Factor Analytic Mixed Model Analysis for Multi-Environmental Trials Data

Tarekegn Argaw[1] and Melkamu Demelash[2]

[1]*Computational Science Research Program, EIAR, Climate and Computational Science Research Directorate, Addis Ababa, Ethiopia,* [2]*Climate and Geospatial Research Program, Climate and Computational Science Research Directorate, EIAR, Addis Ababa, Ethiopia;*

Corresponding Author Email: tare.aragaw@gmail.com

Abstract

The analysis of multi-environment trials (MET) data is a critical component of plant breeding and agricultural research, providing essential insights into genotype-by-environment (GxE) interactions. However, as the complexity of MET experiments grows, conversional analysis of variance (ANOVA)-based methods can exhibit limitations in accurately capturing the underlying variance-covariance structure of genetic and non-genetic effects. This study presents a factor analytic mixed model (FAMM) approach to the analysis of MET data, using a dataset of grain yield from ten common bean variety trials conducted in Ethiopia. This study investigated the modeling of variance-covariance structure for genotype-by-environment (GxE) effects and residual error in a multi-environment field trial. The inclusion of a model with heterogeneous error variance resulted in a significant improvement in model fit compared to a base GxE model with heterogeneous genetic variance and constant error variance. Factor Analytic (FA) models of increasing order were then fitted, and the first three orders (FA1, FA2, and FA3) showed remarkable improvements in the percentage of variance explained and statistical significance. The FA3 model, which explained 78.12% of the total variance, was determined to provide the best fit between model complexity and explanatory power. Across the ten trial environments, the estimates of genetic variance, error variance, and heritability ranged widely, from 0.008 to 0.984, 0.053 to 0.695, and 65.40 to 89.86, respectively. This highlighted the substantial variability in the underlying genetic and environmental factors influencing the traits of interest. The genetic correlations between environments also varied from negative to positive values, indicating differing levels of consistency in the genetic factors across experimental conditions. These results demonstrate the importance of properly modeling the variance-covariance structure and considering the complex genotype-by-environment interactions when analyzing multi-environment trial data. It is strongly recommended to scale up the utilization of this efficient analysis method to enhance varietal evaluation across diverse environments, and facilitating the identification of superior varieties..

Keywords: Factor analytic mixed model, *multi-environment trials,* genetic variance, heritability, error variance

Introduction

Crop variety development is a fundamental component of modern agriculture. Advancements in this field over the past century have played a vital role in enhancing global food security, improving farmer livelihoods, and promoting sustainable farming practices (Zenad *et al*., 2021; Begna *et al*. 2020). By creating high-performing, adaptable cultivars that are resilient to biotic and abiotic stresses, researchers have enabled stable and abundant crop yields, bolstered food supplies, and supported the economic well-being of farming communities, while also facilitating the adoption of other sustainable agricultural innovations (Renard *et al*., 2022 ; Zsögön *et al*. 2022).

Continued investment and innovation in crop variety development will be crucial as the world navigates the complex challenges of ensuring long-term food security and environmental sustainability. This domain has been a driving force behind the remarkable progress in agricultural productivity observed over the past decades, and further advancements will be essential for addressing the growing global demand for food, feed, and other agricultural products (Atlin *et al*., 2017)

The analysis of multi-environment trials (METs) in plant breeding and variety testing presents a key challenge, which is the need to appropriately model and exploit genotype–environment interaction (Smith *et al*., 2001a and 2021b). METs are a crucial element of the crop variety development pipeline, where newly bred genotypes are evaluated across a range of agro-ecological environments to capture the influence of diverse environmental factors on genotypic performance. The assessment of genotype-by-environment (G×E) effects is a critical consideration, as it allows researchers to identify superior and stable crop varieties that can perform well under a range of conditions (Verbyla, 2023; Lee *et al*., 2023).

By testing new genotypes across multiple environments, breeders can gain valuable insights into how a variety's traits manifest and interact with the local environmental context. This information is essential for selecting cultivars that exhibit both high productivity and reliability, making them suitable for deployment across a wide geographical area. Classical ANOVA-based methods, such as AMMI and GGE analysis, have enabled researchers to obtain insights into MET data and this benefit must be acknowledged. However, these approaches exhibit several limitations, particularly in handling unbalanced and incomplete data structures (Beeck *et al*., 2010; Zhang *et al*., 2020).

The linear mixed model (LMM) approach has emerged as a more efficient and versatile methodology for the analysis of various types of data, including multi-environment trial (MET) data. Compared to conventional statistical techniques, LMMs provide a flexible modeling framework that can effectively handle a wide range of data structures and assumptions. This flexibility is a key advantage of the LMM approach (Smith *et al*., 2001a).

One critical aspect of the LMM framework is its ability to easily accommodate incomplete or unbalanced data, which is a common issue in many research studies, including MET experiments. MET data often contains missing observations due to various reasons, such as loss of experimental units, failed measurements, or uncontrolled environmental factors (Piepho, 1997; Smith *et al*., 2001a; Kelly *et al*., 2007). The LMM approach can analyze this type of incomplete data without the need to discard entire observations or environments, leading to more efficient use of the available information. Beyond handling missing data, the LMM framework also offers advantages in modeling complex variance structures (Piepho *et al*., 20012; Smith *et al*., 2005). Depending on the research context, LMMs can be extended to incorporate various random effects and covariance structures to appropriately model the sources of variability in the data. This flexibility is particularly useful in studies involving multiple levels of hierarchy or repeated measurements.

One such advanced mixed model approach is the Factor Analytic Multiplicative Mixed (FAMM) model, which builds upon the general LMM framework. The FAMM model offers specific advantages in estimating the variance structure of genotype-by-environment (GxE) effects, which is a crucial aspect in the analysis of MET data. The FAMM model can provide a more informative and interpretable visualization of the GxE patterns, enabling researchers to better understand and leverage the complex interactions between genotypes and environments (Smith *et al*., 2001b and 2005).

Overall, the LMM approach, and its extensions like the FAMM model, have emerged as more efficient and effective methodologies for the analysis of a wide range of data types, including MET data, compared to commonly used statistical techniques. The flexibility and modeling capabilities of LMMs make them invaluable tools for researchers across various disciplines (Kelly *et al*., 2007).

The objective of this research is to explore the potential of FAMM models for effectively analyzing MET data and extracting meaningful insights from the complex G×E interactions. The application of FAMM models is explored, as these approaches provide advantages in estimating the variance structure of GxE effects and enabling more informative visualizations. The ultimate goal is to provide researchers and breeders with a robust and efficient analytical tool for extracting meaningful insights from MET data, ultimately supporting the development of superior and adaptable crop varieties.

The remainder of the paper is structured as follows. Section 2 begins by describing the multi-environment trial (MET) dataset used in this study. It then introduces the statistical models employed, including ANOVA-based models, linear mixed models, and the formula for calculating heritability. This section also provides a detailed account of the data analysis procedures implemented. The results from the MET data analysis are presented in Section 3. Section 4 offers a comprehensive discussion of the results. Finally, the study is concluded in Section 5.

Data and Methods

Motivating data

The FAMM model analysis is illustrated using grain yield data from the 2019 and 2020 common bean variety trials conducted across five locations (Arsinegele, Bako, Goffa, Hawasa, and Melkassa) by the Ethiopian lowland pulses research program. Table 1 presents a summary of 10 trials (location by year combination). All the remaining trials were set up as randomized complete block (RCB) experiments with three replications, arranged in a rectangular (row x column) plot layout. The level of trial connectedness, as indicated by the number of common entries across trials, is high (Table 2), enabling doable genotype-by-environment (GxE) analysis.

Table 1. Summary of trials: Trial location, year replication, number of entries, trial mean yield (t/ha), and number of missing values

Trial Name	Location	Year	Replication	Genotype	Trial Mean	Missing
AN19CBN2	Arsinegele	2019	3	100	2.86	2
AN20CBN2	Arsinegele	2020	3	66	2.82	0
BK19CBN2	Bako	2019	3	40	1.63	0
BK20CBN2	Bako	2020	3	30	1.28	0
GF19CBN2	Goffa	2019	3	40	2.32	0
GF20CBN2	Goffa	2020	3	30	4.1	7
HW19CBN2	Hawasa	2019	3	40	2.47	1
HW20CBN2	Hawasa	2020	3	30	2.08	0
MK19CBN2	Melkassa	2019	3	100	3.18	0
MK20CBN2	Melkassa	2020	3	66	3.68	1

Table 2 Common entries between trials

Site	AN19CBN2	AN20CBN2	BK19CBN2	BK20CBN2	GF19CBN2	GF20CBN2	HW19CBN2	HW20CBN2	MK19CBN2	MK20CBN2
AN19CBN2	100									
AN20CBN2	65	66								
BK19CBN2	40	66	40							
BK20CBN2	30	66	30	30						
GF19CBN2	40	66	40	30	40					
GF20CBN2	30	66	30	30	30	30				
HW19CBN2	40	66	40	30	40	30	40			
HW20CBN2	30	66	30	30	30	30	30	30		
MK19CBN2	100	66	40	30	40	30	40	30	100	
MK20CBN2	65	66	30	30	30	30	30	30	65	66

ANOVA based models

The base-line statistical model for MET data analysis can be written as

$$y_{ikj} = \eta_{ij} + \beta_{kj} + \varepsilon_{ikj}$$
$$\eta_{ij} = \mu + \alpha_i + \delta_j + \gamma_{ij} \qquad (1)$$

where y_{ijk} is yield of the i^{th} entry of replicate block *k* in environment *j* (*i*=1, 2…*m*, *j*=1,2…*t*, *k*=1,2…r), η_{ij} is the empirical/ least-square mean effect of entry *i* in environment *j*, μ is an overall mean effect, α_i is the main effect for genotype i, β_{kj} is the block effect at trial *j*, γ_{ij} is the interaction effect for genotype i in trial j, ε_{ikj} is the random error effect for genotype *i* in replicate block *k* of trial *j*, assumed to be $N(0,\sigma^2)$. The analysis of this model follow the approaches of two stage data analysis, in which the two-way table means η_{ij} are estimated first from the individual trial's analysis, and then the G×E analysis using GGE or AMMI model. The models for the second stage analysis can be written as

$$\gamma_{\text{ij}} = (\eta_{ij} - (\mu + \alpha_i + \delta_j)) = \sum_{i=1}^{c} \lambda_l \tau_{il} \theta_{jl} + \zeta_{ij} \qquad (2)$$

where l = 1, 2, . . . , *c*, λ_l is the singular value of the l^{th} multiplicative or principal component (PC), with $c \leq \min(m-1, t)$, τ_{il} is the eigenvector of genotype i for PC *l*, θ_{jl} is the eigenvector of environment *j* for PC *l*, and ζ_{ij} is the residual associated with genotype *i* in environment *j*, assumed to be NID(0, σ^2/r) where r is the number of replications within an environment. The models are subject to the constraints λ1 ≥λ2, ..., λ*c* ≥ 0 and orthogonally constraints on the τ_{il} scores, that is $\sum_{i=1}^{c} \lambda_l \tau_{il} \tau_{i'l}$ = 1 if $i = i'$ and $\sum_{i=1}^{c} \lambda_l \tau_{il} \tau_{i'l}$ = 0 if $i \neq i'$ with similar constraints on the θ_{jl} scores by replacing symbols (i, m, τ) with (*j*, *s*, *θ*). AMMI analysis uses the model in equation 2

Linear mixed models

A general form of linear mixed model for the n×1 vector y of individual plot yields combined across trials can be written as

$$\mathbf{y} = X\tau + Z_g u_g + Z_o u_o + \varepsilon \qquad (3)$$

where τ is the *a × 1* vector of fixed effects, u_g is an *mt × 1* vector of random *G×E* effects with associated design matrix Z_g, u_o is a *b x 1* vector of (non-genetic) random effect with corresponding design matrix Z_o, ε is the *n × 1* vector of residual error across all trials. Some statistical assumptions are made about the random terms of the general linear mixed models. Thus, we assume that u_g, u_e and ε are mutually independent and have a multivariate normal distribution with zero means vectors and variance matrices $var(u_g)$ = ***Gg***, $var(u_e)$ = ***Go*** and $var(\varepsilon)$ = **R**.

The random non-genetic effects u_o can be considered as sub- vectors $u_{oj}^{(b_j \times 1)}$

for each trial, where b_j is the number of random terms for trial *j*. These random terms are based on the terms for the blocking structure (e.g. replicate blocks or rows and columns of the field). In the analysis of MET data, the sub-vectors of u_o are typically assumed to be mutually independent, with variance matrix G_{oj} for trial j that has a block diagonal form. Thus, there is a variance matrix $G_o = \oplus^t{}_j G_{oj}$ for the set of none-genetic effects at each trial j.

Smith *et al*. (2001b, 2005) presented an alternative parsimonious model for u_g using a factor analysis model to provide a variance structure for the genetic variance matrix G_g. This approach aims to simplify the understanding of genetic variance by reducing the number of parameters while still effectively capturing the underlying relationships among genetic traits. By employing factor analysis, the we sought to identify latent factors that account for the observed genetic variance, thereby enhancing the interpretability of genetic data in their research.

This model can adequately represent the nature of heterogeneous variances and covariances found to occur in most MET data. Thus, the u_g can be modelled with multiplicative terms. That is

$$\begin{aligned} u_g &= (\lambda_1 \otimes I_m) f_1 + \ldots + (\lambda_d \otimes I_m) f_d + \xi \\ &= (\Lambda \otimes I_m) f + \xi \end{aligned} \qquad (4)$$

where λ_h is the $t \times 1$ vector of loadings, f_h is the $m \times 1$ vector of factor scores $(h = 1 \ldots d)$, ξ is the $mt \times 1$ vector of residuals, Λ is the $t \times d$ matrix of loadings $\{\lambda_1 \;.\;.\;.\; \lambda_d\}$ and f is the $md \times 1$ vector of factor scores $(f_1{}' f_2{}' \ldots f_d{}')'$. The random effects f and ξ are assumed to follow a normal distribution with zero mean vector and variance-covariance matrix

$$\begin{bmatrix} G_f \otimes I_m & 0 \\ 0 & \Psi \otimes I_m \end{bmatrix} \qquad (5)$$

where Ψ is a diagonal matrix of specific variances represents the residual variance not explained by the factor model, that is Ψ = *diag* (Ψ_1 . . . Ψ_t). The factor scores are commonly assumed to be independent and scaled to have unit variance, so that $G_f = I_d$. The genetic effects u_g can be considered as a two dimensional (genotype by environment) array of random effects, and can be assumed to have a separable variance structure for the (*mt* × *mt*) variance matrix G_g which can be written as

$$G_g = \mathrm{G_e} \otimes G_g \qquad (6)$$

where G_e is the $t \times t$ genetic variance matrix representing the variances at each trial and covariances between trials, and G_g is the $m \times m$ symmetric positive definite matrix represents variances of environment effects at each genotype and the covariances of environment effects between genotypes. It is typically assumed that the varieties are independent and that

$G_g = I_m$. However, if the pedigree information of the varieties is available, other forms of G_g can be applicable (Smith *et al.*, 2001b; Oakey *et al.*, 2006 and 2007). Based on equation 2 the variance of genetic effects would be

$$\begin{aligned} \mathrm{var}(u_g) &= (\Lambda\Lambda' + \Psi) \otimes I_m \\ &= \mathrm{G_e} \otimes I_m \end{aligned} \qquad (7)$$

Thus, the FA model approach results in the following form for G_e

$$G_e = \Lambda\Lambda' + \Psi \qquad (8)$$

In the model, the variance parametric in these variance matrices are directly estimated using REML estimation method.

Heritability formula

According to the methodology outlined by Cullis *et al.* (2006), the heritability (H_j^2) value for the jth trial can be calculated from a generalized formula that is employed within the context of linear mixed model analysis. This formula is as follows:

$$H_j^2 = 1 - \frac{A_j}{2{\sigma_{gj}}^2} \qquad (9)$$

where A_j is the average pairwise prediction error variance of genetic effects for the j^{th} environment and ${\sigma_{gj}}^2$ is the genetic variance at environment j

Statistical inferences, analysis procedures and software

Fitting a linear mixed model involves estimating the values of the fixed effects (τ), a random GxE effects (u_g), the random non-genetic effects (u_o), as well as the variance-covariance parameters in G_g, G_o, and R. This estimation process comprises two interconnected steps. First, the variance parameters of the model are estimated using Residual Maximum Likelihood (REML), an approach introduced by Patterson and Thompson (1971). Second, the fixed and random effects are estimated using distinct techniques - Best Linear Unbiased Estimation (BLUE) is employed for the fixed effects, while Best Linear Unbiased Prediction (BLUP) is used for the random effects.

To assess the statistical significance of the random effects in the linear mixed model, the Residual Maximum Likelihood Ratio Test (REMLRT) can be utilized. However, it is important to note that the REMLRT is only applicable when comparing the fit of two nested models that share the same fixed effects structure.

The modeling of genotype-by-environment (G×E) effects was carried out using the model fitting procedures demonstrated by De Faveri (2013) and Smith (1999). In this analysis, a combined model was first fitted, which is a combined form of the individual trial models constructed in in the individual trials analysis. This combined model forms the basis of a sequence of models to be fitted for the G×E analysis, and it helps to organize the trial-specific models in a combined form and to confirm the presence of genetic variance in each trial. If any trial is found to have no genetic variance, it would be excluded from the multi-environment trial (MET) data analysis.

Factor Analytic (FA) models were then considered. The adequacy of the FA models with several factors (h) was formally tested, as they were fitted within a mixed model framework. A model with h factors, denoted as FA-h, is nested within a model with h+1 factors. The models were compared, such as FA-1 versus FA-2, FA-2 versus FA-3, and so on. Both the Residual Maximum Likelihood Ratio Test (REMLRT) and the total percentage of the G×E variance (%var) explained by factor components were used to identify the final plausible FA models.

The licensed version of the ASReml-R statistical software package was used to fit all models analyzed in this study (Butler, 2009). ASReml-R is a specialized software application designed for fitting linear mixed models, which was well-suited for the data and research questions addressed here.

Results and Discussion

Modeling variance covariance structure for GxE effects and residual error

The inclusion of a model with heterogeneous error variance on top of the base GxE model with heterogeneous genetic and constant error variance resulted in a significant improvement in model fit. The likelihood ratio test yielded a test statistic of 73.62 with a p-value less than 0.001, indicating strong statistical evidence to support the inclusion of the heterogeneous error variance component (Orellana *et al.*, 2024 ; Smith *et al.*, 2019). We then proceeded to fit Factor Analytic (FA) models up to order 4, and the first three orders of the FA model showed remarkable improvements in the percentage of variance explained and statistical significance.

The FA1 model accounted for approximately 62.65% of the total variance, with a highly significant p-value of less than 0.001 (Argaw *et al.*, 2024). The FA2 model improved upon this, explaining around 70.85% of the total variance, with a p-value of 0.009. The FA3 model further enhanced the explanation, capturing approximately 78.12% of the total variance, again with a p-value less than 0.001. While the FA4 model explained an impressive 96.89% of the total variance, the statistical significance of this model was not significant (p-value <0.169), suggesting that the additional latent factors in the FA4 model did not provide a substantial improvement in the model fit compared to the FA3 model (Smith *et al.*, 2005 ; Kelly *et al.*, 2009). Ultimately, we determined that the FA model of order 3 with heterogeneous error variance provided the best fit to the data, as it offered a favorable balance between model complexity and explanatory power, with significant improvements in the percentage of variance explained and statistical significance.

By incorporating a model with heterogeneous error variance, we were able to achieve a significant improvement in the model fit, indicating that the assumption of constant error variance was not appropriate for such types of data (Smith *et al.*, 2019). The Factor Analytic (FA) models of increasing order provided valuable insights into the underlying structure of the data. The first three orders of the FA model (FA1, FA2, and FA3) demonstrated significant improvements in the percentage of variance explained and statistical significance, suggesting that the data had a complex covariance structure that could not be adequately captured by the base model alone.

Thus, the FA model of order 3 effectively captures the genetic correlations between environments, leading to more accurate estimates of genotype performance. This capability greatly facilitates the selection of lines in breeding programs, enabling breeders to make more informed decisions that enhance the overall effectiveness of their selection strategies

Table 3. Variance covariance model comparisons for GxE effects and residual error: the total percentage of the G×E variance (%var) explained by the FA components, residual log-likelihoods (LR), and residual maximum likelihood ratio tests (REMLRT)

Variance covariance models	%var	LR	REMLRT	Final model
H.Gvar and C.Evar	-	-188.535		
H.Gvar and H.Evar	-	-73.6252	<0.001	
FA1 and H.Evar	62.65	-33.6854	<0.001	
FA2 and H.Evar	70.85	-26.422	0.009	
FA3 and H.Evar	78.12	-18.3014	0.001	FA3 and H.Evar
FA4 and H.Evar	96.89	-16.021	0.169	

H.Gvar =Heterogeneous genetic variance; C.Evar=constant error variance; H.Evar=heterogeneous error variance

Estimates of genetic parameters

Across the trials, the estimates of genetic variance, error variance, and heritability ranged from 0.008 to 0.984, 0.053 to 0.695, and 65.40 to 89.86, respectively, as presented in Table 4. Notably, the trial GF20CBN2 had relatively high genetic and error variance estimates compared to the other trials, while the trial BK20CBN2 had relatively low genetic and error variance estimates.

The results revealed a wide range of genetic variance, error variance, and heritability estimates across the trials, suggesting substantial variability in the underlying genetic and environmental factors influencing the traits of interest (Smith *et al*., 2001; Kelly *et al*., 2007). The high genetic and error variance observed in the GF20CBN2 trial indicates that this trial site has high potential for discriminating between genotypes, but also a high degree of residual or unexplained variation. The wide range of variance components and heritability estimates across trials highlights the complex and context-dependent nature of the genotype-by-environment interactions influencing the target traits.

Conversely, the low genetic and error variance estimates for the BK20CBN2 trial suggest a more homogeneous genetic background and a more controlled experimental environment, leading to less pronounced genetic effects and lower residual variation. These highlight the importance of considering the specific trial conditions when interpreting the genetic and environmental contributions to the observed traits (Beeck *et al*., 2010,).

The genetic correlations between environments, presented in Table 5, ranged from negative to positive values, indicating varying degrees of genetic relationships across the experimental conditions. The estimates also varied from strong to relatively weaker correlations, suggesting differing levels of consistency in the genetic factors influencing the traits.

Table 4 Summary of genetic variance, error variance and heritability from the final fitted model

Site	Genetic variance	Error variance	Heritability
AN19CBN2	0.209	0.684	69.24
AN20CBN2	0.281	0.36	81.81
BK19CBN2	0.062	0.125	78.06
BK20CBN2	0.008	0.053	76.4
GF19CBN2	0.166	0.105	89.86
GF20CBN2	0.984	0.695	88.43
HW19CBN2	0.28	0.394	88.71
HW20CBN2	0.197	0.142	89.42
MK19CBN2	0.117	0.279	65.4
MK20CBN2	0.07	0.265	70.97

Table 5 Summary of genetic correlation from the final fitted model analysis

Trial	AN19CBN2	AN20CBN2	BK19CBN2	BK20CBN2	GF19CBN2	GF20CBN2	HW19CBN2	HW20CBN2	MK19CBN2	MK20CBN2
AN19CBN2	1									
AN20CBN2	0.675	1								
BK19CBN2	-0.523	0.187	1							
BK20CBN2	-0.349	0.315	0.974	1						
GF19CBN2	0.622	0.946	0.182	0.277	1					
GF20CBN2	0.397	0.671	0.245	0.329	0.661	1				
HW19CBN2	0.473	0.894	0.383	0.481	0.891	0.65	1			
HW20CBN2	0.651	0.926	0.217	0.353	0.901	0.654	0.868	1		
MK19CBN2	0.431	0.077	-0.134	0.062	-0.09	0.043	-0.013	0.149	1	
MK20CBN2	0.479	0.668	0.397	0.595	0.545	0.504	0.645	0.711	0.68	1

The visualization techniques from the final fitted GxE model

The visualization techniques from the final fitted GxE model analysis are presented in Fig. 1, Fig 2 and Fig. 3. These visualizations provide valuable insights into the underlying structure and patterns within the data (Tesfaye *et al*., 2023; Argaw *et al*., 2024).

The genetic correlation heat map (Fig. 1) provides a visualization of the correlations between trials. The coloration of the heat map indicates the strength and direction of the correlations. The deep red coloration corresponds to strong positive correlations among certain trials. This suggests a high degree of similarity in the genetic responses of these trials to the environmental factors, indicating a close relationship between them. In contrast, the yellow hues in the heat map represent weak positive and negative correlations between other trial pairings (Tesfaye *et al*., Beeck *et al*., 2010). This indicates more complex or nuanced relationships between these trials, where the genetic effects are less straightforward. Finally, the deep blue shading in the heat map highlights strong negative correlations between specific trials (Cullis *et al*., 2010 ; Tesfaye *et al*., 2023). This suggests that these trials have contrasting genetic responses to the environmental conditions, implying that they represent significantly different growing environments.

The dendrogram representation of the dissimilarity matrix for the dataset is presented in Figure 4. This hierarchical clustering visualization groups the trials based on their similarity, providing insights into the underlying relationships between the environments (Cullis *et al*., 2010; Argaw *et al*., 2023). The key observation from the dendrogram is that a dissimilarity cut-off at 0.5 delineates three distinct trial clusters. The first cluster comprises AN19CBN2, AN20CBN2, GF19CBN2, GF20CBN2, HW19CBN2, and HW20CBn2, the second cluster includes MK19CBN2 and MK20CBN2, and the third cluster consists of BK19CBN2 and BK20CBN2. This suggests that the trials within each of these clusters exhibit relatively strong correlations, indicating a high degree of similarity in the genetic responses of the genotypes to the environmental factors (Cullis *et al*., 20210 ; Beeck *et al*. 2010). The dendrogram visualization allows us to identify the most closely related trials, which can inform the selection of genotypes or varieties. By focusing on the trials within the tightly clustered groups, we can make more informed decisions about which genotypes or varieties to choose based on their consistent performance across the various environments. This information can be valuable for breeders and agronomists in developing and deploying well-adapted cultivars (Beeck *et al*., 2010; Burgueño *et al*., 2011).

The FAMM analysis also generates bi-plots, which provide further insights into the relationships between the trial environments (Cullis *et al*., 2010; Argaw *et al*., 2023). The bi-plot shown in Fig 3 reveals patterns of correlated trials and highlights the discriminating power of individual trials. Trials with longer arms extending from the center of the bi-plot had higher genetic variance compared to the others, indicating a greater ability to differentiate between genotypes (Tesfaye *et al*., 2023). In this case, the trials GF20CBN2, HW19CB2, and AN20CBN2 had the highest genetic variance and, consequently, the greatest discriminating power. Conversely, the trials BK20CBN2, BK19CBN2, and MK20CBN2 had relatively low genetic variance, and therefore exhibited lower discriminating power for genotypes (Beeck *et al*., 2010). This information can help researchers identify the most informative trial environments for evaluating and selecting superior genotypes. The insights gained from the bi-plot analysis, in conjunction with the dendrogram visualization, provide a comprehensive understanding of the relationships between the trial environments and their suitability for genotype assessment and selection (Burgueño *et al*., 2011). Leveraging these insights can lead to more efficient and effective breeding programs

Overall, the combination of the genetic correlation heat map (Fig. 1), the dendrogram visualization (Fig. 2) and the bi-plot (Fig. 3) provide a comprehensive understanding of the complex genotype-environment interactions within the data. These visualizations help reveal the underlying patterns and relationships, enabling more informed decision-making regarding the selection of genotypes or varieties.

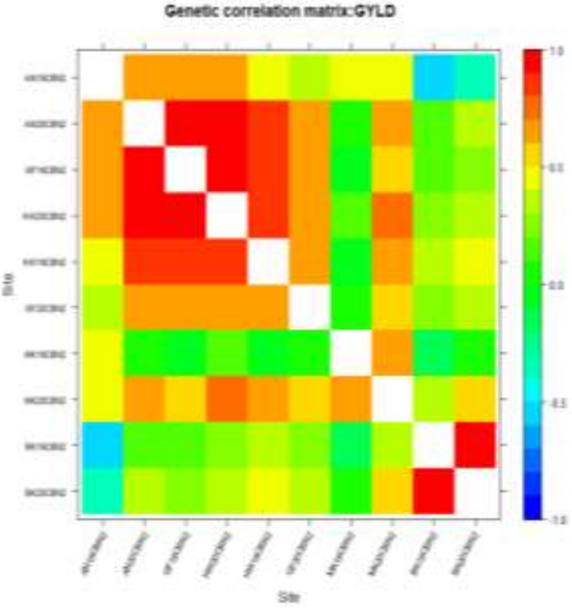

Fig. 1 Heat map representation of the genetic correlation matrix from the final fitted GxE mode analysis

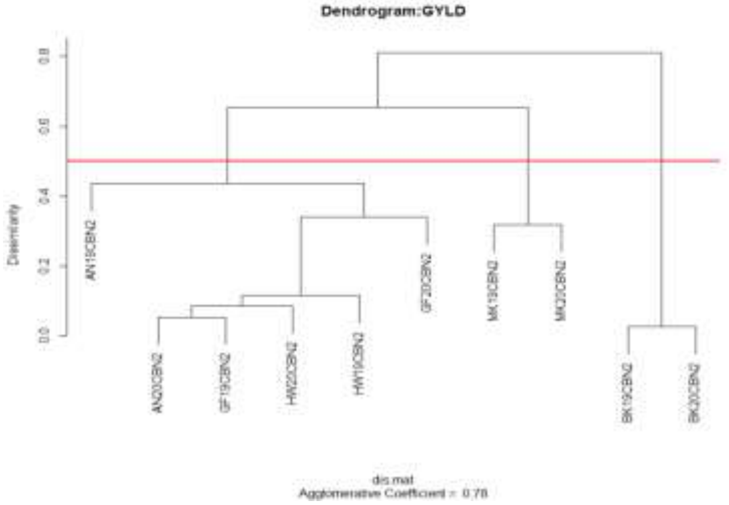

Fig. 2 Dendrogram representation of the dissimilarity matrix from final fitted GxE model analysis

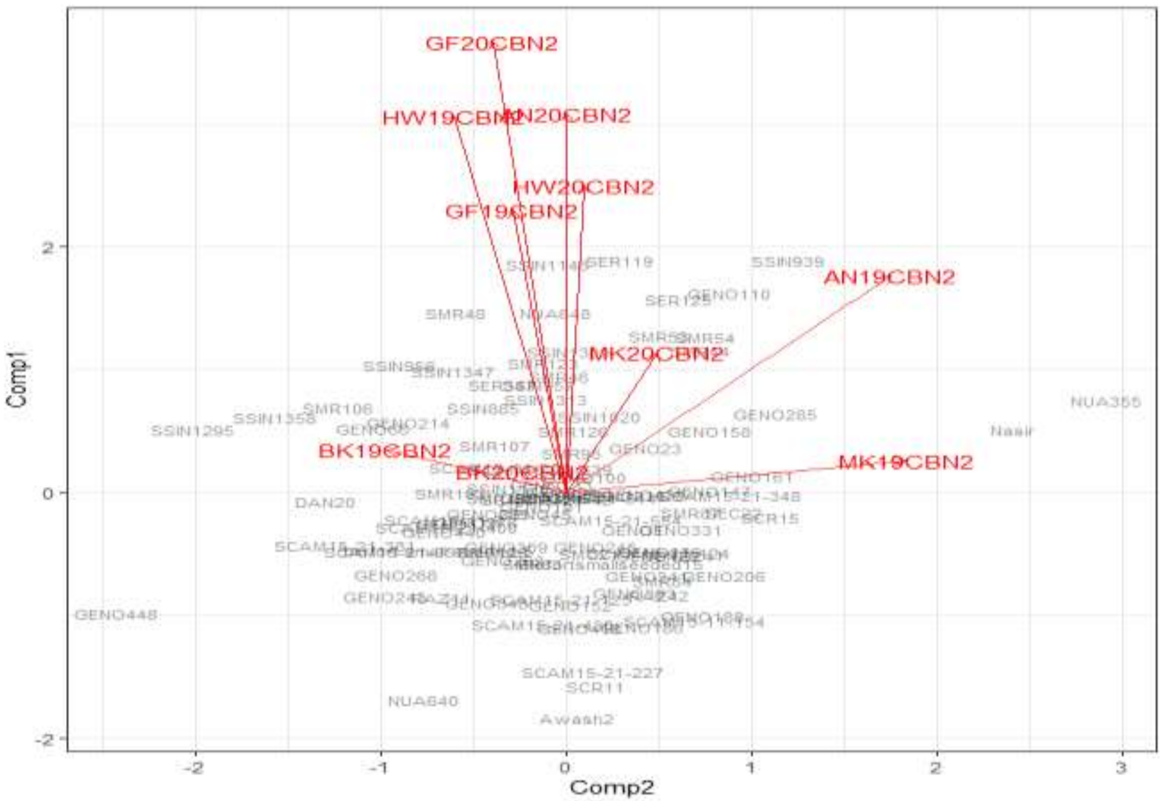

Fig. 3 Bi-plot representation of FA components from final fitted GxE model

Conclusion

The study employed advanced statistical modeling approaches to investigate the genotype-by-environment (GxE) effects and residual error structure in the data. The inclusion of a model with heterogeneous error variance resulted in a significant improvement in model fit. This indicated that the assumption of constant error variance was not appropriate for multi-environmental trials datasets. The Factor Analytic (FA) models of increasing order provided valuable insights into the underlying covariance structure of the data. The first three orders of the FA model (FA1, FA2, and FA3) demonstrated significant improvements in the percentage of variance explained and statistical significance, suggesting a complex covariance structure that could not be adequately captured by the base model alone. The FA3 model, explaining approximately 78.12% of the total variance, was determined to provide the best balance between model complexity and explanatory power.

The genetic parameter estimates, including genetic variance, error variance, and heritability, revealed substantial variability across the different trial environments. This highlights the context-dependent nature of the genotype-by-environment interactions and the importance of considering the specific trial conditions when interpreting the genetic and environmental contributions to the observed traits. The genetic correlations between environments also ranged from negative to positive values, indicating varying degrees of genetic relationships across the experimental conditions. The estimates varied from strong to relatively weaker correlations, suggesting differing levels of consistency in the genetic factors influencing the traits. Overall, the findings from this study emphasize the complex and nuanced nature of GxE interactions and the importance of using appropriate statistical models to capture the underlying structures in the data. The insights gained can contribute to a better understanding of the genetic and environmental factors influencing the traits of interest and inform future research and breeding strategies.

Acknowledgements

We gratefully acknowledge the support of the Ethiopian Institute of Agricultural Research (EIAR), National Lowland Pulse Research Program, for providing the necessary facilities. We also thank the National Lowland Pulse researchers for generating and supplying the field trials data used in this study.

References

Argaw, T., Fenta, B.A. and Assefa, E. 2024. Application of factor analytic and spatial mixed models for the analysis of multi-environment trials in common bean (Phaseolus vulgaris L.) in Ethiopia. Plos One 19(4): e0301534.

Atlin, G.N., Cairns, J.E. and Das, B. 2017. Rapid breeding and varietal replacement are critical to adaptation of cropping systems in the developing world to climate change. Global Food Security 12: 31–37.

Beeck, C.P., Cowling, W.A., Smith, A.B. and Cullis, B.R. 2010. Analysis of yield and oil from a series of canola breeding trials. Part I. Fitting factor analytic models with pedigree information. Genome 53(11): 992–1001.

Begna, T. and Begna, T. 2021. Role and economic importance of crop genetic diversity in food security. Int. J. Agric. Sci. Food Technol. 7(1): 164–169.

Butler, D.G., Cullis, B.R., Gilmour, A.R. and Gogel, B.J. 2009. ASReml-R reference manual. The State of Queensland, Department of Primary Industries and Fisheries, Brisbane.

Cullis, B.R., Smith, A.B., Beeck, C.P. and Cowling, W.A. 2010. Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. Genome 53: 1002–1016.

Cullis, B.R., Smith, A.B. and Coombes, N.E. 2006. On the design of early generation variety trials with correlated data. J. Agric. Biol. Environ. Stat. 11: 381–393.

De Faveri, J. 2013. Spatial and temporal modelling for perennial crop variety selection trials. Doctoral dissertation, University of Adelaide.

Kelly, A.M., Cullis, B.R., Gilmour, A.R., Eccleston, J.A. and Thompson, R. 2009. Estimation in a multiplicative mixed model involving a genetic relationship matrix. Genet. Sel. Evol. 41(1): 1–9.

Kelly, A.M., Smith, A.B., Eccleston, J.A. and Cullis, B.R. 2007. The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. Crop Sci. 47(3): 1063–1070.

Kenward, M.G. and Roger, J.H. 1997. The precision of fixed effects estimates from restricted maximum likelihood. Biometrics 53(3): 983–997.

Lee, S.Y., Lee, H.S., Lee, C.M., Ha, S.K., Park, H.M., Lee, S.M. and Mo, Y. 2023. Multi-environment trials and stability analysis for yield-related traits of commercial rice cultivars. Agriculture 13(2): 256.

Oakey, H., Verbyla, A., Pitchford, W., Cullis, B. and Kuchel, H. 2006. Joint modeling of additive and non-additive genetic line effects in single field trials. Theor. Appl. Genet. 113: 809–819.

Oakey, H., Verbyla, A.P., Cullis, B.R., Wei, X. and Pitchford, W.S. 2007. Joint modeling of additive and non-additive (genetic line) effects in multi-environment trials. Theor. Appl. Genet. 114(8): 1319–1332.

Piepho, H.P., Möhring, J., Schulz-Streeck, T. and Ogutu, J.O. 2012. A stage-wise approach for the analysis of multi-environment trials. Biometrical J. 54(6): 844–860.

Patterson, H.D. and Thompson, R. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58(3): 545–554.

Piepho, H.P. 1997. Analyzing genotype-environment data by mixed models with multiplicative terms. Biometrics 53(2): 761–766.

Renard, D., Mahaut, L. and Noack, F. 2023. Crop diversity buffers the impact of droughts and high temperatures on food production. Environ. Res. Lett. 18(4): 045002.

Smith, A., Cullis, B. and Gilmour, A. 2001. The analysis of crop variety evaluation data in Australia. Aust. New Zeal. J. Stat. 43: 129–145.

Smith, A., Cullis, B. and Thompson, R. 2001. Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. Biometrics 57: 1138–1147.

Smith, A., Cullis, B. and Thompson, R. 2005. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. J. Agric. Sci. 143(6): 449–462.

Smith, A.B. 1999. Multiplicative mixed models for the analysis of multi-environment trial data. Doctoral dissertation, University of Adelaide.

Smith, A.B., Borg, L.M., Gogel, B.J. and Cullis, B.R. 2019. Estimation of factor analytic mixed models for the analysis of multi-treatment multi-environment trials.

Tesfaye, K., Alemu, T., Argaw, T., de Villiers, S. and Assefa, E. 2023. Evaluation of finger millet (Eleusine coracana (L.) Gaertn.) in multi-environment trials using enhanced statistical models. Plos One 18(2): e0277499.

Verbyla, A. 2023. On two-stage analysis of multi-environment trials. Euphytica 219(11): 121.

Zenda, T., Liu, S., Dong, A. and Duan, H. 2021. Advances in cereal crop genomics for resilience under climate change. Life 11(6): 502.

Zhang, W., Hu, J., Yang, Y. and Lin, Y. 2020. One compound approach combining factor-analytic model with AMMI and GGE biplot to improve multi-environment trials analysis. J. For. Res. 31(1): 123–130.

Zsögön, A., Peres, L.E., Xiao, Y., Yan, J. and Fernie, A.R. 2022. Enhancing crop diversity for food security in the face of climate uncertainty. The Plant J. 109(2): 402–414.