F.A. van Eeuwijk, J.-B. Denis & M.S. Kang (1996) Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables. In: Genotype-by-environment interaction: New perspectives, pp. 15-49. Eds. M.S. Kang and H.G. Gauch Jr. CRC Press, Boca Raton, Florida.

1

Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables

I. INTRODUCTION

For genotypes and environments making up the factor levels in two-way genotype by environment tables, often substantial additional information is either available or easily obtainable. For genotypes, additional quantitative information may be present from laboratory and greenhouse tests bearing on the physiology of the plants, while additional qualitative information may be present from various categorizations, like those on basis of genealogy. For environments, quantitative information can consist in edaphic and climatological data, whereas a minimum in qualitative information consists in year and location groupings. The additional information on genotypes and environments includes more than direct measurements. More remotely, statistics calculated from previous or comparable trials, concerning the variable under study as well as other variables, may be used.

For the additional information, more or less clear-cut hypotheses may be entertained regarding its relation with the structure of the genotype by environment interaction in the variable to be analyzed. To test these hypotheses statistically, models are necessary that allow the incorporation of that information. In plant breeding, the emphasis has long been on models not offering this opportunity. We feel that for a broad spectrum of hypotheses, between exploratory and inferential, it is imperative to pay more attention to regression based statistical methods. As a consequence, more parsimonious models may be built, providing more accurate tools to decide and act on. Similar ideas have been expressed by Hinkelmann (1974), Denis and Vincourt (1982), Tai (1990), van Eeuwijk (1993), and Federer and Scully (1993).

The main point of this paper is to give a survey of the most important regression based models for two-way tables and to illustrate the interpretation of their interaction parameters. Three families of models will be presented. After some details on notation (Section II), fixed factorial regression is introduced first (Section III), and an account is given of how quantitative as well as qualitative covariates may be included. Secondly, reduced rank factorial regression, based on bilinear descriptions of the interaction, will be dealt with (Section IV). Lastly, mixed factorial regression is presented, in which either the genotypes or the environments are supposed to represent a random sample from a population (Section V). Estimation and testing, together with software availability are briefly discussed in Section VI. Considerations playing a role in model choice form the subject of Section VII. Finally, Section VIII presents alternatives and further extensions to the models presented in the sections III, IV and V.

II. DATA AND NOTATION

A. THE DATA TABLE TO BE INTERPRETED

The topic of the paper will be restricted to the interpretation of the joint effects of the factors 'genotype' and 'environment' on a continuous variable Y. The subscript i (1,...,I) will be used to indicate the genotype, and j (1,...,J) to indicate the environment. Typically, Y_{ij} represents

yield, but many other quantitative variables are equally substitutable. Averaging over replicates makes Y_{ij} to comply closer with distributional assumptions.

For a number of statistical tests with regard to the structure of interaction an estimate of error is required. This estimate may be obtained from the mean intra-block error, or from the part of the interaction not modeled. We will not consider the question of which estimate to use, but assume that a non-controversial estimate of error is available.

B. ADDITIONAL INFORMATION.

Besides the data, Y_{ij}, additional information is assumed to be present at the levels of the genotypes and/or environments. The value of the k-th covariate ($k = 1,...,K$) for the i-th genotype will be denoted by x_{ik}. Covariates are either quantitative, as in multiple regression, or qualitative, as in analysis of variance. Examples of quantitative genotypic covariates are physiological characterizations, such as earliness, and disease susceptibilities. Examples of qualitative genotypic covariates are genetic and geographic origin. When only one covariate is considered, the subscript in question is dropped, *i.e.* x_i instead of x_{i1}. Quantitative covariates are throughout supposed to be centered.

Similarly, we will denote the h-th covariate ($h = 1,...,H$) for the j-th environment by z_{jh}. For environments, we can think of humidity and soil pH as quantitative covariates, and location (region, country) and cultivation regimes as qualitative covariates.

To indicate that the covariate values are considered known, they are written with lower case letters. This does not mean that the corresponding factor may not be random, but only that the analyses are done conditionally on the values of the covariates.

A popular environmental covariate is $y_{.j}$, the mean over genotypes (Finlay and Wilkinson, 1963). $y_{.j}$ can be treated as any other covariate (Mandel, 1961), because of its statistical independence of the interaction estimates. Also the genotypic main effect, $y_{i.}$ can be used in that way. A further possibility is to use the product of $y_{.j}$ and $y_{i.}$, as a covariate for the entire table. A simple extension uses covariates of the type $o_{i.}$ and $o_{.j}$, *i.e.* main effects of another variable on the same set of genotypes and environments. For even further extensions, see Baril (1992). Though mostly covariates are used in a linear fashion only, higher order terms as squares, cubes, and cross-products can be considered equally well.

In some models, pseudo covariates are estimated. When they are defined as linear combinations of measured covariates, they will be designated as synthetic covariates. When they are only subject to statistical/numerical construction rules, they will be called artificial covariates.

C. DESCRIPTION OF THE MODELS.

The models presented below consist of sums of model terms. Terms are related to the expectation or the variance. Fixed parameters are represented by lower case Greek letters, random parameters and variates by standard upper case letters, and observations on (co)variates by standard lower case letters. The error term is written E_{ij}. Unless stated otherwise, E_{ij}'s are assumed to have zero mean, constant variance, and to be uncorrelated. For fixed effects models, the decomposition of the degrees of freedom (parametric dimension) corresponding to the different model terms is displayed via recapitulative tables. These tables are two-dimensional depictions showing the composition of the models, in which each model term corresponds with a zone in the table, and where the area of this zone is proportional to the associated degrees of freedom (Denis, 1991).

D. EXAMPLE DATA SET.

To enhance understanding, for some of the presented models a numerical example will be used. The data set used is a modified and rounded version of the data used by Kang and Gorman (1989), including yield figures for 17 genotypes in 12 environments (Table 1). For each genotype, an associated fictitious resistance measure, Rstc, is available (Table 2). The 12 environments are characterized by four climatological variables (Table 3). An independent estimate for the error variance is also present.

III. FIXED FACTORIAL REGRESSION

A general formal treatment of the models presented in this section is given by Denis (1980, 1988).

A. THE ADDITIVE MODEL AS BASE LINE

It is common to define interaction in two-way tables relative to the two-way additive model,

$$Y_{ij} = \mu + \alpha_i + \beta_j + E_{ij}\,. \qquad (1)$$

The additive model provides a first, rough approximation to the data. Analysis of variance (ANOVA) on our example data showed that 82% of the total sum of squares could be explained by only 13% of the degrees of freedom (Table 4). Nevertheless, interaction was highly significant and could not be omitted.

For many purposes it is useful to express (1) as a double regression model with the constant covariates 1_i=1_j=1:

$$Y_{ij} = \mu + \alpha_i 1_j + 1_i \beta_j + E_{ij}\,. \qquad (1')$$

The main effects (α_i,β_j) thus are the regression coefficients for these non-informative constant covariates. The structure of the model is displayed in Table 5.

B. INCLUDING ONE QUANTITATIVE ENVIRONMENTAL COVARIATE.

Perhaps the simplest way of introducing a covariate associated with the environments, is to write the interaction as a regression on this covariate with the coefficient depending on the genotype. Early applications of this type of model include Knight (1970), and Freeman and Perkins (1971). More recent applications are Fakorede and Opeke (1986), and McGraw *et al.* (1986). The model can be written as

$$Y_{ij} = \mu + \alpha_i + \beta_j + \rho_i z_j + E_{ij}\,. \qquad (2)$$

For illustration we choose the rainfall data from Table 3 as environmental covariate. The estimates for the genotypic regression coefficients, ρ_i, are given in Table 2. These coefficients can be interpreted as underlying a differential genotypic response to rainfall. For example, for *G15* yield increases with rainfall relative to what might have been expected on the basis of an additive model. Under dry circumstances, this yield decreases. Recall that the covariates were all

centered.

The partitioning of the interaction sum of squares according to model (2) is given in Table 4. The sum of squares due to heterogeneity of genotypic slopes amounted to 38.3, with 16 degrees of freedom. The corresponding mean square, 2.395, is clearly greater than the mean square for the total interaction, 0.853. Therefore, rainfall can be considered to be a good explanatory covariate.

Table 6 reveals the structure of the model by showing its recapitulative table. Because the environmental main effect was fitted before the regression on z_j, the $\rho_i z_j$ term of the interaction corresponds to I-1 degrees of freedom for I parameters.

C. INCLUDING ONE QUANTITATIVE GENOTYPIC COVARIATE.

The counterpart of model (2), including one genotypic covariate is

$$Y_{ij} = \mu + \alpha_i + \beta_j + x_i \tau_j + E_{ij}. \qquad (3)$$

An untimely use of this model can be found in Freeman and Crisp (1979). For our example, x_i is a resistance measure for genotype i, as given in Table 2. Now, τ_j can be interpreted as the potential of environment j to favor the spread of the disease. If the environment is beneficial to the spread of the disease, *i.e.* τ_j is large and positive, and if the genotype is susceptible, *i.e.* x_i is large and negative, then the correction term $x_i \tau_j$ will be large and negative, implying a decrease in yield. Estimates for τ_j are given in Table 3, and the explained sum of squares is given in Table 4. The mean square amounted to 2.816, again much greater than the total interaction mean square. The number of degrees of freedom attributed to a term is not determined by the factor with which the covariate is associated, but by the opposite factor (Table 7).

D. INCLUDING SEVERAL QUANTITATIVE ENVIRONMENTAL COVARIATES

A generalization of (2), including two environmental covariates leads to

$$Y_{ij} = \mu + \alpha_i + \beta_j + \rho_{i1} z_{j1} + \rho_{i2} z_{j2} + E_{ij}, \qquad (4)$$

where z_{j1} and z_{j2} can be rainfall and average maximum temperature over the growing season in environment j. The structure of the model is given in Table 8. When covariates are correlated, inclusion of more than one covariate complicates interpretation of the coefficients, just as for multiple regression. Coefficients are conditional upon the values of the other included covariates, so one should be cautious in interpretations. Examples of the application of model (4) can be found in Hardwick (1972), Hardwick and Wood (1972), Rameau and Denis (1992), and van Eeuwijk and Elgersma (1993).

E. INCLUDING ONE QUALITATIVE GENOTYPIC COVARIATE

Qualitative genotypic covariates attribute group membership to genotypes. Let x_i be a qualitative variable that indicates to which of three groups with a common ancestor a genotype belongs. For example, x_i=3 would mean that genotype i belongs to the third group of genotypes. Including this variable x_i in model (3) does not result in anything sensible, because the numbering of the groups is arbitrary and does not refer to something inherent to that group of

genotypes. What we can do is replace the qualitative variable x_i by indicator variables (valued 0 or 1), just as when ANOVA models are presented in multiple regression form. Now x_{i1}, x_{i2}, and x_{i3} attribute membership when they have value 1, *e.g.* x_{i3}=1 means genotype *i* belongs to group 3. Of course, if x_{i3}=1, then x_{i1}=x_{i2}=0, therefore x_{i1}+x_{i2}+x_{i3} is always one. This redundancy can be removed by leaving out one of the indicator variables, or imposing an additional constraint. Another possibility is to remove the environmental main effect, as is done in

$$Y_{ij} = \mu + \alpha_i + x_{i1}\,\tau_{j1} + x_{i2}\,\tau_{j2} + x_{i3}\,\tau_{j3} + E_{ij}\,. \qquad (5)$$

The parameters τ_{jk} represent the environmental 'main' effects for each of the three groups of genotypes separately. Table 9 displays the structure of the model.

F. INCLUDING GENOTYPIC AND ENVIRONMENTAL COVARIATES

1. Quantitative-quantitative

The simplest extension of the additive model including one genotypic and one environmental covariate is

$$Y_{ij} = \mu + \alpha_i + \beta_j + x_i\nu\ z_j + E_{ij}\,. \qquad (6)$$

It can be derived from models (2) or (3) by imposing the restriction of ρ_i=$x_i\nu$, or τ_j=νz_j, respectively. In Table 10 it is shown how the single parameter ν represents one degree of freedom in the interaction space.

The model has been fitted for all combinations of genotypic and environmental covariates at our disposal (Table 4). The combination of genotypic resistance and environmental rainfall produced the highest mean square, as might have been expected from the previous results.

One step further than model (6), a simple combination of (2), (3), and (6) gives

$$Y_{ij} = \mu + \alpha_i + \beta_j + x_i\nu\ z_j + x_i\,\tau_j + \rho_i\,z_j + E_{ij}\,. \qquad (7)$$

The recapitulative table (Table 11) for this model shows how $x_i\nu z_j$ is common to both $x_i\,\tau_j$ and ρ_i z_j. To estimate ν, supplementary constraints have to be imposed on τ_j and ρ_i. The ANOVA table (Table 12) shows that a significant amount of interaction was left unexplained by model (7). For a more telling example, see Paul *et al.* (1993).

Model (6) can straightforwardly be extended to include several genotypic as well as environmental covariates, to give

$$Y_{ij} = \mu + \alpha_i + \beta_j + \sum_{k=1}^{K}\sum_{h=1}^{H} x_{ik}\ \nu_{kh}\ z_{jh} + E_{ij}\,. \qquad (8)$$

Good illustrations of applications of model (8) are given by Charmet *et al.* (1993), and Baril *et al.* (1995).

2. Qualitative-quantitative

Taking x_i qualitative in model (6) leads to the model

$$Y_{ij} = \mu + \alpha_i + \beta^*_j + x_{i1}\nu_1 z_j + x_{i2}\nu_2 z_j + x_{i3}\nu_3 z_j + E_{ij}\,, \qquad (9)$$

and Table 13. The parameters β* represent the environmental main effects after adjustment for the general mean and the regressor z_j. The β*'s may be interpreted as a type of residuals. For applications see Saeed and Francis (1984), and Royo *et al.* (1993).

3. Qualitative-qualitative

When in model (9) both the genotypic covariate and the environmental covariate are qualitative, we arrive at

$$\begin{aligned} Y_{ij} = \alpha^*_i + \beta^*_j + x_{i1}\nu_{11} z_{j1} + x_{i2}\nu_{21} z_{j1} + x_{i3}\nu_{31} z_{j1} + \\ x_{i1}\nu_{12} z_{j2} + x_{i2}\nu_{22} z_{j2} + x_{i3}\nu_{32} z_{j2} + E_{ij}\,, \end{aligned} \qquad (10)$$

and Table 14. The environmental covariate z_j may indicate one of two regions, and is represented in the model by two indicator variables, z_{j1} and z_{j2}. In addition to the β*'s of model (9), α*'s appear, representing the genotypic main effects after adjustment for cross-product terms involving x_i. The parameters ν_{kh} represent the mean for the genotypes of the genotypic group k (descendance) in the environments of the environmental group h (region).

Model (10) can be reparametrized giving

$$Y_{ij} = \mu + \alpha_i + \beta_j + \nu^*_{[x_i],[z_j]} + E_{ij}\,, \qquad (10')$$

and Table 15. In (10') the usual main effects are included, and the $\nu^*_{[x_i],[z_j]}$ 12's have to sum to zero over genotypes (sum over i) and environments (sum over j). Being adjusted for the main effects, they are interaction parameters. Interaction is exclusively of the 'between by between' type. One might think of classifying the original data in the six groups following from the intersection of the three genotypic groups with the two environmental groups. Interaction is present only between these six groups.

Many authors have studied models of the type exemplified by (10'). Although the use of a priori groupings is inferentially superior over the use of a posteriori groupings, most references relate to the latter (Horner and Frey, 1957; Abou-el-Fittouh *et al.*, 1969; Lin and Thompson, 1975; Byth *et al.*, 1976; Denis, 1979; Seif *et al.*, 1979; Berbigier *et al.*, 1980; Brennan *et al.*, 1981; Brown *et al.*, 1983; Lefkovitch, 1985; Lin and Butler, 1988; Corsten and Denis, 1990; Crossa *et al.*, 1990; Arntzen and van Eeuwijk, 1992; Muir *et al.*, 1992; Oliveira and Charmet, 1992). With a priori grouping, the procedure is fully inferential, otherwise it is more exploratory. The inferential value when using a posteriori groupings remains a point of discussion. Certainly, the type of testing needs more consideration in these cases.

IV. REDUCED RANK FACTORIAL REGRESSION

Theory on general reduced rank regression models has been developed over time by a number of authors belonging to very different disciplines. Among the major contributions we list Rao (1964), Izenman (1975), van den Wollenberg (1977), Gabriel (1978), Obadia (1978), Tso (1981), Davies and Tso (1982), Sabatier *et al.* (1989), van der Leeden (1990), and Velu (1991).

As a solution to genotype by environment interaction problems in plant breeding, reduced rank *factorial* regression models have been proposed. Important contributions are due to Wood (1976), Denis (1991), van Eeuwijk (1992a), and van Eeuwijk (1995a).

A. ONE-WAY REDUCED RANK REGRESSION WITH ONE TERM

Considering model (4) and Table 8, we see that up to *J*-1 covariates are conceivable. For the case of *J*-1 covariates, the interaction described would be equal to the total non-additivity remaining from the additive two-way model (1). A number of covariates would then very likely be modeling mere noise, as in most situations with large numbers of covariates. A method allowing the incorporation of substantial amounts of covariates, while using fewer degrees of freedom than a comparable factorial regression model, is reduced rank (factorial) regression. Basically, a so-called synthetic covariate is formed as a linear combination of the available covariates, *i.e.* the most explanatory linear combination that can be constructed according to a least squares criterion. A synthetic covariate can be incorporated in a model like (2) without further complications. Define the synthetic covariate

$$\zeta_j = \sum_{h=1}^{H} \lambda_h z_{jh} . \qquad (11)$$

The coefficients λ_h are unknown parameters to be estimated from the data.
The model becomes:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \rho_i \left(\sum_{h=1}^{H} \lambda_h z_{jh} \right) + E_{ij} . \qquad (12)$$

Table 16 shows the distribution of the degrees of freedom over the various terms. It is obvious that substantial amounts of degrees of freedom can be won. As an illustration, compare Table 16 with Table 8. Table 8 gives the degrees of freedom for the interaction in a factorial regression model with 2 environmental covariates (H=2), 2(I-1). The comparable reduced rank regression model (12) uses I degrees of freedom. In general, the difference between a reduced rank model as (12) and the corresponding full rank model amounts to (I-2)(H-1) degrees of freedom. The difference increases with I and H. This increase in parsimony can express itself in greater accuracy and stability. For interpretational purposes, one should try to integrate the synthetic covariate in subject matter knowledge about the environments. For the genotypic sensitivities, ρ_i, a physiological basis should be sought.

Model (12) is not linear in its parameters, but bilinear. Least squares estimates are no longer linear combinations of the observations, but can come from a singular value decomposition of the fitted values matrix of the factorial regression model including the same set of covariates.

Wood (1976) contains an example of a reduced rank factorial regression model with one synthethic covariate.

B. ONE-WAY REDUCED RANK REGRESSION WITH SEVERAL TERMS

There is no need to restrict the number of synthetic covariates in reduced rank models to just one,

$$Y_{ij} = \mu + \alpha_i + \beta_j + \sum_{r=1}^{R} \rho_{ir} \left(\sum_{h=1}^{H} \lambda_{hr} \; z_{jh} \right) + E_{ij} \; . \qquad (13)$$

Model (12) follows from (13) by taking R=1. Table 17 represents the recapitulative table.

Illustrations of the use of model (13) are presented in van Eeuwijk (1992a), and van Eeuwijk *et al.* (1995).

C. TWO-WAY REDUCED RANK REGRESSION

Synthetic covariates may be used on both the genotypic as well as the environmental dimension of the table. We define the genotypic synthetic covariate as

$$\xi_i = \sum_{k=1}^{K} \pi_k \; x_{ik} \; , \qquad (14)$$

with the π_k as unknown parameters to be estimated. This leads to the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \left(\sum_{k=1}^{K} \pi_k \; x_{ik} \right) \left(\sum_{h=1}^{H} \lambda_h \; z_{jh} \right) + E_{ij} \; , \qquad (15)$$

with the recapitulative table given in Table 18.

D. REDUCED RANK REGRESSION INDEPENDENT OF COVARIATES

When I-1 linearly independent genotypic covariates are used to create a synthetic covariate, there is no restriction on the ξ_i's of having to be a linear combination of the x_k's. The same holds true for the ζ_j's when there are J-1 environmental covariates. Model (15) can thus be defined without reference to covariates as

$$Y_{ij} = \mu + \alpha_i + \beta_j + \xi_i \zeta_j + E_{ij} \; . \qquad (16)$$

Model (16) is known under various names, like AMMI model (Gauch, 1988) and bilinear model (Denis, 1991). Recently it was placed in the biadditive model family by Denis and Gower (1992, 1994a), in an attempt to create a more unified nomenclature for models for two-way tables. Table 19 gives the distribution of the degrees of freedom over the model terms.

For completeness we give the extension of (16) to more than one term,

$$Y_{ij} = \mu + \alpha_i + \beta_j + \sum_{r=1}^{R} \xi_{ir} \zeta_{jr} + E_{ij} \; . \qquad (17)$$

The recapitulative table of model (17) is shown in Table 20.

Models (16) and (17) are extensively used in plant breeding. A good review that emphasizes prediction can be found in Gauch (1992). A brief exposition emphasizing interpretation is given by van Eeuwijk (1992b). Generalized bilinear models are described in van Eeuwijk (1995b).

V. MIXED FACTORIAL REGRESSION

General presentations of the models proposed in the subsections A and B can be found in Goldstein and McDonald (1988), and Denis and Dhorne (1989).

A. GENOTYPES FIXED AND ENVIRONMENTS RANDOM

Sometimes, the environments included in an experiment can be assumed to represent a random sample from a population of environments, thereby fulfilling a sufficient condition for a mixed model approach. Model (2) can be changed into a mixed model by replacing the fixed environmental parameters indexed by *j*, by random parameters;

$$Y_{ij} = \mu + \alpha_i + B_j + \rho_i z_j + E_{ij} . \qquad (18)$$

Although the environments are considered random, z_j is not considered to be random, as the analysis proceeds conditional on the value of z_j. The random parameters B_j can be obtained as best linear unbiased predictions, after estimation of the variance component σ_{BB}=var(B_j) (Searle *et al.*, 1992).

B. ENVIRONMENTS FIXED AND GENOTYPES RANDOM

In the early phases of the selection process, plant breeders tend to work with groups of genotypes that are considered to be samples from larger populations, whose performance needs to be estimated in a number of well defined environments. By inserting a random genotype in model (2) we obtain

$$Y_{ij} = \mu + A_i + \beta_j + R_i z_j + E_{ij} . \qquad (19)$$

Variance components to be estimated are σ_{AA}=Var(A_i), σ_{RR}=Var(R_i), and σ_{AR}=Cov(A_i,R_i). For individual genotypic performances, again best linear unbiased predictions can be calculated. A noticeable feature of model (19) is that the variance of Y_{ij} depends on *j*, the environment,

$$Var\left(Y_{ij}\right) = \sigma_{AA} + 2\sigma_{AR} z_j + \sigma_{RR}\left(z_j\right)^2 + \sigma_{EE} . \qquad (20)$$

When one wants to use model (19) to predict future genetic gain, one should be aware that the gain depends on the particular environment *j*.

C. GENOTYPES FIXED, ENVIRONMENTS RANDOM, AND RANDOM INTERACTION DEPENDING ON GENOTYPE

Shukla (1972a,b) introduced a model that included fixed genotypes and random environments, besides a genotypic specific error component. Interesting applications are present in Kang and Miller (1984), Gorman *et al.* (1989), Kang and Gorman (1989), Gravois *et al.* (1990), Helms (1993), Magari and Kang (1993), and Kang (1993). The model formulation is very similar to (18);

$$Y_{ij} = \mu + \alpha_i + B_j + \rho_i z_j + E_{ij} , \qquad (21)$$

with the variance of E_{ij} depending on the genotype;

$$Var\left(E_{ij}\right) = \sigma_{EE}(i). \qquad (22)$$

The variance $\sigma_{EE}(i)$ is usually interpreted as a stability associated with genotype *i*. Inclusion of more than one covariate is straightforward;

$$Y_{ij} = \mu + \alpha_i + B_j + \sum_{h=1}^{H} \rho_{ih}\, z_{jh} + E_{ij}\,. \qquad (23)$$

Results of the application of models (21) and (23) to the example data are presented in Table 21.

Deleting the regression term $\rho_i z_j$ from model (21) produces a well-known, particularly simple type of heteroscedastic model, often discussed in literature (Russel and Bradley, 1958; Shukla, 1982; Snee, 1982; Denis, 1983; Vincourt *et al.*, 1984; Longford, 1987; Searle *et al.*, 1992; Mudholkar and Sarkar, 1992).

VI. ESTIMATION AND TESTING

A. FIXED FACTORIAL REGRESSION

Fixed factorial regression models fall in the class of fixed linear models and therefore no special problems arise with regard to estimation of parameters and testing of hypotheses.

B. REDUCED RANK FACTORIAL REGRESSION

The inclusion of bilinear terms complicates estimation and testing. Closed form least squares estimators and asymptotic variances are only known for orthogonal cases, *i.e.* without missing values and with proportional numbers of replications (Denis and Gower, 1992, 1994b). In non-orthogonal cases, numerical approaches are inevitable, and tests and confidence intervals will be approximate.

C. MIXED FACTORIAL REGRESSION

With the exception of the models developed by Shukla, which seem to need a specific procedure, estimation for mixed factorial regressions can be done using restricted maximum likelihood.

D. SOFTWARE

Most of the models presented can be processed with the main statistical packages that include programming facilities, for example Genstat (1993), SAS (1992), and S-plus (1994). The most important Genstat statements for fixed full and reduced rank factorial regression have been added as an Appendix to van Eeuwijk *et al.* (1995). Special purpose packages also have been developed. We mention first MatModel (Gauch, 1990), which deals mainly with AMMI models. An attractive feature of this package is the cross-validation procedure for assessing the number of interaction terms, when replicates are present. INTERA (Decoux and Denis, 1991) offers facilities for a wide range of fixed factorial regression models and AMMI models, applicable to balanced and unbalanced data. Furthermore, INTERA can fit models combining features of both factorial regression and AMMI. Computer programs to calculate the ecovalence (Wricke, 1962) and Shukla's stability statistics, σ and s (Shukla, 1972a), are described in Kang (1988, 1989). Presently a new program is available that calculates, in addition to the above mentioned statistics, the YS_i statistic, which combines yield and stability into a single selection criterion

(Kang, 1993).

VII. SOME CONSIDERATIONS WITH RESPECT TO MODEL CHOICE

The basic question for the experimenter is, which model to choose out of all the possibilities enumerated above? No definite answer is possible. The choice strongly depends on the desired goal. Various choices accompany the model selection process. We briefly address four issues.

A. CONSTRAINTS

All the models described are overparameterized. Supplementary constraints can be imposed to solve this indeterminacy. Various possibilities exist. Natural extensions of the sum-to-zero constraints were proposed by Denis (1991). These lead to orthogonal decompositions, that are convenient for the construction of recapitulative tables. However, we feel that mathematical convenience should always be made subordinate to biological knowledge, also in choosing identification constraints.

B. COVARIATE SELECTION

The most difficult point in the application of factorial regression models seems to be the choice of a good subset of covariates for genotypes as well as environments. It is a variable selection problem having the square of the complexity of that of variable selection in the standard `one-way' multiple regression context. It is important to keep in mind that the size of the sample for factorial regressions is not *IJ*, but *I*-1 for regressions with genotypic covariates, and *J*-1 for regressions with environmental covariates.

Denis (1988) contains a discussion of variable selection strategies for factorial regression models. It is shown how nesting relationships between models can be used to test for the inclusion of covariates and the possibility of rank reduction.

In the absence of subject matter knowledge, exhaustive variable searches may be used as exploratory analyses. One should then be cautious against over-interpretation, and correct for selection bias by using an appropriate experiment-wise error rate. If possible, it is, however, always preferable to work inferentially, *i.e.* test specific hypotheses following from subject matter knowledge about the interaction of physiological processes in the plant with defined environmental factors. The relevancy of selected covariates can be further investigated in future trials, as a safeguard against conclusions based on chance correlations.

C. FIXED OR RANDOM

Another important question is the choice of terms as fixed or random. Two main types of arguments can be distinguished. A first type of argument is based on sampling considerations. Do the genotypes and/or environments in the experiment constitute a sample from a population to which the inference is directed? The second kind of arguments is more pragmatic, and involves the desirability of shrinkage and recovery of information, and the convenience of choosing a model term random when many parameters are associated with the term. With regard to shrinkage we may question whether it is reasonable to shrink estimates deviating from the mean of the sample back towards that mean? Or, should relatively good genotypes pay for being an element of a relatively bad sample, while relatively bad genotypes benefit from being an

element of a relatively good sample? Considerations concerning recovery of information play a role when data are unbalanced. At all times it must be possible to assess whether the random effects indeed could have come from the assumed distribution. For example, for the estimation of a variance component, at least 10 degrees of freedom should be available, otherwise it is preferable to take the term fixed. The same remark applies to Shukla's approach, many environments are needed for accurate estimates of individual genotypic variances.

D. PARSIMONY

In model building and model choice, one should always take into account the parsimony principle (Gauch, 1988), *i.e.* avoid-over fitting. By including ever more covariates, the amount of interaction described will keep on increasing. However, as a consequence, more noise will be fitted, leading to less robust models. From this perspective, reduced rank regression models are attractive as they allow more covariates for the same number of degrees of freedom. A word of caution should be given for too uncritically accepting the degrees of freedom attributed to synthetic and artificial covariates. When pattern does not clearly dominate noise, these degrees of freedom will be too low, thus declaring the influence of synthetic and artificial covariates significant, when it is not (Gauch, 1992; Williams and Wood, 1993; Cornelius, 1993).

VIII. ALTERNATIVES AND EXTENSIONS

Despite the long enumeration of models given above, the possibilities of modeling interaction using additional information are not exhausted. In this final section, we give some further ideas on the subject.

A. DECOMPOSING MAIN EFFECTS

Use of covariates for decomposition of variation need not be kept restricted to interaction effects. Main effects can be split into a part due to regression on a covariate and a residual

$$Y_{ij} = \mu + x_i \alpha_0 + \alpha_i^* + \beta_0 z_j + \beta^*_j + x_i \nu z_j + E_{ij} . \qquad (6')$$

The residual from a main effects regression almost always is strongly significant because of the dominant role of the main effects in the description of the total variation.

B. ANALYSIS OF COVARIANCE

Some authors (Snedecor and Cochran, 1976; Searle, 1979) have used the following model under the name of analysis of covariance;

$$Y_{ij} = \mu + \alpha_i + \beta_j + \rho o_{ij} + E_{ij} , \qquad (24)$$

where o_{ij} is a covariate whose value depends specifically on the cell (i,j). As previously indicated, covariates defined on cell level can be subsumed under the factorial regression models by defining a genotypic covariate $x_i=o_{i.}-o_{..}$ and an environmental covariate $z_j=o_{.j}-o_{..}$, where a dot means averaging.

C. PARTIAL LEAST SQUARES REGRESSION

Multivariate partial least squares regression models have been proposed to model interaction in dependence on covariates (Aastveit and Martens, 1986; Talbot and Wheelwright,

1989). These models can be interpreted in a way reminiscent of reduced rank regression. Partial least squares can be viewed as a robust estimation procedure.

D. BIADDITIVE MIXED MODELS

An interesting conjunction of model classes is given by allowing the multiplicative covariates in biadditive models, to which the Finlay-Wilkinson and AMMI model belong, to be random. Some preliminary work has been done here by Oman (1991).

E. PIECEWISE REGRESSION

Genotypic responses to many environmental factors will reach an upper limit. A simple way to model this kind of response is;

$$Y_{ij} = \mu + \alpha_i + \beta_j + \rho_i \, Min(\phi_i, z_j) + E_{ij} . \qquad (25)$$

In model (25) the covariate z_j is replaced by the minimum of a threshold φ_i and the covariate z_j. Each genotype has its own threshold after which the response cannot increase any more.

F. GENERALIZED LINEAR AND BILINEAR MODELS

All fixed factorial regression models dealt with so far assume that the expectation can be modeled linearly in the parameters and that the variance is constant. Deviations from these assumptions sometimes can be cured by transformation of the response. However, the optimal transformation for achieving linearity need not be the same as the optimal transformation for achieving homogeneity of variance. For the models in the class of generalized linear models it is not necessary to find a transformation of the response as a compromise between first (expectation linear in the parameters) and second order (homogeneous variance) requirements. In generalized linear models a suitable transformation of the expectation can be combined with a convenient choice for a variance function, expressing the dependence of the variance on the mean (McCullagh and Nelder, 1989). Generalized factorial regression models extend considerably the range of application for factorial regression models. A further elaboration in the form of generalized bilinear models is discussed in van Eeuwijk (1995b).

G. HIGHER WAY FACTORIAL REGRESSION

Genotype by environment problems often involve more than two factors. Environments are usually cross-classifiable by years and locations. This fact does not complicate the use of factorial regression models for the fixed and mixed cases. Somewhat harder to make are the extensions to the class of biadditive models, although progress is made also here. In van Eeuwijk and Kroonenberg (1995) quadri-additive models are introduced for three-way interaction.

ACKNOWLEDGEMENT

Thanks are due to L.C.P. (Paul) Keizer for his help in the preparation of the manuscript.

REFERENCES

Aastveit, A.H., and H. Martens. 1986. ANOVA interactions interpreted by partial least squares regression. *Biometrics* 42:826-844.

Abou-el-Fittouh, H.A., J.O. Rawlings and P.A. Miller. 1969. Classification of environments to control genotype by environment interactions with an application to cotton. *Crop Sci* 9:135-140.

Arntzen, F.K., and F.A. van Eeuwijk. 1992. Variation in resistance levels of potato genotypes and virulence level of potato cyst nematode populations. *Euphytica* 62:135-143.

Baril, C.P. 1992. Factor regression for interpreting genotype-environment interaction in bread-wheat trials. *Theor. Appl. Genet.* 83:1022-1026.

Baril, C.P., Denis, J.-B., Wustman, R. and van Eeuwijk, F.A. 1995. Analysing genotype by environment interaction in Dutch Potato Variety Trials using factorial regression. *Euphytica*, in press.

Berbigier, A., J.-B. Denis and C. Dervin. 1980. Interaction variété-lieu: analyse du rendement d'orges de printemps. *Ann. Amélior. Plantes* 30:79-94.

Brennan, P.S., D.E. Byth, D.W. Drake, I.H. DeLacy and D.G. Butler. 1981. Determination of the location and number of test environments for a wheat cultivar evaluation program. *Aust. J. Agric. Res.* 32:189-201.

Brown, K.D., M.E. Sorrells and W.R. Coffman. 1983. Method for classification and evaluation of testing environments. *Crop Sci* 23:889-893.

Byth, D.E., R.L. Eisemann and I.H. DeLacy. 1976. Two-way pattern analysis of a large data set to evaluate genotypic adaptation. *Heredity* 37:215-230.

Charmet, G., F. Balfourier, C. Ravel and J.-B. Denis. 1993. Genotype x environment interactions in a core collection of French perennial rye grass populations. *Theor. Appl. Genet.* 86:731-736.

Cornelius, P.L. 1993. Statistical tests and retention of terms in the additive main effects and multiplicative interaction model for cultivar trials. *Crop Sci* 33:1186-1193.

Corsten, L.C.A., and J.-B. Denis. 1990. Structuring interaction in two-way tables by clustering. *Biometrics* 46:207-215.

Crossa, J., W.H. Pfeiffer, P.N. Fox and S. Rajaram. 1990. Multivariate analysis for classifying sites: application to an international wheat yield trial. p. 214-233. *In* M.S. Kang (ed.) *Genotype-by-environment interaction and plant breeding.* Louisiana State Univ., Baton Rouge.

Davies, P.T., and M.K.-S. Tso. 1982. Procedures for reduced-rank regression. *Appl. Statist.* 31:244-255.

Decoux, G., and J.-B. Denis. 1991. INTERA, version 3.3, Notice d'utilisation, logiciel pour l'interprétation statistique de l'interaction entre deux facteurs. Laboratoire de Biométrie, INRA, route de St-Cyr, F-78026 Versailles Cédex, 175pp.

Denis, J.-B. 1979. Structuration de l'interaction. *Biom. Praxim.* 19:15-34.

Denis, J.-B. 1980. Analyse de régression factorielle. *Biom.Praxim.* 20:1-34.

Denis, J.-B. 1983. Extension du modèle additif d'analyse de variance par modélisation multiplicative des variances. *Biometrics* 39:849-856.

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

Denis, J.-B. 1991. Ajustement de modèles linéaires et bilinéaires sous contraintes linéaires avec données manquantes. *Rev. Statist. Appl.* 39:5-24.

Denis, J.-B., and T. Dhorne. 1989. Modelling interaction by regression with random coefficients. *Biuletyn Oceny Odmian* 21-22:63-73.

Denis, J.-B., and J.C. Gower. 1992. Biadditive models. Technical report of the Laboratoire de Biométrie, INRA, F-78026 Versailles, 33 pp.

Denis, J.-B., and J.C. Gower. 1994a. Biadditive models. Letter to the editor. *Biometrics* 50:310-311.

Denis, J.-B., and J.C. Gower. 1994b. Asymptotic covariances for the parameters of biadditive Models. *Utilitas Mathematica* 46: 193-205.

Denis, J.-B., et P. Vincourt. 1982. Panorama des méthodes statistiques d'analyse des interactions génotype x milieu. *Agronomie* 2:219-230.

Fakorede, M.A.B., and B.O. Opeke. 1986. Environmental indices for the analysis of genotype x environment interaction in maize. *Maydica* 31:233-243.

Federer, W.T., and B.T. Scully. 1993. A parsimonious statistical design and breeding procedure for evaluating and selecting desirable characteristics over environments. *Theor. Appl. Genet.* 86:612-620.

Finlay, K.W., and G.N. Wilkinson. 1963. The analysis of adaptation in a plant-breeding program. *Aust. J. Agric. Res.* 14:742-754.

Freeman, G.H., and P. Crisp. 1979. The use of related variables in explaining genotype-environment interactions. *Heredity* 42:1-11.

Freeman, G.H., and J.M. Perkins. 1971. Environmental and genotype-environmental components of variability. VIII. Relations between genotypes grown in different environments and measures of these environments. *Heredity* 27:15-23.

Gabriel, K.R. 1978. Least squares approximation of matrices by additive and multiplicative models. *J. R. Statist. Soc.* B 40:186-196.

Gauch, H.G.Jr. 1988. Model selection and validation for yield trials with interaction. *Biometrics* 44:705-715.

Gauch, H.G.Jr. 1990. *MATMODEL version 2.0, AMMI and related analyses for two-way data matrices*. Cornell University, Ithaca, New-York 14853, USA, 69pp.

Gauch, H.G.Jr. 1992. *Statistical analysis of regional yield trials: AMMI analysis of factorial designs*. Elsevier, Amsterdam, 278pp.

Genstat 5 Comittee. 1993. *Genstat 5 release 3, reference manual*. Clarendon Press, Oxford.

Goldstein, H., and R.P. McDonald. 1988. A general model for the analysis of multi level data. *Psychometrika* 53:455-467.

Gorman, D.P., M.S. Kang and M.R. Milam. 1989. Contribution of weather variables to genotype x environment interaction in grain sorghum. *Plant Breeding* 103:299-303.

Gravois, K.A., K.A.K. Moldenhauer and P.C. Rohman. 1990. Genotype-by-environment interaction for rice yield and identification of stable, high-yielding genotypes. p. 181-188. *In* M.S. Kang (ed.) *Genotype-by-environment interaction and plant breeding*. Louisiana State Univ., Baton Rouge.

Hardwick, R.C. 1972. Method of investigating genotype environment and other two factor interactions. *Nature New Biology* 236:191-192.

Hardwick, R.C., and J.T. Wood. 1972. Regression methods for studying genotype-environment interactions. *Heredity* 28:209-222.

Helms, T.C. 1993. Selection for yield and stability among oat lines. *Crop Sci* 33:423-426.

Hinkelmann, K. 1974. Genotype-environment interaction: Aspects on statistical design, analysis and interpretation. Invited paper to the 8th International Biometric Conference, Constanta, Romania.

Horner, T.W., and K.J. Frey. 1957. Methods for determining natural areas for oat varietal recommendations. *Agron. J.* 49:313-315.

Izenman, A.J. 1975. Reduced-rank regression models for the multivariate linear model. *J. Mult. Anal.* 5:248-264.

Kang, M.S. 1988. Interactive BASIC program for calculating stability-variance parameters. *Agron. J.* 80:153.

Kang, M.S. 1989. A new SAS program for calculating stability-variance parameters. *J. Hered.* 80:415.

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 D.P. Gorman. 1989. Genotype x environment interaction in maize. *Agron. J.* 81:662-664.

Kang, M.S., and J.D. Miller. 1984. Genotype x environment interactions for cane and sugar yield and their implications in sugar cane breeding. *Crop Sci* 24:435-440.

Knight, R. 1970. The measurement and interpretation of genotype-environment interactions. *Euphytica* 19:225-235.

Lefkovitch, L.P. 1985. Multi-criteria clustering in genotype-environment interaction problems. *Theor. Appl. Genet.* 70:585-589.

Lin, C.S., and G. Butler. 1988. A data-based approach for selecting locations for regional trials. *Can. J. Plant. Sci.* 68:651-659.

Lin, C.S., and B. Thompson. 1975. An empirical method of grouping genotypes based on a linear function of the genotype-environment interaction. *Heredity* 34:255-263.

Longford, N.T. 1987. A fast algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika* 74:817-827.

Magari, R., and M.S. Kang. 1993. Genotype selection via a new yield-stability statistic in maize yield trials. *Euphytica* 70:105-111.

Mandel, J. 1961. Non-additivity in two-way analysis of variance. *J. Am. Statist. Ass.* 56:878-888.

McCullagh, P., and J.A. Nelder. 1989. *Generalized linear models*. Chapman and Hall, London.

M**cGraw, R.L., P.R. Beuselinck and R.R. Smith.** 1986. Effect of latitude on genotype x environment interactions for seed yield in birdsfoot trefoil. *Crop Sci* 26:603-605.

Mudholkar, G.S., and I.C. Sarkar. 1992. Testing homoscedasticity in a two-way table. *Biometrics* 48:883-888.
Muir, W., W.E. Nyquist and S. Xu. 1992. Alternative partitioning of the genotype-by-environment interaction. *Theor. Appl. Genet.* 84:193-200.
Obadia, J. 1978. L'analyse en composantes explicatives. *Rev. Statist. Appl.* 26:5-28.
Oliveira, J.A., and G. Charmet. 1992. Genotype by environment interaction in Lolium perenne: grouping wild populations by cluster analysis. *Invest. Agr.: Prod. Prot. Veg.* 7:117-128.
Oman, S.D. 1991. Multiplicative effects in mixed model analysis of variance. *Biometrika* 78:729-739.
Paul, H., F.A. van Eeuwijk and W. Heijbroek. 1993. Multiplicative models for cultivar by location interaction in testing sugar beets for resistance to beet necrotic yellow vein virus. *Euphytica* 71:63-74.
Rameau, C., and J.-B. Denis. 1992. Characterization of environments in long-term multi-site trials in Asparagus, through yield of standard varieties and use of environmental covariates. *Plant Breeding* 109:183-192.
Rao, C.R. 1964. The use and interpretation of principal component analysis in applied research. *Sankhya* B 26:329-358.
Royo, C., A. Rodriguez and I. Romogosa. 1993. Differential adaptation of complete and substituted triticale. *Plant Breeding* 111:113-119.
Russel, T.S., and R.A. Bradley. 1958. One-way variances in a two-way classification. *Biometrika* 45:111-129.
Sabatier, R., J.-D. Lebreton and D. Chessel. 1989. Principal component analysis with instrumental variables as a tool for modelling composition data. p. 341-352. *In* R. Coppi and S. Bolasco (eds) *Multiway Data Analysis.* Elsevier, North-Holland.
Saeed, M., and C.A. Francis. 1984. Association of weather variables with genotype x environment interactions in grain sorghum. *Crop Sci* 24:13-16.
SAS Institute Inc., SAS*. 1992. *Software: changes and enhancements, release 6.07*. Technical report P-229, SAS/STAT*, Cary, NC: SAS Institute Inc., 620 pp.
Searle, S.R. 1979. Alternative covariance models for the 2-way crossed classification. *Comm. St.* A 8:799-818.
Searle, S.R., G. Casella and C.E. McCulloch. 1992. *Variance Components*. Wiley, New York.
Seif, E., J.C. Evans and L.N. Balaam. 1979. A multivariate procedure for classifying environments according to their interaction with genotypes. *Aust. J. Agric. Res.* 30:1021-1026.
Shukla, G.K. 1972a. An invariant test for the homogeneity of variances in a two-way classification. *Biometrics* 28:1063-1072.
Shukla, G.K. 1972b. 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.
Snedecor, G.W., and W.G. Cochran. 1976. *Statistical Methods* (6th edn., 8th printing). Iowa State Univ. Press, Ames.
Snee, R.D. 1982. Nonadditivity in a two-way classification: is it interaction or nonhomogeneous variance?. *J. Am. Statist. Ass.* 77:515-519.
S-plus. 1994. *S-plus for Windows version 3.2 supplement*. Seattle: StatSci, a division of MathSoft, Inc.
Tai, G.C.C. 1990. Path analysis of genotype-environment interactions. p. 273-286. *In* M.S. Kang (ed.) *Genotype-by-environment interaction and plant breeding*. Louisiana State Univ., Baton Rouge.
Talbot, M., and A.V. Wheelwright. 1989. The analysis of genotype x environment interactions by partial least squares regression. *Biuletyn Oceny Odmian* 21-22:19-25.
Tso, M.K.-S. 1981. Reduced-rank regression and canonical analysis. *J. R. Statist. Soc.* B 43:183-189.
van den Wollenberg, A.L. 1977. Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika* 42:207-219.
van der Leeden, R. 1990. Reduced rank regression with structured residuals. DSWO Press, Leiden.
van Eeuwijk, F.A. 1992a. Interpreting genotype-by-environment interaction using redundancy analysis. *Theor. Appl. Genet.* 85:89-100.
van Eeuwijk, F.A. 1992b. Multiplicative models for genotype-environment interaction in plant breeding. *Statistica Applicata* 4:393-406.
van Eeuwijk, F.A. 1993. Genotype by environment interaction; Basic ideas and selected topics. p. 91-104. *In* Johan H.L. Oud and Rian A.W. van Blokland-Vogelesang (eds.) *Advances in longitudinal and multivariate analysis in the behavioral sciences*. ITS, Nijmegen.
van Eeuwijk, F.A. 1995a. Linear and bilinear models for the analysis of multi-environment trials: I. An inventory of models. *Euphytica*, in press.
van Eeuwijk, F.A. 1995b. Multiplicative interactions in generalized linear models. *Biometrics* in press.

van Eeuwijk, F.A., and A. Elgersma. 1993. Incorporating environmental information in an analysis of genotype by environment interaction for seed yield in perennial ryegrass. *Heredity* 70:447-457.

van Eeuwijk, F.A., L.C.P. Keizer and J.J. Bakker. 1995. Linear and bilinear models for the analysis of multi-environment trials : II. An application to data from the Dutch Maize Variety Trials. *Euphytica*, in press.

van Eeuwijk, F.A., and P.M. Kroonenberg. 1995. Multiplicative decompositions of interactions in three-way analysis of variance, with applications to plant breeding. Submitted.

Velu, R.P. 1991. Reduced rank models with two sets of regressors. *Appl. Statist.* 40:159-170.

Vincourt, P., M. Derieux and A. Gallais. 1984. Quelques méthodes de choix des génotypes à partir d'essais multilocaux. *Agronomie* 4:843-848.

Williams, E.R., and J.T. Wood. 1993. Testing the significance of genotype-environment interaction. *Aust. J. Agric. Res.* 35:359-362.

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. Pflanzenzücht.* 47:92:96.