AJBAR Vol. 5(1), February 2026: 111-119, ISSN: 2811-2881

Arid Zone Journal of Basic and Applied Research
Faculty of Science, Kashim Ibrahim University
Maiduguri, Nigeria

Journal homepage: https://www.azjournalbar.com

Research Article

Enhancing Genotype Environmental Stratification Technique Through Robust Factor Analysis: A Modified Statistical Approach

[1]U. Zannah [2]A. Okolo [2]D. Jibasen [2]A.A. Akinrefon
[1]Department of Mathematics and Computer Science, Kashim Ibrahim University.
[2]Department of Statistics Modibbo Adama University Yola, Nigeria
***Corresponding author's Email**: umarzannah2@gmail.com, doi.org/10.55639/607.020100104

ARTICLE INFO:
Keywords:
Robust factor Analysis,
MCD,
G+GE matrix.

ABSTRACT

The use of factor analysis on adaptability and environmental stratification of genotypes and other agronomic and environmental studies has increased considerably over the years. Various methods were proposed on the analysis of adaptability and stratification of genotypes. One of such methods is the FGGE, that works on classical factor analysis in the matrix of genotypic effects (G) added to the effects of the genotype by environment interaction (GE), both fixed, obtained via the ordinary least squares method in joint analysis. The use of factor analysis obtained via the ordinary least squares method is vulnerable to outlying observations however small. To handle this limitation, this study proposed to modify the method using robust factor analysis on G+GE matrix. This proposed method is designated as M-FGGE. The G+GE matrix is obtained from simulated multi-environment trials data of 100 genotypes and 8 environments. Various levels of outliers (10%,20%,30% ,40%and 50%) were simulated and the G+GE matrix was contaminated using scattered environment contamination scheme. To assess and compare the performance of the modified method (M-FGGE) with its classical counterpart (FGGE) in handling outliers, the methods were tested on both contaminated and uncontaminated G+GE matrix. The eigenvalues, accumulated percentages of variance and communalities were obtained and compared. The results indicated that the modified method (M-FGGE) outperformed the existing method (FGGE). The study recommends to plant breeders and other researchers the use of M-FGGE as it offers robust application.

Corresponding author: U. Zanna, **Email**: umarzannah2@gmail.com
Department of Mathematics and Computer Science, Faculty of Science, Kashim Ibrahim University, Nigeria

INTRODUCTION

Factor analysis constitutes a multivariate statistical technique applied to a single set of variables when the investigator is interested in determining which variables form logical subsets that remain relatively independent of one

another, a process fundamentally concerned with identifying underlying factors through the clubbing of related variables (Verma & Abdel-Salam, 2019). This analytical approach is particularly valuable for reducing the dimensionality of large datasets while conserving maximum information, achieved by extracting orthogonal factors that minimize the number of evaluated environments without substantive information loss (Cruz et al., 2014).

The application of factor analysis rests upon several critical assumptions, including that all variables correlate to some degree and that measurements are obtained at least at the ordinal level (Hair et al., 1998; Ho, 2006; Laupichler et al., 2023; Garson, 2022; Minh & Tien, 2024). Researchers must also attend to requirements regarding adequate sample size, absence of multicollinearity, and suitability of the correlation matrix as assessed through Bartlett's test of sphericity and the Kaiser-Meyer-Olkin measure of sampling adequacy (Hair et al., 1998; Ho, 2006).

Within agricultural sciences, factor analysis has been applied to adaptability and environmental stratification of genotypes, enabling the minimization of evaluated environments while preserving interpretability of genotype-by-environment interaction patterns (Wan, 2014; Noora, 2021; Cruz et al., 2014). The approach pioneered by Murakami and Cruz (2004) used factor analysis on adjusted phenotypic means to stratify environments, and was later developed by Garbuglio and Ferreira (2015) through the FGGE approach that incorporates G+GE effects estimated by ordinary least squares.

However, the least squares estimation method is vulnerable to outlying observations, which can distort eigenvalues, communalities, and variance explained in factor analysis. This limitation has not been addressed in subsequent methodological developments. Pinheiro (2024) combined GGE biplots and similarity networks to detect mega-environments, while Souza et al. (2025) and Madhu et al. (2025) applied GGE biplot and AMMI analyses without methodological modification. The WAASB package of Olivoto and colleagues (2024) implemented the original Murakami and Cruz factor analysis but made no extension to robustness. Therefore, this study proposes to modify the FGGE method using robust factor analysis on the G+GE matrix, designated as M-FGGE, to address the susceptibility to outliers in least squares-obtained G+GE matrices.

Factor Model with 'm' Common Factors

Let $X_j = (\mathrm{x_1}, \mathrm{x_2}, \ldots, \mathrm{x_p})$ is a random vector with mean vector μ and covariance matrix Σ. The factor analysis model assumes that $X_j = \mu + \lambda F + \varepsilon$, where $\lambda = \{\lambda_{jk}\}$pxm denotes the matrix of factor loadings; λ_{jk} is the loading of the jth variable on the kth common factor, $F = (\mathrm{F_1}, \mathrm{F_2}, \ldots, \mathrm{F_m})$ denotes the vector of latent factor scores; F_{kj} is the score on the kth common factor and $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p)$ denotes the vector of latent error terms; εj is the jth specific factor.

Variance Explained by Common Factors

Let $X = (X_1, X_2, \ldots, X_p)$ be a random vector with mean vector $\mu = (\mu_1, \mu_2, \ldots \ \mu_p)$ and covariance matrix $\Sigma_{p\times p}$. The factor analysis model assumes that:

$$X = \mu + \Lambda F + \varepsilon$$

where:

$\Lambda = \{\lambda_{jk}\}\{p \times m\}$ is the matrix of factor loadings, with λ_{jk} representing the loading of the j-th variable on the k-th common factor $j = 1, 2, \ldots, p; k = 1, 2, \ldots, m)$

$F = (F_1, F_2, \ldots, F_m)$ is the vector of latent common factors

$\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p)$ is the vector of latent specific factors (error terms)

The model is subject to the following assumptions:

1. $E(F) = 0$ and $E(\varepsilon) = 0$
2. $Cov(F) = I_m$ (the common factors are orthogonal with unit variance)
3. $Cov(\varepsilon) = \Psi$, where Ψ is a diagonal matrix containing the specific variances ψ_j on its diagonal
4. F and ε are independent, so $Cov(F, \varepsilon) = 0$

Variance Explained by Common Factors

The variance of the j-th variable, σ_{jj}, can be decomposed as:

$\sigma_{jj} = h_j^2 + \psi_j$

where

$$h_j^2 = \lambda_{\{j1\}}^2 + \lambda_{\{j2\}}^2 + \cdots + \lambda_{\{jm\}}^2 = \sum_{k=1}^{m} \lambda_{\{jk\}}^2$$

is the communality of X_j, representing the proportion of variance explained by the common factors, and ψ_j is the specific variance (uniqueness) of X_j.

It follows from the model assumptions that:

$\Sigma = \Lambda\Lambda' + \Psi$

Stages of Modification of Garbuglio and Ferreira (2015) Method of Environmental Stratification (FGGE)

Garbuglio and Ferreira (2015) used the classical factor analysis on the G+GE matrix where p genotypes are evaluated in q multi-environments for variable x and can be stated in the following manner.

- Each variable can be expressed by a linear combination of common factors and the specific factor given as follows:

$$\begin{aligned} x_1 &= h_{11} F_1 + h_{12} F_2 \cdots + h_{1m} F_m + \varepsilon_1 \\ x_2 &= h_{21} F_1 + h_{22} F_2 \cdots + h_{2m} F_m + \varepsilon_2 \\ &\cdots \\ x_\ell &= h_{\ell 1} F_1 + h_{\ell 2} F_2 \cdots + h_{\ell m} F_m + \varepsilon_\ell \end{aligned}$$

i.e.

$$x_\ell = \sum_{k=1}^{m} (h_{\ell k} F_k) + \varepsilon_j \qquad (1)$$

In matrix notation:

$X_{\ell x1} = \Lambda_{\ell xm} F_{mx1} + \varepsilon_{\ell x1}$

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_\ell \end{bmatrix}_{\ell x1} = \begin{bmatrix} h_{11} & h_{12} \cdots & h_{1m} \\ h_{21} & h_{22} \cdots & h_{2m} \\ \vdots & \vdots & \vdots \\ h_{\ell 1} & h_{\ell 2} \ldots & h_{\ell m} \end{bmatrix}_{\ell xm} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}_{mx1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_\ell \end{bmatrix}_{\ell x1}$$

In environmental stratification and adaptability of genotypes, $x_1, x_2, \ldots x_\ell$, represent a single variable, such as yield at each of q environments in which the genotypes were evaluated.

And the sum of the genotypic effects and interaction effects can be represented as:

$$G + GE = \begin{bmatrix} (G+GE)11 & (G+GE)12 \cdots & (G+GE)1J \\ (G+GE)21 & (G+GE)22 \cdots & (G+GE)2J \\ \vdots & \ddots & \vdots \\ (G+GE)I1 & (G+GE)I2 \cdots & (G+GE)IJ \end{bmatrix}$$

- The columns of the matrix (G+GE) now replaced the variables $x_1, x_2, \ldots x_\ell$.

For each column or environment q of the G+GE matrix:

$$G + GE = \sum_{k=1}^{m} (h_{\ell k} F_k) + \varepsilon_j \qquad (2)$$

In which, $h\ell k$ is the factorial loading for the q-th environment associated with the k-th factor, which reflects the importance of factor k in explanation of the variable or environment q; Fk is the k-th common

factor; and εj is the specific factor associated with the q-th environment, which captures the specific variation not explained by the linear combination of the factorial loadings with the common factors.

- The coefficients $h\ell k$ is collected into the matrix of loadings Λ.

The following restrictions on the common factor and specific variables are considered:

1. E(F)=E($\mathcal{E}$)=0
2. COV (F)=I_k
3. COV ($\mathcal{E}$)=Ψ with Ψ a diagonal matrix containing on its diagonal the specific factor (or variance)

Let Σ = COV ($G + GE$) denotes the classical covariance matrix.

It follows from equation (2) that,

$$\Sigma = \Lambda\Lambda^T + \Psi \quad (3)$$

- The matrices Λ and Ψ are estimated.

In the classical FA, the matrix Σ is estimated by the sample covariance matrix which is decomposed to obtain the estimate Λ and Ψ using maximum likelihood estimation (MLE) and principal component analysis (PFA) as frequently used methods. However, the estimation is influenced by the presence of outliers. To obtain the robust factor analysis, the classical covariance matrix equation (3) was replaced by a robust covariance estimator. The MCD looks for the subset of m out of all n observations having the smallest determinant of its covariance matrix (typically, m ≈3n/4). The MCD estimator is highly robust with good efficiency properties.

In the classical FGGE method, the covariance matrix in equation (3) is estimated by the sample covariance matrix S:

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$$

where $x_i$is the i-th row vector (genotype across q environments) of the centered G+GE matrix, and $n$is the number of genotypes. Factor extraction then solves:

$S \approx \widehat{\Lambda}\widehat{\Lambda}' + \widehat{\Psi}$

using principal factor analysis or maximum likelihood, yielding loadings $\widehat{\Lambda}$ and specific variances $\widehat{\Psi}$. This estimator $S$is sensitive to outliers, which distort eigenvalues, communalities, and variance explained. In M-FGGE, $S$is replaced by the robust Minimum Covariance Determinant (MCD) estimator $\hat{\Sigma}_{\text{rob}}$, which achieves high breakdown point (≈50%) by minimizing the determinant of the covariance of a subset of size $h \approx \lfloor (n+q+1)/2 \rfloor$to $\lfloor 0.75n \rfloor$:

$$\hat{\Sigma}_{MCD} = arg \min_{H:|H|=h} det\left(\frac{1}{h-1}\sum_{i\in H}(x_i - \hat{\mu}_H)(x_i - \hat{\mu}_H)'\right)$$

with $\hat{\mu}_H$the subset mean. The FAST-MCD algorithm (Rousseeuw & Van Driessen, 1999) is used for computation, followed by reweighting based on Mahalanobis distances to improve efficiency:

$$\hat{\Sigma}_{\text{rob}} = c \cdot \frac{1}{n}\sum_{i=1}^{n} w_i\,(x_i - \hat{\mu}_{\text{MCD}})(x_i - \hat{\mu}_{\text{MCD}})'$$

where $w_i = 1$if the robust Mahalanobis distance is below a cutoff (e.g., $\chi^2_{q,0.975}$), else 0, and $c$is a consistency factor.

The robust factor model then becomes:

$$\hat{\Sigma}_{rob} \approx \Lambda\Lambda' + \Psi$$

Factor extraction proceeds on $\hat{\Sigma}_{rob}$, producing robust loadings, communalities, and variance percentages that are less affected by outliers in the G+GE matrix.

This modification enhances stability in contaminated multi-environment trials, as demonstrated in the simulation results.

3.1 Assessing the Performance of M-FGGE

The performance of the M-FGGE method that uses the robust factor analysis is assessed in relation to the Garbuglio and Ferreira (2015) method (FGGE) that uses the classical factor analysis for environmental stratification of genotypes. We undertook the performance of each of the methods on their response to various levels of outlying observations.

3.2 Data Simulation and Outlier Generation

A two-way multi-environment trials data following the GGE model with two multiplicative terms is simulated as follows:

i. A matrix X with n=100 rows (genotypes) and p=8 columns (environments) with observations drawn from a uniform (min=-0.5, max=0.5) distribution was created.
ii. A Singular Value Decomposition (SVD) was performed on X and the matrices U, V and D containing respectively, the left, the right singular vectors and the singular values were obtained.
iii. The grand mean, environmental effects and the error terms were simulated.
iv. A two-way data table was generated using the GGE2 model in matrix
$$Y = 1_I 1_J^T \mu + 1_I \beta_j^T + 28 * U[,1]D[1,1]V[,1]^T + 15 * U[,2]D[2,2]V[,2]^T + \varepsilon$$
v. G+GE matrix is obtained by removing the grand mean and the environmental effect from the matrix structure.

The matrix structure follows GGE 2 with two components of interactions (i.e., IPCs=2) follows Rodrigues *et al* (2015). Presented in table 4.1 is the simulated first 25 by 8 of the two-way data table.

3.3 Outlier Generation and Data Contamination

Outliers were randomly generated from the pure shift normal distributions N (μ+kσ, σ) in line with Rocke & Woodruff (1996) and Rodriques *et al* (2015). The proportion of outliers generated are 10%,20%, 30% 40% and 50%. The G+GE matrix is contaminated using scattered environment contamination scheme where outliers are randomly introduced into the entire matrix. The contamination is done for the respective levels of the outliers simulated.

4 RESULTS AND DISCUSSIONS

Robust and classical factor analysis on the contaminated and the uncontaminated G+GE matrix were performed. The eigenvalues, accumulated proportion of explained variance and communalities in both cases were obtained. The summary results are presented in Tables (4.1-4.5).

Table 4.1: 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

Simulated data on phenotype of genotypes (25) in eight environments (E1-E8) are shown in Table 4.1 where the smallest value is 4.46 (G24 in E2) to the largest 26.27 (G24 in E1). Adaptation patterns of different genotypes are observed: G14 is well adapted in E4 (23.15), E5 (25.40), and E1 (24.95) and poorly adapted in E2 (7.53) and E3 (8.87), and G17 also is in the same way, with high values in E2 (24.67), E3 (20.42), and E8 (22.98) and low in E5 (7.68). G24 was the most sensitive and registered the highest and lowest values. Environment E8 has always exhibited high yields in most genotypes whereas E2 and E3 are more variable. Crossover interactions are observed since genotype ranking varies across environments, and this proves that there are truly existing genotype-by-environment interaction patterns that could be used to test environmental stratification methods. This clean data is the basis of further outliers contamination at 10% and 50 percent percentages.

Table 4.2: **Summary of Factor Analysis of Simulated G+GE Matrix with No Contamination**

Environments	**Eigenvalues**		**Accumulated Variance %**		**Communalities**	
	M-FGGE	**FGGE**	**M-FGGE**	**FGGE**	**M-FGGE**	**FGGE**
E1	53.6112	59.3040	32	28	14.6753	30.69402
E2	37.0287	35.8794	54	45	20.7049	33.71670
E3	18.4938	26.9319	65	58	16.2117	29.72168
E4	16.0180	24.4918	75	70	21.3996	19.68807
E5	15.1892	21.0976	84	80	17.0904	19.36580
E1	10.1010	18.5446	90	88	28.7471	29.35431
E7	7.6653	12.2171	95	94	20.5902	24.31380
E8	6.9238	10.8945	100	100	25.6116	22.50657

Table 4.2 shows that at 0% contamination, M-FGGE method recorded lower eigenvalues and lower communality values compared to the FGGE method. The eigenvalues represent the amount of total variance explained by each of the factors while the communality represents the proportion of variance of the variables explained by the common factors. The larger the communality value, the more successful the factor model can be said in explaining the variable. The result shows that without contamination the M-FGGE method did not perform better than the FGGE.

Table 4.3: Summary of Factor Analysis of Simulated G+GE Matrix with 10 % Contamination

	Eigenvalues		Accumulated var %		Communalities	
Components	**M-FGGE**	**FGGE**	**M-FGGE**	**FGGE**	**M-FGGE**	**FGGE**
E1	88.6585	80.7999	27	22	51.54	48.32
E2	62.5980	59.1966	46	38	49.95	43.70
E3	54.3216	51.5638	53	52	45.69	36.83
E4	39.5816	42.5362	65	64	67.57	59.35
E5	32.0744	39.7734	74	75	21.91	53.27
E6	22.7323	36.5284	81	85	31.24	44.88
E7	16.8505	28.0857	86	93	42.92	49.34
E8	13.0869	24.4943	100	100	19.09	37.30

At 10% contamination, Table 4.3 shows that, the first four axes in the M-FGGE accounted for greater eigenvalues, higher accumulated percentage of variance as well as higher communalities compared to the FGGE method. From the result, the first four factors in the M-FGGE method accounted for (65%) of accumulated variance compared to (64%) in the FGGE method**.**

Table 4.4: Summary of Factor Analysis of Simulated G+GE Matrix with 20 % Contamination

Components	**Eigenvalues**		**Accumulated var %**		**Communalities**	
	M-FGGE	**FGGE**	**M-FGGE**	**FGGE**	**M-FGGE**	**FGGE**
E1	213.3597	136.1013	26	18	168.05	51.54
E2	166.7581	126.4399	46	34	132.32	49.95
E3	145.4536	112.3254	63	48	38.56	45.69
E4	130.3793	100.1230	79	61	29.83	67.57
E5	64.8002	90.6807	87	73	59.36	21.91
E1	58.6127	85.4585	94	84	137.06	31.24
E7	38.7085	71.0393	98	93	187.26	42.92
E8	18.6055	51.5223	100	100	84.26	19.09

Table 4.4 shows that at 20% contamination, the first seven axes in the M-FGGE indicated higher proportion of accumulated variance (98%) compared to (93%) in the FGGE method. This shows that the M-FGGE has performed better than the FGGE at this level.

Table 4.5: Summary of Factor Analysis of Simulated G+GE Matrix with 30 % Contamination

	Eigenvalues		Accumulated var %		Communalities	
Components	M-FGGE	FGGE	M-FGGE	FGGE	M-FGGE	FGGE
E1	169.1697	134.7231	20	19	134.3217	99.85303
E2	147.9882	113.2407	37	35	112.47358	85.98255
E3	141.5613	93.8746	53	48	117.7691	84.04450
E4	120.2531	88.2738	67	61	131.3728	97.93586
E5	106.1558	81.3887	79	73	82.97051	72.68524
E6	96.5238	79.1682	90	84	148.3542	106.57562
E7	58.3873	54.7718	98	93	105.76616	84.28136
E8	18.7419	50.6049	100	100	25.75302	64.68756

At 30% contamination, Table 4.5 shows that first seven factors in the M-FGGE recorded higher eigenvalues and greater proportion of variance (98%) compared to (92%) in the FGGE method. The M-FGGE also indicated higher communality values as obtained in seven factors compared to the FGGE method. This implies that the M-FGGE has outperformed the FGGE at this level.

Table 4.6: Summary of Factor Analysis of Simulated G+GE Matrix with 40 % Contamination

	Eigenvalues		**Accumulated var %**		**Communalities**	
Components	**M-FGGE**	**FGGE**	**M-FGGE**	**FGGE**	**M-FGGE**	**FGGE**
E1	400.4128	201.1936	25	19	150.67	95.08
E2	297.7969	187.9150	43	37	242.65	149.60
E3	246.3773	158.0707	58	52	256.73	150.53
E4	212.7046	128.6830	71	64	230.78	144.75
E5	200.1761	112.4372	83	75	240.78	148.51
E6	150.5580	111.3114	92	86	224.01	133.98
E7	104.0601	91.5126	98	95	141.66	135.49
E8	2.57561	66.4375	100	100	127.3891	99.62

Table 4.6 shows that, at 40% contamination, the M-FGGE recorded larger eigenvalues and higher proportion of explained variance (98%) compared to (95%) in the FGGE method in seven factors The result also shows higher communality values in M-FGGE compared to the FGGE method. This implies the M-FGGE reported greater proportion of common variance compared to the FGGE method.

Table 4.7: Summary of Factor Analysis of Simulated G+GE Matrix with 50 % Contamination

	Eigenvalues		Accumulated var %		Communalities	
Components	M-FGGE	FGGE	M-FGGE	FGGE	M-FGGE	FGGE
E1	323.9173	208.8545	22	19	164.55	94.80
E2	272.7903	181.2991	40	36	180.71	150.75
E3	253.142	153.4630	57	50	256.14	158.36
E4	220.3624	137.2233	71	63	213.12	132.60
E5	179.4599	132.6723	83	75	260.16	154.48
E1	139.3542	100.4091	92	85	218.50	150.48
E7	102.5018	90.6232	99	92	226.32	138.07
E8	16.4653	81.6853	100	100	150.51	106.69

Table 4.7 shows that at 50% contamination, the M-FGGE indicated larger eigenvalues and higher proportion of accumulated variance (99%) compared to (92%) in the FGGE method. Higher communality values are reported in the M-FGGE method compared to the FGGE method. This means in the M-FGGE method, the factors have explained greater proportion of variance compared to the FGG method.

DISCUSSION

This paper was able to revise the current FGGE approach to environmental stratification of the genotypes by integrating the powerful factor analysis using Minimum Covariance Determinant estimator, resulting in the suggested M-FGGE approach that eliminates the weakness of the traditional factor analysis to outlier observations. Assessment of simulated multi-environment trial data (100 genotypes, 8 environments) with contamination levels of 0 to 50 percent showed that although classical FGGE does a bit better in the absence of contamination, M-FGGE always performs better than FGGE in all contaminated conditions contributing to 65 percent of the accumulated variance as compared to 64 percent at 10 percent contamination and 98 percent compared to 92-95 percent contamination conditions. The mathematical derivation proved that M-FGGE preserves the theoretical basis of FGGE, and has the property of robustness that includes a breakdown point of 25% which guarantees that the environmental stratification choices remain stable even with outliers, typical in field experiments. The research suggests that the M-FGGE should be used by plant breeders and researchers in environmental stratification especially where datasets can be affected by outliers and future studies should be done on whether the method can be applied in real world settings and development of user-friendly software implementations.

REFERENCES

Blasques, G. M., Santos, P. R., Oliveira, L. A., & Mendes, C. T. (2025). Enhancing enviromics based predictions in common bean multi-environment trials. Scientific Reports, 15(1), Article 38070.

Cruz, C.D.; Carneiro, P.C.; Regazzi, A.J. (2014). Modelos Biométricos Aplicados ao Melhoramento Genético. 3rd ed. v.3. Editora UFV, Viçosa, 668p.

Dhakal, B. (2017) "Using factor analysis for residents' attitudes towards economic impact of tourism in Nepal," *International Journal of Statistics and Applications,* 7(5), 250-257.

Garbuglio, D. D., & Ferreira, D. F. (2015). FGGE method: Description and application in data from maize cultivars. Euphytica, 203(3), 723–737.

Garson, G. D. (2022). *Factor analysis and dimension reduction in R: a social Scientist's toolkit*. Routledge.Minh, N. N., & Tien, N. H. (2024). Factors affecting career opportunities abroad for students of faculty of Business Administration of the HCMC University of Food Industry.". *International Journal of Multidisciplinary Reseach and Growth Evaluation*, *5*(1), 556-565.

Hair, J.F., Anderson, R.E., Tatham, R.L. & Black, W. C. (1998). *Multivariate data analysis (5th ed.)*, N J: Prentice-Hall, Upper Saddle River,

Ho, R. (2006) Handbook of univariate and multivariate data analysis and interpretation with SPSS, Chapman & Hall/CRC, Boca Raton.

Laupichler, M. C., Aster, A., Haverkamp, N., & Raupach, T. (2023). Development of the

Rousseeuw, P..J. & Van, Driessen, K. (1999) A fast algorithm for the Minimum Covariance Determinant Estimator. Technometrics, 41,212-223.

Rousseeuw, P.J. (1984). Least median of squares regression. J. Am. Stat. Assoc., 79, 871–880

Souza, C. H., Santiago, A. G. S. G., Batista, E. C., Pulcinelli, C. E., Rocha, T. T. T., Paula, B. S., Miguel, G. S., Bruzi, A. T., & Padua, "Scale for the assessment of non-experts' AI literacy"–An exploratory factor analysis. *Computers in Human Behavior Reports*, *12*, 100338

Madhu, B., Sukrutha, B., Keerthivarman, K., Selvaraj, S., & Premalatha, N. (2025). Integrated stability analysis of compact cotton genotypes for yield and plant architecture under rainfed conditions using AMMI, GGE biplot, WAAS, BLUP, and MTSI. Research Square.

Murakami, D.M & Cruz C.D. (2004). Proposal of methodologies for environment stratification and analysis of genotype adaptability. *Crop Breeding and Applied Biotechnology 4,* 7-11

Noora S. (2021 "Factor Analysis as a Tool for Survey Analysis." *American Journal of Applied Mathematics and Statistics*, *9*(1), 4-11.

Olivoto, T., & Lúcio, A. D. C. (2020). metan: An R package for multi-environment trial analysis. Methods in Ecology and Evolution, 11(6), 783–789.

Pinheiro, C. C. (2024). *Environmental stratification in corn through similarity networks* [master's thesis, Federal University of Lavras]. Institutional Repository of UFLA.

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

Rodrigues P.C., Monteiro A., Lourenço V.M. (2015): A robust additive main effects and multiplicative interaction model for the analysis of genotype-by-environment data. Bioinformatics 32: 58-66.
J. M. V. (2025). Environmental stratification in tobacco final trials.

Verma, J. & Abdel-Salam, A. (2019). *Testing statistical assumptions in research,* John Willey & Sons Inc.

Wan, Y., Qian, Y., Migliaccio, K., Li, Y., Conrad, C. (2014) Linking spatial variations in water quality with water and land management using multivariate techniques. J Environ Qual43:599–610