Theor Appl Genet (1990) 80:153–160

Full and reduced models for yield trials*

H. G. Gauch, Jr.

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

Received November 15, 1989; Accepted March 23, 1990
Communicated by A. R. Hallauer

Summary. Empirical results routinely demonstrate that the reduced Additive Main effects and Multiplicative Interaction (AMMI) model achieves better predictive accuracy for yield trials than does the full treatment means model. It may seem mysterious that treatment means are not the most accurate estimates, but rather that the AMMI model is often more accurate than its data. The statistical explanation involves the Stein effect, whereby a small sacrifice in bias can produce a large gain in accuracy. The corresponding agricultural explanation is somewhat complex, beginning with a yield trial's design and ending with its research purposes and applications. In essence, AMMI selectively recovers pattern related to the treatment design in its model, while selectively relegating noise related to the experimental design in its discarded residual. For estimating the yield of a particular genotype in a particular environment, the AMMI model uses the entire yield trial, rather than only the several replications of this particular trial, as in the treatment means model. This use of more information is the source of AMMI's gain in accuracy.

Key words: Accuracy – AMMI – Prediction – Stein effect – Yield trials

Introduction

Yield estimates can be based upon the full treatment means or cell means model, which simply uses the averages over replications, or alternatively upon a reduced model, such as the Additive Main effects and Multiplicative Interaction (AMMI) model, which produces adjusted means because of a discarded residual. By splitting the yield data at random, using part for model construction and part for model validation, various models can be compared in terms of their predictive success by a statistic such as the root mean square difference between model estimates (predictions) and validation observations. Routinely empirical results demonstrate that the reduced AMMI model achieves better predictive accuracy than does the full treatment means model (Gauch 1988; Gauch and Zobel 1988, 1989).

It may seem mysterious that treatment means are not the best, most predictively accurate estimates. Nevertheless, the reasons underlying AMMI's predictive success are well understood, although they are somewhat complex, because the explanation necessarily touches upon numerous concepts and issues from the yield experiment's design forward. Also, a meaningful explanation must clarify agricultural as well as statistical issues. This paper attempts to provide a clear theoretical explanation for the documented empirical fact that AMMI yield estimates are often more predictively accurate than treatment means. Often the AMMI model is more accurate than its data.

Yield trial design

A designed experiment, such as a yield trial, has two components: treatment design and experimental design. Given R replications for T treatments, there are RT experimental units – here referred to as yield plots – for which yield is observed. The treatment design concerns the choice of imposed treatments of agricultural interest – here a two-way factorial design concerning yield for G genotypes grown in E environments (site-year combina-

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

tions), for a total of $T = GE$ treatments. On the other hand, the experimental design concerns the allocation of experimental units (yield plots) to the treatments, ordinarily involving randomization and replication of some sort. It serves to validate the subsequent application of an appropriate statistical analysis for the potential purpose of drawing significant inferences from the data (and of making rational decisions).

This fundamental two-fold design – treatment design and experimental design – is reflected in the first, most fundamental partition in an analysis of variance (ANOVA) table. The source for the "Total" variation with $RT-1$ degrees of freedom (df) after customarily removing the grand mean is first partitioned into "Treatments" with $T-1$ df and "Error" with $(R-1)\,T$ df. The source for treatments reflects variation due to the treatment design, whereas the source for error reflects variation due to the experimental design (replication).

If the treatment design has some structure (such as the two-way factorial here of genotypes by environments), or if certain treatment comparisons have special interest, then the $T-1$ df for treatments may be further partitioned by an appropriate statistical model. Likewise, if the experimental design has some special structure (such as a randomized complete block or a lattice design, rather than merely the completely randomized design), then the $(R-1)\,T\,df$ for error may be further partitioned (such as $B-1$ df for "Blocks" and the remainder for "Pure error" with the randomized complete block design).

The error mean square

For the moment, however, let us focus on the simplest possible case involving no further partitioning: the treatments are merely modelled by their treatment means (or "cell means"), and likewise the error is not partitioned (presumably reflecting a completely randomized design).

Five kinds of quantities are of interest. (1) Y_{ger} are the individual yield observations for genotype g, environment e, and replicate r of replications 1 to R used for constructing the model. (2) $\bar{Y}_{ge}$ is the empirical treatment mean, namely, the averages over replications 1 to R, for genotype g in environment e. (3) μ_{ge} is the true or population mean for genotype g in environment e. It is, of course, a theoretical, unobservable quantity. (4) $\hat{Y}_{ge}$ is the yield estimate calculated by a statistical model, based upon some or all of the GER observations Y_{ger} allocated for model construction. Of course, several different statistical models may be used, producing several such estimates. (5) Y_{gef} are the individual yield observations for genotype g, environment e, and future or new replicate f of replications 1 to F, used for validating the model. The total number of replications, for both modelling and validating purposes, is $R+F$.

Note that quantities 1, 2, and 5 are empirical and observed, whereas quantity 3 is theoretical and unobservable. Quantity 4 involves a combination of observation and theory in that the model uses the observations Y_{ger}, but then statistical calculations lead to an estimate $\hat{Y}_{ge}$ which, in general, does not equal either the observations Y_{ger} or their unadjusted mean $\bar{Y}_{ge}$. That is, both the data and the model (theory) are required in order to calculate $\hat{Y}_{ge}$.

Incidentally, the R modelling replications are sometimes called the "old" data and the F validation replications the "new" data, reflecting the time sequence of model construction followed by model validation. These terms do not necessarily require or imply that the new data are collected literally in the future, that is, not until after model construction has been completed. In the present context of yield trials, the modelling and validation data are collected concurrently, which is necessary given the impossibility of exactly duplicating the weather and other conditions from year to year. Hence, the modelling and validation data, or "old" and "new" data, are distinguished here in the present application solely by their allocation to different purposes, ordinarily on the basis of a randomization.

Since all T treatments involving the yields $\hat{Y}_{ge}$ are to be analyzed in the same way here, the remainder of this section and its mathematical notation may be simplified by dropping the subscripts g and e and focusing on a single empirical mean, here denoted $\bar{x}$, and its R individual replicates x_r and true mean μ.

The R replicates are understood to constitute a random and finite sample from a larger population of possible replicates or observations. Hence, their variance is estimated by the error mean square (EMS), calculated using the equation for samples, namely:

$$\mathrm{EMS} = \hat{\sigma}_s^2 = \left[\sum_{r=1}^{R} (x_r - \bar{x})^2\right] \Big/ (R-1)\,.$$

Note that the corresponding equation for populations (exhaustive samples), denoted $\hat{\sigma}_p^2$, instead divides by R, but is not used here.

The corresponding variance for a sample mean based on R replicates is estimated by $\hat{\sigma}_{\bar{x}}^2 = \hat{\sigma}_s^2/R$, and the standard error of this mean is the square root of this quantity. Homogeneity of variances between treatments is commonly assumed. Accordingly, a single pooled estimate of the variance exploits the entire data set and applies to all GE treatments.

The critical comment may now be offered that this estimation of the variance $\hat{\sigma}_s^2$ implements a distinctively predictive interest in the accuracy of the yield observations, in that it provides an unbiased estimate of the mean square difference between an individual empirical observation x_r and its true mean μ. Although this point is amply clear in statistics texts (such as Snedecor and

Cochran 1980), nevertheless confusion is common. This confusion probably arises from the subtlety that the estimation of σ^2 necessarily involves empirical quantities, namely, differences of the sort $x_r - \bar{x}$, whereas the meaning or definition of σ^2 involves a theoretical quantity, namely, the (unobserved) true mean μ and differences of the sort $x_r - \mu$. The relationship between the true and empirical means, μ and $\bar{x}$, as regards the EMS, is as follows:

$$\sigma^2 = \left[\sum_{r=1}^{R} (x_r - \mu)^2\right] \Big/ R$$

$$\text{is estimated by EMS} = \left[\sum_{r=1}^{R} (x_r - \bar{x})^2\right] \Big/ (R-1) .$$

Note that division by $R-1$, rather than R, in the equation for estimating EMS corrects for sample bias, and causes the variance estimate to reflect differences from the true population mean μ rather than the sample mean $\bar{x}$.

Also note that this equation for σ^2 on the left side of the above equation line is not a recipe for calculating σ^2, since it contains the unobservable parameter μ, and furthermore σ^2 is itself a parameter. Rather, it is an informal tool for explaining the meaning of σ^2, approaching this parameter's value as the number of replications R goes to infinity.

Likewise, for the empirical mean $\bar{x}$,

$$\sigma_{\bar{x}}^2 \text{ is estimated by EMS}/R = \left[\sum_{r=1}^{R} (x_r - \bar{x})^2\right] \Big/ [R(R-1)] .$$

Again note that although these calculations necessarily involve empirical quantities, nevertheless EMS/R estimates a theoretical quantity or parameter $\sigma_{\bar{x}}^2$, namely, the mean square difference between an empirical sample mean $\bar{x}$ and its theoretical and true, but unobserved, population mean μ.

Error and noise

A statistical model serves to interrelate empirical, observable quantities (here given Roman symbols) and theoretical, unobservable quantities or parameters (given Greek symbols). The simplest model is the "Treatment Means" or "Cell Means" model:

$$Y_{ger} = \mu_{ge} + \varepsilon_{ger}$$

where Y_{ger} is the yield observation for genotype g, environment e, and replicate r, μ_{ge} is the true mean, and ε_{ger} is the error or deviation from the true mean. The treatment means model is a full model since it estimates all T means with $T\,df$ (or equivalently $T-1$ deviations from the grand mean with $T-1$ df plus the grand mean with 1 df), where $T = GE$ presuming that there are no missing data. Similarly, the simplest yield estimates, using this treatment means model, are the averages over replications:

$$\hat{Y}_{ge} = \bar{Y}_{ge} = \left(\sum_{r=1}^{R} Y_{ger}\right) \Big/ R .$$

Here error is defined as discrepancies between individual observations Y_{ger} and their true mean μ_{ge}, that is, as ε_{ger}. This is the conventional definition.

However, ordinarily greater interest focuses on the yield estimates $\hat{Y}_{ge}$ than on the individual replicates Y_{ger}; so it is useful to also have terminology pertaining to these estimates. Accordingly, here noise is defined as discrepancies between yield estimates $\hat{Y}_{ge}$ and their true mean μ_{ge}, denoting noise as v_{ge}:

$$\hat{Y}_{ge} = \mu_{ge} + v_{ge} .$$

Both error and noise involve comparisons with the same true mean μ_{ge}. The salient distinction is that error involves discrepancies between individual replicates and their true mean, whereas noise involves discrepancies between yield estimates and their true mean. The existence of error means that an individual replicate is not perfect, whereas noise means that a yield estimate (such as an average over several replicates) is also not perfect.

The treatment means model offers the average over replicates $\bar{Y}_{ge}$ for a yield estimate $\hat{Y}_{ge}$, but there also exist countless alternative models, and hence countless alternative estimates. Therefore, the concept of noise is inherently relative to both a particular data set and a particular model. This makes sense because the noise v_{ge}, or the discrepancy between $\hat{Y}_{ge}$ and μ_{ge}, is caused both by sampling variations and by model inadequacies.

An analogous remark applies to error because an experimental design, such as a randomized incomplete block design, leads to adjusted empirical means $\bar{Y}'_{ge}$, which differ from the usual raw or unadjusted mean $\bar{Y}_{ge}$, and hence give an adjusted error ε'_{ger} instead of the original error ε_{ger}. Therefore, the concept of error is also inherently relative to both a particular data set and a particular model.

Note in particular that for a given yield trial data set, the errors (and noise) could be relatively large with one statistical model, but relatively small with another model. This raises the issue, to be addressed momentarily, of empirical and theoretical considerations bearing upon model choice.

The above definitions of error and noise are purely statistical. However, they apply to a designed experiment serving agricultural purposes, and therefore it is fitting next to also seek an agricultural interpretation of these concepts. Hopefully, the statistical and agricultural dimensions of these concepts will correspond to some usable degree.

The fundamental practical problem here is that yield data are noisy: "data = pattern + noise" (Freeman 1973).

We desire to know the population or true means μ_{ge}, that is, the true pattern; however, the yield estimates $\hat{Y}_{ge}$ and sample means $\bar{Y}_{ge}$ fall short because of noise.

Postdiction concerns a model's accuracy in fitting its own data, whereas prediction concerns a model's accuracy in fitting new or independent data (Gauch 1988; Gauch and Zobel 1988). When the data are split, the modelling data can be regarded as "$\text{data}_1 = \text{pattern} + \text{noise}_1$" and the validation data as "$\text{data}_2 = \text{pattern} + \text{noise}_2$." Postdiction makes no distinction between pattern and noise, simply rewarding recovery of the data's total sum of squares (SS). But prediction distinguishes pattern from noise, rewarding recovery of pattern (since it is constant and predictively useful across data subsets), but penalizing recovery of noise (since it is idiosyncratic and predictively deleterious across data subsets). "The pattern is stable, agronomically meaningful, and has predictive value, whereas the noise is idiosyncratic, uninterpreted, and of no predictive value" (Gauch and Zobel 1988). Additional discussion of the difference between postdiction and prediction follows momentarily.

Insofar as a yield trial successfully allocates agriculturally significant, interpretable, and repeatable factors in its treatment design, while relegating agriculturally insignificant, uninterpreted, and unrepeatable factors to its experimental design, to that extent statistical patterns will correspond with agricultural patterns and statistical noise will correspond with agricultural noise.

In practice, there is never a completely clean separation between treatment and experimental designs. Imposed treatments may affect unrecognized factors, such as a manure containing some needed microelement or, alternatively, some deleterious bacterium. The control of factors may not be complete or exact, as when variations in residual nitrogen confound the effects of imposed fertilizer applications. Some presumably "random" variation may actually be understood agronomically, such as a well-known, excessively sandy corner in a test field, or recognized spots with serious soil compaction problems. Also complex interactions among controlled or uncontrolled factors or both may elude the agronomist's grasp.

Nevertheless, the well-established history of successful yield trial research demonstrates that some measure of separation between treatment and experimental designs is accomplished routinely. In other words, treatments are significant enough, and controlled well enough, to obtain useful agricultural information from yield trial experiments, despite uncontrolled factors and despite not perfectly equivalent replicates.

Insofar as a controlled yield trial experiment succeeds, there exists, from its design, the prospect that a good statistical model will recover pattern with agricultural significance and utility, while relegating noise to a discarded residual.

Prediction and postdiction

A yield estimate $\hat{Y}_{ge}$ can be used for two purposes: postdiction of an individual observation Y_{ger}, where the subscript r indexes this old observation already used for model construction and for estimation of $\hat{Y}_{ge}$; and prediction of other or future observations Y_{gef}, where the subscript f indexes this new or future observation to be used for model validation but not used for model construction or for estimation of $\hat{Y}_{ge}$ (Gauch and Zobel 1989). Obviously prediction is harder than postdiction. On average, the new observations Y_{gef} will lie farther from the old sample mean $\bar{Y}_{ge}$ than will the old observations Y_{ger}, since this sample mean is based upon the old observations Y_{ger} for replicates 1 to R and, hence, is correlated with these old observations. Accordingly, the prediction mean square difference (estimated as the mean squared difference between Y_{gef} and $\hat{Y}_{ge}$ values) is larger than the postdiction mean square difference (estimated as the mean squared difference between Y_{ger} and $\hat{Y}_{ge}$ values).

The Stein effect

If a recommendation predicts some particular yield or, equivalently, the actually best variety gives this yield, but regrettably the farmer's actual yield is smaller by an amount x, then the farmer suffers a yield reduction of x relative to the expected or the best possible yield. This yield reduction can be associated through a loss function $l(x)$ with some practical loss, such as a farmer's income loss resulting from an inaccurate yield estimate leading to a suboptimal variety choice.

Statisticians are concerned with partitioning accuracy into precision and bias. Consider a random variable Y, its population mean μ, and expected value based upon some specified model $\mathrm{E}(Y)$. Then accuracy $(Y-\mu)$ equals precision $(Y-\mathrm{E}(Y))$ plus bias $(\mathrm{E}(Y)-\mu)$. That a model be unbiased is desirable.

Regardless of the form of the loss function, however, note that it is simply a function of the yield reduction x. For example, a yield reduction of 300 kg/ha has the same economic impact for a farmer, regardless of whether a statistician would attribute this inaccuracy to imprecision or bias or to some mixture of imprecision and bias. Hence, inaccuracy is to be minimized.

Now statistical situations can arise in which a small increase in bias allows for a large decrease in inaccuracy. Thus, a biased estimator may be more accurate and give a smaller (better) loss than does an unbiased estimator and, hence, be superior for agricultural and economical purposes. This statistical phenomenon is called the Stein effect [first reported in Stein (1955), but also see James and Stein (1960), a general account by Berger (1985), and interesting remarks in Howson and Urbach (1989)].

The estimates $\bar{Y}_{ge}$ of the treatment means full model are unbiased (Snedecor and Cochran 1980:44). By contrast, the AMMI reduced model produces yield estimates $\hat{Y}_{ge}$, which differ from those of the full model, and furthermore there is no biological or agricultural reason to expect the AMMI model to fit a yield trial exactly. Accordingly, the AMMI model is guaranteed to have some bias, suffering some measure of model misspecification.

Therefore, were unbiasedness taken as a critical requirement in model choice, then the full model would automatically surpass any reduced model, including AMMI. But because of the Stein effect, a requirement for unbiasedness can contradict the above commitment to minimization of inaccuracy and, hence, loss. If both considerations (minimal bias and minimal inaccuracy) cannot be met simultaneously, then the minimization of inaccuracy (and, hence, minimization of loss) must prevail over unbiasedness, because loss is the criterion that relates best to agricultural economic considerations. By comparison, whether a given inaccuracy results from imprecision or bias or both is rather academic.

Hence, when the Stein effect causes a reduced model to gain more in accuracy than it loses in bias, the reduced model may outperform the full model overall, as judged by a criterion of, e.g., minimizing the squared loss. This situation is impossible for infinite sample sizes, since in this case the treatment means would already be perfect and hence could not be improved upon. However, the Stein effect occurs commonly with small sample sizes, such as the two to ten replications characterizing yield trial experiments.

An estimator that exhibits the Stein effect shrinks its estimates toward the means, or shrinks deviations from the means toward zero. This is exactly what happens when an AMMI model discards the residual SS or variation – the remaining model variance is smaller than the original data variance by exactly this difference of the discarded residual variance.

Agronomists and breeders often feel that the best estimate of a yield is necessarily the data, that is, the average over replications for a given yield trial. Accordingly, a model can be worse than its data, not recovering quite all of its variance, but a model cannot be better. Insofar as a model differs from its data, according to this conventional perspective, it is worse than the data.

But the Stein effect says exactly the opposite – a reduced model can be better than the full model. Since the individual replicates have error, their treatment means have noise. That is, if σ^2 is greater than zero, then $\sigma^2_{\bar{x}} = \sigma^2/R$ is also greater than zero for finite R. A postdictive outlook does not distinguish pattern from noise, and hence automatically considers the full model most (postdictively) accurate. However, a predictive outlook seeks to recover the pattern in the data but to relegate the noise to a discarded residual, and hence may judge a reduced model most (predictively) accurate.

The accuracy of AMMI

In general statistical terms, AMMI estimates can be more predictively accurate than their data (treatment means) because of the Stein effect (as in Gauch 1988 and Gauch and Zobel 1988). Nevertheless, it may be helpful to supplement this general, somewhat abstract, explanation with a more specific explanation concerning the AMMI model and its application to yield trials.

So, how does an AMMI reduced model produce more predictively accurate estimates than does the treatment means full model that constitutes AMMI's data? The explanation has two main components.

First, the treatment means model and the AMMI model represent two different statistical perspectives (Gauch and Zobel 1988). Consider a yield trial with G genotypes in E environments with R replications, and now focus upon one particular treatment, genotype g in environment e, and the estimation of its yield $\hat{Y}_{ge}$. What data are relevant for this estimation?

The treatment means modelling perspective considers relevant only the R yield replicates for genotype g in environment e, and uses their average as the estimate. So of the GER observations, R are relevant and $(GE-1)\,R$ are irrelevant.

By constrast, the AMMI modelling perspective considers the entire yield trial to constitute the relevant data. Of GER observations, all GER are relevant. Every datum in the entire yield trial bears upon and can alter this estimation of each $\hat{Y}_{ge}$.

AMMI estimates can be more predictively accurate than treatment means, simply because AMMI considers and uses more data. This AMMI perspective is not explicitly or formally Bayesian, but it has a decidedly Bayesian flavor. In estimating $\hat{Y}_{ge}$, the R direct observations have the Bayesian flavor of likelihood information, while the $(GE-1)\,R$ indirect observations have the Bayesian flavor of prior information. In any case, more data makes AMMI better.

Because the AMMI model misspecifies the true model to some degree, this use of indirect information introduces some bias. However, in many cases the problem with variance outweighs this problem with bias, particularly given a small number of replications and a sizeable EMS. Consequently the Stein effect occurs. But AMMI's improvement does not come from nowhere or from sheer computation, but rather from using more data in making each yield estimate $\hat{Y}_{ge}$.

Second, although the treatment MS combines a variance component for error (or noise) and a variance component for treatments (or pattern), the AMMI model

selectively recovers pattern while its discarded residual selectively recovers noise (Gauch 1988). In brief, the explanation for this selectivity is that a relatively small number of principal agricultural causal factors affects numerous yields in correlated or coordinated major patterns. These factors write a low-dimensional story into the high-dimensional genotypes by environments multivariate data matrix. AMMI models this high-dimensional data structure by a parsimonious, low-dimensional summary in terms of several vectors (namely, genotype and environment means plus scores for one to a few interaction principal components analysis IPCA axes; Gauch 1988). The data contain $G \times E$ numbers, but the model contains only several (usually 2 or 3) times $G+E$ numbers. The low-dimensional AMMI model recovers the low-dimensional pattern originally produced by the principal agricultural causal factors.

By contrast, the uncontrolled and essentially random yield fluctuations that cause replicate observations to be unequal involve idiosyncratic, high-dimensional structure. The low-dimensional AMMI model simply cannot possibly capture inherently high-dimensional information effectively; therefore, it appears instead in the model's residual, which is high dimensional.

Insofar as a yield experiment's treatment design successfully imposes agriculturally important factors, whereas its experimental design randomizes agriculturally unimportant factors, the agricultural distinction between pattern and noise will coincide with its statistical counterpart. Consequently, the EMS can be used to provide rough guidance on how much of the treatment MS represents agricultural pattern worth modelling.

For example, consider Table 1 in Gauch and Zobel (1990). For this soybean yield trial, the error MS is 107,884, while the treatment MS is 2,286,664 with 384 *df.* But the treatment MS includes the error variance component, so the treatment MS may be regarded as being composed of 107,884 noise and 2,178,780 pattern, or 4.7% noise and 95.3% pattern. Hence an ideal, predictively accurate model will not recover 100% of the treatment SS, but rather 95.3% or 836,651,520, and more specifically that variance associated with pattern (while relegating the noise to a discarded residual, presumably with a SS of about 41,427,456). Incidentally, were the treatment effects negligible, then the treatment MS would be expected to nearly equal the error MS, giving rise to an F ratio of nearly 1, which indicates a nonsignificant treatment effect.

Now this 4.7% noise may seem rather insignificant. However, even just a few percent noise is consequential, because it can alter genotype rankings within an environment considerably (Gauch and Zobel 1989). The error MS is 107,884, so the root error MS is 328, in this case in units of kg/ha. For comparison, the grand mean is 2,606 kg/ha. Most treatments have four replicates, so the standard error of the treatment mean is $328/4^{0.5}$ or 164 kg/ha, giving a coefficient of variation for the treatment means of 164/2,606 or 6.3% (which is fairly good for a yield trial). Therefore, were AMMI to discard a residual containing this noise, the resulting adjusted treatment means would differ from the unadjusted raw treatment means by about 164 kg/ha in terms of their root mean square difference, or by about 6.3% of the grand mean.

Another way to understand the significance of AMMI adjustments is to equate their impact to an equivalent number of free observations (Gauch and Zobel 1988). The same gain in predictive accuracy from the AMMI model could also be achieved by providing the treatment means model with a larger number of replications or observations. The adjusted means from AMMI often move the yield estimates toward the true means, to an extent equivalent to the far costlier alternative of supplying the treatment means model with hundreds or thousands of additional yield observations.

Discussion

After splitting the data for validation purposes, the data may then be recombined in order to produce a final, most accurate model based upon all of the data (Gauch and Zobel 1988). This final model is ordinarily more predictively accurate than its predecessor modelling only part of the data. However, this final model's predictive accuracy cannot be measured empirically because no independent validation observations exist at this stage. Nevertheless, some estimate or extrapolation is desired for this final accuracy, even if it must be only approximate or even cautiously conservative.

Several possibilities for this extrapolation merit further investigation. Statistical theory might help, including comparing validation procedures other than simple data splitting. Empirical studies of yield trials conducted with a few more replications than customary could characterize accuracy trends throughout the usual numbers of replications.

One suggestion has been to assume that the statistical gain factor obtained by data splitting approximates the factor for the complete data (Gauch and Zobel 1988). However, subsequent research has indicated that this rule may be too generous. As AMMI models are based upon more and more replications, the supplied (input) treatment means become better and better (less noisy), and consequently the opportunity diminishes for AMMI to improve the yield estimates.

Until this extrapolation is better understood, the most conservative approach is to measure the predictive accuracy for the penultimate model, using only one replication for validation and the remainder for modelling, and to claim that the final model using the complete data

is somewhat more accurate without specifying exactly how much.

Another possible extrapolation is less conservative, but probably not excessive in most cases. Consider a yield trial with four replications, having a penultimate AMMI model using three replications (and reserving one replication for validation) and a final model using the complete four replications. Assume that the AMMI statistical efficiency factor for the penultimate model with three replications is 2.5 (as in Gauch and Zobel 1989), meaning that AMMI's predictive accuracy is equivalent to $2.5 \times 3 = 7.5$ replications (analyzed by the treatment means model). Then a moderately conservative extrapolation for the final model is to add one replication for this final increment in supplied data, so the final accuracy is estimated to be equivalent to $7.5 + 1.0 = 8.5$ replications.

AMMI and blocking provide fundamentally different approaches for controlling error, addressing orthogonal *df*. Thus both can be used. Two objectives are involved in error control: detecting significant effects and adjusting estimates.

Blocking addresses the error *df*, partitioning it into, e.g., blocks and pure error. The objective is to obtain a smaller EMS, thereby increasing *F* ratios and significance. By contrast, AMMI addresses the treatment *df* and, more specifically, the *GE* interaction *df* in order to partition it into interaction principal components with larger MS (than *GE*), thereby increasing *F* ratios and significance. Since AMMI and blocking address orthogonal *df*, both can be used to increase *F* ratios. Increasing a source MS and decreasing an error MS both have the desired consequence of increasing an *F* ratio.

Furthermore, some experimental designs, such as randomized incomplete block designs, derive block adjustments and thereby produce adjusted estimates for the individual observations (replicates) and treatment means. Blocking and AMMI may be combined by supplying these adjusted treatment means to AMMI rather than the unadjusted raw means, and likewise validating the adjusted observations. This use of two orthogonal error-control strategies, addressing both the treatment *df* and the error *df* (i.e., both the treatment and experimental designs), may in some instances achieve remarkable statistical efficiencies.

Although AMMI and blocking are ordinarily complementary error-control strategies, the situation is different for an unreplicated trial. Bradley et al. (1988) document historical trends over the most recent few decades toward trials with fewer replications, and even no replication. Of course, blocking is impossible without replication, but AMMI is still applicable.

More generally, with rather few replications, AMMI will frequently provide more error control than blocking, because AMMI improvements are greater for noisier data and blocking has relatively little information to exploit. By contrast, with rather numerous replications, blocking will sometimes provide more error control than AMMI, because AMMI improvements are smaller for better data and blocking has relatively abundant information to exploit. Hence, recent trends toward fewer replications and more test environments imply decreasing benefits from blocking but increasing benefits from AMMI.

Advantages associated with the full treatment means model, including unbiasedness and 100% variance accounted for, presume a postdictive context. By contrast, advantages associated with the reduced AMMI model, including predictive accuracy and statistical gains in efficiency, presume a predictive context. But the real purpose of an agricultural experiment is not merely to know what happened on experimental plots in the past, but rather to improve the reliability of variety recommendations or to increase the genetic gains from superior selections (Gauch 1988; Gauch and Zobel 1988, 1989). Hence, the agricultural purposes of yield trial research are thoroughly predictive. This predictive context motivates and justifies the advantages of the reduced AMMI model.

On balance, the postdictive context is not wrong or useless, although it is secondary to the predictive context. *F* tests and other postdictive analyses are useful, although their limitations and approximations are best understood from a basically predictive outlook. At any rate, it must be clearly understood that the choice between a postdictive or predictive outlook, or between a primarily postdictive or primarily predictive outlook, turns upon agricultural rather than statistical considerations. Off-the-shelf statistical methods provide the concepts and calculations for either the postdictive or predictive analysis, so there is no problem here. However, the choice of research questions and purposes is fundamentally an agricultural matter. Mere mathematics and statistics are essentially neutral regarding a postdictive or predictive stance, but agricultural questions and applications clearly indicate a primarily predictive stance.

Finally, it may be reiterated that other papers present the empirical basis for claiming statistical efficiency gains from the AMMI reduced model for a variety of crops and situations. This paper's objectives are limited to supplementing these empirical results with a clearer theoretical basis.

Acknowledgements. I appreciate insightful discussions with R. Zobel and helpful suggestions from the reviewers.

References

Berger JO (1985) Statistical decision theory and Bayesian analysis, 2nd edn. Springer, Berlin Heidelberg New York

Bradley JP, Knittle KH, Troyer AF (1988) Statistical methods in seed corn product selection. J Prod Agric 1:34–38

Freeman GH (1973) Statistical methods for the analysis of genotype-environment interactions. Heredity 31:339–354
Gauch HG (1988) Model selection and validation for yield trials with interaction. Biometrics 44:705–715
Gauch HG, Zobel RW (1988) Predictive and postdictive success of statistical analyses of yield trials. Theor Appl Genet 76:1–10
Gauch HG, Zobel RW (1989) Accuracy and selection success in yield trial analyses. Theor Appl Genet 77:473–481
Gauch HG, Zobel RW (1990) Imputing missing yield trial data. Theor Appl Genet 79:753–761
Howson C, Urbach P (1989) Scientific reasoning: the Bayesian approach. Open Court, LaSalle/IL
James W, Stein C (1960) Estimation with quadratic loss. In: Proc 4th Berkeley Symp Math Stat Prob. University of California Press, Berkeley/CA, pp 361–380
Snedecor GW, Cochran WG (1980) Statistical methods, 7th edn. Iowa State University Press, Ames/IO
Stein C (1955) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Proc 3rd Berkeley Symp Math Stat Prob. University of California Press, Berkeley/CA, pp 197–206