Paper SD15

Statistical Analysis of Genotype X Environment Interaction in Agricultural Research

K. Bondari
Experimental Statistics, Coastal Plain Station, University of Georgia, Tifton, GA 31793-0748

ABSTRACT

In agricultural experimentation, a large number of genotypes are normally tested over a wide range of environments (locations, years, growing seasons, etc.) and the underlying statistical and genetical theories used to model this system may be rather complicated. The occurrence of the genotype (G) X environment (E) interaction effect further complicates the selection of superior genotypes for a target population of environments. In the absence of G X E interaction, the superior genotype in one environment may be regarded as the superior genotype in all, whereas the presence of the G X E interaction confirms particular genotypes being superior in particular environments. A variety of statistical procedures are available to analyze the results of multi-environment trials. One of the most common methods in a G X E interaction study is to compute the simple averages across replications for a genotype in an environment and then analyzing the means. An alternative method of analyzing the data in a two-way table of means is the Additive Main Effects and Multiplicative Interaction (AMMI) model which combines the conventional analyses of variance for additive main effects with the principal components analysis (PCA) for the non-additive residuals. AMMI is frequently applied in yield trials in agricultural research when both main effects and interaction are important. Other methods (e.g., the regression of genotype means on the environment means) and selected SAS codes will be presented.

Keywords: Genotype-environment interaction

INTRODUCTION

Genotype X Environment interaction (GXE) is a common phenomenon in agricultural research. Differences between genotypic values may increase or decrease from one environment to another which might cause genotypes to even rank differently between environments. The GXE studies are somewhat complicated as they require integrated approaches which combine many fields including agriculture, biology, statistics, computer, and genetics.

A genotype or the genetic makeup of an organism is defined by Falconer and Mackay (1996) as the combination of alleles at a single autosomal locus in a diploid organism. The physical or visible characteristics resulting from the interaction between the genetic makeup and the environment are referred to as phenotype. Phenotypes can be observed, measured, classified, or counted.

Environmental factors (non-genetic factors) such as locations, growing seasons, years, rainfall, the amount of precipitation received in each season, temperature, etc. may have positive or negative impacts on genotypes. Mather and Jinks (1982), Mukai (1988), and Wu and OíMalley (1998) report on two types of environmental variations: (1) micro-environmental which cannot easily be identified or predicted (e.g., year-to-year variation in rainfall, drought conditions, extent of the insect damage) and (2) macro-environmental variances which can be identified or predicted (e.g., soil type, management practices, controlled temperatures). According to these investigators, the GXE interaction variance can only be estimated for the macro-environmental condition indicating that some variables that explain experiment differences are often unknown or canít be measured.

The association between the environment and the phenotypic expression of a genotype constitute the GXE interaction. The GXE interaction determines if a genotype is widely adapted for an entire range of environmental conditions or separate genotypes must be selected for different sub-environments. When GXE interaction occurs, factors present in the environment (temperature, rainfall, etc.), as well as the genetic constitution of an individual (genotype), influence the phenotypic expression of a trait. The impact of an environmental factor on different genotypes may vary implying that the productivity of an animal or plant may also vary from one environment to the next. Breeding plans may focus on the GXE interaction to select the best genotype for a target population of environments.

A basic principle indicated by the GXE interaction is that even if all animals or plants were created equal (same genotypes), they will not necessarily express their genetic potential in the same way when environmental conditions (drought, temperature, disease pressure, stress, etc.) vary. This important concept may require genetic engineering of plants or animals specifically tailored to their environmental conditions.

Genotypes are normally tested over a wide range of diverse environments (e.g., locations, years, growing seasons) and agricultural experiments involving GXE interactions may involve a large number of genotypes. For instance a cross between parental lines differing at n loci involves 3^n different genotypes (2^n different homozygous genotypes). The number of possible genotypes exceeds 80 when n=4. According to Stroup (1990), such an experiment in plant breeding may commonly have more than a dozen varieties and one having as many as a hundred or more varieties would not be unusual.

Several SAS programs related to the analyses of GXE interaction have been developed (e.g., Kang, 1989; Fernandez, 1991; Shafii and Price, 1998). Comstock and Moll (1963) considered environmental effects and their interactions with genotypes random. However, Funnah and Mak (1980) assumed both genotypes and environments (seasons and locations) as random effects in their analysis. Piepho (1994) assumed environments, blocks, and GXE interactions effects as random but genotype effects as fixed. Although the random assumption of both environmental and genotypic effects is often debatable, it is frequently assumed that environmental effects are random.

Interaction illustration

Statistically, GXE interactions occur if the performance of genotypes varies significantly across environments. Assuming 2 genotypes (G1 and G2) tested in 2 environments (E1 and E2), Fig.1 indicates the presence of GXE interaction since G1 is phenotypically superior to G2 in Environment 1 (E1) but inferior to G2 in E2. The phenotypic difference between G1 and G2 remains the same in the two environments representing no interaction between the genotype and the environment in Fig.2.

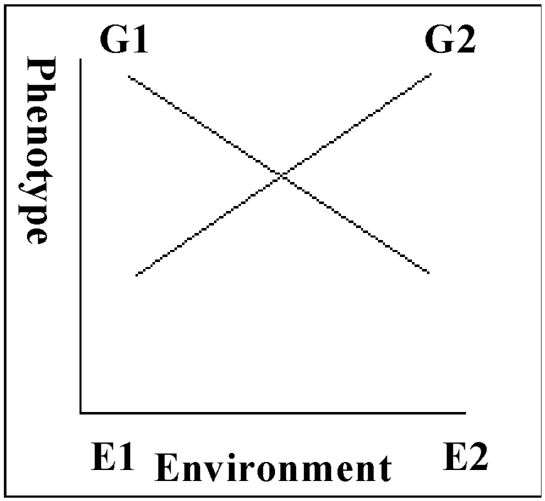

Fig. 1. Shows GXE interaction present

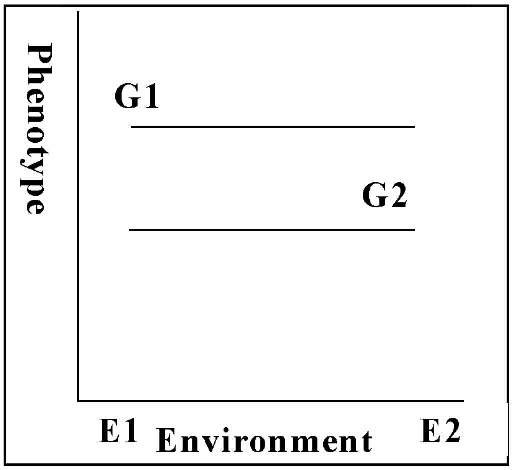

Fig. 2. Shows no GXE interaction

Considering 3 environmental conditions (E1, E2, and E3) and 2 genotypes (G1 and G2), interpretation of the results could be more complicated. Fig. 3 shows one type of GXE interactions for this situation where G1 is superior in performance to G2 in E1 and E3, but is inferior to G2 when exposed to E2. Agricultural researches have demonstrated that a genotype resulting in a good phenotype in one environment might not necessarily result in a good phenotype in another environment.

The Model

Expressing phenotypic value (P) as a function of the genotype (G) and the environment (E), the equation, P= G + E indicates the situation when environmental factors influence each genotype equally (Fig. 2). However, when environment influences some genotypes more than others (Fig. 1), the phenotypic relationship changes to $P=G + E + I_{GE}$ and the expression includes the GXE interaction term I_{GE} . The variance (V) of the effects follows V(P) = V(G) + V(E) + 2 Cov(GE) showing that variance components analyses (the VARCOMP or MIXED procedures of SAS/STAT) could be used to partition the phenotypic variance into its genotypic, environmental, and their interaction components. One way to determine the importance of V(G) or V(E) is to experimentally minimize one of the two effects (minimizing V(G) by using identical genotypes, or minimizing V(E) by using controlled

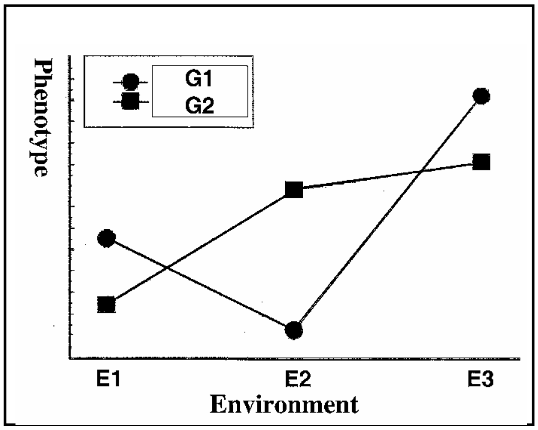

Fig. 3. GXE interaction involving 2 Genotypes (G1 & G2) and 3 Environments (E1-E3)

environment chambers and random allocation of genotypes to environmental conditions). Genotype-environment covariance (Cov) occurs when better genotypes are provided better environments.

For a simple analysis of variance of a randomized complete block design the model:

$$Y_{ijk} = \mu + G_i + E_j + GE_{ij} + B_{jk} + \varepsilon_{ijk} \quad (1)$$

can be applied where μ is the mean, G_i is the effect of the ith genotype, E_j is the effect of the jth environment, GE_{ij} is the interaction of the ith genotype with the jth environment, B_{jk} is the effect of the kth replication in the jth environment, and ε_{ijk} is the random error.

METHODS

A variety of statistical methods have been proposed to analyze GXE interaction data. These methods include Analysis of Variance (e.g., Least Squares, Restricted Maximum Likelihood=REML), Regression (e.g., Joint Regression Analysis, Partial Least-Squares Regression, Factorial Regression), Shifted Multiplicative Model (SHMM), Variance Components, Cluster Analysis, Factor Analysis, and Additive Main Effects and Multiplicative Interaction effects (AMMI model). The GXE data can often be arranged in a two-way layout designating genotypes in rows and environments in columns (Table 1). To apply AMMI model, the conventional analysis of variance for the additive main effects ($\mu + G_i + E_j$) is combined with the Principal Component Analysis (PROC PRINCOMP in SAS/STAT, 2001) for the multiplicative interaction (non-additive residual) effects to analyze the matrix of two-way means. One requirement here is that the matrix of means should be larger than 3X3, requiring more than 3 genotypes to be tested in 3 or more environments (Gauch and Zobel, 1989). Plotting of means and coefficient of variation [CV= (Standard deviation X 100)/Mean)] desirable genotypes having high mean performances and low CV values (Fig. 4). The regression methods of analyses involve the regression of genotype means on the environment means and the regression coefficient is used as a measure of the consistency of the genotypic performance over environments. The Shifted Multiplicative Model (SHMM) proposed by Cornelius *et al.* (1993) clusters genotypes into groups within which crossover interactions do not exist. Within such groups, the genotype with the best mean would be best.

Table 1. A two-way table of means for 4 Genotypes (G1-G4) tested in 3 Environments (E1-E3).

	E1	**E2**	**E3**
G1	50.00	55.50	23.75
G2	91.25	41.50	53.25
G3	70.50	56.25	47.00
G4	67.00	92.75	41.75

Application

Consider the hypothetical experiment in which a researcher wishes to test 4 Genotypes in 3 Environments with 4 replications to illustrate the GXE interaction. To conduct this test 48 experimental units are assumed to have been randomly divided into 3 groups (replicates) and it is assumed that the 4 genotypes are allocated at random to each environment within each replicate. The phenotypic measurement is recorded for each of the 48 experimental units. The statistical model 1 described above may be applied to perform the preliminary data analysis.

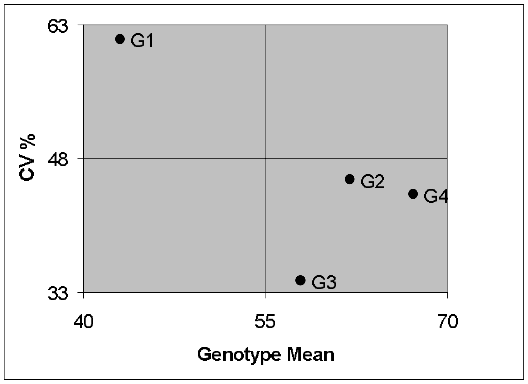

Fig. 4. CV-mean plot of genotypes

Data and SAS Syntax

Analysis and construction of two-way table of means:

```
Data GXE;
Title 'GXE Interaction';
Input Rep Gene Y1 Y2 Y3 @@;
Env=1; Y=Y1; Output;
Env=2; Y=Y2; Output;
Env=3; Y=Y3; Output;
Drop Y1 Y2 Y3;
Datalines;
1 1 64 17 8 1 2 76 20 59 1 3 38 51 66 1 4 90 110 30
...
4 1 47 52 55 4 2 117 64 68 4 3 81 45 58 4 4 58 73 33
Proc Sort Data=GXE; By Gen Env Rep;
Proc MIXED Data=GXE;
Class Env Rep Gen;
Model Y=Env Rep(Env) Gen Gen*Env;
Test H=Env E=Rep(Env);
Means Gen*Env;
Run; Quit;
```

RESULTS

PROC GLM or PROC MIXED of SAS/STAT (SAS, 2001) can be applied to construct the two-way table of means (Table 1). The analysis of variance indicated significant effects of genotype, environment, and GXE interaction (Table 2). The Restricted Maximum Likelihood (REML) estimates of variance components for Environment, Genotype, and GXE Interaction using PROC MIXED are σ^2_E = 127.74, σ^2_G = 6.19, and σ^2_{GXE} = 197.67. The residual component of variance was σ^2_{Error} = 421.89. These variance components may also be computed from PROC VARCOMP or PROC GLM of the SAS/STAT (SAS, 2001). These components indicate a situation controlled by a much stronger environmental contribution than genetic contribution. The PROC VARCOMP of SAS/STAT (SAS, 2001) assumes random effects of genotype and environmental factors.

TABLE 2. Analysis of Variance results

Source	DF	MS	Pr > F
Env (E)	2	3380.27	0.0203
Rep(Env)	9	545.81	0.2854
Gen (G)	3	1286.81	0.0456
G*E	6	1212.58	0.0269
Error	27	421.89	

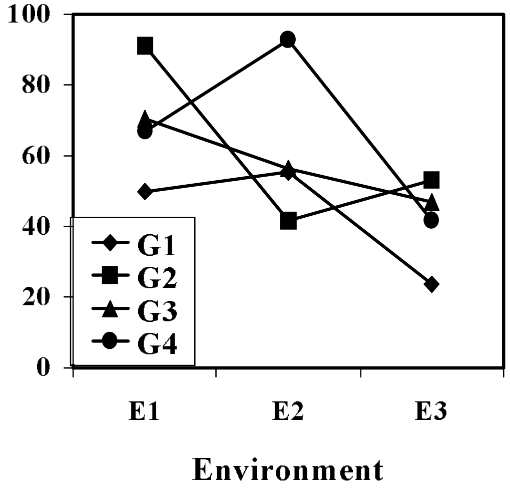

Fig. 5. The plot of 2-way means for 4 Genotypes (G1-G4) and 3 Environments (E1-E3)

Since the GXE interaction is significant (Table 2), the interaction becomes the focus of the study (Bondari, 1999) and follow-up analyses will have to be conducted to determine which levels of the genotype, environment, or the combinations of the two factors are responsible for the differences in the phenotypic response. A plot of the interaction means (Fig. 5) could help study the performance trends of each genotype exposed to the 3 diverse environmental conditions.

DISCUSSION

Presence of the GXE interaction indicates that the phenotypic expression of one genotype might be superior to another genotype in one environment but inferior in a different environment (Falconer and Mackay, 1996). The traditional analysis of variance determines the value of each variance source and the significance of the contribution of each component, but it does not partition the interaction into several components and thus regression or other types of analyses are performed. The issue of fixed or random effects of genotype and environment has not been resolved for the analysis of GXE interaction. The statistical distinction is based on whether an experiment contains all possible levels of genotypes and environments (fixed effects) or only a representative sample of all possible levels of these two factors (random effects). In some cases descriptions of the experimental data might prove useful in determining whether the GXE interaction effects should be viewed as fixed or random (see Bondari, 1999 and Baker, 2002 for further details). The review of literature on GXE interaction indicates that the environmental effects are frequently assumed random as representing a random sample from large populations of environments.

The hypothetical example presented here shows that except for G1the other 3 genotypes have desirable performances (high means, low CV%) in the 3 diverse environmental conditions (Fig. 4). The slopes of the regression of the 4 genotypes on the environmental index in the Joint Regression method of analysis were 1.05 for G1, 0.98 for G2, 0.76 for G3, and 1.20 for G4. Because of the small DF in this hypothetical example, none of the 4 coefficients was significantly different from zero. The slope of the regression line =1 indicates a genotype with an average performance and stability, or a genotype which responds to improved environments. The slope=0 represents a stable genotype which has performed consistently over diverse environments. Positive slope values for genotypes show better performances in superior environments than in inferior environments, and the negative slopes show the contrary performances (Freeman, 1973; Finlay and Wilkinson, 1963; Baker, 1988, 2002). The regression slopes for the 4 genotypes in this example were positive and not significantly different from zero indicating stable genotypes performing better in the superior environment.

Analysis of Genotype X Environment interaction using AMMI Model is more complex but several web sites offer AMMI analysis including:
(1) Quick results from genotype X Environment interaction analysis demo, by George Fernandez (2002), http://www.ag.unr.edu/gf/gei_tai_demo.htm and http://www.ag.unr.edu/gf/sasmac.htm
(2) A University of Idaho program based on Shafii et al. (1992) and Shafii and Price (1998) publications http://www.uidaho.edu/ag/statprog/ammi/index.html
(3) Baker (2002) Genotype-environmental interaction. http://duke.usask.ca/~rbaker/gxe.html
(4) GGEbiplot, http://www.ggebiplot.com/. SAS codes for AMMI analysis are also given by Fernandez (1991) and Macchiavelli and Beaver (1999).

Various methods exist for statistical analyses of GXE interaction data and one should use different approaches to properly interpret the results. Different approaches may lead to similar interpretation of the GXE interaction but disagreements could occur which would require closer inspection of the data. Conditions for the application of AMMI model have been summarized by Gauch and Zobel (1989) which include having a two-way layout with dimensions of at least 3X3 and containing only one kind of data (e.g., yield). Cornelius et al. (1993) proposed the Shifted Multiplicative Method (SHMM) based on clustering genotypes into groups within which crossover interactions (interactions involving rank changes) do not exist. This was in relation to AMMI analysis resulting in too many significant components. Crossover interactions are frequently found to be an important component of GXE interactions. Experimental conditions also play an important role in deciding which methods of analyses should be applied. Lin and Butler (1990) proposed cluster analysis for analyzing two-way GXE interaction data and Piepho (1998) suggested BLUP.

Genotype and environment may interact in a variety of ways and in some cases the GXE interaction can be removed by a change of scale such as logarithmic transformation. However, Mather and Jinks (1982) report that all apparent interactions of genotype and environment are not direct metrical relations which could be removed by a simple change of scale. The choice of making suitable data transformations is often difficult and would require previous knowledge of the variables under study. Plotting the data in various ways may also help in choosing a suitable transformation. For further details, consult Bondari (1999), Mather and Jinks (1982), and Draper and Smith (1981). A researcher should seriously consider those GXE interactions which repeatedly occur in experiments and pay less attention to those that occur only occasionally. Further, environmental factors are complex and similarities in environments can be deceptive and non-repeatable. For instance, two environmental conditions may be similarly poor, one due to poor soil fertility and the other due to rainfall shortage. Also, same genotype may express different phenotypes in different environments.

REFERENCES

Baker, R. J. 1988. Differential response to environmental stress. *Proceedings of the 2nd* Figure 1*International Conference on Quantitative Genetics* (Editors: B. S. Weir, E. J. Eisen, M. M. Goodman, and G. Namkoong): 492-504.

Baker, R. J. 2002. Genotype-environmental interaction. http://duke.usask.ca/~rbaker/gxe.html

Bondari, K. 1999. Interactions in entomology: Multiple comparisons and statistical interactions in entomological experimentation. *Journal of Entomological Science* 34: 57-71.

Comstock, R. E. and R. H. Moll. 1963. Genotype-environment Interactions. *In* Statistical Genetics and Plant Breeding. NAS-NRC No. 982, pp. 164-196.

Cornelius, P. L., D. A. Van Sanford, and M. S. Seyedsadr. 1993. Clustering cultivars into groups without rank-change interactions. *Crop Science* 33: 1193-1200.

Draper , N. R. and H. Smith. 1981. *Applied Regression Analysis*. John Wiley & Sons, New York, 709 pp.

Falconer, D. S. and Mackay, T. F. C. 1996. *Introduction to Quantitative Genetics*. 4th edition, Longman, New York, P.132-133.

Fernandez, G. C. J. 1991. Analysis of genotype X environment interaction by stability estimates. *HortScience* 26: 947-950.

Finlay, K. W. and G. N. Wilkinson. 1963. The analysis of adaptation in a plant breeding program. *Australian Journal of Agricultural Research* 14: 742-754.

Freeman, G. H. 1973. Statistical methods for the analysis of genotype-environment interactions. *Heredity* 31: 339-354.

Funnah, S. M. and C. Mak. 1980. Genotype X environment interactions on grain yield and other characters of soybeans (*Glycine max*). *Experimental Agriculture* 16: 269-273.

Gauch, Jr., H. G. and R. W. Zobel. 1989. Using interaction in two-way tables. *Proceedings of the 1989 Kansas State University Conference on Applied Statistics in Agriculture* (G. A. Milliken and J. R. Schwenke, Editors): 205-213.

Kang, M. S. 1989. A new SAS program for calculating stability-variance parameters. *Journal of Heredity* 80: 415.

Lin, C. S. and G. Butler. 1990. Cluster analyses for analyzing two-way classification data. *Agronomy Journal* 82: 344-348.

Macchiavelli, R. E. and J. S. Beaver. 1999. Analysis of genotype-by-environment interaction with AMMI Model using SAS PROC MIXED. *Proceedings of the 1999 Kansas State University Conference on Applied Statistics in Agriculture*: 171-183.

Mather, K. and J. L. Jinks. 1982. *Biometrical Genetics/ The Study of Continuous Variation.* Chapman and Hall, London, New York, 396 pp.

Mukai, T. 1988. Genotype-environment interaction in relation to the maintenance of genetic variability in populations of *Drosophila melanogaster*. *Proceedings of the 2nd International Conference on Quantitative Genetics* (Editors: B. S. Weir, E. J. Eisen, M. M. Goodman, and G. Namkoong): 21-31.

Piepho, H. P. 1994. Application of a generalized Grubbsí model in the analysis of genotype-environment interaction. *Heredity* 73: 113-116.

Piepho, H. P. 1998. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. *Theoretical and Applied Genetics* 97: 195-201.

SAS Institute Inc. 2001. *SAS/STAT*Æ *User's Guide*, Version 8.2. SAS Institute Inc., Cary, NC.
Shafii, B., K. A. Mahler, W. J. Price, and D. L. Auld. 1992. Genotype X environment interaction effects on winter rapeseed yield and oil content. *Crop Science* 32:922-927.

Shafii, B., and W. J. Price. 1998. Analysis of genotype-by-environment interaction using the additive main effects and multiplicative interaction model and stability estimates. *Journal of Agricultural, Biological, and Environmental Statistics* 3: 335-345.

Stroup, W. W. 1990. Nearest neighbor adjusted best linear unbiased prediction in field experiments. *Proceedings of the 1990 Kansas State University Conference on Applied Statistics in Agriculture* 169-184.

Wu, R. L. and D. M. OíMalley. 1998. Nonlinear genotypic response to macro- and micro-environments. *Theoretical and Applied Genetics* 96: 669-675.

CONTACT INFORMATION

Contact the author at:

K. Bondari, Experimental Statistics, Coastal Plain Experiment Station, University of Georgia, Tifton, GA 31793-0748.
Phone: (229) 386-3385 Fax: (229) 391-2501
E-mail: Bondari@Tifton.UGA.EDU