Monte Carlo Approach To Genotype By Environment Interaction Models

Oyamakin S. Oluwafemi[1,*], Durojaiye M. Olalekan[1]

[1]Biometry Unit, Department of Statistics, Faculty of Science, University of Ibadan, Nigeria

Abstract

Understanding the implication of Genotype-by-Environment (GXE) interaction structure is an important consideration in plant breeding programs. Traditional statistical analyses of yield trials provide little or no insight into the particular pattern or structure of the GXE interaction. In this study, efforts were made to solve these problems under different level of data occurrence. We employed the simulation process of Monte Carlo in generating since use of a real-life data may pose a serious difficulty. In this paper, we simulated for two data Types of Balance and Unbalance designs with different Levels of generations (3X3, 7X7, 10X10, and 3X7, 7X3, 7X10, 10X7 , , respectively). We therefore check the performance of GXE interaction on four different models (AMMI, FW, GGE and Mixed model), and also their stability and adaptability. The findings revealed that, when the assumption was maintained, AMMI outperformed Finlay-Wilkinson model, GGE Biplot model and Mixed model.

Introduction

Food insecurity is a big challenge in Africa [8]. Sub-Saharan Africa is the only region in the world currently facing both widespread chronic food insecurity and threats of famine [2]. This challenge can be addressed through focusing on a crop that requires low input and at the same time can meet major nutritional needs of the people in this region.

Genotype-by-Environment Interaction (GEI)

Multi-location trials play an important role in plant breeding and agronomic research. A number of parametric statistical procedures have been developed over the years to analyze genotype by environment interaction and especially yield stability over environments. A number of different approaches have been used to describe the performance of genotypes over environments. Therefore, the function that described the phenotypic performance of a genotype in relation to an environmental characterization is called the "norm of reaction" (Griffiths et al., 1996).

Figure 1A shows the case where there is no GEI, the genotype and the environment behave additively (this will be developed later) and the reaction norms are parallel. The remaining plots show different situations in which GEI occurs: divergence (Figure 1B), convergence (Figure 1C), and the most critical one, crossover interaction (Figure 1D). Crossover interactions are the most important for breeders as they imply that the choice of the best genotype is determined by the environment.

Crossa [1] pointed out that data collected in multi-location trials are intrinsically complex having three fundamental aspects: structural patterns, nonstructural noise, and relationships among genotypes, environments, and genotypes and environments considered jointly. Plant Breeders generally agree on the importance of high yield stability, but there is less accord on the most appropriate definition of "stability" and the methods to measure and to improve yield stability (Becker and Leon, 1988). Finlay et al. (2007) tested six spring wheat cultivars at five locations across Manitoba and Saskatchewan over two years to examine genotypic and environmental variation in grain, flour, dough and bread-making characteristics. They reported that the relative magnitude of the environmental contribution to wheat variance, depending on the trait (including yield), was considerably larger (14 to 89%) than the variance contribution of either genotype (0 to 33%) or G x E interaction (0 to 17%). Rodrigues, Monteiro and Lourenco [7] also reviewed the performance of the robust extensions of the AMMI model is assessed through a Monte Carlo simulation study where several contamination schemes are considered. Applications to two real plant datasets are also presented to illustrate the benefits of the proposed methodology, which was broadened to both animal and human genetics studies.

The general aim of this study is to determine which of these models best suit GEI using Monte Carlo simulated data. The specific objectives are: (i) to compare the various statistical methods and determine the most suitable parametric procedure that best describe genotype performance under multi-location trials, (ii) to determine the efficiency of each method (AMMI, Finlay-Wilkinson, GGE and Mixed model) in detecting GEI and (iii) also to determine the adaptability and specificities of the methods.
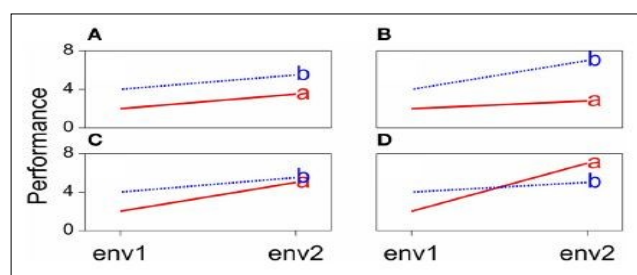

Figure 1. GEI in terms of changing mean performances across environment

Materials and Methods

A combined analysis of variance procedure is the most common method used to identify the existence of GEI from replicated multi-location trials. If the GEI variance is found to be significant, one or more of the various methods for measuring the stability of genotypes can be used to identify the stable genotype (s). A wide range of methods is available for the analysis of GEI and can be broadly classified into four groups: the analysis of components of variance, stability analysis, multivariate methods and qualitative methods.

The methods to be adopted in this study are suitable for the plant breeders in estimating Genotype by Environment Interaction (GEI) parameters. The methods are as follows;

AMMI Model

The AMMI model combines the features of ANOVA and SVD as follows: first, the ANOVA estimates the additive main effects of the two-way data table; then the SVD is applied to the residuals from the additive ANOVA model, estimating $N \leq \min(I-1, J-1)$ interaction principal components (IPCs). The model can be written as [5, 6]

$$y_{ijk} = \mu + \alpha_i + \beta_j + \sum_{n=1}^{N} \lambda_n \gamma_{n,i} \delta_{n,j} + \rho_{i,j} + e_{ijk} \quad \ldots(1)$$

where y_{ijk} is the phenotypic trait (yield or some other quantitative trait of interest) of the ith genotype in the jth environment for replicate k;

μ is the grand mean;

α_i are the genotype deviations from μ;

β_i are the environment deviations from μ;

λ_n is the singular value of the IPC analysis axis n;

$\gamma_{n,i}$ and $\delta_{n,j}$ are the ith and jth genotype and environment IPC scores (i.e. the left and right singular vectors, scaled as unit vectors) for axis n, respectively;

$\rho_{i,j}$ is the residual containing all multiplicative terms not included in the model;

e_{ijk} is the experimental error; and N is the number of principal components retained in the model.

In matrix formulation the AMMI model can be written as:

$$Y = 1_I 1_J^T \mu + \alpha_I 1_J^T + 1_I \beta_J^T + UDV^T + \varepsilon \quad \ldots(2)$$

where Y is the (IXJ) two-way table of genotypic means across environments. The interaction part of the model $Y^* = Y_{-I} 1_J^T \mu - \alpha_I 1_J^T - 1_I \beta_J^T$ is approximated by the product of matrices UDV^T, with U an (IXN) matrix whose columns contain the left singular vectors interactions of n, D a (NXN) diagonal matrix containing the singular values of Y^*, and V a (JXN) matrix whose columns contain the right singular vectors of Y^*

Finlay-Wilkinson Model

A more attractive alternative is to extend the additive model:

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \quad \ldots(3)$$

by incorporating terms that explain as much as possible of the GEI. A popular strategy in plant breeding is that proposed by Finlay and Wilkinson [4], which describes GEI as a regression line on the environmental quality. In the absence of explicit environmental information, the biological quality of an environment can be reflected in the average performance of all genotypes in that environment. The GEI part is then described by genotype-specific regression slopes on the environmental quality, and the model can be written in the following equivalent ways:

$$y_{ij} = \mu + \alpha_i + \beta_j + b_i \beta_j + e_{ij} \quad \ldots(4)$$

$$y_{ij} = \alpha_i' + b_i' \beta_j + e_{ij} \quad \ldots(5)$$

Model (5) follows from model (4) by taking $\mu + \alpha_i = \alpha_i'$ and $\beta_j + b_j \beta_i = (1+b_j) \beta_j = b_t' \beta_j$ Model (5) is easier to interpret because it looks as a set of regression lines; each genotype has a linear reaction norm with intercept α_i' and slope b_i'. The explanatory environmental variable in these reaction norms is simply the environmental main effect β_j. Model (4) shows more clearly how GEI is captured by a regression on the environmental main effect, with the hope that as much as possible of the

GEI signal will be retained by the term $b_t \beta_j$. Note that in model (5) the average value of b'is 1, meaning that b' > 1 for genotypes with a higher than average sensitivity, and b' > 1 for genotypes that are less sensitive than average.

GGE Model

Plant breeders are interested in the total genetic variation and not exclusively in the GEI part. For that reason, it is useful to have a modification of model (1) that considers the joint effects of the genotypic main effect and the GEI as a sum of interpretation procedures hold as for model (1). Because genotypic scores now describe genotypic main effects G and GEI together, this type of model is also known as the "GGE model" and the Biplots are called "GGE Biplots" (Yan et al., 2000). The model reads:

$$y_{ij} = \mu + \beta_j + \sum_{n=1}^{N} \lambda_n \gamma_{n,i} \delta_{n,j} + \rho_{i,j} + e_{ij} \quad .....(6)$$

In GGE, the result of SVD is often presented in a "Biplot illustration". Its approximate overall performance (G + GEI).

Mixed Model

The REML/BLUP method allows the consideration of different structures of variance and covariance for the genotypes by environments effects, which makes the model more realistic. For the GEI evaluation by mixed model, the following statistical model was used:

$$y = Z\alpha + X\beta + W\eta + \varepsilon \quad .....(7)$$

Where, y is the vector of observed data; α is the vector of genotype effects (assumed as random); β is the vector of block effects within each environment (assumed as fixed); β is the vector of GEI effect (assumed as random); and ε is the error vector (random). The uppercase letters represent the matrices of incidence for the referred effects. The distribution of the random effects were:

$$\alpha \mid \sigma_\alpha^2 \sim N(0, I\sigma_\alpha^2), \eta \mid \sigma_\eta^2 \sim N(0, I\sigma_\eta^2) \text{ and }$$
$$\varepsilon \mid \sigma_\varepsilon^2 \sim N(0, I\sigma_\varepsilon^2)$$

Setting up Monte Carlo Experiment

We simulate two-way data tables for balanced and unbalanced design with 3 replications each, where the interaction is explained by two multiplicative terms (i.e. two IPCs; k = 2 components to be retained). Without loss of generality, the two-way data tables are simulated in the following way:

Balance Design

Create a matrix X with *(NxP)* data design;

(3x3) data design, where n = 3 rows (Genotypes) and p = 3 columns (Environments)

(7x7) data design, where n = 7 rows (Genotypes) and p = 7 columns (Environments).

(10x10) data design, where n = 10 rows (Genotypes) and p = 10 columns (Environments).

with observations drawn from a Unif [0, 0.5] distribution.

Do the SVD of X and obtain the matrices U, V and D, containing, respectively, the left and right singular vectors and the singular values of X;

Simulate the grand mean, the genotypic and environmental main effects, considering: $\mu \sim N(15,3)$ $\alpha \sim N(5,1)$ and $\beta \sim N(8,2)$ (Rodrigues et al.(2015)).

Unbalanced Design

Create a matrix X with *(NxP)* data design;

(3x7)data design, where n = 3 rows (Genotypes) and p = 7 columns (Environments)

(7x3)data design, where n = 7 rows (Genotypes) and p = 3 columns (Environments).

(7x10) data design, where n = 7 rows (Genotypes) and p = 10 columns (Environments).

(10x7) data design, where n = 10 rows (Genotypes) and p = 7 columns (Environments).

with observations drawn from a Unif[0, 0.5] distribution.

Do the SVD of X and obtain the matrices U, V and D, containing, respectively, the left and right singular vectors and the singular values of X;

Simulate the grand mean, the genotypic and environmental main effects, considering: $\mu \sim N(15,3)$ $\alpha \sim N(5,1)$ and $\beta \sim N(8,2)$ (Rodrigues et al.(2015)).

Results and Discussion

Model Stability and Adaptability

Balance Design

Comparison of stability of different models using different stability parameters

Table 1 shows the model stability for balance design of which we observed that among all the models, AMMI and FW are the most stable models for 7X7 simulated design showing the highest stability ranked mean of 24.18 and regression coefficient deviation from 1 respectively. Similarly, on the same table, GGE and mixed model claimed to be stable at 3X3simulated design. That is, the complete GGE model contained 98.5% of the Sum of Square, and the residual 1.5%. Also, the Mixed Model showed the lowest ranked stability variance (i.e.σ^2 = 1.919)).

The biplot analysis system showing in Figure 2 are the visual inspection plots that show the most adaptable models.

Therefore, it was observed that the closer the concentric circles to the center point, the more adaptable the models. Similarly, in the second plot, the closer the model to the thick blue arrow line, the more adaptable the model. It can be deduced that from the balance design simulated data, AMMI model is more stable and better adaptable.

Unbalance Design

Comparison of Stability of Different Models Using Different Stability Parameters

Table 2 shows the model stability for Unbalance design of which we observed that among all the models, AMMI and FW are the most stable models for 7X3

simulated design showing the highest stability ranked mean of 24.5 and regression coefficient deviation from 1 respectively. Similarly, on the same table, GGE and mixed model claimed to be stable at 3X7 and 7X10 simulated design. That is, the complete GGE model contained 94.5% of the Sum of Square, and the residual 5.5%. Also, the Mixed Model showed the lowest ranked stability variance (i.e. σ^2 = 28.19).

In the same vein, the biplot analysis system showing in Figure 3 are the visual inspection plots that show the most adaptable models. Therefore, it was observed that the closer the concentric circles to the center point, the more adaptable the models. Similarly, in the second plot, the closer the model to the thick blue arrow line, the more adaptable the model. It can be deduced that from the Unbalance design simulated data, AMMI model is more stable and better adaptable.

Conclusion

In this study, efforts were made to solve these problems under different level of data occurrence. We employed the simulation process of Monte Carlo in generating since use of a real-life data may pose a serious difficulty.

In this research work, we simulated for two data Types of balance and unbalance designs with different Levels of generations (3X3, 7X7, 10X10 and 3X7, 7X3, 7X10, 10X7 respectively).

The findings revealed that, when the assumption was maintained, AMMI outperformed Finlay-Wilkinson model, GGE Biplot model and Mixed model. We therefore check the performance of GXEinteraction on four different models (AMMI, FW, GGE and Mixed model), and also their stability and adaptability.

Table 1. Model stability for Balance simulated data design

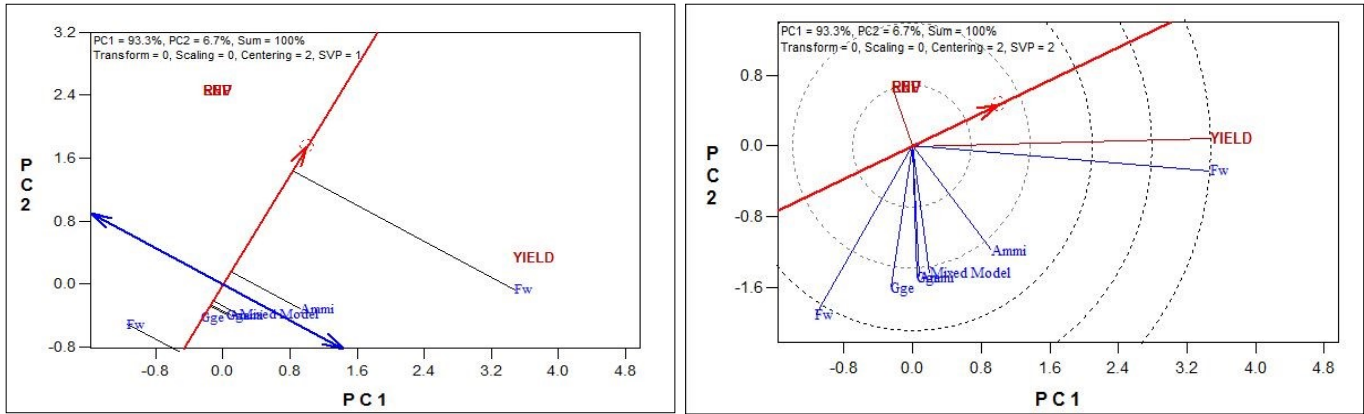| Balance Design		AMMI		FW		GGE		Mixed Model	
Design	Mean	ASV	Rank	b_t	Rank	IPCs	Rank	σ_ε^2	Rank
3X3	18.73	16.80	2	-0.8375	2	98.5%	1	1.919	1
7X7	24.18	6.08	1	-1.6375	1	79.7%	2	28.29	2
10X10	23.70	3.86	3	-0.7419	3	67.5%	3	25.57	3

Figure 2. Model Adaptability for Balance Design

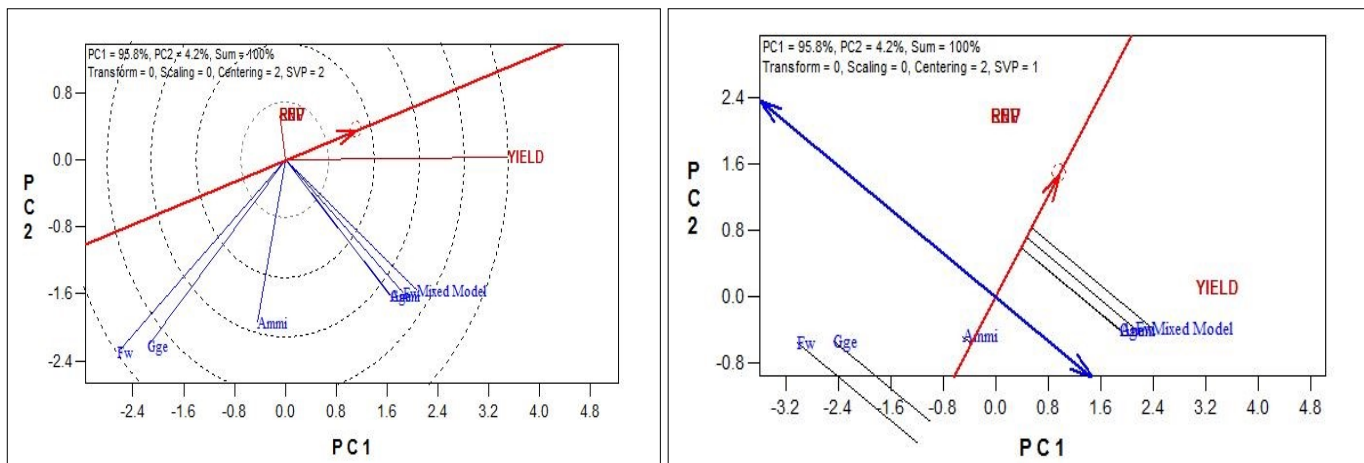Figure 3. Model Adaptability for Unbalance Design

Table 2. Model stability for Unbalance simulated data design

Unbalance Design		AMMI		FW		GGE		Mixed Model	
Design	Mean	ASV	Rank	b_t	Rank	IPCs	Rank	σ_ε^2	Rank
3X7	23.15	23.19	2	-0.7079	4	94.5%	1	30.42	3
7X3	24.5	3.17	1	-4.4698	1	62.3%	4	47.78	4
10X7	22.83	4.34	3	-1.0957	3	81.9%	2	30.18	2
7X10	21.90	2.43	4	-1.4761	2	72.5%	3	28.19	1

Table 3. Model Evaluation of Balance and Unbalance simulated data design

| Balance	RMSE				MSE				Abs. Bias			
Data Design	AMMI	FW	GGE	Mixed Model	AMMI	FW	GGE	Mixed Model	AMMI	FW	GGE	Mixed Model
3X3 Data	1.1312	1.2218	1.7874	1.1374	0.0370	1.9194	1.9190	1.2938	0.6319	4.4565	2.5617	0.7907
7X7 Data	2.7233	4.9308	4.7120	4.3430	18.2120	26.8717	28.2920	22.2025	0.3931	3.0206	2.3156	2.4673
10X10 Data	2.9672	4.8729	4.7044	4.1288	23.4850	25.4414	25.5710	23.1311	0.2982	3.6605	2.1024	1.8547
Unbalance	RMSE				MSE				Abs. Bias			
Data Design	AMMI	FW	GGE	Mixed Model	AMMI	FW	GGE	Mixed Model	AMMI	FW	GGE	Mixed Model
3X7 Data	4.0414	5.8680	4.7957	4.5036	27.1070	38.0586	30.4240	22.9984	0.9037	4.8829	3.1856	2.7243
7X3Data	3.6666	6.4907	6.4199	5.6436	39.1170	54.1660	47.7760	41.2155	0.8199	5.6584	1.9236	2.5613
10X7Data	2.1601	4.7352	4.9967	5.6436	24.2270	24.7819	28.1930	24.9669	0.2600	3.6762	3.2005	1.7961
7X10 Data	3.0695	5.2520	5.1482	5.6436	27.8110	29.5536	30.1800	28.5039	0.3695	4.4930	3.2565	1.9173
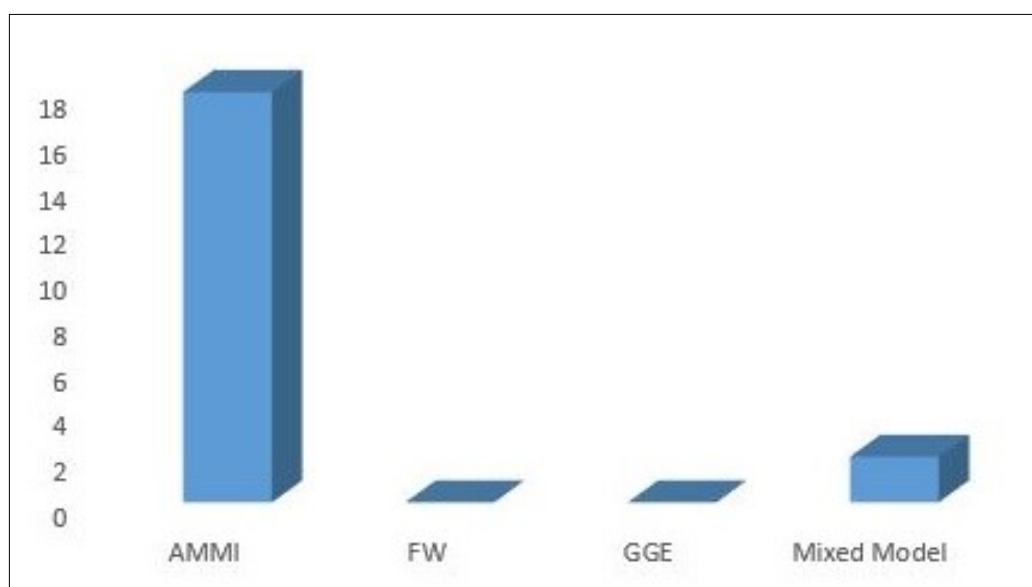
Figure 4. Simulated data rank performance

Finally, the study has clearly shown that the four models considered detects the GXE interaction effect in a different way. We were able to evaluate and described GXE interaction performance by their stability and adaptability using multi-location trials. Also, this study confirmed the suitability of AMMI in detecting GXE when the assumptions are maintained or kept. That is, when outlier is not influential, AMMI can be used. (Table 3, Figure 4).

References

1. Crossa, J. (1990) Statistical Analyses of Multilocation Trials. Advances in Agronomy, 44, 55-85. http://dx.doi.org/10.1016/S0065-2113(08)60818-4

2. Devereux, S., and Maxwell, S., 2001. Food Security in Sub-Saharan Africa. ITDG Publishing.

3. Finlay G J, Bullock PR, Sapirstein HD, Naeem HA, Hussain A, Angadi SV, De-Pauw RM. 2007. Genotypic and environmental variation in grain, our, dough and bread-making characteristics of western Canadian spring wheat. Can. J. Plant Sci. 87: 679-690 Link.

4. Finlay, K.W., Wilkinson, G.N., 1963.The Analysis of Adaptation in a Plant breeding Program. pp. 742-54.

5. Gauch, H. G., 1988. Model selection and validation for yield trials with interaction. Biometrics. pp. 705-715.

6. Gauch, H.G. (2006) Statistical Analysis of Yield Trials by AMMI and GGE. Crop Science, 46,1488-1500. http://dx.doi.org/10.2135/cropsci2005.07-0193

7. Rodrigues, P. C., Andreia M., Vanda M. L. (2015). A robust AMMI model for the analysis of genotype-by-environment data. Bioinformatics, 32(1), 2016, 58-66. doi:10.1093/bioinformatics/btv533

8. UNESCO, 2009. Economic Commission for Africa. Committee on Food Security and Sustainable Development. Sixth Session. Regional Implementation Meeting for CSD-18. The Status of Food Security in Africa.