*Heredity* **79** (1997) 162–171 *Received 5 August 1996*

# Modelling expectation and variance for genotype by environment data

JEAN-BAPTISTE DENIS*, HANS-PETER PIEPHO† & FRED A. VAN EEUWIJK‡

*Laboratoire de Biométrie, INRA, Route de Saint-Cyr, F-78026 Versailles, France, †Faculty of Agricultural and Environmental Sciences, University of Kassel, D-37213 Witzenhausen, Germany and ‡DLO-Center for Plant Breeding and Reproduction Research, CPRO-DLO, PO Box 16, NL-6700AA Wageningen, The Netherlands*

An integration of two types of models for the analysis of genotype by environment interaction is presented. On the one hand, the expectation of G × E interaction is frequently modelled by regression models; on the other hand, for deviations from these regressions, either separate stability parameters are defined or extra components of variance are introduced. A class of mixed models is described that contains facilities for modelling expectation by regression and, in addition, has extensive possibilities for dealing with heteroscedasticity. Practical aspects of the use of these mixed models are illustrated on a data set involving sugar yield in beet.

**Keywords:** covariate, factorial regression, genotype × environment interaction, heteroscedasticity, interaction, mixed model.

## Introduction

This paper presents a number of models that can account for interaction and heteroscedasticity in genotype by environment tables. These models can be viewed as generalizations of both the classical model by Shukla (1972) and the mixed factorial regression model by Denis & Dhorne (1989). The models can be used for the analysis of replicated and unreplicated tables alike, as no estimate for error is required. Modelling heteroscedasticity is especially relevant for genotype by environment interaction (Kang & Gorman, 1989; Kang, 1993), but similar models may be used to analyse, for example, repeated measures data accruing in sociological and psychological research (Crowder & Hand, 1990; Longford, 1993).

For selecting genotypes, a plant breeder uses assessments of the phenotypic value under different environmental conditions. These assessments are collected in genotype by environment tables. Inferences follow from adequate statistical models for these tables, and decisions are made regarding the selection and rejection of varieties. We will consider environments to be either locations or years, i.e. there is no factorial structure in the environments. Of course, in some cases, the environments comprise location by year combinations, and it may be worthwhile exploiting this factorial structure (Piepho, 1994a). In this paper, we will take genotypes as fixed and environments as random. A partial justification for this choice is that we are studying a given set of genotypes and are not interested in testing the environments themselves; they are considered only to provide information about the genotypes.

Later some classical models will be described, after which their common structural features will be discussed, leading to the delineation of a coherent family of models for the analysis of genotype by environment data; some of its more interesting members are presented. To illustrate the practical aspects of interpreting model parameters, a set of sugar beet data is analysed. GENSTAT and SAS source codes for running some of the presented models are given in the Appendix.

*Correspondence. E-mail: denis@versailles.inra.fr

## Review of current models

### *Additive model*

The additive two-way mixed model provides a baseline against which other more elaborate models can be compared. Let $Y_{ij}$ be a typical entry for a genotype by environment table, where $i \in \{1 \ldots I\}$ corresponds to the $i$th genotype and $j \in \{1 \ldots J\}$ corresponds to the $j$th environment. $Y_{ij}$ is taken as the sum of a (fixed) parameter depending on the genotype ($\alpha_i$), a random parameter depending on the environment ($B_j$) and an independent residual term ($E_{ij}$):

$$Y_{ij} = \alpha_i + B_j + E_{ij}.$$

This model has an obvious interpretation. Its first two moments are:

$$e(Y_{ij}) = \alpha_i;\ V(Y_{ij}) = \sigma_B + \gamma_E;$$

$Cov(Y_{ij}, Y_{i'j'}) = \sigma_B$ for $j = j'$, 0 otherwise.

The similarity in performance of different genotypes grown in the same environment is represented by a constant positive correlation, identical for every pair of genotypes:

$$Cor(Y_{ij}, Y_{i'j}) = \frac{\sigma_B}{\sigma_B + \gamma_E}.$$

Between performances in different environments, this correlation is zero and this basic assumption will be true for all models presented in this paper. Thus, a convenient notation is to introduce $\mathbf{Y}_j$, the vector of the $I$ performances of the genotypes in the $j$th environment. Covariances between different $\mathbf{Y}_j$ are null and models can be defined by their expectations and variances. For the additive model, it turns out that

$$\mathcal{E}(\mathbf{Y}_j) = \boldsymbol{\alpha};\ \mathcal{V}(\mathbf{Y}_j) = \sigma_B \mathbf{J} + \gamma_E \mathbf{I}, \quad (1)$$

where $\mathbf{J}$ is the $I \times I$ matrix with all components equal to 1, $\mathbf{I}$ is the identity matrix of size $I$ and $\boldsymbol{\alpha}$ is the vector of $\alpha_i$.

### *General heteroscedastic model*

The additive model may be extended by attributing a different variance to each genotype. The model formulation is identical, but the variance structure is now different; each genotype is considered to have its own variance, $\gamma_i$. Shukla (1982) suggested the term *stability variance* for $\gamma_i$. Earlier, Wricke (1962) had proposed the term *ecovalence* for the contribution of a genotype to the interaction sum of squares, and this quantity is directly related to $\gamma_i$. Expectation and variance structures are given by

$$\mathcal{E}(\mathbf{Y}_j) = \boldsymbol{\alpha};\ \mathcal{V}(\mathbf{Y}_j) = \sigma_B \mathbf{J} + \mathrm{dg}\,(\boldsymbol{\gamma}), \quad (2)$$

where dg $(\boldsymbol{\gamma})$ is the diagonal matrix whose terms are $\gamma_i$, the components of vector $\boldsymbol{\gamma}$. The interpretation is straightforward: the variance depends on the genotype and the correlation differs among pairs of genotypes:

$$Cor(Y_{ij}, Y_{i'j}) = \frac{\sigma_B}{\sqrt{(\sigma_B + \gamma_i)(\sigma_B + \gamma_{i'})}}.$$

The more variable a genotype is, the less correlated it will be with other genotypes. This model is much more flexible than the additive model (1), as the number of variance parameters increases from 2 to $1+I$.

The above type of model appears to have been used first by Grubbs (1948) for the analysis of measurement errors. Subsequently, it has been reconsidered by several authors, e.g. Russell & Bradley (1958) and Shukla (1972, 1982). Some extensions of Shukla's stability variance concept were given by Piepho (1994a,b,c, 1995). A recent review may be found in Piepho (1996a).

### *Scheffé model*

The mixed model proposed by Scheffé (1959, p. 266) provides a further generalization by allowing any covariance structure between performances from the same environment. As a consequence, the $B_j$ term (environment main effect) becomes redundant and the model may be written as:

$$Y_{ij} = \alpha_i + E_{ij}.$$

In contrast to Scheffé, we cannot include a residual term, as we are addressing the non-replicated case. The $E_{ij}$ components are correlated within environments:

$$\mathcal{E}(\mathbf{Y}_j) = \boldsymbol{\alpha};\ \mathcal{V}(\mathbf{Y}_j) = \boldsymbol{\Gamma}. \quad (3)$$

$\boldsymbol{\alpha}$ is a column vector of size $I$ and $\boldsymbol{\Gamma} = \{\gamma_{ii'}\}$ is any covariance matrix of size $I$. The model is very flexible, but at the cost of $I(I+1)/2$ variance components that need to be estimated. Many environments are required to obtain good estimates of the covariance parameters. The correlations within each environment may be negative, whereas model (2) constrains the correlations to be positive and of a defined structure. Model (3) has been extensively studied and used by Calinski *et al.* (1987a,b) for interpreting genotype–environment data. Piepho (1996b) considered the problem of genotypic mean comparisons under this general model.

### *Mixed factorial regression model*

Mixed factorial regression incorporates covariates associated with genotypes and covariates associated with environments. This type of model was described by Denis & Dhorne (1989) (see also Denis, 1994). It is an extension of the factorial regression approach developed earlier by Denis (1979, 1988), elaborating on initial work carried out by, among others, Hardwick & Wood (1972) and Wood (1976). An extensive review of such models can be found in van Eeuwijk *et al.* (1996).

The main feature is the introduction of regression terms, including covariates corresponding to the levels of one factor or both factors. Let us consider here one covariate for environments ($z_j$ for the $j$th environment) and one covariate for genotypes ($x_i$ for the $i$th genotype). The model can be written

$$Y_{ij} = \alpha_{i1} + \alpha_{i2} z_j + B_{j1} + x_i B_{j2} + E_{ij},$$

where $\alpha_{i2}$ and $B_{j2}$ are regression coefficients relating to genotypes and environments, respectively. The

term $\alpha_{i2}z_j$ then becomes fixed because $z_j$ values are known, whereas the term $x_iB_{j2}$, embodying the environmental regressions on a genotypic covariate, remains random. For $\alpha_{i1}$ and $B_{j1}$ to keep their usual main effect interpretation, the covariates must be centred. The first two moments for this model are:

$$\mathcal{E}(\mathbf{Y}_j) = \boldsymbol{\alpha}(1, z_j)'; \; \mathcal{V}(\mathbf{Y}_j) = (\mathbf{1}, \mathbf{x})\boldsymbol{\Sigma}(\mathbf{1}, \mathbf{x})' + \gamma_E \mathbf{I}, \qquad (4)$$

where $\boldsymbol{\alpha}$ is an $I \times 2$ matrix of fixed parameters, $\mathbf{1}$ is the $I$ vector of ones, $\mathbf{x}$ is the $I$ vector of genotypic covariate $\{\mathbf{x}_i\}$ and $\boldsymbol{\Sigma}$ is any covariance matrix of size 2. Although the residual term $E_{ij}$ is homoscedastic, the $Y_{ij}$ are heteroscedastic, depending on the genotypic covariate $\mathbf{x}$. As a consequence, the correlations between the $Y_{ij}$s in a particular environment can be positive or negative:

$$Cor(Y_{ij}, Y_{i'j}) = \frac{\sigma_{11} + (x_i + x_{i'})\sigma_{12} + x_i x_{i'}\sigma_{22}}{\sqrt{(\sigma_{11} + 2x_i\sigma_{12} + x_i^2\sigma_{22} + \gamma_E)(\sigma_{11} + 2x_{i'}\sigma_{12} + x_{i'}^2\sigma_{22} + \gamma_E)}}.$$

In Denis & Dhorne (1989), this model was developed for any number of environmental and genotypic covariates (see also the next section on general models).

### *Shukla's model*

A mixture between the completely heteroscedastic model (2) and the factorial regression model (4) was proposed by Shukla (1972). This model provided a main inspiration for this paper. Applications can be found in Kang & Gorman (1989) and Kang (1993). The assumptions for the random parameters in this model produce the following expectation and variance for the observed random variates:

$$\mathcal{E}(\mathbf{Y}_j) = \boldsymbol{\alpha}(1, z_j); \; \mathcal{V}(\mathbf{Y}_j) = \sigma_B \mathbf{J} + \mathrm{dg}(\gamma) \qquad (5)$$

The non-null correlations between the $Y_{ij}$s are identical to those of (2).

The information about the genotypes conveyed by Shukla's model is concentrated in triplets of parameters: a general level of performance ($\alpha_{i1}$), a measure of sensitivity to the environmental covariate ($\alpha_{i2}$) and a stability variance ($\gamma_i$).

## A general model

### *Description*

The models proposed in the previous sections can be expressed in a unified way, which in turn generates more useful models. Each model can be presented as the sum of three components: the fixed terms, the random terms and the residual term. Mathematically, the distinction between the last two terms is not always obvious. We already saw that for Scheffé's model (3) only one term remains; nevertheless, for interpretation and software application, it is convenient to make this distinction.

The fixed part is based on $H$ covariates in each environment. These are collected in a $J \times H$ matrix $\mathbf{z}$. The regression on these covariates involves $IH$ fixed terms $\alpha_{ih}$:

$$\sum_{h=1}^{H} \alpha_{ih} z_{jh}; \; \boldsymbol{\alpha}\mathbf{z}_j, \qquad (6)$$

where $\mathbf{z}_j$ is the $j$th row vector of matrix $\mathbf{z}$. The first covariate is usually the constant covariate ($z_{j1} = 1$ for every $j$), producing the main effect. It is also convenient to centre the other covariates ($\Sigma_j z_{jh} = 0$ for every $h > 1$) to obtain the standard separation of main effects and interaction terms (for complete tables).

The random part of the model consists of $J$ environmental regressions on $K$ genotypic covariates, the latter collected in the matrix $\mathbf{x}$ ($I$ by $K$). Thus, in total there are $JK$ random regression coefficients (parameters). We can express the random part in the following way:

$$\sum_{k=1}^{K} x_{ik} B_{jk}; \; \mathbf{x}\mathbf{B}_j$$

which produces, as variance component of $\mathbf{Y}_j$,

$$\mathbf{x}\boldsymbol{\Sigma}\mathbf{x}', \qquad (7)$$

where $\boldsymbol{\Sigma}$ is the variance matrix of vector $\mathbf{B}_j$ of size $K$. Again, the first covariate is usually the constant covariate ($x_{i1} = 1$ for every $i$), producing the main effect. Centring of the covariates, $\Sigma_i x_{ik} = 0$ for every $k > 1$, partitions the variation between main effects and interactions. The $B_{jk}$ are random variates whose variance–covariance matrix must be specified. In classical models, zero correlations are assumed to exist between the random coefficients in different environments. This may be justified by thinking of the environments as being randomly sampled from a large population of environments.

The residual part comprises not only the experimental error, but also the interaction not yet accounted for by the fixed and random terms already included. It appears reasonable to employ a simple model for the residual term when the covariates account for most of the heteroscedasticity (if any) in the data. If the covariates remove little heteroscedasticity or no covariates are available, it may be useful to choose a more flexible model for the residual term. In the previous models, three possibilities occurred. If $\mathbf{E}_j$ is the counterpart of $\mathbf{Y}_j$

for the residual term, let $\mathbf{\Gamma}$ be its variance matrix. We can distinguish three forms of $\mathbf{\Gamma}$:

$$\begin{cases} \textit{simple} & : \mathbf{\Gamma} = \gamma_E \mathbf{I}, \\ \textit{diagonal} & : \mathbf{\Gamma} = \mathrm{dg}(\gamma), \\ \textit{unstructured} & : \mathbf{\Gamma} = \{\gamma_{ii'}\}. \end{cases} \quad (8)$$

Finally, the general model is

$$\mathbf{Y}_j = \alpha \mathbf{z}_j + \mathbf{x}\mathbf{B}_j + \mathbf{E}_j$$

and the general forms of expectation and variance–covariance structures are, respectively,

$$\mathcal{E}(\mathbf{Y}_j) = \alpha \mathbf{z}_j;\ \mathcal{V}(\mathbf{Y}_j) = \mathbf{x}\mathbf{\Sigma}\mathbf{x}' + \mathbf{\Gamma}. \quad (9)$$

Table 1 gives the numbers of parameters for the three terms of all models presented in this paper. The important point is that all these models are, in fact, mixed linear models and that standard classical methods for estimation, testing and model selection can be applied. In the following, we propose some new models pertaining to the family we have just identified, combining their possibilities or adding similar ones.

### *Structured heteroscedastic model*

A difficulty with the completely heteroscedastic model (2) is the large number of variance components to be estimated; poor estimates may be the consequence. Sometimes the breeder is able to distinguish groups of genotypes with *a priori* different variabilities. This leads to a simplified version of the completely heteroscedastic model by assigning the same residual variance to all genotypes belonging to a group. Let $g(i) \in \{1 \ldots G\}$ be the numbering of these groups. $\gamma_i$ is supposed to be equal to $\gamma_{g(i)}$. The $g(i)$s represent discrete covariates associated to genotypes, and the interesting novelty is that they are applied to the residual random component, producing an intermediate model between (1) and (2). Correlations between genotypes in an environment are positive and depend only on the groups to which the genotypes belong. The number of variance components is $1+G$.

### *Correlated structured heteroscedastic model*

Another possibility for deriving a structured heteroscedastic model is to retain a classical homoscedastic residual and replace the main effect $B_j$ by group-specific effects $B_{g(i)j}$, corresponding to groups $g(i)$. This means that the random environment effect is different from one group to another. Variances depend on the group and covariances on the pair of groups involved. The model can be written as

$$Y_{ij} = \alpha_i + B_{g(i)j} + E_{ij},$$

implying

$$\mathcal{E}(\mathbf{Y}_j) = \alpha;\ \mathcal{V}(\mathbf{Y}_j) = \mathbf{x}\mathbf{\Sigma}\mathbf{x}' + \gamma_E \mathbf{I}, \quad (10)$$

where $\mathbf{x}$ is a $I \times G$ matrix of binary covariates indicating group membership. The correlation structure is more sophisticated than that for the simple structured heteroscedastic model of the previous section, allowing negative correlations:

$$Cor(Y_{ij}, Y_{i'j}) = \frac{\sigma_{g(i)g(i')}}{\sqrt{(\sigma_{g(i)g(i)} + \gamma_E)(\sigma_{g(i')g(i')} + \gamma_E)}} \quad [i \neq i'],$$

where $\{\sigma_{g(i)g(i')}\}$ is matrix $\mathbf{\Sigma}$. We have here a difference that is similar to that between the completely heteroscedastic model (2) and the Scheffé model (3), but now at the level of groups of genotypes instead of genotypes.

**Table 1** Numbers of parameters of the presented models

| Model | Fixed | $\mathbf{\Sigma}$ | $\mathbf{\Gamma}$ |
|---|---|---|---|
| Additive | $I$ | 1 | 1 |
| Heteroscedastic | $I$ | 1 | $I$ |
| Scheffé | $I$ | — | $I(I+1)/2$ |
| Mixed factorial regression | $HI = 2I$ | $K(K+1)/2 = 3$ | 1 |
| Shukla | $2I$ | 1 | $I$ |
| General (simple) | $HI$ | $K(K+1)/2$ | 1 |
| General (diagonal) | $HI$ | $K(K+1)/2$ | $I$ |
| General (unstructured) | $HI$ | — | $I(I+1)/2$ |
| Structured heteroscedastic | $I$ | 1 | $G$ |
| Correlated structured heteroscedastic | $I$ | $G(G+1)/2$ | 1 |
| Extended Shukla's model | $HI$ | 1 | $I$ |
| Heteroscledastic mixed factorial regression | $HI$ | $K(K+1)/2$ | $I$ |

### *Extending Shukla's model*

Shukla's model can be generalized by introducing more than one environmental covariate. Although this extension was mentioned by Shukla (1972), no one seems to have elaborated upon it since then. The variance structure, and therefore the correlations, are identical to those of Shukla's model given in (5), as is the interpretation of the model. The only difference is that more information about environments is taken into account for modelling the expectation.

### *Heteroscedastic mixed factorial regression model*

The mixed factorial regression model presented earlier can be generalized by supposing that the variance of the residual part is a function of the genotypes. Obviously, this is also a generalization of the model suggested in the previous section, adding genotypic covariates.

### *Identifiability and estimability*

When constructing models, one has to be cautious of overparameterization. For the variance components, the possibility of determining all the parameters uniquely, i.e. the identifiability problem, is equivalent to the question of whether they are uniquely determined for the covariance matrix for $Y_{ij}$ (Jöreskog, 1981). Hence, a necessary (but not sufficient) condition for identification is that the number of functionally independent variance components is less than or equal to $I(I+1)/2$. This is the reason why, under the assumption of an unstructured $\mathbf{\Gamma}$, no $\mathbf{\Sigma}$ can be added; this was the case for Scheffé's model (3). Still, overparameterization can occur, even when this necessary condition is fulfilled. Identifiability is a prerequisite for estimability, but it turns out, using the theory given in Rao & Kleffe (1988, Chapter 4), that for the models presented in this paper it also ensures estimability. Therefore, it is sufficient to check whether overparameterization occurs.

For the fixed parameters, it can be easily checked that the sufficient and necessary condition for estimability is that the matrix of environment covariates $\mathbf{z}$ be full column rank. If it is not the case, standard supplementary constraints can be used, or some covariates can be dropped, according to the preferences of the user.

### *Estimation*

Estimation of fixed parameters and variance parameters of the models presented in this paper is a special case of mixed model analysis. For most of them, the analysis can be performed using standard mixed model software; for instance, GENSTAT and SAS have special procedures, which allow estimation of variance components by common methods, such as Minimum Norm Quadratic Unbiased Estimation (MINQUE), Maximum Likelihood (ML) or REstricted Maximum Likelihood (REML). SAS is presently more flexible than GENSTAT (version 5.3.1) because it allows nonzero covariances in $\mathbf{\Sigma}$. Some hints are given in the Appendix.

When the data are complete, the generalized least squares estimates of the fixed parameters are identical to the ordinary least squares estimates. However, the variances of the estimates obtained by standard ordinary least squares programs will be incorrect, because a wrong variance structure will be used.

## Example: sugar yield in sugar beet in relation to infection with beet necrotic yellow virus

This section demonstrates the use of mixed models in field trial analysis. It will be shown how genotypic slopes can be used to model the expectation for differential genotypic responses in relation to an environmental covariate, and how remaining heteroscedasticity can be removed by including either an additional variance component or a genotypical covariate. We use a small data set, which allows a simple and meaningful interpretation and for which computations are easy to verify. Genotypes are fixed, environments are random and genotypic and environmental covariates are present.

The data concern sugar yields (ton/ha) in sugar beet. Ten cultivars with varying levels of resistance to beet necrotic yellow vein virus were evaluated in 1990 at six locations in the Netherlands, which varied in infestation level. Table 2 gives the sugar yields ($Y_{ij}$) together with a resistance indicator for the cultivars ($x_i$; low is resistant, high is susceptible) as obtained from a greenhouse test, and an infestation indicator for the locations ($z_j$; low is non-infested, high is heavily infested). For experimental details and phytopathological background see Paul *et al.* (1993). For the analysis, GENSTAT 5 committee (1993) was used (see Appendix).

Three models denoted (a), (b) and (c), were fitted (Table 3). Their respective variance component estimates can be found in Table 4. Model (a) contains fixed intercepts and slopes for each genotype with respect to the infestation pressure. From previous research, this model can be considered as adequate for modelling the expectation (Paul *et al.*, 1993). Estimated genotypic means are given in Table 5 together with standard errors and standard errors of differences. The means represent the sugar yields in an average infested environment. Because the table was complete, all genotypic means have the same

**Table 2** Sugar yield in beet and concomitant information

| Cultivar | Location | | | | | | Resistance |
|---|---|---|---|---|---|---|---|
| | W | O | LZ | N I | N II | A | |
| Roxane | 12.28 | 9.46 | 10.88 | 13.40 | 11.71 | 11.61 | 1.68 |
| Samba 2 | 11.56 | 8.51 | 10.30 | 13.30 | 10.64 | 9.49 | 1.71 |
| Rizo 92 | 11.37 | 8.63 | 10.11 | 11.58 | 10.17 | 10.01 | 1.72 |
| Rima | 12.11 | 9.26 | 11.22 | 12.92 | 10.85 | 10.50 | 1.87 |
| Rizofort | 12.33 | 9.25 | 10.82 | 12.52 | 11.23 | 9.53 | 1.91 |
| Donna | 11.03 | 9.04 | 9.58 | 11.16 | 10.34 | 9.24 | 2.17 |
| M 8917 | 13.75 | 9.51 | 11.47 | 11.83 | 10.26 | 9.59 | 2.31 |
| Univers | 13.45 | 9.84 | 11.76 | 11.10 | 8.93 | 6.64 | 2.40 |
| Regina | 13.35 | 9.96 | 10.89 | 9.98 | 8.15 | 6.73 | 2.49 |
| Accord | 13.65 | 10.61 | 10.66 | 10.58 | 7.95 | 6.32 | 2.51 |
| Infestation | 0.00 | 0.01 | 0.37 | 0.71 | 1.51 | 2.10 | |

standard error for model (a). For environments that are more or less infested than the average environment, the expected sugar yield can only be obtained by taking into account the differential susceptibility of the genotypes to infestation given by the slope (Table 5). All but one genotype had negative sensitivities, i.e. with higher infestation they did relatively worse.

For the fixed effects, hypotheses of the type $\alpha = 0$ can be tested by the use of Wald statistics defined as $\hat{\alpha}' [V(\hat{\alpha})]^{-1}\hat{\alpha}$; the treatment sum of squares divided by an estimate for the error. These Wald statistics

**Table 3** Models fitted to sugar yield in beet

| Formula | Expectation | Variance | Model |
|---|---|---|---|
| $Y_{ij} = \alpha_{i1} + \alpha_{i2}z_j + B_{j1} + E_{ij}$ | $\alpha_{i1} + \alpha_{i2}z_j$ | $\sigma_1 + \gamma_E$ | (a) |
| $Y_{ij} = \alpha_{i1} + \alpha_{i2}z_j + B_{j1} + E_{ij}$ | $\alpha_{i1} + \alpha_{i2}z_j$ | $\sigma_1 + \gamma_{Eg(i)}$ | (b) |
| $Y_{ij} = \alpha_{i1} + \alpha_{i2}z_j + B_{j1} + x_i B_{j2} + E_{ij}$ | $\alpha_{i1} + \alpha_{i2}z_j$ | $\sigma_1 + \sigma_2 x_i^2 + \gamma_E$ | (c) |

**Table 4** Variance component estimates, deviances and Wald's statistics for the three models (a), (b) and (c)

| | Model | | |
|---|---|---|---|
| | (a) | (b) | (c) |
| $\sigma_1$ | 1.62 | 1.54 | 1.63 |
| $\sigma_2$ | — | — | 1.29 |
| $\gamma_E$ | 0.30 | — | 0.16 |
| $\gamma_{E1}$ | — | 0.24 | — |
| $\gamma_{E2}$ | — | 0.84 | — |
| Deviances | 81.58 | 78.29 | 66.74 |
| Degrees of freedom | 39 | 38 | 38 |
| Wald test for intercepts (9 d.f.) | 59.6 | 73.8 | 75.1 |
| Wald test for common slope (1 d.f.) | 1.8 | 2.1 | 1.8 |
| Wald test for different slopes (9 d.f.) | 143.6 | 172.1 | 73.0 |

**Table 5** Beet cultivar means and slopes, with their standard errors, and minimum and maximum standard errors of differences between cultivars, estimated for the three models (a), (b) and (c)

| Cultivar | Mean | SE(a) | SE(b) | SE(c) | Slope | SE(a) | SE(b) | SE(c) |
|---|---|---|---|---|---|---|---|---|
| Roxane | 11.56 | 0.565 | 0.545 | 0.576 | 0.3784 | 0.725 | 0.699 | 0.739 |
| Samba 2 | 10.63 | 0.565 | 0.630 | 0.572 | −0.1760 | 0.725 | 0.808 | 0.733 |
| Rizo 92 | 10.31 | 0.565 | 0.545 | 0.570 | −0.0105 | 0.725 | 0.699 | 0.731 |
| Rima | 11.14 | 0.565 | 0.545 | 0.554 | −0.1433 | 0.725 | 0.699 | 0.710 |
| Rizofort | 10.95 | 0.565 | 0.545 | 0.551 | −0.4093 | 0.725 | 0.699 | 0.707 |
| Donna | 10.06 | 0.565 | 0.545 | 0.547 | −0.1999 | 0.725 | 0.699 | 0.702 |
| M 8917 | 11.07 | 0.565 | 0.545 | 0.556 | −1.0005 | 0.725 | 0.699 | 0.713 |
| Univers | 10.29 | 0.565 | 0.545 | 0.566 | −2.3533 | 0.725 | 0.699 | 0.726 |
| Regina | 9.84 | 0.565 | 0.545 | 0.578 | −2.3591 | 0.725 | 0.699 | 0.742 |
| Accord | 9.96 | 0.565 | 0.545 | 0.582 | −2.7349 | 0.725 | 0.699 | 0.746 |
| Min SED | | 0.317 | 0.285 | 0.228 | | 0.407 | 0.365 | 0.292 |
| Max SED | | 0.317 | 0.425 | 0.448 | | 0.407 | 0.401 | 0.574 |

have an asymptotic $\chi^2$ distribution with the degrees of freedom equal to those of the model term $\alpha$ (GENSTAT 5 committee, 1993). But before calculating Wald statistics, first the variance structure should be satisfactorily modelled, i.e. no pattern should be apparent in the residual effects. The residuals from model (a) seemed to contain some heteroscedasticity. Two approaches were used to model a more appropriate variance structure. Either extra stability variances can be added for less stable genotype [model (b)], or genotypic covariates can be introduced to account for the heteroscedasticity [model (c)].

Inspection of genotypic ecovalences revealed that Samba 2 behaves differently from the others. So, a separate variance component was added for this genotype distinguishing two groups of genotype $g(i) = 2$ for Samba 2 and $g(i) = 1$ for the other genotypes. A test for inclusion of variance components can be based on the differences in deviance, i.e. minus two times the log likelihood, between models (GENSTAT 5 committee, 1993). The reduction in deviance is approximately $\chi^2$-distributed with degrees of freedom equal to the difference in the number of parameters between the two models (one degree of freedom here). The deviance decreased by 3.3 after inclusion of the new component for Samba 2 (Table 4), which corresponds to a $P$-value less than 0.10. Because the data were balanced, the estimates for genotypic slopes and intercepts did not change, but their standard errors did (Table 5): all genotypes had slightly decreased standard errors, whereas Samba 2 had increased standard errors.

Alternatively, a genotypic covariate may be used to account for the heterogeneity of the residual variance; here we introduced a measure for disease resistance [model (c)]. This caused a strong decrease in the deviance (14.8 for one degree of freedom; $P$-value less than 0.001; Table 4). Again, the estimates for intercepts and slopes did not change, but each genotype now has its own standard errors. This result, together with the outcome of Wald's test for slope (see Table 4), leads us to prefer model (c).

## References


CALINSKI, T., CZAJKA, S. AND KACZMAREK, Z. 1987a. A model for the analysis of a series of experiments repeated at several places over a period of years. I. Theory. *Biuletyn Oceny Odmian*, **17–18**, 7–33.

CALINSKI, T., CZAJKA, S. AND KACZMAREK, Z. 1987b. A model for the analysis of a series of experiments repeated at several places over a period of years. II. Example. *Biuletyn Oceny Odmian*, **17–18**, 35–71.

CROWDER, M. J. AND HAND, D. J. 1990. *Analysis of Repeated Measures*. Chapman & Hall, London.

DENIS, J.-B. 1979. L'analyse de régression factorielle. *Biométrie-Praximétrie*, **10**, 1–34.

DENIS, J.-B. 1988. Two-way analysis using covariates. *Statistics*, **19**, 123–132.

DENIS, J.-B. 1994. *COBORU – Mixed Factorial Regression. Description of the Model and the Statistical Estimation Process via the Genotype–Environment Example*. Technical report, Laboratoire de Biométrie. INRA, Versailles.

DENIS, J.-B. AND DHORNE, T. 1989. Modelling interaction by regression with random coefficients. *Biuletyn Oceny Odmian*, **21–22**, 65–73.

GENSTAT 5 COMMITTEE, 1993. *Genstat 5, Release 3, Reference Manual*. Clarendon Press, Oxford.

GRUBBS, F. E. 1948. On estimation of precision of measuring instruments and product variability. *J. Am. Stat. Ass.*, **43**, 243–264.

HARDWICK, R. C. AND WOOD, J. T. 1972. Regression methods for studying genotype–environment interactions. *Heredity*, **28**, 209–222.

JÖRESKOG, K. G. 1981. Analysis of covariance structures. *Scand. J. Stat.*, **8**, 65–92.

KANG, M. S. 1993. Simultaneous selection for yield and stability in crop performance trials: consequences for growers. *Agron. J.*, **85**, 754–757.

KANG, M. S. AND GORMAN, D. P. 1989. Genotype × environment interaction in maize. *Agron. J.*, **81**, 662–664.

LONGFORD, N. T. 1993. *Random Coefficient Models*. Clarendon Press, Oxford.

PAUL, H., VAN EEUWIJK, F. A. AND HEIJBROEK, W. 1993. Multiplicative models for cultivar by location interaction in testing sugar beet for resistance to beet necrotic yellow vein virus. *Euphytica*, **71**, 63–74.

PIEPHO, H.-P. 1994a. Partitioning genotype–environmental interaction in regional yield trials via a generalized stability variance. *Crop Sci.*, **34**, 1682–1685.

PIEPHO, H.-P. 1994b. Missing observations in the analysis of stability. *Heredity*, **72**, 141–145. Correction in *Heredity*, (1994), **73**, 458.

PIEPHO, H.-P. 1994c. Application of a generalized Grubbs' model in the analysis of genotype–environment interaction. *Heredity*, **73**, 113–116.

PIEPHO, H.-P. 1995. Detecting, interpreting and handling heteroscedasticity in yield trial data. *Commun. Stat. B*, **24**, 243–274.

PIEPHO, H.-P. 1996a. Analysis of genotype-by-environment interaction and phenotypic stability. In: Kang, M. S. and Gauch H. G. (eds) *Genotype by Environment Interaction: New Perspectives*, pp. 151–174. CRC-Press, Boca Raton.

PIEPHO, H.-P. 1996b. Comparing cultivar means in multilocation trials when the covariance structure is not circular. *Heredity*, **76**, 198–203.

RAO, C. R. AND KLEFFE, J. 1988. *Estimation of Variance Components and Applications*. North-Holland, New York.

RUSSELL, T. S. AND BRADLEY, R. A. 1958. One-way variances in a two-way classification. *Biometrika*, **45**, 111–129.

SAS INSTITUTE INC. 1992. *SAS Technical Report P-229, SAS/STAT Software: changes and enhancements, release 6.07*. SAS Institute, Cary, NC.

SAS INSTITUTE INC. 1994. *SAS/STAT Software: changes and enhancements, release 6.10*. SAS Institute, Cary, NC.

SCHEFFÉ, H. 1959. *The Analysis of Variance*. John Wiley & Sons, New York.

SHUKLA, G. K. 1972. Some statistical aspects of partitioning genotype-environmental components of variability. *Heredity*, **29**, 237–245.

SHUKLA, G. K. 1982. Testing the homogeneity of variances in a two-way classification. *Biometrika*, **69**, 411–416.

VAN EEUWIJK, F. A., DENIS, J.-B. AND KANG, M. S. 1996. Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables. In: Kang, M. S. and Gauch H. G. (eds) *Genotype by Environment Interaction: New Perspectives*, pp. 15–49. CRC Press, Boca Raton.

WOOD, J. T. 1976. The use of environmental variables in the interpretation of genotype–environment interaction. *Heredity*, **37**, 1–7.

WRICKE, G. 1962. Über eine Methode zur Erfassung der ökologischen Streubreite in Feldversuchen. *Z. PflZücht.*, **47**, 92–96.

## Appendix

Most models discussed in this paper may be fitted using GENSTAT or SAS statistical packages. Here are given as illustration the corresponding commands. In the following program codes, GEN and LOC are classification variables for genotypes and locations, respectively. COV_GEN and COV_LOC are covariates for genotypes and locations.

### *GENSTAT code*

The mixed model analysis facilities in GENSTAT are centred around the statements VCOMPONENTS and REML (see Genstat 5 committee, 1993). With VCOMPONENTS, the structure of the fixed and random model is specified. For the additive model with fixed genotypes, random locations and equal stability variances for the genotypes (model 1), the required specification is

```
VCOMPONENTS [FIXED = GEN]\
  RANDOM = LOC.
```

The factors (qualitative covariates) GEN and LOC have length $IJ$. They indicate for each entry in the $I$ by $J$ genotype by location table the corresponding genotype and location. Covariates (for the interaction) can be included as follows:

```
VCOMPONENTS\
  [FIXED = GEN+GEN.COV_LOC[1]]\
  RANDOM = LOC+LOC.COV_GEN[1],
```

where COV_LOC[1] and COV_GEN[1] represent covariates for the locations and genotypes, respectively (model 4). Just like the factors GEN and LOC, these covariates have length $IJ$. COV_LOC[1] has the same value for all cells corresponding to a particular environment. Likewise, COV_GEN[1] changes value only when the genotype changes. The parameters estimated for GEN and LOC can be interpreted as intercepts, those for GEN.COV_LOC[1] and LOC.COV_GEN[1] as slopes. Including more than one covariate for genotypes and locations is straightforward:

```
VCOMPONENTS\
  [FIXED = GEN+GEN.COV_LOC[1...H]]\
  RANDOM = LOC+LOC.COV_GEN[1...K].
```

In the present release of GENSTAT (version 5.3.1), it is neither possible to specify correlations between intercepts and slopes nor between slopes mutually for the random model, but this situation will be remedied in the next release. Thus, at the moment, only the diagonal option for $\Sigma$ is available to GENSTAT users.

After the declaration of the fixed and random model the REML statement performs the analysis. In its most simple form, the mixed model, analysis for the response variable Y based on REML estimation of the variance components is

```
REML Y.
```

Default printed output contains, among other things, the estimates of the variance components plus their standard errors, and the sum of BLUEs (fixed effects) and BLUPs (random effects) plus their standard errors of differences as predictors of observations. The options PRINT, PTERMS and PSE provide ample facilities for printing other information as well, where PRINT controls general printing of all kinds of information, PTERMS selects the model terms for which printing of effects and means is wanted and PSE controls the type of standard errors that will be printed alongside the tables of effects and means. For example:

```
VCOMPONENTS\
  [FIXED = GEN+GEN.COV_LOC[1]]\
  RANDOM = LOC+LOC.COV_GEN[1]
REML [PRINT = Components,Effects,Means;\
  PTERMS = GEN+GEN.COV_LOC[1]+\
  LOC+LOC.COV_GEN[1];\
  PSE = Allestimates] Y
```

would produce estimates of the variance components with standard errors, BLUPS for the genotype and location means with standard errors and the estimates for all fixed and random effects with standard errors.

The default estimation procedure in GENSTAT is REML, but it is not difficult to obtain MIVQUE0 (MINQUE0) estimates. All we have to do is give the appropriate initial values for the variance components, i.e. all zero except for the error, which should have unit value, and allow only one iteration of the REML estimation algorithm. For example, MIVQUE0 estimates for model (4) are obtained by

```
VCOMPONENTS\
  [FIXED = GEN+GEN.COV_LOC[1]]\
  RANDOM = LOC+LOC.COV_GEN[1];
INITIAL = 0,0,1
REML [MAXCYCLE = 1] Y
```

The INITIAL parameter list must contain a value for each component specified in the RANDOM parameter list plus a value for the error.

To model different stability variances for individual genotypes, additional factors (qualitative variables) have to be declared and incorporated in the random model. For each genotype, except the last one, a factor must be declared. These factors have the levels (values) 1 to $J$ for the cells of (the row of) the genotype by location table corresponding to the genotype of interest, and have a missing value (*) for the $I(J-1)$ other cells. To fit Shukla's model (5), using REML estimation, we can use

```
VCOMPONENTS\
  [FIXED = GEN+GEN.COV_LOC[1]]\
  RANDOM = LOC+SVGEN[1...I_1]
REML Y
```

The factors SVGEN[1] to SVGEN[I_1] are needed to model the differences in variance between the stability variance of the genotype $I$ with the genotypes 1 to $I-1$. To get the stability variances, the estimated differences must be added to the variance of the $I$th genotype (the error). In the present version of GENSTAT, only the simple and diagonal options of eqn (8) are available. However, the next release will also provide the facilities to fit models with unstructured residual covariance matrices.

When groups of genotypes are required to have a common variance, group-specific factors have to be declared analogous to the genotypic-specific factors above. For example, when genotypes 1 to 5 differ from genotypes 6 to $I$ with respect to their residual variance, a factor SVGROUP[1] can be declared having the values 1 to 5$J$ for the cells corresponding to the genotypes 1 to 5, whereas this factor has missing values (*) elsewhere. The statements

```
VCOMPONENTS\
  [FIXED = GEN+GEN.COV_LOC[1]]\
  RANDOM = LOC+SVGROUP[1]
REML Y
```

fit a model that is similar to model (5), but now there are only two different stability variances, one for the genotypes 1 to 5 and the other for the genotypes 6 to $I$.

### *SAS code*

Most models discussed in this paper may be fitted using PROC MIXED of the SAS statistical package. For details see SAS Institute Inc. (1992, 1994). The following code fits a simple additive model with fixed genotypes and random locations; the residuals are assumed to have a common variance (model 1).

```
PROC MIXED METHOD = REML;
  CLASS GEN LOC;
  MODEL Y = GEN/SOLUTION NOINT;
```

```
RANDOM INT/SUB = LOC SOLUTION;
RUN;
```

The option METHOD = REML specifies the REML method for estimating covariance components. Alternative methods are ML and MIVQUE0. The CLASS statement is used to indicate the factors GEN for genotypes and LOC for locations. The MODEL statement specifies the fixed part of the linear model. Y is the dependent variable. The SOLUTION option in the MODEL statement prints the estimates of the fixed effects. The NOINT prevents fitting of a general mean. The random part of the model is specified in the RANDOM statement. The statement given above fits a random intercept term (INT) for each location (SUB = LOC); this is equivalent to fitting simple main effects. Note that locations may be regarded as subjects in the repeated measure terminology. The SOLUTION option in the RANDOM statement produces empirical BLUPS of the random effects.

In order to fit a linear regression of genotypes on a location covariate COV_LOC1 (say), add the term GEN*COV_LOC1 to the model:

```
MODEL Y = GEN GEN*COV_LOC1/SOLUTION
  NOINT;
```

More covariates (COV_LOC2...) can be added in a similar fashion.

A (random) linear regression of locations on a genotypic covariate (COV_GEN1) is fitted by stating the covariate with the RANDOM statement. In order to allow for a covariance between intercept and slope, the TYPE = UN (UN means unstructured covariance matrix) option must be invoked:

```
RANDOM INT COV_GEN1/SUB = LOC
  TYPE = UN SOLUTION;
```

Observe that the SUB = LOC option ensures that a separate slope is fitted for each location.

The modified code fits a mixed factorial regression model (model 4):

```
PROC MIXED METHOD = REML;
  CLASS GEN LOC;
  MODEL Y = GEN GEN*COV_LOC1/
    SOLUTION NOINT;
  RANDOM INT COV_GEN1/SUB = LOC
  TYPE = UN SOLUTION;
RUN;
```

So far, we have assumed that the residuals are independently distributed with common variance. The variance structure of the residual term can be modified by using the REPEATED statement. Heteroscedastic model (2) involves heterogeneity of the residual variances among genotypes. The appropriate SAS statement is:

```
REPEATED /SUB = LOC TYPE = UN(1);
```

SUB = LOC invokes a block-diagonal covariance matrix for the residual term, where blocks correspond to subjects = locations. The TYPE = UN(1) option produces diagonal blocks with a different diagonal element ('stability variance') for each genotype. For example, the generalized Shukla model could be fitted by SAS with the following code:

```
PROC MIXED METHOD = REML;
  CLASS GEN LOC;
  MODEL Y = GEN GEN*COV_LOC1 ...
GEN*COV_LOCH/SOLUTION NOINT;
  RANDOM INT/SUB = LOC SOLUTION;
  REPEATED/SUB = LOC TYPE = UN(1);
RUN;
```

An alternative statement to obtain the residual variance structure of heteroscedastic models is

```
REPEATED/GROUP = GEN;
```

By this statement, each level of GEN is assigned a different residual variance. If we define a new variable GENGROUP, which specifies groups of genotypes with homogeneous residual variance, we can fit a simple structured heteroscedastic model by

```
REPEATED/GROUP = GENGROUP;
```

Correlated structured heteroscedastic models could be fitted by defining appropriate dummy covariates associated with groups.