Using Partial Least Squares Regression, Factorial Regression, and AMMI Models for Interpreting Genotype × Environment Interaction

Mateo Vargas, José Crossa,* Fred A. van Eeuwijk, Martha E. Ramírez, and Ken Sayre

ABSTRACT

Partial least squares (PLS) and factorial regression (FR) are statistical models that incorporate external environmental and/or cultivar variables for studying and interpreting genotype × environment interaction (GEI). The Additive Main effect and Multiplicative Interaction (AMMI) model uses only the phenotypic response variable of interest; however, if information on external environmental (or genotypic) variables is available, this can be regressed on the environmental (or genotypic) scores estimated from AMMI and superimposed on the AMMI biplot. The objectives of this study with two wheat [*Triticum turgidum* (L.) var. *durum*] field trials were (i) to compare the results of PLS, FR, and AMMI on the basis of external environmental (and cultivar) variables, (ii) to examine whether procedures based on PLS, FR, and AMMI identify the same or a different subset of cultivar and/or environmental covariables that influence GEI for grain yield, and (iii) to find multiple FR models that include environmental and cultivar covariables and their cross products that explain a large proportion of GEI with relatively few degrees of freedom. Results for the first trial showed that AMMI, PLS, and FR identified similar cultivar and environmental variables that explained a large proportion of the cultivar × year interaction. Results for the second wheat trial showed good correspondence between PLS and FR for 23 environmental covariables. For both trials, PLS and FR complement each other and the AMMI and PLS biplots offered similar interpretations of the GEI. The FR analysis can be used to confirm these results and to obtain even more parsimonious descriptions of the GEI.

Multi-environment trials play an important role in selecting the best cultivars (or agronomic practices) to be used in future years at different locations and in assessing a cultivar's stability across environments before its commercial release. When the performance of cultivars is compared across sites, several cultivar attributes are considered, of which grain yield is one of the most important. Cultivars grown in multi-environment trials react differently to environmental changes. This differential response of cultivars from one environment to another is called genotype × environment interaction (GEI).

Genotype × environment interaction has been studied, described, and interpreted by means of several statistical models (Crossa, 1990). Some models, such as analysis of variance, regression on the environmental mean models (Yates and Cochran, 1938; Finlay and Wilkinson, 1963; Eberhart and Russell, 1966), and the Additive Main effects and Multiplicative Interaction (AMMI) models (Gollob, 1968; Mandel, 1971; Kempton, 1984; Gauch, 1988) use only the phenotypic response variable of interest. The AMMI model is more parsimonious than the conventional analysis of variance model in describing GEI and provides greater scope for modeling and interpreting GEI than the simple regression on the site mean because GEI can be modeled in more than one dimension.

When information on external environmental (or genotypic) variables such as meteorological data or soil information is available, these variables can be correlated to or regressed on the environmental (genotypic) scores estimated by AMMI. Information from these regressions can be superimposed on the AMMI biplot along with cultivar and environmental scores (van Eeuwijk, 1995), so that interpretation of the grain yield GEI is possible. External environmental information cannot be used directly in the AMMI model, however. When additional information on external cultivar variables is available (physiology, maturity, disease reaction, genetic markers, etc.), other statistical models, including factorial regression models (FR) (Denis, 1988; van Eeuwijk et al., 1996) and partial least squares (PLS) regression (Aastveit and Martens, 1986; Talbot and Wheelwright, 1989; Vargas et al., 1998) can be used to determine which of these external environmental or cultivar variables influence GEI of grain yield.

Factorial regression models are ordinary linear models that explain GEI by differential cultivar sensitivity to explicit external environmental variables (environmental characterization) and have the advantage that hypotheses about the influence of those external variables on GEI of grain yield can be statistically tested. As with all linear regression models, factorial regression models become difficult to deal with when there are many explanatory variables that are highly correlated—the multi-collinearity problem. PLS regression models are appropriate for these situations. As in factorial regression, PLS regression describes GEI in terms of dif-

M. Vargas, Programa de Estadística del Instituto de Socioeconomía, Estadística e Informática (ISEI), Colegio de Postgraduados, CP 56230, Montecillo, Mexico, Universidad Autónoma Chapingo, CP 56230, Chapingo, Mexico, and International Maize and Wheat Improvement Center (CIMMYT), Lisboa 27, Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico; J. Crossa, Biometrics and Statistics Unit, CIMMYT, Lisboa 27, Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico; F.A. van Eeuwijk, Dep. of Agricultural, Environmental and Systems Technology, Wageningen Agricultural Univ., Dreijenlaan 4, 6703 HA Wageningen, the Netherlands; M.E. Ramírez, Programa de Estadística del Instituto de Socioeconomía, Estadística e Informática (ISEI), Colegio de Postgraduados, CP 56230, Montecillo, Mexico; Ken Sayre, Wheat Program, CIMMYT, Lisboa 27, Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico. Received 2 July 1998. *Corresponding author (JCROSSA@CIMMYT.MX).

Published in Crop Sci. 39:955–967 (1999).

Abbreviations: AMMI, Additive Main effect and Multiplicative Interaction; GEI, genotype × environment interaction; FR, factorial regression; PLS, partial least squares.

ferential sensitivity of cultivars to environmental variables. The difference is that the explanatory variables are hypothetical, synthetic variables (linear combinations of the complete set of measured environmental and/or cultivar variables) and there is no limit to the number of explanatory covariables that can be used. The PLS regression models are not linear models, so standard linear regression theory for testing cannot be used; however, good alternatives are available.

The advantages and/or disadvantages of the above mentioned statistical models for studying and interpreting GEI with a large number of external environmental and/or cultivar variables have not been compared. Therefore, the objectives of this study were to: (i) compare the results from AMMI, PLS, and FR in two wheat trials when a large set of external environmental (and cultivar) covariables are available, (ii) examine whether procedures based on AMMI, FR, and PLS identify the same or different subsets of cultivar and/or environmental covariables that explain GEI for grain yield, and (iii) find more parsimonious multiple FR models that include environmental and cultivar covariables and their cross products that explain large proportion of GEI with relatively few degrees of freedom.

MATERIALS AND METHODS

Theory

van Eeuwijk (1996) gave a comprehensive description of the AMMI and FR models and how to apply them to assess, study, and interpret GEI. Vargas et al. (1998) described the theory of PLS in the context of GEI and detailed its algorithm. The AMMI, FR, and PLS models are briefly described here.

AMMI Model and the Biplot. A basic model for the analysis of the two-way table of cultivar yield by environment data is the analysis of variance model:

$$E(y_{ij}) = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} \qquad [1]$$

where E stands for expectation, μ is the grand mean, τ_i is the main effect of the ith cultivar, β_j is the main effect of the jth site, and $(\tau\beta)_{ij}$ is the GEI effect of the ith cultivar in the jth environment.

Model [1] can be written in matrix notation as:

$$E(\mathbf{Y}) = \mu\mathbf{1}_I\mathbf{1}_J' + \boldsymbol{\tau}\mathbf{1}_J' + \mathbf{1}_I\boldsymbol{\beta}' + \boldsymbol{\tau\beta} \qquad [2]$$

where $\mathbf{Y} = (y_{ij})$ is the data matrix of size I×J of grain yield of I cultivars in J environments, μ is a scalar representing the grand mean, $\boldsymbol{\tau} = (\tau_i)$ is a I×1 vector of main effects of cultivars, $\boldsymbol{\beta} = (\beta_j)$ is a J×1 vector of main effects of sites, and $\boldsymbol{\tau\beta} = (\tau\beta)_{ij}$ is the I×J interaction matrix (not a vector product) where each element of the matrix specifies the interaction effect for the ith cultivar in the jth site. $\mathbf{1}_I$ and $\mathbf{1}_J$ are unit vectors of size I×1 and J×1, respectively. The common constraints are $\mathbf{1}_I'\ \boldsymbol{\tau} = \mathbf{1}_J'\boldsymbol{\beta} = 0$ and $\mathbf{1}_I'\boldsymbol{\tau\beta}\ \mathbf{1}_J' = 0$.

As mentioned previously, a commonly used procedure for modeling GEI is the simple regression of cultivar performance on the environment mean such that $(\tau\beta)_{ij} = \zeta_i\beta_j + d_{ij}$, where ζ_i measures the sensitivity of the ith cultivar to prevailing environmental conditions in the multiplicative (bilinear) term $\zeta_i\beta_j$ and d_{ij} is the residual term (Yates and Cochran, 1938; Finlay and Wilkinson, 1963; Eberhart and Russell, 1966). This model can be depicted as a set of straight lines with different slopes, one for each cultivar, where the heterogeneity of slopes accounts for the GEI. Since heterogeneity of slopes in this model generally explains only a small proportion of the usually complex GEI, a more elaborate model is often necessary for an adequate description of GEI.

A generalization of the regression on the site mean model is the multiplicative (bilinear) model

$$E(y_{ij}) = \mu + \tau_i + \beta_j + \textstyle\sum_{k=1,\,K} \lambda_k\ \theta_{ik}\ \gamma_{jk} \qquad [3]$$

also called Principal Component Analysis (PCA) of the GEI or Additive Main effect and Multiplicative Interaction (AMMI) model (Gollob, 1968; Mandel, 1971; Kempton, 1984; Gauch, 1988) or biadditive model (Denis and Gower, 1994). The parameters μ, τ_i, and β_j are the same as in the analysis of variance model. K being the number of multiplicative (bilinear) terms in the model. The λ_k are scaling constants obtained from the singular value decomposition of the residual matrix consisting of the two-way table of means corrected for cultivar and site main effects (residual from additivity), $w_{ij} = \bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$, (Gabriel, 1978) and are ordered such that $\lambda_k \geq \lambda_{k+1}$. The θ_{ik} are cultivar interaction parameters (or scores) that measure cultivar sensitivity to hypothetical environmental factors denoted by environmental interaction parameters γ_{jk} (or scores). Orthonormality constraints for the cultivar and environmental scores are $\Sigma_i\ \theta_{ik}^2 = \Sigma_j\ \gamma_{jk}^2 = 1$ and $\Sigma_i\theta_{ik}\ \theta_{ik'} = \Sigma_j\ \gamma_{jk}\ \gamma_{jk'} = 0$ for $k \neq k'$.

For determining the number of multiplicative terms to be retained in a multiplicative model, various tests can be used. The F-test of Gollob (1968) uses the ratio between the mean square for axis k against an estimate of the error term. The mean squares of axis k is calculated by taking the square of λ_k, and the corresponding degrees of freedom, computed by $(I - 1) + (J - 1) - (2k - 1)$. Using simulation studies, Cornelius (1993) showed that the Gollob's F-test is very liberal. Thus, in this study we have used the approximate F-tests F_{GH1}, F_R, and F_1 (Cornelius et al., 1993, 1996; Cornelius, 1993).

Model [3] written in matrix notation is

$$E(\mathbf{Y}) = \mu\mathbf{1}_I\mathbf{1}_J' + \boldsymbol{\tau}\mathbf{1}_J' + \mathbf{1}_I\boldsymbol{\beta}' + \boldsymbol{\Theta}\ \boldsymbol{\Lambda}\ \boldsymbol{\Gamma}' \qquad [4]$$

where the first three terms on the right side are the same as in Eq. [2]. The fourth term represents the GEI, where $\boldsymbol{\Theta} = (\theta_{ik})$ is a $I\times K$ matrix , $\boldsymbol{\Lambda} = (\lambda_{kk})$ is a $K\times K$ diagonal matrix and $\boldsymbol{\Gamma} = (\gamma_{jk})$ is a $J\times K$ matrix. The normalization and orthogonality constraints are $\mathbf{1}_I'\boldsymbol{\tau} = \mathbf{1}_J'\ \boldsymbol{\beta} = 0$, $\mathbf{1}_I'\boldsymbol{\Theta} = \mathbf{1}_J'\ \boldsymbol{\Gamma} = \mathbf{0}$, where $\mathbf{0}$ is a matrix of zeros of size $1\times K$, and $\boldsymbol{\Theta}'\boldsymbol{\Theta} = \boldsymbol{\Gamma}'\boldsymbol{\Gamma} = \mathbf{I}_K$. The kth bilinear term, $k = 1, ..., K$, is formed by a score θ_{ik} specific to Cultivar i, a scale constant factor λ_{kk} and a score γ_{jk} specific to Site j.

The results of AMMI analysis can be presented graphically in the form of biplots (Gabriel, 1971) in which the cultivar and environment scores of the first two or three bilinear (multiplicative) terms are represented by vectors in a space, with starting points at the origin and end points determined by the scores. Usually the environmental and cultivar scores of the first and second bilinear terms are plotted. The distance between two cultivar vectors (their end points) is indicative of the amount of interaction between the cultivars. The cosine of the angle between two cultivar (or environment) vectors approximates the correlation between the cultivars (or environments) with respect to their interaction. Acute angles indicate positive correlation, with parallel vectors (in exactly the same directions) representing a correlation of 1. Obtuse angles represent negative correlations, with opposite directions indicating a correlation of −1. Perpendicularity of directions indicates a correlation of 0. The relative amounts of interaction for a particular cultivar over environments can be obtained from orthogonal projections of the environmental vectors on the line determined by the direction of the corresponding cultivar vector. Environmental vectors having the same direction as the cultivar vectors have positive interactions (that is, these environments favored these cultivars), whereas vectors in the opposite direction have negative interactions.

The biplot obtained from AMMI can be enriched by a procedure described by van Eeuwijk (1995). Information on external environmental (or cultivar) variables can be correlated to or regressed on the environmental or cultivar interaction parameters (θ_{ik} or γ_{jk}) estimated from AMMI and incorporated into the biplot so that a better interpretation of the GEI of grain yield can be attempted. Once it has been decided that the AMMI solution has, for example, two axes for interaction, the squared correlation coefficients from the regressions of the covariables on the scores of both axes simultaneously (regression through the origin) are computed. When this squared correlation is sufficiently high, information for the covariables can be drawn in the biplot by giving them a direction that is determined by the regression coefficients (van Eeuwijk, 1995). For example, if one environmental covariable is regressed on the AMMI environmental scores of Axes 1 and 2, then coefficients b_{axis1} and b_{axis2} serve as coordinates for that covariable in the biplot. When the environments are projected in the direction determined by the regression coefficients (b_{axis1}, b_{axis2}), it gives the rank-order of the environments on this environmental covariable. The same can also be done for cultivar covariables.

Factorial Regression Models. Factorial Regression models also have multiplicative structure for the interaction, like the AMMI model. The main difference between FR models and multiplicative models such as AMMI model is that in FR the GEI (residual matrix consisting of the two-way table of means corrected for cultivar and site main effects, $\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{j.} + \bar{y}_{..}$) is modeled directly as a function of the cultivar and environmental variables. A factorial regression model for the mean of the ith cultivar in the jth environment, for which the interaction includes the cultivar covariables x_{i1} to x_{iG}, is

$$E(y_{ij}) = \mu + \tau_i + \beta_j + \textstyle\sum_{g=1,\,G} x_{ig}\, \xi_{jg} \qquad [5]$$

The parameters μ, τ_i, and β_j are the same as in Eq. [1] and [3]. The GEI consists of the products of the environmental factors ξ_{j1} to ξ_{jG} with respect to cultivar covariables x_{i1} to x_{iG} ($G \leq I - 1$). The cultivar covariables are known, but the environmental potentialities have to be estimated. In matrix notation, Model [5] can be written as:

$$E(\mathbf{Y}) = \mu\mathbf{1}_I\mathbf{1}_J' + \boldsymbol{\tau}\mathbf{1}_J' + \mathbf{1}_I\boldsymbol{\beta}' + \mathbf{X}\boldsymbol{\Xi}' \qquad [6]$$

where the first three terms on the right side are the same as before, $\mathbf{X} = (x_{ig})$ is a $I{\times}G$ matrix of known cultivar covariables and $\boldsymbol{\Xi} = (\xi_{ig})$, a $J{\times}G$ matrix of unknown environmental constants, and G is the number of cultivar covariables.

An FR model in which the interaction part contains the environmental covariables z_{j1} to z_{jH} can be written as:

$$E(y_{ij}) = \mu + \tau_i + \beta_j + \textstyle\sum_{h=1,\,H} \zeta_{ih}\, z_{jh} \qquad [7]$$

The GEI term in this model allows the cultivars to have different sensitivities, ζ_{i1} to ζ_{jH} ($H \leq J - 1$), to the environmental covariables. The values of the environmental variables are known, but the cultivar sensitivities need to be estimated. Similar to Model [5], Model [7] can be written as:

$$E(\mathbf{Y}) = \mu\mathbf{1}_I\mathbf{1}_J' + \boldsymbol{\tau}\mathbf{1}_J' + \mathbf{1}_I\,\boldsymbol{\beta}' + \boldsymbol{\zeta}\,\mathbf{Z}' \qquad [8]$$

where $\mathbf{Z} = (z_{jh})$ is the $J{\times}H$ matrix of environmental covariables and $\boldsymbol{\zeta} = (\zeta_{jh})$ is a $I{\times}H$ matrix of cultivar sensitivities, and H is the number of environmental covariables.

The structure of the FR model including both cultivar and environmental covariables simultaneously is similar to that of Models [5] and [7] (Denis, 1988; van Eeuwijk et al., 1996). In matrix notation it can be written as:

$$E(\mathbf{Y}) = \mu\mathbf{1}_I\mathbf{1}_J' + \boldsymbol{\tau}\mathbf{1}_J' + \mathbf{1}_I\boldsymbol{\beta}' + \mathbf{X}\boldsymbol{\nu}\mathbf{Z}' + \mathbf{X}\,\boldsymbol{\Xi}' + \boldsymbol{\zeta}\mathbf{Z}', \qquad [9]$$

where in the new term $\mathbf{X}\boldsymbol{\nu}\mathbf{Z}'$, $\boldsymbol{\nu}$ is a $G{\times}H$ matrix of regression coefficients to cross-products of cultivar and environmental covariables. General identification constraints for factorial regressions with already centered covariables are $\boldsymbol{\zeta}'\mathbf{X} = \mathbf{Z}'\boldsymbol{\Xi} = \mathbf{0}$, where $\mathbf{0}$ is now a matrix of zeros of order $H{\times}G$. Covariables may be quantitative and qualitative, and more complicated FR models are possible by combining quantitative and qualitative covariables.

In this study the FR procedure was implemented in GENSTAT version 5, release 3.2 (GENSTAT, 1995). The stepwise procedure implemented in Genstat for the multiple linear regression, selects a term to be included or excluded from the model based on an F-test. For example, for X_1, X_2, X_3, and X_4, explanatory variables, the procedure starts by fitting a model containing variable X_1. Then it attempts to drop X_1 and to add, one at the time, X_2, X_3, and X_4. The procedure permanently modifies the current model according to the change that was most successful; if dropping X_1 improves the model, then X_1 is permanently removed; or, when no removals are worthwhile, if adding X_2 gives the biggest improvement, then X_2 is permanently included. The stepwise procedure allows for forward selection or backward elimination.

Partial Least Squares Regression. In many situations, when the number of variables (S) is much larger than the number of observations (N), and there is high collinearity among variables, the usual methods for fitting regressions based on ordinary least squares are not adequate. In this situation, partial least squares regression seems to be a more appropriate alternative. Details of PLS theory (Helland, 1988) and its similarities to principal components regression and stepwise multiple linear regression are described in Aastveit and Martens (1986). A description of univariate and multivariate PLS and their algorithms was given in Vargas et al. (1998). In this paper, the multivariate PLS algorithm, the cross validation procedure, and the F-test were applied by a procedure implemented in GENSTAT version 5, release 3.2 (GENSTAT, 1995).

For the standard situation where multivariate PLS is used to model cultivar responses ($\mathbf{Y}$) over environments on environmental covariables ($\mathbf{Z}$), the corresponding bilinear forms are $\mathbf{Z} = \mathbf{TP}' + \mathbf{E}$ and $\mathbf{Y} = \mathbf{TQ}' + \mathbf{F}$, respectively, where matrix $\mathbf{T}$ contains the Z-scores, matrix $\mathbf{P}$ contains the Z-loadings, matrix $\mathbf{Q}$ contains the Y-loadings, and $\mathbf{E}$ and $\mathbf{F}$ are the residual matrices. It is easiest to work with the transpose of $\mathbf{Y}$: $\mathbf{Y}'$such that the columns of $\mathbf{Y}'$ (i.e., the rows of $\mathbf{Y}$) contain cultivar responses over environments. Then $\mathrm{E}(\mathbf{Y}') = (\mathbf{TQ}')' = \mathbf{QW}'\mathbf{Z}' = \boldsymbol{\zeta}\mathbf{Z}'$, where $\mathbf{T}$ contains the Z-scores (indexed by environments), $\mathbf{W}$ the Z-loadings (or weights, indexed by environmental variables), and $\mathbf{Q}$ the Y-loadings (indexed by genotypes). $\boldsymbol{\zeta}$ contains the PLS approximation to the regression coefficients of the responses in $\mathbf{Y}'$ (genotypic responses) to the explanatory variables in $\mathbf{Z}$ (environmental variables). Note that $\boldsymbol{\zeta}\mathbf{Z}'$ is the same as the last term of Eq. [8]. From this formulation, it can be deduced which biplots can be constructed to summarize PLS analyses. The set $\mathbf{T}$, $\mathbf{W}$, and $\mathbf{Q}$ can be depicted in the same biplot; the rows of matrix $\mathbf{T}$ contain the coordinates for environments, the rows of $\mathbf{W}$ contain the coordinates for environmental covariables and the rows of $\mathbf{Q}$ contain the coordinates for cultivars. Projection of the jth row of $\mathbf{T}$ on the ith row of $\mathbf{Q}$ (or vice versa) approximates the interaction of the ith genotype on the jth environment: $\mathbf{Y}' = (\mathbf{TQ}')'$. Projection of the hth row of $\mathbf{W}$ on the ith row of $\mathbf{Q}$ (or vice versa) approximates the regression coefficient of the ith genotype on the hth environmental covariable: $\mathbf{QW}' = \boldsymbol{\zeta}$. Thus, the PLS biplot including representations of cultivars, environments and covariables allows the same types of interpretation to be made as the enriched AMMI biplot introduced earlier.

When environmental responses over cultivars are modeled on cultivar covariables, $E(\mathbf{Y}) = \mathbf{TQ}' = \mathbf{XWQ}' = \mathbf{X}\boldsymbol{\Xi}'$ (the same as the last term of Eq. [6]) the rows of $\mathbf{T}$ will contain coordinates for the cultivars, the rows of $\mathbf{W}$ will contain coordinates for the cultivar covariates, and the rows of **Q** will contain coordinates for the environments. $\boldsymbol{\Xi}'$ contains the PLS approximation to the regression coefficients of the responses in **Y** to the explanatory variables in **X**. Biplot relations follow from $\mathbf{Y} = \mathbf{TQ}'$ and $\mathbf{WQ}' = \boldsymbol{\Xi}'$.

Experimental Data

Durum Wheat Variety Trial (Data Set 1). This data set, used by Vargas et al. (1998), consisted of one experiment with seven durum wheat cultivars tested for 6 yr (1990–1995) in Ciudad Obregón, Mexico. The cultivars included were a historical set released from the early 1960s to the late 1980s; the order of Numbers 1 to 7 is the order of cultivar releases over time (Sayre et al., 1997).

The cultivar variables, ***X***, were days to anthesis after emergence (ANT), days to maturity after emergence (MAT), days of grainfill (GFI), plant height (cm) (PLH), above-ground biomass (kg ha^{-1}) (BIO), harvest index (HID), straw yield (kg ha^{-1}) (STW), number of spikes per square meter (NSM), number of grains per square meter (NGM), number of grains per spike (NGS), thousand-kernel weight (g) (TKW), weight per tiller (g) (WTI), spike grain weight (g) (SGW), vegetative growth rate (kg ha^{-1} d^{-1}) (VGR), and individual kernel growth rate (mg kernel^{-1} d^{-1}) (KGR) during the grainfill period. The environmental variables, **Z**, measured from December to March of each year were mean daily maximum temperature (°C) (MT), mean daily minimum temperature (°C) (mT), monthly total precipitation (mm) (PR), and sun hours per day (SH).

Wheat Agronomic Trial (Data Set 2). This data set consisted of one wheat experiment including several treatments for cultural practices, conducted over 10 yr (1988–1997) in Ciudad Obregón, Mexico. Each year the experiment was arranged in a randomized complete block design with three replicates. Treatments were obtained by combining four factors at the following levels: tillage at 2 levels (1 = with deep knife, 2 = without deep knife), summer crop at 2 levels [1 = sesbania, (*Sesbania* spp.) 2 = soybean (*Glycine max* (L.) Merr.)], manure at 2 levels (1 = with chicken manure, 2 = without chicken manure), and nitrogen fertilization rate at 3 levels (1 = 0 kg N ha^{-1}, 2 = 100 kg N ha^{-1} and 3 = 200 kg N ha^{-1}), resulting in 2 × 2 × 2 × 3 = 24 treatments. Therefore, Treatment 1 is [1–1–1–1], Treatment 2 is [2–1–1–1], Treatment 3 is [1–2–1–1], and so on up Treatment 24 [2–2–2–3].

Data matrix **Y** had 24 rows (treatments) and 10 columns (years). Matrix **Y** had grain yield interaction residuals ($\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{j.} + \bar{y}_{..}$). The 27 explanatory variables in the **Z** matrix of size 10 × 27 (years × environmental variables) were mean minimum temperature sheltered (°C) (mT), mean minimum temperature unsheltered (°C) (mTU), mean maximum temperature sheltered (°C) (MT), total monthly precipitation (mm) (PR), mean sun hours per day (SH), and total monthly evaporation (mm) (EV). Environmental variables were measured from December through April of each year. All covariables were centered and standardized prior to analysis. For reasons of consistency with earlier analyses (Vargas et al., 1998), the columns of the **Y** matrix were standardized.

RESULTS

Data Set 1

AMMI Analysis, Biplot and Correlations

The combined analysis of variance across years showed that 66% of variation among the 42 cultivar × year combinations was explained by differences among cultivar means, 22% by differences between year means, and 5% by cultivar × year interaction (Table 1). AMMI analysis of variance indicated that the first multiplivative term was significant ($P < 0.05$) by the F_{GH1} test (Cornelius et al., 1993, 1996; Cornelius, 1993) and the first two multiplicative terms were significant ($P < 0.05$) by the F_1 and F_R tests (Cornelius et al., 1993, 1996; Cornelius, 1993). The first bilinear interaction term of the AMMI analysis of variance accounted for 65% of the cultivar × year interaction sum of squares with 10 degrees of freedom and the second for 15% with 8 degrees of freedom. The first two bilinear terms accounted for 80% of the interaction, indicating that with only 18 degrees of freedom, from the 30 degrees of freedom contained in the analysis of variance cultivar × year interaction, a considerable amount of the GEI was explained (Table 1).

To investigate relationships between additive and multiplicative parameters in the AMMI model and the values of the cultivar and environmental covariables, correlations coefficients were calculated between the cultivar mean grain yields and each of the cultivars covariables. Similarly, the environment means for grain yield were correlated with each of the environmental covariable. The coefficients of determination (R^2) for the regression of the standardized cultivar and environmental covariables on the cultivar and environmental scores of the first two bilinear terms (scores of Axes 1 and 2) were also computed. The cultivar main effect was highly positively correlated with number of grains per square meter (NGM), number of grains per spike (NGS), harvest index (HID), spike grain weight (SGW), and above-ground biomass (BIO), and was highly negatively correlated with individual kernel growth rate (KGR) (Table 2). The R^2 values of the regressions of these cultivar variables on the scores of AMMI Axes 1 and 2 were also high. The environmental main effect was negatively correlated with the environmental variables, minimum temperature in December (mTD), January (mTJ), and February (mTF) and precipitation in February (PRF), and positively correlated with maximum temperature

Table 1. AMMI analysis of variance for Data Sets 1 and 2. Data Set 1 consisted of one experiment with seven durum wheat cultivars tested for 6 yr (1990–1995) at Ciudad Obregón, Mexico. The cultivars included were a historical set released from the early 1960s to the late 1980s. Data Set 2 consisted of one wheat experiment including several treatments for cultural practices, conducted over 10 yr (1988–1997) at Ciudad Obregón, Mexico.

Source	df	Sum of squares (× 10^6)†	Mean squares (× 10^5)†	Prob
		Data set 1		
Cultivar	6	183.740	306.233	0.0001
Year	5	62.624	125.248	0.0001
Cultivar × year	30	14.547	4.849	0.0001
Bilinear term 1	10	9.549	9.549	0.0001
Bilinear term 2	8	2.238	2.797	0.0679
Deviation	12	2.760	2.300	0.1169
Error	72	10.410	1.446	
		Data set 2		
Treatment	23	773.970	336.508	0.0001
Year	9	373.260	414.733	0.0001
Year × treatment	207	279.520	13.503	0.0001
Bilinear term 1	31	151.130	48.751	0.0001
Bilinear term 2	29	39.112	13.486	0.0001
Bilinear term 3	27	36.781	13.622	0.0001
Deviations	120	52.497	4.374	0.0001
Error	460	110.870	2.410	

† Actual values multiplied by this factor to obtain reported values.

in December (MTD) and January (MTJ) and sun hours in January (SHJ) and February (SHF).

The AMMI biplot (Fig. 1a) separates the high yielding years, 1990, 1991, and 1994, from the low yielding years, 1993 and 1995, along the fist axis from left to right. With respect to cultivars, along the horizontal first axis, earlier released cultivars, 1 and 2, are separated from intermediate and later released cultivars, 3, 4, 5, 6, and 7. Cultivars 1 and 2 were positively influenced by environmental conditions in 1990, 1991, and 1994 and negatively influenced by environmental conditions in 1993 and 1995. Cultivars 5 and 6 were favored in 1995 and, to some extent, in 1993, while they were negatively influenced by environmental conditions prevailing in 1990, 1991, 1992, and 1994.

Table 2. Correlations coefficients (*r*) between cultivar covariables versus cultivar mean grain yield and environmental covariables versus environmental mean grain yield, and the proportion of variation explained in each cultivar and environmental variable by the regression of the cultivar and environmental covariables on the scores of the first two AMMI bilinear terms (R^2) for Data Sets 1 and 2.

Data set 1					
Cultivar†			Environmental‡		
Covariable	Correlation	R^2	Covariable	Correlation	R^2
HID	0.94**	0.91	PRF	−0.66	0.95
NGS	0.96**	0.90	MTM	−0.28	0.87
NGM	0.99**	0.88	mTJ	−0.81*	0.85
SGW	0.90**	0.85	SHM	−0.43	0.78
KGR	−0.82*	0.84	SHF	0.68	0.75
PLH	0.71	0.74	SHJ	0.70	0.75
BIO	0.89**	0.66	MTJ	0.55	0.72
ANT	−0.48	0.61	PRJ	−0.54	0.67
WTI	0.47	0.58	PRM	−0.36	0.64
MAT	−0.45	0.58	SHD	0.45	0.46
TKW	−0.67	0.41	MTD	0.76	0.43
VGR	0.63	0.41	MTF	0.12	0.32
NSM	0.10	0.36	mTF	−0.75	0.30
STW	0.01	0.18	mTD	−0.60	0.29
GFI	0.14	0.10	PRD	−0.34	0.08
			mTM	−0.31	0.04

Data set 2					
Environmental‡			Environmental‡		
Covariable	Correlation	R^2	Covariable	Correlation	R^2
EVA	0.17	0.68	MTA	−0.15	0.32
mTM	−0.33	0.67	mTD	−0.30	0.28
mTUM	−0.24	0.62	SHD	0.23	0.25
EVD	0.54	0.58	mTJ	−0.20	0.25
MTF	−0.01	0.53	mTUD	−0.20	0.25
EVJ	0.42	0.52	MTD	0.59*	0.24
SHJ	0.28	0.51	MTJ	0.31	0.24
mTUF	−0.34	0.49	PRM	−0.29	0.22
mTF	−0.41	0.48	mTUA	0.27	0.14
EVF	0.47	0.45	PRD	−0.50	0.14
EVM	0.19	0.36	mTA	0.43	0.08
PRF	−0.37	0.35	SHF	−0.30	0.20
PRJ	−0.25	0.33	MTM	−0.13	0.00
mTUJ	−0.03	0.32			

*, ** Significantly different from zero at the 0.05 and 0.01 levels of probability, respectively.

† HID = harvest index, NGS = number of grains per spike, NGM = number of grains per square meter, SGW = spike grain weight, KGR = individual kernel growth rate, PLH = plant height, BIO = above-ground biomass, ANT = days to anthesis after emergence, WTI = weight per tiller, MAT = days to maturity after emergence, TKW = thousand kernel weight, VGR = vegetative growth rate, NSM = number of spikes per square meter, STW = straw yield, GFI = days for grain filling.

‡ PR = total monthly precipitation, MT = mean maximum temperature sheltered, mT = mean minimum temperature, SH = sun hours per day; EV = total monthly evaporation; mTU = mean minimum temperature unsheltered; D = December, J = January, F = February, M = March, A = April.

To help interpret the AMMI biplot patterns, the directions of greatest changes for six cultivar covariables ($R^2 > 0.73$, Table 2), as obtained from regressions of the standardized covariables on the first and second AMMI axes, were added to the biplot. These were, in decreasing order with respect to R^2, HID, NGS, NGM, SGW, KGR, and PLH. Cultivars 1 and 2 had above average values for KGR (i.e., they were positively associated with covariable KGR) and had below average values for PLH, SGW, NGS, NGM, and HID. The intermediate-released cultivars, 3 and 4, and later released cultivar, 7, had above average values for HID and NGM, and below average values for KGR. Cultivars 5 and 6 had above average values for SGW, PLH, and NGS. Cultivars 1 and 2 had the largest values for KGR; Cultivars 3, 4, and 7 showed the largest values for HID and NGM; and Cultivars 5 and 6 had the largest values for NGS, PLH, and SGW.

The direction of the greatest changes for six environmental covariables ($R^2 > 0.75$) was also added to the AMMI biplot (Fig. 1a). These were in decreasing order with respect to R^2 PRF, maximum temperature in March (MTM), mTJ, sun hours per day in march (SHM), SHF, and SHJ. Minimum temperature in December ($R^2 = 0.286$) was included in the AMMI biplot for reasons that will be explained later. Years 1993 and 1995 had above average values for MTM and SHM but below average values for SHF. Years 1990, 1991, 1992, and 1994 had high SHF, but low SHM.

Apparently SHM and MTM in 1993 and 1995 caused Cultivars 5 and 6 to develop relatively more NGS and to have heavier SGW than the other cultivars. Years 1990, 1991, 1992, and 1994 had lower values of SHM and MTM and, Cultivars 5 and 6 had correspondingly lower NGS and SGW in those years. Sun hours in January and February (SHJ and SHF) in 1990, 1991, 1992, and 1994 helped Cultivars 1 and 2 develop high individual KGR; however, low SHJ and SHF values in 1993 and 1995 were not conducive for Cultivars 1 and 2 to develop this trait.

Explaining Genotype × Environment Interaction Using Partial Least Squares and Individual Factorial Regression with Cultivar Explanatory Variables

Results from the PLS procedure showed that the first and second factors explained 56 and 13% of the cultivar × year interaction, respectively. Table 3 shows the X-loadings (weights) for each cultivar covariable sorted by the first PLS factor, as well as the complete set of individual factorial regressions on each cultivar covariable, ranked by their contribution to the total cultivar × year sum of squares.

With respect to the FR analysis, there were 15 cultivar covariables available, but only a maximum of $G \leq I - 1$ of them can be used simultaneously, where G is the number of cultivar covariables and I is the number of cultivars (7). All the cultivar covariables were included in the individual FR models and those that explained most of the cultivar × year interaction sum of squares were the same as those that had the highest R^2 values with AMMI scores (Table 2), namely, NGS, NGM, SGW, HID, PLH, and KGR. The rank order of the

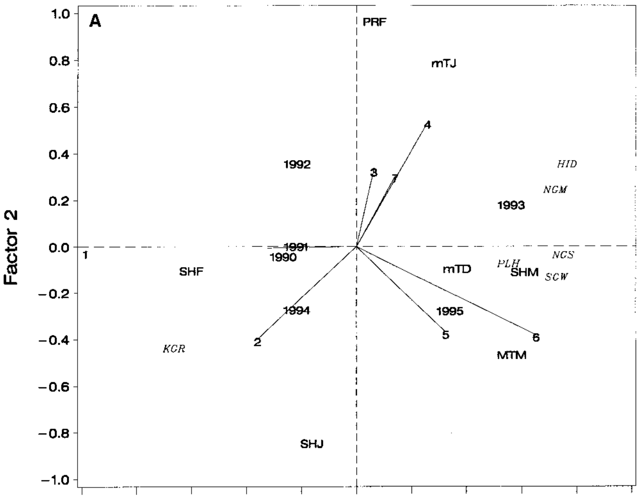

Fig. 1 (A). Biplot of the first and second AMMI axes representing the environmental and cultivar scores of 6 yr (1990–1995) and seven durum wheat cultivars (1–7) enriched with the direction of greatest change of selected cultivar and environmental covariables from Data Set 1. Scaling constant c = 0.5. Cultivar covariables are PLH = plant height, KGR = individual kernel growth rate, HID = harvest index, NGS = number of grains per spike, NGM = number of grains per square meter, SGW = spike grain weight. Environmental covariables are SH = sun hours per day, MT = mean maximum temperature; D = December, J = January, F = February, M = March.

cultivar covariables in relation to how much they contributed to explaining the cultivar × year interaction was practically identical for the PLS approach and the FR model.

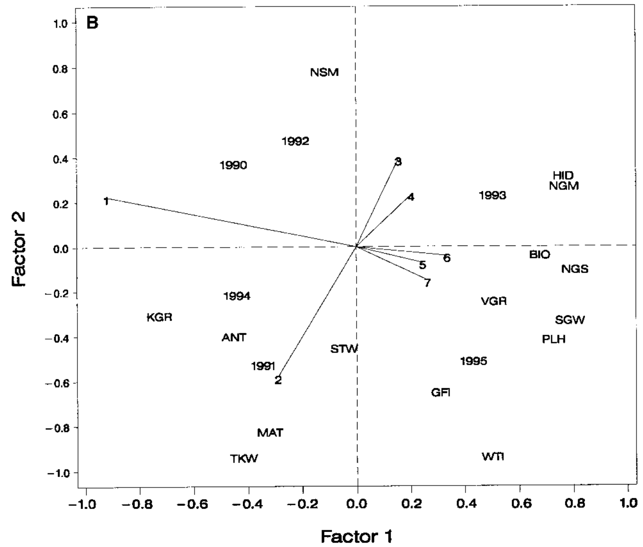

Fig. 1 (B). Biplot of the first and second PLS factors representing the X-scores of seven durum wheat cultivars (1–7), the Y-loadings of 6 yr (1990–1995) enriched with the X-loadings of 15 cultivar covariables from Data Set 1. Scaling constant c = 0.5. Cultivar variables are NGS = number of grains per spike, HID = harvest index, SGW = spike grain weight, NGM = number of grains per square meter, KGR = individual kernel growth rate, PLH = plant height, BIO = above-ground biomass, VGR = vegetative growth rate, ANT = days to anthesis after emergence, WTI = weight per tiller, TKW = thousand kernel weight, MAT = days to maturity after emergence, GFI = days for grain filling, STW = straw yield, NSM = number of spikes per square meter.

The PLS biplot of environmental responses over cultivars versus cultivar covariables is depicted in Fig. 1b. Similarities to the AMMI biplot (Fig. 1a) are evident. The PLS biplot shows that subsets of correlated cultivar covariables can be distinguished (the angle between the variables is important): (HID, NGM), (BIO, NGS, VGR, SGW, PLH), (GFI, WTI), (STW, MAT, TKW, ANT, KGR), and NSM. In contrast to the AMMI biplot, the PLS biplot separated Cultivar 7 from Cultivars 3 and 4, and grouped it with the Cultivars 5 and 6. Cultivar 1 had high KGR and ANT and low HID, NGM, BIO, NGS, VGR, SGW, and PLH. It yielded relatively well in 1990, 1991, 1992, and 1994 and yielded poorly in 1993 and 1995. Cultivars 5, 6, and 7 behaved exactly the opposite. Cultivar 2 had high KGR, ANT, MAT, TKW, and STW, while being low for NSM, HID, and NGM. It yielded relatively well in 1991, 1994, and 1995 and yielded poorly in 1990, 1992, and 1993. Performance of Cultivars 3 and 4 was the reverse of Cultivar 2 with respect to years.

Explaining Genotype × Environment Interaction Using Partial Least Squares and Individual Factorial Regression with Environmental Explanatory Variables

The first PLS factor explained 40% of the cultivar × year interaction sum of squares, while the second PLS factor explained 26%. Table 3 shows the Z-loadings (weights) of the environmental explanatory variables sorted by the first PLS factor, and the individual factorial regression for each of the environmental variables, ranked by their contribution to the analysis of variance GEI sum of squares. The maximum number of covariables that can be used simultaneously in factorial regressions with centered environmental variables is $H \leq J -$

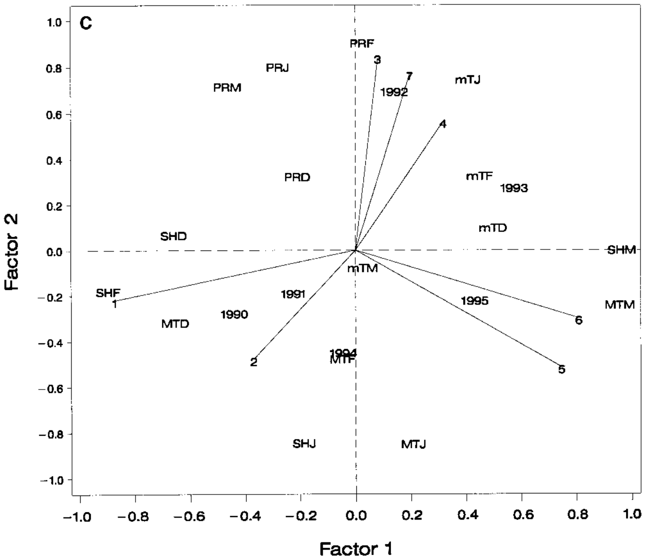

Fig. 1 (C). Biplot of the first and second PLS factors representing the Z-scores of 6 yr (1990–1995) and the Y-loadings of seven durum wheat cultivars (1–7) enriched with the Z-loadings of 16 environmental covariables from Data Set 1. Scaling constant c = 0.5. Environmental variables are mT = mean minimum temperature, MT = mean maximum temperature, PR = total monthly precipitation, SH = sun hours per day; D = December, J = January, F = February, M = March.

1 (see Eq. [7]). As $J = 6$ yr in this data set, $H \leq 5$. The environmental variables that explained most of the cultivar × year interaction sum of squares were SHM, SHF, MTM, MTD, and SHD (Table 3). SHM, SHF, and MTM had values for $R^2 > 0.75$ (Table 2).

The PLS method determined that environmental covariables SHM, SHF, SHD, MTM, and MTD were associated with Factor 1, which explained a large proportion of the cultivar × year interaction (i.e., they had the highest absolute loadings) (Table 3). The FR method also considered these covariables individually to be the five most important environmental covariables in explaining cultivar × year interaction. All other environmental covariables were ranked similarly by both methods. Both procedures considered precipitation to be less important for explaining cultivar × year interaction.

The PLS biplot depicted in Fig. 1c is similar to the AMMI biplot (Fig. 1a) enriched with the directions of greatest change for the environmental covariables with $R^2 > 0.75$. Unlike the AMMI biplot, the PLS biplot separated the low-yielding year 1992 from the other two low yielding years 1994 and 1995. Years 1990 and 1991 had high SHF and MTD and low mTJ and mTF. In contrast to 1990 and 1991, 1993 had high mTJ and mTF, and low MTD and SHF. The year 1992 had high precipitation in general (PRD, PRJ, PRF, PRM) and high mTJ, while 1992 had low MTF, SHJ, and MTJ. In contrast to 1992, 1994 had high MTF, SHJ and MTJ, and low precipitation, low mTJ and mTF. Precipitation during the months of the growing cycle (PRD, PRJ, PRF, and PRM) formed a subset of correlated covariables located in the upper left quadrant of the PLS biplot, whereas a subset of minimum temperatures during the growing cycle (mTD, mTJ, and mTF) is located in the upper right quadrant (mTM is at the center of the PLS biplot). Variables SHM and MTM were positively correlated and formed a subset of environmental variables with high loadings for the first PLS factor.

Cultivars 1 and 2 were favored by SHJ, SHF, MTD, and MTF. This led to higher yields in 1990, 1991, and 1994. Lower mTD, mTJ, and mTF and greater PRF did not favor Cultivars 1 and 2, most notably in 1992 and 1993; however, these environmental conditions during the 1992 and 1993 growing cycles favored Cultivars 3,

Table 3. X-loadings and Z-loadings of the cultivar and environmental covariables, respectively, of the first two PLS factors (sorted by the first PLS factor), and mean squares of all individual factorial regressions for Data Set 1.

	Partial least squares		Factorial regression			
Covariable	Factor 1†	Factor 2	Source	df	Mean square (× 10^5)	Prob > F
			Cultivar covariables‡			
	X-loadings		Year × Cultivar	30	4.849	0.0001
NGS	0.36	0.04	NGS	5	17.298	<0.0001
SGW	0.35	0.14	NGM	5	16.349	<0.0001
NGM	0.34	−0.14	SGW	5	16.339	<0.0001
HID	0.34	−0.14	HID	5	16.058	<0.0001
PLH	0.32	0.18	PLH	5	14.469	<0.0001
KGR	−0.32	0.13	KGR	5	14.290	<0.0001
BIO	0.30	0.01	BIO	5	12.825	<0.0001
VGR	0.23	0.11	VGR	5	8.574	0.0001
WTI	0.22	0.41	WTI	5	8.156	0.0001
ANT	−0.20	0.17	TKW	5	7.291	0.0005
TKW	−0.18	0.41	ANT	5	7.108	0.0006
MAT	−0.14	0.36	MAT	5	5.466	0.0042
GFI	0.14	0.29	GFI	5	3.972	0.0251
NSM	−0.05	−0.50	NSM	5	3.367	0.0512
STW	−0.02	0.20	STW	5	1.563	0.3785
			Error	72	1.446	
			Environmental covariables¶			
	Z-loadings		Year × Cultivar	30	4.849	0.0001
SHM	0.47	0.01	SHM	6	12.594	<0.0001
MTM	0.44	−0.11	SHF	6	12.112	<0.0001
SHF	−0.42	−0.08	MTM	6	11.540	<0.0001
SHD	−0.31	0.04	MTD	6	7.745	0.0001
MTD	−0.30	−0.14	SHD	6	7.086	0.0003
mTD	0.23	0.05	mTJ	6	5.815	0.0016
PRM	−0.22	0.34	mTF	6	5.786	0.0017
mTF	0.21	0.16	mTD	6	5.484	0.0024
mTJ	0.19	0.35	PRM	6	4.582	0.0082
PRJ	−0.13	0.38	SHJ	6	3.850	0.0218
PRD	−0.10	0.15	PRF	6	3.748	0.0249
MTJ	0.10	−0.39	PRJ	6	3.446	0.0373
SHJ	−0.09	−0.39	MTJ	6	3.156	0.0547
MTF	−0.02	−0.22	MTF	6	2.616	0.1111
mTM	0.01	−0.03	mTM	6	2.362	0.1512
PRF	0.01	0.42	PRD	6	1.805	0.2962
			Error	72	1.446	

† PLS results extracted from Tables 1 and 2 of Vargas et al. (1998).

‡ NGS = number of grains per spike, SGW = spike grain weight, NGM = number of grains per square meter, HID = harvest index, PLH = plant height, KGR = individual kernel growth rate, BIO = above-ground biomass, VGR = vegetative growth rate, WTI = weight per tiller, ANT = days to anthesis after emergence, TKW = thousand kernel weight, MAT = days to maturity after emergence, GFI = days for grain filling, NSM = number of spikes per square meter, STW = straw yield.

¶ SH = sun hours per day, MT = mean maximum temperature, mT = mean minimum temperature, PR = total monthly precipitation; D = December, J = January, F = February, M = March.

Table 4. Analysis of variance tables for stepwise multiple factorial regression models with environmental and cultivar covariables, for Data Sets 1 and 2. Terms in factorial regression models appear in the order of inclusion.

Source	d.f.	Sum of squares ($\times 10^6$)	Mean squares ($\times 10^5$)	F	Prob > F
			Data set 1		
			MFR1†		
Cultivar × SHM‡	6	7.557	12.595	8.71	<0.0001
Cultivar × mTD	6	2.655	4.425	3.06	0.0100
Cultivar × PRF	6	2.350	3.916	2.70	0.0198
Deviations	12	1.986	1.655	1.14	0.3428
			MFR2		
Year × NGS	5	8.649	17.298	11.96	<0.0001
Year × NGM	5	1.782	3.564	2.46	0.0409
Deviations	20	4.115	2.057	1.42	0.1489
			MFR3		
NGS × SHM	1	6.484	64.840	44.84	<0.0001
NGS × mTD	1	1.947	19.470	13.46	0.0004
NGS × PRF	1	0.748	7.480	5.17	0.0259
Deviations	27	5.364	1.986	1.37	0.1464
			Data set 2		
Treat × mTF¶	23	78.53	34.143	14.16	0.0001
Treat × EVF	23	54.75	23.804	9.87	0.0001
Treat × mTJ	23	40.50	17.608	7.30	0.0001
Treat × mTUM	23	27.62	12.008	4.98	0.0001
Deviations	115	78.12	6.793	2.81	0.0001

† MFR1 = Multiple Factorial Regression model with environmental covariables; MFR2 = Multiple Factorial Regression model with cultivar covariables; MFR3 = Multiple factorial regression model with cross products of cultivar and environmental covariables from MFR1 and MFR2.

‡ SHM = sun hours per day in March; mTD = minimum temperature in December; PRF = precipitation in February; NGS = number of grains per spike; NGM = number of grains per square meter.

¶ mTF = minimum temperature sheltered in February; EVF = evaporation in February; mTJ = minimum temperature sheltered in January; mTUM = minimum temperature unsheltered in March.

4, and 7. Cultivars 5 and 6 had high yields in year 1995, probably because of higher MTJ, MTM, and SHM. Covariables SHM, MTM and SHF had high loading values for the first PLS factor, whereas covariables PRF, MTJ, SHJ, and PRJ had high loading values for the second PLS factor. These seven covariables had the highest R^2 values (Table 2).

Explaining Genotype × Environment Interaction Using Multiple Factorial Regression with Cultivar and Environmental Explanatory Variables Simultaneously

Multiple factorial regression coupled with a stepwise procedure for variable selection was used to search for informative sets of environmental and/or cultivar covariables. When the set of independent variables from which a selection had to be made consisted of all environmental covariables, a model for the interaction was found that included the terms cultivar × SHM, cultivar × mTD, and cultivar × PRF (Table 4). This model explained 86% of the GEI with 18 df. This model, called multiple factorial regression 1 (MFR1), appeared be slightly better than the AMMI_2 (with two bilinear terms) model, which explained 81% of the GEI with the same 18 df (Table 1).

When the set of candidate variables consisted of all cultivar variables, a model was found that included the terms Year × NGS and Year × NGM (Table 4). This model, MFR2, explained 72% of the GEI with 10 df and appeared to be superior to AMMI_1 (with only one bilinear term), which accounted for 66% of the GEI with 10 df (Table 1). Thus, AMMI models with one or two bilinear terms were possibly less effective than the MFR1 and MFR2 models. It should be pointed out that although MFR1 and MFR2 represent the best sets of environmental and cultivar covariables that were found by stepwise regression, a number of other sets were equally good.

Relating significant cultivar and environmental covariables obtained in the stepwise multiple factorial regression (Table 4) to the clusters of environmental and cultivar covariables previously described in the PLS biplots (Fig. 1b and 1c) can be informative. For example, in Fig. 1c, environmental covariables, SHM and MTM, formed a cluster, and the stepwise procedure selected SHM as the more important. Covariables, mTD, mTJ, mTF, and mTM, formed another cluster, and the stepwise procedure selected mTD as a representative candidate for describing GEI. All precipitation covariables formed another cluster, and stepwise regression found PRF to be the best candidate among them. Although mTD and PRF were not among the best covariables for explaining GEI when performing individual FR, they were important when considered together with other environmental covariables. The stepwise variable selection procedure with factorial regression selected, in order, SHM, mTD, and finally PRF. In the PLS biplot (Fig. 1b), roughly four clusters of cultivar covariables may be distinguished, one for each quadrant. The stepwise procedure selected, as significant contributors to explaining GEI, first NGS and then NGM.

After having found ‘best’ multiple factorial regression (MFR) models for cultivar and environmental variables, we investigated whether further parsimony could be achieved by fitting a multiple FR model that included compound covariables consisting of the products of cultivar and environmental covariables (i.e., we fit a multiple FR model in which only the term $\mathbf{X}\boldsymbol{\nu}\mathbf{Z}'$ from Eq. [9] is maintained). A multiple FR model (MFR3) including the cross products of the cultivar covariable NGS with the environmental covariables SHM, mTD, and PRF gave a very good fit and, on the basis of the differences in sums of squares and df, was not significantly different from MFR1. This gave a very efficient description explaining 63% of the interaction with only 3 df. It is worthwhile to take a look at the interpretation of these cross products. The most important term, NGS × SHM, explained 45% of the total cultivar × year interaction with 1 df. The sign of the estimated coefficient for this term was positive. Thus, cultivars with above average NGS (covariables were all centered, so that positive values after centering mean above average values, and negative values mean below average values) did relatively well in years with above average SHM [(+ NGS in $\mathbf{X}$) × (+ coefficient in $\boldsymbol{\nu}$) × (+ SHM in $\mathbf{Z}'$) = +], as did cultivars with below average NGS in years with below average SHM [(− NGS in $\mathbf{X}$) × (+ coefficient in $\boldsymbol{\nu}$) × (− SHM in $\mathbf{Z}'$) = +). High NGS in years with low SHM and low NGS in years with high SHM are associated with relatively poor performance [(+ NGS in $\mathbf{X}$) × (+ coefficient in $\boldsymbol{\nu}$) × (− SHM in $\mathbf{Z}'$) = −].

The interactions due to NGS × mTD (positive coefficient) and NGS × PRF (negative coefficient) can be interpreted in the same way, although it should be noted that these terms were far less important than the NGS × SHM term.

Data Set 2

AMMI Analysis, Biplot and Correlations

The main effect of treatments (cultural practices) explained 50% of the total sum of squares, whereas differences among year means contributed 24% and the interaction term, 18% (Table 1). The F_R and F_{GH1} tests (Cornelius et al., 1993, 1996; Cornelius, 1993) indicated that the first 5 multiplicative terms were significant ($P <$ 0.05) (the first six multiplicative terms were significant by the F_1 test). The first bilinear interaction term of the AMMI model accounted for 54% of the GEI sum of squares, the second 14%, and the third 13%, using 31, 29, and 27 df, respectively. The first two bilinear terms used 60 df of the total of 207 available in the interaction.

Year main effect was not highly correlated with any environmental variable, except maximum temperature sheltered in December (MTD, $r = 0.59$) (Table 2). Total monthly evaporation and mean precipitation in December (EVD and PRD, respectively) showed relatively high correlations with environmental main effects ($r =$ 0.54 and − 0.50, respectively), indicating the influence of the prevailing climatic conditions on grain yield. In general, values of R^2 obtained from the regression of the standardized environmental variables on the first two bilinear factor scores were relatively low. Only seven variables out of 27 had $R^2 > 0.50$.

The AMMI biplot (Fig. 2a) shows that the axis for the first bilinear term separated the four of the highest yielding years (1994, 1988, 1997, and 1993) from the four lowest yielding years (1995, 1992, 1989, and 1996), althougth 1991, the second highest yielding year, was located with the lowest yielding years; and 1990, the eighth highest yielding year, was grouped with the highest yielding years. Regarding cultural practices, the first axis separated the nine highest yielding treatments (9 [1–1–1–2], 19 [1–2–1–3], 21 [1–1–2–3], 17 [1–1–1–3], 11 [1–2–1–2], 12 [2–2–1–2], 10 [2–1–1–2], 23 [1–2–2–3], and 18 [2–1–1–3]) (five treatments had 200 kg N ha^{-1} and four had 100 kg N ha^{-1}) from the nine treatments with the lowest grain yield (1 [1–1–1–1], 2 [2–1–1–1], 3 [1–2–1–1], 4 [2–2–1–1], 5 [1–1–2–1], 6 [2–1–2–1], 7 [1–2–2–1], 8 [2–2–2–1], and 16 [2–2–2–2]). All had 0 kg N ha^{-1}, except Treatment 16 which had 100 kg N ha^{-1}. The remaining treatments did not show any apparent pattern. The highest yielding treatments were positively related to the highest yielding years, while the lowest yielding treatments were associated with the lowest yielding years.

The AMMI biplot was enriched with the directions of greatest changes for the seven environmental covariables with $R^2 > 0.50$. These covariables were total monthly evaporation in December, January, and April (EVD, EVJ, and EVA, respectively), mean minimum temperature sheltered and unsheltered in March (mTM and mTUM, respectively), mean maximum temperature in February (MTF), and sun hours per day in January (SHJ) (Table 2). Years, 1988, 1990, 1991, and 1996, had above average values (i.e., were positively associated with the covariables EVD, EVJ, EVA, SHJ, and MTF) and had below average values for mTM and mTUM. Years, 1989, 1992, 1994, and 1995, had above average values for covariables, mTM and mTUM, and below average values for the other environmental covariables (Fig. 2a).

Explaining Genotype × Environment Interaction Using Partial Least Squares and Individual Factorial Regression with Environmental Explanatory Variables

The cross validation assessment and Osten's (1988) *F*-test for the number of significant PLS factors indicated that the first factor was significant for prediction, explaining 19% of the year × treatment interaction (data not shown). The second factor explained 28% of the interaction and was found not significant. However, the cross validation gave a PRESS (Predicted Residual Sum of Squares) that was lower than that obtained for the first factor indicating that the second PLS factor improved the prediction accuracy of the model. The first PLS factor had relatively high negative Z-loadings for environmental variables EVD, EVJ, EVF, EVM, and MTD (Table 5) and relatively high positive Z-loadings for mTUF, mTF, mTD, mTM, MTA and PRF. The second PLS factor had high negative loadings for the covariables MTF, mTF and MTA and positive loadings for mTUJ and mTJ.

The maximum number of covariables that could have been used simultaneously in the FR analysis was $H \leq J - 1$ (see eq. [7]), where $J = 10$ (years) so that $H \leq$ 9. Although all individual factorial regressions for the

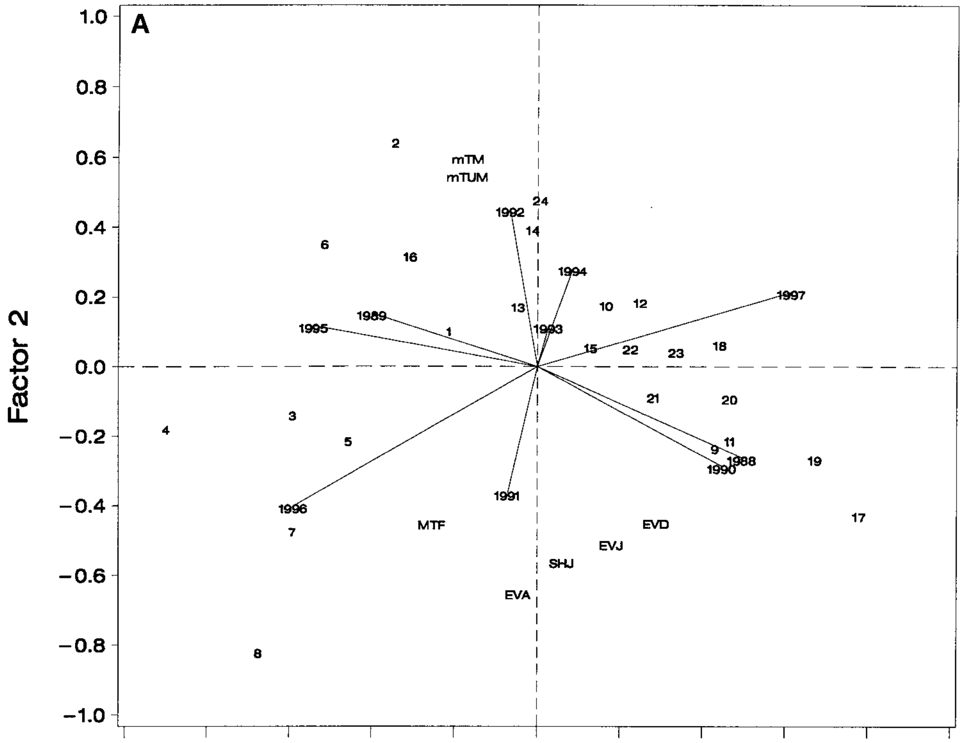

Fig. 2. (A) Biplot of the first and second AMMI axes representing the environmental and treatments scores for 10 yr (1988–1997) and 24 cultural practice treatments (1–24) enriched with the direction of greatest changes of selected environmental covariables from Data Set 2. Scaling constant c = 0.5. Environmental covariables are mT = mean minimum temperature sheltered, mTU = mean minimum temperature unsheltered, EV = total monthly evaporation, MT = mean maximum temperature; D = December, J = January, F = February, M = March, A = April.

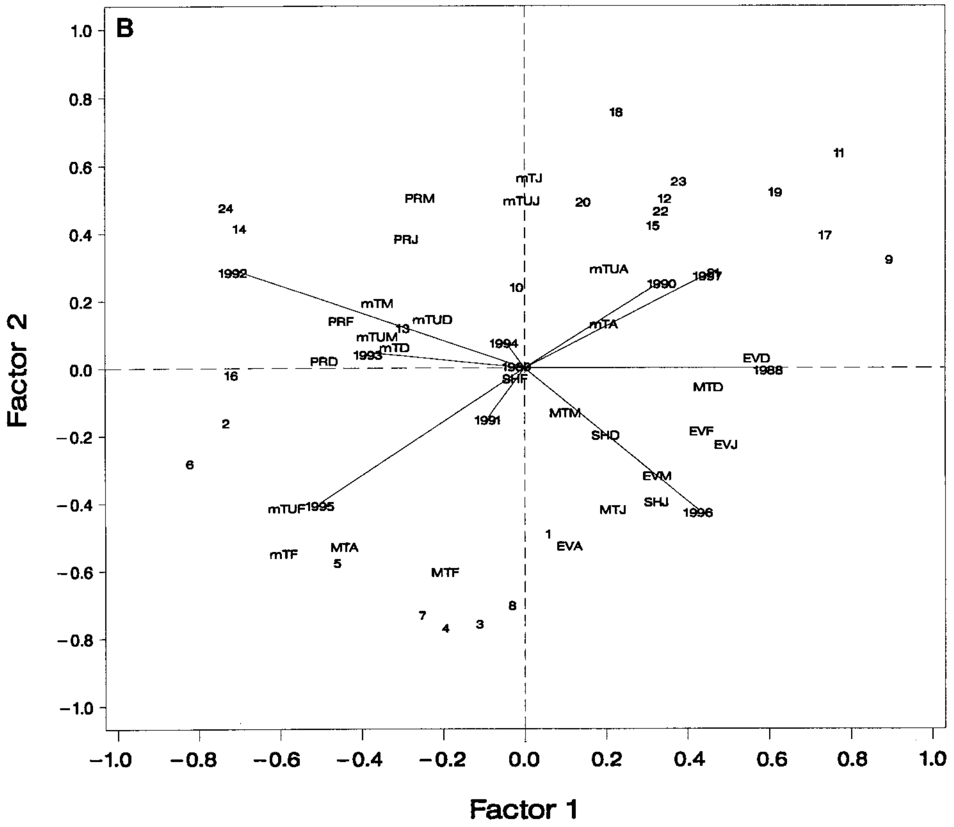

Fig. 2. (B) Biplot of the first and second PLS factors representing the Z-scores of the 10 yr (1988–1997), and the Y-loadings of the 24 cultural practice treatments (1–24) enriched with the Z-loadings of 27 environmental variables from Data Set 2. Environmental variables are mT = mean minimun temperature sheltered, mTU = mean minimum temperature unsheltered, MT = mean maximum temperature sheltered, SH = sun hours per day, EV = total monthly evaporation, PR = total monthly precipitation; D = December, J = January, F = February, M = March, A = April.

centered environmental covariables were significant (each with 23 df), the most interesting were those with the largest sum of squares. The FR model showed that environmental variables, mTF, mTUF, EVD, MTA, MTF, EVJ, mTUM, mTM, and EVF, were important in explaining year × treatment interaction; these variables also had the highest R^2 values for the addition to the AMMI biplot (Table 2). Evaporation in April (EVA) had the largest R^2 value but ranked 14th in FR and 21th in PLS. The rank order of the environmental variables with respect to their contribution to explaining the year × treatment interaction showed good correspondence between PLS and FR for 23 of the covariables (Table 5) (they are ranked at distances lower than four places apart). The most divergent ranking was for MTF, which ranked fifth by FR and 26th by PLS; however, MTF had the highest Z-loading for the second PLS factor (−0.4452). Other variables that differed markedly in ranking were mTUM, EVA, and SHD.

The PLS biplot (Fig. 2b) showed that, for treatments, the results were similar to those obtained with the AMMI biplot (Fig. 2a). The first two PLS factors clearly separated eight of the nine highest yielding treatments (9, 19, 21, 17, 11, 12, 23, and 18) from the nine lowest yielding treatments (1, 2, 3, 4, 5, 6, 7, 8, and 16) (Fig. 2b); however, the separation of years was not as distinct as it was in the AMMI biplot. The low-yielding treatments, 1, 2, 3, 4, 5, 6, 7, 8, and 16, had positive interaction in years with high mTF and mTUF and with high MTF and MTA. This positive interaction was most noticeable in 1995. The year 1995 can be further characterized as being low in mTJ, mTUA, mTA, EVD, MTD, EVF, and EVJ. Negative interactions occurred for the low-yielding treatments in 1988, 1990, and 1997. These years scored just the opposite on the variables enumerated for 1995. In contrast, the eight highest-yielding treatments did relatively well in 1988, 1990, and 1997 and relatively poorly in 1995.

Explaining Genotype × Environment Interaction Using Multiple Factorial Regression with Environmental Explanatory Variables

At least eight covariables were found to be significant by the stepwise selection procedure. The FR model including mTF, EVF, and mTJ had 69 df and explained 62% of the GEI (Table 4), whereas the AMMI_2 (with two bilinear terms) accounted for 68% of the GEI with 60 df (Table 1). The factorial regression model with mTF, EVF, mTJ, and mTUM had 92 df and explained 72% the GEI, whereas AMMI_3 accounted for 81% of the interaction using 87 df. For this data set, AMMI with two or three bilinear terms was slightly more efficient in describing GEI than FR with three or four of the most significant environmental covariables; however, PLS analysis and stepwise FR are still useful for investigating the influence of different environmental covariables.

The PLS biplot (Fig. 2b) contains roughly four clusters of environmental covariables (one for each quadrant). For example, the first cluster is in the lower left quadrant of Fig. 2b and includes correlated variables mTF, mTUF, MTA, and MTF (in decreasing order according to the sum of squares in the individual FR). The second cluster is in the lower right quadrant and comprises correlated variables EVD, EVJ, EVF, MTD, EVA, SHJ, EVM, SHD, MTJ, and MTM. The third cluster involves mTUA, mTUJ, mTJ, and mTA and the fourth cluster is composed of mTUM, mTM, mTD,

Table 5. Z-loadings of environmental variables sorted by the first PLS factor and mean squares of all individual factorial regressions for Data Set 2.

Environmental covariable	Partial least squares		Factorial regression			
	Factor 1	Factor 2	Source	df	Mean square (× 10^6)	Prob >F
	Z-loadings		Year × Treat	207	1.350	0.0001
EVD†	−0.33	0.05	mTF	23	3.414	<0.0001
EVJ	−0.28	−0.07	mTUF	23	3.122	<0.0001
mTUF	0.27	−0.28	EVD	23	2.634	<0.0001
EVF	−0.27	−0.07	MTA	23	2.522	<0.0001
mTF	0.26	−0.35	MTF	23	2.182	<0.0001
MTA	0.24	−0.31	EVJ	23	1.813	<0.0001
mTD	0.24	0.01	mTUM	23	1.729	<0.0001
mTM	0.23	0.04	mTM	23	1.685	<0.0001
MTD	−0.23	0.01	EVF	23	1.633	<0.0001
PRF	0.22	0.04	mTD	23	1.474	<0.0001
EVM	−0.21	−0.15	PRD	23	1.382	<0.0001
SHJ	−0.20	−0.19	MTD	23	1.342	<0.0001
PRD	0.20	−0.02	PRF	23	1.293	<0.0001
mTUM	0.19	0.02	EVA	23	1.272	<0.0001
mTUD	0.19	0.05	SHJ	23	1.248	<0.0001
SHD	−0.18	−0.10	EVM	23	1.235	<0.0001
PRJ	0.16	0.19	mTUA	23	1.234	<0.0001
PRM	0.14	0.20	mTUD	23	1.091	<0.0001
mTUA	−0.12	0.19	mTUJ	23	1.054	<0.0001
mTA	−0.11	0.10	mTJ	23	1.049	<0.0001
EVA	−0.10	−0.28	PRJ	23	1.034	<0.0001
MTJ	−0.09	−0.19	PRM	23	1.031	<0.0001
mTUJ	0.08	0.29	SHD	23	1.003	<0.0001
MTM	−0.07	−0.07	mTA	23	0.887	<0.0001
mTJ	0.07	0.29	MTJ	23	0.770	<0.0001
MTF	0.03	−0.45	SHF	23	0.610	0.0001
SHF	−0.03	−0.01	MTM	23	0.456	0.0079
			Error	460	0.241	

† EV = total monthly evaporation; mTU = mean minimum temperature unsheltered; mT = mean minimum temperature sheltered; MT = mean maximum temperature sheltered; PR = total monthly precipitation; SH = sun hours per day; D = December; J = January; F = February; M = March; A = April.

PRD, PRF, mTUD, PRJ, PRM, and SHF. It is interesting to note that the stepwise FR selected one covariable from each cluster in the following order: mTF, EVF, mTJ, and mTUM, from the first to the fourth clusters, respectively. The next four covariables selected by the stepwise procedure were from the first (MTA), fourth (PRF and PRJ), and third (mTUA) clusters, respectively. These results indicate that PLS was effective in grouping correlated covariables and that stepwise FR was sensitive enough to detect these groups of correlated covariables and to select the most representative from each cluster.

DISCUSSION

Results of this study indicated that FR and PLS were effective in detecting the environmental and cultivar covariables that explained a sizeable proportion of the total GEI variability in two complex data sets. The AMMI biplot enriched with the covariables that showed high R^2 values was also useful for interpreting GEI of grain yield; however, FR and PLS directly incorporate the external variables into their models, whereas AMMI does not. For Data Set 1, the three procedures identified similar cultivar and environmental covariables that explained most of the GEI. For Data Set 2, results were not as clear as those for Data Set 1, but there was a relatively good correspondence between PLS and FR for 23 of 27 environmental covariables.

In general, the AMMI biplot and the PLS biplot offered similar interpretations of the results for both data sets. The AMMI biplot was very similar to the PLS biplot. Interpretation of these biplots is useful for researchers because it helps to identify major environmental (or cultivar) variables that cause positive or negative interactions between subsets of cultivars with subsets of environments. One advantage of the PLS approach is that a large number of environmental (or cultivar) covariables can be used. Furthermore, PLS is insensitive to multicollinearity; for example, for Data Set 2, minimum and maximum temperatures (sheltered and unsheltered), sun hours per day, and total monthly evaporation are correlated. In the AMMI-enriched biplots, multicollinearity is not a problem but when a large number of genotypic or/and environmental covariables are included, none of them may have sufficiently high R^2 to be drawn in the AMMI biplot. In PLS, the cross validation assessment and Osten's (1988) *F*-test can be used to test for the significance of the number of components that must be retained. Although the X- or Z-loadings for each covariable for a given PLS factor are not statistical tests of significance, they do provide a measure of their relative importance for explaining GEI.

The main advantage of the FR is that parameters are estimated and hypotheses are tested in relation to the available external covariables. When environmental and cultivar covariables are considered simultaneously, multiple FR with a stepwise variable selection procedure provides a useful tool for selecting the most relevant covariables, and their cross products, for explaining GEI. For both data sets, selected covariables obtained from stepwise FR represented each of the covariable clusters observed in the PLS biplots. While the PLS analysis is done separately on the set of environmental variables and the set of genotypic covariables, FR and the enriched AMMI-biplot perform a simultaneous analysis on both sets of covariables.

When a large number of correlated environmental (and/or cultivar) covariables is available, an important question that researchers face is how to select a set of relevant environmental and cultivar covariables that effectively explain most of the GEI variability. On the basis of the results obtained in this study, a possible strategy for selecting the most important covariables affecting GEI would be to use, first, the PLS analysis with the PLS biplot. It would also be useful to enrich the AMMI biplots with the relevant environmental and cultivar covariables to compare and confirm results obtained by the PLS approach. Results concerning the relevant covariables affecting GEI obtained by PLS and AMMI can always be confirmed by computing factorial regressions. It is therefore advisable to include in the selected subset covariables that are only slightly correlated. An option would be to select the covariables with the largest explained sum of squares in each of the PLS clusters. After arriving at a satisfactory FR model, one could try to reduce further the model by studying just the cross products of the selected environmental and cultivar covariables.

This study indicated that AMMI, PLS, and FR are useful tools for interpreting GEI in the context of multi-environment trials when a large number of external environmental and cultivar covariables are included. The PLS and FR analyses complement each other and offer an aid to researchers not only for determining the importance of individual environmental and cultivar covariables in explaining GEI, but also for finding subsets of covariables that adequately describe GEI in terms of understandable covariables.

ACKNOWLEDGMENTS

The authors thanks the valuable comments and suggestions given by three anonymous reviewers and the associate editor that significantly improved the quality of the manuscript.

REFERENCES

Aastveit, H., and H. Martens. 1986. ANOVA interactions interpreted by partial least squares regression. Biometrics 42:829–844.

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.

Cornelius, P.L., J. Crossa, and M. Seyedsard. 1996. Statistical tests and estimators of multiplicative models for genotype-by-environment interaction. p. 199–234. *In* M.S. Kang and H.G. Gauch (ed.) Genotype-by-environment interaction, CRC Press, Boca Raton, FL.

Crossa, J. 1990. Statistical analyses of multilocation trials. Adv. Agron. 44:55–85.

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

Denis, J.-B., and J.C. Gower. 1994. Biadditive models. Biometrics 50:310–311.

Eberhart, S.A., and W.A. Russell. 1966. Stability parameters for comparing varieties. Crop Sci. 6:36–40.

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

Gabriel, K.R. 1971. Biplot display of multivariate matrices with application to principal component analysis. Biometrika 58:453–467.

Gabriel, K.R. 1978. Least squares approximation of matrices by additive and multiplicative models. J. Roy. Stat. Soc. Series B. 40: 186–96.

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

GENSTAT. 1995. Genstat 5 release 3.2, reference manual supplement. Clarendon Press, Oxford, UK.

Gollob, H.F. 1968. A statistical model which combines features of factor analysis and analysis of variance techniques. Psychometrika 33:73–115.

Helland, I.S. 1988. On the structure of partial least squares regression. Commun. Statist. Simula. 17(2):581–607.

Kempton, R.A. 1984. The use of biplot in interpreting variety by environment interactions. J. Agric. Sci. (Cambridge) 103:123–135.

Mandel, J. 1971. A new analysis of variance model for non-additive data. Technometrics 13:1–18.

Osten, D.W. 1988. Selection of optimal regression models via cross-validation. J. Chemometrics 2:39–48.

Sayre, K.D., S. Rajaram, and R.A. Fischer. 1997. Yield potential progress in short bread wheats in northwest Mexico. Crop Sci. 37:36–42.

Talbot, M., and A.V. Wheelwright. 1989. The analysis of genotype × environment interactions by partial least squares regression. Biuletyn Oceny Odmian Zeszyt 21–22:19–25.

van Eeuwijk, F.A. 1995. Linear and bilinear models for the analysis of multi-environment trials: I. An inventory of models. Euphytica 84:1–7.

van Eeuwijk, F.A. 1996. Between and beyond additivity and non-additivity; the statistical modelling of genotype by environment interaction in plant breeding. Ph.D. Diss., Wageningen University, Wageningen, the Netherlands.

van Eeuwijk, F.A., J.-B. Denis, and M.S. Kang. 1996. Incorporating additional information on gentoypes and environments in models for two-way genotype by environment tables. p. 15–49. *In* M.S. Kang and H.G. Gauch (ed.) Genotype-by-environment interaction, CRC Press, Boca Raton, FL.

Vargas, M., J. Crossa, K. Sayre, M. Reynolds, M.E. Ramírez, and M. Talbot. 1998. Interpreting genotype × environment interaction in wheat using partial least squares regression. Crop Sci. 38:679–689.

Yates, F., and W.G. Cochran. 1938. The analysis of groups of experiments. J. Agric. Sci. (Cambridge) 28:556–580.

Minimum Sample Size and Optimal Positioning of Flanking Markers in Marker-Assisted Backcrossing for Transfer of a Target Gene

Matthias Frisch, Martin Bohn, and Albrecht E. Melchinger*

ABSTRACT

In recurrent backcrossing designed for introgression of a target allele from a donor into the genetic background of a recurrent parent (RP), molecular markers can accelerate recovery of the recurrent parent genome (RPG). The objectives of this study were to determine in marker-assisted backcrossing (MAB) (i) the optimum distances (d_1, d_2) between the flanking markers and the target locus and (ii) the minimum number of individuals (*n*) required for obtaining with a certain probability a given number of individuals that carry the donor allele at the target locus and have a minimum proportion of donor genome on the carrier chromosome. Analytical solutions and tabulated results are given for relevant parameters (d_1, d_2, *n*) required to obtain, with a specified probability of success, at least one desired individual. They depend on the length of the carrier chromosome, the chromosomal position of the target locus, its distance to the flanking marker loci, and the number of individuals evaluated. Our approach can increase the efficiency of MAB by reducing the number of individuals and marker data points required.

Recurrent backcrossing is a breeding method commonly employed to transfer alleles at one or more loci from a donor to a recurrent parent (Allard, 1960). Examples include the transfer of resistance alleles from a wild or unimproved form into elite breeding materials and cultivars or the transfer of a target allele introduced by genetic transformation into a line that is easy to handle in tissue culture but otherwise of no agronomic value (Ragot et al., 1995). Besides transfer of the target allele(s), the main goal is to recover the RPG as completely and as quickly as possible.

Molecular markers are used in recurrent backcrossing for two purposes: (i) as a diagnostic tool for tracing the presence of a target allele, for which direct selection is difficult or impossible (e.g., recessive alleles expressed at a late stage in plant development or quantitative trait loci) and/or (ii) for identifying individuals with a low proportion of the undesirable genome from the donor parent. Adopting the terminology of Hospital and Charcosset (1997), we refer to the first approach as *foreground selection* (for review see Melchinger, 1990) and to the second approach as *background selection* (for review see Visscher et al., 1996). As demonstrated by Tanksley et al. (1989) with computer simulations, use of molecular markers for background selection can accelerate recovery of the RPG by two or three generations.

Background selection has two goals: (i) reduction of the proportion of the donor genome on the carrier chromosome of the target allele and (ii) reduction of the donor genome on the non-carrier chromosomes. The length of the chromosome segment from the donor that is linked to the target allele (linkage drag) is reduced by selecting individuals that carry the target allele and are homozygous for the RP alleles at tightly linked marker loci. In practical implementations of MAB, two crucial questions are How should the flanking markers by positioned? and How many individuals must be generated and genotyped with molecular markers to reduce the undesirable donor genome below a certain threshold?

Hospital et al. (1992) determined optimum distances d_1 and d_2 between the target locus and the flanking marker loci to recover a maximum amount of the RPG on the carrier chromosome by applying equation

Institute of Plant Breeding, Seed Science, and Population Genetics, Univ. of Hohenheim, 70593 Stuttgart, Germany. Received 31 March 1998. *Corresponding author (melchinger@uni-hohenheim.de).

Published in Crop Sci. 39:967–975 (1999).

Abbreviations: BC, backcross; BC*t*, *t*-th backcross generation; cM, centimorgan; MAB, marker-assisted backcrossing; NRP, non-recurrent parent; QTL, quantitative trait loci; RP, recurrent parent; RPG, recurrent parent genome; RFLP, restriction fragment length polymorphism.