Kansas State University Libraries

New Prairie Press

Conference on Applied Statistics in Agriculture

1989 - 1st Annual Conference Proceedings

USING INTERACTION IN TWO-WAY DATA TABLES

Hugh G. Gauch Jr.

Richard W. Zobel

Follow this and additional works at: <https://newprairiepress.org/agstatconference>

Part of the [Agriculture Commons](#), and the [Applied Statistics Commons](#)

This work is licensed under a [Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License](#).

Recommended Citation

Gauch, Hugh G. Jr. and Zobel, Richard W. (1989). "USING INTERACTION IN TWO-WAY DATA TABLES," *Conference on Applied Statistics in Agriculture*. <https://doi.org/10.4148/2475-7772.1467>

This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.

USING INTERACTION IN TWO-WAY DATA TABLES

Hugh G. Gauch, Jr. and Richard W. Zobel
Department of Agronomy and USDA-ARS,
Cornell University, Ithaca, New York 14853

Abstract

Agronomists and breeders frequently collect yield data for a number of genotypes in a number of environments (site-years), resulting in a two-way data table. The Additive Main effects and Multiplicative Interaction (AMMI) model combines regular analysis of variance (ANOVA) for additive main effects with principal components analysis (PCA) for multiplicative structure within the interaction (that is, within the residual from ANOVA). AMMI is effective for (1) understanding genotype-environment interaction, (2) improving the accuracy of yield estimates, (3) increasing the probability of successfully selecting genotypes with the highest yields, (4) imputing missing data, and (5) increasing the flexibility and efficiency of experimental designs. Ultimately these advantages imply larger selection gains in breeding research and more reliable recommendations in agronomy research. AMMI is ordinarily the statistical method of choice when main effects and interaction are both important.

1 Introduction

Yield trials frequently have significant main effects and significant genotype-environment (GE) interaction. Interaction complicates an agronomist's or breeder's research because then yields are not understandable or predictable on the basis of simple additive effects of genotype means and environment means, and furthermore genotype rankings differ from one environment to another. Traditional statistical analyses are frequently unsatisfactory in handling such complex data, whereas AMMI often provides excellent results (Zobel *et al.* 1988). Effective use of interaction information can provide important insights into the system under study, and can increase the accuracy of yield estimates.

2 Data Requirements

In order for AMMI to be applicable to a given data set, three structural requirements must be met.

(1) The data must be organized in a two-way table, such as genotypes and environments, or more generically rows and columns — not one-way and not three-way or more. The ANOVA part of AMMI is flexible, but the PCA part demands a two-way data structure since eigenanalysis is defined only for a two-way matrix. However, a three-way table (such as genotypes, sites, and years) can often be approached fruitfully as one or more two-way sub-problems (such as combining sites and years to form environments). For modelling purposes and hypothesis generation, the

experiment may be replicated or not, but if F-tests are desired then the error mean square is required and hence replication is needed. An expectation maximization (EM) version of AMMI can fit the model to data sets with missing cells and can impute these missing cells.

(2) The data matrix dimensions (number of rows and number of columns) must be at least 3 by 3 since anything less would not allow the interaction to be decomposed by AMMI. However, since much of the practical value of AMMI arises from discarding a residual with many degrees of freedom but a relatively small sum of squares, larger minimal dimensions of 5 by 5 or preferably 10 by 10 characterize analyses generating truly useful results.

(3) The data must be of one kind, such as yields. It is not allowable for various matrix rows to contain different data and units, such as soil nutrient concentrations, moisture, and temperatures. Such a mixture would cause model parameters for columns to have meaningless units. Also enormous differences in numerical ranges within rows, as typically encountered with such data, would cause rows with very small variances to be practically ignored in the analysis. Also the data must be quantitative — not mere presence or absence data, and not qualitative or categorical data (such as colors or nationalities). A rough scale, such as 0 to 5 for increasing levels of insect damage, is acceptable when increasing values signify increasing levels of a single thing (in contrast to different values coding for different entities, such as nationalities, which do not have a single or simple logical relationship).

In summary, data for analysis by AMMI must have a two-way layout either replicated or not, with dimensions of at least 3 by 3, and contain only one kind of data. A moment's inspection should suffice to determine whether or not a given data set satisfies these structural requirements.

In order for AMMI to be useful, two further conditions are required.

(1) The data structure must conform, to some substantial degree, to the AMMI model equation. Most pointedly, the data must exhibit significant main effects and significant interaction. Whether or not this condition is met usually cannot be determined by mere inspection of the data, but rather AMMI must be applied to the data and the output scanned. However, even when AMMI is not appropriate and a different model is better, an initial analysis by AMMI is ordinarily the easiest means for diagnosing the appropriate sub-model or other model (Bradu and Gabriel 1978).

(2) Research purposes must call for parameters, displays, estimates, or insights of the sort provided by AMMI. However, this condition is almost always met because AMMI serves a remarkably rich variety of purposes (as detailed later).

This discussion features agricultural yield trials, and more specifically multi-location variety trials. However, the AMMI statistical model is applicable and useful for a tremendous diversity of experiments since the two-way layout with one kind of data is a common data structure in science and technology.

3 The AMMI Model

Consider yield data Y_{ge} for G genotypes in E environments, either unreplicated or as averages over R replications. (More generically, Y may be any one kind of data for each row and column treatment combination for a two-way matrix or table.) The AMMI model equation for a given genotype g and environment e is as follows.

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

where

- Y_{ge} is the yield of genotype g in environment e ,
- μ is the grand mean,
- α_g is the genotype mean deviation (genotype mean minus grand mean),
- β_e is the environment mean deviation,
- N is the number of interaction PCA axes retained in the model,
- λ_n is the singular value for IPCA axis n ,
- γ_{gn} is the genotype eigenvector value for IPCA axis n ,
- δ_{en} is the environment eigenvector value for IPCA axis n , and
- ρ_{ge} is the residual.

Note that $\sum \alpha = \sum \beta = 0$. The γ and δ eigenvector values for each PCA axis are scaled to unit vectors such that $\sum \gamma^2 = \sum \delta^2 = 1$. The eigenvalue for a given PCA axis is the sum of squares (SS) accounted for by that axis, and it equals λ^2 or the square of the singular value λ . The sum of the eigenvalues $\sum \lambda^2$ for N axes, plus the residual SS of $\sum \rho^2$ for a reduced model, equals the genotype-environment (GE) interaction SS. A convenient scaling for tabulating the multiplicative part of the AMMI model results from expressing genotype scores as $\lambda^{0.5} \gamma_g$ and environment scores as $\lambda^{0.5} \delta_e$ since multiplication of a genotype score by an environment score then gives the estimated interaction directly (without need of an additional multiplication by λ). Note that AMMI applies PCA to the interaction values, not the original data, and this distinction can be emphasized by calling these "interaction PCA" or "IPCA" axes. The simple method of Gollob (1968) may be used to assign $G+E-2n-1$ degrees of freedom (df) for IPCA axis n (Gauch 1988). The AMMI model with n IPCA axes is designated AMMI n , so AMMI1 has one IPCA axis, and AMMI0 is the special case of no IPCA axes, namely the additive ANOVA sub-model. The units for μ , α , β , λ , and ρ are exactly the same units of yield as for Y , but γ and δ are dimensionless, and therefore the genotype scores $\lambda^{0.5} \gamma_g$ and environment scores $\lambda^{0.5} \delta_e$ are in the units of the square root of yield (and hence the product of a genotype score times an environment score is in the units of yield).

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 $\bar{Y}_{ge}$ mean, as follows.

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

The least-squares fit for balanced data is obtained by first fitting the additive part of the AMMI model (μ , α_g , and β_e) with the ordinary analysis of variance (ANOVA; Snedecor and Cochran 1980), and then analyzing the nonadditive residual or interaction by fitting the multiplicative part (λ_n , γ_{gn} , and δ_{en}) with principal components analysis (PCA; Gabriel 1978). The computations are unproblematic, allowing a linear workload (so twice as much data requires about twice as much computing time, not four or eight times or even more; Gauch 1990b).

4 Related Statistical Models

AMMI may be compared with several of the more familiar statistical models (Zobel *et al.* 1988).

ANOVA is identical with the additive part of the AMMI model, but the interaction term in ANOVA is not partitioned further. Because the interaction has a large number of df, namely $(G-1)(E-1)$, its mean square (MS) is frequently too small to generate a significant F-test, even though its SS may be quite large (even comparable to, or larger than, the genotype SS or environment SS). By partitioning the interaction SS, AMMI frequently finds statistically significant (and agriculturally meaningful) structure within the interaction, even in cases where an F-test fails to declare the interaction as a whole to be significant.

PCA is identical computationally with the multiplicative part of the AMMI model. However, regular PCA is applied to the original *yield data* directly (namely Y_{ge}), whereas in the AMMI model PCA is applied to the *interaction values*, that is, to the residuals from the additive ANOVA (namely $Y_{ge} - \mu - \alpha_g - \beta_e$). Alternatively sometimes PCA is applied to yield deviations from the grand mean (namely $Y_{ge} - \mu$) rather than to the original yield data Y_{ge} .

Finlay and Wilkinson (1963) linear regression is related to AMMI, but the environment scores calculated by PCA in AMMI are instead constrained to equal the environment mean deviations. Because of this constraint, the SS recovered by linear regression can at most equal the SS of IPCA 1 in AMMI, but commonly it is considerably smaller. Hence AMMI always does as well as, but frequently much better than, Finlay-Wilkinson linear regression.

Joint regression is like ANOVA in modelling genotype and environment additive parameters, but it also multiplies these two parameters together with a joint regression constant requiring 1 df (Tukey 1949; Marasinghe and Johnson 1981, 1982a, b). Joint regression is a sub-model of the more general Finlay-Wilkinson linear regression model and the yet more general AMMI model.

In summary, AMMI largely integrates and subsumes other more familiar models. *AMMI is ordinarily the analysis of choice when main effects and interaction are both important.* This is the commonest case for yield trials. In some cases a sub-model of AMMI or else some different model may be best, and yet in such cases an initial analysis by AMMI usually provides the easiest means for diagnosing the appropriate model (Bradru and Gabriel 1978).

5 Purposes and Results

Theoretical considerations and empirical results combine to demonstrate the effectiveness of the AMMI model for five important research purposes.

(1) **Understand GE Interaction.** Frequently agriculturally important features of genotypes (such as maturity group or pedigree) and environments (such as latitude, elevation, rainfall, and soil classification) impact not only main (additive) effects, but also interaction effects. Commonly some of these features mostly impact main effects, whereas others mostly impact interaction effects.

For example, AMMI analysis of a New York soybean yield trial relates overall genetic merit to the genotype additive effect and site quality to the environment additive effect. By contrast, the interaction, well summarized by IPCA 1, relates to maturity group for genotypes (from group 0 through group II cultivars) and correspondingly to length of the growing season for environments (with Chazy at the northern, short-season extreme and Riverhead at the southern extreme).

Unfortunately, three challenges often conspire to obscure interaction. First, if an F-test of the entire GE interaction is insignificant, a researcher may too hastily dismiss the interaction from further consideration, even if the interaction SS exceeds the genotype SS which does receive attention. Second, even if the interaction is studied by means of traditional statistical analyses, such analyses may frequently fail to fit the interaction well. Third, the interaction typically contains hundreds or even thousands of df, so its inherent complexity presents a serious barrier to human comprehension.

AMMI results can be used to construct a biplot with a point for each genotype and for each environment, located in a graph which shows the main effects (α_g and β_e) on the abscissa and the interaction scores ($\lambda^{0.5}\gamma_g$ and $\lambda^{0.5}\delta_e$ for IPCA axis 1) on the ordinate. Such a graph shows, at a glance, both the main effects and the interaction effects for both genotypes and environments. It can readily provide deep insights into a large, complex experiment (Bradu and Gabriel 1978; Kempton 1984; Zobel *et al.* 1988). Such results are useful for generating hypotheses about the genotypes, environments, and GE interaction. If the experiment is replicated, they are also useful for testing hypotheses.

(2) **Accurate Yield Estimates.** AMMI partitions the total variance into a model (μ , α , β , λ , γ , and δ) and a residual (ρ). Statistical theory and empirical demonstrations show that this model selectively recovers the pattern in the data, whereas the discarded residual selectively recovers the noise in the data (Gauch 1988, 1990a). Validation studies typically show that adjusted means from AMMI are as predictively accurate as would be unadjusted means (raw means over replicates) based upon 2 to 5 times as many replicates (Gauch and Zobel 1988). Therefore an inexpensive computation can often improve the accuracy of yield estimates as much as would the expensive collection of data from hundreds or thousands of additional yield plots.

(3) **Selection Success.** A common purpose in yield trial research is to select the best one or few genotypes (or fertilizers, insecticides, or whatever kind of treatments are studied in a given yield trial). Order statistics deals with ranked or ordered means, and can be used to calculate the probability of successfully selecting that genotype with the highest true mean on the basis of imperfect empirical data. Expected upward biases in the highest-yielding genotypes can also be calculated. Such calculations show that selection tasks are frequently far more difficult, and contain larger biases, than an agronomist's or breeder's intuition may suspect. Greater accuracy of yield estimates from AMMI implies substantially greater selection success (Gauch and Zobel 1989). Better selections increase the speed and effectiveness of breeding programs, and increase the reliability of variety recommendations.

(4) **Impute Missing Data.** Missing data may arise accidentally from problems encountered while conducting an experiment. Alternatively, a partial factorial treatment design, including some genotype and environment combinations but not others, may be intended from the outset. Indeed, as the years go by, researchers often intentionally add or drop both genotypes and environments (sites) from their yield trials. Whatever the cause, the result is a genotypes-by-environments two-way matrix with some missing cells.

The expectation-maximization (EM) algorithm (Little and Rubin 1987:127-141) works well for AMMI with missing data (Gauch 1990b; Gauch and Zobel 1990). Computer time is about an order of magnitude greater than without missing data, but this increase is not at all problematic since the workload is still approximately linear.

One experiment concerned a soybean yield trial with 7 genotypes, 55 environments, and 4 replications (occasionally only 2 or 3). Two replicates, chosen at random for each of the $7 \times 55 = 385$ treatments, were used for modelling, and the remainder of the data for a validation study of predictive success. An EM version of AMMI was used 385 times, removing the data from each matrix cell in turn, in order to impute each cell on the basis of the *other* data (namely $770 - 2 = 768$ observations). Comparison of the predictive success of actual treatment means based on 2 replicates with the predictive success of the imputed means based on 768 other observations showed that these two approaches gave comparable accuracy. *Remarkably, the indirect information from 768 other observations equalled the direct information from 2 replicates.* Precisely because the AMMI model includes interaction, it appears to be excellent for imputing missing data.

(5) **Flexible and Efficient Experimental Design.** The scientific value of a yield trial experiment increases with the number of treatments (genotype and environment combinations), whereas the cost of the experiment increases with the number of yield plots. Therefore, for a given cost, there is a tradeoff between the number of treatments and the number of replications. By increasing accuracy without increasing replication, AMMI provides new options when considering this tradeoff. When 2 replicates with AMMI produce yield estimates as accurate as 4 replicates without AMMI, then twice as many treatments can be explored with the same number of yield plots. Twice as many genotypes, or twice as many

environments, explored with nearly the same cost implies much more flexible, efficient, informative experiments. Such advantages are particularly significant given current, well-motivated trends toward fewer replications and more test environments (Bradley *et al.* 1988).

Furthermore, the option to impute missing cells accurately opens even greater possibilities for clever, efficient experiments. For example, an international yield trial could be structured intentionally to measure yields for hundreds of corn varieties at just several international research centers, while measuring yields for only ten or so varieties at numerous smaller cooperating research centers. Calculation and inspection of imputed yields could then identify specific varieties likely to perform well at each of the smaller centers, and a few of these promising varieties could then be added to future yield trials. By this means, the information and progress attained from small trials at most sites would almost equal that attainable by large trials (which are economically impossible at most sites).

The preceding four benefits from AMMI (namely understand GE interaction, accurate yield estimates, selection success, and impute missing data) thus contribute so powerfully to achieving the research purposes of yield trials that they even open up new options at the level of the initial selection of an experimental design. AMMI benefits not only the analysis of an experiment, but also the design of an experiment.

6 Discussion and Conclusions

The value of AMMI modelling is best understood in relation to other modelling options.

An alternative approach for predicting crop yields is to model yield as a function of, say, daily temperatures and rainfall, soil nutrients and physical structure, and management practices. Such models capable of respectable accuracy are characterized by (a) large rosters of input variables and (b) complex, special-purpose computer models containing numerous empirical values specific to a particular crop and location. In addition to collecting yield data, numerous environmental factors must also be measured.

Relative to such a model, AMMI represents a minimal modelling effort in that (a) the only input variable is yield and (b) the model is a standard, off-the-shelf statistical model requiring absolutely no theoretical or empirical basis whatsoever regarding any particular crop or location. No environmental data are needed. Consequently AMMI can be used to re-analyze historical yield data even if no concomitant environmental data were collected and the opportunity to collect such data is forever gone.

Detailed crop models and minimal AMMI models have different costs, purposes, and rewards. These options are not competitive, but rather complementary. However, particularly where economic constraints prohibit collection of extensive environmental data, the alternative of a minimal modelling effort merits serious consideration.

An efficient research strategy must balance resources devoted to experimental design, execution, and analysis. Every research option has associated costs, benefits, and perhaps risks. Every option faces competing options, given overall constraints of time and resources. An obvious and certain way to improve accuracy and selection success is to increase replications, but this option is costly, it faces diminishing returns from a standard error improving reluctantly with the square root of the number of replications, and in any event it may be impossible because of limited seed supplies or other resources. Remarkably, AMMI calculations costing only as much as a yield plot or two frequently can improve the accuracy of yield estimates as much as would planting and harvesting hundreds to thousands of additional yield plots.

References

- Bradley, J.P., Knittle, K.H., and Troyer, A.F. 1988. Statistical methods in seed corn product selection. *Journal of Production Agriculture* 1:34-38.
- Bradu, D., and Gabriel, K.R. 1978. The biplot as a diagnostic tool for models of two-way tables. *Technometrics* 20:47-68.
- Finlay, K.W., and Wilkinson, G.N. 1963. The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research* 14:742-754.
- Gabriel, K.R. 1978. Least squares approximation of matrices by additive and multiplicative models. *Journal of the Royal Statistical Society, Series B* 40:186-196.
- Gauch, H.G. 1988. Model selection and validation for yield trials with interaction. *Biometrics* 44:705-715.
- Gauch, H.G. 1990a. Full and reduced models for yield trials. *Theoretical and Applied Genetics* (in press).
- Gauch, H.G. 1990b. MATMODEL Version 2.0. Microcomputer Power, 111 Clover Lane, Ithaca, New York 14850.
- Gauch, H.G., and Zobel, R.W. 1988. Predictive and postdictive success of statistical analyses of yield trials. *Theoretical and Applied Genetics* 76:1-10.
- Gauch, H.G., and Zobel, R.W. 1989. Accuracy and selection success in yield trial analyses. *Theoretical and Applied Genetics* 77:473-481.
- Gauch, H.G., and Zobel, R.W. 1990. Imputing missing yield trial data. *Theoretical and Applied Genetics* (in press).
- Gollob, H.F. 1968. A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika* 33:73-115.
- Kempton, R.A. 1984. The use of biplots in interpreting variety by environment interactions. *Journal of Agricultural Science, Cambridge* 103:123-135.
- Little, R.J.A., and Rubin, D.B. 1987. *Statistical Analysis with Missing Data*. John Wiley, New York, New York.
- Marasinghe, M.G., and Johnson, D.E. 1981. Testing subhypotheses in the multiplicative interaction model. *Technometrics* 23:385-393.
- Marasinghe, M.G., and Johnson, D.E. 1982a. A test of incomplete additivity in the multiplicative interaction model. *Journal of the American Statistical Association* 77:869-877.
- Marasinghe, M.G., and Johnson, D.E. 1982b. Estimation of σ^2 in the multiplicative interaction model. *Communications in Statistics, Theory and Methods* 11:315-324.

- Snedecor, G.W., and Cochran, W.B. 1980. Statistical Methods, Seventh Edition. Iowa State University Press, Ames, Iowa.
- Tukey, J.W. 1949. One degree of freedom for non-additivity. Biometrics 5:232-242.
- Zobel, R.W., Wright, M.J., and Gauch, H.G. 1988. Statistical analysis of a yield trial. Agronomy Journal 80:388-393.