Theor Appl Genet (1988) 76: 1–10

Predictive and postdictive success of statistical analyses of yield trials*

H. G. Gauch, Jr. and R. W. Zobel

Department of Agronomy, Cornell University, Ithaca, NY 14853, USA

Received June 30, 1987; Accepted December 17, 1987
Communicated by A. R. Hallauer

Summary. The accuracy of a yield trial can be increased by improved experimental techniques, more replicates, or more efficient statistical analyses. The third option involves nominal fixed costs, and is therefore very attractive. The statistical analysis recommended here combines the Additive main effects and multiplicative interaction (AMMI) model with a predictive assessment of accuracy. AMMI begins with the usual analysis of variance (ANOVA) to compute genotype and environment additive effects. It then applies principal components analysis (PCA) to analyze nonadditive interaction effects. Tests with a New York soybean yield trial show that the predictive accuracy of AMMI with only two replicates is equal to the predictive accuracy of means based on five replicates. The effectiveness of AMMI increases with the size of the yield trial and with the noisiness of the data. Statistical analysis of yield trials with the AMMI model has a number of promising implications for agronomy and plant breeding research programs.

Key words: AMMI – Genotype-environment interaction – Prediction – Soybean – Yield trials

Introduction

Yield trials constitute a major experimental effort in crop production and breeding research. However, the usefulness of yield data is affected greatly by their predictive accuracy. Three options exist for increasing the predictive accuracy of a yield trial:

1 Improved experimental techniques. Accuracy can be increased by larger plots, selection of sites with more uniform soil and more uniform management (application of seed, herbicide, fertilizer and so on). Also, more precise duplication of farmer's fields and management procedures can increase the pertinence of experimental results.

2 More replicates or more sophisticated layout of the replicates. The accuracy of a mean improves, for the completely randomized experimental design, with the square root of the number of replicates (Snedecor and Cochran 1980). For more sophisticated experimental designs, such as the randomized complete block or randomized incomplete block designs, the rate of improvement may be somewhat different.

3 Better statistical analyses. Effective statistical analysis can filter noise from the data pattern, resulting in greater accuracy. Noise is selectively recovered in a model's residual, which is then disregarded when computing the model's expected values.

This paper considers the third option for improving the predictive accuracy of yield trials by means of a specific statistical analysis. In comparison to the first two options, the third option is not as frequently explored by field researchers. Nevertheless, it offers considerable cost effectiveness. A one-time nominal cost may be incurred to obtain and implement a new computer program, but thereafter the computing costs are essentially unchanged (presuming, as is reasonable in this case, that the old and new programs use comparable amounts of computer time). Hence, gains in accuracy by this third option would be practically free.

As already noted, with any given number of replications better statistical analyses can increase accuracy. However, the other way to look at this is that better

* This research was supported by the Rhizobotany Project of the USDA-ARS

statistical analyses can potentially achieve a given level of accuracy with fewer replicates and hence lower attendant costs. In contrast, the first two options invariably increase costs. For example, doubling the number of replicates will generally double costs. Furthermore, limitations in resources often preclude significant advances by the first two options.

Two statistical perspectives

Consider a yield trial with G genotypes grown in E environments (site-year combinations), replicated R times. A genotype-environment combination will be called a "treatment". We now focus on one particular treatment, that is, on one particular genotype, *g*, in one particular environment, *e*, and ask, "What is the yield of genotype *g* in environment *e*?"

Two statistical perspectives are available for estimating this yield: using the full model, namely treatment means, or using some reduced model.

Treatment means

Under this perspective, the only information considered relevant for estimating the yield of genotype *g* in environment *e* is the R yield observations of genotype *g* in environment *e*. Thus, the best yield estimate is the average of these R replicates, that is, the treatment mean (perhaps with an adjustment for block effects if a randomized incomplete block design is used).

The error mean square (EMS) is an unbiased estimate of the mean square difference between individual replicates and the true population means. The standard error of the mean, equalling EMS divided by the square root of the number of replicates, estimates the root mean square difference between treatment means and the true population means (Snedecor and Cochran 1980). The EMS is estimated with GE(R-1) degrees of freedom (*df*). Ordinarily the number of *df* is rather large, and consequently the standard error of the treatment means is estimated accurately.

The most common format for reporting the results of a yield trial is to present the treatment means, perhaps ranked, and often accompanied by the standard error of the mean, a least significant difference (LSD) value, or a multiple range test. This format exemplifies the treatment means perspective.

Modelling

The modelling perspective takes a broader view of yield estimation. A multivariate statistical model relates all other yields to the yield of interest. Consequently, the information relevant to estimate the yield of genotype *g* in environment *e* is the entire yield trial. Not only are the data on genotype *g* in environment *e* relevant, but also the data on genotype *g* in other environments, the data on other genotypes in environment *e*, and even the data on other genotypes in other environments. Every datum from the entire yield trial bears upon the yield estimate of genotype *g* in environment *e*. The model produces a residual that selectively recovers the noise, which is discarded in order to reduce the deleterious influence of noise. Several statistical models and two success criteria are considered here in diagnosing the best model for a given data set.

Three similarities between these two perspectives should be recognized before proceeding to note dissimilarities. First, both statistical perspectives recognize the fundamental problem that yield data are noisy: "data = pattern + noise" (Freeman 1973). We desire to know the population or true means, that is, the true pattern; however, the observational data fall short because of error or noise. Regardless, whether the treatment means or modelling perspective is taken, the fundamental problem of noisy data exists.

Secondly, both statistical perspectives may use various experimental designs (such as randomized complete blocks) to identify and remove sources of variation not of interest. Likewise, both may use outlier detection to remove exceptionally bad data. The choice between the treatment means or modelling perspective is independent of the options concerning experimental design and outlier detection.

Thirdly, both statistical perspectives differentiate treatment design from experimental design. Treatment design specifies the treatments included in an experiment, whereas experimental design assigns treatments to the available experimental units. Treatment design is directed at efficient pursuit of interesting research questions, whereas experimental design is directed at error control. For both perspectives, the need remains to choose a worthwhile treatment design – to test promising genotypes at relevant experimental sites, to design factorials of fertilizers or other treatments that include agriculturally significant research objectives, and so on. The significance of this work for treatment design is its suggestion that in appropriate situations, effective statistical analysis can reduce the needed amount of replication by a factor of about 2–5. Consequently, with the same resources (number of experimental units), treatment designs having 2–5 times as many treatments may be considered. The result is the potential for an increase in the breadth or efficiency of a research program.

Two criteria of successful analysis

Model evaluation can be based upon either postdictive or predictive success. "Prediction by its derivation (L. *praedicere,* to say before) means literally the stating beforehand of what will happen at some future time" (Aitchison and Dunsmore 1975). In contrast, postdiction is making "an assertion or deduction about something in the past" (Burchfield 1982). As applied to the context of statistical models, the essential distinction is not the literal matter of whether an event occurs in the future or the past, but rather whether assessment of the statistical model's success involves the same data as were used to construct the model or alternatively new data not used in model constuction (Blackburn 1973; Aitchison and Dunsmore 1975; Gauch 1988).

In postdiction, data are used to construct a model and then the model's expected values are compared with the same data. Postdictive success can be measured by statistics such as the average absolute difference on the root mean square difference between observed and expected values, or the percentage of the total sum of squares (SS) captured by the model. For example, the F-test in an analysis of variance (ANOVA) measures postdictive success: a source, such as genotype means, is declared significant at a given level if its mean square exceeds the error mean square by the appropriate F-test ratio. The sources included in an ANOVA model thus account for a given percentage of the total SS.

A postdictively successful model captures as much of the total SS as possible with as few *df* as possible. This constraint of parsimony is required because otherwise the winner is always the trivial full model of treatment means having GE-1 *df*. Postdictive success requires both accuracy and parsimony.

In prediction as implemented here, the data are first partitioned into two parts: the model data and the reserved or validation data. The model data are used to construct a model and then the expected values predicted by the model are compared with the validation data. Predictive success can be measured by statistics such as the average absolute difference or the root mean square difference between validation data and model predictions. For example, a predictively successful model of a yield trial might have a predictive root mean square error of 150 kg/ha representing 5% of a grand mean of 3,000 kg/ha. Predictive success also requires both accuracy and parsimony.

The distinction between postdiction and prediction is not merely philosophical, but is also practical. That the data are noisy implies, both philosophically and practically, that these two criteria are different. The model data can be regarded as "$\text{data}_1 = \text{pattern} + \text{noise}_1$" and the validation data as "$\text{data}_2 = \text{pattern} + \text{noise}_2$". The pattern is stable, agronomically meaningful and has predictive value, whereas the noise is idiosyncratic, uninterpreted and of no predictive value.

For example, if the treatment SS is composed of 70% pattern and 30% noise, the goal of statistical analysis should not be to recover 100% of the SS, but rather only 70%, and more specifically the 70% that represents pattern. Recovery of pattern increases the accuracy of predictions but recovery of noise decreases accuracy. Consequently, a model which selectively recovers pattern, while selectively relegating noise to a disregarded residual, can have better predictive accuracy than its own data (Gauch 1982, 1985, 1988). Such a model's predicted values are closer than its data to the true population means. In contrast, the goal in postdictive success is to recover as close to 100% of the treatment SS as possible while still maintaining some regard for parsimony, and therefore the model can have *worse* postdictive accuracy, but never better.

Another important consequence of the choice between postdictive and predictive criteria of success is that, in general, they lead to diagnosis or selection of different models (Gauch 1985). Thus, the model selected for the best postdictive success is usually not also the best model for prediction.

A thorough statistical explanation for the ability of multivariate models, such as the Additive main effects and multiplicative interaction model (AMMI), to selectively recover pattern must be left to the references (see Gauch 1982, 1988, and the citations therein). In essence, selectivity occurs because noise produces idiosyncratic variations in individual treatment means, whereas pattern originating from properties of the genotypes and sites produces coordinated variations in numerous treatment means. From a mathematical viewpoint, idiosyncratic variations represent high-dimensional information that eigenanalysis recovers with a large number of small eigenvalues, whereas coordinated variations represent low-dimensional information that eigenanalysis recovers with a small number of large eigenvalues. Hence the early (large) eigenvalues of the principal components analysis (PCA) component of AMMI selectively capture pattern, whereas the late (small) eigenvalues selectively recover noise. Actually, the additive ANOVA component of AMMI can recover only relatively low-dimensional information or pattern from the treatment means, because the genotype and environment additive main effects contain only G+E-2 *df*, which is small relative to the GE-1 *df* contained in treatments.

Given that postdiction and prediction are genuinely different tasks, which success criterion is most relevant for yield trial analyses? As Student observed in 1923: "the object of testing varieties of cereals is to find out which will pay the farmer best." As D. V. Lindley asks more recently: "what is an agricultural experiment ... but and aid to forecasting the yield that a farmer might get from a new variety or insecticide?" (in Harrison and Stevens 1976; also see Aitchison and Dunsmore 1975). Yield trials actually measure past yields on experimental plots, but the purpose is to improve future yields on farmer's fields. Clearly the purpose of yield trials is predictive. Consequently, analysis of yield data should emphasize predictive success.

Materials and methods

A New York soybean yield trial is analyzed as an example. The yield data and details of the field methods are available in reports from the Department of Agronomy, Cornell University. Dr. M. Wright kindly made available the original data on individual replicates. The subset of these data analyzed here has 7 genotypes grown in 40 environments (certain combinations of 9 sites and 9 years). Yields are expressed in kg/ha at 13% moisture content. Most trials had 4 replicates, but due to occasional problems some had only 2 or 3; of the 1,120 plots planted, 1,044 plots were harvested. This represents 93% of the possible total, or an average of 3.73 replications. A randomized complete block design was used. However, the data were analyzed as a completely randomized design because of missing observations and the lack of software to fit some of the statistical models to unbalanced data. The resultant approximation should be very close. Furthermore, interest focuses here upon predictive success rather than postdictive success, which uses a random partition between model data and validation data and hence discards blocking.

The effectiveness of several statistical strategies was examined for variations in three factors:

1 Statistical model. The treatment means model, which is the full model of treatment averages over replicates, served as the baseline model. Several reduced models were compared: the additive ANOVA model, the multiplicative principal components analysis (PCA) model, the Finlay and Wilkinson (1963) regression model (amended to consider not only the original regressions of genotype yields on environment means, but also the complementary regressions of environments on genotype means, as well as the 1 *df* for joint regression or concurrence; Wright 1971), and the Additive main effects and multiplicative interaction (AMMI) model (Gollob 1968; Mandel 1971; Bradu and Gabriel 1978; Kempton 1984; Gauch 1985, 1988; Zobel et al. 1988). The AMMI algorithm is used frequently to produce biplots (graphs showing relationships among both the genotypes and the environments; Bradu and Gabriel 1978; Kempton 1984). All models were computed by the FORTRAN77 program MATMODEL (Gauch 1987). These statistical models are regarded as fixed effects models because all inferences pertain to specific soybean genotypes and specific New York sites, rather than to genotypes or sites in general.

2 Success criterion. Postdictive success was measured by the root mean square difference between the observed and expected values (that is, by the square root of the EMS), and by the percentage of the treatment SS accounted for by the model. Predictive success was measured by the predictive root mean square difference between validation data and model predictions. In both cases, parsimony (a small number of *df* in the model) was also a relevant consideration.

3 Number of replications. Models were constructed using 1 replicate, 2 replicates, and all the data or 3.73 replicates. Models with 1 replicate used 280 yield observations for constructing the model and 764 for validation; 2 replicates provided 560 observations for modelling and 484 for validation; and 3.73 replicates used all 1,044 observations for modelling. In this last case, no calculation of predictive success was possible.

To check the stability of the predictive accuracy calculations, five random partitions between model data and validation data were formed and the results compared. For example, for models using 1 replicate, 1 replicate was selected at random from each treatment and the remaining replicates were set aside as validation data. This procedure was repeated 5 times.

Results

First consider the postdictive success of various analyses.

Table 1 shows the ANOVA for the entire soybean yield trial, consisting of 1,044 yield observations on 7 genotypes in 40 environments, averaging 3.73 replicates. These data have a total of 1,043 *df.*

The treatment means model partitions this total *df* into only two sources: treatments (combinations of 40 environments and 7 genotypes) with 279 *df,* and error with 764 *df* (that is, between and within treatments, as shown in the second and last lines of Table 1). The error MS is 104,748, meaning that the root mean square difference between yield observations and the true means is estimated to be 323.6 kg/ha. This represents 12.9% of the grand mean of 2,518 kg/ha. For means without missing data and hence based on 4 replications, the standard error of the mean is 161.8 kg/ha. The treatment means are significant, with an F-ratio of 5.67. The treatment means model cannot possibly be improved for postdictive accuracy, but because this full model lacks parsimony it is of interest to consider other models. A parsimonious, reduced model that uses fewer *df* and yet retains most of the SS will have greater heuristic value than the full model.

The ANOVA model partitions the treatment *df* and SS into three sources: additive environment effects, additive genotype effects and the genotype-environment GE interaction (that is, the non-additive residual from the additive ANOVA model). These sources

Table 1. ANOVA for soybean yield trial

Source	*df*	SS	MS	F
Total	1043	245722327		
Treatment	279	165694661	593888	5.67
Environment	39	129469970	3319743	31.69
Genotype	6	9722960	1620493	15.47
G×E	234	26501731	113255	1.08
PCA 1	44	18632502	423466	4.04
Residual	190	7869229	41417	0.40
Error	764	80027666	104748	

contain 78%, 6% and 16% of the treatment SS, respectively (see Table 1). The environment and genotype additive effects are highly significant, but the interaction, with an F-ratio of only 1.08, is not significant at the 5% level. Temporarily ignore the entries in Table 1 for PCA 1 and Residual which pertain to the AMMI analysis discussed momentarily.

The PCA multiplicative model has 46, 44 and 42 *df* in the first 3 PCA axes, accounting for 79%, 12% and 6% of the treatment SS. These 3 axes are highly significant, leaving a residual (consisting of the 4th and higher axes) with 147 *df* and an insignificant F-ratio of 0.40. By comparison, the PCA model with 1 axis and the ANOVA are comparable in parsimony (46 and 45 *df*, respectively), but the ANOVA model is slightly more effective in extracting SS (79% and 84%, respectively). For these data the ANOVA model is better. Probably more important than this marginal statistical superiority is the consideration that the ANOVA identifies genotype means as a source whereas PCA does not (since the PCA model has no additive terms), and yet genotype means are ordinarily the most interesting result emerging from a yield trial.

Finlay-Wilkinson regression analysis is ineffective for this data set. The genotype regressions have a mean square (MS) of 20,472, giving an F-ratio of only 0.20. The environment regressions have a MS of 40,278, which again is much less than the error MS of 104,748. Only the joint regression (for concurrence), with 1 *df* and a MS of 520,839, is significant at the 5% level. However, since the joint regression accounts for only 2% of the interaction SS, it is not of interest. The resulting residual with 190 *df* has a larger MS than does the interaction itself, and its F-ratio of 1.22 is significant at the 5% level. For a model to produce a residual more significant than the source from which the residual is taken constitutes the ultimate in model failure.

One feature of the ANOVA analysis in Table 1 is disturbing. The interaction is judged insignificant because of its low MS. However, the interaction SS is almost three times as large as the genotype SS (and the genotype MS is highly significant). This disparity between MS and SS viewpoints arises because the genotype and interaction sources have 6 and 234 *df* respectively, thus differing in *df* by an enormous factor of 39. Although the Finlay-Wilkinson regressions happen not to partition useful sources from the interaction, this large interaction SS should stimulate other attempts to partition this interaction. Even if only a third of this interaction SS were modelled successfully, this third would have as large an impact upon model predictions as does the entire main effect for genotypes.

Failure of the Finlay-Wilkinson model to effectively partition the interaction does not prove that no analysis of the interaction can work. Only if the interaction SS is so small that it would be insignificant even if captured completely in a source with 1 *df* can one conclude from the outset that an effective analysis of the interaction cannot be found.

AMMI partitions the 234 *df* in the GE interaction, producing a first interaction PCA axis with 44 *df* containing 70% of the interaction SS, giving a MS of 423,466 and a highly significant F-ratio of 4.04 (Table 1). The remaining interaction PCA axes have F-ratios of less than 1, so they may be combined into a residual with 190 *df*, a MS of 41,417 and an F-ratio of 0.40. Consequently, the partitioning of the interaction SS by AMMI is extremely effective in finding structure within the interaction. The SS of this significant first interaction PCA axis is almost twice as large as is the SS for genotype means, so it greatly affects the AMMI model's predicted values.

The AMMI analysis is statistically efficient and agronomically interpretable. The genotype and environment main effects are identical with ANOVA and have the usual interpretations of overall yield or productivity. The first interaction PCA axis for genotypes is clearly related to maturity groups. Soybean cultivars with higher maturity group ratings require longer growing seasons. Group 0 cultivars were at one end of the first interaction PCA axis, group II cultivars at the opposite end, and group I cultivars in between. Correspondingly, the environments (site-years) are arranged from cool and short growing seasons to warm and long seasons.

The postdictive results for data subsets based on only 1 or 2 replicates, instead of all 3.73 replicates as considered thus far, are much the same. Hence they are not described here.

Next consider the predictive success of various statistical analyses as applied to various subsets of this data set.

The analyses considered are (1) treatment means, (2) ANOVA, and (3) AMMI with 1 or 2 replicates used to construct the model. The Finlay-Wilkinson model's predictions are nearly equal to the ANOVA predictions simply because the Finlay-Wilkinson model was so ineffective, and are not considered further.

A sequence of models is produced by AMMI in that 1 to many PCA axes may be retained in the model and the remaining axes relegated to the residual (Gauch 1985). For the 5 random subsets using only 1 replicate, predictive success was best for AMMI with 1 PCA axis. For the 5 random subsets using 2 replicates, in 4 cases 1 PCA axis was best, and in 1 case 2 PCA axes were best. However, in this last case the improvement over 1 PCA axis was under 1%, so for simpicity 1 PCA axis was used throughout.

The results are shown in Table 2. With unreplicated data the ANOVA model gave a small improvement

Table 2. Predictive root mean square error in kg/ha for treatment means, ANOVA and AMMI models, using 1 or 2 replicates

Model	1 Replicate	2 Replicates
Treatment means	455.2	391.9
ANOVA	438.4	424.6
AMMI	382.1	354.0

Table 3. Estimated root mean square in kg/ha between model predictions and the true population means

Model	1 Replicate	2 Replicates
Treatment means	320.0	220.9
ANOVA	295.7	274.8
AMMI	203.0	143.3

over treatment means, but with 2 replicates ANOVA was worse. The explanation derives from a positive and a negative feature of ANOVA. It improves estimation of the additive component of the data, but it eliminates the nonadditive interaction component of the data. Unreplicated data are so noisy that ANOVA improves the estimation of the additive component more so than it hinders the estimation of the poorly-defined interaction component. However, with 2 replicates the treatment means are more accurate and therefore less improvement of the additive component can occur, whereas the data are now accurate enough to carry important interaction information, which ANOVA discards.

A striking feature of Table 2 is that AMMI with unreplicated data was more predictively accurate than treatment means based on 2 replicates. More predictive accuracy was gained by using the AMMI model than by doubling the amount of data!

The predictive errors of Table 2 reflect differences between single replicate validation observations and model values, or model-to-validation differences. However, the validation observations and the model predictions are both imperfect, departing from the true population mean μ for each treatment. Hence the model-to-validation differences have two error components: validation-to-μ and model-to-μ.

The variance rule says that the model-to-validation variance equals the validation-to-μ variance plus the model-to-μ variance. The quantity of real interest is the model-to-μ variance, telling how well the model itself is doing apart from errors in the validation data. This model-to-μ variance can be estimated as the model-to-validation variance (from Table 2) minus the validation-to-μ variance.

The validation-to-μ difference, expressed in terms of root mean square, is the square root of the EMS because the EMS is an unbiased estimator of the mean square difference between individual replicates (such as the individual validation observations) and the true population mean μ. Here the error mean square of 104,748 is estimated rather accurately with 764 *df* (Table 1), and its square root is 323.6 kg/ha.

The model-to-μ differences can thus be estimated by removing the validation-to-μ root mean square of 323.6 from the model-to-validation differences in Table 2 (by taking the square root of the result of the model-to-validation MS minus the validation-to-μ or error MS). The results (Table 3) show that AMMI with 2 replicates has a model-to-validation mean square error (from Table 2) of 354.0, whereas the corresponding model-to-μ error (from Table 3) is only 143.3. In terms of variance, the model-to-μ variance of 20,535 and the validation-to-μ (or error mean square) of 104,748 combine to give the model-to-validation total variance of 125,283. Of this total variance, 84% is due to the validation data and only 16% to the model itself. The validation observations are used individually without calculating any means, so these unreplicated observations are naturally noisier than the AMMI model based on 2 replicates.

Given data for 2 replicates, AMMI has a predictive root mean square error of 143.3 kg/ha (Table 3). Relative to the yield trial's grand mean of 2,518 kg/ha, this represents an error of only 5.7%.

The theoretical standard error of the mean for various numbers of replicates is based on a root error mean square of 323.6 (Table 4). These values are simply 323.6 divided by the square root of the number of replicates. Table 4 also shows the theoretical prediction root mean square, obtained by adding in the variance of single replicate validation observations (the exact reverse of the process used in deriving Table 3 from Table 2).

Note that theoretical values in Table 4, based merely on the error MS, are close to their corresponding observed values in Table 2. For example, for unreplicated data the theoretical 457.6 is close to the observed 455.2, and for 2 replicates 396.3 is close to 391.9.

Other observed values of predictive error in Table 2 can be compared with the last column in Table 4 to determine how many replicates are needed to obtain the same predictive accuracy as if replicates are merely averaged to estimate treatment means (in contrast to the alternative of applying the AMMI model). For example, Table 2 shows that unreplicated data analyzed by AMMI has a predictive error of 382.1, which from Table 4 is equivalent to treatment means based on 2.5 replicates. Likewise, 2 replicates with AMMI are

Table 4. Theoretical standard error of the mean and prediction root mean square error in kg/ha for different numbers of replicates, based on root error mean square of 323.6

Replicates	SE of mean	Prediction RMS
1	323.6	457.6
2	228.8	396.3
2.5	204.7	382.9
3	186.8	373.7
3.73	167.6	364.4
4	161.8	361.8
5	144.7	354.5
8	114.4	343.2
16	80.9	333.6
∞	0.0	323.6

equivalent in predictive accuracy to 5 replicates without AMMI. Application of AMMI to this soybean yield trial gives an increase in predictive accuracy equivalent to that resulting from increasing the number of replications by a factor of 2.5. Thus, the number of replicates could be reduced by a factor of 2.5 and the same predictive accuracy maintained simply by running the data through the AMMI analysis. To state the advantage of using AMMI in another way, predictive errors were reduced to 63% (since the reciprocal of the square root of 2.5 is 0.63). For example, for 2 replicates the treatment means have a predictive error of 8.8% of the grand mean, compared to only 5.7% with AMMI (based on Table 3).

One last matter is the predictive accuracy of the best model for these data, based on AMMI analysis of all 3.73 replicates. No reserved validation data remain, however, to measure this accuracy empirically. All that can be said for sure is that this AMMI model using 3.73 replicates is better than the AMMI model based on only 2 replicates, shown earlier to have a root mean square predictive error of 5.7%. However, from the last line of Table 3 it appears that the AMMI model is, like the treatment means model, improving with the square root of the number of replications. If this pattern can be extrapolated, then the AMMI model with 3.73 replications has a predictive root mean square error of about 4.2% of the grand mean. However, there is no theoretical requirement that the AMMI model must behave just like the statistically simple treatment means model, so the most cautious evaluation is merely that this best model has less than 5.7% error.

Discussion

Results presented here concern a single case: a New York soybean yield trial. The AMMI model based on unreplicated data produced an improvement in predictive accuracy equivalent to increasing by a factor of 2.5 the number of replicates when using treatment means. Statistical theory leads to two generalizations regarding what might be expected on other cases:

1. The benefit from AMMI will increase with the size of the yield trial, that is, with increase in the number of genotypes and number of environments. More data improves the performance of a multivariate analysis like AMMI.

2. The benefit from AMMI will increase with the noisiness of the data. Were the data perfect (noiseless), no improvement in predictive accuracy could result from AMMI (or from any other analysis), but the greater the noise, the greater the opportunity for improvement. The greater the noise, the more divergent are the postdictive and predictive perspectives, and hence the greater the importance of using a distinctively predictive assessment of success.

The factor of 2.5 observed here is thus not a statistical constant, but rather the factor observed in this particular yield trial. For a more noisy international corn yield trial, a factor of 4.3 was observed (unpublished results). Experience with a number of yield trials will be necessary to refine the generalizations, but the above two points of statistical theory indicate what may be expected.

One feature that increases noise is conducting a yield trial with a small number of replicates, especially the extreme case of an unreplicated trial. However, in this case no measurement of predictive success is possible because no reserved data are available, and consequently it is not possible to empirically select the best number of interaction PCA axes to include in the AMMI model. Inspection of the interaction PCA axes in terms of agronomic interpretability is a reasonably good substitute. Likewise, a decision could be based on extrapolation from other experiences with replicated trials having similar crops and conditions. When interaction is obviously important, at least 1 interaction PCA axis is needed. However, agricultural data will very rarely support more than 2 axes (as indicated in the studies cited by Gauch 1985). Furthermore, the differences in yield estimates between these two adjacent models, having 1 or 2 interaction PCA axes, will often be very small. Consequently, adequate model diagnosis is relatively easy. AMMI applied to a large and noisy yield trial without replication frequently can be expected to result in a predictive accuracy equal to means based on several replicates.

It might be objected that it wastes data to use only part of it for modelling, since part is reserved for validation. Three responses are relevant.

1. If the interest in yield trials (or more generally in science) includes predictive accuracy, then by definition

there is no option to using part of the data for validation. By definition, predictive accuracy is assessed by exposing the model to new data – to data not used in making the model. The alternative is to abandon interest in predictive accuracy, but this goes against the purposes of yield trials in particular and of science in general.

2. The data must be partitioned into modelling data and validation data for the purpose of diagnosing or selecting the predictively best model. However, after this step is done a final model may be fitted to the entire data set. Consequently no data are wasted. The diagnosis based on part of the data can be retained for the entire data set because the modest increase in the data's accuracy can make, at most, a small, inconsequential change in the ideal model diagnosis. The predictive accuracy of the final model cannot be measured, but it is known to be better than the measurement resulting from the previous model on part of the data.

3. The real problem is that inappropriate statistical analyses waste data. For example, assume that a yield trial has 4 replicates and that the AMMI increase in accuracy is equal to a factor of 2.5 increase in the number of replicates. The AMMI predictions based on 4 replicates are thus as accurate as mere treatment means based on 10 replicates. Effectively used, the AMMI analysis gives an extra 6 replicates for free. Compared to AMMI, the treatment means analysis effectively uses only 4 of 10 replicates and wastes 6 replicates.

It may be objected that measuring predictive success by comparing replicates from the same yield trial is different from the ultimate goal of extending predictions to new sites and new years. This objection is entirely valid. Inference to new sites and new years, rather than merely to new replicates, encounters additional problems and hence is more challenging (Cady and Allen 1972; Wood and Cady 1981). However, four matters of perspective may be mentioned.

1. Within-trial predictive success to new replicates is the penultimate rather than the ultimate goal of between-trial predictive success to new sites and new years. Nevertheless, it is moving in the right direction as compared with the traditional postdictive assessment. Furthermore, although the ultimate goal entails conducting involved experiments, this penultimate goal merely requires replication, a condition fulfilled in most yield trials for other reasons. Hence this first step in the right direction is virtually cost free.

2. The recommended estimate of predictive error is calculated from a model using only part of the data, and hence constitutes a large (conservative) estimate of the predictive error for the final and more accurte model using all of the data. This compensates in part or in whole for the increase in predictive error associated with the extension of the inference from new replicates to new fields.

3. Inference from a researcher's yield trial to a farmer's yield involves a new site and a new year. Many experimental sites are selected in order to well represent the range of soils and conditions encountered on farms within the given study area. Therefore a farmer can refer to results from a relevant experimental site. In comparison to these site-to-site variations in yield, the year-to-year variations are often significantly larger (Talbot 1984). Therefore, even if the experimental site-to-farm site difference were reduced to zero, effectively making the measurements into replicates, the consequence may be an insignificant reduction in the total year and site-to-year and site variability. Having accepted the inherent and often considerable unpredictability of next year's weather, the differences between farmers' fields and the most closely matching experimental fields are of little practical interest unless this match is quite poor. If the match is evidently poor, then no inference is justified regardless of the statistical analysis chosen.

4. Yield trials are used primarily to select the best genotypes (or fertilizer levels or whatever). Clearly the achieved accuracy of yield estimates affects success in this selection process (Talbot 1984). The greater accuracy and more appropriate assessment of GE interaction provided by AMMI greatly affects selections. Soybean and corn yield trials, as well as simulation studies, show that the AMMI model selects a different highest-yielding genotype than does treatment means in over half of the trials' environments (unpublished results). Better selections will increase the rate of progress in breeding programs and will also increase the reliability of variety recommendations given to farmers. A subsequent paper will present this research in detail.

Blocking is a routine statistical tool for removing variability due to a field's heterogeneity, and can be especially effective when the blocks are placed in accordance with known gradients in slope or soil texture. The randomized complete block (RCB) is doubtless the most popular experimental design for yield trials. However, blocking is a dubious means of error control in the presence of GE interaction (Gusmão 1985, 1986). In contrast, AMMI frequently provides excellent error control when GE interaction is present. Furthermore, the typical gain in accuracy from AMMI is shown to be much larger than that from blocking. Researchers consider blocking successful when, compared to the completely randomized design, blocking gives a relative efficiency factor of say 1.2. This means that using 5 replicates with blocking is as accurate as $5 \times 1.2 = 6$ replicates without blocking (Snedecor and Cochran 1980). Likewise, a factor of 1.4 is very good and 1.6 is remarkable, equating to 7 or 8 replicates. By comparison, the efficiency factor of 2.5 observed here using AMMI with soybean data would equate to $5 \times 2.5 = 12.5$ replicates, and the factor of 4.3 with the noisier corn data would equate to 21.5 replicates. Hence blocking typically gives 1–3 "free" replicates, whereas AMMI gives 7–16.

AMMI and blocking provide radically different approaches for error control, as is best appreciated by understanding that they concern different sources or *df*. A yield trial with G genotypes, E environments and R replications has GER-1 *df*. The most fundamental partition attributes GE-1 *df* to treatments (genotype and environment combinations) and the remaining GE(R-1) *df* to error (originating from replication). The treatment and error *df* and SS are statistically orthogonal. AMMI partitions a residual from the treatment *df*, or more specifically from the interaction *df*. Error control is provided by discarding this noise-rich residual. In contrast, blocking partitions the blocks *df* from the error *df*. Error control may be provided by removing interblock variation, making the error mean square smaller and hence increasing the F-ratio test statistics used to test model sources for significance. Because these two error control options are applied to different and orthogonal sources of variation, either or both options may be used as desired.

A statistical model of a yield trial may be visualized as a response surface, where yield is the response plotted as a function of genotype and environment (Bradu 1984; Gauch 1985; Gregorius and Namkoong 1986). Model diagnosis then amounts to finding a response surface which fits well the empirical data points. Gregorius and Namkoong (1986) present an alternative approach for model diagnosis to that of Bradu and Gabriel (1978), but we prefer the latter approach because its statistical theory has been developed extensively.

The modelling approach presented here involves two statistically independent issues: (1) diagnosing an appropriate statistical model for a given data set; and (2) assessment of predictive success using data splitting. If an ANOVA or Finlay-Wilkinson model best fits a given data set, then this model, rather than AMMI, will give superior predictive success. However, as emphasized by Bradu and Gabriel (1978) and Zobel et al. (1988), even if some other model is best for a given case, an initial AMMI analysis is ideally suited for rapidly diagnosing this best model. The encouraging results with AMMI presented here are not to be understood as a general recommendation of AMMI for all yield trial data sets, although it is expected that AMMI is best for a majority of yield trials having a significant interaction. On the other hand, the conclusions regarding the distinction between postdictive and predictive success apply in general, assuming only that the data are noisy.

The various advantages of AMMI and the variety of yield trial situations imply a variety of applications, but one application is expected to be most prevalent. Given a large yield trial conducted annually, with significant GE interaction and with 4 or 5 replicates, use of AMMI will frequently give superior predictive accuracy despite a reduction to only 2 replicates. Consequently, the breadth of future research programs can be doubled without loss of accuracy. In cases where currently available data are still in use for purposes of making recommendations, reanalysis of the data by AMMI should improve the predictive accuracy and hence the merit of the recommendations.

Acknowledgements. We appreciate helpful suggestions from Chr. McVoy, M. Smith, M. Sorrells, D. Wallace, and the reviewers.

References

Aitchison J, Dunsmore IR (1975) Statistical prediction analysis. Cambridge University Press, Cambridge

Blackburn S (1973) Reason and prediction. Cambridge University Press, Cambridge

Bradu D (1984) Response surface model diagnosis in two-way tables. Commun Stat Theory Methods 13:3059–3106

Bradu D, Gabriel KR (1978) The biplot as a diagnostic tool for models of two-way tables. Technometrics 20:47–68

Burchfield RW (1982) A supplement to the Oxford English Dictionary. Oxford University Press, Oxford

Cady FB, Allen DM (1972) Combining experiments to predict future yield data. Agron J 64:211–214

Finlay KW, Wilkinson GN (1963) The analysis of adaptation in a plant-breeding programme. Aust J Agric Res 14: 742–754

Freeman GH (1973) Statistical methods for the analysis of genotype-environment interactions. Heredity 31:339–354

Gauch HG (1982) Noise reduction by eigenvector ordinations. Ecology 63:1643–1649

Gauch HG (1985) Integrating additive and multiplicative models for analysis of yield trials with assessment of predictive success, Mimeo 85-7. Department of Agronomy, Cornell University, Ithaca/NY

Gauch HG (1987) MATMODEL. Microcomputer Power, Ithaca/NY

Gauch HG (1988) Model selection and validation for yield trials with interaction. Biometrics (in press)

Gollob HF (1968) A statistical model which combines features of factor analytic and analysis of variance techniques. Psychometrika 33:73–115

Gregorius HR, Namkoong G (1986) Joint analysis of genotypic and environmental effects. Theor Appl Genet 72: 413–422

Gusmão L (1985) An adequate design for regression analysis of yield trials. Theor Appl Genet 71:314–319

Gusmão L (1986) Inadequacy of blocking in cultivar yield trials. Theor Appl Genet 72:98–104

Harrison PJ, Stevens CF (1976) Bayesian forecasting. J R Stat Soc Ser B 38:205–247

Kempton RA (1984) The use of biplots in interpreting variety by environment interactions. J Agric Sci 103:123–135

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

Snedecor GW, Cochran WG (1980) Statistical Methods, 7th edn. Iowa State University Press, Ames/IA, pp 44–45, 264–265

Student (1923) On testing varieties of cereals. Biometrika 15: 271–293

Talbot M (1984) Yield variability of crop varieties in the UK. J Agric Sci 102:315–321

Wood CL, Cady FB (1981) Intersite transfer of estimated response surfaces. Biometrics 37:1–10

Wright AJ (1971) The analysis and prediction of some two factor interactions in grass breeding. J Agric Sci 76: 301–306

Zobel RW, Wright MJ, Gauch HG (1988) Statistical analysis of a yield trial. Agron J (in press)