Theor Appl Genet (1998) 96: 803–811

J. Moreno-Gonzalez · J. Crossa

Combining genotype, environment and attribute variables in regression models for predicting the cell-means of multi-environment cultivar trials

Received: 15 August 1997 / Accepted: 28 October 1997

Abstract The main objectives of this study were: (1) to develop models which combine variables of genotype, environment and attribute in regression models (GEAR) for increasing the accuracy of predicted cell-means of the genotype × environment two-way table, and (2) to compare GEAR models with the additive main effects and multiplicative interaction (AMMI) model. GEAR models were developed by regressing the observed values on principal components of genotypes (PCG) and environments (PCE). Genetic and environmental attributes were also added to the GEAR models. GEAR and AMMI models were applied to multi-environment trials of triticale (trial 1), maize (trial 2) and broad beans (trial 3). The random data-splitting and cross-validation procedure was used and the root mean square-predicted difference (RMSPD) was computed to validate each model. GEAR models increased the accuracy of predicted cell-means. Attribute variables, such as soil pH, rainfall, altitude and class of genotype, did not improve the best GEAR model of trial 1, but they increased the predictive value of other models. Two iterations of the computer program further refined the best GEAR model. Based on the RMSPD criterion, GEAR models were as good as, or better than, some AMMI truncated models for predicting cell-means. The approximate accuracy gain factors (GF) of the best GEAR model over the raw data were 2.08, 3.02 and 2.22, for trials 1, 2 and 3, respectively. The GF of the best AMMI model were 1.74, 2.28 and 2.32 for trials 1, 2 and 3, respectively. The analysis of variance of the predicted cell means showed that the genotype × environment interaction (GEI) variance was reduced by about 20% in trial 1 and 81% in trial 2. A bias associated with the predicted cell reduced the GEI variability. Advantages of using GEAR models in muti-environment cultivar trials are that they: (1) increase the precision of cell-mean estimates and (2) reduce the GEI variance and increase trait heritability.

Key words Genotype × environment interaction (GEI) · AMMI · GEI variance

Communicated by P. M. A. Tigerstedt

J. Moreno-Gonzalez (✉)
Centro de Investigaciones Agrarias de Mabegondo,
Xunta de Galicia, Apdo 10, 15080 La Coruña, Spain
Fax: +34 81 673656
E-mail: moreno_ciam@igatel.igape.es

J. Crossa
Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Lisboa 27,
Apdo. Postal 6-641, 06600 Mexico D.F., Mexico

Introduction

Conventional additive models for assessing genotype × environment interaction (GEI) and estimating the realized performance of a genotype in environments include the overall mean, the genotype and environment effects, the GEI effect and the error term associated with the observation. Experimental evidence has shown that GEI is important in estimating genetic variability (Hallauer and Miranda 1981). Increasing the number of testing environments and replications has been suggested as a way to reduce the non-genetic component of variance in phenotypic means and to increase the heritability of traits in selection programs.

Response levels of crop cultivars in a given environment can be better predicted using multiplicative models such as the additive main effects and multiplicative interaction model (AMMI) (Gauch 1988; Gauch and Zobel 1988, 1989), the complete multiplicative model (COMM), the genotypes regression model (GREG), the sites regression model (SREG), and the shifted multiplicative model (SHMM) (Cornelius et al. 1992; Cornelius 1993; Cornelius et al. 1993; Crossa and Cornelius 1993; Cornelius and Crossa 1995; Cornelius et al. 1996). One way to determine the number of multiplicative terms to

be retained in the models is by random data-splitting and cross-validation (Gauch 1988; Gauch and Zobel 1988) by which r_m replicates of each genotype × environment combination are used for modelling and r_v replicates are used for validation. However, Cornelius et al. (1993) and Cornelius and Crossa (1995) suggested that an hypothesis test using *F*-type statistics to determine the optimal number of significant multiplicative terms and/or shrinkage estimates of multiplicative models would eliminate the need for cross-validation as a criterion for model choice.

The cross-validation procedure predicts a true cell-mean by combining the direct information given by the empirical mean of all observations in a cell with indirect information which can be extracted from the other cells. The PRESS statistics (Allen 1971), on the other hand, predict the value of missing cells one at a time by using all the information available from the other cells, and should identify the multiplicative model which most effectively extracts indirect information from the other cells (Cornelius et al. 1993).

A different approach for predicting the true value of a two-way table of genotype × environment cell-means would be to characterize the environments and estimate the effects of environmental variables on the genotype in such way that the genotype value could be predicted by the environmental effects. These variables may be environmental attribute variables (e.g., temperature, rainfall, soil pH, disease and insect attacks, etc.), but can also be the differential behavior of a set of genotypes across environments. These genotype variables could be transformed to principal components (PCG) to condense most of their variability into a few manageable variables. One problem, of course, is to find the set of genotypes that best characterize the environments. Using the same reasoning as above, the genotypes can also be characterized by pedigree, by restricted fragment length polymorphism (RFLP) bands, and by quantitative trait locus (QTL) effects, but also by the differential performance of a set of environments on the genotypes. The environment variables can be also transformed to principal components (PCE).

This approach, which attempts to characterize genotypic behavior by environmental variation and environmental performance by genotypic variation, also utilizes the indirect information in other cells to predict the true values of genotype × environment cell-means of the two-way table. The advantage of this approach over the estimation using the multiplicative models such as AMMI is that environmental and/or genotypic attribute variables can be included in the predictive model.

The objectives of this paper were to: (1) develop statistical models that allow the characterization of genotypic behavior, based on environmental variation, and reciprocally explain environmental performance according to genotypic variation by combining variables of genotypes, environments and attributes in a regression model (GEAR); and (2) compare the predictive assessment of GEAR with AMMI models, using the random data-splitting and cross-validation procedure on data from three multi-environment crop trials.

Materials and methods

Theory of the genotype-environment attribute-regression (GEAR) models

Consider a two-way table where m rows and n columns are the genotypes and environments, respectively, from a multi-environment trial and where the elements of the matrix are the cell means. Principal component (PC) analysis is applied to the matrix where genotypes and environments can be optionally considered either as observations (rows) or variables (columns).

First, consider genotypes as observations and environments as variables; the matrix is m × n. It follows, from principal component algebra, that

$$y_{ij} = \bar{y}_{\cdot j} + \sum_{k=1}^{p} a_{jk} PCE_{ik} + \varepsilon_{ij} \quad (1)$$

(for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$).

Transposing the matrix (n × m) and considering environments as observations and genotypes as variables gives

$$y_{ij} = \bar{y}_{i\cdot} + \sum_{k=1}^{q} b_{ik} PCG_{jk} + \varepsilon_{ij} \quad (2)$$

(for $j = 1, 2, \ldots, n$ and $i = 1, 2 \ldots, m$), where y_{ij} is the value of the i^{th} genotype at the jth environment; $y_{i\cdot}$ and $y_{\cdot j}$ are the means of the i^{th} genotype and the j^{th} environment, respectively; PCE_{ik} is the value of the k^{th} component (axis) for the i^{th} genotype when environments are considered variables; PCG_{jk} is the value of the k^{th} axis for the j^{th} environment when genotypes are taken as variables; a_{jk} is the k^{th} element of the eigenvector for the j^{th} environment variable; b_{ik} is the k^{th} element of the eigenvector for the i^{th} genotype variable; ε_{ij} is the error term associated with the i^{th} genotype in the j^{th} environment; $k = 1, \ldots, p$, and $k = 1, 2, \ldots, q$ are the number of multiplicative terms (principal components) used in Eqs. 1 and 2, respectively.

Model 1

The first multiplicative term of Eq. 1 ($a_{j1} PCE_{i1}$) is the regression of genotypes (tested in the j^{th} environment) on the first PCE. Similarly, the first multiplicative term of Eq. 2 ($b_{i1} PCG_{j1}$) is the regression of environments (across the i^{th} genotype) on the first *PCG*. Equations 1 and 2 estimated two different values of the cell-mean (y_{ij}); however, a unique predicted value could be obtained if the two equations were simultaneously combined in one model.

Consider for simplicity that only one cell-mean, corresponding to the i^{th} genotype (row) in the j^{th} environment (column), is to be predicted. In the following combined regression model all observations of row *i* and column *j*, except that from the element itself (y_{ij}), are used for prediction

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{x}\boldsymbol{\beta} + \mathbf{Gb} + \mathbf{Ea} + \mathbf{e}, \quad (3)$$

where column vector **y** with dimensions $(n + m) \times 1$ is formed by the *n* observations of row *i* followed by the *m* observations of column *j*, except for the two y_{ij} observations which were assigned values of 0 (see Appendix); $\boldsymbol{\mu}$ is a vector formed by the mean of genotype *i* repeated *n* times, followed by the mean of environment *j* repeated *m* times; **x** is a vector where all elements have value 0, except the two elements corresponding to the y_{ij} observations which have a value of

-1; $\boldsymbol{\beta}$ is a vector with a single element for the predicted value of y_{ij} (y_{ij}); **G** is a $(n+m)\times k$ matrix where the first n rows are the first k *PCs* taking genotypes as variables, and the following m rows have value 0; E is a $(n+m)\times q$ matrix where the first n rows have the value 0 and the following m rows are the first q *PCs* taking environments as variables; **a** and **b** are vectors of the partial regression coefficients of the variables included in **E** and **G**, respectively; **e** is the vector of random error effects for the observations.
Combining appropriate matrices, Eq. 3 can be also written as

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Xv} + \mathbf{e}, \tag{4}$$

where

$$\mathbf{X} = [\mathbf{x}:\mathbf{E}:\mathbf{G}] \text{ and } \mathbf{v}' = [\boldsymbol{\beta}':\mathbf{a}':\mathbf{b}'].$$

Since y_{ij} is included in **v**, the predicted value of y_{ij} can be estimated from Eq. 4 by applying least-squares algebra such that

$$\mathbf{v} = [\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}(\mathbf{y} - \boldsymbol{\mu})$$

Note that in model 1 only *PCE* and *PCG* variables are used for prediction and no environmental- or genotypic-attributes are added to the model.

Model 2

If environmental-attribute variables, such as soil pH, soil fertility levels, rainfall, daily temperature, altitude, etc., and genetic-attributes such as RFLP bands, QTL or pedigree relationships are available, Model 1 can be expanded to include these variables. Then, Eq. 4 becomes:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Xv} + \mathbf{Tu} + \mathbf{e} \tag{5}$$

where $\mathbf{T} = [\mathbf{H}:\mathbf{K}]$; $\mathbf{u}' = [\mathbf{h}':\mathbf{k}']$; **H** is a $(n+m)\times$ r matrix formed with the values of the r environmental-attribute variables for the first n rows and the value 0 for the following m rows (see Appendix); **K** is a $(n+m)\times s$ matrix which has the value 0 for the first n rows and the corresponding values of the genetic-attribute variables for the following m rows; **h** is a $r\times 1$ vector of environmental-attribute effects; **k** is a $s\times 1$ vector of genetic-attribute effects.

In general, variables in the **H** matrix are of the continuous type, while variables in the **K** matrix are generally from the class type. If continuous variables are transformed into principal components (PC), only the first few PC will be considered in the model. Class variables are treated as dummy variables.

The predicted value y_{ij} included in **v** can be estimated by applying least-squares analysis to Eq. 5.

Several regression strategies are possible:
Strategy A: only PCE and PCG variables from the **X** matrix are entered into the regression model. This strategy is equivalent to applying GEAR Model 1 (Eq. 4) to the data, since it does not include any attribute-variables.
Strategy B: variables from **X** are retained in the model, and only the attribute-variables from matrix **T** (Model 2, Eq. 5) are further added to the model using stepwise regression (Draper and Smith 1966).
Strategy C: the PCE and PCG variables from matrix **X** and the attribute variables from matrix **T** are all entered in the model using stepwise regression.

Computer programs were written in SAS (1989) for Eqs. 4 and 5 to make predictions of true cell values. Each observed value was predicted under strategies A, B and C. Several predicted values were computed for strategy A using different combinations of the first PCG and PCE. Predicted values for strategies B and C were only programmed for the first three PCG and PCE, but predictions for other combinations can be easily added to the program. The environmental-attribute variables are transformed to principal components (PCEAV).

Random data-splitting and cross-validation

Cross-validation of the models was performed by partitioning the data into model (r_m) and validation data (r_v). Two replicates were randomly selected for each genotype at each site to form the model data ($r_m = 2$) for each trial, and the remaining two replicates from trials 1 and 2 ($r_v = 2$) and one replicate from trial 3 ($r_v = 1$) comprised the validation data. The root mean square difference of the predicted values (RMSPD), as described by Gauch and Zobel (1988) and Crossa et al. (1990), was used as the criterion for validating the models. The RMSPD over several data sets was calculated as the difference between the predicted value of model data and the replication means of validation data squared and summed over all genotype and environment (GE) combinations and data sets. This sum was divided by the number of GE combinations and data sets, and the square root was taken.

In addition to the GEAR models, the following AMMI models were fitted to the data of each trial: $AMMI_0$, which includes only the genotype and environment main effects, and $AMMI_1$, $AMMI_2$ and $AMMI_3$ which combine the main effects from $AMMI_0$ with part of the GEI effect estimated from the first PC axes 1, 2 and 3 of the AMMI analysis, respectively (Gauch and Zobel 1988, 1996; Crossa et al. 1990). The cell-means which incorporate the main genotype and environment and the GEI effects were also used to predict true cell values. The RMSPD criterion was also utilized to validate the AMMI model. The prediction assessment of the AMMI and GEAR models were compared on the basis of the least RMSPD criterion for the same data generated.

Analysis of variance (ANOVA) on raw- and predicted-data sets

ANOVAs were performed on two-replicate and four-replicate raw- and predicted-data sets in both trials 1 and 2. For each trial, four independent raw-data subsets were formed by randomly assigning a replicate from each cell to a subset. The four subsets made a four-replicate raw-data set. A four-replicate predicted-data set was obtained by applying a GEAR model to each raw-data subset. Then an ANOVA was performed on the four-replicate raw- and predicted-data sets.

Likewise, two independent-data subsets were formed with the mean of two random replications from each cell. The two subsets made a two-replicate raw-data set. A two- replicate predicted-data set was obtained by applying a GEAR model to the two raw-data subsets. Then an ANOVA was performed on the two-replicate raw- and predicted-data sets.

Data

Three sets of yield data were used for testing the models. Trial 1 is an official multi-environment trial coordinated by the Spanish National Seed Institute (INSPV) which comprises 16 triticale cultivars with four replications evaluated at ten environments in Spain during 1989. These data were utilized by Royo et al. (1993) for an AMMI analysis. Besides the yield data, three environmental-attribute variables (soil pH, annual rainfall and altitude) were recorded at each site. One genetic-attribute variable, the presence of either a complete or substituted rye genome, was also recorded for each triticale genotype. Trial 2 is a CIMMYT maize international trial where eight maize genotypes were arranged in a randomized complete block design with four replications at each of 33 sites scattered over the tropical region in 1987. Trial 3 comprises 11 broad bean (*Vicia faba*) genotypes arranged in a randomized complete block design with three replications grown at ten environments in Southern Spain. Data of trial 3 were extracted from Cubero and Flores (1995). No attribute-variables were available for trials 2 and 3.

Results and discussion

Trial 1

The range of cell-means varied from 620 kg ha^{-1} for genotype 8 at environment 10 up to 9340 kg ha^{-1} (about 15-times larger) for genotype 2 at environment 1. The range of environment-means was also large (from 823 to 6753 kg ha^{-1}) but that for genotype-means was smaller (from 3564 to 4735 kg ha^{-1}). The relative magnitude of these ranges was reflected in the analysis of variance where the mean squares for environments were 37 times larger than for genotypes (data not shown). Genotype, environment, and GEI effects were all significant. The component of variance for the GEI, $\sigma^2_{ge} = 371388$, was larger than the error component, $\sigma^2_e = 326878$.

The RMSPD obtained from 100 random data-splitting and cross-validation approaches using $r_m = 2$ and $r_v = 2$ for several GEAR and AMMI models are shown in Table 1. The best AMMI model was $AMMI_2$ and the worst $AMMI_0$; in $AMMI_0$ the GEI variance component was not involved in the prediction model. Of the GEAR models (Strategy A, Eq. 4), G3S1, which involves the first three PCG and the first PCE, was the best predictor. The G3S1 model was also a slightly better predictor than the best AMMI model, $AMMI_2$ (504.8 kg ha^{-1} vs 507.2 kg ha^{-1}). Other GEAR models such as G2S2, G3S1 and G4S1 were superior to $AMMI_1$, the second best AMMI model. One advantage of the GEAR models is the flexibility of combining different numbers of PCs from the genotype and environment variables. All GEAR models, as well as the $AMMI_1$–$AMMI_3$ models, were better predictors than the cell-means model (573.1 kg ha^{-1}).

Strategy B consists of entering the attribute variables by stepwise regression into a pre-selected GEAR model while the PCG and PCE variables are retained in the model. In this case the best GEAR model, G3S1, was pre-selected. Two options of strategy B were examined; one included the first PC of the three environmental-attribute variables (soil pH, altitude and annual rainfall) in the G3S1 model using stepwise regression; the other involved entering both environmental-attribute variables and the genetic-attribute variable (complete or substituted rye genome) in the G3S1 model using stepwise regression. The results from random data-splits and cross-validation show that neither strategy B, including only environmental attributes, nor strategy B, including both environmental and genetic attributes, improved the original G3S1 model (data not shown). This suggests that the differential performance of genotypes across environments, and reciprocally the differential performance of environments through genotypes, were sufficient to explain the true values of cell means, and no margin was left for attribute-variables to improve the prediction.

Strategy C includes the first three PCGs and PCEs, along the environmental- and genetic-attribute variables, via a stepwise regression procedure. This strategy improved predictability over the model alone (with no attribute-variables included). This seems to be a result of the attribute-variables adding significant structural information to the original model. Royo et al. (1993) using data of trial 1 further explained GEI by incorporating soil pH values as a co-variate to AMMI models.

Table 1 Root mean square-predicted difference (RMSPD) of trials 1, 2 and 3 based on 100 random data splits ($r_m = 2$, $r_v = 2$) for different truncated AMMI models and GEAR models obtained using strategy A (Eq. 4)

Models	kg ha^{-1}		
	Trial 1	Trial 2	Trial 3
AMMI model			
$AMMI_0$	705.2	649.5	514.1
$AMMI_1$	530.0	665.7	459.1
$AMMI_2$	507.2	682.8	485.3
$AMMI_3$	546.2	701.2	495.0
GEAR model[a]			
G3S3	526.7	–	–
G2S2	524.6	–	469.2
G3S2	527.8	–	475.9
G3S1	504.8	646.0	472.2
G4S1	512.3	–	471.1
G2S1	–	641.5	467.8
G1S2	–	634.9	–
G1S3	–	632.2	–
Cell means	573.1	765.1	510.8

[a] Numbers after G and S refer to the number of the first PC retained in the model using genotypes and environments as variables, respectively

Trial 2

The 33 environment-yield means ranged from 2233 to 6950 kg ha^{-1} for environments 27 and 2, respectively. This is reflected in the highly significant F value (88.1) for environment effects. The genotype and GEI effects were also significant, but the F value for GEI was relatively small. The GEI variance component ($\sigma^2_{ge} = 98005$) was about six times smaller than the error variance ($\sigma^2_e = 586577$).

The best AMMI model for predicting cell means was $AMMI_0$ (Table 1), due in part to the small σ^2_{ge}/σ^2_e ratio. $AMMI_0$ values are estimated as the mean of several observations and do not involve GEI. Thus, the absence of the GEI component (with a small value) is compensated for by a larger reduction in the error variance associated with the observation means that estimate the $AMMI_0$ cells.

The best GEAR model was G1S3 (632.2 kg ha^{-1}) which includes the first PCG and the first three PCEs (Table 1). All GEAR models were better predictors

than any AMMI model. The cell-means model was the poorest predictor of true values of genotype × environment combinations (765.1 kg ha^{-1}), most likely due to the large error-variance associated with the observations.

Trial 3

The range of environmental and genotypic means was relatively large, a fact which is reflected in the large significant F values of those effects. The GEI effects were also significant and the GEI variance component was $\sigma^2_{ge} = 132950$, slightly smaller than the error variance ($\sigma^2_e = 173376$).

$AMMI_1$ was the best AMMI model and better than all GEAR models (459.1 kg ha^{-1}) (Table 1). However, the RMSPDs of the five GEAR models were smaller than the RMSPD of the second best AMMI model ($AMMI_2$), which suggests a good degree of robustness of the GEAR models for cell-means prediction. In this trial, predictions by the cell-means model were better than those obtained by $AMMI_0$. Comparison of the efficiency of the $AMMI_0$ vs the cell-means model depends on the $r_m\sigma^2_{ge}/\sigma^2_e$ ratio. The cell-means model is better than $AMMI_0$ if the ratio is higher than 1, and $AMMI_0$ is better if the ratio is less than 1.

Improving the predictive-values of the GEAR models

One way to improve the performance of the GEAR models is to use the combined average of the predicted-values of the best two or three GEAR models as a new predictor for each cell. Only models having close enough RMSPD values should be combined; otherwise there is no improvement over the best model.

The RMSPD of the best models and their weighted combination for each trial is shown in Table 2. A slight improvement over the best model was achieved in trials 2 and 3 by assigning equal weights to the predicted values of GEAR models having close RMSPDs. Note that the RMSPDs of (G1S3 + G1S2)/2 and (G2S1 + G2S2 + G4S1)/3 were smaller than those of the best models G1S3 in trial 2 and G2S1 in trial 3, respectively. No improvement was found in trial 1 or over the AMMI models of trial 3, most likely because the RMSPDs of the combined models were not close enough.

Another way to improve the predictive-value of the GEAR models is by iterating the procedure. Cell predictions are calculated by executing the computer regression program only once. However, improvements in prediction were obtained by repeating the procedure, assigning predicted values from first iteration to each cell. The regression is then performed on the new predicted cells, keeping values of the PCG and PCE variables from the original data constant. Table 3 shows the RMSPDs of the best one-iteration and two-iteration models, and a sizeable prediction improvement was achieved via the latter in all trials. The first three PC variables were retained in the second iteration of the best models, most likely because they explained further significant variation. These results suggest that a two-iteration model starting with the best one-iteration model is a useful approach for prediction assessment.

Table 2 Root mean square-predicted difference (RMSPD) of trials 1, 2 and 3 based on 100 random data splits ($r_m = 2$, $r_v = 2$) for the best truncated AMMI models, GEAR models, and the weighted combination

Trial	Best models[a]	Weighed combination of best models	RMSPD (kg ha^{-1})
1	G3S1		504.8
	G4S1		512.3
		(G3S1 + G4S1)/2	506.6
		(4*G3S1 + G4S1)/5	505.0
2	G1S3		632.2
	G1S2		634.9
		(G1S3 + G1S2)/2	630.6
3	$AMMI_1$		459.1
	$AMMI_2$		485.3
		($AMMI_1$ + $AMMI_2$)/2	467.1
3	G2S2		469.2
	G4S1		471.1
	G2S1		467.8
		(G2S1 + G2S2 + G4S1)/3	464.3

[a] Numbers after G and S refer to the number of the first PC retained in the model using genotypes and environments as variables, respectively

Table 3 Root mean square-predicted difference (RMSPD) of trials 1, 2 and 3 based on 100 random data splits ($r_m = 2$, $r_v = 2$) for one-iteration and two-iteration (first and second iteration) of the best GEAR models

Trial	One iteration best model[a]	First + second iteration models[a]	RMSPD (kg ha^{-1})
1	G3S1		499.2
		G3S1 + G3S3	492.1
		G3S1 + G4S3	493.4
		G3S1 + G3S2	497.5
		G3S1 + G3S1	504.6
2	G1S3		629.6
		G1S3 + G3S3	624.8
		G1S3 + G3S2	625.5
		G1S3 + G3S1	627.6
		G1S3 + G2S2	625.2
3	G2S1		467.5
		G2S1 + G3S3	461.6
		G2S1 + G3S2	460.9
		G2S1 + G2S2	461.2

[a] Numbers after G and S refer to the number of the first PCG and PCS retained in the model, respectively

Gain factor of GEAR and AMMI models

The mean square error of the model, MSE(model), is the difference between the square of the RMSPD and the error variance of the validation-data [VAR(validation)] (Crossa et al. 1990). Since means rather than observations were used for validation, the VAR(validation) was estimated as the error-variance of the trial divided by the number of replications for validation. Estimates of VAR(validation) were 163 439, 293 288 and 173 376 for trials 1, 2 and 3, respectively. Thus, it follows from Table 3 that the MSE(model) estimates of the best two-iteration GEAR models were 78 723, 97 087 and 39 053 for trials 1, 2 and 3, respectively. The approximate gain factor (GF) in the precision of the model relative to the raw data has been established (Gauch and Zobel 1988) as the error variance/r_mMSE(model) ratio. The computed GF for the best two-iteration GEAR models in trials 1, 2 and 3 were 2.08, 3.02 and 2.22, respectively. Likewise the GF for the best AMMI models in trials 1, 2 and 3 were 1.74, 2.28 and 2.32. The predictive-value of the GEAR models was better than that of the AMMI in trials 1 and 2, but was slightly inferior in trial 3. These results show that the efficiency of the GEAR models for predicting cell-means is better than that of the AMMI models.

Analysis of variance (ANOVA)

Unlike the AMMI models, which partition out the GEI variation into orthogonal components, the GEAR models predict each cell individually by fitting the regression model to a set of observations and do not provide an orthogonal partition of the variation. However, ANOVAs were performed on predicted-data sets with two and four replicates.

Table 4 shows the averages of mean squares (MS) and variance components for genotypes, environments, GEI and error effects from ANOVAs performed on 90 four-replicate and two-replicate raw- and predicted-data sets in both trials 1 and 2. The error MS of the predicted data were much smaller than those of the raw data. This confirms the increase in precision of the GEAR models which had also been shown by the RMSPD criterion.

According to Crossa et al. (1990), the MSE(model) = VAR(model) + $(BIAS_{model})^2$, where MSE(model) is the mean square of the difference between the predicted and the true values of each cell, VAR(model) is the variance of the error associated with the model (i.e., the error MS of the predicted data), and $BIAS_{model}$ is the bias of the model. In the two-replicate data sets of trial 1, the VAR(model) = 50 300 (Table 4) and the MSE(model) was estimated as 78 723, thus the estimate of $(BIAS_{model})^2$ = 28 423 = likewise for trial 2, the VAR(model) = 65 194 (Table 4) and the MSE(model) = 97 087, thus the estimate of $(BIAS_{model})^2$ = 31 893. In

Table 4 Averages of pertinent mean squares (MS) (kg ha^{-1}) and variance components[a] from the ANOVA performed on 90 data-sets of four and two replicates predicted by two-iteration GEAR models in trials 1 and 2

Sources of variation	Four-replicate data sets[b]					Two-replicate data sets[c]				
	df	Raw data		Predicated data[d]		*df*	Raw data		Predicted data	
		MS	Variance component	MS	Variance component		MS	Variance component	MS	Variance component
Trial 1										
Genotype (G)	15	5 762 335	98 748	5 237 147	99 391	15	2 881 168	98 748	2 704 620	103 111
Environment (E)	9	213 708 091	3 310 870	213 874 420	3 322 077	9	106 854 046	3 310 870	106 928 537	3 321 442
G × E	135	1 812 431	371 388	1 261 501	288 890	135	906 216	371 190	642 393	296 046
Error	480	326 877	326 877	105 940	105 940	160	163 836	163 836	50 300	50 300
Trial 2										
Genotype (G)	7	10 770 993	74 185	9 739 608	72 834	7	5 385 497	74 185	5 247 626	78 166
Environment (E)	32	51 665 972	1 583 980	51 648 472	1 610 219	32	25 832 985	1 583 980	25 836 398	1 609 355
G × E	224	978 599	98 005	121 461	1503	224	489 300	98 900	87 355	11 080
Error	792	586 577	586 877	115 448	115 448	264	291 500	291 500	65 194	65 194

[a] Estimated from expected mean squares assuming random effects
[b] Each replicate of the four-replicate raw-data sets was formed with a random single replicate from each cell of the original data
[c] Each replicate of the two-replicate raw-data sets was formed with the mean of two random replicates from each cell of the original data
[d] Predicted data were obtained by applying the G4S1 + G3S3 and G3S1 + G3S3 two-iteration models to the four-replicate and two-replicate raw-data sets of trial 1, and the G1S1 + G3S3 and G1S3 + G3S3 models to the four-replicate and two-replicate raw data sets in trial 2, respectively

the four-replicate data sets of trial 1, the estimate of VAR(model) = 105 940 (Table 4) and the MSE(model) = 146 975 (data not shown), so that the $(BIAS_{model})^2$ = 41 035; likewise for trial 2, the estimate of VAR(model) = 115 448 (Table 4) the MSE(model) = 163 365 (data not shown), so that the $(BIAS_{model})^2$ = 48 187. These biases are associated with the predicted cell-means. The genotype and environment variance components of the predicted data were similar to those of the raw data. The GEI variances were smaller for the predicted data than for the raw data (Table 4). The reduction in GEI variances for the two-replicate and four-replicate data sets was 20 and 22% in trial 1 and 98 and 89% in trial 2, respectively. The reduction of the GEI variance can be partially explained by the bias associated with the predicted cell-means. The bias made the predicted cell-means approach the additive model without interaction ($AMMI_0$), and this favored the reduction of the GEI. Furthermore, the bias did not adversely effect model precision, since the deviation from the true cell-means was compensated for by a drastic decline in the error variance (Table 4), which resulted in a more favorable RMSPD.

The MSE(model) associated with the mean of replicates from predicted-data sets can be estimated as $VAR(model)/r + (BIAS_{model})^2$, where r is the number of replicates. Thus, the MSE(model) of predicted means from the two-replicate and four-replicate data sets were 53 573 and 67 520 in trial 1, and 64 490 and 89 028 in trial 2, respectively. Therefore, when all observations were used for predictions, the mean of the two predicted replicates (each based on the mean of two random original replicates) was more precise than the mean of the four predicted replicates (each based on a single replicate). Taking this idea further, the best data set for model prediction should be the complete data set involving the mean of all replications. Since at least one replication is needed to validate the model, selection of the best model should be done by testing models on data sets formed with the mean of $r - 1$ random replicates, leaving the remaining replicate for cross-validation. Cornelius and Crossa (1995) and Cornelius et al. (1996) have also indicated that if a data-splitting and cross-validation procedure is used, r_m should be $r - 1$ ($r_v = 1$) to decrease the noise in the modelling data. The model selected in this way should be applied to the data formed with the cell-means involving all replications. Furthermore, an ANOVA can be performed on the predicted-data set.

Table 5 shows the ANOVAs of raw-data sets formed with the cell means involving all replicates and the ANOVAs of predicted-data sets obtained by applying the G3S1 + G3S3 and G1S3 + G3S3 two-iteration models to the raw-data sets in trials 1 and 2, respectively. No direct estimate of the error MS exists because of the lack of replications; however, it can be approximated by the error MS of the two-replicate predicted data (Table 4) divided by two, because four replicates instead of two were used for the model data. The genotype and environment variance components of the predicted data were slightly higher than those of the raw data in trials 1 and 2, while the GEI variance of the predicted data was reduced by 20% in trial 1 and 81% in trial 2. The broad-sense heritability (h_b^2) based on phenotypic means can be easily estimated from Table 5. The h_b^2 estimates of predicted data increased from 0.685 to 0.767 in trial 1 and from 0.908 to 0.983 in trial 2.

In trial 1, the genotype and environment predicted means, obtained by applying the G3S1 + G3S3 two-iteration model to the data of raw cell-means, were similar to the genotype and environment observed means (data not shown).

Table 5 Pertinent mean squares (MS) and variance components (kg ha^{-1}) from the ANOVAs performed on the raw data of cell means and predicted data in trials 1 and 2

Sources of variation	*df*	Raw data of cell means[a]		Predicted data[b]	
		MS	Variance component	MS	Variance component
Trial 1					
Genotype (G)	15	1 440 584	98 748	1 379 799	105 804
Environment (E)	9	53 427 023	3 310 870	53 469 249	3 321 718
G × E	135	453 108	371 388	321 761	296 611
Error	480	81 719[c]	81 719	25 150[d]	25 150[d]
Trial 2					
Genotype (G)	7	2 692 748	74 185	2 634 959	78 470
Environment (E)	32	12 916 493	1 583 980	12 925 260	1 609 971
G × E	224	244 650	98 005	45 449	12 852
Error	794	146 644[c]	146 644	32 597[e]	32 597[e]

[a] The raw data were formed with the cell means involving all replicates
[b] Predicted data were obtained by applying the G3S1 + G3S3 and G1S3 + G3S3 two-iteration models to the raw data of cell means in trials 1 and 2, respectively
[c] The error MS associated with raw cell means is the error MS of the four-replicate raw-data sets (Table 4) divided by 4
[d] Estimated from the error MS of the two-replicate predicted data sets of trial 1 (Table 4) divided by 2
[e] Estimated from the error MS of the two-replicate predicted data sets of trial 2 (Table 4) divided by 2

Conclusions

The results of the GEAR models applied to three multi-environment cultivar trials differing in respect of experiment size, the ratio of error variance to GEI variance, the magnitude of genotype and environment variances, the cultivated crop, and the area of cultivation suggest that the applicability of the GEAR models may be broad. In addition, the best AMMI models for prediction were also different for trials 1, 2 and 3 ($AMMI_2$, $AMMI_0$ and $AMMI_1$, respectively). The results of this study indicate that the main advantages of using GEAR and iterated GEAR models for predicting the values (cells) of multi-environment cultivar trials experiments are: (1) the accuracy of prediction using cell-means is improved by allowing variables of genotypes, environments and attributes to be combined in the model, (2) GEI variance is reduced, and (3) the precision of estimates of heritability for a trait is increased.

Although not done in this study, the interaction between PCGs and PCEs could have been included in the GEAR model for predicting cell-means. Further research is required to assess the efficiency of the GEAR models for imputing values of cell-means that are accidentally or intentionally missing from the two-way table of genotype × environment. More research is also needed to compare the prediction assessment of the GEAR models with that of other multiplicative models and with the best linear unbiased predictor (BLUP).

Appendix

The matrices and vectors of Eq. 3, $\mathbf{y} = \boldsymbol{\mu} + \mathbf{x}\boldsymbol{\beta} + \mathbf{Gb} + \mathbf{Ea} + \mathbf{e}$, of Model 1 are:

$$\begin{bmatrix} y_{i1} \\ \cdot \\ \cdot \\ y_{ij-1} \\ 0 \\ y_{ij+1} \\ \cdot \\ \cdot \\ y_{in} \\ y_{1j} \\ \cdot \\ \cdot \\ y_{i-1\,j} \\ 0 \\ y_{i+1\,j} \\ \cdot \\ \cdot \\ y_{mj} \end{bmatrix} = \begin{bmatrix} \bar{y}_{i\cdot} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ . \\ \bar{y}_{i.} \\ \bar{y}_{\cdot j} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \bar{y}_{\cdot j} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \beta + \begin{bmatrix} PCG_{11} \ldots PCG_{1k} \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ PCG_{j1} \ldots PCG_{jk} \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ PCG_{n1} \cdots PCG_{nk} \\ 0 \cdots 0 \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ 0 \cdots 0 \end{bmatrix} [b] + \begin{bmatrix} 0 \cdots 0 \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ 0 \cdots 0 \\ PCE_{11} \cdots PCE_{1q} \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ PCE_{i1} \cdots PCE_{iq} \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ PCE_{m1} \cdots PCE_{mq} \end{bmatrix} [a] + \begin{bmatrix} e_{i1} \\ \cdot \\ \cdot \\ e_{ij-1} \\ e_{ij} \\ e_{ij+1} \\ \cdot \\ \cdot \\ e_{in} \\ e_{1j} \\ \cdot \\ \cdot \\ e_{i-1\,j} \\ e_{ij} \\ e_{i+1\,j} \\ \cdot \\ \cdot \\ e_{mj} \end{bmatrix},$$

where the predicted value has, for this case, only one element $\beta = y_{ij}$; elements of $\mathbf{a}' = [a_{i1}, a_{i2}, \ldots, a_{iq}]$ are partial regression coefficients of the first q PCE variables for genotype i; elements of vector $\mathbf{b}' = [b_{j1}, b_{j2}, \ldots, b_{jk}]$ are the partial regression coefficients of the first k PCG variables for environment j:

The matrices included in **T** of Model 2 (Eq. 5, $\mathbf{y} = \boldsymbol{\mu} + \mathbf{Xv} + \mathbf{Tu} + \mathbf{e}$) are:

$$\mathbf{H} = \begin{bmatrix} EA_{11} \cdots EA_{1r} \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ EA_{j1} \cdots EA_{jr} \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ EA_{n1} \cdots EA_{nr} \\ 0 \cdots 0 \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ 0 \cdots 0 \end{bmatrix}; \quad \mathbf{K} = \begin{bmatrix} 0 \cdots 0 \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ 0 \cdots 0 \\ GA_{11} \cdots GA_{1s} \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ GA_{i1} \cdots GA_{is} \\ \cdots\cdots \\ \cdots\cdots \\ \cdots\cdots \\ GA_{m1} \cdots GA_{ms} \end{bmatrix},$$

where EA denotes the rth environmental attribute measured in the j^{th} environment and GA is the s^{th} genotypic attribute measured in the i^{th} genotype.

References

Allen DM (1971) Mean square error of prediction as a criterion for selecting variables. Technometrics 13:469–495

Cornelius PL (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 PL, Crossa J (1995) Shrinkage estimators of multiplicative models for crop cultivar trials. Technical Report No 352, Department of Statistics, University of Kentucky, Lexington, Kentucky

Cornelius PL, Seyedsadr MS, Crossa J (1992) Using the shifted multiplicative model to search for "separability" in crop cultivar trials. Theor Appl Genet 84:161–172

Cornelius PL, Crossa J, Seyedsadr MS (1993) Tests and estimators of multiplicative models for variety trials. Proc 5th Annu Kansas State Univ Conf, on Appl Statistics in Agric. Manhattan, Kansas

Cornelius PL, Crossa J, Seyedsadr MS (1996) Statistical tests and estimators of multiplicative models for genotype-by-environment interaction. In: Kang, MS, Gauch, HG (eds). Genotype-by-environment interaction, CRC Press, Boca Raton. Florida, USA, pp 199–234

Crossa J, Cornelius PL (1993) Recent developments in multiplicative models for cultivar trials. In: Buxton, DR et al. (eds). International Crop Science I, CSSA, Madison, Wisconsin, USA, pp 571–577

Crossa J, Gauch HG, Zobel RW (1990) Additive and multiplicative interaction analysis of two international maize cultivar trials. Crop Sci 30:493–500

Cubero JI, Flores F (1995) Métodos estadísticos. Junta de Andalucia, (ed) Consejería de Agricultura y Pesca, Sevilla, Spain

Draper NR, Smith H (1966) Applied regression analysis. John Wiley and Sons, New York

Gauch HG (1988) Model selection for yield trials with interaction. Biometrics 44:705–715

Gauch HG, Zobel RW (1988) Predictive and postdictive success of statistical analysis of yield trials. Theor Appl Genet 76:1–10

Gauch HG, Zobel RW (1989) Accuracy and selection success in yield trials. Theor Appl Genet 77:473–481

Gauch HG, Zobel RW (1996) AMMI analysis of yield trials. In: Kang, MS, Gauch HG (eds) Genotype by environment interaction, CRC press, Inc., pp 85–122

Hallauer AR, Miranda JB (1981) Quantitative genetics in maize breeding. Iowa State University press. Ames, Iowa, USA

Royo C, Rodriguez A, Romagosa I (1993) Differential adaptation of complete and substituted triticale. Plant Breed 111:113–119

SAS Institute Inc (1989) SAS/STAT user's guide, ver 6, 4th edn. Cary, North Carolina, USA