ROBUST MODIFICATION OF GENOTYPE -BY -ENVIRONMENTS INTERACTION MODEL BY MONTE-CARLO SIMULATION

[1] Zannah, U. [2] Okolo, A. [2]Jibasen, D.[2] Akinrefon, A.A.

[1] Department of Mathematics and Computer Science, Kashim Ibrahim University. [2] Department of Statistics, Modibbo Adama University Yola, Nigeria

ABSTRACT

Genotype Main Effects and Genotype -by -Environments Interaction (GGE) model is one of the frequently used models to capture and analyze Genotype- by-environment Interaction (GEI). The primary concern of most plant breeders and biometricians is to accurately model and analyze GEI, However. This could not be achieved in the GGE model as the model works on singular value decomposition (SVD), a method severely vulnerable to outlying observations. By a Monte Carlo simulation, this study modified the classical GGE model using three (3) robust SVD/PCA methods and obtained three (3) candidates GGE models namely: H-GGE, G-GGE and L-GGE. A simulated GGE multi-environment data was contaminated using pure shift scheme at various levels of generated outliers (2%,5%,10%,15%,20%,25% and 30%) to test and compare the performance of the models. The results revealed the vulnerability of the classical GGE model and further demonstrated robust performance of the modified models at the levels of the outliers used. The models were successfully tested on real multi-environment trials data involving twelve (12) genotypes of wheat grown in nine (9) environments obtained from Lake Chad Research Institute Maiduguri, Nigeria. We recommend to biometricians and plant breeders the use of the modified models for the robust analysis and interpretations of multi-environments data.

Keywords: GGE model, Simulation, multi-environment, vulnerability, Monte-Carlo, Outliers, Contamination

1. Introduction

Multi-Environment Trials (METs) are field trials in which series of genotypes (genetic makeup) are evaluated across multiple environments and over time. The analysis of MET data has evolved over time, beginning with traditional statistical approaches like ANOVA and progressing toward the development of numerous advanced modeling techniques (Piepho et al., 2021). The data from these trials are usually summarized in a two-way table with genotypes in the rows and environment (location /year combination) in the column or vice versa.

In multi-environment trial (MET) data, variations in phenotypic traits such as yield across different environments often reveal that genotype and environment effects do not act independently or additively. This non-additive behavior indicates the presence of Genotype-by-Environment

ISSN NUMBER: 1116-249X

Interaction (GEI), where the performance of genotypes varies depending on the environmental conditions. In other words, a genotype that performs well in one environment may not necessarily do so in another, highlighting the complexity of breeding decisions and the need for models that can capture these interactions effectively. (Eeuwijk et al., 2016). The phenotypic variation due to changing environment is commonly referred to as $G \times E$ interaction and this is important in stability analysis of genotypes in breeding program (Danakumara et al., 2023).

One of the biggest challenges in statistical genetics and plant breeding is identifying top-performing genotypes that consistently show excellent results across varying environmental conditions and over time (Rodrigues, 2018). To handle this challenge, biometricians have developed models (parametric and nonparametric) to model and analyse GEI for stability analysis of genotypes (Mohammadi & Amri, 2008). These models have been applied to data from large-scale plant breeding experiments conducted in diverse environments and over multiple years (Bruno & Balzarini, 2024; Mendes *et al*., 2024).

The parametric models may not perform well if any or all their assumptions are unfulfilled. Specifically, the assumptions of homogeneity of mean square error (MSE), non-mixture of normal distribution and data having no outliers (Huehn,1996). MSEs are rarely homogenous in such multi-environment trials as they are influenced by specific circumstances such as topography, climatic conditions, and soil fertility (Bowman & Watson, 1997; Oliveira et al., 2023).

Parametric models like AMMI and GGE use singular value decomposition (SVD) on residuals from a linear model. Since SVD is a least squares method, it is highly sensitive to outliers, and a single extreme value can distort the leading principal component, leading to misinterpretation and poor decisions. (Rodrigues et al., 2015; Sofi et al., 2022). To address this issue and create a more reliable GGE model for analyzing genotype-environment interactions, this study introduces a robust alternative, aligning with recent calls for more resilient statistical methods in agricultural science (Gauch, 2023). The proposed method replaces the standard linear fit with a robust M-regression and substitutes the conventional SVD with a robust SVD approach, making the model resilient to outliers.

2.1: Genotype Main Effect Plus Genotype-by Environment Interaction (GGE)

ISSN NUMBER: 1116-249X

The GGE model captures both genotype main effects and genotype-by-environment interaction, modifying the traditional AMMI model introduced by Yan et al. (2000). It combines the additive genotype effect from AMMI with the multiplicative GEI effect. A key advantage of GGE over AMMI is its consistent explanation of an intermediate portion of the combined sum of squares for genotype and GEI (Yan & Tinker, 2006; Gauch, 2023). Like AMMI, GGE uses singular value decomposition to analyze multi-environment trial data (Yan & Kang, 2002, 2003; Yang et al., 2009), and remains widely applied in modern breeding programs (Mendes et al., 2024). The GGE biplot (Gabriel, 1971) visually represents relationships among environments, genotypes, and GEI. It highlights three key aspects (Yan, 2001): (i) GEI structure, enabling grouping of genotypes and environments with similar patterns, useful for identifying mega-environments (Bhatta et al., 2023); (ii) environment interrelationships, helping pinpoint optimal and less favorable environments for genotype evaluation; and (iii) genotype interrelationships, supporting comparisons for yield and stability, essential for cultivar selection (Yan, 2014; Oliveira et al., 2023).

2.2: Robust PCA Methods

To improve the GGE model, this study uses a few strong PCA techniques that help deal with messy or complex data. One method, suggested by Hubert and others in 2005, mixes two ideas: looking for useful directions in the data and using a smart way to measure spread that ignores outliers. This works well when there are lots of variables. Another method, from Croux and team in 2007, also looks for patterns but does the searching in a flat grid instead of a big space, which makes it faster. The third method, by Locantore and colleagues in 1999, puts the data on a round surface and then runs regular PCA. This trick helps get good results when the data follows a certain shape and is also very quick to run.

2.3 M-Huber Estimation

Huber (1964) introduced M-estimators within regression analysis, establishing a foundation for robust estimation methods. These estimators have since been generalized to apply across all probability distributions, extending the maximum likelihood approach while yielding consistent and asymptotically normal results. (Heritier et al., 2009). The utility of M-estimation remains highly relevant, particularly in modern high-dimensional data analysis where robustness is critical (Maronna et al., 2019). In the context of agricultural statistics, robust regression techniques like

ISSN NUMBER: 1116-249X

M-estimation provide a reliable foundation for analyzing data from multi-environment trials, which are often prone to outliers and heteroscedasticity (Olivares et al., 2022). Here we use the Huber M estimator to estimate the additive main effects of the robust GGE model, leveraging its well-established properties to shield the initial model fitting stage from anomalous observations.

3.1 GGE Model

The GGE model proposed by Yan and Hunt (2002) is given by:

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

The matrix form in Equation (2) is a direct representation of the element-wise model in Equation (1). The scalar grand mean $\boldsymbol{\mu}$ from (1) is expanded into a constant matrix $\mathbf{1_i 1_j^T \mu}$ in (2). The environment main effects $\mathbf{E_j}$ are collected into a vector $\boldsymbol{\beta_j}$ and spread across rows via $\mathbf{1_i \beta_j^T}$. Most importantly, the entire summation term $\sum \boldsymbol{\lambda_n \gamma_{in} \delta_{jn}}$ from (1), which defines the Genotype-by-Environment Interaction, is exactly equivalent to the product of matrices $\mathbf{UDV^T}$ in (2), where **U** contains the genotype scores (γ), **V** contains the environment scores (δ), and **D** is the diagonal matrix of singular values (λ). Finally, the residual errors $\boldsymbol{\varepsilon_{ij}}$ are simply collected into the full error matrix $\boldsymbol{\varepsilon}$.

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

The matrix (μ), with dimensions ($I \times J$), contains the overall average value repeated in all its entries. The vector (β) shows the main environmental effects. Matrix (U), sized ($I \times N$), holds the left singular vectors that describe the interaction patterns. Matrix (D) is a diagonal ($N \times N$) matrix filled with singular values. Matrix (V), which is ($J \times N$), includes the right singular vectors linked to the interaction. Lastly, (ε) is an ($I \times J$) matrix that accounts for the random experimental errors.

Singular Value Decomposition on the residual matrix produces $\boldsymbol{UDV\text{^}T}$ which is the product of the matrix **U**, of the left singular vectors, that has I rows and r columns $r \leq min\ (I, J)$ with matrix **D** which is a diagonal matrix containing the r singular values and with the transposed matrix **V**, of the right singular vectors, that has r rows and J columns (Yan & Kang, 2003; Yan & Tinker, 2006).

ISSN NUMBER: 1116-249X

3.1.1 Modification of GGE model

The GGE model, equation (1), comprises of a linear fit (additive part) and an interaction (multiplicative part). The additive part is estimated using ANOVA method while the multiplicative part is estimated using a standard SVD applied on the residual matrix. The model is modified in terms of its estimations as follows:

i. The ANOVA method used in estimating the linear fit (additive part) is replaced with a robust regression method (M- estimation).
ii. The standard SVD used in estimating the interaction (multiplicative part) of the model is replaced with robust SVD/PCA methods.

$$y_{ij} = \underbrace{\mu + E_j}_{\text{ANOVA method}} + \underbrace{\sum_{n=1}^{N} \lambda_n \gamma_{i,n} \delta_{j,n}}_{\text{Standard SVD}} + \varepsilon_{i.j} \quad (3)$$

$$\downarrow \qquad\qquad \downarrow$$

Robust Linear fit — Robust SVD/PCA

In these modifications, three robust SVD/PCA methods were used. Thus, a total of three robust GGE candidate models are proposed which are described below.

a) **H-GGE model**: This model was developed by substituting the standard singular value decomposition (SVD) with a robust principal component analysis (PCA) approach proposed by Hubert et al. (2005). The method integrates projection-pursuit (PP) and robust covariance estimation using the minimum covariance determinant (MCD) to derive robust loadings. The implementation of this modification is carried out through a robust linear fit (rlm).

$$y_{ij} = \underbrace{\mu + E_j}_{\text{Robust Linear fit}} + \underbrace{\sum_{n=1}^{N} \lambda_n \gamma_{i,n} \delta_{j,n}}_{\text{Robust SVD/PCA PCAHubert}} + \varepsilon_{i.j} \quad (4)$$

ISSN NUMBER: 1116-249X

b) **G-GGE** model: This model was derived from a robust grid algorithm defined by Croux et al (2007) that uses projection-pursuit (PP) via the grid search algorithm instead of the p-dimensional space to compute PCA estimators. This is implemented using the robust linear fit (*rlm*) and the robust *pcaGrid* in R programming.

$$y_{ij} = \underbrace{\mu + E_j}_{\text{Robust Linear fit (M-Regression)}} + \underbrace{\sum_{n=1}^{N} \lambda_n \gamma_{i,n} \delta_{j,n}}_{\text{Robust SVD/PCA (PCAGrid)}} + \varepsilon_{i.j} \qquad (5)$$

c) L-GGE model: Here, we have replaced the standard SVD used by the GGE model with the robust spherical PCA (Locantore et. al, 1999). The modification is implemented using the robust linear fit (*rlm*) and the robust *pcaLocantore* in R programming.

$$y_{ij} = \underbrace{\mu + E_j}_{\text{Robust Linear fit (M-Regression)}} + \underbrace{\sum_{n=1}^{N} \lambda_n \gamma_{i,n} \delta_{j,n}}_{\text{Robust SVD/PCA PCALocantore)}} + \varepsilon_{i.j} \qquad (6)$$

3.2: Data Simulation of GGE Model

The study simulated a two-way data table following the GGE model with two multiplicative terms using the matrix structure presented as follows:

$$Y = 1_I 1_J^T \mu + 1_I \beta_J^T + 28 * U[,1] D[1,1] V[,1]^T + 15 * U[,2] D[2,2] V[,2]^T + \varepsilon \qquad (7)$$

The matrix structure follows Rodrigues *et al* (2015). The data table comprises of eight Environments and hundred Genotypes. Presented in table 1 the simulated first 25 by 8 of the two-way data table.

ISSN NUMBER: 1116-249X

Table 1: A Simulated two-way data table of GGE2 model (first 25 by 8 observations)

	E1	E2	E3	E4	E5	E6	E7	E8
G1	15.57491	16.527177	11.337308	18.338735	13.856614	19.525371	17.613944	21.89756
G2	24.69064	8.863903	12.791207	17.959041	21.772571	14.087325	16.260611	16.99394
G3	24.32653	11.408724	14.224000	18.449952	21.144411	15.594146	17.178318	18.61575
G4	23.37650	12.792951	13.582161	20.100223	21.124593	17.723432	18.546859	20.33493
G5	24.56568	8.767831	13.733090	16.071433	20.588246	12.520876	14.954802	15.89346
G6	19.88290	14.777139	16.488039	13.124897	14.095750	13.555447	14.035648	17.71983
G7	22.82155	13.851822	17.197071	14.662337	17.226742	13.790070	15.092740	17.87138
G8	23.36735	14.037043	17.049438	15.896880	18.279445	14.759105	16.015222	18.61759
G9	17.80135	15.994676	13.580140	16.849089	14.600141	17.582438	16.666866	20.60802
G10	20.76082	9.340334	8.661909	20.256698	20.385649	17.057532	17.596331	18.84383
G11	23.48402	15.360794	17.943437	16.314851	18.254632	15.608028	16.627740	19.52827
G12	20.76534	16.029587	14.209037	19.839800	18.359629	19.394302	18.938361	22.07113
G13	19.46779	14.296573	15.093020	14.276762	14.632989	14.405862	14.660651	18.15135
G14	24.94667	7.534178	8.866572	23.153754	25.397284	17.778977	19.460186	19.20442
G15	19.53440	12.484446	14.800083	12.438761	14.063232	12.149259	12.988013	16.17542
G16	26.01784	10.612035	16.579125	15.694281	20.844613	12.622167	15.269279	16.52060
G17	14.65484	24.669075	20.420891	12.643279	7.675535	18.301225	15.673869	22.98399
G18	19.85138	12.855097	15.998520	11.339243	13.520786	11.320037	12.363429	15.72422
G19	17.78852	14.450643	11.760939	17.839093	15.593167	17.782512	16.958428	20.36625
G20	20.32762	15.078146	18.797983	10.228215	12.585079	11.188016	12.193722	16.21258
G21	19.24777	14.275311	14.406690	15.109786	14.988366	15.133610	15.198494	18.62592
G22	22.43997	7.339540	10.009907	17.575286	20.442487	13.666827	15.444242	16.16821
G23	19.81086	16.026354	15.529462	16.316061	15.619040	16.703127	16.483189	20.16524
G24	26.26962	4.461997	10.038636	18.855643	24.536190	12.706269	15.928959	15.09417
G25	10.54498	24.420604	16.360542	13.612095	5.568164	19.925940	15.923202	23.72633

3.3 Pure Shift Outliers Generation scheme and Data Contamination

To assess the performance of the modified models in relation to the existing GGE model, the study conducted a pure shift outlier scattered environment contamination scheme in Monte Carlo simulation study. Here, the outliers were generated from the pure shift scheme normal distributions N ($\mu+k\sigma$, σ) in line with Rocke and woodruff (1996). The percentages of outliers generated are 2,5%, 10%, 15%, 20% and 30%. These are the proportion of entries used to contaminate the simulated two-way data table.

In this scheme, we used the scattered environments contamination, as the generated outliers were randomly replaced with entries in the whole simulated two-way data table. These random generation of the outliers and their subsequent replacements in the simulated data table were repeated for m=1000 times. Consequently, we obtained 1000 contaminated two-way data table for every percentage level of outliers. Presented in Tables 2 are the first 50 by 8 of the simulated data with 2% contamination.

ISSN NUMBER: 1116-249X

Table 2.: A simulated data with 2% contamination (first 25 by 8 observations)

	E1	E2	E3	E4	E5	E6	E7	E8
G1	15.57491	16.527177	11.337308	18.338735	13.856614	19.525371	17.613944	21.89756
G2	24.69064	8.863903	12.791207	32.981732	21.772571	14.087325	16.260611	16.99394
G3	24.32653	11.408724	14.224000	18.449952	21.144411	15.594146	17.178318	18.61575
G4	23.37650	12.792951	13.582161	20.100223	21.124593	17.723432	18.546859	20.33493
G5	24.56568	8.767831	13.733090	16.071433	20.588246	12.520876	14.954802	15.89346
G6	19.88290	14.777139	16.488039	13.124897	14.095750	13.555447	14.035648	17.71983
G7	22.82155	13.851822	17.197071	14.662337	17.226742	13.790070	15.092740	17.87138
G8	23.36735	14.037043	17.049438	15.896880	18.279445	14.759105	16.015222	18.61759
G9	17.80135	15.994676	13.580140	16.849089	39.981007	17.582438	16.666866	20.60802
G10	20.76082	9.340334	8.661909	20.256698	41.858344	17.057532	17.596331	18.84383
G11	23.48402	15.360794	17.943437	16.314851	18.254632	15.608028	16.627740	19.52827
G12	20.76534	16.029587	14.209037	19.839800	18.359629	19.394302	18.938361	22.07113
G13	19.46779	14.296573	15.093020	14.276762	14.632989	14.405862	14.660651	18.15135
G14	24.94667	7.534178	8.866572	23.153754	25.397284	17.778977	19.460186	19.20442
G15	19.53440	12.484446	14.800083	12.438761	14.063232	12.149259	12.988013	16.17542
G16	26.01784	10.612035	16.579125	15.694281	20.844613	12.622167	15.269279	16.52060
G17	14.65484	24.669075	20.420891	12.643279	7.675535	18.301225	15.673869	22.98399
G18	19.85138	12.855097	15.998520	11.339243	13.520786	11.320037	12.363429	15.72422
G19	17.78852	14.450643	11.760939	17.839093	15.593167	17.782512	39.423360	20.36625
G20	20.32762	15.078146	18.797983	10.228215	12.585079	11.188016	12.193722	16.21258
G21	19.24777	14.275311	14.406690	15.109786	14.988366	15.133610	15.198494	18.62592
G22	22.43997	7.339540	10.009907	36.576524	20.442487	13.666827	15.444242	16.16821
G23	19.81086	16.026354	15.529462	16.316061	15.619040	16.703127	16.483189	20.16524
G24	26.26962	4.461997	10.038636	18.855643	24.536190	12.706269	15.928959	15.09417
G25	10.54498	24.420604	16.360542	13.612095	5.568164	19.925940	15.923202	23.72633

3.4 Assessing and Comparing Performance of the Models

The performance of each of the modified models in relation to the existing GGE model were assessed using the Monte Carlo estimations provided by Hubert *et al* (2005) below.

a) Mean squared Error (MSE; Hubert *et al (*2005)

$$MSE(\hat{\lambda}_j) = \frac{1}{m}\sum_{l=1}^{m}\left(\hat{\lambda}_j^{(l)} - \lambda_j\right)^2 \quad (8)$$

$\widehat{\lambda_j}$ is the estimated singular value for the *j*-th principal component from the fitted model from the standard or robust GGE model.

ISSN NUMBER: 1116-249X

λj: This is the true population singular value for the *j*-th principal component. In a simulation study, this is a known value set when the data is generated.

m: This is the total number of Monte Carlo replications or simulations run in the experiment.

l: This is an index that specifies a single simulation run out of the total m*m* replications.

- $\lambda_j^{(l)}$ is the estimated singular value for the *j*-th principal component from the *l*-th specific simulation run.

Here, for m= 1000 simulations, we computed the deviation of eigenvalues for both contaminated and uncontaminated simulated data in accordance with the estimation procedure of each model. The mean square error equation (8), which is the Monte Carlo estimate of the simulations were obtained. This process was performed for all the models across all levels of contaminations.

b) Mean proportion of explained variability (MPEV)

$$MPEV = \frac{1}{m}\sum_{l=1}^{m}\frac{\hat{\lambda}_1^{(l)}+\hat{\lambda}_2^{(l)}+\cdots+\hat{\lambda}_k^{(l)}}{\lambda_1+\lambda_2+\cdots+\lambda_p} \quad (9)$$

λ^1(l),$\hat{\lambda}_1^2, \hat{\lambda}_2^2 \ldots, \hat{\lambda}_k^2$: These are the estimated singular values for the first k principal components from the l-th simulation run. The value k is the number of components retained by the model 2 for a biplot.

$\lambda_1, \lambda_2, \ldots, \lambda_p$: These are the true population singular values for all pp principal components, where pp is the maximum possible number of components (the rank of the matrix). This sum represents the total variability in the true underlying model.

m is the total number of Monte Carlo replications or simulations.

l: This is an index that specifies a single simulation run.

For each of the models, the proportion of explained variability was computed in each run of simulation for m=1000 times, and the Monte Carlo estimate is obtained, which is the mean proportion of explained variability (MPEV). The procedure is carried out for the various percentages of contaminations under the study.

c) Biplot Interpretation

ISSN NUMBER: 1116-249X

Biplots help us visually explore how genotypes and environments relate to one another. When genotypes are similar, they appear close together on the plot, while those that differ are farther apart. The same pattern applies to environments those with similar characteristics tend to group or cluster in the same area of the plot.

To illustrate the differences between the models in terms of their biplots, we compared the biplots produced with contaminated data with the biplots produced without contamination for each of the models. For the classical GGE model, as the contaminations increased, the positions of the genotypes (scores) and environmental (loadings) continued to change in plots, as observed in Figures (1-2). This demonstrated that biplots produced with contaminations behaved differently from the ones produced without contamination. For instance, at 15% contamination, Figure (2), we obtained a biplot harder to interpret, and with completely different behavior from the classical GGE model without contamination. This revealed the impact of the outlying observations on the classical GGE model.

The modified models produced biplots in Figure (3-5) that are not difficult to interpret even at extreme contamination levels. This demonstrated the reduced impact of the outlying observations in the data.

ISSN NUMBER: 1116-249X

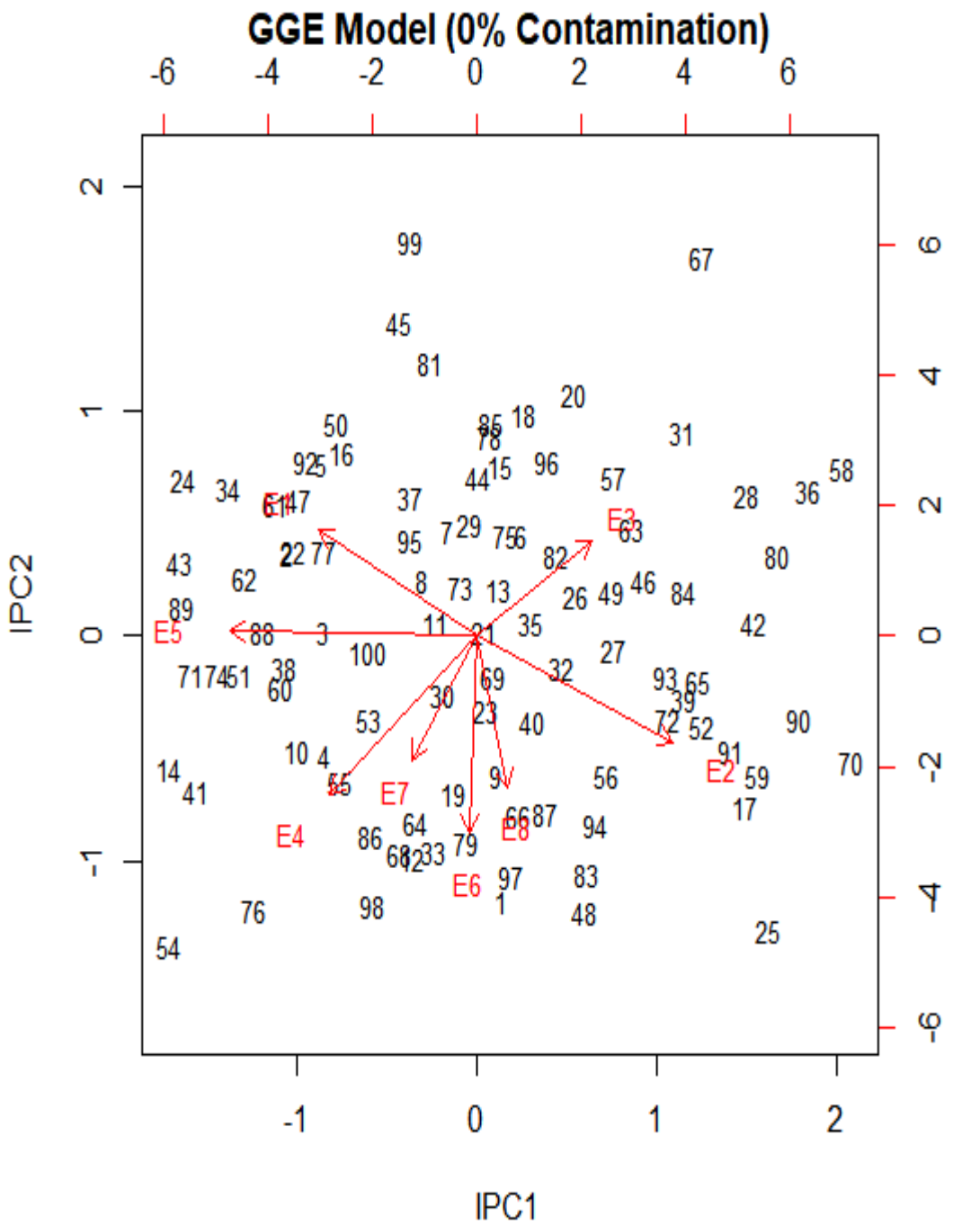


Figure 1: A biplot of GGE model with no contamination

ISSN NUMBER: 1116-249X

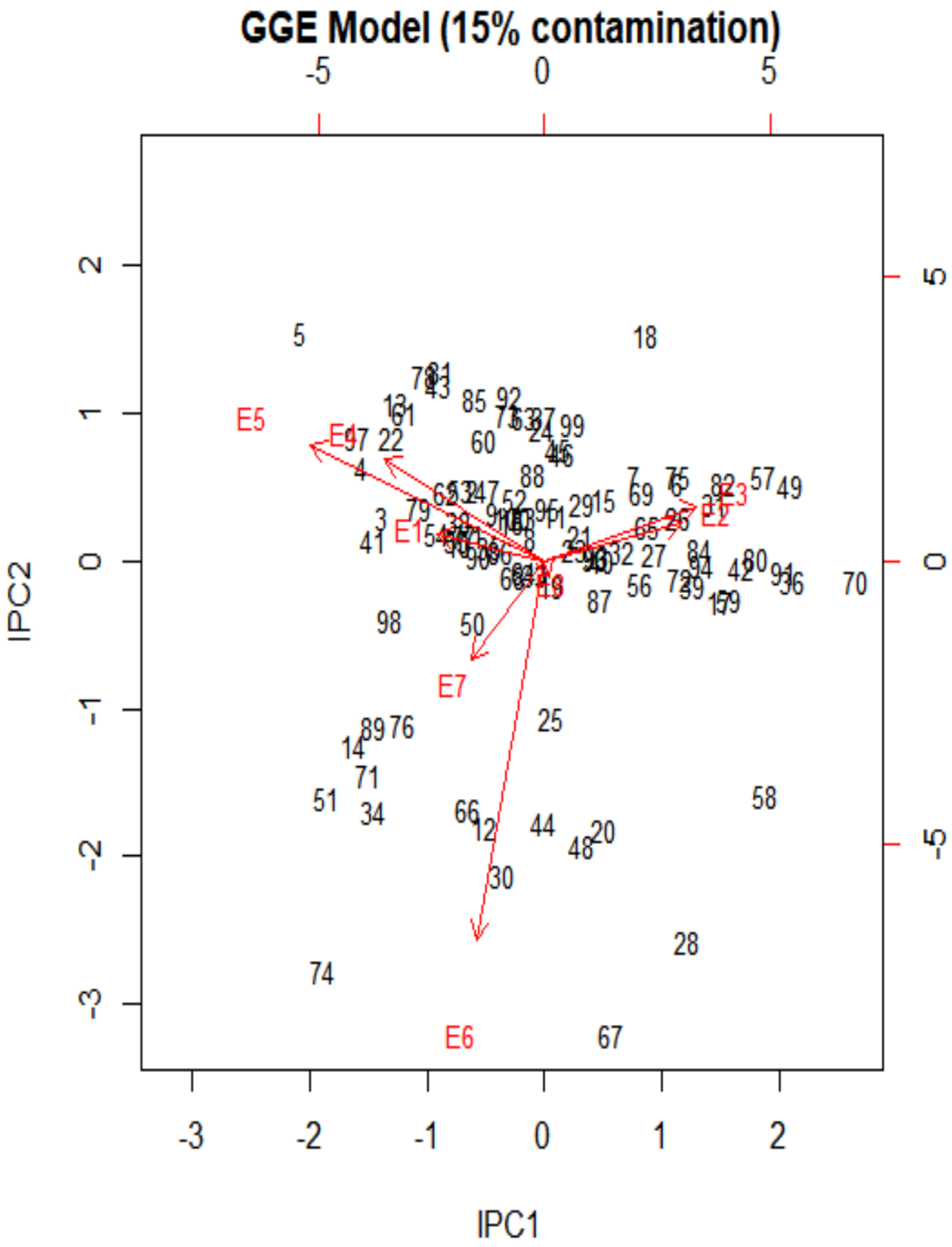


Figure 2: A biplot of GGE model with 15% contamination

ISSN NUMBER: 1116-249X

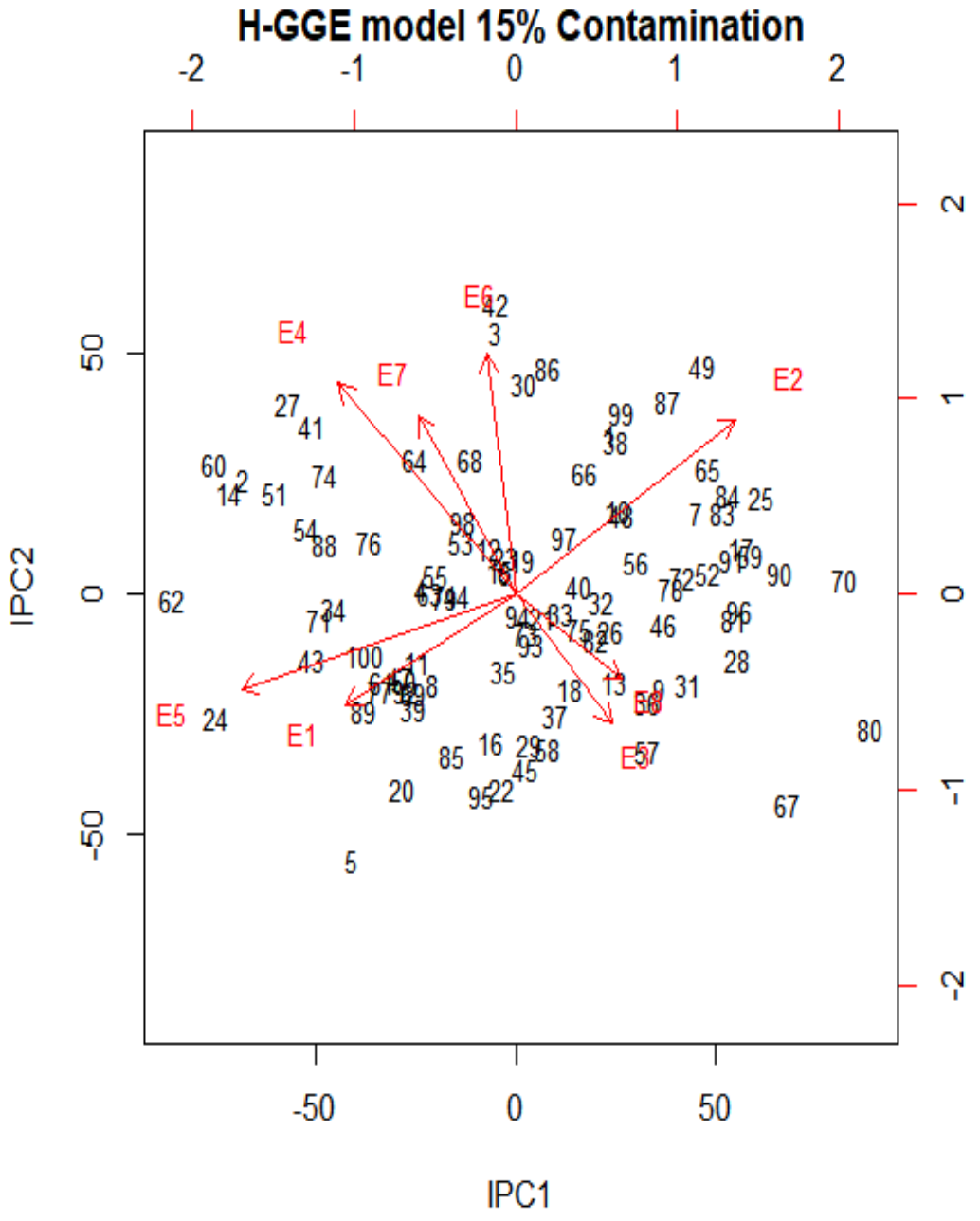


Figure 3: A biplot of H-GGE model with 15% contamination

ISSN NUMBER: 1116-249X

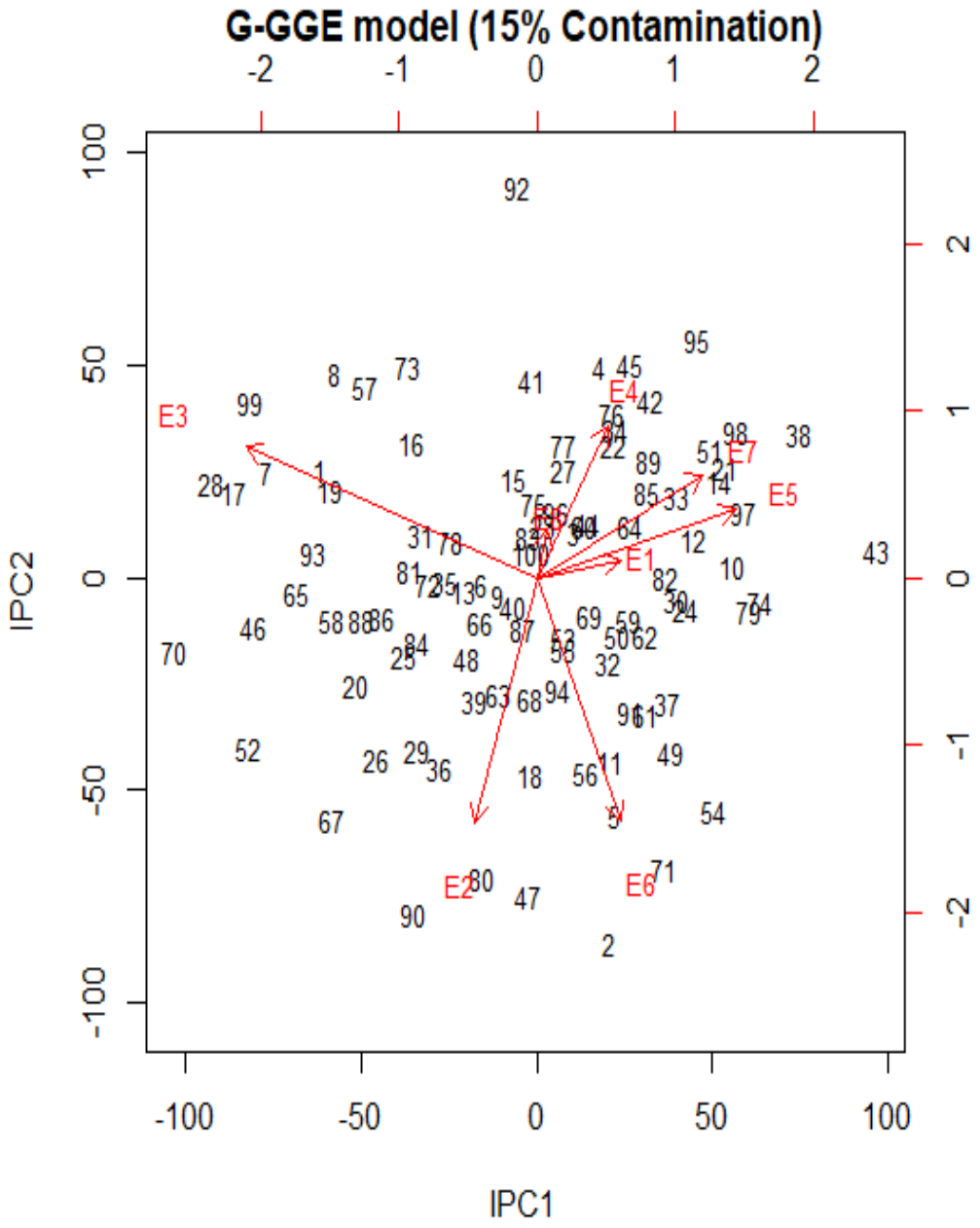


Figure 4: A biplot of G-GGE model 15% contamination

ISSN NUMBER: 1116-249X

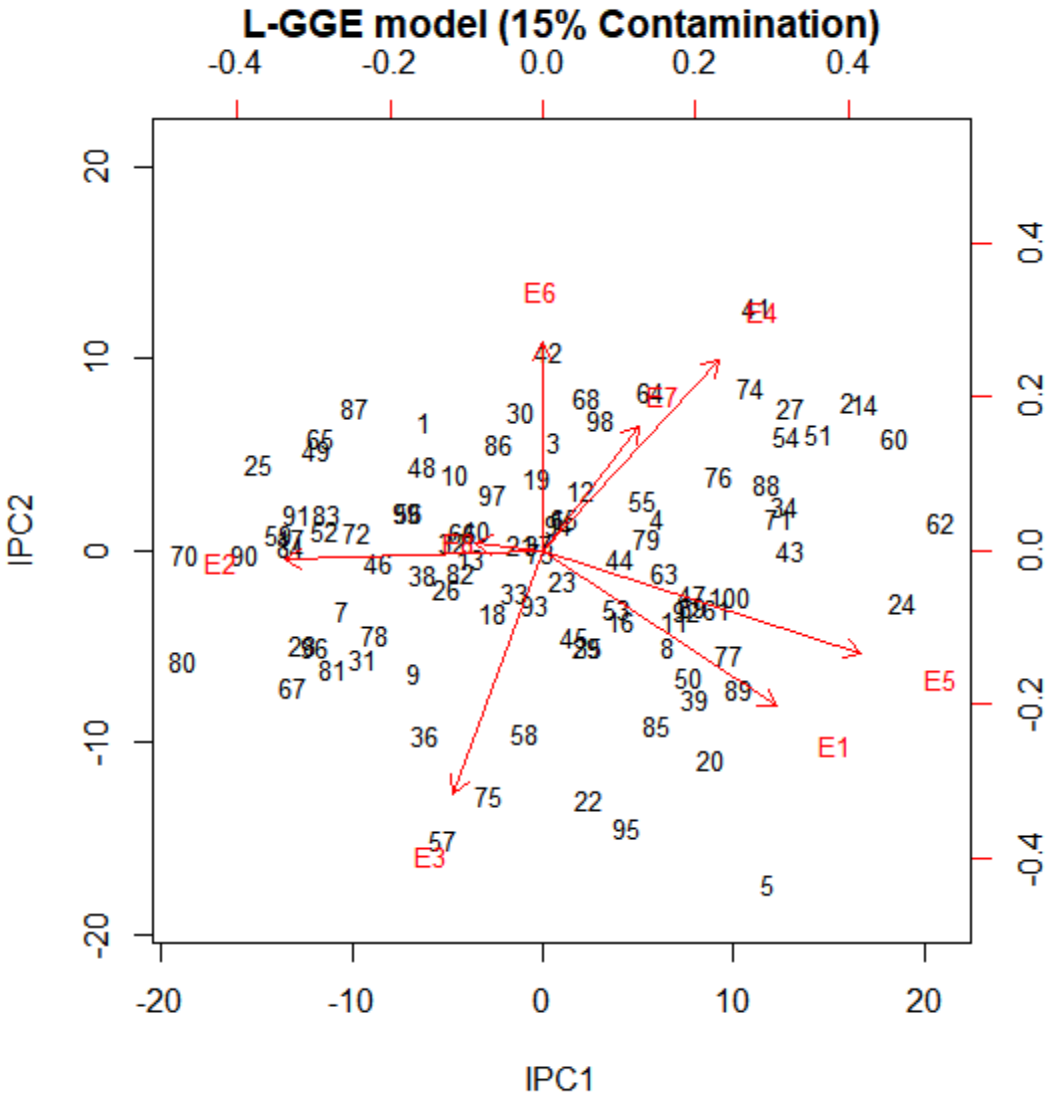


Figure 5: A biplot of L-GGE model with 15% contamination

3.5 Application

We tested our models with multi-environment trials data obtained from Lake Chad Research Institute, Maiduguri, Borno State, involving twelve (12) genotypes of wheat, in nine (9) environments of three major locations of the institute across Northeast and Northwest Nigeria, from 2007-2009.

ISSN NUMBER: 1116-249X

Here, we obtained the biplots and the proportion of variation captured by the first two interactions principal components (IPCs) as accounted by the eigenvalues of each of the models. We presented the biplots in **Figures (4.21-4.24).**

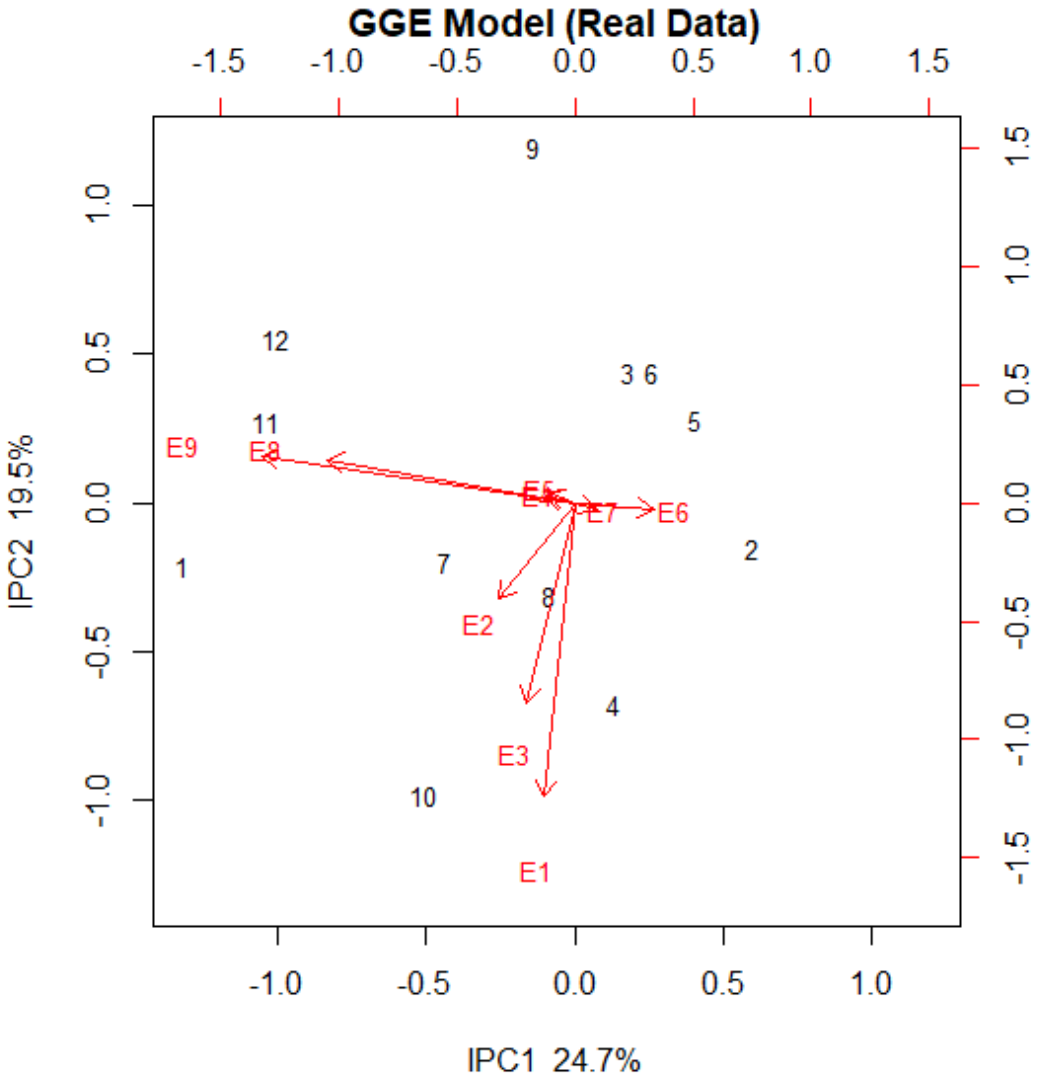
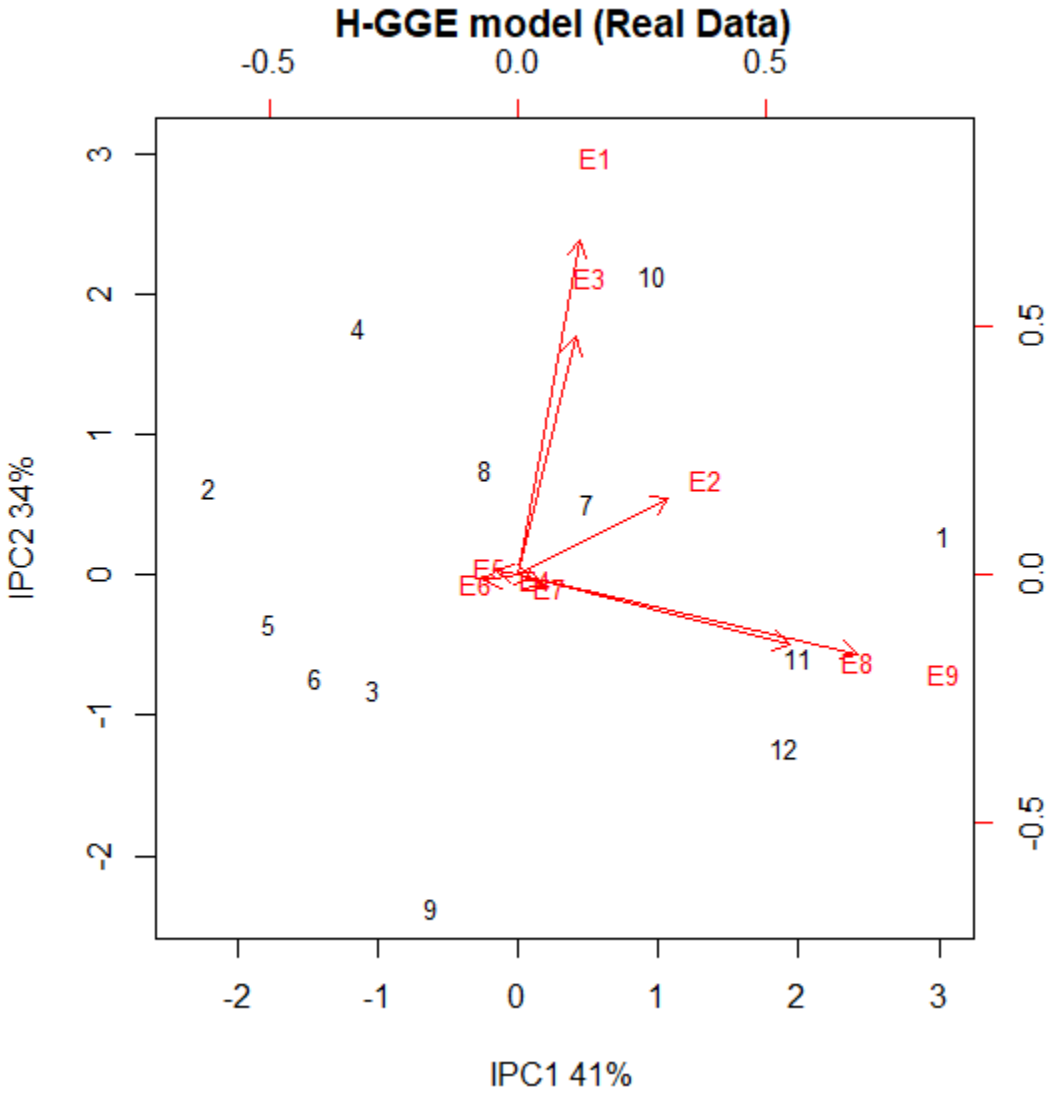


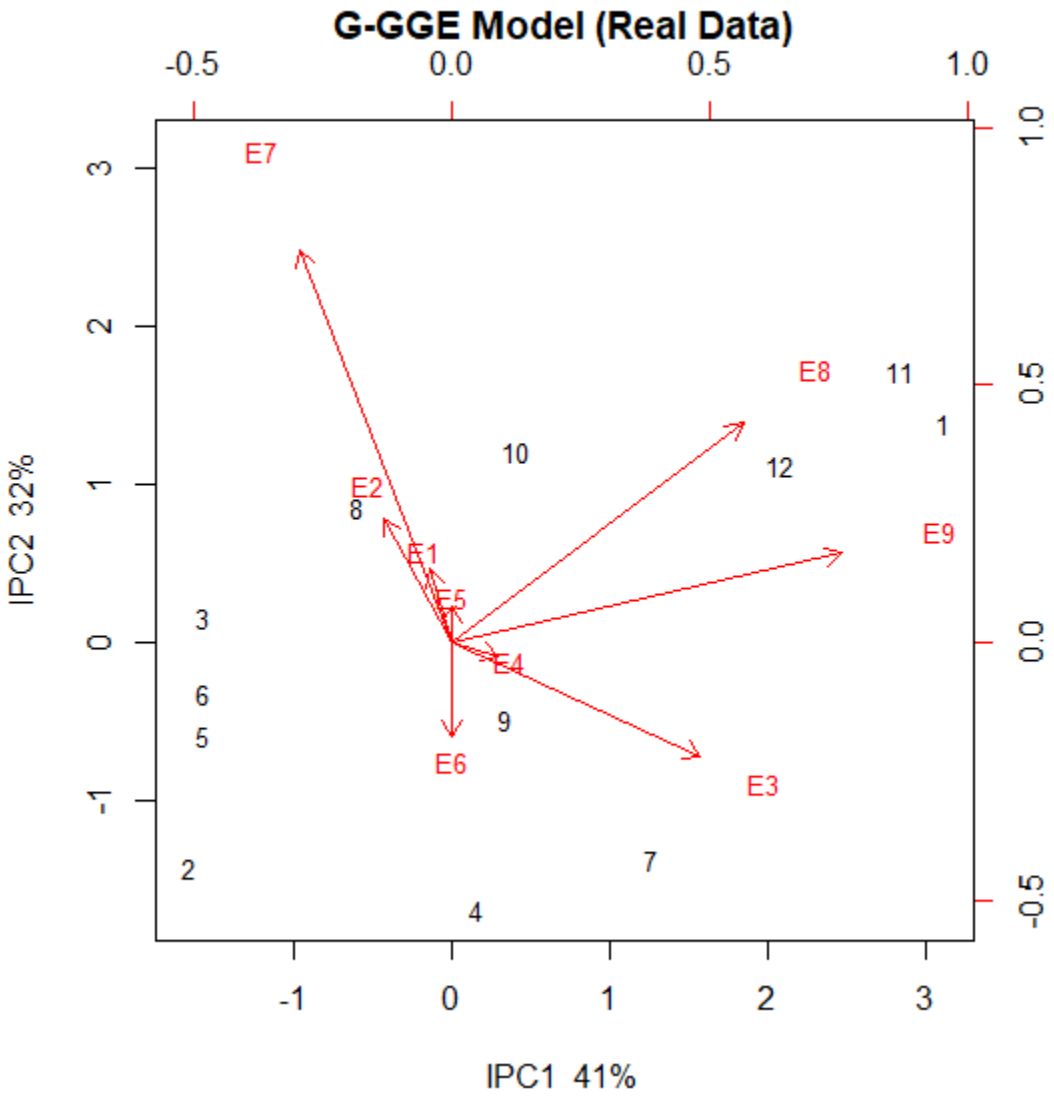
Figure 4.21: A biplot of classical GGE model obtained from real multi-environment data

ISSN NUMBER: 1116-249X



Figur4.22: A biplot of H-GGE model obtained from real multi-environment data

ISSN NUMBER: 1116-249X



Figur4.23: A biplot of G-GGE model obtained from real multi-environment data

ISSN NUMBER: 1116-249X

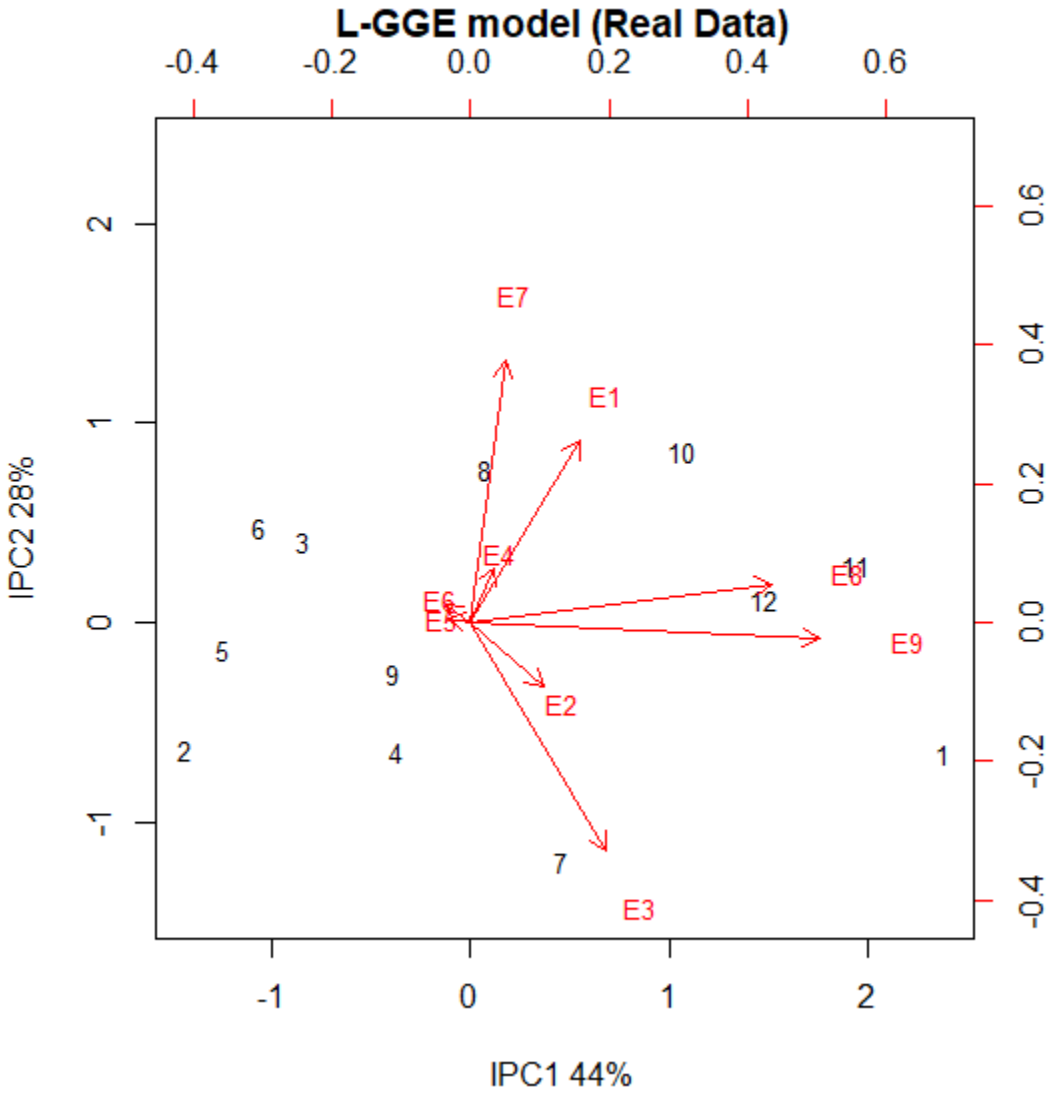


Figure 4.24: A biplot of L-GGE model from real multi-environment data

The biplots Figures (4.21-4.24) illustrate the visual relationship between the genotypes and the environments. Figure 4.21: the biplot of the classical GGE model, indicates that the first two IPCs accounted for 44.2% of the total variation, Figure 4.22: the biplot of H-GGE model, shows that 75% of the total variation are explained by the first two IPCs, Figure 4.23: biplot of GGE model, reveals that 74% of the total variation captured by the first two IPCs and Figure 4.24: biplot of L-GGE model, indicates 72% of the total variation explained by the first two IPCs. The higher the proportion of variation captured by the IPCs the better the performance of the model. Therefore, it implies that all the modified models performed reasonably better than the classical GGE model.

ISSN NUMBER: 1116-249X

3.5: Results

All statistical simulations and estimations were performed using functions in R 4.1.1 software (R core Team, 2018). The summary results are presented below.

Table 3: Mean Square Error for 1st and 2nd Interaction Principal Components for 1000 Simulations using Pure Shift Contamination Scheme

Contamination	Components	Models			
		GGE	H-GGE	G-GGE	L-GGE
2%	IPC 1	8.507	0.0915	0.3130	0.01158
	IPC2	47.138	0.1492	1.1349	0.0089
5%	IPC 1	36.0593	0.8788	1.0455	0.0692
	IPC2	229.406	1.1117	3.7342	0.0198
10%	IPC 1	105.081	6.5369	2.6719	0.1912
	IPC 2	725.587	8.5389	11.1132	0.019
15%	IPC 1	202.5658	7.262	5.0528	0.2999
	IPC 2	1332.077	5.589	21.8858	0.0523
20%	IPC 1	309.741	9.1951	7.6695	0.3775
	IPC 2	1941.498	28.3247	34.8571	0.5546
25%	IPC 1	405.3602	9.7745	10.6521	0.4209
	IPC 2	2488.94	34.2291	48.2464	0.5237
30%	IPC 1	501.0446	8.7501	14.3902	0.4486
	IPC 2	2958.746	36.8501	61.7421	0.0496

The results in Tables 3 shows that the L-GGE model consistently recorded the smallest mean square error for IPCs 1 and 2 compared to the other models at each percentage level of data contamination. This clearly indicates that the L-GGE not only outperforms the existing GGE model but all the other modified models. Closely following the L-GGE model is the H-GGE model which recorded smaller MSE at 2%, 5%,25% and 30% data contaminations. In the overall, the modified models recorded smaller MSE compared to the existing GGE model.

ISSN NUMBER: 1116-249X

Table 4: Mean Proportion of Explained Variability (MPEV) under Pure-shift outlier Scattered environments contamination Scheme in 1000 simulations

Contamination	GGE	H-GGE	G-GGE	L-GGE
2%	106	102	108	90
5%	114	111	115	79
10%	125	135	267	99
15%	134	134	139	121
20%	142	154	149	122
25%	148	114	158	127
30%	153	159	167	131

From Table 4, the result shows that the L-GGE model outperformed the other models by not overestimating the MPEV at 2%, 5% and 10% data contamination. Though, the model overestimated the MPEV at the other percentage levels of contamination, it performed reasonably better than the other models.

3.6: Discussions

The study has examined the fragility of the classical GGE model when fitted with contaminated two-way data. The contamination of such data occurs either due to measurement errors or influence of pest/disease on genotypes in some given environments in multi-environments trials often resulting to lower yields than expected. The classical GGE model works on ANOVA method applied on the additive part of the model and singular value decomposition (SVD) applied on the residual matrix. SVD being a least square method is highly sensitive to contamination and in extreme cases produces misleading results as reported by Rodriques *et al* (2015)

To handle this situation, this study modified the classical GGE model by replacing the ANOVA with a robust fit (M-Regression) and have replaced the standard SVD with robust SVD/PCA. Thus, three (3) robust GGE models (H-GGE, G-GGE, L-GGE) were obtained. These modifications were performed with the help of robust PCA functions in Multi-Carlo simulation study.

The fragility of the classical GGE model and the performance of the modified models were iinvestigated using the pure -shift scattered environment contamination schemes in Monte-Carlo study. The results in Table (3) indicated that across the contamination levels, the mean square error (MSE) increased for the classical GGE model in higher proportion compared to the modified models. The lower the MSE, the better the performance of a model and the closer to 100% the

ISSN NUMBER: 1116-249X

MPEV, the better estimation as reported in Hubert *et al* (2005). It therefore follows that the classical GGE model did not perform better than any of the modified models.

The results in Table (3) also indicated that the L-GGE model consistently recorded a lower MSE compared to other models across the levels of contaminations. In addition, the model recorded MPEV closer to 100% as shown in Table (4) which means it did not overestimate the mean proportion of variation. Closely following the L-GGE in performance is the H-GGE model with lower MSE across the contamination levels.

REFERENCES

Bhatta, M., Gutierrez, L., Cammarota, F., Cardozo, M., & Condori, B. (2023). Multi-environment trials and stability analysis for yield-related traits in commercial wheat cultivars. *Scientific Reports, 13,* 15044.

Bowman, D. T., & Watson, C. E. (1997). The stability of cotton genotypes in the southeastern USA. *Journal of Cotton Science, 1*(1), 30-35.

Bruno, C., & Balzarini, M. (2024). Comparison of additive main effect ‐ multiplicative interaction model and factor analytic model for genotypes ordination from multi ‐ environment trials. *Agronomy Journal, 3*(4), 11-23.

Croux, C., Filzmoser, P., & Oliveira, M. (2007). Algorithms for projection-pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 87(*2), 218–225.

Daba, S. D., Kiszonas, A. M., & McGee, R. J. (2023). Selecting high-performing and stable pea genotypes in multi-environmental trial (MET): Applying AMMI, GGE-biplot and BLUP procedures. *Plants, 12*(12), 2343.

Danakumara, T., Kumar, T., Kumar, N., Patil, B. S., Bharadwaj, C., Patel, U., & Chaturvedi, S. K. (2023). A multi-model based stability analysis employing multi-environmental trials (METs) data for discerning heat tolerance in chickpea (Cicer arietinum L.) landraces. *Plants, 12*(21), 3691.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika, 58*(3), 453–467.

ISSN NUMBER: 1116-249X

Gauch, H. G. (2023). Statistical analysis of yield trials by AMMI and GGE: Further considerations. *Crop Science, 63*(1), 37–58.

Heritier, S., Cantoni, E., Copt, S., & Victoria-Feser, M. P. (2009). Robust methods in biostatistics. John Wiley & Sons.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics, 35*(1), 73–101.

Huber, P. J. (1981). Robust statistics. John Wiley & Sons.

Hubert, M., Rousseeuw, P. J., & Branden, K. V. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics, 47*(1), 64–79.

Huehn, M. (1996). Nonparametric analysis of genotype x environment interactions by ranks. In M. S. Kang & H. G. Gauch (Eds.), Genotype by environment interaction, 213-228.

Lin, C. S., & Binns, M. R. (1994). Concepts and methods for analyzing regional trial data for cultivar and location selection. *Plant Breeding Reviews, 12*, 271-297.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., & Cohen, K. L. (1999). Robust principal component analysis for functional data. *Test, 8*(1), 1–73.

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2019). Robust statistics: Theory and methods (with R) (2nd ed.). Wiley.

Mendes, M. P., de Oliveira, R. A., & Daros, E. (2024). Advanced statistical models for genotype-by-environment interaction in plant breeding: A review. *Agronomy Journal, 116*(2), 455–470.

Mohammadi, R., & Amri, A. (2008). Comparison of parametric and non-parametric methods for selecting stable and adapted durum wheat genotypes in variable environments. *Euphytica, 159*(3), 419–432.

Najafian, G., Kaffashi, A. K., & Jafar-Nezhad, A. (2010). Analysis of grain yield stability in hexaploid wheat genotypes grown in temperate regions of Iran using additive main effects and multiplicative interaction. *Journal of Agricultural Science and Technology, 12*, 213-222.

ISSN NUMBER: 1116-249X

Olivares, B., Cortez, A., Muñoz, E., García, P., & Parra, R. (2022). A robust approach to estimate the stability of agricultural crops under different climate change scenarios. *Agronomy, 12*(9), 2196.

Oliveira, I. C. M., dos Santos, A., & Ferreira, D. F. (2023). Heterogeneity of variances in multi-environment trials: Implications for selection and recommendation of cultivars. *Pesquisa Agropecuária Brasileira, 58*, e03245.

Piepho, H. P., Boer, M. P., & Williams, E. R. (2021). Tensor p-spline smoothing for spatial analysis of plant breeding trials. BioRxiv.

Rocke, D. M., & Woodruff, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association, 91*(435), 1047–1061.

Rodrigues, P. C. (2018). An overview of statistical methods to detect and understand genotype-by-environment interaction and QTL-by-environment interaction. *Biometrical Letters, 55*(1), 1-26.

Rodrigues, P. C., Monteiro, A., & Lourenço, V. M. (2015). A robust AMMI model for the analysis of genotype-by-environment data. *Bioinformatics, 32*(1), 58–66.

Sofi, P., Rather, A. G., & Wani, S. A. (2022). Robust statistical methods for analysis of genotype × environment interaction in crop breeding. In Advances in Crop Breeding, *Springer,* 215–240.

van Eeuwijk, F. A., Bustos-Korts, D. V., & Malosetti, M. (2016). What should students in plant breeding know about the statistical aspects of genotype × environment interactions. *Crop Science, 56*(5), 2119 ‑ 2140.

Yan, W. (2001). GGEbiplot-A Windows application for graphical analysis of multi-environment trial data and other types of two-way data. *Agronomy Journal, 93*(5), 1111–1118.

Yan, W. (2014). Crop variety trials: Data management and analysis. Wiley-Blackwell.

Yan, W., & Hunt, L. A. (2002). Biplot analysis of multi-environment trial data. In M. S. Kang (Ed.), Quantitative genetics, genomics and plant breeding (pp. 289–303). CABI Publishing.

ISSN NUMBER: 1116-249X

Yan, W., & Kang, M. S. (2003). GGE biplot analysis: A graphical tool for breeders, geneticists, and agronomists. CRC Press.

Yan, W., & Tinker, N. A. (2006). Biplot analysis of multi-environment trial data: Principles and applications. *Canadian Journal of Plant Science, 86*(3), 623–645.

Yan, W., Hunt, L. A., Sheng, Q., & Szlavnics, Z. (2000). Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Science, 40*(3), 597–605.

Yang, R. C., Crossa, J., Cornelius, P. L., & Burgueño, J. (2009). Biplot analysis of genotype × environment interaction: Proceed with caution. *Crop Science, 49*(5), 1564–1576.

Yau, S. K. (1995). Regression and AMMI analysis of genotype environment interactions: An empirical comparison. *Agronomy Journal, 87*(1), 121-126.

Zali, H., Farshadfar, E., & Sabaghpour, S. H. (2011). Non-parametric analysis of phenotypic stability in chickpea (Cicer arietinum L.) genotypes in Iran. *Crop Breeding Journal, 1*(1), 89-100.

ISSN NUMBER: 1116-249X