Theor Appl Genet (1990) 79:753–761

THEORETICAL AND APPLIED GENETICS

Imputing missing yield trial data *

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

Department of Agronomy and USDA-ARS, Cornell University, Ithaca, NY 14853, USA

Received August 8, 1989; Accepted November 30, 1989
Communicated by A.R. Hallauer

Summary. The Additive Main effects and Multiplicative Interaction (AMMI) statistical model has been demonstrated effective for understanding genotype-environment interactions in yields, estimating yields more accurately, selecting superior genotypes more reliably, and allowing more flexible and efficient experimental designs. However, AMMI had required data for every genotype and environment combination or treatment; i.e., missing data were inadmissible. The present paper addresses the problem. The Expectation-Maximization (EM) algorithm is implemented for fitting AMMI depite missing data. This missing-data version of AMMI is here termed "EM-AMMI". EM-AMMI is used to quantify the direct and indirect information in a yield trial, providing theoretical insight into the gain in accuracy observed and into the process of imputing missing data. For a given treatment, the direct yield data are the replicates of that treatment, and the indirect data are all the other yield data in the trial. EM-AMMI is used to inpute missing data for a New York soybean yield trial. Important applications arise from both unintentional and intentional missing data. Empirical measurements demonstrate good predictive success, and statistical theory attributes this success to the Stein effect.

Key words: AMMI – Missing data – Prediction – Soybean – Yield trials

Introduction

Previous studies have shown the Additive Main effects and Multiplicative Interaction (AMMI) statistical model to be useful for analyzing yield trial data (Kempton 1984; Gauch 1988; Zobel et al. 1988; Gauch and Zobel 1988, 1989). AMMI combines the usual additive analysis of variance (ANOVA) with principal components analysis (PCA) of the interaction, i.e., of the ANOVA's residual. Hence, AMMI models both the main effects and the interaction. It serves research purposes of understanding genotype-environment (GE) interaction, estimating yields more accurately, selecting superior genotypes more reliably, and allowing more flexible and efficient experimental designs. A major practical limitation to date, however, has been a requirement of no missing cells. Data for every genotype and evironment combination or treatment have been necessary.

This paper has two purposes. (1) It implements the Expectation-Maximization (EM) algorithm for imputing missing cells, thereby fitting the AMMI model despite missing cells. This missing-data version of AMMI is here denoted EM-AMMI. Missing cells may originate intentionally or accidentally, and both cases present important applications for EM-AMMI. (2) It examines theoretical implications regarding the direct and indirect information contained in yield data. For a given treatment, the direct yield data are the replicates of that treatment, and the indirect data are all the other yield data in the trial. The effectiveness of EM-AMMI for imputing missing data depends upon it being a reasonably realistic model, and more specifically to its modelling the GE interaction in addition to the main effects, thereby allowing the Stein effect to occur. Results are presented for a New York soybean yield trial.

Preliminary considerations

Several definitions, distinctions, and concepts merit preliminary clarification regarding direct and indirect infor-

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

mation, the AMMI and treatment means models, Stein effect, effective replications, prediction and postdiction, and data splitting.

Consider yield data, Y_{ger}, for G genotypes, E environments (site-year combinations), and R replications, totalling GER observations. Each genotype and environment combination will be termed a treatment. Now consider the objective of estimating the yield, Y_{ge}, for a particular genotype g grown in a particular environment e. Both direct and indirect information is available for this estimation.

The direct information in a yield trial is the R replications for genotype g grown in environment e. Accordingly, the yield, Y_{ge}, may be estimated directly, by the treatment means or cell means model, as the average over these R replications (Searle 1987). The indirect information in a yield trial is the other GER-R observations, i.e., all the yield data except the R replications for genotype g grown in environment e. Thus, the direct and indirect information together constitute a complete partitioning of the total information containing all GER yield observations.

For the objective of estimating Y_{ge}, these GER-R indirect observations are informative if and only if a theoretical or statistical model exists to interrelate the yield trial's data – however formal or informal, and complete or incomplete, that model may be. Without an encompassing model, the indirect information is irrelevant for estimating Y_{ge}. For interrelating all of the yields, of particular interest here is the AMMI or biplot model, as follows:

$$Y_{ge} = \mu + \alpha_g + \beta_e + \sum_{n=1}^{N} \lambda_n \gamma_{gn} \delta_{en} + \varrho_{ge}$$

where Y_{ge} is the yield of genotype g in environment e,
μ is the grand mean,
α_g are the genotype mean deviations (means minus grand mean),
β_e are the environment mean deviations,
N is the number of PCA axes retained in the model,
λ_n is the singular value for PCA axis n,
γ_{gn} are the genotype eigenvector values for PCA axis n,
δ_{en} are the environment eigenvector values for PCA axis n, and
ϱ_{ge} are the residuals.

If the experiment is replicated, the individual observation Y_{ger} for replicate r may be modelled by adding to the above equation an error term ε_{ger} which equals Y_{ger} minus the Y_{ge} mean.

Ordinarily the number N of interaction principal components axes retained in the model is chosen with empirical considerations of F-tests of significance, predictive accuracy, agricultural interpretability of the associated interaction PCA scores, and so on. Usually N is 0–3, and most frequently 1, producing a reduced model with a residual (which combines the discarded axis $N+1$ and all higher axes).

The salient feature of the AMMI model, relative to the present objective of estimating Y_{ge} with good predictive accuracy, is that each and every observation in the entire yield trial has some influence upon every model parameter and hence upon every model estimate of Y_{ge} values. Consequently, the AMMI estimate of Y_{ge} is influenced not only by the R replicates of genotype g grown in environment e, but also by all of the remaining GER-R other observations. Every observation has some influence upon every estimation.

The treatment means or cell means model (Searle 1987) is also considered here:

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

where Y_{ge} is estimated by the average over the R replications $\left(\sum_{r=1}^{R} Y_{ger}\right)/R$,
μ_{ge} is the true mean for genotype g in environment e, and
ε_{ger} is the errror or difference between replicate r and the true mean μ_{ge}.

Note that this estimate of Y_{ge} exploits only the direct information.

The treatment means model is unbiased (Snedecor and Cochran 1980; Searle 1987), but the AMMI model cannot be expected to fit the data perfectly and hence is biased (Gauch 1990). The accuracy of a yield estimate depends upon both the variance of the estimate (or precision) and the magnitude of bias. Sometimes the gain from reduction of variance resulting from use of a biased estimator more than offsets the inaccuracy due to bias. This is termed the "Stein effect" (Stein 1955; James and Stein 1960; Berger 1985), which has been observed with AMMI (Gauch 1988; Gauch and Zobel 1988, 1989).

The most straightforward strategy for increasing the accuracy of a Y_{ge} yield estimate is to use the treatment means model with an increased number of replications R, since the standard error of a mean decreases with the square root of R. However, economic and other practical factors obviously limit this possibility, and furthermore the square-root dependency implies diminishing returns. Hence, alternative strategies are important, such as blocking and AMMI analysis.

It is helpful to be able to express improvements in accuracy achieved by a variety of strategies in terms of a common currency. Effective replications can serve this role, expressing improvements in terms of the total number of replications that would achieve the same standard error or accuracy. Likewise, free replications are the effective replications in excess of the actual replications.

For example, Snedecor and Cochran (1980, p. 265) compare a simple completely randomized (CR) experi-

mental design with a more sophisticated randomized block (RB) design, observing for a particular experiment that: "If a CR plan had been used, about six replications instead of five would have been needed to obtain the same standard error of a treatment mean as with RB." In other words, the RB experiment has five actual, physical replications, but the subsequent statistical analysis improves the accuracy of adjusted Y_{ge} estimates to a degree that would be equivalent from CR unadjusted means based upon six replications. There are five actual replications, six effective replications and, hence, one free replication.

Note that blocking involves a special allocation of treatments to experimental units together with a special statistical analysis, which is a fundamentally different strategy than is increasing the number of replications. Nevertheless, the resulting increases in accuracy for both strategies can be expressed in the same terms of effective replications. In effect, this yield experiment supplies five replications and its statistical analysis supplies a free sixth replication.

Likewise, AMMI can be used to obtain biased but more accurate Y_{ge} estimates, and this improvement can also be expressed in terms of effective replications (Gauch 1988, 1990; Gauch and Zobel 1988). By using effective replications as a common currency, numbers of "replications" can be attributed both to a physical experiment and to a statistical or theoretical analysis.

Blocking is a familiar, routine method for adjusting means in order to increase accuracy and effective replications, whereas AMMI is a fairly novel approach. However, blocking and AMMI differ fundamentally in concept, and substantially in potential. Blocking is aimed at the error df, whereas AMMI is aimed at the orthogonal treatment df. Blocking partitions the error df into sources for blocks and pure error. If the block sum of squares (SS) is relatively large, then the pure error SS may be decreased, reducing the pure error mean square (MS) and hence making F-tests more significant (presuming the usual case in which the concomitant decrease in the pure error df has not had a larger deleterious effect upon significance). However, AMMI partitions the treatment df into a pattern-rich model and a discarded, noise-rich residual (Gauch 1988; Gauch and Zobel 1988; Gauch 1990). If much of the GE interaction SS is concentrated into relatively few df in the first one or few interaction PCA axes, then these sources may have a large MS, hence making F-tests more significant. Since treatments and error are orthogonal, both strategies may be exploited simultaneously.

Another important distinction concerns prediction and postdiction (Gauch 1988, 1990; Gauch and Zobel 1988). In postdiction, the selfsame data set is used to construct and to evaluate a model. Hence, in ANOVA an F-test is a postdictive test. On the other hand, in prediction one data set is used to construct a model, while different and independent data are used to validate the model. For example, an AMMI model can be fitted to some yield data, and its expected values can then be evaluated by calculating the root mean square prediction difference between the model and validation observations not used previously in modelling. This use of independent validation data precludes bias.

Predictions vary in scope. Within-trial predictions concern the same genotypes and environments as the model, using data splitting or something else to partition the data into modelling data and validation data. Between-trial predictions concern new genotypes or new environments not in the original experiment and model, such as inferences from past yields to future yields, or from experimental fields to farmer's fields. The scope of this paper is limited to within-trial prediction.

The root mean square predictive difference (RMSPD) between an AMMI model and validation observations is simply the square root of the quantity of the sum of squared differences between AMMI estimates and validation observations, divided by the number of validation observations (Gauch and Zobel 1988). RMSPD is in the same units as the yield measurements, and a small value indicates predictive success or accuracy.

A model's predictive accuracy may be estimated as follows. Consider the variance of a model, σ_M^2, the variance of validation observations, σ_V^2, and the variance of differences between the model and validation obserations, σ_{MV}^2. By the variance rule, $\sigma_{MV}^2 = \sigma_M^2 + \sigma_V^2$. Now σ_{MV}^2 can be estimated empirically as the mean square difference between the model's estimates and validation observations (i.e., as the square of RMSPD). Likewise, σ_V^2 is estimated empirically by the error MS. Thus, the model's accuracy may be assessed by $\sigma_M^2 = \sigma_{MV}^2 - \sigma_V^2$. This estimate is unbiased because σ_{MV}^2 and σ_V^2 are both unbiased. In other words, σ_M^2 really does assess accuracy, not merely precision. The square root of this quantity is in the original units of yield, and it estimates the root mean square difference between the model's estimates and the true means.

Furthermore, σ_M^2 can be equated with a number of effective replications, namely, the error MS divided by σ_M^2. When the effective replications exceed the actual replications supplied to a model, then the model exhibits the Stein effect. The model estimates are more predictively accurate than treatment means – the model is better than its data.

With the above definitions and concepts in mind, the theoretical significance of EM-AMMI can be appreciated (while practical implications are pursued later). EM-AMMI allows an exact, complete decomposition of the total information in a yield trial into direct information and indirect information, as follows.

Data splitting is used to partition the data into model data and validation data. For each treatment, some replicates are chosen at random for AMMI modelling, while the remaining replicates are reserved for validation. For concreteness, presume that an experiment has four replications total, with two replicates from each treatment chosen at random and used for modelling, and the remaining two replicates used for validation. Also assume that there are 385 treatments (such as would result from 55 genotypes grown in 7 environments), so there are 770 modelling observations and 770 validation observations. (Actually, only 684 validation observations were available because of occasional missing replicates, but for the moment this discussion will use the ideal number of 770 for the sake of simplicity.)

Estimates are then constructed using three different data sets: (1) the direct data, (2) the indirect data, and (3) the total data. These three estimates are then analyzed and compared in terms of their predictive success, i.e., in terms of their accuracy in predicting the validation observations.

(1) The direct data estimates are trivial, using the treatment means model. For each treatment, merely calculate the cell mean or treatment average from the two modelling observations. This mean is then used to predict the two validation observations for each treatment. This procedure is applied to all 385 treatments in turn, and hence to all of the 770 validation observations. The accuracy of these predictions can be assessed from RMSPD, as already discussed. Obviously, the effective replications can be expected to be close to two.

(2) The indirect data estimates are possible given an implementation of EM-AMMI. For each treatment in turn, EM-AMMI is given only the indirect modelling data, namely, the $770-2=768$ other observations, and the model is used to impute a yield value for the missing cell. The result is a complete matrix of imputed yield values for all 385 treatments, in which not a single computation has used a single direct yield observation. The RMSPD for these imputed yields is then computed, and their accuracy is expressed in terms of equivalent replications.

Note that the direct and indirect predictions are based upon absolutely no data in common. For any given treatment, the direct prediction uses only this treatment's 2 modelling observations, whereas the indirect prediction uses only the other treatments' 768 modelling observations. Yet both predictions are assessed in terms of predictive success with the same validation data, and the results are comparable in the common currency of effective replications.

(3) The total data estimates give regular AMMI the total 770 modelling observations, so it has the total information, both direct and indirect. This model's predictive success (with the 770 validation observations) can also be measured with RMSPD, and the results compared with the direct (cell means) and the indirect (EM-AMMI) models.

The thesis to be developed here, supported by empirical results with soybean yields, is that yield trial data often contain as much or more indirect information as direct information. Therefore, much is to be gained by using a reasonably realistic model to extract the indirect information or the total information.

This proposal to use the entire yield trial or the total *GER* observations for each yield estimate may appear odd at first, seeming different from customary practice using only the direct R observations for each estimate. In fact, however, even traditional analyses make limited use of indirect information, although far from as vigorously as is proposed here.

For example, consider testing a given genotype's mean to determine whether it is significantly different from the yield trial's grand mean. This test entails three numbers: the genotype's mean, the grand mean, and a pooled estimate of the error. Two of these three numbers involve the entire data set. Likewise, multiple comparisons also involve the entire data set, at the least because of the estimate of the error. Furthermore, adjusted means from incomplete block and related designs also use all of the data in calculating each yield estimate. Likewise, various spatial statistics and nearest-neighbor methods also use much or all of the data.

Materials and methods

A New York soybean yield trial was analyzed as an example. The data and details of the field methods are available in reports from the Department of Agronomy, Cornell University. Dr. Madison Wright kindly made available the original data on individual replicates. A subset was chosen to avoid missing data (other than an occasional missing replicate or two within a given treatment), so that experiments with missing data algorithms could withhold validation data used later to assess the predictive accuracy of imputed values. The seven cultivars, with maturity groups in parentheses, were: Evans (0), Wilkin (0), Chippewa 64 (I), Hodgson (I), Corsoy (II), SRF 200 (II), and Wells (II). The 55 environments were certain combinations of 10 New York sites and the 12 years from 1977 to 1988. Yields were expressed in kg/ha at 13% moisture. Most treatments had four replicates, but occasional problems reduced this to two or three. Of the 1,540 yield plots planted, 1,454 (or 94.4%) were harvested. However, for this data set there were no missing cells, i.e., no treatments with zero replicates.

One algorithm for implementing EM-AMMI was considered, but rejected. The additive part of EM-AMMI can be fitted despite missing data as described by Searle (1987), and then the multiplicative part as described by Gabriel and Zamir (1978). However, this approach has two problems. First, the usual AMMI algorithm fits the additive parameters and then the multiplicative parameters, but this approach is valid only for balanced data (Bradu and Gabriel 1978; Gabriel 1978). The present unbalanced case requires a simultaneous solution for all model parameters. Second, this multiplicative part has serious conver-

gence problems, particularly with over 5% missing data (K.R. Gabriel, personal communication). Also, Freeman (1975) presents an algorithm related to the approach chosen here.

The algorithm chosen for implementing EM-AMMI was the Expectation-Maximization (EM) algorithm. "In many important cases," including the AMMI model, "the EM algorithm is remarkably simple, both conceptually and computationally" (Little and Rubin 1987, p. 129). In essence, EM involves "filling in missing values and iterating" in such a manner that the starting values do not affect the solution and hence are arbitrary and inconsequential, apart from some affect upon the number of interactions required for convergence. Little and Rubin (1987, p. 129) summarize the computations: "the EM algorithm formalizes a relatively old ad hoc idea for handling missing data: (1) replace missing values by estimated values, (2) estimate parameters, (3) reestimate the missing values assuming the new parameter estimates are correct, (4) reestimate parameters, and so forth, iterating until convergence." A suitable implementation of the EM algorithm for EM-AMMI works as follows.

First, compute cell means for every cell with data. Then initialize EM-AMMI's additive parameters by computing the unweighted genotype means, environment means, and grand mean. Then initialize the interaction residuals as usual for cells with data (namely, the interaction equals the cell mean minus the genotype mean minus the environment mean plus the grand mean), but impute an interaction residual of zero for missing cells. Now the interaction matrix has no unspecified cells, so perfectly ordinary PCA calculations (such as the power method, Acton 1970) solve for EM-AMMI's multiplicative parameters, continuing for as many interaction PCA axes as desired. Note that missing cells are initialized here by the unweighted additive model (since their interaction residuals are imputed by zero), but a still simpler initialization with the grand mean would lead to identical results, although requiring a somewhat larger number of iteractions to reach convergence.

Next, reestimate and revise each missing cell with the current EM-AMMI model. Then fit EM-AMMI to these revised data, treating imputed values the same as actual data. Iterate this process until convergence, i.e., until the imputed values for missing cells show acceptably small changes.

Upon convergence, the EM-AMMI model "fits" the imputed cells perfectly, with a residual of zero (within numerical precision), whereas actual data have finite residuals as usual. Hence, the EM algorithm fits a model to the actual data, while ignoring missing cells in the sense that they receive imputed values that fit the model perfectly.

For regular AMMI with balanced data, successive interaction PCA axes are orthogonal. Hence each PCA axis stays the same regardless of how many other axes are or are not considered in the model. The situation, however, is otherwise for EM-AMMI with missing cells. Each PCA axis affects the imputed data and hence the data set itself, thereby altering every model parameter from the grand mean up. Therefore the first PCA axis (as well as the additive parameters) for the EM-AMMI model with one interaction PCA axis is not the same as is the first PCA axis for EM-AMMI with two interaction PCA axes. Therefore, missing cells require that each EM-AMMI model be computed from scratch, without allowing the results from lower-order models to be used.

Clearly, EM-AMMI requires more computation than AMMI. However, the PCA calculations are ordinary because PCA never sees any missing cells, since imputed values are always inserted before calculating. If only one interaction PCA axis is required for a data matrix of a given size EM-AMMI tends to take about ten times as much computer time as AMMI. However, this factor depends upon, and increases with, the percentage of missing cells (Little and Rubin 1987, p. 129). No problems with numerical instability or local minima have been noted, and they seem unlikely or insignificant for ordinary, nonpathological data sets. Nevertheless, further theoretical and empirical study of stability would be desirable. At any rate, when empirical results demonstrate that EM-AMMI achieves good predictive success for a given data set, then little concern is merited for that data set.

These AMMI and EM-AMMI analyses were performed by program MATMODEL Version 2.0 (Gauch 1989). It allows each observation to be marked for modelling or validation, and computes statistics on predictive success.

Results

Table 1 gives the analysis of variance for the AMMI model, approximating these data as balanced data by computing the treatment SS as if all of the 385 treatments had the full 4 replicates. The unweighted grand mean is 2605.69 kg/ha.

Note that genotypes, environments, and interaction have 5.1%, 76.6%, and 18.3% of the treatment SS, respectively. The interaction is important, having a SS which is 4.17 times as large as the genotype SS. The first interaction principal component axis (IPCA 1) alone, which captures 69.6% of the interaction SS in only 18.2% of the interaction df, has a SS that is 2.52 times as large as the genotype SS. Clearly any realistic or accurate analysis of these soybean data must consider this large interaction. Incidentally, the interaction pattern in the IPCA 1 scores is clearly related to maturity groups for the genotypes and to growing season length (or warmth or longitude) for the environments, so it has a straightforward agricultural interpretation.

Note in Table 1 that the SS for axis 2 is only one-seventh that for axis 1. Furthermore, all subsequent analyses of predictive success with these data showed IPCA 1 to have predictive value, but not IPCA 2 and the higher axes. Because axis 1 gave the best predictive success, this AMMI model was used for this particular yield trial. This partitions the treatment SS and df into a model with a SS of 829202037 in 119 df and a residual with a SS of 48876832 in 265 df. Hence, this model contains 94.4% of the treatment SS.

Table 1. Analysis of variance for a soybean yield trial

Source	df	SS	MS
Total	1453	993406457	683693
Treatment	384	878078869	2286664
Genotype	6	44439308	7406551
Environment	54	672971054	12462427
$G \times E$	324	160668507	495890
IPCA 1	59	111791675	1894774
IPCA 2	57	15752327	276357
Residual	208	33124505	159252
Error	1069	115327588	107884

Although the residual with only 5.6% is discarded, nevertheless the residual is important because it is precisely this discarding that causes AMMI estimates to differ from mere cell means, and hence to have potential for greater predictive success (Gauch 1988, 1990; Gauch and Zobel 1988). Indeed, this residual SS of 48876832 divided by the number of treatments (385) and by the number of replications (4) implies a root mean square difference between the AMMI model and the cell means model of 178.15 kg/ha, which is 6.8% of the grand mean. This is not negligible. Indeed, these adjustments change the genotype rankings within each environment considerably (Gauch and Zobel 1989). Incidentally, although this AMMI 1 model with a 5.6% residual is selected here on the basis of its predictive success, this amount of residual corresponds well with the 4.7% noise indicated by the error MS (Gauch 1990).

The direct, indirect, and total information in the soybean data was assessed as described above. Each of the 385 treatments provided 2 validation observations, or occasionally only 1 or 0 observations because of missing replicates, for a total of 684 validation observations. Also, each treatment provided exactly 2 modelling observations, for a total of 770. For each treatment's model predictions, the direct model used cell means based on that treatment's 2 modelling observations, the indirect model used EM-AMMI's imputed value from the other 768 modelling observations, and the total model used AMMI given the total 770 modelling observations. In order to obtain the most accurate estimate of error mean square (EMS), all 1,454 observations were used to supply 1,069 df, and the resulting root EMS estimate was 328.4564 kg/ha.

The root mean square difference between these 385 imputed values and their withheld corresponding actual cell means (based on the two replications used for modelling) was 335.39 kg/ha. This value assesses the general postdictive accuracy of these imputed values. But how much can a particular, individual imputed value be trusted? The individual differences follow an approximately normal distribution with this standard deviation, as shown in Table 2. The differences are tabulated in intervals of 100 kg/ha so, e. g., the first interval of 0–100 kg/ha had a count of 103 differences in this interval. Note that 203 imputed values or 53% were within 200 kg/ha of the withheld actual data, which is within only 7.7% of the grand mean. The distribution is approximately normal (or more exactly a folded or half normal one, since absolute values of the differences are used).

This preliminary postdictive experiment had no actually missing cells; rather, missing cells were generated by temporarily withholding data, allowing subsequent empirical assessment of the imputed values. Obviously in a real missing-data problem, the exact error in an individual imputed value cannot be calculated as in Table 2 (or else the exact estimate would follow immediately). The more penetrating question is not how far the imputed values are from imperfect cell means based on two replications, but rather how far the imputed values are from the true means, as considered next. This question requires a predictive rather than a postdictive outlook.

Table 2. Absolute difference between imputed and actual data in kg/ha

Difference	Count
0– 100	103
100– 200	100
200– 300	59
300– 400	51
400– 500	22
500– 600	22
600– 700	10
700– 800	5
800– 900	2
900–1000	4
1000–1100	3
1100–1200	4

The RMSPD values for the direct, indirect, and total models were 397.8451, 395.4022, and 361.0331 kg/ha, respectively. Removing the validation observations' variance, the root mean square predictive errors between these three models and the true means were 224.4930, 220.1347, and 149.8709 kg/ha, respectively. These predictive errors correspond to 2.14, 2.27, and 4.80 effective replications, respectively.

The direct model actually has two modelling replictions and, accepting some variability from statistical fluctuations, the predictive success estimate of 2.14 effective replications constitutes good agreement. Remarkably, the indirect model does about as well, and actually slightly better with 2.27 effective replications. Allowing for some inaccuracy due to statistical fluctuations, it suffices to conclude that the indirect model does about as well as the direct model.

Hence, these data contain as much indirect EM-AMMI-derived information as direct information. Of course, without a model, there is no such thing as indirect information. The concept of indirect information necessarily presupposes a model that relates the various yields to one another.

The total model has 4.80 effective replications, for a statistical gain factor of $4.80/2 = 2.40$. The direct effective replications of 2.14 and indirect of 2.27 are roughly additive in this instance, giving 4.80. However, there is no reason to expect these values to always be additive (although obviously the total model is always expected to take the top rank).

Another way to look at this is that given two replicates, the AMMI model supplies 2.80 free replicates; and

given zero replicates, the EM-AMMI model supplies 2.27 free replicates. It is not surprising that the EM-AMMI model based on 768 modelling observations does almost as well in terms of free replications as does the AMMI model based on 770 modelling observations, since there is rather little difference in available modelling data.

Since there are 385 treatments and the total-data AMMI model gives 2.80 free extra effective replications, AMMI analysis improves the predictive accuracy of yield estimates as much as would collecting field data on an extra 1,078 yield plots. Obviously, this statistical analysis offers a very cost-effective strategy for improving accuracy.

Each EM-AMMI calculation of an imputed treatment value is based on 768 indirect observations, and results on average in an estimate equivalent to 2.27 effective replicates. Hence, about $768/2.27 = 338$ indirect observations are as informative as one direct observation. In other words, for estimating the yield of genotype g in environment e, one direct replicate of this particular treatment is as informative as are 338 indirect observations of other treatments.

The indirect information is, of course, much more dilute than is the direct information. However, from the perspective of each treatment, the indirect information is much more abundant, in fact 384 times as abundant. Likewise, from the perspective of a given datum, it serves 1 time as direct information, but 384 times as indirect information. Therefore, although the indirect information is dilute, it is also very abundant, and consequently it is worth extracting. Indeed, for this particular case, the indirect information alone is slightly superior to the direct information alone. Furthermore, combining the direct and indirect information, the AMMI model based upon all of the data has a predictive accuracy equal to treatment means based upon the 770 actual modelling observations plus 1,078 free observations.

This factor of 384 in this instance should be understood as an average. Clearly, the indirect observations are themselves variously informative, with relatively repetitious observations less influential and relatively novel observations more influential. Also, presumably observations with either their genotype or environment in common with an imputed cell are more influential on average than are observations with neither in common.

The above results are aimed at quantifying the direct and indirect information content in yield data. However, these results concern data matrices with 384 filled cells and only 1 missing cell. Next, EM-AMMI results are reported for data matrices with substantial fractions of missing cells.

As the focus now moves away from quantifying direct and indirect information, and toward solving missing-data problems with more substantial amounts of missing data, data splitting is organized differently. In the above experiments, about half of the data across all treatments was used for validation. In the following experiments, a given treatment is either used entirely for modelling, or else entirely for validation.

Data matrices were generated retaining three genotypes from each environment for modelling, selecting at random one genotype from each of the three maturity groups, and using the remaining four genotypes for validation. Hence, there were 165 filled cells and 220 missing cells, or 57.1% missing cells. This entire process was repeated ten times and the results were averaged. EM-AMMI models with 0–3 IPCA axes were computed, and axis 1 gave the best predictive success.

RMSPD for these imputed cells was 445.7864 kg/ha. Removing the validation observations' variance, the root mean square predictive error between the EM-AMMI model and the true means was 301.3998 kg/ha, or 11.6% of the grand mean (whereas, by comparison, an actual mean based on four replicates had a standard error of 164.2285 kg/ha, or 6.3% of the grand mean). This corresponds to 1.19 effective replications. There were actually roughly $165 \times 4 = 660$ indirect observations, so about 555 indirect observations were as informative as 1 direct observation.

Hence, the indirect information is now more dilute than before. Apparently, the informativeness of a given indirect observation is enhanced by the presence of other indirect observations, particularly observations from otherwise unsampled or poorly sampled treatments. This dilution probably partly reflects diminishing returns from repetitious indirect observations within a givent treatment (here 4 replicates instead of only 2 as before), and partly reflects the diminishing predictive accuracy of an EM-AMMI model supplied data for only 165 treatments (instead of 384 of the 385 total, as before).

Nevertheless, it is remarkable that with less than half of the cells filled, EM-AMMI can impute missing cells with an accuracy equivalent to about 1.19 effective replications. This is a lot more than zero replicates. This equates to 262 free observations for these 220 missing cells.

Discussion

Consider an experiment with two replications, measuring a response at ten levels of some factor, for which the relationships between levels and responses happens to be nearly linear, so that linear regression is a good model. Now focus on the estimation of the response at level 5. This response has 2 direct observations, and 18 indirect observations. It is quite possible that a regression line based upon the 18 indirect observations will give an equal or even greater predictive accuracy for this level's response than would an estimate based upon averaging the

2 direct observations. Of course, the linear model based upon the total data, all 20 observations, would do best of all. Now the AMMI or EM-AMMI model is a complex multivariate model and is therefore harder to envision than simple regression, but the general principles concerning direct and indirect information are the same.

Assume for the moment that a particular yield trial's data fits perfectly the additive ANOVA model, leaving no residual, that is, no *GE* interaction. This scenario is well understood to imply that the rankings of the genotypes will be constant over environments, greatly simplifying research for breeding or for variety recommendations.

However, this scenario also has additional but rarely considered implications for experimental design. Were the additive model truly valid, then it would suffice to measure yields for all genotypes in only one environment (presumably a convenient experimental site), and in all other environments to measure only one genotype (any genotype will do, but for simplicity presume that the same genotype is used throughout).

Therefore only $G+E-1$ measurements are needed in order to estimate all $G \times E$ yields. For example, given 100 genotypes and 50 environments, 149 measurements suffice to fit the additive model and hence to estimate all 5,000 yields. Hence, we could tolerate abundant missing data, in fact 4,851/5,000 or 97% missing data, and yet still estimate all of the 5,000 yields perfectly well, if only the additive model were true.

Furthermore, assuming homogeneity of variance, replication can provide the usual pooled estimate of the error MS, and hence allow ordinary statistical inferences and tests. There is no need for a full replication of the experiment, nor even for replication of more than one treatment. About 20–30 replicates would provide a sufficiently accurate estimate of the error MS for most purposes. Nothing further need be mentioned here regarding replication.

The problem that invalidates the above wonderfully efficient experimental design is, of course, interaction. In all yield trials, *GE* interaction might occur, and in most yield trials interaction actually does occur. It is precisely the existence of *GE* interaction that causes genotype rankings to vary from environment to environment, and causes yield estimates based upon only additive effects to be inaccurate and unreliable. Consequently, missing cells pose a substantial problem, not solved adequately by a merely additive model.

This complication of substantial interaction in yield trials is likely to increase, if anything, in the future. For example, Bradley et al. (1988) review trends in corn breeding in the past (before 1980) and present (1980s), and offer projections into the future (1990 and beyond). They observe trends toward fewer replications and more environments, with researchers deliberately distributing experimental effort in order to maximize sampling over diverse environmental conditions. Contrasting the present with the past, they say that "a smaller share of the researcher's budget is devoted to error reduction at an individual location; a larger share is spent on measuring genotype × environment interaction across locations." This choice reflects an ultimate objective of predictive success, that is, "maximum emphasis on precision across locations and years." Accordingly, agronomists and breeders do forego, and must forego, the above simplistic scenario, which reduces experimental design and data analysis to an additive model – despite its wonderful efficiency and supposed tolerance of missing data. It just will not work.

Hence, the serious difficulty when imputing missing yield data originates from the interaction, not the additive effects. Of course, if even the additive effects were also absent, then only a grand mean would remain, requiring but a single observation for its estimation. This trivial case is practically nonexistent in yield trial research, however, and it would indicate a very dull experiment, to say the least. At any rate, fitting the additive parameters requires only $G+E-1$ observations, which is ordinarily not difficult. The challenge comes from the interaction.

However, EM-AMMI offers estimates of missing cells that do take the interaction into account, as well as the additive effects. The merely additive model requires rank 1 data, that is one observation for each genotype and environment (namely, $G+E-1$ observations). The EM-AMMI model with one interaction PCA axis requires rank 2 data, that is, two observations for each genotype and environment (namely, $2G+2E-4$ observations). For example, given 100 genotypes and 50 environments as before, EM-AMMI requires 296 suitably chosen measurements in order to fit the EM-AMMI model, and hence to estimate all 5,000 yields. This amounts to 94% missing data, and yet the EM-AMMI imputed values *do* take into account both additive effects and interaction.

A simple suitable choice of measurements for the rank 2 model (EM-AMMI with one interaction PCA axis) contains data for all genotypes in two environments, and in all other environments data for two genotypes (the same two throughout). Likewise, the rank 3 model (EM-AMMI with two interaction PCA axes) requires all genotypes in three environments and three genotypes in all other environments, and so forth for higher ranks.

Therefore, the EM-AMMI model can impute missing cells, despite considerable missing data. The real question is how good or how bad these imputed values are in a given particular instance.

This question is not a matter for theoretical statistical speculation, but rather for empirical measurement. The procedure is simple. Consider a yield trial with G genotypes and E environments, but having data for only N treatments and, hence, missing data for $(G \times E)-N$ cells. The trial may be unreplicated or replicated (partially or fully, and balanced or unbalanced). Then, apply EM-

AMMI to these data N times, for each treatment in turn reserving its data for validation, computing RMSPD. The predictive success thus measured for matrices with $N-1$ filled cells will be virtually equal to that for the original matrix with N cells (and, in fact, will be slightly conservative, giving a slightly larger or worse RMSPD). Assuming that the filled and missing cells do not differ systematically or significantly in estimation difficulty, the RMSPD empirical value stands as a good assessment of predictive success for the original data matrix. The root mean square difference between EM-AMMI imputed values and the true means is thus measured. If the experiment is replicated so that the error MS can be estimated, then this RMSPD value can be equated to a number of effective replications.

Surely the accuracy of EM-AMMI imputed values thus measured is not merely a function of the number or percentage of filled cells, but also of which particular cells are filled. In general, given only a small fraction of a complete data set, it is best to spread around the observations to cover the range of variation in genotypes and environments as representatively as possible (rather than oversampling one situation to the exclusion of other situations). The AMMI biplot of genotypes and environments, showing additive effects on one axis and the first interaction PCA scores on the other axis, is ideal for selecting representative genotypes or environments (Kempton 1984; Zobel et al. 1988).

An interesting prospect to be explored in future research is to analyze data that by deliberate treatment design have a large portion of missing cells. More specifically, yield data for hundreds of genotypes would be available for just a few large international breeding centers, plus data for only a dozen or so genotypes at numerous small research centers. EM-AMMI can then impute the missing cells, indicating which untested genotypes are likely to have done well at each of the small centers.

Often, small centers receive hundreds of seed packets, but can only manage to plant perhaps 10 or 20, perhaps with little or no guidance in selecting this subset. However, they could plant just several representative genotypes as a basis for EM-AMMI calculations, which would then provide imputed yields for hundreds of additional, untested genotypes. Remaining resources could then be focused on the most promising genotypes for the local environment, as indicated by rankings of EM-AMMI yield estimates.

Finally, EM-AMMI can promote remarkable experimental efficiency. The AMMI or EM-AMMI gain factor varies from one data set to another, with two to five being typical. However, EM-AMMI applied to large data sets with mostly missing cells may produce yet greater efficiency. For example, if a yield trial has only 20% filled cells, then EM-AMMI will estimate five times as many cells as have actual data. Even if the statistical gain factor is only two or three, this additional factor of five combines to imply that each observation is about 10–15 times as informative with the EM-AMMI model as it would be with merely the cell means model. Experimentation is largely a matter of efficiency – of maximizing informativeness relative to effort or cost. AMMI or EM-AMMI can help agricultural researchers get more information out of expensive yield trial data, enabling breeding to progress more quickly, and making variety recommendations more reliable.

Acknowledgements. We appreciate helpful suggestions from K. Basford, D. Bradu, R. Gabriel, J. Hwang, D. Lansky, and the reviewers.

References

Acton FS (1970) Numerical methods that work. Harper & Row, New York

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

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

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

Freeman GH (1975) Analysis of interactions in incomplete two-way tables. Appl Stat 24:46–55

Gabriel KR (1978) Least squares approximation of matrices by additive and multiplicative models. J R Stat Soc Ser B 40:186–196

Gabriel KR, Zamir S (1978) Lower rank approximation of matrices by least squares with any choice of weights. In: Corsten LCA, Hermans J (eds) COMPSTAT 1978, Proceedings in Computational Statistics. Physica, Wien, pp 304–310

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

Gauch HG (1989) MATMODEL Version 2.0. Microcomputer Power, Ithaca/NY

Gauch HG (1990) Full and reduced models for yield trials. Theor Appl Genet (in press)

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

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

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

Little RJA, Rubin DB (1987) Statistical analysis with missing data. Wiley and Sons, New York

Searle, SR (1987) Linear models for unbalanced data. Wiley and Sons, New York

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

Zobel RW, Wright MJ, Gauch HG (1988) Statistical analysis of a yield trial. Agron J 80:388–393