_s = [\tilde{g}_1 \ldots \tilde{g}_s]'$. The explicit fitting of genotype main effects term is achieved by including another random vector for these main effects in the mixed model equations. After that, the $\tilde{g}_s$ effects in the mixed model equations will represent *g* x *e* interactions.

Solving the mixed model equations above provides BLUPs of genotype effects in individual environments. The BLUPs of the genotype's factorial scores *f* can then be obtained from $\tilde{g}_s$ as

$$\tilde{f}_s = \hat{\text{var}}(f)[Z\ (\hat{\Lambda} \otimes I_g)]' \hat{P}y = [\hat{\Lambda}'(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})^{-1} \otimes I_g]\tilde{g}_s.$$

The estimates are:

$\hat{\Lambda}$: matrix of estimated loadings;

$\hat{\Psi}$: matrix of estimated specific variances.

The BLUPs of the residuals of the *g* x *e* interactions can be obtained by

$$\tilde{\delta} = [\hat{\Psi}(\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})^{-1} \otimes I_g]\,\tilde{g}_s.$$

It can be seen that the factor analytic model requires calculations of parameters Λ and Ψ which compose the variance-covariance matrix G_0, and can be estimated by REML (Patterson and Thompson, 1971) through the algorithm average information (Gilmour et al., 1995; Johnson and Thompson, 1995). A specific REML algorithm for factor analytic models was developed by Thompson et al. (2003).

With assumption of model $y = X\beta + Z[(\Lambda \otimes I_g)f + \delta] + \varepsilon$, the predicted effects of genotypes in an average environment ($\tilde{g}_{\bar{s}}$) can be given by the formula:

$$\tilde{g}_{\bar{s}} = \bar{\hat{\beta}} + [(\bar{\hat{\lambda}}_1\, \bar{\hat{\lambda}}_2 \ldots\ldots \bar{\hat{\lambda}}_k) \otimes I_g]\tilde{f}\,.$$

Quantities $\bar{\hat{\lambda}}_r$ and $\tilde{f}$ are the mean across environments of the estimated loadings for the *r*th factor, and the estimated factorial scores for genotypes, respectively. This is a prediction at the average values of the loadings. By definition of the loadings, these are predictions of genotype means for an environment that is average in the sense of having average covariance with all other environments. The prediction of overall genotype performance is the same irrespective of the inclusion of genotype main effects on the model. The question concerning the interpretation of the genotype main effects included is important. These are not main effects in the usual sense,

namely a measure of overall genotype performance, but merely intercepts in the regression. They, therefore, reflect genotype performance in an environment that has zero values of the loadings. That inclusion would provide results of genotype main effects which are identical to the predicted values for an average environment ($\tilde{g}_{\bar{s}}$) (Smith et al. 2001).

One form of obtaining the overall performance of genotypes is by forming the two-way table of predicted genotype means for each environment and then averaging across environments to obtain the overall genotype means. These predicted means are also given by the formula:

$$\tilde{g}_{\bar{s}m} = \bar{\hat{\beta}} + [(\bar{\hat{\lambda}}_1\ \bar{\hat{\lambda}}_2 .....\bar{\hat{\lambda}}_k) \otimes I_g]\tilde{f} + \bar{\hat{\delta}}$$

This formula differs from $\tilde{g}_{\bar{s}}$ only by the addition of the unexplained *g* x *e* effects, which refer to the lack of fit from the factor analysis. This overall performance is only likely to be a good predictor if the correlation of genotype in different environments is high.

2.3 Constraints and rotation on loadings and interpretation of environmental loadings and factorial scores

When the number *k* of factors is greater than 1, constraints must be imposed on the factor analytic parameters in order to ensure identifiability. This arises because the distribution of $(\Lambda \otimes I_g) f$ is singular. It can be shown that *k*(*k*-1)/2 independent constraints must be imposed on the elements of Λ. According to Mardia et al. (1988), the factor analytic model is not unique under rotation so the constraints must be chosen to ensure uniqueness. A set of constraints that fulfils this requirement is to set all *k*(*k*-1)/2 elements in the upper triangle of Λ are zero, i.e., $\lambda_{jr} = 0\ for\ j < r = 2...k$ (Jennrich and Schluchter, 1986). The implication of the constraints is that the number of variance parameters in the factor analytic model with k terms is given by *pk*+*p*-*k*(*k*-1)/2 (Smith et al., 2001).

The nonuniqueness of Λ when $k > 1$ introduces ambiguity in the interpretation of the environmental loadings and genotype scores. The constrained form of Λ is merely for computational ease and has no biological basis. So, rotation of loadings is advocated for generating meaningful results. Lawley and Maxwell (1971) describe a number of useful rotations. In MET data the required rotation is $\Lambda^* = \Lambda T$,where T is an orthogonal matrix. According to Johnson and Wichern (1988), the axes can then be rotated in a certain angle ϕ and the rotated loadings can be given by $\Lambda^* = \Lambda T$, with $T = \begin{bmatrix} \cos\phi & sen\phi \\ -sen\phi & \cos\phi \end{bmatrix}$.

The loadings from factor analytic models are useful for clustering environments in terms of genetic correlations. The graphical display of loadings from a model with $k > 1$ can be very informative in this respect.

In factor analysis, the main interest is centred on the parameters of the factor model. Nevertheless, the predicted values of the common factors, named factor scores, are particularly useful in cluster analysis. Besides their utility in predicting genotype averages, the genotype's factorial scores can also be plotted for factors 1 and 2 for example, allowing for inference about the grouping of genotypes based on their similarity.

2.4 Selection of FAMM models

In a search for parsimonious models the adequacy of the FAMM models of several *k* orders can be formally tested, as it is fitted within a mixed model framework. The model with k factors, denoted FAk, is nested within the model with k + 1 factors. Models including the main genotype effect (g) are intermediate between the factor analytic models of order k (FAk) and of order FAk+1. Model FA1 + *g* is intermediate to models FA1 and FA2. Residual maximum likelihood ratio tests (REMLRT) can be used to compare such models. Other approaches for testing the goodness-of-fit of factor analytic models involve comparisons with the unstructured covariance matrix (Mardia et al., 1988), which is very hard to obtain with a great number of environments. All models were fitted using the ASREML software (Gilmour and Thompson, 1998, Gilmour et al., 2002) which uses the REML procedure through the average information algorithm (Gilmour et al., 1995; Johnson and Thompson, 1995; Thompson et al., 2003).

2.5 Application

Two large unbalanced data sets were used. The first one concerned 200 eucalypt treatments (progenies) evaluated for trunk circumference on six sites in lattice designs with different replication numbers in each trial. The total number of plants evaluated was 65000. The second data set concerned 60 tea plant treatments (progenies) evaluated in complete block designs for leaf weight in three consecutive years and in two trials. Trial 1 provided 5400 observations (60 treatments x 5 replications x 6 plants per plot x 3 annual measures) and trial 2 provided 4050 observations (45 treatments, 5 replications, 6 plants per plot and 3 annual measures). The 45 treatments in trial 2 are also in trial 1.

3 Results and discussion

3.1 Eucalypt data set

Results concerning several models applied to the eucalypt data set on six environments are presented in Table 1.

The first part of Table 1 contains only models (1 to 6) fitted with assumption of homogeneous error variance. Model 1 fitted treatment effects on each environment and considered a common error variance for all environments. Model 2 fitted treatment effects on an average environment and considered a common error variance for all environments. Model 3 fitted treatment effects on an average environment plus *g* x *e* interaction and considered a common error variance for all environments. Model 4 fitted a factor analytic structure of order 1 for treatment effects and considered a common error variance for all environments. Model 5 fitted a factor analytic structure of order 2 for treatment effects and considered a common error variance for all environments. Model 6 fitted a full multivariate unstructured for treatment effects and considered a common error variance for all environments. The second part of the same table contains only models (7 to 10) with assumption of heterogeneous error variance. Models 7 and 9 fitted a factor analytic structure of order 1 and 2, respectively, for treatment effects. Model 8 fitted a factor analytic structure of order 1 for treatment effects plus treatment main effects. Model 10 fitted a full multivariate unstructured for treatment effects.

Table 1 - Residual log-likelihoods (Log L) and likelihood ratio statistic (LRT) for the sequence models fitted to the eucalypt data

Model for *G*	Log L	LRT in relation to the previous model	Number of Variance parameters in *G*	Total number of variance parameters
1.Uniform for *g* in *e*	-151100	-	1	3
2.Uniform for *g*	-149228	-	1	3
3.Uniform for *g* + *g* x *e*	-147892	2672	2	4
4.FA1, var. homog.	-147619	546	12	14
5.FA2, var. homog.	-147562	114	17	19
6.Multiv.var. homog.	-147556	12	21	23
7.FA1, var. heterog.	-146381	-	12	19
8.FA1+g, var.heterog.	-146381	0	13	20
9.FA2, var. heterog.	-146325	112	17	24
10.Multiv. var. heterog.	-146318	14	21	28

Contrasting the two parts in terms of Log L, it can be seen that the models allowing error variance heterogeneity are far better than the models assuming variance homogeneity. This shows the superiority of FAMM models over AMMI models, which do not consider the error variance heterogeneity. Common error variance for all trials is implicit in the AMMI approach. Even the full multivariate model (6) for G_0 (21 parameters) with homogeneous variance is worse than the FA1 model (7) for G_0 (12 parameters) with heterogeneous variance. This confirms the great importance of considering error variance heterogeneity in MET analysis. And this can only be done in the mixed modelling framework. So, the factor analytic models being embedded in this framework, is a great advantage.

Another important feature of the FAMM models is the provision of parsimonious models in relation to the full unconstrained multivariate approach. The multivariate approach is prohibitive with a large (usually > 5) number of environments, generating over-parameterised and hard-to-converge models. Results from Table 1 reveal that model FAMM with two factors (FA2) is close (REMLRT of 14 and 12 on 4 degrees of freedom) to the full multivariate model in both situations, with and without allowing for variance heterogeneity. So, in practice, a model with four less parameters can be used. It is worth mentioning that all the FAMM models converged without a need for constraining the G_0 matrix.

Model including the main genotype effect (g) is intermediate between the factor analytic models of order k (FAk) and of order FAk+1, as it is FAk+1 with constraints. Model FA1 + *g* is intermediate to models FA1 and FA2. In the present data set models FA1 and FA1 + *g* were equivalent, giving the same Log L. In fact, the estimate of the variance component for genotype effects was on the boundary; that is, it was estimated as zero. The role of genotype main effects in an FA model is purely in terms of the search for a parsimonious variance structure between a given FAk model and a FAk+1 model. The approach for prediction of overall genotype means across environments is the same irrespective of the inclusion of genotype main effects (Smith et al., 2001). In a factor analytic context, the model without genotype main effects is equivalent to a model for genotype effects in each environment.

Overall, the best parsimonious model was FA2 with heterogeneous variance for errors (model 9 in Table 1). Results concerning loadings, common, specific and error variances provided by this model are presented in Table 2.

It can be seen that the FA2 model explained a large amount (almost 90%) of the total genotypic variance. The first factor explained 77.3% of the variation and the second factor added 11.9%. The specific variances (in percentage of the total) were low, except for environments 2 and 6, which were 22% and 16%, respectively. The high values of the common variance (or communality) show that the two factors explained a great percentage of the variance of each environment and that the FA2 model fitted well to the data set (Table 2).

Table 2 - Estimated loadings (on the correlation scale), common (communality), specific and error variances for model FA2 fitted to the eucalypt data

	Original Loadings and (Rotated)		Common	Specific.	Error
Location	Factor 1	Factor 2	Variance	Variance	Variance
1. L1	0.845 (0.433)	0.498 (0.880)	0.962	0.038	20.0422
2. L2	0.791 (0.443)	0.398 (0.767)	0.784	0.216	20.5270
3. L3	0.837 (0.450)	0.454 (0.839)	0.907	0.093	22.6041
4. L4	0.907 (0.596)	0.295 (0.745)	0.910	0.090	44.5751
5. L5	0.979 (0.761)	0.104 (0.624)	0.969	0.031	38.0380
6. L6	0.904 (0.837)	-0.149 (0.372)	0.839	0.161	28.9856
Eigenvalues	4.639	0.710			
Accu. Var. Explained	0.773	0.892			

The genotypic variance-covariance matrix and the correlations (obtained by $\Lambda\Lambda' + \Psi$ from model FA2 on the correlation scale) involving the several environments are presented in Table 3.

Table 3 - Estimated genotypic covariance (below the diagonal), variance (diagonal) and correlation (above the diagonal) matrix associated to model FA2 applied to eucalypt data set

	L1	L2	L3	L4	L5	L6
L1	**6.312**	0.867	0.933	0.914	0.879	0.689
L2	6.964	**10.225**	0.843	0.835	0.812	0.655
L3	7.375	8.481	**9.905**	0.893	0.867	0.689
L4	8.132	9.463	9.959	**12.555**	0.919	0.776
L5	6.566	7.754	8.108	9.682	**8.837**	0.869
L6	5.135	6.207	6.425	8.148	7.659	**8.784**

It can be observed that there is heterogeneity among the specific variances concerning several environments (diagonal of Table 3). This justifies the use of models with heterogeneous specific variances. Piepho (1997, 1998) proposed the use of a factor analytic model with common specific variance for all sites. However, Smith et al. (2001) noted that models with heterogeneous specific variances were significantly better. It can be seen that there is also heterogeneity of covariance between the several combinations of environments. These covariances represent the genotypic variance free from interaction effects between every two sites. This heterogeneity explains the better fit of FAk and multivariate models over model 3, which includes $g + g \times e$. When there are only two environments, the bivariate and model 3 tend to give the same fitting (see results from the tea plant data set).

Correlations results reveal that the first four environments have smaller correlations with the environment 6, which has higher correlations with environment

5 (Table 3). It can be observed that factor analysis put greater emphasis on environments 5 and 6 in factor 1 (rotated loadings higher than 0.76) and higher emphasis on sites 1, 2, 3 and 4 in factor 2 (rotated loadings higher than 0.74) (Table 2). This is the logic of factor analysis: to separate groups of traits with high correlations between them in each group and then put higher weights in traits of a group in one factor (factor 1) and higher weights in traits of another group in another factor (factor 2). Plotting the first set of loadings against the second will show the clustering of environments: L1, L2, L3 and L4 close together in one group and L5 and L6 in a second group. Another advantage of FAMM models over AMMI is that they provide an estimate of the full correlation structure, facilitating practical decisions to be made.

FAMM and AMMI models are also useful for the clustering of environments based on their similarity in terms of genetic correlations. This can be done through biplots (AMMI) or plot of loadings from the first factor against the loadings from the second factor (FAMM). The full structure of correlation provided by FAMM models can be also subjected to methods of cluster analysis or other multivariate methods. Such methods traditionally operate on correlations estimated by pairs of environments through balanced ANOVA. FAMM models use the information on all environments simultaneously to give the correlation for pairs of environments, thus providing more precise estimates.

3.2 Tea plant data set

Multi-environment spatial analysis for each trait

The two trials contained 45 treatments in common, so it was possible to analyse all data simultaneously. Although not all progenies were represented in the two environments, FAMM models were applied. An important remark is that the factor analysis under the mixed model can be done with incomplete data sets.

Firstly, multi-environment spatial analyses were made for each trait in a combination of the two trials. Three objectives pursued by breeders were considered: selection for specific environments (multivariate multi-environment spatial model), selection for an average environment (univariate multi-environment spatial model), selection for a non-tested environment (univariate multi-environment spatial models, including genotype x environment interaction effects).

Results concerning the first objective are presented in Table 4. The plot effect was not fitted because it was non-significant with spatial analysis.

The genetic correlations between environments were about 0.48, 0.57 and 0.57 for leaf yield in years 1, 2 and 3, respectively. The magnitudes of these correlations reveal a need for specific selection for each site. The bivariate model involving the two sites was fitted also assuming variance homogeneity across sites and independent errors. The deviance values obtained were –3756.5, 1127.04 and 6883.92, for the three traits, respectively. These are much higher than the -3966.24, 833.70 and 6393.74 obtained with the model allowing heterogeneity of variance and spatial errors. Such results reinforce that FAMM models could be more adequate than AMMI models, which do not allow for heterogeneity of variance and spatial errors. The residual auto-correlation coefficients were very high for the site 2 and spatial analysis could be abdicated for this site without efficiency loss.

Table 4 - Estimates of the variance parameters: genetic among treatments (progenies) in environment 1 ($\hat{\sigma}^2_{\tau_1}$) and in environment 2 ($\hat{\sigma}^2_{\tau_2}$), genetic covariance among treatments across sites ($\hat{\sigma}_{\tau_{12}}$), correlated residual in site 1 ($\hat{\sigma}^2_{\xi_1}$) and in site 2 ($\hat{\sigma}^2_{\xi_2}$), non-correlated residual in site 1 ($\hat{\sigma}^2_{\eta_1}$) and in site 2 ($\hat{\sigma}^2_{\eta_2}$), narrow sense heritability in site 1 ($\hat{h}^2_1$) and in site 2 ($\hat{h}^2_2$), respective adjusted heritabilities ($\hat{h}^2_{adj_1}$ and $\hat{h}^2_{adj_2}$) and residual auto-correlation coefficients between columns (AR Column *i*) and rows (AR Row *i*), in the specific trial or site *i*

Parameters estimates	First year	Second year	Third year
$\hat{\sigma}^2_{\tau_1}$	0.0157 ± 0.004	0.1074 ±0.02	0.3573 ±0.08
$\hat{\sigma}^2_{\tau_2}$	0.0214 ± 0.005	0.0978 ±0.02	1.1526±0.27
$\hat{\sigma}_{\tau_{12}}$	0.0087± 0.003	0.0585±0.02	0.3669 ±0.12
$\hat{\sigma}^2_{\xi_1}$	0.0296± 0.006	0.1439±0.03	0.9032±0.18
$\hat{\sigma}^2_{\xi_2}$	0.0183 ± 0.018	0.1286±0.04	1.9108 ±0.62
$\hat{\sigma}^2_{\eta_1}$	0.0948±0.004	0.4326±0.02	1.7135±0.07
$\hat{\sigma}^2_{\eta_2}$	0.0797±0.003	0.3531±0.02	3.2352±0.14
$\hat{h}^2_1$	0.4492	0.6283	0.4806
$\hat{h}^2_2$	0.7163	0.7017	0.7261
AR Column 1	0.8073±0.05	0.8463±0.04	0.8875±0.03
AR Row 1	0.8000±0.05	0.7967±0.05	0.8137±0.05
AR Column 2	0.9816±0.03	0.9192±0.04	0.9603±0.02
AR Row 2	0.9960±0.01	0.9482±0.02	0.9100±0.03
Deviance	-3966.24	829.13	6393.20
$\hat{h}^2_{adj_1} = (4\hat{\sigma}^2_{g_1})/(\hat{\sigma}^2_{g_1} + \hat{\sigma}^2_{\eta_1})$	0.5683	0.7956	0.6902
$\hat{h}^2_{adj_2} = (4\hat{\sigma}^2_{g_2})/(\hat{\sigma}^2_{g_2} + \hat{\sigma}^2_{\eta_2})$	0.8466	0.8676	1.04

Results concerning the second objective are presented in Table 5.

This model (Table 5), albeit more parsimonious than the full multivariate (Table 4), gave a significant higher deviance and higher AIC value. So, the multivariate is preferred and selection for an average environment can be made by taking means of predicted genetic values in each environment. The superiority of the multivariate model can be explained by the heterogeneity of genetic variance across sites (Table 4). Data standardisation should correct this and make the univariate (for an average environment) model suitable.

Table 5 - Estimates of the variance parameters: genetic among treatments (progenies) in an average environment ($\hat{\sigma}^2_{\tau}$), correlated residual in site 1 ($\hat{\sigma}^2_{\xi_1}$) and in site 2 ($\hat{\sigma}^2_{\xi_2}$), non-correlated residual in site 1 ($\hat{\sigma}^2_{\eta_1}$) and in site 2 ($\hat{\sigma}^2_{\eta_2}$) and respective residual auto-correlation coefficients between columns (AR Column *i*) and rows (AR Row *i*), in the specific trial or site *i*

Parameters estimates	First year	Second year	Third year
$\hat{\sigma}^2_{\tau}$	0.01397±0.003	0.0797±0.02	0.4310 ±0.09
$\hat{\sigma}^2_{\xi_1}$	0.0338± 0.007	0.1606±0.03	0.9470±0.18
$\hat{\sigma}^2_{\xi_2}$	0.0182 ± 0.005	0.1350±0.04	1.9259±0.62
$\hat{\sigma}^2_{\eta_1}$	0.09757±0.004	0.4423±0.02	1.7335±0.07
$\hat{\sigma}^2_{\eta_2}$	0.07531±0.004	0.3580±0.02	3.4773±0.15
AR Column 1	0.8049±0.05	0.8154±0.05	0.8766±0.03
AR Row 1	0.8365±0.05	0.8094±0.05	0.8169±0.05
AR Column 2	0.8893±0.05	0.9000±0.04	0.9487±0.02
AR Row 2	0.7336±0.08	0.9290±0.03	0.9103±0.03
Deviance	-3917.40	883.21	6483.50

Results concerning the third objective are presented in Table 6.

By comparing results from Tables 5 and 6, it can be seen by the deviance values that the model with interaction (Table 6) fits to the data better, revealing the significance of the *g* x *e* interaction effects.

This model gave approximately the same deviance and smaller AIC values in relation to the full multivariate (Table 4). Then it should be preferred. The *g* x *e* component encompassed all the heterogeneity of genetic variance. From this model, predicted genetic values can be derived for each treatment (parent or individual) in each environment by summing the correspondent *g* and *g* x *e* predicted effects. Then, the mean of predicted genetic values of each treatment over several environments can be taken aiming at the selection of an average environment.

Another alternative is to obtain treatment effects in each environment directly by fitting only the *g* x *e* component, i.e., overlooking *g* main effects. Applying this approach for the measure in the first year, the variance component for *g* x *e* obtained was 0.01858, which is approximately equivalent to the sum of the variance component for *g* and *g* x *e* presented in Table 6, as expected. The deviance obtained was –3957.20, which is significantly (by LRT) higher then the –3965.28 reported in Table 6. This shows that the model with *g* is better.

Table 6 - Estimates of the variance parameters: genetic among treatments (progenies) free of *g* x *e* interaction effects ($\hat{\sigma}^2_{\tau}$), *g* x *e* interaction effects ($\hat{\sigma}^2_{ge}$), correlated residual in site 1 ($\hat{\sigma}^2_{\xi_1}$) and in site 2 ($\hat{\sigma}^2_{\xi_2}$), non-correlated residual in site 1 ($\hat{\sigma}^2_{\eta_1}$) and in site 2 ($\hat{\sigma}^2_{\eta_2}$) and respective residual auto-correlation coefficients between columns (AR Column *i*) and rows (AR Row *i*), in the specific trial or site *i*

Parameters estimates	First year	Second year	Third year
$\hat{\sigma}^2_{\tau}$	0.00865±0.003	0.0588±0.02	0.3305±0.12
$\hat{\sigma}^2_{ge}$	0.00976±0.003	0.0442±0.01	0.3412±0.09
$\hat{\sigma}^2_{\xi_1}$	0.0298± 0.007	0.1437±0.03	0.9047±0.18
$\hat{\sigma}^2_{\xi_2}$	0.0183 ± 0.02	0.1286±0.04	1.8799±0.64
$\hat{\sigma}^2_{\eta_1}$	0.09469±0.004	0.4327±0.02	1.7111±0.07
$\hat{\sigma}^2_{\eta_2}$	0.07979±0.003	0.3505±0.01	3.2502±0.14
AR Column 1	0.8078±0.05	0.8466±0.04	0.8888±0.03
AR Row 1	0.8004±0.06	0.7968±0.05	0.8144±0.05
AR Column 2	0.9817±0.03	0.9187±0.04	0.9591±0.02
AR Row 2	0.9959±0.09	0.9485±0.02	0.9161±0.04
Deviance	-3965.28	829.22	6409.64

Factor analytic models (spatial and non-spatial) for multivariate and multi-environment data

Although the univariate model with *g* and *g* x *e* for treatment effects is sufficient for the multi-site analysis of individual traits, the univariate approach is not appropriate for all six measures together due to the great variance heterogeneity between measures in each site. So, a multivariate approach for the six traits together with fit of individual permanent effects in each site was adopted. The fit of permanent effects aimed at the elimination of the residual covariance between measures in each site. The model is an extension (increasing the number of traits to six and including permanent effects) of that concerning selection for specific environments.

However, the fit of this model not converged with spatial errors and a non-spatial model was fitted. Results are presented in the sequence together with the factor analytic models, which were fitted as alternative parsimonious models.

Results concerning factor analytic models for the six repeated measures in two environments in comparison with the multivariate model are presented in Table 7. In all models the individual permanent effects were fitted as a mean of eliminating the residual correlation between repeated measures in each site.

Table 7 - Log-REML (Log L) and REMLRT (LRT) for comparing models of fitting covariances structures involving six traits. Models fitted were multivariate for treatments and non-spatial for residuals (MNS), factor analytic of order 1 for treatments and non spatial for residuals (FA1NS), factor analytic of order 1 for treatments and spatial (including both the correlated and independent term) for residuals (FA1S)

	Number of Variance parameters				
Model for *G*	*G*	Total	LogL	LRT(P value)	%Variance
MNS	21	28	-2335.67	-	
FA1NS	12	19	-1848.10	975.14(0.001)	
FA1S	12	37	-585.31	2525.58(0.001)	71.5

It can be seen that the best model was the factor analytic with spatial error (FA1S). This model was superior to that with non-spatial error (FA1NS). This fact is sufficient to show the superiority of factor analytic multiplicative mixed models (FAMM) over the additive main and multiplicative interaction effects (AMMI), which assumes fixed treatment effects and do not permit to model separate spatial errors. The proportion of genetic variance explained by FA1S was 71.5%. This value is sufficient for the purpose of the analysis, i.e., genetic selection.

The non-spatial factor analytic model showed to be superior to the non-spatial multivariate model (MNS), revealing the advantages of the factor analytic models in terms of parsimony and ability of fitting. The MNS model, although with more parameters, showed a smaller Log L and was hard to converge demanding restriction on G to be positively definite. Even so, the convergence was not so reliable, as ASREML fixed some variance components on the boundaries. In fact, it might have not converged to a maximum likelihood solution. Other models like the full multivariate model with spatial error and factor analytic of order 2 did not converge.

Results concerning genetic correlation for the best model (FA1S) are presented in Table 8.

Table 8 - Estimated genetic correlations obtained from the FA1S modelling

Trait	1	2	3	4	5	6
1	1	0.982	0.999	0.665	0.852	0.745
2		1	0.982	0.585	0.794	0.653
3			1	0.664	0.851	0.744
4				1	0.870	0.862
5					1	0.935
6						1

The estimated correlations are relatively coherent with previous estimates and expectation: higher correlation between repeated measures within sites and lower correlations across sites. This, together with the suitable proportion of genetic variance explained by the FA1S model reveals the adequacy of the factor analytic model for analysis of this sort of data. Otherwise, the whole data set could not be analysed simultaneously. The variograms showed adequate behavior.

Gilmour and Thompson (2002) reported the computational aspects of analysing six traits in an animal breeding context, when some traits are highly correlated. They conclude that the Factor Analytic and Cholesky models appear best in this situation. We confirm the adequacy of FA models. The Cholesky appear to be inadequate for

our data set with errors non-correlated across traits, as we fit the permanent effect to account the correlation across traits within sites and the errors are non-correlated across sites.

Practical experiments with several perennial plants annually generate, a large amount of data on repeated measures throughout the world. These measures are usually taken only three or four times before selection, since more than that, leads to less genetic gain per unit of time (Resende, 2002). Suitable models should be found for application in such type of data in one or several experiments simultaneously. For analysing multi-environment data sets with longitudinal data, the factor analytic multiplicative mixed model proved to be a very useful tool, mainly when applied together with spatial analysis. Software ASReml showed to be essential for modeling the complex data structure involving repeated measures, spatial dependency and multi-environment data sets in perennial plants. FAMM and FAMMS models can also be used for studies concerning QTL (quantitative trait loci) x environment interaction. This approach can be better than that advocated by Romagosa et al. (1996), based on AMMI analysis.

Conclusions

- Parsimonious FAMM models were selected for the two data sets: FA2 for an eucalyptus data set and FA1 for a tea plant data set.
- There were great advantages of heterogeneous variance FAMM models over homogeneous variance FAMM models. This reveals the superiority of FAMM models over AMMI models for the data sets considered in this study.
- Heterogeneity was noted among the specific variances in individual environments, so factor analytic models with common specific variances for all sites were not suitable.
- FAMM models provided estimates of the full correlation structure, facilitating practical decisions to be made.
- FAMM models with heterogeneous variance among traits and spatial errors within traits were advantageous over FAMM models with variance homogeneity and non-spatial error. This also shows the superiority of FAMM models over AMMI models, which do not allow for dependent or spatial errors.
- For analysing multi-environment data sets with longitudinal data, FAMM models proved to be a very useful tool, mainly when applied together with spatial analysis.

RESENDE, M. D. V. de; THOMPSON, R. Modelos mistos multiplicativos "fator-analítico" na análise de múltiplos experimentos. *Rev. Mat. Est.*, São Paulo, v.22, n.2, p.31-52, 2004.

▪ *RESUMO: A análise de grupos de experimentos ou de experimentos conduzidos em múltiplos ambientes (MET) tem sido tradicionalmente baseada em modelos simples, os quais assumem homogeneidade de variância residual entre os ensaios, independência de erros dentro de ensaio, efeitos da interação genótipo x ambiente (g x e) como um grupo de efeitos aleatórios independentes. A análise de dados de grupos de experimentos por meio de modelos realísticos é um problema estatístico complexo que demanda extensões ao modelo linear misto padrão. A suposição referente a independência dos efeitos de g x e pode ser eliminada através do uso de modelos multiplicativos. Tais modelos foram popularizados como modelos aditivos para os efeitos principais e multiplicativos para os efeitos da*

interação (AMMI) e tiveram grande aplicação. Entretanto, a análise AMMI apresenta pelo menos cinco grandes limitações: considera os efeitos de genótipo e de g x e como fixos; é adequado apenas para dados balanceados; não considera a variação espacial dentro de ensaios; não considera a heterogeneidade de variância entre ensaios; não considera o diferente número de repetições através dos ensaios. Estas características não são realísticas na análise de experimentos de campo. No contexto dos modelos mistos, Piepho (1998) apresentou um modelo misto multiplicativo fator - analítico (FAMM) com efeitos aleatórios de genótipo e de g x e, o qual é conceitualmente e funcionalmente melhor que o AMMI. No mesmo contexto, Smith et al. (2001) apresentou uma classe geral de modelos FAMM que abrange a abordagem de Piepho (1998) e inclui erros espaciais para cada ensaio (FAMMS). Esta classe geral de modelos propicia uma abordagem realística completa para análise de dados de múltiplos experimentos. Este trabalho lida com a aplicação dos modelos FAMM e FAMMS em dois grandes conjuntos de dados desbalanceados (de eucalipto e de erva-mate) visando enfatizar suas vantagens sobre os modelos AMMI em termos das suposições de homogeneidade de variâncias entre ensaios e independência de erros dentro de ensaios. Adicionalmente, enfatiza-se a capacidade dos modelos FAMM em propiciar modelos parcimoniosos. Modelos parcimoniosos foram selecionados para os dois conjuntos de dados. Foram constatadas grandes vantagens dos modelos FAMM com variâncias heterogêneas sobre modelos FAMM com variâncias homogêneas. Isto revela a superioridade dos modelos FAMM sobre os modelos AMMI. Grande heterogeneidade entre variâncias específicas entre ambientes individuais foi observada. Assim, modelos fator - analíticos com variâncias específicas comuns a todos os ensaios não foram adequados. Os modelos FAMM propiciaram estimativas da completa estrutura de correlação, facilitando a tomada de decisões práticas. Modelos FAMM com heterogeneidade de variâncias entre caracteres e erros espaciais dentro de caracteres mostraram-se vantajosos sobre modelos FAMM com homogeneidade de variância e erros não espaciais. Isto revela a superioridade de modelos FAMM sobre modelos AMMI, os quais não permitem o ajuste de erros com dependência espacial. Para a análise de múltiplos experimentos com dados longitudinais, os modelos FAMMS mostraram ser uma ferramenta muito útil.

- *PALAVRAS-CHAVE: Modelos mistos multiplicativos fator-analíticos; modelos mistos multiplicativos fator - analíticos espaciais; modelos AMMI, máxima verossimilhança restrita; melhor predição linear não viciada;, experimentos em múltiplos ambientes; análise de estabilidade.*

4 References

CROSSA, J. Statistical analysis of multi-location trials. *Adv. Agron.*, San Diego, v.44, p.55-85, 1990.

CROSSA, J.; GAUCH, H. G.; ZOBEL, R. W. Additive main effects and multiplicative interaction analysis of two international maize cultivars trials. *Crop Sci.*, Madison, v.30, n.3, p.493-500, 1990.

CULLIS, B. R.; GOGELL, B.; VERBYLA, A.; THOMPSON, R. Spatial analysis of multi-environment early generation variety trials. *Biometrics*, Washington, v.54, p.1-18, 1998.

CULLIS, B. R.; GLEESON, A. C. Spatial analysis of field experiments-an extension at two dimensions. *Biometrics*, Washington, v.47, p.1449-60, 1991.

CRUZ, C. D.; TORRES, R. A. A.; VENCOVSKY, R. An alternative to the stability analysis proposed by Silva an Barreto. *Braz. J. Genet.*, Ribeirão Preto, v.12, p.567-80, 1989.

DUARTE, J. B.; VENCOVSKY, R. *Interacao genotipos x ambientes: uma introducao a analise AMMI*. Ribeirao Preto: Sociedade Brasileira de Genetica, 1999. 60p. (Serie Monografias, n.9)

EBERHART, S. A.; RUSSELL, W. A. Stability parameters for comparing varieties. *Crop Sci.*, Madison, v.6, p.36-40, 1966.

EEUWIJK, F. A. van; KEIZER, L. C. P.; BAKKER, J. J. Linear and bilinear models for the analysis of multi-environment trials. II. An application to data from Dutch Maize Variety Trials. *Euphytica*, Dordrecht, v.84, p.9-22, 1995.

FINLAY, K. W.; WILKINSON, G. N. The analysis of adaptation in a plant breeding programme. *Aust. J. Agric. Res.*, Collingwood, v.14, p.742-54, 1963.

FISHER, R. A. ; MACKENZIE, W. A. Studies in crop variation. II. The manurial response of different potato varieties. *J. Agric. Sci.*, Cambridge, v.13, p.311-20, 1923.

GAUCH, H. G. Model selection and validation for yield trials with interaction. *Biometrics*, Washington, v.44, p.705-15, 1988.

GAUCH, H. G. *Statistical analysis of regional yield trials*: AMMI analysis of factorial designs. Amsterdam: Elsevier, 1992. 172p.

GILMOUR, A. R.; CULLIS, B. R.; WELHAM, S. J.; THOMPSON, R. *ASReml reference manual.* Release 1.0. 2 ed. Harpenden: Biomathematics and Statistics Department - Rothamsted Research, 2002. 187p.

GILMOUR, A. R.; THOMPSON, R. Estimating parameters of a singular variance matrix in ASREML. In: *AUSTRALASIAN GENSTAT CONFERENCE*, 2002, Busselton. *Proceedings*... Busselton: Atlas Conferences, 2002.

GILMOUR, A. R.; THOMPSON, R. Modelling variance parameters in ASREML for repeated measures. In: WORLD CONGRESS ON GENETIC APPLIED TO LIVESTOCK PRODUCTION, 6., 1998, Armidale. *Proceedings*... Armidale: AGBU / University of New England, 1998. v.27, p.453-54.

GILMOUR, A. R.; THOMPSON, R.; CULLIS, B. R. Average information REML: an efficient algorithm for parameter estimation in linear mixed models. *Biometrics*, Washington, v.51, p.1440-50, 1995.

GOLLOB, H. F. A statistical model which combines features of factor analytic and analysis of variance technique. *Psychometrika*, Baltimore, v.33, n.1, p.73-115, 1968.

HENDERSON, C. R. *Aplications of linear models in animal breeding*. Guelph: University of Guelph, 1984. 462p.

HILL, R. R.; ROSENBERGER, J. L. Methods for combining data from germplasm evaluation trials. *Crop Sci.*, Madison, v.25, p.467-70, 1985.

JENNRICH, R. L.; SCHLUCHTER, M. D. Unbalanced repeated measures models with structured covariance matrices. *Biometrics*, Washington, v.42, p.805-20, 1986.

JOHNSON, D. L.; THOMPSON, R. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *J. Dairy Sci.*, Savoy, v.78, p.449-56, 1995.

JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis*. Englewood : Prentice Hall, 1988. 594 p.

LAWLEY, D. N.; MAXWELL, A. E. *Factor analysis as a statistical tool*. 2nd. ed. London: Butterworths, 1971. 448p.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. *Multivariate analysis*. London: Acadenic Press, 1988. 521p.

MARTIN, R. J. The use of time-series models and methods in the analysis of agricultural field trials. *Commun. Stat. Theory Methods*, New York, v.19, n.1, p.55-81, 1990.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. *Biometrika*, London, v.58, p.545-54, 1971.

PIEPHO, H. P. Best linear unbiased prediction (BLUP) for regional yield trials: a comparison to additive main effects and multiplicative interaction (AMMI) analysis. *Theor. Appl. Genet.*, Berlin, v.89, p.647-54, 1994.

PIEPHO, H. P. Analysing genotype-environment interaction data by mixed models with multiplicative terms. *Biometrics*, Washington, v.53, p.761-7, 1997.

PIEPHO, H. P. Empirical best linear unbiased prediction in cultivar trials using factor analytic variance-covariance structures. *Theor. Appl. Genet.*, Berlin, v.97, p.195-201, 1998.

RESENDE, M. D. V. *Genética biométrica e estatística no melhoramento de plantas perenes*. Brasilia: Embrapa Informação Tecnológica, 2002. 975p.

ROMAGOSA, I.; ULLRICH, S. E.; HAN, F.; HAYES, P. M. Use of additive main effects and multiplicative interaction model in QTL mapping for adaptation in barley. *Theor. Appl. Genet.*, Berlin, v.93, p.30-7, 1996.

SMITH, A.; CULLIS, B. R.; THOMPSON, R. Analysing variety by environment data using multiplicative mixed models and adjustment for spatial field trend. *Biometrics*, Washington, v.57, p.1138-47, 2001.

STROUP, W. W.; MULITZE, D. K. Nearest neighbour adjusted best linear unbiased prediction. *Am. Stat.*, Washington, v.45, p.194-200, 1991.

THOMPSON, R.; WELHAM, S. J. REML analysis of mixed models. In: Payne, R. (Ed.). *GenStat 6 Release 6.1. The guide to GenStat, v.2 – Statistics*. Harpenden: Rothamsted Research, 2003. p.469-560.

THOMPSON, R.; CULLIS, B. R.; SMITH, A. B.; GILMOUR, A. R. A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. *Aust. N. Z. J. of Stat.*, Carlton South, v.45, n.4, p.445-59, 2003.

TOLER, J. E.; BURROWS, P. M. Genotype performance over environmental arrays: a non-linear grouping protocol. *Journal of Applied Statistics*, London, v.25, n.1, p.131-43, 1998.

Recebido em 08.10.2003.

Aprovado após revisão em 01.07.2004.