Accuracy of prediction from multi-environment trials to new locations using pedigree information and environmental covariates

Diriba Tadese Gudata (diriba.tadese@uni-hohenheim.de)
Universitat Hohenheim https://orcid.org/0000-0002-9400-4621

Hans-Peter Piepho
University of Hohenheim: Universitat Hohenheim

Jens Hartung
University of Hohenheim: Universitat Hohenheim

Research Article

Keywords:

Posted Date: December 20th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3760192/v1

Accuracy of prediction from multi-environment trials to new locations using pedigree information and environmental covariates

Diriba Tadese*, Hans-Peter Piepho and Jens Hartung

Diriba Tadese, Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstraße 23, 70599 Stuttgart, Germany.

Email: diriba.tadese@uni-hohenheim.de

Hans-Peter Piepho, Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstraße 23, 70599 Stuttgart, Germany.

Email: hans-peter.piepho@uni-hohenheim.de

Jens Hartung, Biostatistics Unit, Institute of Crop Science, University of Hohenheim, Fruwirthstraße 23, 70599 Stuttgart, Germany.

Email: jens.hartung@uni-hohenheim.de

Author contributions: **DT**: Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; writing—original draft. **HPP**: Methodology; resources; software; reading; editing; supervision and validation of the manuscript. **JH**: Methodology; resources; software; analysis; editing; reading and validation of the manuscript.

Conflict of interest: The authors declare that there is no conflict of interest.

Key messages

We investigate a method of extracting and fitting synthetic covariates and pedigree information in multi-location trials data analysis to predict genotype performances in untested locations

Abstract Plant breeding trials are usually conducted across multiple testing locations to predict genotype performances in the targeted population of environments. The predictive accuracy can be increased by the use of adequate statistical models. We compared models with and without synthetic covariates (SC) and pedigree information under the identity, the diagonal and the factor-analytic variance-covariance structures of the genotype-by-location interactions. The model comparison was made to evaluate predictive accuracy of different models in predicting genotype performances in untested locations using the mean squared error of predicted differences (MSEPD) and the Spearman rank correlation between predicted and adjusted means. A multi-environmental trial (MET) dataset evaluated for yield performance in the dry low-land sorghum (*Sorghum bicolor* (*L.*) *Moench*) breeding program of Ethiopia was used. For validating our models, we followed a leave-one-location-out cross-validation strategy. A total of 65 environmental covariates (ECs) obtained from the sorghum test locations were considered. From the actual ECs, SC were first extracted using multivariate partial least squared analysis. Then, the model was fitted accounting for pedigree information by linear mixed models. According to MSEPD, our results indicate that models accounting for SC improve prediction precision of genotype performances in the three of the variance-covariance structures compared to others without SC. The rank correlation was also higher for the model with the SC. When the SC was fitted, the rank correlation was 0.58 for the factor-analytic, 0.51 for the diagonal and 0.46 for the identity variance-covariance structure.

Introduction

In plant breeding, genotypic selection for a given target population of environments (TPE) involves testing of several genotypes across multiple environments and therefore across locations and / or years (Piepho, 1998). Different statistical approaches for multi-environment trial (MET) data analysis were proposed over time to increase prediction precision of genotype performances accounting for genotype-by-environment interaction (GEI) effects (Gilmour et al., 1997; Piepho, 1998; Smith & Cullis, 2018). The common statistical methods used for MET analysis rely on randomization-based models considering different variance-covariance structures for GEI. The simplest variance-covariance structure assumes an identity matrix for the GEI effects, multiplied by a constant variance, implying independence between environments where the genotype main effect is also considered in the model. The identity variance-covariance structure can be extended to assume different variances at each environment and / or dependence between environments like the diagonal and factor-analytic variance-covariance structures of GEI (Gollob, 1968; Yates & Cochran, 1938). The classical approach of modeling the MET during data analysis is based on the fixed effects model (van Eeuwijk, 1992; Vargas, 1999). Later on, the multiplicative fixed effect model has been extended to its random-effects equivalent related to factor-analytic variance-covariance structures in the context of linear mixed model (LMM) (Piepho, 1997, 1998; Smith et al., 2001, 2005).

In plant breeding, field trials are usually conducted at a limited number of locations in the TPE and breeders use statistical methods allowing borrowing information through correlation between genotypes to predict genotypes performances using LMM (Li et al., 2021). In self-pollinated plants, sister lines are correlated through genetic kinship and the covariance of their breeding values are equal to the additive genetic covariance among the individual lines (Crossa et al., 2006). Therefore, breeders have used methods to incorporate pedigree information and/or marker data in an LMM for predicting breeding values (Buntaran et al., 2022; Crossa et al., 2010, Henderson, 1991, Mrode, 2005). In the standard LMM, BLUP of breeding values of random genotypes allows borrowing of information among relatives through the coefficient of parentage. Hence, closely related genotypes tend to contribute more to an estimated breeding value than less related lines (Jarquín et al., 2014). Hence, considering the kinship matrix is expected to improve prediction precision of genotypes.

Recent work on MET data analysis has focused on incorporating environmental covariates (ECs) in predicting genotype performances (Li et al., 2021, 2022; Piepho, 2022; Piepho & Blancon, 2023). In most cases, the main ECs include weather data (for example, rainfall and temperature) and soil information (for example, soil texture, pH, total nitrogen, organic carbon content). The ECs can be fitted as regressor variables for the main effects of locations and for genotypes-by-location interactions effects (Li et al., 2021, 2022; Piepho, 2022; Resende et al., 2021). Furthermore, breeders may be interested in genotypic prediction for new locations, where the trials were yet not conducted using ECs. Even though the test locations are expected to be representative of the TPE, it can be hard to find a perfect match for new locations among the tested locations. The common ECs include weather data and soil information like soil texture, organic carbon content, nitrogen content, etc. measured at different soil depths. Often, a large number of ECs is available and regressing the main effect of locations and GEI on several ECs using multiple regression, also known as factorial regression (Denis, 1988), may not be practical (Buntaran et al., 2021). One solution for such challenges could be extracting a smaller number of synthetic covariates (SCs) that represent the actual ECs, and then fitting the model using extracted SCs (Piepho, 2022; Piepho and Blancon, 2023). Our focus here is to investigate and compare different modeling strategies that provide precise genotypic predictions for untested locations through pedigree information and a large number of ECs. In order to mimic prediction scenarios in new locations, we followed a leave-one-location-out cross-validation mechanism.

In this study, we propose a modeling strategy for predicting genotype performances in untested locations using MET and evaluate their predictive ability. The general objective was to compare the predictive accuracy of models using MET data analysis (i) without pedigree and ECs, (ii) with pedigree or/and ECs, under three different variance-covariance structure of the GEI.

Material and Methods

Data source

We used sorghum (*Sorghum bicolor* (*L.*) *Moench* data from the Melkasa Agricultural Research Center (MARC), which is located near Adama City, southeast of Addis Ababa. MARC is responsible for the national sorghum breeding program in Ethiopia under the coordination of Ethiopian Institutes of Agricultural Research (EIAR). In this study, we used six MET data of the year 2019. These trials represent a dry low-land sorghum breeding program of Ethiopia and yield performance of genotypes measured in kilogram per hectare was considered. The field trials were laid out as resolvable row-column designs. At each location, the plot arrangement was 25 rows by four columns per replicate. 100 genotypes including two checks ('Melkam' and 'Argiti') were replicated twice. All genotypes have pedigree information except for one check ('Melkam'). Trials description and the environmental parameters used in this paper were described in Table 1 and Table 2 respectively. The environmental parameters include soil information taken at different soil layers and weather data. For each location, we obtained 65 ECs of weather station data and soil information.

Table 1. Description of the dry low-land sorghum breeding locations

Locations	Longitude	Latitude	Altitude (m.a.s.l)	Minimum T (C°)	Maximum T(C°)	Rain fall (mm)
Erer	42˚15'E	9˚10'N	1323	17	37	778
Kobo	39˚38'E	12˚09'N	1495	14.8	32	678
Mehoni	39˚68'E	12˚51'N	1777	12.81	23.24	539
Mieso	39˚21'E	8˚30'N	1317	16	31	571
Shiraro	39˚9'E	14˚6'N	1034	20.4	34	615
Shewarobit	39˚93'E	10˚35'N	1265	17.7	33	713

Source: National meteorology, m.a.s.l =meters above sea level, T = temperature

Table 2. Description of the soil information at different soil layers and weather data taken from each of the test locations

Description of the covariates	Layers
Soil organic carbon concentration	6
Soil pH	6
Coarse fragments volumetric	6
Soil texture fraction sand	6
Soil texture fraction silt	6
Soil texture fraction clay	6
Cation Exchange Capacity	6
Total nitrogen	1
Aluminium concentration	2
Exchangeable acidity	3
Exchangeable Calcium	2
Exchangeable Magnesium	2

Exchangeable Sodium	2
Sum of exchangeable bases	3
Electrical conductivity	6
Temperature	-
Rainfall	-

Source: ISRIC and EthioSIS map

Statistical methods

Our modeling strategy follows a stage-wise approach, where information from Stage I will be forwarded to Stage II (Piepho et al., 2012a). In Stage I, individual locations were subjected to an LMM analysis producing adjusted genotype means and associated variance-covariance matrix of genotype means. Adjusted means and their precision measure from Stage I were forwarded to the second-stage analysis, where the combined analysis was conducted across locations.

Stage I analysis

The LMM used for each location in Stage I (Diriba and Piepho, 2023) can be expressed as:

$$y_{ijkl} = \mu + a_i + h_j + r_{jk} + c_{jl} + e_{ijkl} \qquad [1]$$

where y_{ijkl} is the observed yield of the *i*-th genotype in the *k*-th row and *l*-th column within replicate *j*, μ is the intercept, a_i is the fixed effect of the *i*-th genotype, $h_j \sim N(0, \sigma_h^2)$ is the random effect of the *j*-th replicate, $r_{jk} \sim N(0, \sigma_r^2)$ is the random effect of the *k*-th row nested in the *j*-th replicate, $c_{jl} \sim N(0, \sigma_c^2)$ is the random effect of the *l*-th column nested in the *j*-th replicate and $e_{ijkl} \sim N(0, \sigma_e^2)$ is the error associated with y_{ijkl}. Using Eq. [1], we estimated genotype means ($\bar{y}_{im}$) for the *i*-th (*i*=1, 2, …, *I*+1) genotype at each of *M* locations (*m*=1, 2, …, *M*). To forward information of Stage I analysis, genotype means of the *m*-th location sorted by genotype were put in to a vector $\boldsymbol{y}_m = (\bar{y}_{1m}, \bar{y}_{2m}, \ldots, \bar{y}_{Im})$. Note that only *I* genotype means per location were forwarded to Stage II as standard check 'Melkam' was dropped after estimating means in Stage I due to missing pedigree information. Furthermore, we defined the vector $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M)$ as a vector of genotype means across locations. To forward precision of genotype means, weights were calculated from the estimated variance-covariance structure of genotype means $\boldsymbol{y}_m$ denoted as $\boldsymbol{\Omega}_m$. The inverse $\boldsymbol{\Omega}_m^{-1}$ was approximated by a diagonal matrix formed by the diagonal elements of $\boldsymbol{\Omega}_m^{-1}$ (Damesa et al., 2017, Smith et al., 2001). Hence, these diagonal elements were used as weights in the second stage. The matrix

with the inverses of the weights down the diagonal approximates $\boldsymbol{\Omega}_m$ and will be denoted as $\boldsymbol{\Omega}_m^{(d)}$.

Stage II analysis

In Stage II, we considered genotype as the random factor. This allows to include pedigree information. In addition, location was considered as a random factor, too, so that prediction for new locations is possible. Therefore, with vector $\boldsymbol{a} = (a_1, a_2, \ldots, a_I)$ of genotype effects, vector $\boldsymbol{l} = (l_1, l_2, \ldots, l_M)$ of location effects and vector $\boldsymbol{s} = (s_{11}, s_{12}, \ldots, s_{1M}, s_{21}, s_{22} \ldots, s_{2M}, \ldots, s_{IM})$ of genotype-by-location interaction effects, the basic model considered in Stage II analysis can be written as

$$\boldsymbol{y} = \mathbf{1}_{IM}\mu + \boldsymbol{Z}_1\boldsymbol{a} + \boldsymbol{Z}_2\boldsymbol{l} + \boldsymbol{Z}_3\boldsymbol{s} + \boldsymbol{f} \qquad [2]$$

where $\boldsymbol{y}$ is a vector of estimated genotype-by-location means from Stage I, $\mathbf{1}_{IM}$ is a vectors of ones, μ is the intercept, $\boldsymbol{a} \sim N(\mathbf{0}, \mathbf{I}_I\sigma_a^2)$, $\boldsymbol{l} \sim N(\mathbf{0}, \mathbf{I}_M\sigma_l^2)$ and $\boldsymbol{s} \sim N(\mathbf{0}, \mathbf{I}_{IM}\sigma_s^2)$ are vectors of genotype, location and genotype-by-location interaction parameters where $\mathbf{I}$ is identity matrix with the subscript represent the dimension, $\boldsymbol{Z}_1, \boldsymbol{Z}_2$ and $\boldsymbol{Z}_3$ are the corresponding design matrices, respectively and $\boldsymbol{f}$ is a vector of error containing sub-vectors $\boldsymbol{f}_m$, with var($\boldsymbol{f}_m$)$= \boldsymbol{\Omega}_m^{(d)}$. The total variance-covariance matrix of $\boldsymbol{f}$ is given by

$$var(\boldsymbol{f}) = \begin{pmatrix} \boldsymbol{\Omega}_1^{(d)} & \mathbf{0} \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_2^{(d)} & \vdots \\ \vdots & \ddots & \\ \mathbf{0} & \cdots & \boldsymbol{\Omega}_M^{(d)} \end{pmatrix} = \oplus_{m=1}^{M} \boldsymbol{\Omega}_m^{(d)} = \boldsymbol{\Omega}^{(d)}$$

Modelling covariance structures and pedigree information

The basic model assumes homogeneous variances and independence between genotype, location and genotype-by-location interactions effects, fitting identity matrices with constant variance in the corresponding variance-covariance structures. The baseline model was modified in two ways. First, the identity matrix for the genotype-by-location interactions was modified to allow for diagonal or FA variance-covariance structures. Second, the independence between genotypes was modified to allow for a kinship matrix. The kinship matrix, multiplied by the genetic variance, contains the genetic variances on the diagonal while the off-diagonal elements are the genetic covariance between pairs of genotypes. Therefore, with $\mathbf{A}$ matrix representing the $I \times I$ numerator relationship for I genotypes, the variance-covariance

structures in Eq. [2] can be redefined as $\boldsymbol{a} \sim N(\mathbf{0}, \boldsymbol{\Gamma}\sigma_a^2)$, $\boldsymbol{l} \sim N(\mathbf{0}, \mathbf{I}_M\sigma_l^2)$ and $\boldsymbol{s} \sim N(\mathbf{0}, \boldsymbol{\Gamma} \otimes \boldsymbol{\Pi})$, where $\boldsymbol{\Gamma} = \mathbf{I}$ or $\mathbf{A}$, $\boldsymbol{\Pi} = \mathbf{I}\sigma_s^2$, $\boldsymbol{\Pi} = \boldsymbol{\Phi}$ (diagonal) or $\boldsymbol{\Pi} = \boldsymbol{\Sigma}$ (FA) with

$$\boldsymbol{\Phi} = \begin{pmatrix} \sigma_{s1}^2 & 0 \cdots & 0 \\ 0 & \sigma_{s2}^2 & \vdots \\ \vdots & \ddots & \\ 0 & \cdots & \sigma_{sM}^2 \end{pmatrix} \qquad [3]$$

and $\otimes$ denotes Kronecker product (Burgueño et al, 2012).

For the FA structure, multiplicative terms for approximating the variance-covariance matrix of the genotypes-by-location interaction effects (Piepho, 1997, 1998; Smith et al., 2001; Crossa et al., 2004, 2006; Burgueño et al., 2011, 2012) were used. In this case, the variance-covariance structure for the genotype-by-location interactions can be expressed as $\mathrm{FA}(K) = \boldsymbol{\Sigma} = (\wedge\wedge' + \boldsymbol{\Phi})$, where K is the number of FA components, $\wedge$ is the $M \times K$ matrix in which the k-th column contains location loadings for k-th latent factor, and $\boldsymbol{\Phi}$ is a $M \times M$ diagonal matrix (Burgueño et al., 2012). When modeling the variance-covariance using FA structures, it is possible to consider more than one component, however, as the number of components gets larger there can be numerical problems of fitting the model (Studnicki et al., 2017). In order to decide on the number of components, one can use information criteria like the Akaike Information Criterion (AIC) (Wolfinger, 1993). In our case, we tried FA(1), FA(2) and FA(3) orders and selected the FA(2) structure based on AIC.

Modeling environmental covariates

Most of the soil data were taken from EthioSIS map while some were obtained from ISRIC (FAO, 2020). We extracted a SCs for each location that represent the actual ECs following a method proposed by Piepho and Blancon (2023). In this method, different genotypes were regarded as variates and a multivariate partial least square (PLS) analysis was applied so that a single linear combination of covariate for each location could be obtained. This technique extracts a set of dependent variables from a set of independent variables where the extraction is achieved through a set of orthogonal factors named as latent variables (Krishnan, 2010). Extraction of SC from the actual ECs can easily be generated using the '*mvr'* function in R package. Before running '*mvr*', the ECs were standardized to mean zero and unit variance. Standardized covariates were then fitted against genotype-environment means to get SC. Then the extension of the model of Eq. [2] when considering the first SC can be written as follows

$$\boldsymbol{y} = \mathbf{1}_{IM}\mu + \boldsymbol{t}\beta + \boldsymbol{Z}_1\boldsymbol{a} + \boldsymbol{Z}_2\boldsymbol{l} + \boldsymbol{Z}_3\boldsymbol{s} + \boldsymbol{Z}_4\boldsymbol{b} + \boldsymbol{f} \qquad [4]$$

where β is the slope of $\boldsymbol{t}$, where $\boldsymbol{t}$ is a vector of SC of different environments, $\boldsymbol{a}$ is a vector of random intercepts and $\boldsymbol{b} = (b_1, b_2, \ldots, b_I)$ is a vector of random slope effects with $\boldsymbol{Z}_4$ is the corresponding design matrix. When allowing for an unstructured variance-covariance between random coefficients $\boldsymbol{a}$ for the intercept and $\boldsymbol{b}$ for the slope of each genotype, $\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} \sim N(\mathbf{0}, \boldsymbol{G} \otimes \boldsymbol{\Gamma})$, where $\boldsymbol{G} = \begin{bmatrix}\sigma_a^2 & \sigma_{ab}\\ \sigma_{ab} & \sigma_b^2\end{bmatrix}$ and $\boldsymbol{\Gamma}$ is either $\boldsymbol{I}$ or $\mathbf{A}$. An extension to consider several SCs is straightforward, using the intercept $\boldsymbol{a}$ and slopes $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots$ for the SCs.

Following Eq. [2] to Eq. [4], a series of 12 models in stage II analysis were fitted and compared for prediction precision of genotype performances in untested locations as summarized in Table 3 considering the first SC.

Table 3. Summary of the 12 models used to predict genotypes performance in the new locations using the first SC and pedigree information

Models	Fixed effects	Random effects	Variance-covariance matrix of		
			$\boldsymbol{a}$	$\boldsymbol{l}$	$\boldsymbol{s}$
M1	μ	$\boldsymbol{a}, \boldsymbol{l}, \boldsymbol{s}$	$\mathbf{I}_I\sigma_a^2$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{I}_I \otimes \mathbf{I}_M\sigma_s^2$
M2	μ	$\boldsymbol{a}, \boldsymbol{l}, \boldsymbol{s}$	$\mathbf{A}\sigma_a^2$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{A} \otimes \mathbf{I}_M\sigma_s^2$
M3	μ, β	$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{l}, \boldsymbol{s}$	$\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} \sim N(\mathbf{0}, \mathbf{I}_I \otimes \boldsymbol{G})$ $\boldsymbol{G} = \begin{bmatrix}\sigma_a^2 & \sigma_{ab}\\ \sigma_{ab} & \sigma_b^2\end{bmatrix}$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{I}_I \otimes \mathbf{I}_M\sigma_s^2$
M4	$\mu,\ \beta$	$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{l}, \boldsymbol{s}$	$\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} \sim N(\mathbf{0}, \mathbf{A} \otimes \boldsymbol{G})$ $\boldsymbol{G} = \begin{bmatrix}\sigma_a^2 & \sigma_{ab}\\ \sigma_{ab} & \sigma_b^2\end{bmatrix}$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{A} \otimes \mathbf{I}_M\sigma_s^2$
M5	μ	$\boldsymbol{a}, \boldsymbol{l}, \boldsymbol{s}$	$\mathbf{I}_I\sigma_a^2$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{I}_I \otimes \boldsymbol{\Phi}$
M6	μ	$\boldsymbol{a}, \boldsymbol{l}, \boldsymbol{s}$	$\mathbf{A}\sigma_a^2$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{A} \otimes \boldsymbol{\Phi}$
M7	$\mu,\ \beta$	$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{l}, \boldsymbol{s}$	$\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} \sim N(\mathbf{0}, \mathbf{I}_I \otimes \boldsymbol{G})$ $\boldsymbol{G} = \begin{bmatrix}\sigma_a^2 & \sigma_{ab}\\ \sigma_{ab} & \sigma_b^2\end{bmatrix}$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{I}_I \otimes \boldsymbol{\Phi}$
M8	μ, β	$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{l}, \boldsymbol{s}$	$\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} \sim N(\mathbf{0}, \mathbf{A} \otimes \boldsymbol{G})$ $\boldsymbol{G} = \begin{bmatrix}\sigma_a^2 & \sigma_{ab}\\ \sigma_{ab} & \sigma_b^2\end{bmatrix}$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{A} \otimes \boldsymbol{\Phi}$
M9	μ	$\boldsymbol{a}, \boldsymbol{l}, \boldsymbol{s}$	$\mathbf{I}_I\sigma_a^2$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{I}_I \otimes \boldsymbol{\Sigma}$
M10	μ	$\boldsymbol{a}, \boldsymbol{l}, \boldsymbol{s}$	$\mathbf{A}\sigma_a^2$	$\mathbf{I}_M\sigma_l^2$	$\mathbf{A} \otimes \boldsymbol{\Sigma}$

M11	μ, β	$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{l}, \boldsymbol{s}$	$\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} \sim N(\mathbf{0}, \mathbf{I}_I \otimes \boldsymbol{G})$ $\boldsymbol{G} = \begin{bmatrix}\sigma_a^2 & \sigma_{ab}\\ \sigma_{ab} & \sigma_b^2\end{bmatrix}$	$\mathbf{I}_M \sigma_l^2$	$\mathbf{I}_I \otimes \boldsymbol{\Sigma}$
M12	μ, β	$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{l}, \boldsymbol{s}$	$\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{bmatrix} \sim N(\mathbf{0}, \mathbf{A} \otimes \boldsymbol{G})$ $\boldsymbol{G} = \begin{bmatrix}\sigma_a^2 & \sigma_{ab}\\ \sigma_{ab} & \sigma_b^2\end{bmatrix}$	$\mathbf{I}_M \sigma_l^2$	$\mathbf{A} \otimes \boldsymbol{\Sigma}$

The matrix **I** represent identity matrix, with the subscript denoting the dimension, **Φ** represent diagonal matrix for the genotype-by-location interactions, ⊗ represent the Kronecker product, **Σ** is the factor-analytic variance-covariance structure, **A** is the kinship matrix, $\boldsymbol{a}$ and $\boldsymbol{b}$ are vectors of random coefficients for genotypes, $\boldsymbol{l}$ is a vector for location main effects, $\boldsymbol{s}$ is a vector for genotype-by-location interactions, β is the slope for the regression on $\boldsymbol{t}$, where $\boldsymbol{t}$ is the vector of the synthetic covariate, *I* is the number of genotypes, *M* is the number of locations.

In Table 3, models M1-M4 were fitted considering the identity variance-covariance structure for the genotype-by-location interactions, M5-M8 were the diagonal and M9-M12 were the FA structures. Within each type of variance-covariance structure, the first model was fitted without SC and pedigree information, the second model with pedigree information, the third model with SC, and the fourth model with SC plus pedigree information.

Model evaluation

To mimic prediction scenarios in untested locations, a leave-one-location-out CV algorism was implemented. According to this CV, we dropped all genotype means of a given location at a time and assigned means from this location as a validation set whereas the genotype means from the remaining locations served as a training set. Prediction of genotype means for the dropped location was then made using the training data set. For the models with the SC, the SC from the dropped locations was also considered. The model predictions for the genotypes-by-location means was expected to benefit from borrowing information like between lines in the same location, between genotypes across location and through correlated locations (Burgueño et al., 2012; Buntaran et al., 2021). The model accuracy in predicting genotype performances for untested locations was then evaluated and compared.

The model comparison was made using Spearman rank correlation and mean squared errors of prediction differences (MSEPD). The correlation between estimated genotype-by-location means from Stage I and predicted values from Stage II analysis ignoring the corresponding location was computed for each location. Then, the correlations obtained with specific model in all locations were averaged to obtain average correlation of the model. The correlation between estimated and predicted value tells the degree of consistence and the sign of

correlation in genotype ranking. Strong correlation is an indication of strong consistence in genotype ranking (Roostaei et al., 2014). In addition, predictive accuracy of the model was also assessed using MSEPD for genotypes (Buntaran et al., 2021; Studnicki et al., 2017). As proposed by Piepho (1998), the MSEPD focuses on comparing the difference between the observed difference between two genotypes in a given location $(\bar{y}_{im} - \bar{y}_{i\prime m})$ and the corresponding predicted difference say $(z_{im} - z_{i\prime m})$. The smallest MSEPD for the differences is an indication of the best model. The MSEPD is computed as

$$MSEPD = \frac{\sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{i'\neq i}^{I}[(\bar{y}_{im} - \bar{y}_{i'm}) - (z_{im} - z_{i'm})]^2}{MI(I-1)} \quad [5]$$

where *M* is the number of locations and *I* is the number of genotypes as defined before. In our case, z_{im} and $z_{i\prime m}$ are predicted values obtained from Stage II using data from all locations except of location *m*.

We used SAS software to estimate genotype means in Stage I and ASReml-R 4.1.0.130 (Butler et al. 2017) for fitting our models in Stage II.

Results

Table 4 indicate the MSEPD and average rank correlations of different models. For models without the SC, fitting pedigree data indicate some improvement in the diagonal and FA variance-covariance structures of the genotype-by-location interactions when compared to the model without pedigree information. The rank correlation resulted in small difference between models with and without pedigree information in the three of the variance-covariance structures. When the first SC is fitted, the diagonal variance-covariance structure resulted in the minimum MSEPD (0.71342 t/ha) followed by the FA (0.90766 t/ha) and the identity (0.95536 t/ha) variance-covariance structures respectively. Based on the rank correlations, the comparison indicates that the FA variance-covariance structure comes first (0.5837) followed by the diagonal (0.509) and the identity (0.4638) variance-covariance structures. When the two SCs are considered, there is no gain in the fitted model compared to fitting only one SC for the identity and diagonal variance-covariance structures but small gain with the FA variance-covariance structure according to MSEPD and average rank correlation. With the FA variance-covariance, the MSEPD changed from 0.90766 t/ha when fitting one SC to 0.879093 t/ha for fitting two SC whereas the average rank correlation changed from 0.58367 to 0.588384.

Table 4. The mean squared error of predicted differences (MSEPD) and average rank correlation between adjusted means and predicted values for each model without SC, with one and two SCs in predicting genotype yield means in new locations

Models	Without SC		With one SC		With two SC	
	MSEPD (t/ha)	Correlation	MSEPD (t/ha)	Correlation	MSEPD (t/ha)	Correlation
M1	1.02248	0.42829	-	-	-	-
M2	1.02311	0.4204	-	-	-	-
M3	-	-	0.95536	0.46375	1.032472	0.425681
M4	-	-	1.02599	0.42105	1.028033	0.420002
M5	1.0464	0.42827	-	-	-	-
M6	1.04629	0.42381	-	-	-	-
M7	-	-	0.71342	0.50913	0.997379	0.47778
M8	-	-	1.0321	0.42548	1.035721	0.423737
M9	1.09375	0.43914	-	-	-	-
M10	1.03831	0.4308	-	-	-	-
M11	-	-	0.90766	0.58367	0.879093	0.588384
M12	-	-	1.00929	0.43857	1.015214	0.437673

Discussion

Evaluation of genotype performances in the TPE is one of the core focus of plant breeders in which trials are conducted at multiple locations. One advantage of METs is that the possibility for allowing borrowing of information among trials during the data analysis (Crossa et al., 2006; Piepho et al., 2008). The MET data analysis can follow either one-stage or stage-wise analysis for predicting genotype performances in the TPEs. In the stage-wise analysis, information from Stage I mainly estimated means and respective variance-covariance components are saved and used during Stage II analysis (Piepho et al., 2012a). According to recent studies, the stage-wise analysis has many advantages over one-stage analysis (Piepho et al., 2012a; Damesa et al., 2017). Three of these advantages are that it is computationally less demanding, that combining trials with different design background is straightforward, and that there are relatively less convergence problems. In our case, advantage of the stage-wise analysis was less computational time and convergence problems. During Stage II analysis, different variance-covariance structures of the genotype-by-location interactions can easily be considered with less computational demand to allows for borrowing information through correlated locations. The FA variance-covariance structure is one of the commonly used in plant breeding (Studnicki et al, 2017).

One of the current advancements in the plant breeding trials to improve prediction accuracy is that linking the ECs in the MET during the data analysis. The main challenge when using the ECs for prediction purpose is that how to deal with the large number of ECs to fit in the MET. For this reason, several studies recommended the use of PLS analysis (Vargas et al., 1998; Vargas et al.,1999; Crossa et al., 1999; Montesinos-López et al., 2022) to approach the challenge. Practically it is possible to considers the actual ECs individually using factorial regression (Denis, 1980) with lower number of ECs, however, with large number of ECs, it is quite difficult to do so (Buntaran et al., 2021; Piepho, 2022; Costa-Neto et al., 2023; Piepho and Blancon, 2023). In our case, we extracted smaller number of SCs from the actual ECs using the multivariate PLS technique. This technique considers different genotypes as variates to obtain a single linear combination of covariates for each location (Piepho and Blancon, 2023).

This study indicated a gain in prediction accuracy of fitting SC in MET compared to the model without SC when predicting genotype means for untested locations which is also confirmed in other studies (Heslot et al., 2014; Buntaran et al., 2021; Montesinos-López et al., 2022). Jarquín et al. (2014) considered ECs as a random regression that each line may face to predict genotype performance in an incomplete trial. In our case, we extracted SC from the actual ECs to fit the model using LMM and make predictions. This method is more advantageous compared to fitting the actual ECs since a large number of ECs can be considered through extracting smaller number of SC and the prediction can easily be made for the new locations. Apart from extracting SC through multivariate PLS, different alternative methods of extracting SC were also illustrated in Piepho and Blancon (2023). Here, we favored PLS because it can deal with a larger number of EC.

More than one SC can be fitted in the MET to make prediction, however, the best way is to start with the first SC and check for if there is a gain in fitting more than one SC. The first SC can capture less than 50%, the first SC capture larger parts of variance compared to all other SC. Our results showed larger MSEPD and rank correlations with fitting two SC compared to only fitting the first SC except for FA as indicated in Table 4. We think this is primarily due to the small number of environments. Fitting two or more SC in the MET is straight forward using the random coefficients of genotypes for the SCs. When fitting the SC, the converge problems may arise which we also faced and approached the problem through setting different initial values of the variance parameters.

The prediction accuracy of pedigree-based models was evaluated, resulting in less gain compared to the one without pedigree and SC. According to recent studies, marker-based models result in more accurate prediction compared to pedigree information (Crossa et al., 2010; Burgueño et al., 2012). Burgueño et al. (2012) compared the prediction accuracy of marker and pedigree-based models. The aim of the comparison was in predicting genotype performances in untested locations using multi-environment mixed models and concluded that marker-based model gives more accurate prediction than pedigree-based model. The basic ideas behind fitting either pedigree or marker-based models is that, the prediction accuracy is expected to be benefit from correlated information among relatives. In our case we do not have a marker data at hand. The results of fitting pedigree information plus SC is encouraging specially when the diagonal and FA variance-covariance structures for the genotype-by-location interactions were considered.

The drop-one-location-out CV mechanism has a great advantage for the model validation when the objective of the prediction is conducted for the new locations. This type of CV mimics the real scenarios on the ground (Montesinos-López et al., 2022). Dropping observations for the model validation was originally used to assess the discrepancy between observed differences and predicted values within environments where observation from some blocks were assigned to a validation set and the others used for modelling (Piepho 1998). Furthermore, the drop-one-location-out CV strategy was also used to validate projecting result of MET to new locations using ECs (Buntaran et al., 2021).

When different models are used in predicting genotype performances, it is necessary to evaluate the prediction accuracy of the candidate models. The prediction precision of different models can be accessed by using MSEPD as proposed by Piepho (1998) since breeders are more interested in the difference between genotypes means rather than the exact value of specific genotype mean. The model that resulted in the minimum MSEPD is considered as the best model. In our case, the smallest MSEPD was obtained with fitting the SC compared other models which indicate the importance of considering ECs for predicting genotype means in the new locations.

Furthermore, the rank correlation between the adjusted means and predicted values of genotype can also be used to evaluate prediction accuracy of different models. The rank correlation ranges between -1 and +1, where -1 is an indication of fully opposite ranking and +1 indicates fully identical ranking between adjusted means and predicted values (Roostaei et al., 2014).

Differences in consistence level among genotype ranking under different models is helpful to select the best model. In all models, the rank correlation indicated a positive correlation between estimated genotype means and predicted values while the strong correlation was obtained for the model with the SC. The strong correlation gained with the SC is an indication of consistence in genotype ranking can be improved through fitting SC.

Our conclusion in this paper was based on one mega-environment. The proposed method works equally for more than one mega-environment. With different mega-environments, locations are clustered to form stratums or zones where each stratum consisting of several environments (Buntaran et al., 2021). The extension of our proposed methods to consider mega-environments is straight-forward, requiring inclusion of a fixed zone effect and its interactions in the Stage II analysis. Based on our findings and other recent studies, we recommend that the use of SC increase prediction accuracy in predicting genotype performances for untested locations given that an appropriate statistical model is used (Costa-Neto et al., 2023; Heslot et al., 2014).

Conclusion

From this study we conclude that fitting SC can increase prediction accuracy in new locations while the model with both SC plus pedigree information is considered as a promising candidate. The fitted SC perform better in the diagonal and the FA variance-covariance structures of genotype-by-location interactions than the identity variance-covariance structure.

Reference

Buntaran, H., Bernal-Vasquez, A. M., Gordillo, A., Sahr, M., Wimmer, V., & Piepho, H. P. (2022). Assessing the response to genomic selection by simulation. *Theoretical and Applied Genetics*, *135*(8), 2891–2905. https://doi.org/10.1007/s00122-022-04157-1

Buntaran, H., Forkman, J., & Piepho, H. P. (2021). Projecting results of zoned multi-environment trials to new locations using environmental covariates with random coefficient models: accuracy and precision. *Theoretical and Applied Genetics*, *134*(5), 1513–1530. https://doi.org/10.1007/s00122-021-03786-2

Burgueño, J., Crossa, J., Miguel Cotes, San Vicente, F., and Das, B. (2011). Prediction assessment of linear mixed models for multienvironment trials. Crop Sci. 51:944–954. doi:10.2135/ cropsci2010.07.0403

Burgueño, J., de los Campos, G., Weigel, K., & Crossa, J. (2012). Genomic prediction of breeding values when modelling genotype × environment interaction using pedigree and dense molecular markers. *Crop Science*, *52*(2), 707–719. https://doi.org/10.2135/cropsci2011.06.0299

Butler, D.G., Cullis, B., Gilmour, A., Gogel, B.J., Thompson, R. (2017). ASReml-R reference manual, version 4. University of Wollongong, Wollongong

Costa-Neto, G., Crespo-Herrera, L., Fradgley, N., Gardner, K., Alison R.B., Dreisigacker, S., Fritsche-Neto, R., Osval A., Montesinos-López, Crossa, J. (2023). Envirome-wide associations enhance multi-year genome-based prediction of historical wheat breeding data. https://doi.org/10.1093/g3journal/jkac313

Crossa, J., Burgueño, J., Cornelius, P. L., McLaren, G., Trethowan, R., & Krishnamachari, A. (2006). Modelling genotype × environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. *Crop Science*, *46*(4), 1722–1733. https://doi.org/10.2135/cropsci2005.11-0427

Crossa, J., De Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., Makumbi, D., Singh, R. P., Dreisigacker, S., Yan, J., Arief, V., Banziger, M., & Braun, H. J. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, *186*(2), 713–724. https://doi.org/10.1534/genetics.110.118521

Crossa, J., Vargas, M., van Eeuwijk, F. A., Jiang, C., Edmeades, G. O., and Hoisington, D. (1999). Interpreting genotype × environment interaction in tropical maize using linked molecular markers and environmental covariables. Theor. Appl. Genet. 99, 611–625. doi:10.1007/s001220051276

Crossa, J., Yang, R.C., and Cornelius, P.L. (2004). Studying crossover genotype × environment interaction using linear-bilinear models and mixed models. J. Agric. Biol. Environ. Stat. 9:362–380. doi:10.1198/108571104X4423

Damesa, T.M., Möhring, J., Worku, M., Piepho, H.P. (2017). One step at a time: stage-wise analysis of a series of experiments. Agron J 109:845–857. https://doi.org/10.2134/agronj2016.07.0395

Denis J.B. (1980). Analyse de régression factorielle. Biométrie Praximétrie 20, 1–34.

Denis, J.B. (1988). Two-way analysis using covariates. Statistics 19:123–132. https://doi.org/10.1080/02331888808802080

Diriba Tadese and Piepho, H.P. (2023). Spatial model selection and design evaluation in the Ethiopian sorghum breeding program. https://doi.org/10.1002/agj2.21450

FAO. (2020). Ten years of the Ethiopian aagricultural transformation aagency. An FAO evaluation of the aagency's impact on agricultural growth and poverty reduction. Rome. https://doi.org/10.4060/cb2422en

Fikret Isik, James Holland, and Christian Maltecca. (2017). Genetic data aanalysis for plant and animal breeding. Springer; 1st ed. 2017 edition (9 September 2017)

Gilmour, A. R., Cullis, B. R., & Verbyla, A. P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. In *Source: Journal of Agricultural, Biological, and Environmental Statistics* (Vol. 2, Issue 3). https://www.jstor.org/stable/1400446?seq=1&cid=pdf-

Gollob, H. F. (1968). A statistical model which combines features of factor analytic and analysis of variance techniques*. Psychometrika, 33(1)

Henderson, C. R. (1991). Contributions to predicting genetic merit. L. R. SCHAEFFER centre for genetic improvement of livestock university of Guelph Guelph

Heslot, N., Akdemir, D., Sorrells, M. E., & Jannink, J. L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theoretical and Applied Genetics*, *127*(2), 463–480. https://doi.org/10.1007/s00122-013-2231-5

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M., Burgueño, J., & de los Campos, G. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, *127*(3), 595–607. https://doi.org/10.1007/s00122-013-2243-1

Krishnan, A., Williams, L.J., McIntosh, A.R., Abdi, H. (2010). Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review: doi:10.1016/j.neuroimage.2010.07.034

Li, X., Guo, T., Bai, G., Zhang, Z., See, D., Marshall, J., Garland-Campbell, K. A., & Yu, J. (2022). Genetics-inspired data-driven approaches explain and predict crop performance fluctuations attributed to changing climatic conditions. In *Molecular Plant* (Vol. 15, Issue 2, pp. 203–206). Cell Press. https://doi.org/10.1016/j.molp.2022.01.001

Li, X., Guo, T., Wang, J., Bekele, W. A., Sukumaran, S., Vanous, A. E., McNellie, J. P., Cortes, L. T., Lopes, M. S., Lamkey, K. R., Westgate, M. E., McKay, J. K., Archontoulis, S. v., Reynolds, M. P., Tinker, N. A., Schnable, P. S., & Yu, J. (2021). An integrated framework reinstating the environmental dimension for GWAS and genomic selection in crops. *Molecular Plant*, *14*(6), 874–887. https://doi.org/10.1016/j.molp.2021.03.010

Montesinos-López OA, Montesinos-López A, Kismiantini, Roman-Gallardo A, Gardner K, Lillemo M, Fritsche-Neto R, Crossa J.(2022). Partial least squares enhances genomic prediction of new environments. Front Genet. 2022 Jul 8;13:920689. doi: 10.3389/fgene.2022.920689. PMID: 36313422; PMCID: PMC9608852.

Mrode, R.A. (2005). Linear models for the prediction of animal breeding values, 2nd edition. Scottish agricultural college Midlothian, UK.

Piepho, H.P. (2023). Extending Finlay-Wilkinson regression with environmental covariates. Plant breeding. **https://doi.org/10.1111/pbr.13130**

Piepho, H.P. (1998). Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. In *Theor Appl Genet* (Vol. 97). Springer-Verlag.

Piepho, H. P. (1998). Methods for comparing the yield stability of cropping systems - A review. In *Journal of Agronomy and Crop Science* (Vol. 180, Issue 4, pp. 193–213). Blackwell Publishing Ltd. https://doi.org/10.1111/j.1439-037X.1998.tb00526.x

Piepho, H. P. (2022). Prediction of and for new environments: What's your model? In *Molecular Plant* (Vol. 15, Issue 4, pp. 581–582). Cell Press. https://doi.org/10.1016/j.molp.2022.01.018

Piepho, H.P. (1997). Analyzing genotype-environment data by mixed models with multiplicative effects. Biometrics 53:761–766. doi:10.2307/2533976

Piepho, H. P., Möhring, J., Melchinger, A. E., & Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. In *Euphytica* (Vol. 161, Issues 1–2, pp. 209–228). https://doi.org/10.1007/s10681-007-9449-8

Piepho, H.P., J. Möhring, T. Schulz-Streeck, and J.O. Ogutu. (2012a). A stage-wise approach for analysis of multi-environment trials. Biometrics 54:844–860. doi:10.1002/bimj.201100219

Piepho, H.P., Ogutu, J.O. (2002). A simple mixed model for trend analysis in wildlife populations. J Agric Biol Environ Stat 7:350. https:// doi.org/10.1198/108571102366

Resende, R. T., Piepho, H. P., Rosa, G. J. M., Silva-Junior, O. B., e Silva, F. F., de Resende, M. D. v., & Grattapaglia, D. (2021). Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. *Theoretical and Applied Genetics*, *134*(1), 95–112. https://doi.org/10.1007/s00122-020-03684-z

Roostaei, M., Mohammadi, R., and Amri, A. (2014). Rank correlation among different statistical models in ranking of winter wheat genotypes. The Crop J. 2:154–163. doi:10.1016/j. cj.2014.02.002.

Russ Wolfinger, (1993). Covariance structure selection in general mixed models. SAS Institute, Inc., SAS. campus drive, cary, North Carolina 27513-2414, U.S.A. DOI: 10.1080/03610919308813143

Smith, A. B., & Cullis, B. R. (2018). Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica*, *214*(8). https://doi.org/10.1007/s10681-018-2220-5

Smith, A., Cullis, B.R., and Thompson, R. (2001). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. Biometrics 57:1138–1147. doi:10.1111/j.0006-341X.2001.01138.x

Smith, A. B., Cullis, B. R., & Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. In *Journal of Agricultural Science* (Vol. 143, Issue 6, pp. 449–462). https://doi.org/10.1017/S0021859605005587

Studnicki, M., Paderewski, J., Piepho, H.P., and Wójcik-Gront, E. (2017). Prediction accuracy and consistency in cultivar ranking for factor-analytic linear mixed models for winter wheat multienvironmental trials. doi: 10.2135/cropsci2017.01.0004

van Eeuwijk, F.A. (1992). Interpreting genotype-environment interaction using redundancy analysis. Theoretical and Applied Genetics, 85, 92–100.

Vargas, M., Crossa, J., Eeuwijk, F. A., Ramírez, M. E., and Sayre, K. (1999). Using partial least squares regression, factorial regression, and AMMI models for interpreting genotype × environment interaction. Crop Sci. 39 (4), 955–967. doi:10.2135/cropsci1999.0011183X003900040002x

Vargas, M., Crossa, J., Sayre, K., Reynolds, M., Ramírez, M.E., & Talbot, M. (1998). Interpreting genotype x environment interaction using partial least squares regression. Crop Science, 38, 679–689.

Yates, F., & Cochran, W. G. (1938). The analysis of groups of experiments. *The Journal of Agricultural Science*, *28*(4), 556–580. https://doi.org/10.1017/S0021859600050978

Acknowledgement: We acknowledge the Ethiopian Institute of Agricultural Research (EIAR), particularly lowland sorghum breeders, agrometeorology and GIS teams for accessibility and clarification of the data. **DT** thanks the Deutscher Akademischer Austauschdienst (DAAD) for supporting this study.

Author contributions: **DT**: Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; writing—original draft. **HPP**: Methodology; resources; software; reading; editing; supervision and validation of the manuscript . **JH**: Methodology; resources; software; analysis; editing; reading and validation of the manuscript.

Funding: This work is funded by the Deutscher Akademischer Austauschdienst (DAAD) for supporting this study (grant 57507871).

Code availability: The SAS and R code are available as electronic Supplementary Materials.

Data availability: The data used in this study are available as electronic Supplementary Materials.

Compliance with ethical standards

Conflict of interest: The authors declare that there is no conflict of interest.

Consent to participate (include appropriate statements): 'Not applicable'.

Consent for publication (include appropriate statements): 'Not applicable'.

Ethics approval (include appropriate approvals or waivers): 'Not applicable'.