國立臺灣大學生物資源暨農學院農藝學系

碩士論文

Department of Agronomy
College of Bioresources and Agriculture
National Taiwan University

Master's Thesis

於多環境試驗中進行基因體選種之訓練集最佳化

Training Set Optimization in Genomic Selection for Multi-environment Trials

劉子捷

Zi-Jie Liu

指導教授：廖振鐸 博士

Advisor: Chen-Tuo Liao, Ph.D.

中華民國 114 年 6 月

June, 2025

doi:10.6342/NTU202501399

Acknowledgements

當我回顧這兩年的碩士生活，深感受益匪淺。除了學習到更深層的專業知識，也逐漸學會如何清楚表達自己的研究內容。將這段歷程濃縮成一篇論文，對我而言既充實又具意義。

感謝浩嘉學長在研究上的提點與建議，冠瑜學長與冠賢的閒話家常，以及三位在我迷惘時的鼓勵與陪伴，讓研究室充滿笑聲與活力，也讓我能以愉悅的心情做研究。也謝謝育種組同學們上課時的歡樂氛圍，與芷瑄、家香、冠賢、青根一起參加的吃西瓜大賽、玩桌遊和爬山，是滿滿的開心回憶。

在學習基因體選種的過程中，感謝寬諺學長教導的基礎知識，也謝謝亮宇與崇安在學業與生活上的支持。從研究到線模、NLP、飆股預測和統計助教，你們總適時給我鼓勵與幫助。感謝學弟妹崑宏、李婕和顯哥的陪伴，整個實驗室一起和樂融融的抽卡、打麻將、吃飯聊天，很不想告別這樣的時光。

感謝大學時期以來一直很好的大家。花花總陪我談心聽我訴苦，也常約我看棒球。麻亞時常關心我的研究，回來讀育種組時也聊了很多，很有感觸。同為統計組的澹思亭與塗頌揚也常來找我出去吃飯、看電影、打麻將、唱歌、喝酒和出遊。跟你們在一起時總感覺像大學生活般悠閒愜意。

感謝父母尊重並支持我讀農藝系的選擇，也感謝兩位在我就讀研究所期間對我無微不至的照顧和關懷，讓我能無後顧之憂地專注於研究上。

最後，衷心感謝振鐸老師給予我安排研究進度上的自由與信任，也很願意傾聽、理解我對研究的想法，因而能在討論中不斷引導我將想法變得更加完善縝密。這樣的討論模式使我更加有熱忱投入研究，使我在碩士期間收穫頗豐。

doi:10.6342/NTU202501399

摘要

基因型與環境交互作用 (Genotype-by-environment Interaction, G×E) 是植物育種中的常見現象，此因子會影響多環境試驗 (Multi-environment Trials, METs) 中的選拔準確性。將基因體最佳線性無偏預測模型 (Genomic Best Linear Unbiased Prediction Model, GBLUP Model) 納入 G×E 效應並進行基因體選種 (Genomic Selection, GS) ，能在多環境試驗之中提升預測準確性。本研究使用兩種決定係數 (Coefficient of Determination, CD) 指標應用於訓練集最佳化中的方法，並針對識別優良品種的能力，將其與隨機抽取訓練集之方法進行比較。本研究使用基因演算法 (Genetic Algorithm, GA) ，從三個作物資料集: 水稻 (*Oryza sativa* L.)、大麥 (*Hordeum vulgare* L.) 和玉米 (*Zea mays* L.) 中選出最佳訓練集。以三項評估指標: 標準化折扣累積增益 (Normalized Discounted Cumulative Gain, NDCG) 、Spearman 等級相關係數 (Spearman's Rank Correlation, SRC)、以及名次總和比率 (Rank Sum Ratio, RS_{ratio}) 評估上述三種方法的表現。結果顯示，基於 CD 指標選出的訓練集在三項評估指標上都表現得較好。將兩個 CD 指標的表現進行比較，結果顯示 $CD_{mean(v2)}$ 於 SRC 與 RS_{ratio} 兩項指標皆優於 $CD_{mean.MET}$，尤其是在使用較大的訓練集規模時。因此，本研究建議使用 $CD_{mean(v2)}$ 在多環境試驗中將訓練集最佳化。

關鍵字： 訓練集最佳化、基因體選種、基因型與環境交互作用、多環境試驗、基因體最佳線性無偏預測模型、決定係數

doi:10.6342/NTU202501399

Abstract

Genotype-by-environment interaction (G×E) is a key factor in plant breeding, impacting multi-environment trials (METs) for accurate selection. Genomic selection (GS) can improve prediction accuracy across environments, especially with Genomic best linear unbiased prediction (GBLUP) models that account for G×E effects. This study evaluates training set optimization using two coefficient of determination (CD) criteria and compares them to random selection based on the ability to identify elite varieties. A genetic algorithm identified optimal training sets from three datasets of rice (*Oryza sativa* L.), barley (*Hordeum vulgare* L.), and maize (*Zea mays* L.), and their performance was assessed using normalized discounted cumulative gain (NDCG), Spearman's rank correlation (SRC), and rank sum ratio (RS_{ratio}). CD-based training sets showed better performance among these evaluation metrics. The performance of the two CD criteria were compared. $CD_{mean(v2)}$ outperformed $CD_{mean.MET}$ in SRC and RS_{ratio} especially in larger training set sizes. Therefore, $CD_{mean(v2)}$ was highly recommended to select training sets in multi-environment trials.

Keywords: Training Set Optimization, Genomic Selection, G×E Interaction, Multi-environment Trials, Genomic Best Linear Unbiased Prediction Models, Coefficient of Determination

doi:10.6342/NTU202501399

Contents

doi:10.6342/NTU202501399

doi:10.6342/NTU202501399

List of Figures

doi:10.6342/NTU202501399

List of Tables

doi:10.6342/NTU202501399

Chapter 1 Introduction

In plant breeding, the main objective is to find the best variety. In most cases, the target variety has the optimized trait value, so the target variety could be selected by phenotyping the entire breeding population and finding the variety with the optimized trait value in traditional plant breeding. Due to globalization and climate change, the research about plant breeding often focus on the ability of adaptation, hence most experimental trials are now conducted in multiple environments. These trials are therefore called multi-environmental trials (METs). In order to analyze phenotypic data from METs, statistical methodologies have been developed, most of which are linear models. When incorporating the trait values into linear models, the trait values can be partitioned into population mean, genetic effects and environmental effects. The genetic effect is the main focus in plant breeding. The environmental effect, however, is seldom viewed to be important and is often treated as a blocking effect. The blocking effect were even subtracted from the trait value to obtain the adjusted trait values in some studies (Wu *et al.*, 2019).

Nonetheless, there is often interaction between genetic effects and environmental effects, often called genotype-by-environment (G×E) interactions, which is a common factor that strongly affects trait values across environments. In practice, when the G×E effect is small, the relative performance among the breeding population is stable, and the elite varieties are often identical across environments. On the other hand, when the level of G×E effect is large, the relative performance among the breeding population varies, so there will likely be a unique set of elite varieties in a specific environment. Traditionally, linear models are commonly applied to incorporate genetic, environment effect and G×E

doi:10.6342/NTU202501399

effect, including the Finlay-Wilkinson regression model (Finlay & Wilkinson, 1963), the additive main effects and multiplicative interaction effects (AMMI) analysis (Gauch Jr, 1988), as well as the genotype main effect and G×E (GGE) biplot (Yan, 2001; Yan *et al.*, 2007).

Most of the earlier analyses considered the genetic, environment effects and G×E effects to be fixed effects. However, when treating the effects mentioned above as random effects, it is more easily to take the heterogeneous genetic variances across trials, the kinship among varieties, the spatial correlation among experiment units and the correlation of the genetic effects across trials into consideration (Cullis *et al.*, 2020). The linear mixed models are thus the most frequently researched and applied statistical methods when analyzing G×E effects (van Eeuwijk *et al.*, 2016).

A dilemma that plant breeders face is the high cost and laboring of phenotyping in field trials, especially in METs. Since the cost of genotyping by high-density molecular markers, namely single nucleotide polymorphisms (SNPs), has dropped significantly, the genomic information of the breeding population could therefore be applied to predict their trait values and select superior varieties. This powerful approach is called genomic selection (GS). This method utilizes molecular markers over an entire genome and the phenotypic data of the varieties in a training set population to construct a predictive linear model, and predicts the breeding values of the other varieties (Meuwissen *et al.*, 2001). The varieties whose values are predicted are called testing set, and the breeding values predicted are called genomic estimated breeding values (GEBVs). In order to select varieties with elite and stable trait values across environments, GS models should be incorporated into breeding programs to speed up the procedure of selection. (Burgueño *et al.*, 2008; Heslot *et al.*, 2014; Malosetti *et al.*, 2016).

The size and the varieties of the training set could strongly impact the prediction

doi:10.6342/NTU202501399

accuracy in GS (Wu *et al.*, 2023), hence selecting an optimal training set is the key to increase the predicting ability of GS, making the breeding program cost-effective. The Pearson's correlation between the true breeding values (TBVs) and GEBVs is a common index to evaluate the predicting ability of the GS model and the training set since it's an essential component of genetic gain in quantifying the progress of breeding programs (Heffner *et al.*, 2010). Therefore, statistical methods for optimizing the training set considering the Pearson's correlation were extensively studied, one of which mainly focused on the generalized coefficient of determination (CD) (Laloë, 1993) and was applied in linear mixed GS models by Rincent *et al.* (2012), Isidro *et al.* (2015) and Rincent *et al.* (2017). These linear mixed GS models were called genomic best linear unbiased prediction (GBLUP) models since these models treated genetic effect as a random effect. When partitioning the genetic effects into multiple fixed marker effects, on the other hand, the whole genome regression (WGR) models are applied and the optimization methods were different from that of the GBLUP models (Akdemir *et al.*, 2015; Ou & Liao, 2019). All of the training set optimization methods mentioned above were later compared by Fernández-González *et al.* (2023), yet few research optimized training sets considering G×E effects in METs.

For the researches about GS in METs, Lopez-Cruz *et al.* (2015) introduced a GBLUP model considering the interaction between marker effects and environments, where the variances of residuals were assumed to be homogenous across trials. Crossa *et al.* (2016) designed a model based on the concept by Lopez-Cruz *et al.* (2015) but incorporated heterogenous variances of residuals into the model. Furthermore, Alves *et al.* (2021) proposed a GBLUP model extended from the model presented in Crossa *et al.* (2016), and added dominance marker effects into the model to analyze the predictive accuracy of maize hybrids. Nevertheless, most of the research about GS in METs mainly focus on

doi:10.6342/NTU202501399

data analysis instead of training set optimization. Rio *et al.* (2022) did the only research that optimized training sets in METs by the CD criteria, the research evaluated several GBLUP models by the prediction accuracy as well.

The main objective of this study is to develop a procedure to optimize training set based on a multi-environment GBLUP model and the generalized coefficient of determination (CD). The optimized training sets from three genome datasets of rice (*Oryza sativa* L.), barley (*Hordeum vulgare* L.), and maize (*Zea mays* L.) were determined by the CD-based criteria. The training sets were later evaluated for their ability to identify superior varieties in a candidate breeding population by three metrics: normalized discounted cumulative gain (NDCG), Spearman's rank correlation (SRC), and rank sum ratio (RS_{ratio}).

doi:10.6342/NTU202501399

Chapter 2 Materials

2.1 Tropical rice dataset

Spindel *et al.* (2015) introduced a rice dataset, which contains 73147 single-nucleotide polymorphism (SNP) markers along with 363 *indica* or *indica*-admixed rice varieties. The yield, plant height and flowering time were phenotyped for 8 growing seasons, from 2009 to 2012. However, plant height wasn't measured in the wet season of 2009, while 35 of all the varieties have missing phenotypic values. Therefore, only 328 varieties were used in this study.

Since Spindel *et al.* (2015) had demonstrated that it was enough to conduct genomic prediction for this set of tropical rice germplasm by only some of the SNP markers, only one marker was chosen randomly every 0.1 centiMorgan (cM) on all of the chromosomes, significantly reducing the number of SNP markers from 73147 to 10772.

2.2 Barley dataset

Oakey *et al.* (2016) used barley height data to demonstrate the procedure of a genomic selection approach. 648 and 856 spring barley cultivars were sown in the growing season of 2010 and 2011, respectively. In both trials, plants were sown in pots in a spatial row-column design with five blocks. Plant height of these plants were measured at full maturity.

500 of the barley lines mentioned above were genotyped using 7864 single nucleotide polymorphism (SNP) markers, 477 of which were chosen for further analyses, including 456 lines grown across both years, one line from 2010 only, and 17 lines from 2011 only.

doi:10.6342/NTU202501399

In the preprocessing step, Non-polymorphic SNPs, SNPs with more than 10% missing values, SNPs with minor allele frequency < 0.05 and SNPs with identical qualitative coding across the lines were removed, leaving 3490 SNP markers. Cockram *et al.* (2010) described the genotyping details of this data.

2.3 DST2 maize dataset

Jarquín *et al.* (2020) applied two maize datasets from CIMMYT's breeding program to conduct genomic selection analyses in Kenya. The dataset used in this study (DST2) includes 453 CIMMYT hybrids generated by crossing tester genotype "T2" with 453 other genotypes. Grain yield of the maize population was measured in three trials.

The dataset includes genotypes on 73,219 SNP markers. After removing SNPs with missing values > 50% and minor allele frequency < 3%, there were 62882 SNP markers left.

doi:10.6342/NTU202501399

Chapter 3 Methods

3.1 A multi-environment GS model

The model used in the following research was a mixed effect model originally proposed by Lopez-Cruz *et al.* (2015) to capture G×E effects, and was later called "multi-environment prediction model" (labelled as MGE model) by Alves *et al.* (2021).

For a list of training set among T trials, let S_j denote the training set of trial j for $j = 1, 2, \cdots, T$, where S_j consists of S_1, S_2, ⋯, and S_T. The total training set size is n_t, where $n_t = n_1 + n_2 + \cdots + n_T$, and n_j is the training size of S_j. It is possible that a variety is allocated to multiple trials.

The MGE model that only consider the additive effects of the training set can be expressed as follows:

$$\boldsymbol{y}_t = \boldsymbol{\mu}_t + \boldsymbol{a}_t + \boldsymbol{a}_{1t} + \boldsymbol{e}_t, \tag{1}$$

where $\boldsymbol{y}_t = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_T \end{bmatrix}$, $\boldsymbol{y}_t$ is the phenotypic values of the training sets among trials, and $\boldsymbol{\mu}_t = \begin{bmatrix} \mu_1 \mathbf{1}_{n_1} \\ \vdots \\ \mu_T \mathbf{1}_{n_T} \end{bmatrix}$, where μ_t is the population mean of all the trials, μ_j is the mean in trial j for $j = 1, 2, \cdots, T$; $\boldsymbol{a}_t$ is the genetic effect regarding only the additive effects of the training set, where $\boldsymbol{a}_t \sim MVN(\mathbf{0}, \boldsymbol{V}_a)$, and

$$\boldsymbol{V}_a = \sigma_G^2 \begin{bmatrix} \boldsymbol{K}_1 & \cdots & \boldsymbol{K}_{1T} \\ \vdots & \ddots & \vdots \\ \boldsymbol{K}_{T1} & \cdots & \boldsymbol{K}_T \end{bmatrix},$$

doi:10.6342/NTU202501399

where σ_G^2 is the genetic variance considering only the additive effects across every trial; $\boldsymbol{K}_j$ is the kinship matrix between every variety in S_j; and $\boldsymbol{K}_{jj'}$ is the kinship matrix between varieties in S_j and $S_{j'}$.

$\boldsymbol{a}_{1t}$ is the genetic effect regarding the interaction of additive effect and environmental effects (G×E) of the training set, where $\boldsymbol{a}_{1t} \sim MVN(\boldsymbol{0}, \boldsymbol{V}_{a1})$, and

$$\boldsymbol{V}_{a1} = \begin{bmatrix} \sigma_{a\times1}^2 \boldsymbol{K}_1 & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \sigma_{a\times T}^2 \boldsymbol{K}_T \end{bmatrix}.$$

$\boldsymbol{V}_{a1}$ is a block diagonal matrix combining the kinship matrices across every trials, and $\sigma_{a\times j}^2$ is the genetic effect regarding the additive G×E effects in trial j.

$\boldsymbol{e}_t = \begin{bmatrix} \boldsymbol{e}_1 \\ \vdots \\ \boldsymbol{e}_T \end{bmatrix}$ is the residual effect, where its assumed that $\boldsymbol{e}_t \sim MVN(\boldsymbol{0}, \boldsymbol{R}_E)$,

$$\boldsymbol{R}_E = \begin{bmatrix} \sigma_{E1}^2 \boldsymbol{I}_{n_1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \sigma_{ET}^2 \boldsymbol{I}_{n_T} \end{bmatrix}.$$

where σ_{Ej}^2 is the residual variance in trial j, $\boldsymbol{I}_{n_j}$ is the identity matrix of size n_j.

To construct the kinship matrix $\boldsymbol{K}_j$ for all the varieties in S_j, the original SNP matrices in the datasets described in Chapter 2 were used. The whole kinship matrix $\boldsymbol{K}_c$ was first constructed by using all of the varieties in the candidate population. The alleles at each SNP locus were coded as 1, 0, and -1 for the homozygote of the major allele, the heterozygote, and the homozygote of the minor allele, respectively. Note that all of the missing data were imputed as 1. After turning the alphabets representing the nucleobases into 1, 0, and -1, numbers of every locus in the SNP matrices $\boldsymbol{X}_c$ were standardized to obtain the standardized marker score matrices ($\boldsymbol{W}_c$). The kinship matrices of the whole candidate population $\boldsymbol{K}_c$ could be obtained by:

doi:10.6342/NTU202501399

$$\boldsymbol{K}_c = \frac{1}{p}(\boldsymbol{W}_c \boldsymbol{W}_c^T), \quad (2)$$

where p is the number of SNP loci. Matrices $\boldsymbol{K}_j$ and $\boldsymbol{K}_{jj'}$, being the kinship matrix in S_j and the kinship matrix between varieties in S_j and $S_{j'}$, are submatrices of $\boldsymbol{K}_c$.

The steps of predicting the additive genetic effect and the additive G×E effect are described as follows. The variance components were first estimated by restricted maximum likelihood (REML) estimation method. The best linear unbiased prediction (BLUP) for $\boldsymbol{a}_t$ and $\boldsymbol{a}_{1t}$ could be then obtained by solving the Henderson's mixed model equations (Henderson, 1975). Last, the BLUPs for the both effects of the candidate population, $\boldsymbol{a}_c$ and $\boldsymbol{a}_{1c}$, could be obtained by:

$$\hat{\boldsymbol{a}}_c = \boldsymbol{V}_{ca} {\boldsymbol{V}_a}^{-1} \hat{\boldsymbol{a}}_t \quad (3)$$

$$\hat{\boldsymbol{a}}_{1c} = \boldsymbol{V}_{ca1} {\boldsymbol{V}_{a1}}^{-1} \hat{\boldsymbol{a}}_{1t} \quad (4)$$

where

$$\boldsymbol{V}_{ca} = \sigma_G^2 \begin{bmatrix} \boldsymbol{K}_{c1} & \cdots & \boldsymbol{K}_{cT} \\ \vdots & \ddots & \vdots \\ \boldsymbol{K}_{c1} & \cdots & \boldsymbol{K}_{cT} \end{bmatrix},$$

$$\boldsymbol{V}_{ca1} = \begin{bmatrix} \sigma_{a\times 1}^2 \boldsymbol{K}_{c1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \sigma_{a\times T}^2 \boldsymbol{K}_{cT} \end{bmatrix}.$$

Here, $\boldsymbol{K}_{cj}$ is the kinship matrix between the whole candidate population and training set.

The MGE model in Eq. (1) could be simplified by combining the additive genetic effect and the G×E effect. Let the total genetic effect $\boldsymbol{g}_t = \boldsymbol{a}_t + \boldsymbol{a}_{1t}$, and $\sigma_{G\times j}^2 = \sigma_{a\times j}^2 + \sigma_G^2$ for $j = 1, 2, \cdots, T$, then the MGE model could also be expressed as:

$$\boldsymbol{y}_t = \boldsymbol{\mu}_t + \boldsymbol{g}_t + \boldsymbol{e}_t, \quad (5)$$

doi:10.6342/NTU202501399

where $\boldsymbol{g}_t \sim MVN(\boldsymbol{0}, \boldsymbol{G}_t)$, the covariance matrix $\boldsymbol{G}_t$ could be written as:

$$\boldsymbol{G}_t = \begin{bmatrix} \sigma^2_{G\times 1}\boldsymbol{K}_1 & \cdots & \sigma^2_G \boldsymbol{K}_{1T} \\ \vdots & \ddots & \vdots \\ \sigma^2_G \boldsymbol{K}_{T1} & \cdots & \sigma^2_{G\times T}\boldsymbol{K}_T \end{bmatrix}.$$

The total genetic effect $\boldsymbol{g}_t$ and environmental effect $\boldsymbol{e}_t$ were assumed to be mutually independent. After simplifying the model, the BLUP for the total genetic effects $\boldsymbol{g}_t$ could be expressed easier by Henderson's mixed model equations (Henderson, 1975):

$$\hat{\boldsymbol{g}}_t = (\boldsymbol{M}_t + \boldsymbol{G}_t^{-1})^{-1}\boldsymbol{M}_t \boldsymbol{y}_t, \tag{6}$$

where

$$\boldsymbol{M}_t = \begin{bmatrix} (\sigma^2_{E1})^{-1}(\boldsymbol{I}_{n_1} - \bar{\boldsymbol{J}}_{n_1}) & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & (\sigma^2_{ET})^{-1}(\boldsymbol{I}_{n_T} - \bar{\boldsymbol{J}}_{n_T}) \end{bmatrix}.$$

Here, $\boldsymbol{M}_t$ is a block diagonal matrix with the j^{th} diagonal sub-matrix $(\sigma^2_{Ej})^{-1}(\boldsymbol{I}_{n_j} - \bar{\boldsymbol{J}}_{n_j})$, where $j = 1, 2, \ldots\ldots, T$. Each sub-matrix is orthogonal to the corresponding j^{th} population mean sub-vector $\mu_j \boldsymbol{1}_{n_j}$ in $\boldsymbol{\mu}_t$. $\bar{\boldsymbol{J}}_{n_j}$ is a matrix of order n_j with every element equal to $1/n_j$.

For the true genotypic values of the candidate population, their genetic effects were denoted as $\boldsymbol{g}_c$, with the genetic effect in each environment being denoted as $\boldsymbol{g}_{cj}$, where $j = 1, 2, \ldots\ldots, T$. The genetic effect $\boldsymbol{g}_c$ was assumed to be $\boldsymbol{g}_c \sim MVN(\boldsymbol{0}, \boldsymbol{G}_c)$, where

$$\boldsymbol{G}_c = \begin{bmatrix} \sigma^2_{G\times 1}\boldsymbol{K}_c & \cdots & \sigma^2_G \boldsymbol{K}_c \\ \vdots & \ddots & \vdots \\ \sigma^2_G \boldsymbol{K}_c & \cdots & \sigma^2_{G\times T}\boldsymbol{K}_c \end{bmatrix}.$$

Since every sub-matrix involved the kinship matrix of the whole candidate population $\boldsymbol{K}_c$, the variance components could be merged as:

$$\boldsymbol{\Omega}_G = \begin{bmatrix} \sigma^2_{G\times 1} & \cdots & \sigma^2_G \\ \vdots & \ddots & \vdots \\ \sigma^2_G & \cdots & \sigma^2_{G\times T} \end{bmatrix}.$$

doi:10.6342/NTU202501399

The covariance matrix could therefore be expressed as:

$$\boldsymbol{G}_c = \boldsymbol{\Omega}_G \otimes \boldsymbol{K}_c.$$

From Henderson (1977), the BLUP for $\boldsymbol{g}_c$ could be obtained as:

$$\hat{\boldsymbol{g}}_c = \boldsymbol{G}_{ct}(\boldsymbol{M}_t\boldsymbol{G}_t + \boldsymbol{I}_{n_t})^{-1}\boldsymbol{M}_t\boldsymbol{y}_t, \quad (7)$$

where the covariance matrix between the candidate population and the training set is:

$$\boldsymbol{G}_{ct} = \begin{bmatrix} \sigma^2_{G\times 1}\boldsymbol{K}_{c1} & \cdots & \sigma^2_G\boldsymbol{K}_{cT} \\ \vdots & \ddots & \vdots \\ \sigma^2_G\boldsymbol{K}_{c1} & \cdots & \sigma^2_{G\times T}\boldsymbol{K}_{cT} \end{bmatrix}.$$

3.2 Coefficient of determination

The coefficient of determination (CD) between the true breeding value (TBV) and the genomic estimated breeding value (GEBV) has always been a typical metric to evaluate the prediction accuracy of genomic selection. The higher the CD value, the higher the prediction accuracy, resulting in better GS predictions. However, CD can only be calculated only after the trait value is measured.

To solve this dilemma, Laloë (1993) first proposed the generalized coefficient of determination as a useful metric to evaluate genetic effects, whereas Isidro *et al.* (2015) and Rincent *et al.* (2017) applied this metric to optimize the training set, in order to improve prediction accuracy.

Rincent *et al.* (2012) originally proposed a CD criterion called CD_{mean}, then Chen *et al.* (2024) improved this metric by saving its long computational time, thus calling the improved metric $CD_{mean(v2)}$.

Let $A_{l,l'}$ denote the element at row l and column l' in the covariance matrix $Var(\hat{\boldsymbol{g}}_c)$ and $B_{l,l'}$ denote the corresponding element of covariance matrix $Var(\boldsymbol{g}_c)$.

doi:10.6342/NTU202501399

Note that $l, l' = 1, \cdots, n_c + 1, \cdots, (T-1) \times n_c + 1, \cdots T \times n_c$, where n_c is the number of the candidate population. Let g_{ij} and $\hat{g}_{ij}$ denote the true and genomic estimated genotypic effect of a variety i in trial j for $i = 1, \cdots, n_c$, and $j = 1, \cdots, T$. What's more, $Var(\hat{\boldsymbol{g}}_c)$ and $Cov(\boldsymbol{g}_c, \hat{\boldsymbol{g}}_c)$ have been proven to be equivalent mathematically (0):

$$Var(\hat{\boldsymbol{g}}_c) = Cov(\boldsymbol{g}_c, \hat{\boldsymbol{g}}_c) = \boldsymbol{G}_{ct}(\boldsymbol{M}_t\boldsymbol{G}_t + \boldsymbol{I}_{n_t})^{-1}\boldsymbol{M}_t\boldsymbol{G}_{ct}^T. \quad (8)$$

After denoting the simplified expressions above and applying Eq. (8), $CD_{mean(v2)}$ could be further simplified:

$$CD_{mean(v2)} = \sum_{i=1}^{n_c}\sum_{j=1}^{T}\frac{[cov(g_{ij}, \hat{g}_{ij})]^2}{var(g_{ij}) \times var(\hat{g}_{ij})} = \sum_{l=1}^{N}(\frac{[A_{l,l}]^2}{B_{l,l} \times A_{l,l}}) = \sum_{l=1}^{N}(\frac{A_{l,l}}{B_{l,l}}). \quad (9)$$

Rio *et al.* (2022) not only applied CD_{mean} in their research, but also proposed another criterion of CD that is also applicable in multi-environment trials. This criterion focuses on predicting the mean genotypic value of a same variety across trials. In this study, this criterion is called $CD_{mean.MET}$.

$$CD_{mean.MET} = \sum_{i=1}^{n_c}\frac{[cov(\bar{g}_{i.}, \bar{\hat{g}}_{i.})]^2}{var(\bar{g}_{i.}) \times var(\bar{\hat{g}}_{i.})} = \sum_{i=1}^{n_c}(\frac{[A_i^*]^2}{B_i^* \times A_i^*}) = \sum_{i=1}^{n_c}(\frac{A_i^*}{B_i^*}). \quad (10)$$

where $\bar{g}_{i.} = \frac{1}{T}\sum_{j=1}^{T} g_{ij}$, and $\bar{\hat{g}}_{i.} = \frac{1}{T}\sum_{j=1}^{T}\hat{g}_{ij}$ are the mean of the true genotypic effect and the genomic estimated genotypic effect for genotype i across the T trials.

$$A_i^* = \sum_{j=1}^{T}\sum_{m=1}^{T} A_{(j-1)\times n_c + i, (m-1)\times n_c + i}$$

and

doi:10.6342/NTU202501399

$$B_i^* = \sum_{j=1}^{T} \sum_{m=1}^{T} B_{(j-1)\times n_c+i,(m-1)\times n_c+i}.$$

When calculating the value of the above CD criteria, the variance components have to be given since there is no phenotypic data, so there is no way to estimate these components via the REML method. In this study, the variance components of $\sigma_{G\times j}^2$ and σ_{Ej}^2 were set to be 1 for all $j = 1, 2, \cdots, T$, and σ_G^2 was set to be 0.5 when calculating these CD criteria. To obtain the optimal training set, an approach for searching and comparing CD values of different training sets proposed by Ou and Liao (2019) is applied.

3.3 Genetic algorithm

The CD values could only be calculated only after the training set has been decided. Since calculating the CD values of all of the possible combinations is belaboring, an efficient algorithm to find the optimal training set is therefore implemented. The genetic algorithm (GA) is an algorithm inspired by the process of natural selection, a vital part of evolution (Holland, 1992). This algorithm obtains a best training set under a given condition, which corresponds to having the highest CD value in this study. This algorithm mimics natural selection, so it has steps called “crossover” and “mutation”, corresponding to the phenomena of chromosomes, crossover and point mutation.

For a candidate population with n_c individuals, and the size of the optimal training set S_t set to be $n_1, n_2, \ldots\ldots\ n_T$, the genetic algorithm could start by using only the kinship matrix of the whole candidate population, and the given variance components mentioned in Chapter 3.2. Phenotypic values are not required since GA should be conducted before the varieties of the training set is sown. The algorithm initiates by

doi:10.6342/NTU202501399

sampling random training sets of the given training set size. The number of training sets m is set to be $m = 30$ if $n \leq 30$ and $m = n$ if $n > 30$. CD values of every training set is calculated, the training sets with the lowest CD values are labeled to be deleted sets, whereas the training sets with the highest CD values are labeled to be elite sets. In the crossover phase, the varieties of the deleted sets are replaced by combinations of elite sets. In the mutation phase, one of the variety is replaced by another variety for every elite set. If the CD value of the mutated set is higher than that of the original one, the original set would be replaced by the mutated set, otherwise the original elite set remains. For the non-elite training sets, one of the variety is also replaced by another variety, yet the original set would automatically be replaced in order to save computational time for calculating CD value. For every iteration, the CD values of every training set are calculated, and the crossover phase and the mutation phase are conducted. The minimum number of iterations is set to be 12000 in this study. Once the number of iteration exceeds 12000, and the highest CD value in that iteration doesn't improve over a threshold value for 500 consecutive iterations, indicating the max CD value converges, the algorithm stops, and the training set with the highest CD value is chosen to be the optimized set.

3.4 Evaluation metrics

Although CD criteria are useful to evaluate GS accuracy, this metric focuses on the whole candidate population. Nonetheless, most of the plant breeders only consider a few elite genotypes, which have the highest TBVs in a specific trial or the highest mean TBVs across all trials, to be essential. Therefore, three evaluation metrics were employed to evaluate the performance of a training set in identifying the top varieties of a specific candidate population. The number of the top varieties, k, was set to be 5% in this study.

doi:10.6342/NTU202501399

Let $v_{(1)} \geq v_{(2)} \geq \cdots \geq v_{(n_c)}$ denote the TBVs of all the candidates sorted in descending order, and $\hat{v}_{(1)}$, $\hat{v}_{(2)}$, …, $\hat{v}_{(n_c)}$ be the GEBVs of the corresponding candidates estimated by GS model. The rank of TBVs, denoted by $\boldsymbol{\pi}_0$, could be simply expressed as $\boldsymbol{\pi}_0 = (1, 2, \ldots\ldots\ n_c)$. By sorting the GEBVs in descending order, $\hat{v}_{(\pi_1)} \geq \hat{v}_{(\pi_2)} \ldots \geq \hat{v}_{(\pi_{n_c})}$ where $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_{n_c})$ denotes the rank of TBVs, and it is a permutation of $\boldsymbol{\pi}_0$. The three metrics could therefore be calculated after denoting the TBVs, GEBVs, and their rank.

3.4.1 Normalized discounted cumulative gain

This metric was proposed by Blondel *et al.* (2015). The discounted cumulative gain (DCG) for the top k varieties predicted by the training set and the GS model could be calculated as follows:

$$DCG@k\left(\boldsymbol{v}, \pi(\hat{\boldsymbol{v}})\right) = \sum_{i=1}^{k} f\left(v_{(\pi_i)}\right) d(i), \tag{11}$$

while the DCG of the ideal ranking could be calculated as:

$$DCG@k\left(\boldsymbol{v}, \pi_0(\boldsymbol{v})\right) = \sum_{i=1}^{k} f\left(v_{(i)}\right) d(i), \tag{12}$$

Here, $f(v)$ has to be a monotonically increasing gain function and $d(i)$ has to be a monotonically decreasing discount function. In this study, $f(v) = v$ and $d(i) = \frac{1}{\log_2(i+1)}$ were applied in Eq. (12) and (13). The normalized DCG (NDCG) could therefore be obtained by dividing DCG of the predicted top k varieties by the DCG of the ideal ranking:

doi:10.6342/NTU202501399

$$NDCG@k = \frac{DCG@k(\boldsymbol{v}, \pi(\hat{\boldsymbol{v}}))}{DCG@k(\boldsymbol{v}, \pi_0(\boldsymbol{v}))}, \quad (13)$$

The value of NDCG falls between 0 and 1. The higher the NDCG value, the better performance the training set could predict.

3.4.2 Spearman's rank correlation

This metric was proposed by Spearman (1904). This metric evaluates the linear relationship between the rank of the TBVs and the GEBVs of the top k varieties.

Spearman's rank correlation (SRC) could be calculated as follows:

$$SRC@k = \frac{\sum_{i=1}^{k}(i - \frac{k+1}{2})(\pi_i - \bar{\pi})}{\sqrt{[\sum_{i=1}^{k}(i - \frac{k+1}{2})^2] \times [\sum_{i=1}^{k}(\pi_i - \bar{\pi})^2]}}, \quad (14)$$

where $\bar{\pi} = \sum_{i=1}^{k} \pi_i / k$. The value of SRC falls between -1 and 1. The higher the value, the higher correlation between the rank of TBVs and the GEBVs of the top k varieties. The SRC could also be interpreted as the Pearson's correlation of the paired values of the rank of TBVs and GEBVs (i, π_i) for $i = 1, 2, \cdots, k$.

3.4.3 Rank sum ratio

SRC evaluates the linear relationship between the rank of TBVs and GEBVs, yet SRC couldn't reflect the number of elite varieties found. Therefore, a metric called rank sum ratio (RS_{ratio}) was proposed to solve this problem.

The sum of the ideal ranking is $\sum_{i=1}^{k} i$, while the sum of the ranks in TBVs of the top k genotypes with the highest GEBVs is $\sum_{i=1}^{k} \pi_i$. The sums of the former ranking

doi:10.6342/NTU202501399

was then divided by that of the latter. The ratio of the sums of the two rankings evaluates the ability to identify the elite genotypes which has the highest TBV values.

$$RS_{ratio}@k = \frac{\sum_{i=1}^{k} i}{\sum_{i=1}^{k} \pi_i}. \quad (15)$$

The RS_{ratio} ranges from 0 to 1 as well. Higher values of RS_{ratio} indicate that the elite varieties ranked by GEBVs has higher rankings of TBVs. The value of RS_{ratio} equals 1 if the top k genotypes with the highest GEBVs exactly include those k genotypes with the highest TBVs.

3.5 Simulation studies

The kinship matrices of every dataset were used to simulate multiple traits under various circumstances in multi-environment trials. The number of trials were set to be 2, 2, and 3 for the tropical rice, barley, and DST2 maize dataset, respectively.

To discuss the performance of three evaluation metrics under different training set sizes, this study set four combinations of training set sizes, described in Table 3.1. The training set sizes in multiple trials for different datasets.

The training sets were determined by three methods: random sampling, conducting GA by using $CD_{mean\ (v2)}$ and $CD_{mean.MET}$ as the objective metric, abbreviated as random sampling method, $CD_{mean\ (v2)}$ method and $CD_{mean.MET}$ method. The randomly sampled training set has 2000 replicates.

Next, based on the MGE model described in Eq. (2), the simulated phenotypic data of the whole candidate population were simulated by R language. The simulated phenotypic data were partitioned into three main components: the population mean, the genotypic effects and the environmental effects.

doi:10.6342/NTU202501399

For the tropical rice and barley dataset, the population means of the two trials were fixed at $\mu_1 = 100$ and $\mu_2 = 150$; while for the DST2 maize dataset, the population means of the three trials were fixed at $\mu_1 = 100$, $\mu_2 = 150$ and $\mu_3 = 200$.

This study set three combinations of variance components for all of the datasets to discuss the performance of three evaluation metrics under different circumstances. The circumstances were quantified by a metric called the correlation of true genotypic values between environments, and it was denoted as ρ (Rio *et al*., 2022). For the MGE model, ρ could be calculated by:

$$\rho = \frac{\sigma_G^2}{\sqrt{(\sigma_{G\times a}^2)(\sigma_{G\times b}^2)}}, \tag{16}$$

where a and $b = 1,2 \ldots\ldots T$, $a \neq b$.

For the tropical rice and barley dataset, the three levels of ρ between trial 1 and 2 were set to be 0.2, 0.5, and 0.8 to represent strong, mediate, and weak level of G×E. In order to obtain these values, the genetic variance components were set to be: $\sigma_G^2 = 10$, $\sigma_{G\times 1}^2 = 13$, and $\sigma_{G\times 2}^2 = 192$, 30 and 12, corresponding to the three values of ρ. As for the DST2 maize dataset, the three levels of ρ between trial 1 and 3 were also set to be 0.2, 0.5, and 0.8. The genetic variance components were set to be: $\sigma_G^2 = 10$, $\sigma_{G\times 1}^2 = \sigma_{G\times 2}^2 = 13$, and $\sigma_{G\times 3}^2 = 192$, 30 and 12, corresponding to the three values of ρ.

The environmental variance components of all the datasets were calculated by $\sigma_{Ei}^2 = \frac{\sigma_{G\times i}^2(1-h^2)}{h^2}$ for $i = 1, 2, 3$, and the narrow sense heritability h^2 being set to be 0.5. After setting the variance components, the kinship matrices of every datasets were obtained as described in Chapter 3.1. Based on the MGE model described in Eq. (5), $\boldsymbol{g}_c$ and $\boldsymbol{e}_c$ could be drawn from the multivariate normal distributions $\boldsymbol{g}_c \sim \text{MV}N(\boldsymbol{0}, \boldsymbol{\Omega}_G \otimes$

doi:10.6342/NTU202501399

$\boldsymbol{K}_c$) and $\boldsymbol{e}_c \sim \text{MV}N\left(\boldsymbol{0}, \boldsymbol{\Omega}_E \otimes \boldsymbol{I}_{n_c}\right)$, and could then be generated by the R function "mvrnorm" (Venables & Ripley, 2002).

The simulated TBVs were then obtained by adding the population mean and the genetic effects, whereas the simulated phenotypic values were obtained by adding the simulated TBVs and the environmental effects. A total of 2000 replicates of simulated traits were generated, and genomic selection was conducted by the REML method using the R function "mmer" in the R package "sommer" (Covarrubias-Pazaran, 2016).

Last, the three evaluation metrics, NDCG, SRC and RS_{ratio} of the simulated data were calculated, to compare the performance of the training sets determined by different methods under different conditions.

3.6 Real data analyses

For the barley and the DST2 maize datasets, the training sets are determined by the same method as that of the simulation studies described in Chapter 3.2, including random sampling, GA conducted by using the $CD_{mean\ (v2)}$ metric and GA conducted by using the $CD_{mean.MET}$ metric. Then the true phenotypic values of the training set were used to conduct GS. The phenotypic values were viewed as TBVs, so that the NDCG, SRC and RS_{ratio} of the top 5% elite varieties in the candidate population could be calculated.

Unfortunately, no multi-environment phenotypic data were available for the tropical rice dataset, so the real trait analysis for this dataset was not conducted.

doi:10.6342/NTU202501399

Table 3.1. The training set sizes in multiple trials for different datasets.

	Datasets		
Combinations	Tropical rice	Barley	DST2 maize
I	$(n_1 = 50, n_2 = 50)$	$(n_1 = 50, n_2 = 50)$	$(n_1 = 50, n_2 = 50, n_3 = 50)$
II	$(n_1 = 50, n_2 = 100)$	$(n_1 = 50, n_2 = 100)$	$(n_1 = 50, n_2 = 50, n_3 = 100)$
III	$(n_1 = 100, n_2 = 100)$	$(n_1 = 100, n_2 = 100)$	$(n_1 = 50, n_2 = 100, n_3 = 100)$
IV	$(n_1 = 100, n_2 = 150)$	$(n_1 = 100, n_2 = 150)$	$(n_1 = 100, n_2 = 100, n_3 = 100)$

doi:10.6342/NTU202501399

Chapter 4 Results

The training sets determined by GA have high CD scores, which would also have high r^2 values theoretically. However, the r^2 values were calculated based on the whole candidate population, yet the evaluation metrics are only calculated by the elite genotypes. Therefore, the values of evaluation metrics are not guaranteed to be higher when the training sets have higher CD values. In this study, the main objective is to find out whether the training sets with high CD values would have higher values of the three evaluation metrics, which indicate that those training sets would have better ability to identify elite varieties.

4.1 Simulation studies

In order to compare the performance of the methods, simulation studies were conducted in case the results displayed were strongly affected by a few specific phenotypic data. The performance of the evaluation metrics was compared in the simulation studies section over various circumstances. The mean and standard deviation over the 2000 simulation replicates for the tropical rice dataset, the barley dataset, and the DST2 maize dataset are shown in Figure 4.1 - Figure 4.3, Figure 4.4 - Figure 4.6, and Figure 4.7 - Figure 4.9, respectively.

The methods of determining the training sets showed much difference in values of NDCG, SRC and RS_{ratio}. the training set determined by $CD_{mean(v2)}$ displayed higher values in all three evaluation metrics than the random sampling method, the training set determined by $CD_{mean.MET}$ displayed higher values in NDCG and RS_{ratio} than the random sampling method as well. Since the results of the tropical rice dataset displayed the most

doi:10.6342/NTU202501399

difference, the main results for the simulation studies would focus on this dataset (Figure 4.1 - Figure 4.3). However, the relative performance in the three datasets were nearly identical.

4.1.1 Normalized discounted cumulative gain

For the tropical rice dataset, the mean NDCG value ranged from 0.8893 to 0.9869, indicating that this metric had relatively small difference between methods, correlation of true genotypic values between environments (ρ), and training set sizes (Figure 4.1).

When correlation of true genotypic values between environments rose, the NDCG value also rose. In the case when the training set size $n_1 = 50, n_2 = 50$ and ρ = 0.2, the NDCG values in the overall environment for the random sampling, $CD_{mean(v2)}$, and $CD_{mean.MET}$ methods are 0.9247, 0.9464, and 0.9449; the corresponding values when ρ = 0.5 were 0.9608, 0.9746 and 0.9742; and those when ρ = 0.8 were 0.9700, 0.9814 and 0.9814. However, the difference between random sampling and the two CD methods were higher when ρ = 0.2, with the difference being 0.0217 and 0.0202, while the difference was lower when ρ = 0.8, with the difference being 0.0138 and 0.0134.

What's more, the genetic variance $\sigma^2_{G\times 2}$ = 192, 30 and 12 when ρ = 0.2, 0.5 and 0.8, respectively. When $\sigma^2_{G\times 2}$ increased, the NDCG value in environment 2 (ENV2) dropped significantly. In the case when the training set size $n_1 = 50, n_2 = 50$ and ρ = 0.2, the NDCG values in ENV2 for the random sampling, $CD_{mean(v2)}$ and $CD_{mean.MET}$ methods were 0.8893, 0.9206, and 0.9184; the corresponding values when ρ = 0.5 were 0.9517, 0.9667 and 0.9663; and those when ρ = 0.8 were 0.9718, 0.9822 and 0.9817.

The training set size also affected NDCG values. When the training set size rose, the NDCG value also rose. In the case when the training set size $n_1 = 50, n_2 = 50$ and ρ

doi:10.6342/NTU202501399

= 0.2, the NDCG values in the overall environment for the random sampling, $CD_{mean(v2)}$, and $CD_{mean.MET}$ methods were 0.9247, 0.9464, and 0.9449; the corresponding values when training set size $n_1 = 100, n_2 = 150$ were 0.9436, 0.9617 and 0.9502. However, the difference between random sampling and the two CD methods were higher when $n_1 = 50, n_2 = 50$, with the difference being 0.0217 and 0.0202, while the difference was lower when $n_1 = 100, n_2 = 150$, with the difference being 0.0181 and 0.0066.

Both the $CD_{mean(v2)}$ method and the $CD_{mean.MET}$ method had better performance than the random sampling method, with the least improvement of the $CD_{mean(v2)}$ method equaled to 0.0085 and that of the $CD_{mean.MET}$ method being 0.0071 when $n_1 = 100, n_2 = 150$ and ρ = 0.8. Additionally, the $CD_{mean(v2)}$ method had non-inferior performance compared to the $CD_{mean.MET}$ method, with the least improvement being 0 when $n_1 = 50, n_2 = 50$ and ρ = 0.8.

4.1.2 Spearman's rank correlation

For the tropical rice dataset, the mean SRC value ranged from 0.2586 to 0.3919. This metric had relatively large difference when compared to NDCG (Figure 4.2). However, the standard deviation mostly has large values.

When correlation of true genotypic values between environments rose, the SRC value also increased. In the case when the training set size $n_1 = 50, n_2 = 50$ and ρ = 0.2, the SRC values in the overall environment for the random sampling, $CD_{mean(v2)}$, and $CD_{mean.MET}$ methods were 0.2586, 0.3067, and 0.2923; the corresponding values when ρ = 0.5 were 0.2634, 0.3254 and 0.3058; and those when ρ = 0.8 were 0.2918, 0.3773 and 0.3573. The difference between random sampling and the two CD methods increased when ρ increased, too. When ρ = 0.2, the difference were 0.0481 and 0.0337; when ρ

doi:10.6342/NTU202501399

= 0.8, the difference were 0.0865 and 0.0665.

The training set size also affected SRC values. When the training set size increased, the SRC value for the random sampling method and the $CD_{mean(v2)}$ method also increased. Yet, the $CD_{mean.MET}$ method didn't show increase. In the case when the training set size $n_1 = 50, n_2 = 50$ and ρ = 0.2, the SRC values in the overall environment for the random sampling, $CD_{mean(v2)}$, and $CD_{mean.MET}$ methods were 0.2586, 0.3067, and 0.2923; the corresponding values when training set size $n_1 = 100, n_2 = 150$ were 0.3062, 0.3535 and 0.2936.

Only the $CD_{mean(v2)}$ method had better performance than the random sampling method under every circumstance, with the least improvement of the $CD_{mean(v2)}$ method equaled to 0.0473 when $n_1 = 100, n_2 = 150$ and ρ = 0.8. The $CD_{mean.MET}$ method had better performance than the random sampling method when the training set size is low, but it has even worse performance than the random sampling method when the training set size is high, when $n_1 = 100, n_2 = 100$ and $n_1 = 100, n_2 = 150$.

In the barley and DST2 maize dataset, however, the difference between the three methods of determining trainings sets were minimal. In some circumstances, the CD-based methods even showed inferior improvements when compared to the random sampling methods, especially when the training set size was higher.

4.1.3 Rank sum ratio

For the tropical rice dataset, the mean RS_{ratio} value ranged from 0.1095-0.2812. This metric had relatively small value and large difference when compared to NDCG (Figure 4.3). The standard deviation was high, but slightly lower than that of SRC.

When correlation of true genotypic values between environments rose, the RS_{ratio} value also increased. When the training set size $n_1 = 50, n_2 = 50$ and ρ = 0.2, the

doi:10.6342/NTU202501399

RS_{ratio} values in the overall environment for the random sampling, $CD_{mean(v2)}$, and $CD_{mean.MET}$ methods were 0.1095, 0.1421, and 0.1375; the corresponding values when ρ = 0.5 were 0.1230, 0.1632 and 0.1598; and those when ρ = 0.8 were 0.1396, 0.1878 and 0.1868. The difference between random sampling and the two CD methods slightly increased when ρ increased, too. When ρ = 0.2, the difference were 0.0326 and 0.0280; when ρ = 0.8, the difference were 0.0482 and 0.0472.

The training set size also affected RS_{ratio} values. When the training set size increased, the RS_{ratio} value also increased. In the case when the training set size $n_1 = 50, n_2 = 50$ and ρ = 0.2, the RS_{ratio} values in the overall environment for the random sampling, $CD_{mean(v2)}$, and $CD_{mean.MET}$ methods were 0.1095, 0.1421, and 0.1375; the corresponding values when training set size $n_1 = 100, n_2 = 150$ were 0.1898, 0.2360 and 0.2090.

Both the $CD_{mean(v2)}$ method and the $CD_{mean.MET}$ method had better performance than the random sampling method, with the least improvement of the $CD_{mean(v2)}$ method equaled to 0.0417 and that of the $CD_{mean.MET}$ method being 0.0111. Additionally, the $CD_{mean(v2)}$ method had non-inferior performance compared to the $CD_{mean.MET}$ method, with the least improvement being 0.0010 when $n_1 = 50, n_2 = 50$ and ρ = 0.8. When the training set size increased and ρ decreased, the difference between the $CD_{mean(v2)}$ method and the $CD_{mean.MET}$ method increased.

In the barley and DST2 maize dataset, however, the difference between the three methods of determining trainings sets were little. However, under all of the circumstances, the overall performance of the CD-based methods showed improvements when compared to the random sampling methods.

doi:10.6342/NTU202501399

4.2 Real data analyses

Simulation studies were conducted to compare the methods in a theoretical inspection. The next step was to apply data in real life into the GS model and compare the results between the methods.

The results of evaluation metrics of NDCG, SRC and RS_{ratio} using the barley and DST2 maize datasets were shown in Figure 4.10 and Figure 4.11, respectively.

The results were not consistent with those displayed in the simulation studies section. Nevertheless, the CD-based methods still showed non-inferior performance than the random sampling method in NDCG and RS_{ratio} in most of the cases.

4.2.1 Normalized discounted cumulative gain

For the NDCG metric, the overall performance of the CD-based methods showed higher values than the random sampling method under most training set sizes. The overall performance of the $CD_{mean(v2)}$ method was higher than the random sampling method in both datasets and all training sizes, except under the training set size $n_1 = 100, n_2 = 150$ in the DST2 maize dataset. However, the difference between the $CD_{mean(v2)}$ method and the random sampling method in the DST2 maize dataset were much lower than that in the barley dataset. The overall performance of the $CD_{mean.MET}$ method was also higher than the random sampling method in most datasets and training sizes, with the exception being under the training set size $n_1 = 50, n_2 = 50$ in the barley dataset and under the training set size $n_1 = 50, n_2 = 100$ in the DST2 maize dataset. The difference between the $CD_{mean.MET}$ method and the random sampling method in the DST2 maize dataset were also much lower than that in the barley dataset.

doi:10.6342/NTU202501399

4.2.2 Spearman's rank correlation

For the SRC metric, the overall performance of the CD-based methods didn't display higher values than the random sampling method under most training set sizes. The overall performance of the $CD_{mean(v2)}$ method was only higher than the random sampling method in the barley dataset under the training set size $n_1 = 50, n_2 = 50$ and $n_1 = 100, n_2 = 150$, and it had lower overall performance under all training set sizes in the DST2 maize dataset. The overall performance of the $CD_{mean.MET}$ method was higher than the random sampling method in the DST2 maize dataset except under the training set size $n_1 = 100, n_2 = 100$. However, the overall performance of the $CD_{mean.MET}$ method was only higher than the random sampling method under the training set size $n_1 = 100, n_2 = 100$ in the barley dataset.

4.2.3 Rank sum ratio

For the RS_{ratio} metric, the overall performance of the CD-based methods displayed non-inferior values than the random sampling method under most training set sizes. The overall performance of the $CD_{mean(v2)}$ method was higher than the random sampling method in both datasets except under the training set size $n_1 = 50, n_2 = 50$ in the DST2 maize dataset, yet its difference was also almost negligible. The overall performance of the $CD_{mean.MET}$ method was not lower than the random sampling method in both datasets, with the DST2 maize dataset being the exception under the training set size $n_1 = 50, n_2 = 100$ and $n_1 = 100, n_2 = 100$. However, the performance of the $CD_{mean.MET}$ method and the random sampling method had nearly identical values in most of the other circumstances.

doi:10.6342/NTU202501399

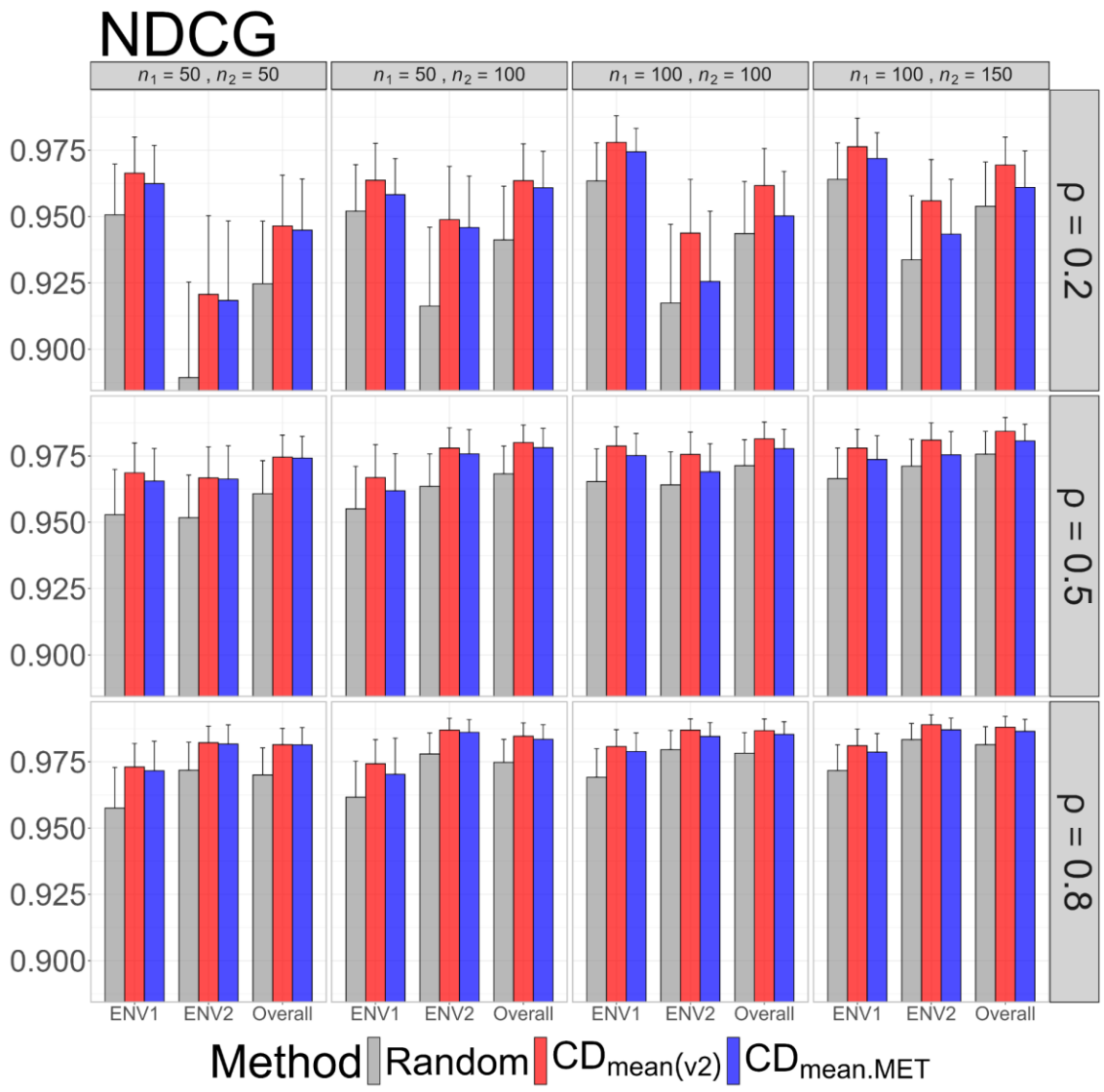

Figure 4.1. The means of the NDCG value for the simulated tropical rice dataset.

doi:10.6342/NTU202501399

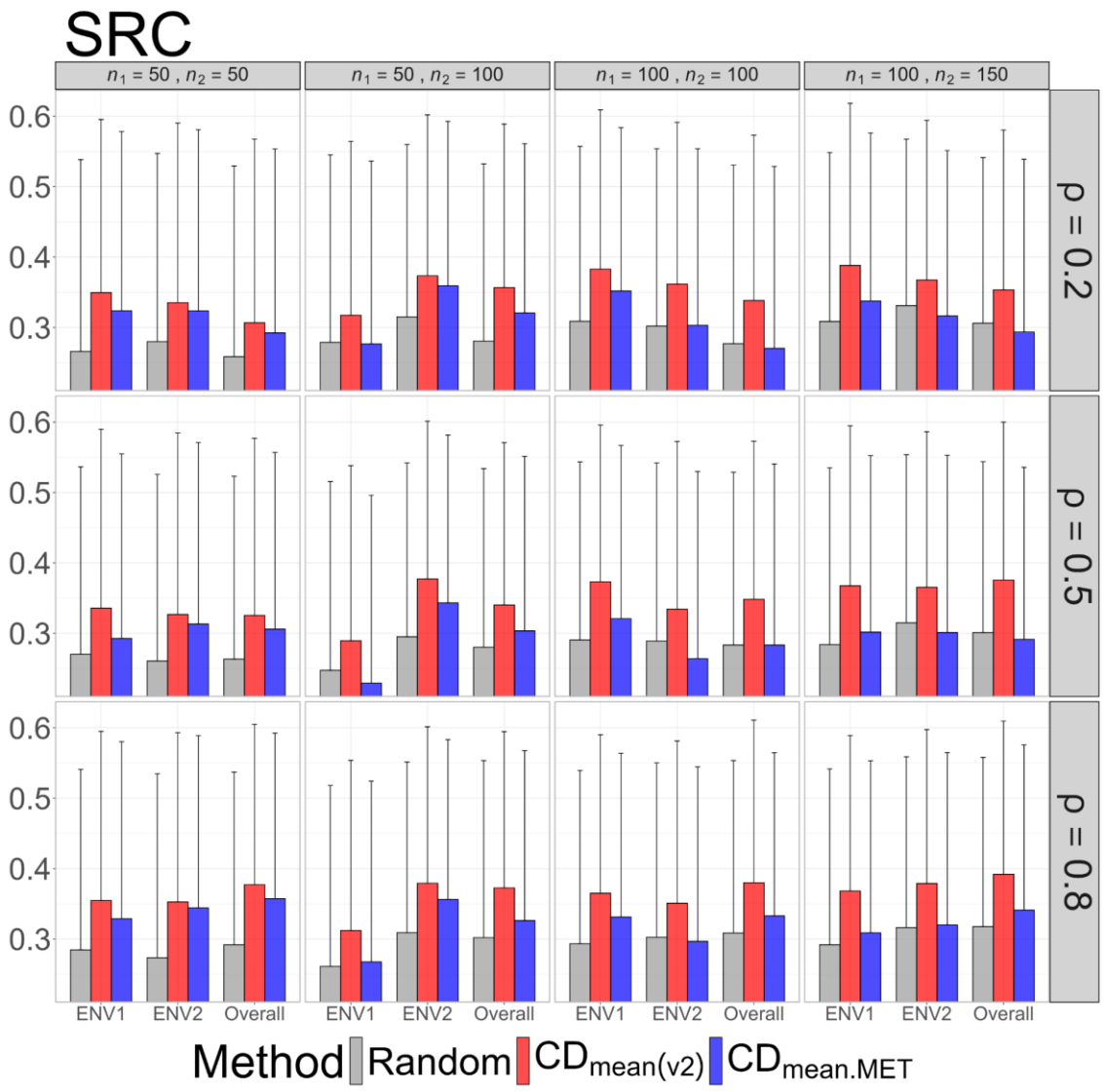

Figure 4.2. The means of the SRC value for the simulated tropical rice dataset.

doi:10.6342/NTU202501399

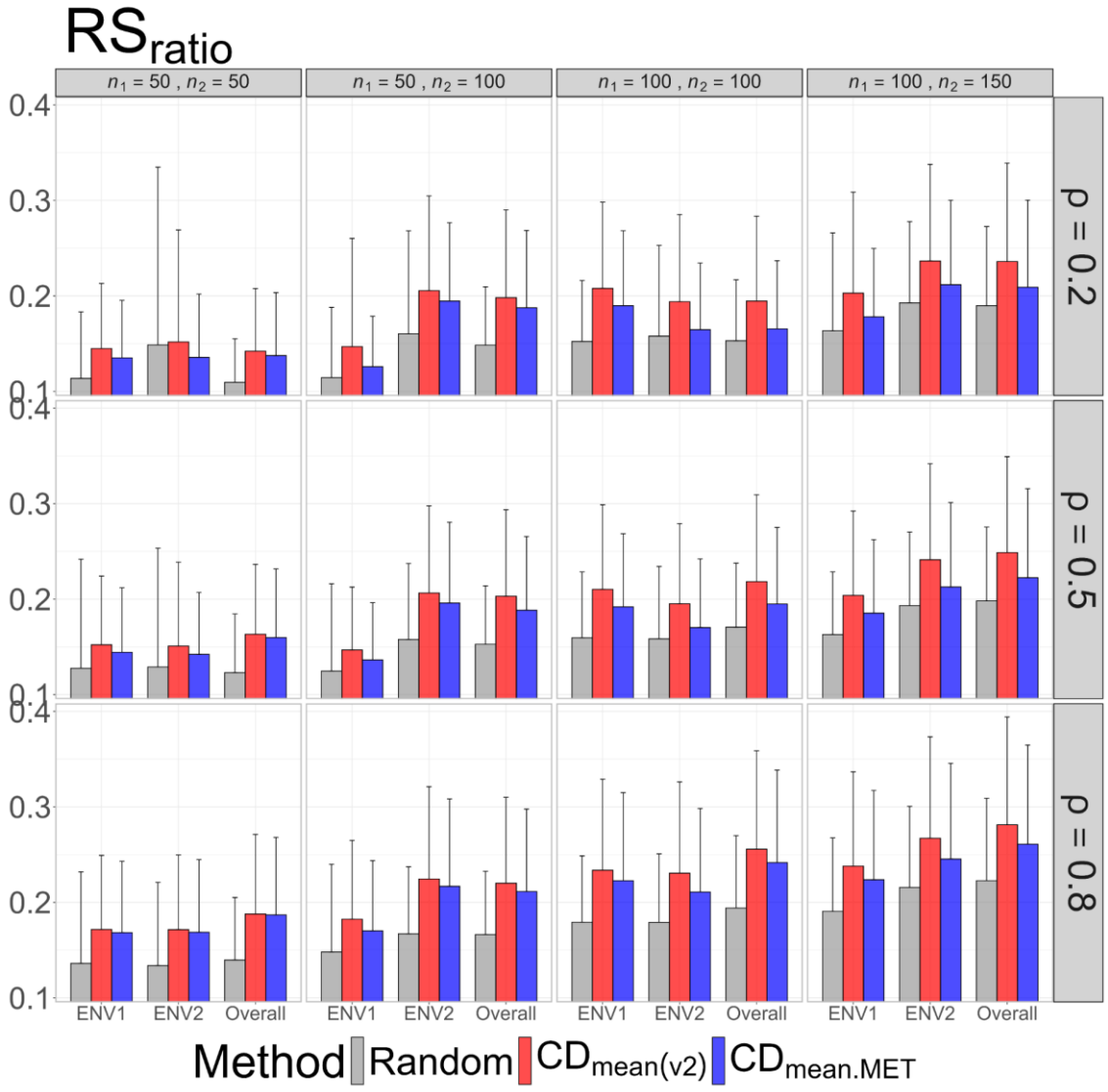

Figure 4.3. The means of the RS_{ratio} value for the simulated tropical rice dataset.

doi:10.6342/NTU202501399

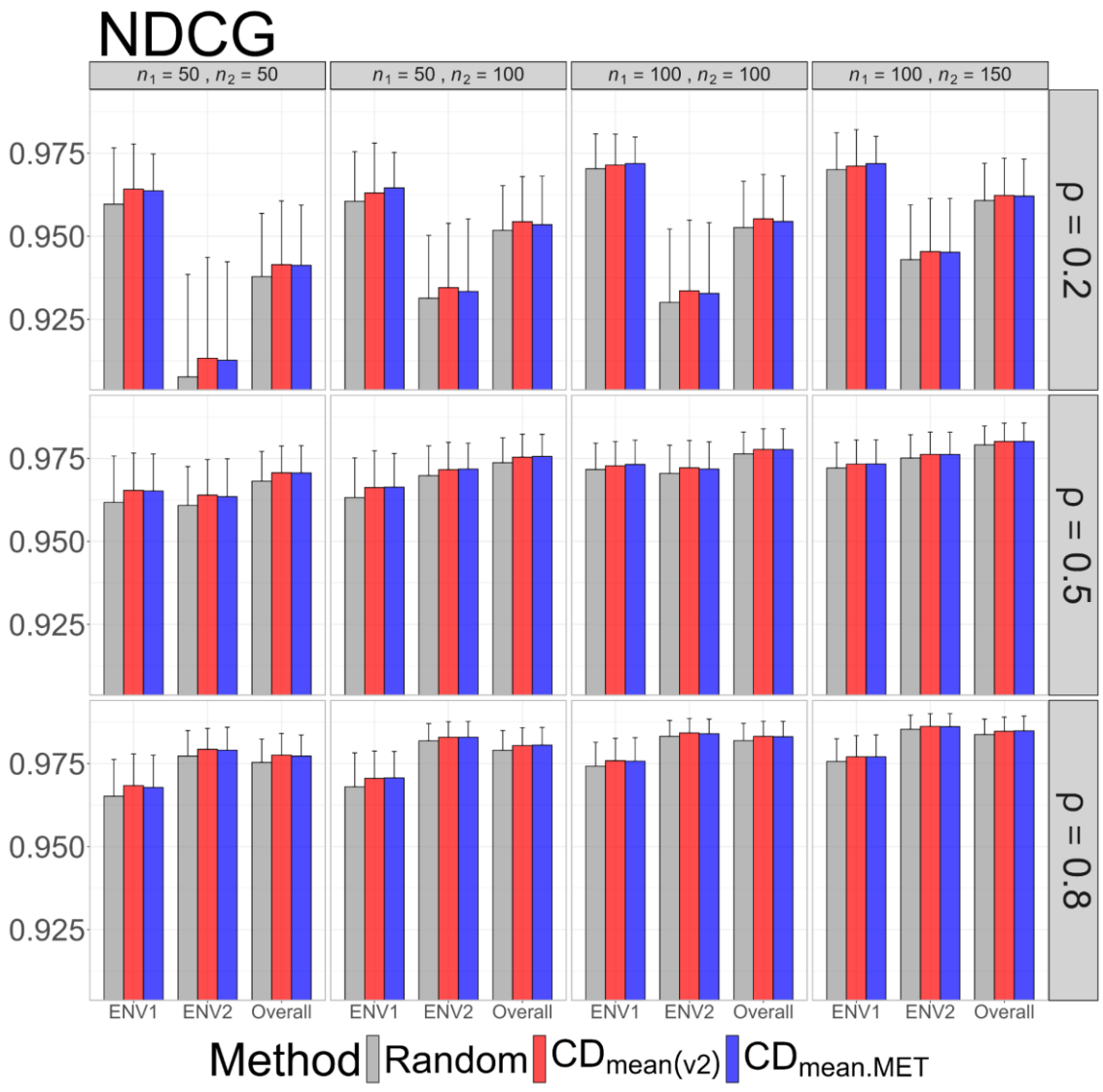

Figure 4.4. The means of the NDCG value for the simulated barley dataset.

doi:10.6342/NTU202501399

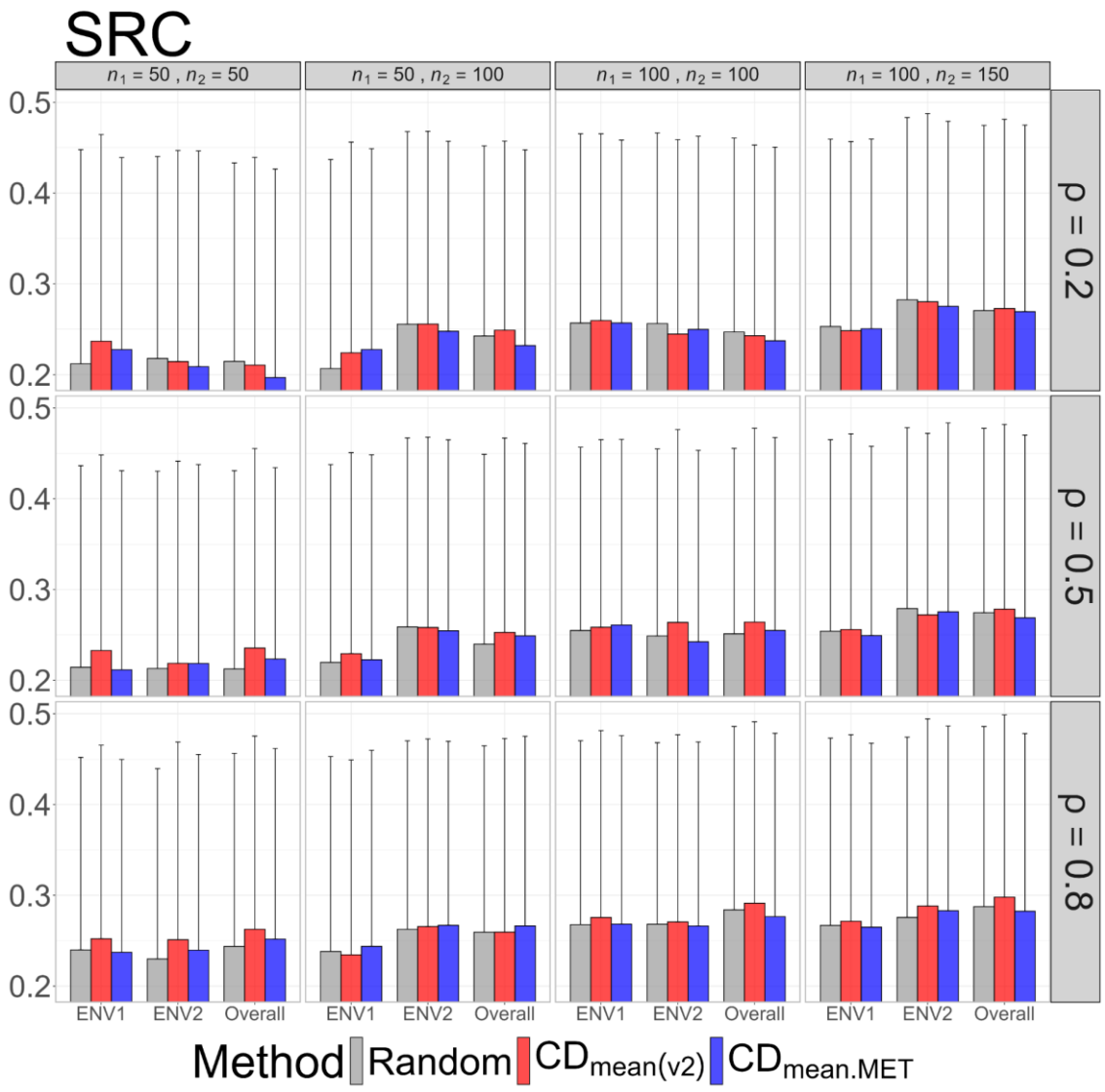

Figure 4.5. The means of the SRC value for the simulated barley dataset.

doi:10.6342/NTU202501399

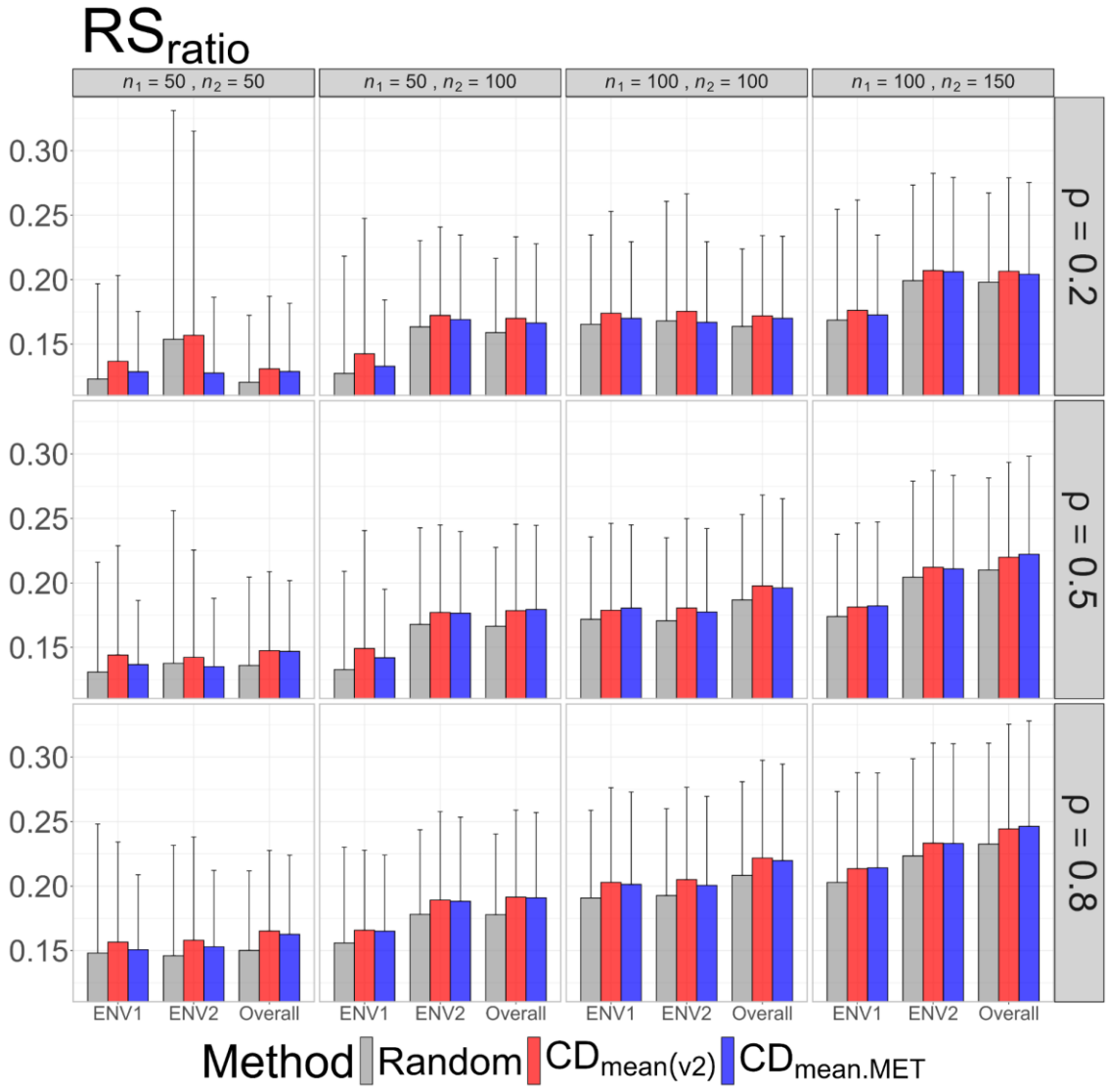

Figure 4.6. The means of the RS_{ratio} value for the simulated barley dataset.

doi:10.6342/NTU202501399

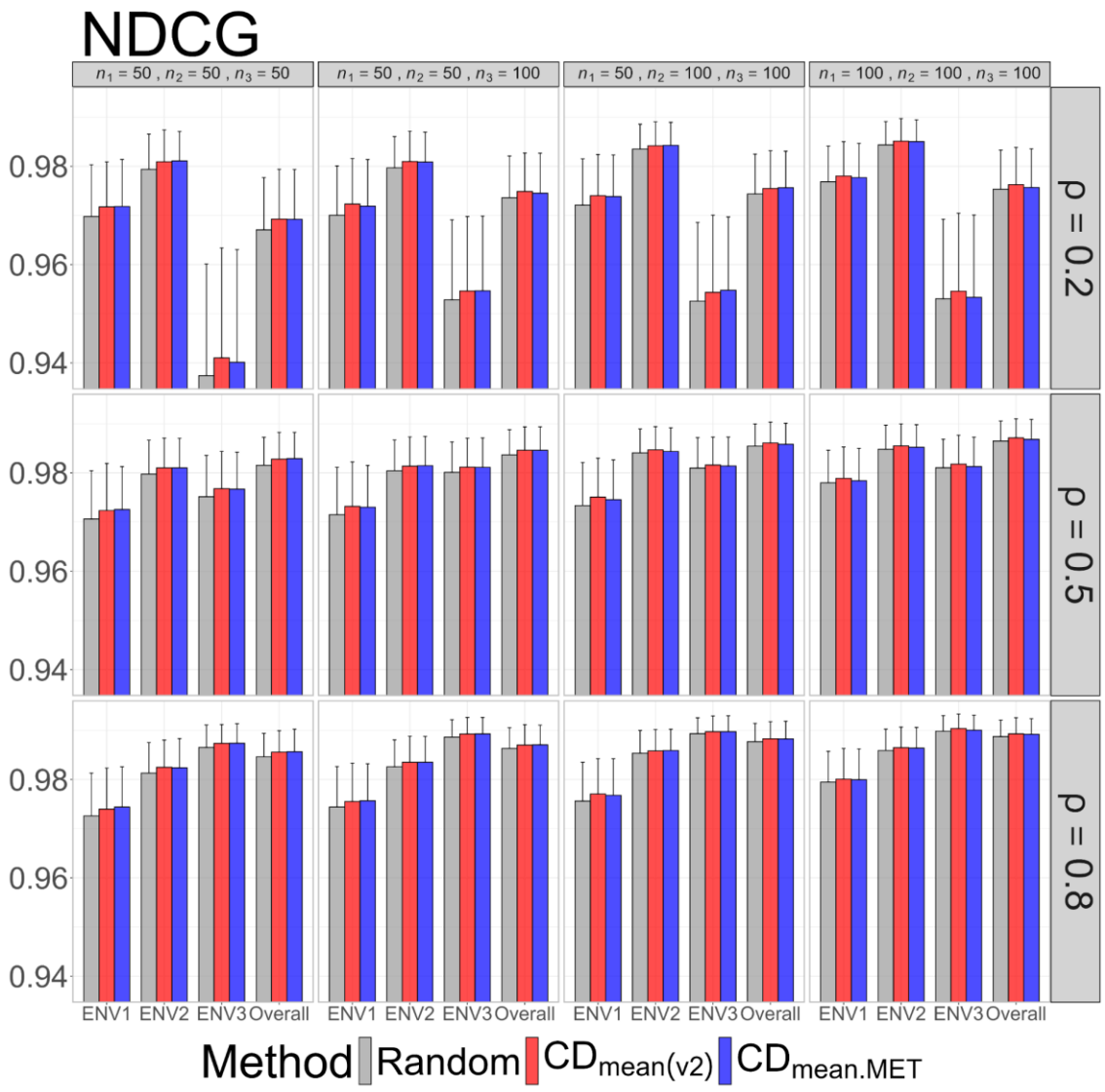

Figure 4.7. The means of the NDCG value for the simulated DST2 maize dataset.

doi:10.6342/NTU202501399

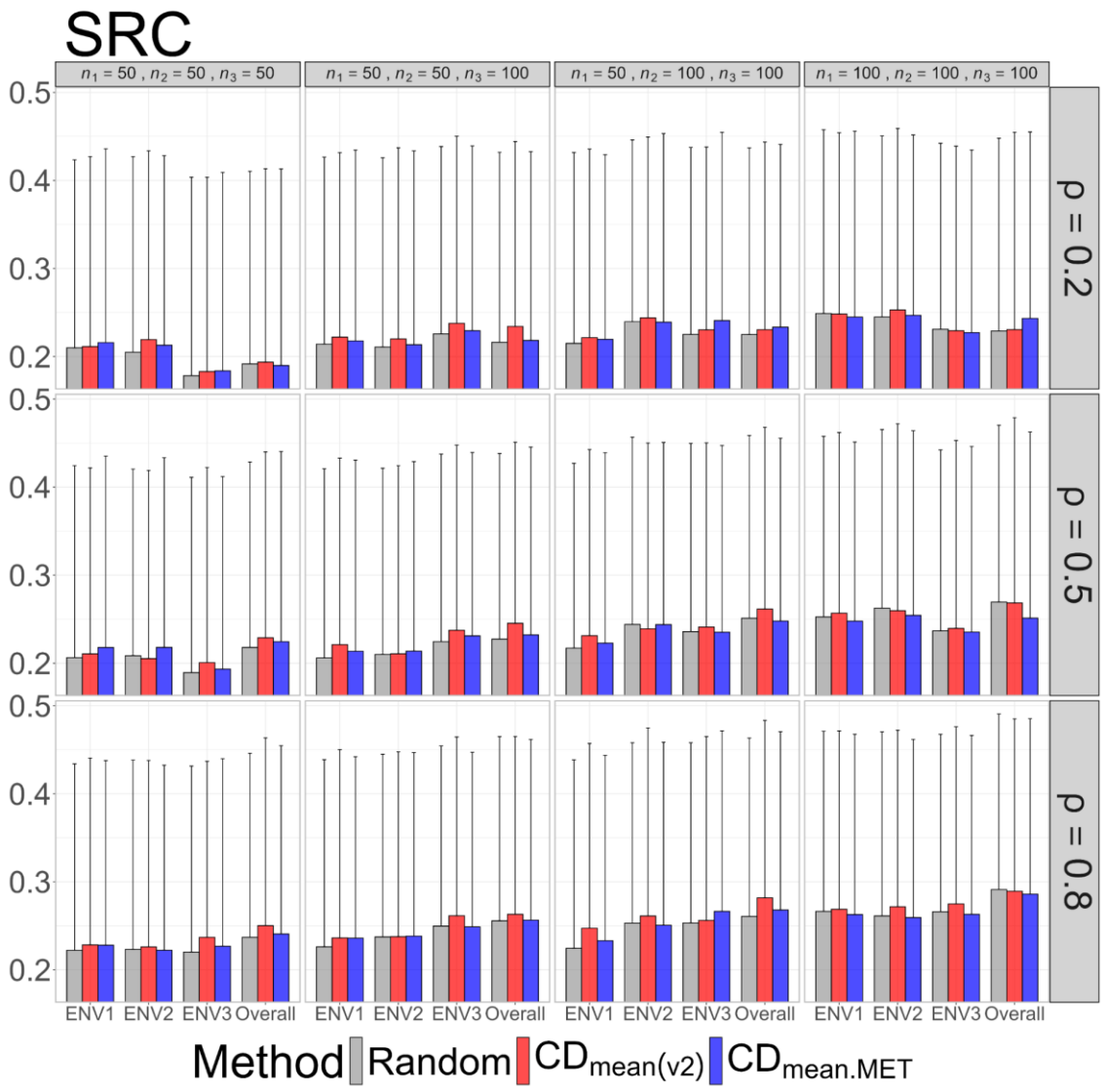

Figure 4.8. The means of the SRC value for the simulated DST2 maize dataset.

doi:10.6342/NTU202501399

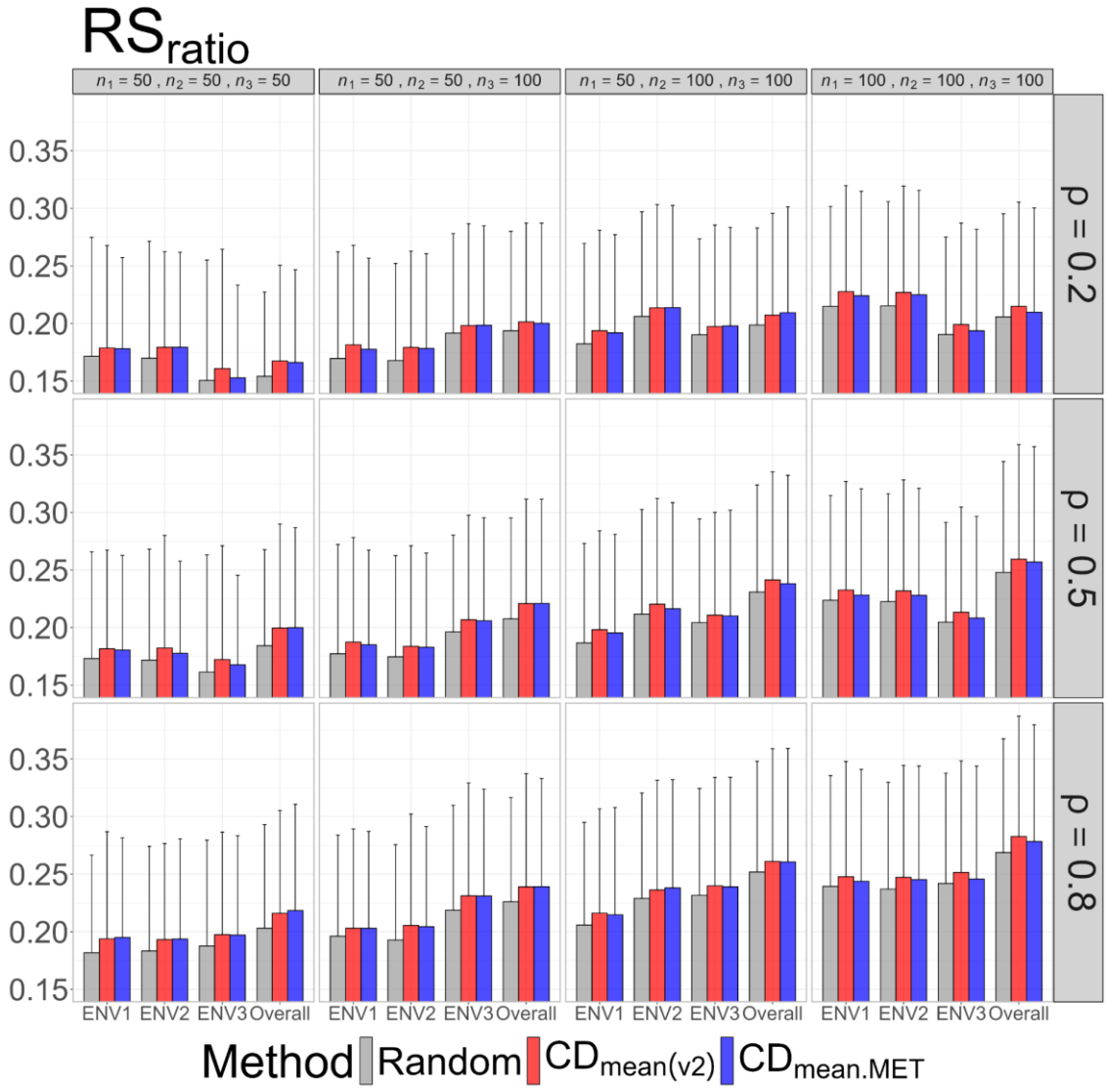

Figure 4.9. The means of the RS_{ratio} value for the simulated DST2 maize dataset.

doi:10.6342/NTU202501399

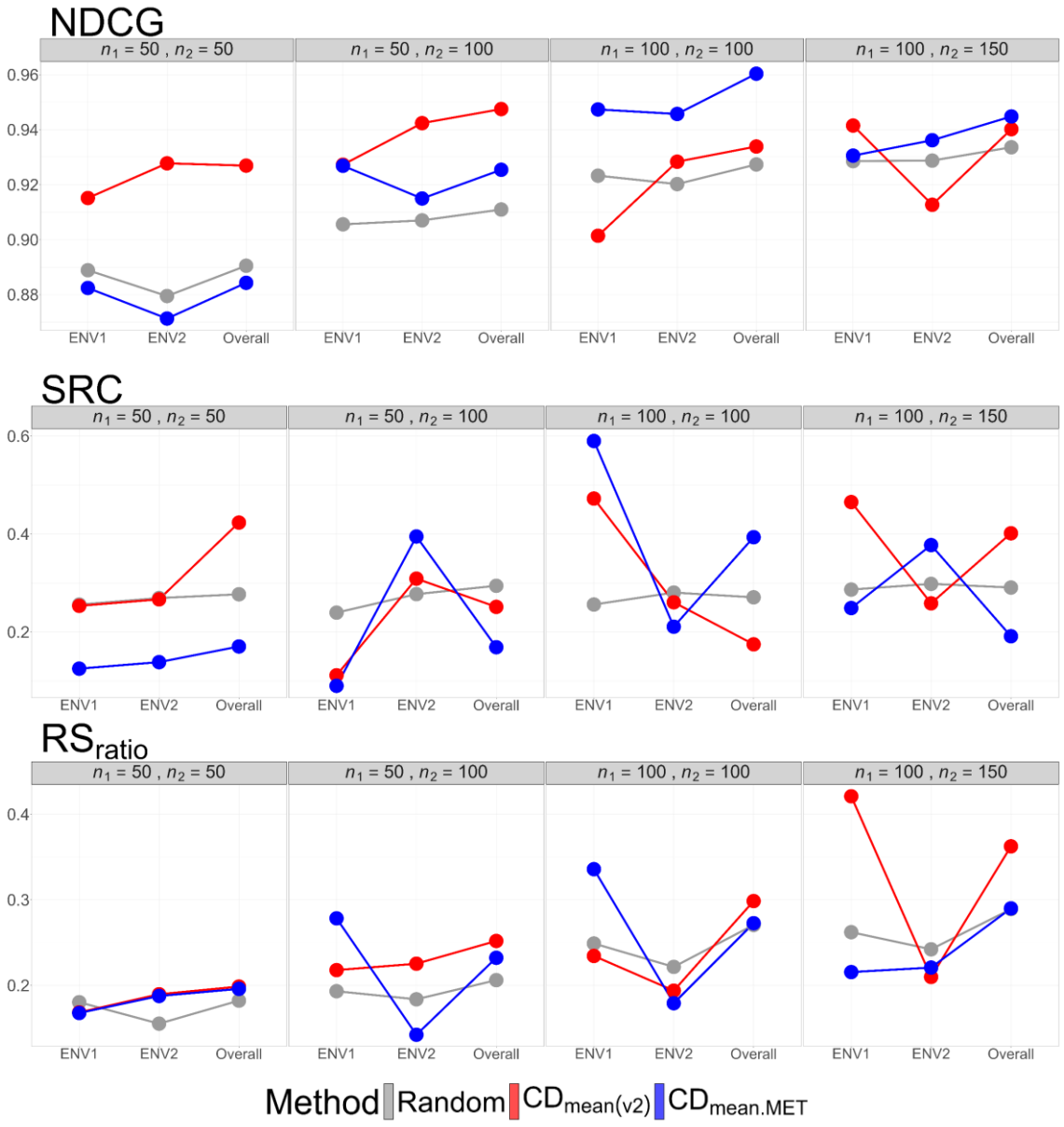

Figure 4.10. The means of the NDCG, SRC and RS_{ratio} value for the real data of the barley dataset.

doi:10.6342/NTU202501399

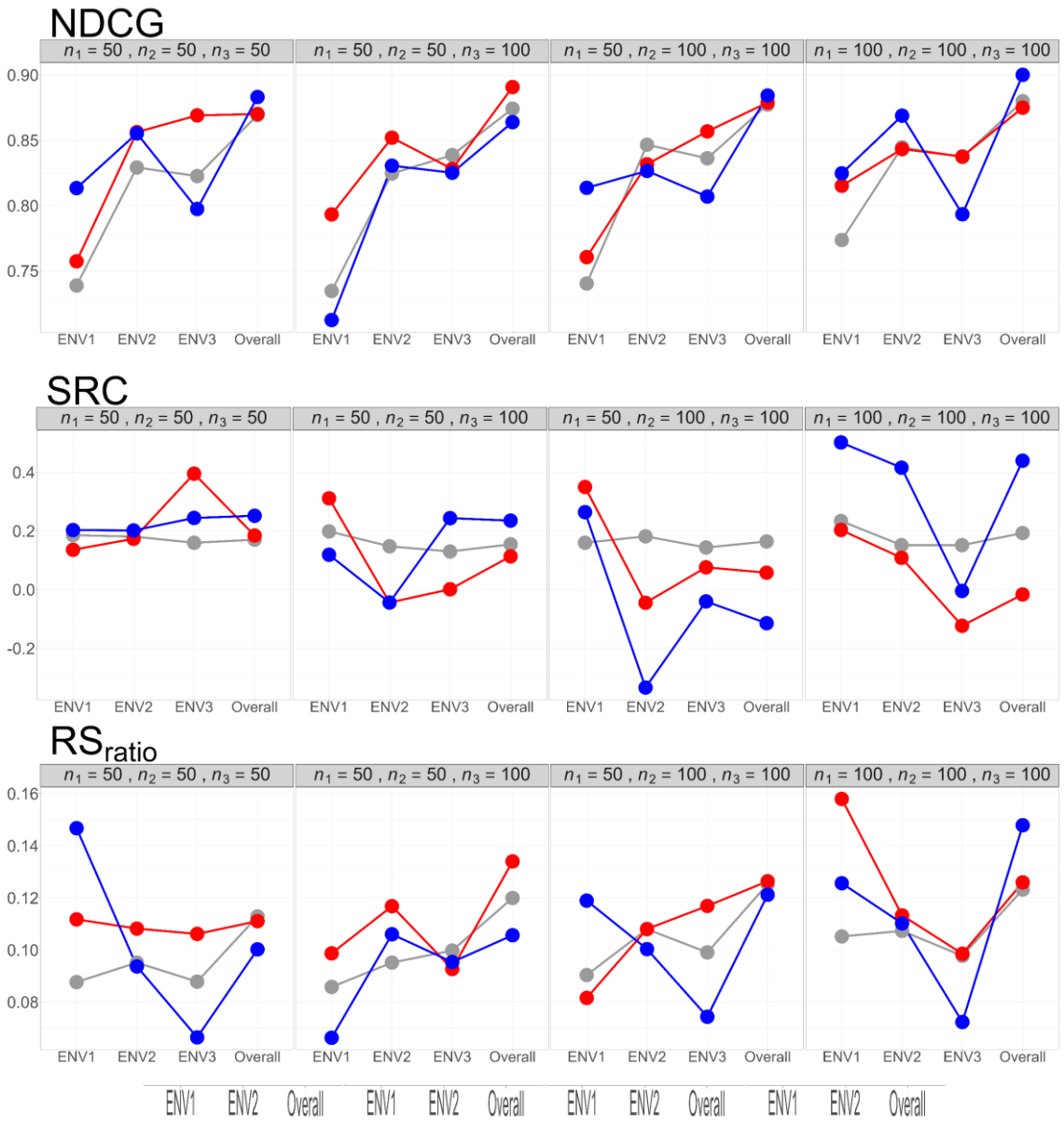

Figure 4.11. The means of the NDCG, SRC and RS_{ratio} value for the real data of the DST2 maize dataset.

doi:10.6342/NTU202501399

Chapter 5 Discussion

5.1 The performance of the three evaluation metrics

From the results of the simulation studies, the main results could be interpreted as follows. The difference between the training set determining methods varied among datasets. The results of the tropical rice dataset displayed significant increase by using the CD-based methods instead of the random sampling method (Figure 4.1- Figure 4.3). This result is consistent with Chen *et al.* (2024), as the tropical rice dataset showed significant advantage of the CD-based methods in the simulation results. However, the results of the barley dataset and the DST2 maize dataset had less improvement by the CD-based methods (Figure 4.4 - Figure 4.9).

Nonetheless, the performance of the CD-based methods had non-inferior values in almost all of the situations. Only the $CD_{mean.MET}$ method had lower mean value for the $SRC@k$ metric in the barley and DST2 maize dataset. The $CD_{mean(v2)}$ method, on the other hand, showed higher overall values in all evaluation metrics among all datasets. Therefore, the proposed $CD_{mean(v2)}$ criterion is recommended to evaluate the ability in predicting elite varieties in METs of a training set.

The performance of the three evaluation metrics could be deeply affected by the variance components. As the genetic variance ($\sigma^2_{G\times2}$) increased in ENV2, the NDCG value of that environment dropped. The increasing genetic variance also led to the decrease of the correlation of genetic effects between environments (ρ), causing the overall performance in every evaluation metric to drop. Furthermore, the increasing training set size also led to better performance of all the evaluation metrics.

The difference between the construction methods in NDCG was by far smaller than

doi:10.6342/NTU202501399

that in SRC and RS_{ratio}, often rounded to two or three decimal places. This was mainly because the value of NDCG is calculated by the TBVs rather than the ranks. In order to obtain non-negative simulated TBVs, the population mean of each environment was set to be larger than 100. However, this would cause the variation of the simulated TBVs to be smaller. The generated TBV values were therefore very close, resulting in close NDCG values in comparison to the two other evaluation metrics.

5.2 Training sets determined by CD criteria have high r^2

Although only the tropical rice dataset showed significant improvement in the evaluation metrics using the CD-based method, all three datasets showed better performance in r^2 using the CD-based method (Figure 5.1 - Figure 5.3). This was because the generalized coefficient of determination was the theorized value of r^2. The two CD criteria both directly derived from the generalized CD, and could therefore be applied in GA to maximize the r^2 value between the TBVs and the GEBVs of a training set. This result was consistent with Chen *et al.* (2024) who proposed $CD_{mean(v2)}$ and Rio *et al.* (2022) who applied the CD criterion to optimizing training sets in GBLUP models for METs.

5.3 Robustness of CD criteria against parameters

As described in Chapter 3.2, the variance components $\sigma^2_{G\times j}$ and σ^2_{Ej} were all fixed at 1 for $j = 1, 2 \ldots\ldots, T$, and the covariance of σ^2_G was fixed at 0.5 in Eq. (9) and (10) when calculating the values of both CD criteria. In order to check whether the values of these variance components could affect the outcoming training set, a comparison study was conducted with 20 training sets of the size $n_1 = 50$ and $n_2 = 50$ being randomly sampled from the tropical rice dataset. Six sets of variance components (Table 5.1) were

doi:10.6342/NTU202501399

applied to calculate the values of both CD criteria of the 20 training sets. The resulting CD values were shown in Table 5.2 and Table 5.3, along with the Pearson's correlation (Cor) and the Spearman's rank correlation (SRC) between the default variance component setting and the remaining settings. All Cor and SRC values for both $CD_{mean(v2)}$ and $CD_{mean.MET}$ were above 0.9. This result was consistent with that showed in Ou and Liao (2019), which also displayed high correlation between various variance component settings. The high correlations between the default variance component setting and the other settings indicate that the optimized training set found by the default setting would also likely be the optimized training set under other settings.

5.4 Correlation of genetic effects between environments

As described in Rio *et al.* (2022), the correlation of genetic effects between environments ρ was used to quantify the level of G×E effect. Simulated scenarios with various ρ values were demonstrated to display the G×E phenomenon. Genetic values of the tropical rice dataset were simulated under various levels of ρ in two environments. The ρ values were fixed at 0, 0.5, and 1; the variance components were set at $\sigma^2_{G\times1} = 10$, $\sigma^2_{G\times2} = 20$, and σ^2_G was calculated by $\sigma^2_G = \rho\sqrt{(\sigma^2_{G\times1})(\sigma^2_{G\times2})}$ in Eq. (16). The resulting mean genetic values of 30 replicates were illustrated in Figure 5.4. In one of the extreme cases where $\rho = 1$, the G×E effect was weak, and the genetic values in environment 1 was completely linearly correlated to those in environment 2. Therefore, the elite varieties in both environments were the same, and it was easier for breeders to select superior and stable varieties. In the other extreme case where $\rho = 0$, the G×E effect was strong. Therefore, the elite varieties for each environment were different, and it was more suitable for breeders to select superior varieties in a specific environment.

doi:10.6342/NTU202501399

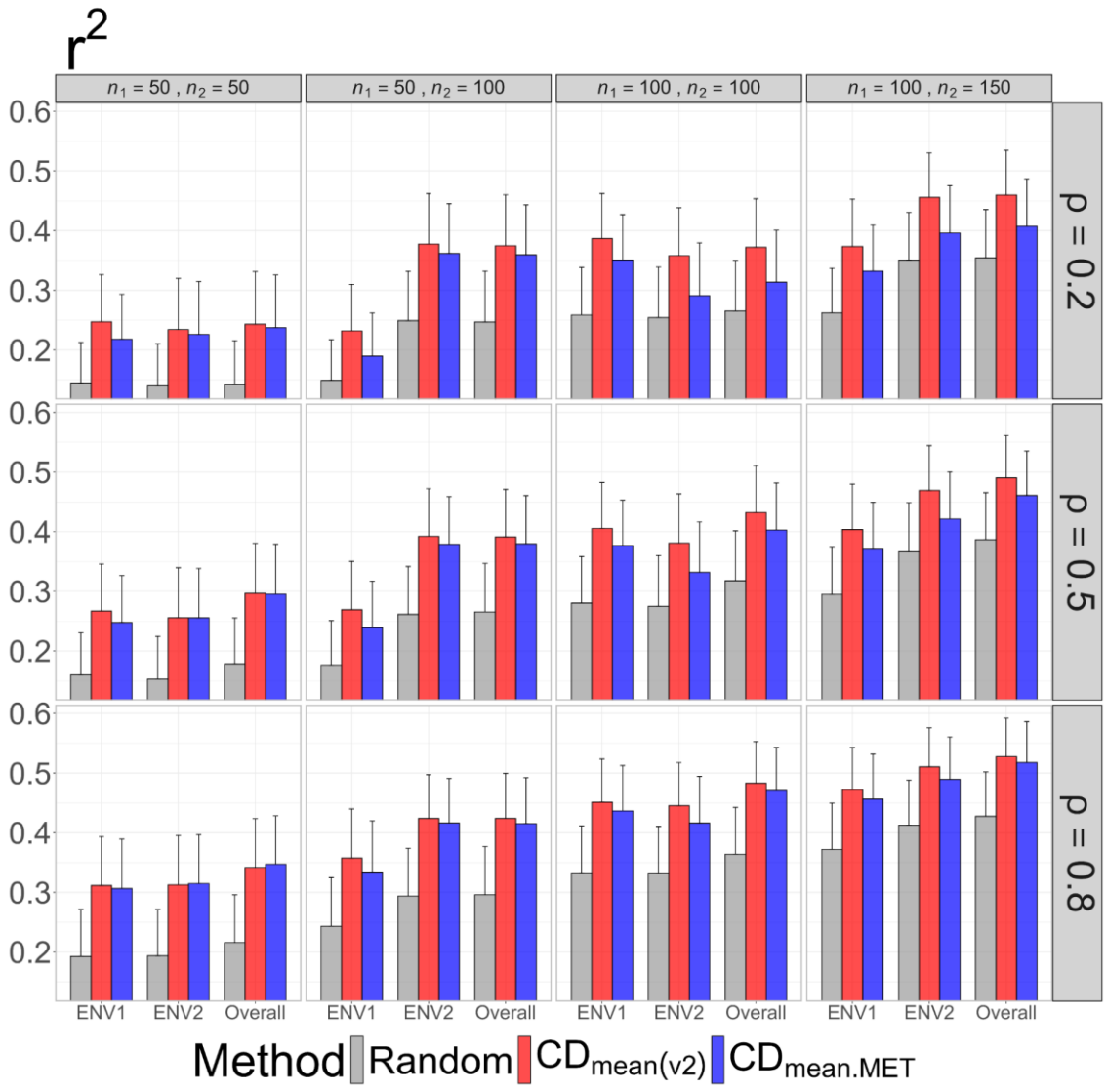

Figure 5.1. The means of the r^2 value for the simulated tropical rice dataset.

doi:10.6342/NTU202501399

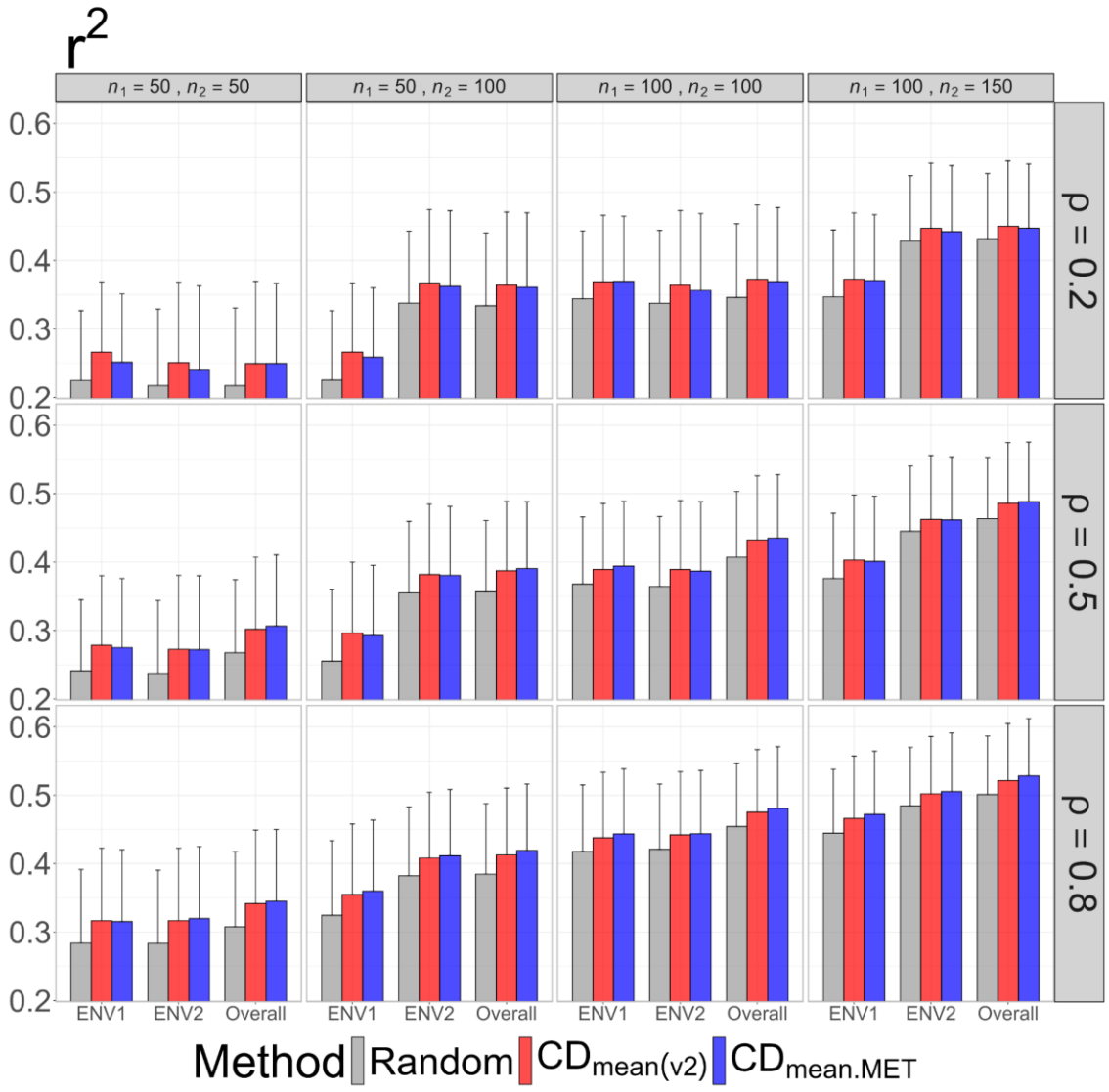

Figure 5.2. The means of the r^2 value for the simulated barley dataset.

doi:10.6342/NTU202501399

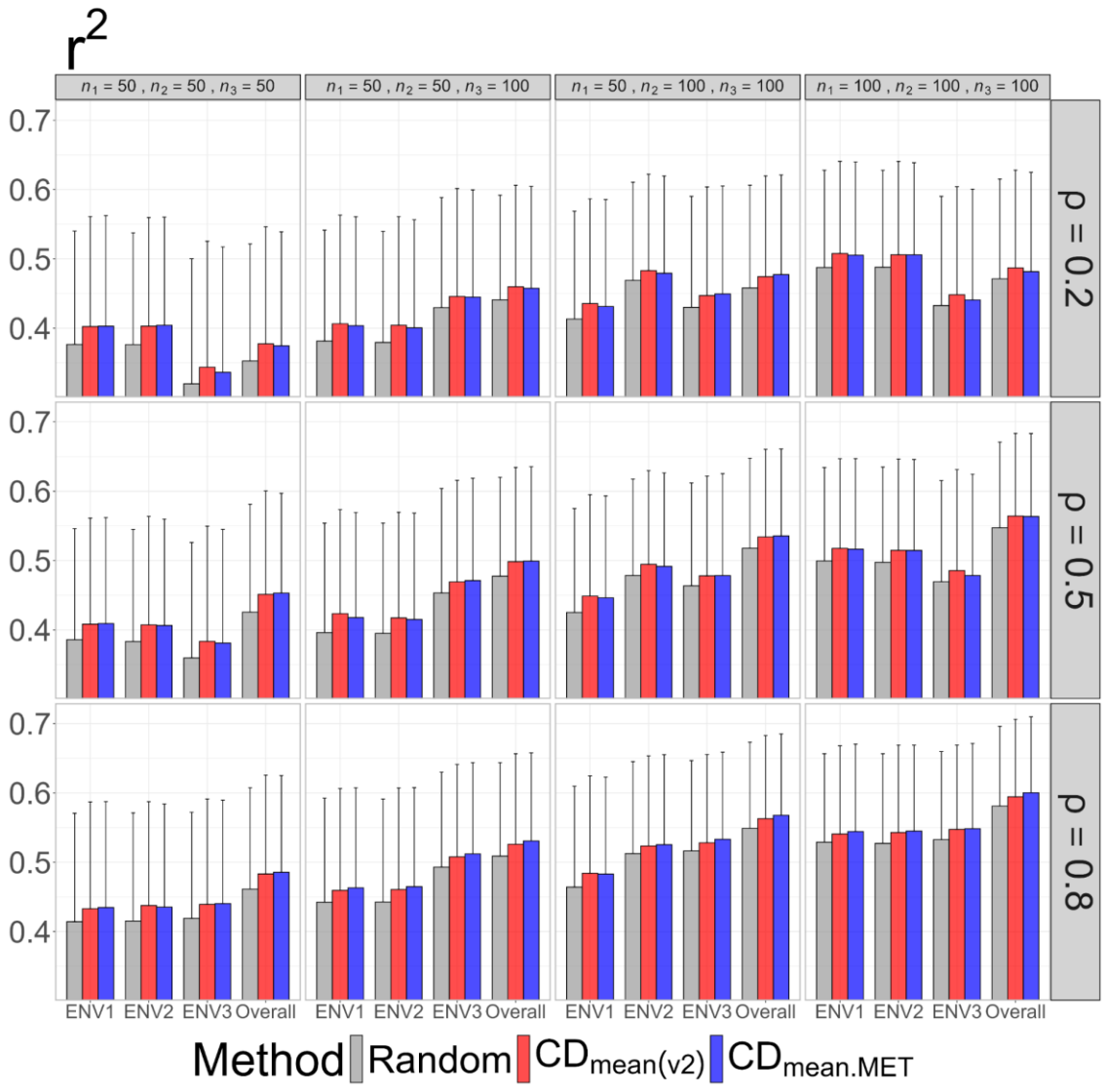

Figure 5.3. The means of the r^2 value for the simulated DST2 maize dataset.

doi:10.6342/NTU202501399

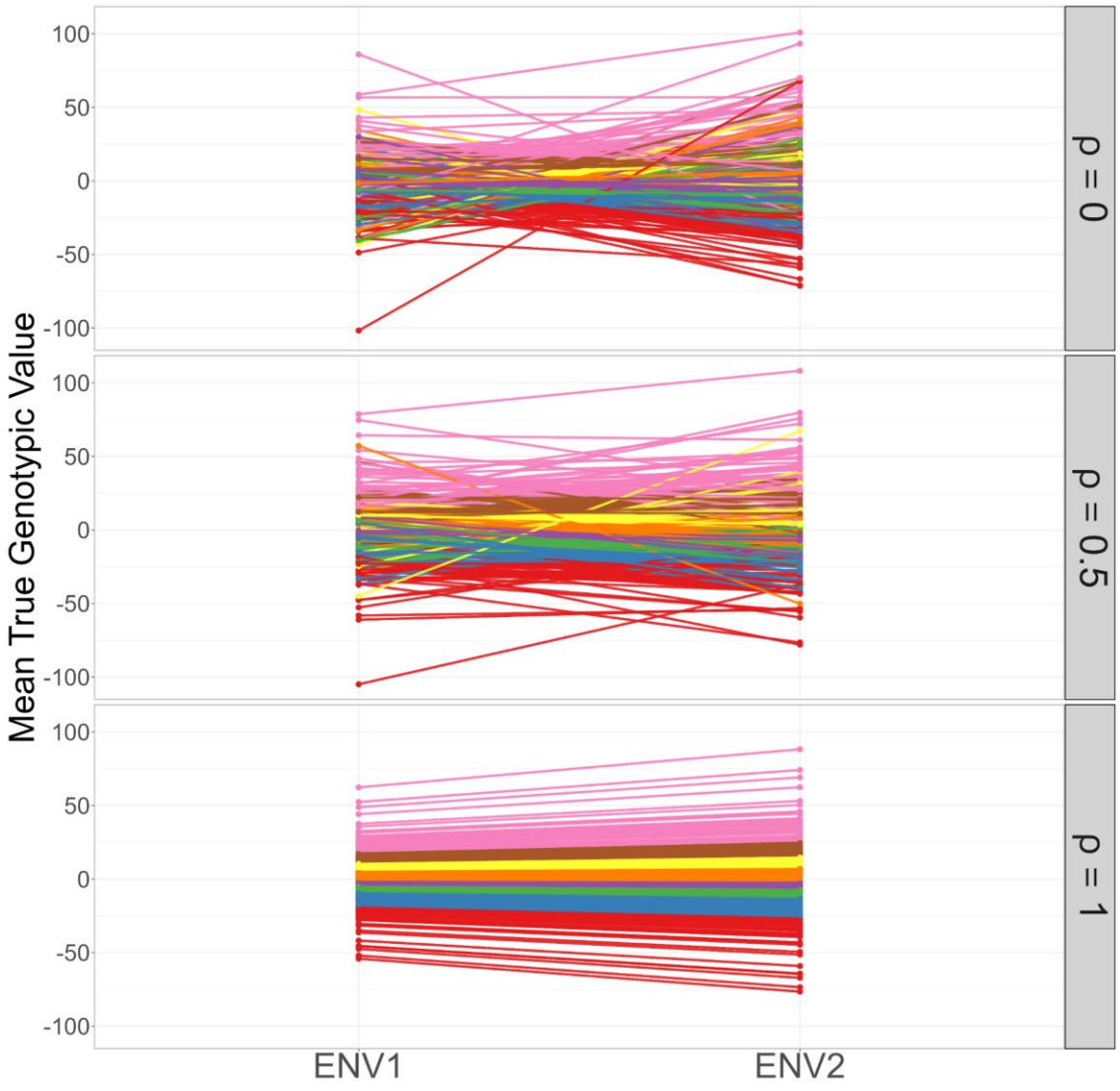

Figure 5.4. The true genetic effect under different levels of correlation between environments for the simulated tropical rice dataset.

doi:10.6342/NTU202501399

Table 5.1. The parameter sets given for the discussion about robustness of CD criteria

	Set1	Set2	Set3	Set4	Set5	Set6
$\sigma^2_{G\times1}$	1	1	1	1	1	1
$\sigma^2_{G\times2}$	1	1	1	2	5	1
σ^2_{G}	0.5	0.2	0.8	0.5	0.5	0.5
σ^2_{E1}	1	1	1	1	1	1
σ^2_{E2}	1	1	1	1	2	5

doi:10.6342/NTU202501399

Table 5.2. The performance of $CD_{mean(v2)}$ for the training sets under each parameter set

(a) The values and rank of $CD_{mean(v2)}$ under each parameter set

Design	Set1	Set2	Set3	Set4	Set5	Set6
D1	0.1873 (1)	0.1649 (1)	0.2075 (1)	0.199 (1)	0.1981 (1)	0.1346 (2)
D2	0.1859 (2)	0.1635 (3)	0.2063 (2)	0.1979 (2)	0.1970 (2)	0.1341 (3)
D3	0.1859 (3)	0.1638 (2)	0.2060 (3)	0.1967 (3)	0.1957 (3)	0.1366 (1)
D4	0.1824 (4)	0.1594 (5)	0.2034 (4)	0.1939 (4)	0.1927 (5)	0.1307 (8)
D5	0.1819 (5)	0.1595 (4)	0.2022 (7)	0.1938 (5)	0.1930 (4)	0.1317 (5)
D6	0.1816 (6)	0.1586 (6)	0.2028 (5)	0.1926 (7)	0.1913 (7)	0.1314 (6)
D7	0.1809 (7)	0.1575 (9)	0.2023 (6)	0.1929 (6)	0.1918 (6)	0.1284 (12)
D8	0.1809 (8)	0.1580 (8)	0.2020 (8)	0.1917 (8)	0.1904 (8)	0.1312 (7)
D9	0.1807 (9)	0.1585 (7)	0.2009 (10)	0.1910 (10)	0.1898 (10)	0.1319 (4)
D10	0.1803 (10)	0.1571 (10)	0.2016 (9)	0.1914 (9)	0.1901 (9)	0.1289 (10)
D11	0.1784 (11)	0.1551 (11)	0.1999 (12)	0.1893 (11)	0.1879 (11)	0.1296 (9)
D12	0.1781 (12)	0.1546 (12)	0.1999 (11)	0.1884 (13)	0.1867 (13)	0.1285 (11)
D13	0.1766 (13)	0.1527 (15)	0.1988 (13)	0.1885 (12)	0.1871 (12)	0.1250 (15)
D14	0.1764 (14)	0.1534 (13)	0.1977 (14)	0.1867 (15)	0.1851 (15)	0.1269 (13)
D15	0.1756 (15)	0.1528 (14)	0.1965 (15)	0.1873 (14)	0.1863 (14)	0.1256 (14)
D16	0.1719 (16)	0.1493 (16)	0.1926 (18)	0.1843 (16)	0.1835 (16)	0.1214 (18)
D17	0.1714 (17)	0.1483 (17)	0.1929 (16)	0.1819 (17)	0.1804 (18)	0.1231 (16)
D18	0.1709 (18)	0.1482 (18)	0.1919 (19)	0.1819 (18)	0.1805 (17)	0.1204 (19)
D19	0.1709 (19)	0.1477 (19)	0.1926 (17)	0.1814 (19)	0.1798 (19)	0.1230 (17)
D20	0.1647 (20)	0.1433 (20)	0.1845 (20)	0.1751 (20)	0.1739 (20)	0.1185 (20)

(b) Pearson's and Spearman's coefficient of correlation of $CD_{mean(v2)}$ values between Set1 (default setting) and all the other parameter sets

	$\rho_{1,2}$	$\rho_{1,3}$	$\rho_{1,4}$	$\rho_{1,5}$	$\rho_{1,6}$
Cor	0.9952	0.9939	0.9944	0.9908	0.9691
SRC	0.9880	0.9835	0.9880	0.9835	0.9429

doi:10.6342/NTU202501399

Table 5.3. The performance of $CD_{mean.MET}$ for the training sets under each parameter set

(a) The values and rank of $CD_{mean.MET}$ under each parameter set

Design	Set1	Set2	Set3	Set4	Set5	Set6
D1	0.2135 (1)	0.1803 (1)	0.2381 (1)	0.2256 (1)	0.2235 (1)	0.1566 (2)
D2	0.2120 (2)	0.1791 (2)	0.2365 (3)	0.2217 (4)	0.2170 (5)	0.1586 (1)
D3	0.2117 (3)	0.1786 (3)	0.2366 (2)	0.2238 (2)	0.2212 (2)	0.1556 (3)
D4	0.2097 (4)	0.1754 (4)	0.2355 (4)	0.2221 (3)	0.2197 (3)	0.1527 (8)
D5	0.2090 (5)	0.1745 (6)	0.2349 (5)	0.2195 (6)	0.2151 (8)	0.1536 (6)
D6	0.2084 (6)	0.1740 (7)	0.2343 (7)	0.2189 (9)	0.2143 (10)	0.1534 (7)
D7	0.2082 (7)	0.1734 (9)	0.2345 (6)	0.2211 (5)	0.2187 (4)	0.1503 (12)
D8	0.2079 (8)	0.1748 (5)	0.2326 (11)	0.2191 (7)	0.2152 (7)	0.1537 (5)
D9	0.2077 (9)	0.1730 (10)	0.2338 (8)	0.2191 (8)	0.2158 (6)	0.1510 (11)
D10	0.2069 (10)	0.1739 (8)	0.2316 (13)	0.2172 (11)	0.2139 (11)	0.1537 (4)
D11	0.2066 (11)	0.1715 (11)	0.2330 (10)	0.2157 (13)	0.2091 (15)	0.1525 (9)
D12	0.2065 (12)	0.1711 (12)	0.2333 (9)	0.2166 (12)	0.212 (12)	0.1511 (10)
D13	0.2052 (13)	0.1693 (14)	0.2325 (12)	0.2177 (10)	0.2146 (9)	0.1473 (14)
D14	0.2042 (14)	0.1696 (13)	0.2304 (14)	0.2153 (14)	0.2120 (13)	0.1490 (13)
D15	0.2020 (15)	0.1682 (15)	0.2275 (15)	0.2134 (15)	0.2098 (14)	0.1472 (15)
D16	0.1992 (16)	0.1640 (19)	0.2259 (16)	0.2087 (19)	0.2029 (19)	0.1451 (16)
D17	0.1991 (17)	0.1644 (16)	0.2255 (17)	0.2092 (18)	0.2043 (18)	0.1449 (17)
D18	0.1981 (18)	0.1640 (18)	0.2240 (18)	0.2106 (16)	0.2086 (16)	0.1415 (19)
D19	0.1974 (19)	0.1641 (17)	0.2227 (19)	0.2103 (17)	0.2082 (17)	0.1421 (18)
D20	0.1903 (20)	0.1582 (20)	0.2145 (20)	0.2004 (20)	0.1963 (20)	0.1390 (20)

(b) Pearson's and Spearman's coefficient of correlation of $CD_{mean.MET}$ values between Set1 (default setting) and all the other parameter sets

	$\rho_{1,2}$	$\rho_{1,3}$	$\rho_{1,4}$	$\rho_{1,5}$	$\rho_{1,6}$
Cor	0.9855	0.9889	0.9812	0.9304	0.9591
SRC	0.9729	0.9744	0.9594	0.9188	0.9203

doi:10.6342/NTU202501399

References

Akdemir, D., Sanchez, J. I., & Jannink, J.-L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genetics Selection Evolution*, *47*, 1-10.

Alves, F. C., Galli, G., Matias, F. I., Vidotti, M. S., Morosini, J. S., & Fritsche-Neto, R. (2021). Impact of the complexity of genotype by environment and dominance modeling on the predictive accuracy of maize hybrids in multi-environment prediction models. *Euphytica*, *217*(3).

Blondel, M., Onogi, A., Iwata, H., & Ueda, N. (2015). A Ranking Approach to Genomic Selection. *PLoS One*, *10*(6), e0128570.

Burgueño, J., Crossa, J., Cornelius, P. L., & Yang, R. C. (2008). Using factor analytic models for joining environments and genotypes without crossover genotype× environment interaction. *Crop Science*, *48*(4), 1291-1305.

Chen, S. P., Sung, W. H., & Liao, C. T. (2024). Constructing training sets for genomic selection to identify superior genotypes in candidate populations. *Theor Appl Genet*, *137*(12), 270.

Cockram, J., White, J., Zuluaga, D. L., Smith, D., Comadran, J., Macaulay, M., Luo, Z., Kearsey, M. J., Werner, P., Harrap, D., Tapsell, C., Liu, H., Hedley, P. E., Stein, N., Schulte, D., Steuernagel, B., Marshall, D. F., Thomas, W. T., Ramsay, L., . . . O'Sullivan, D. M. (2010). Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. *Proc Natl Acad Sci U S A*, *107*(50), 21611-21616.

Covarrubias-Pazaran, G. (2016). Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. *PLoS One*, *11*(6), e0156744.

Crossa, J., de los Campos, G., Maccaferri, M., Tuberosa, R., Burgueño, J., & Pérez-Rodríguez, P. (2016). Extending the marker× environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. *Crop Science*, *56*(5), 2193-2209.

Cullis, B. R., Smith, A. B., Cocks, N. A., & Butler, D. G. (2020). The design of early-stage plant breeding trials using genetic relatedness. *Journal of Agricultural, Biological and Environmental Statistics*, *25*, 553-578.

Fernández-González, J., Akdemir, D., & Isidro y Sánchez, J. (2023). A comparison of methods for training population optimization in genomic selection. *Theoretical and Applied Genetics*, *136*(3), 30.

Finlay, K., & Wilkinson, G. (1963). The analysis of adaptation in a plant-breeding programme. *Australian journal of agricultural research*, *14*(6), 742-754.

doi:10.6342/NTU202501399

Gauch Jr, H. G. (1988). Model selection and validation for yield trials with interaction. *Biometrics*, 705-715.

Heffner, E. L., Lorenz, A. J., Jannink, J. L., & Sorrells, M. E. (2010). Plant breeding with genomic selection: gain per unit time and cost. *Crop Science*, *50*(5), 1681-1690.

Henderson, C. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 423-447.

Henderson, C. (1977). Best linear unbiased prediction of breeding values not in the model for records. *Journal of Dairy Science*, *60*(5), 783-787.

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*, 463-480.

Holland, J. H. (1992). Genetic Algorithms. *Scientific American*, *267*, 66-73.

Isidro, J., Jannink, J. L., Akdemir, D., Poland, J., Heslot, N., & Sorrells, M. E. (2015). Training set optimization under population structure in genomic selection. *Theor Appl Genet*, *128*(1), 145-158.

Jarquín, D., Howard, R., Crossa, J., Beyene, Y., Gowda, M., Martini, J. W. R., Covarrubias Pazaran, G., Burgueno, J., Pacheco, A., Grondona, M., Wimmer, V., & Prasanna, B. M. (2020). Genomic Prediction Enhanced Sparse Testing for Multi-environment Trials. *G3 (Bethesda)*, *10*(8), 2725-2739.

Laloë, D. (1993). Precision and information in linear models of genetic evaluation. *Genetics Selection Evolution*, *25*(6), 557-576.

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J. L., Singh, R. P., Autrique, E., & de los Campos, G. (2015). Increased prediction accuracy in wheat breeding trials using a marker x environment interaction genomic selection model. *G3 (Bethesda)*, *5*(4), 569-582.

Malosetti, M., Bustos-Korts, D., Boer, M. P., & van Eeuwijk, F. A. (2016). Predicting responses in multiple environments: issues in relation to genotype× environment interactions. *Crop Science*, *56*(5), 2210-2222.

Meuwissen, T. H., Hayes, B. J., & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819-1829.

Oakey, H., Cullis, B., Thompson, R., Comadran, J., Halpin, C., & Waugh, R. (2016). Genomic Selection in Multi-environment Crop Trials. *G3 (Bethesda)*, *6*(5), 1313-1326.

Ou, J. H., & Liao, C. T. (2019). Training set determination for genomic selection. *Theor Appl Genet*, *132*(10), 2781-2792.

doi:10.6342/NTU202501399

Rincent, R., Kuhn, E., Monod, H., Oury, F. X., Rousset, M., Allard, V., & Le Gouis, J. (2017). Optimization of multi-environment trials for genomic selection based on crop models. *Theor Appl Genet*, *130*(8), 1735-1752.

Rincent, R., Laloe, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., Rodriguez, V. M., Moreno-Gonzalez, J., Melchinger, A., Bauer, E., Schoen, C. C., Meyer, N., Giauffret, C., Bauland, C., Jamin, P., Laborde, J., Monod, H., Flament, P., Charcosset, A., & Moreau, L. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). *Genetics*, *192*(2), 715-728.

Rio, S., Akdemir, D., Carvalho, T., & Sanchez, J. I. Y. (2022). Assessment of genomic prediction reliability and optimization of experimental designs in multi-environment trials. *Theor Appl Genet*, *135*(2), 405-419.

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72-101.

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redona, E., Atlin, G., Jannink, J. L., & McCouch, S. R. (2015). Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genet*, *11*(2), e1004982.

van Eeuwijk, F. A., Bustos-Korts, D. V., & Malosetti, M. (2016). What Should Students in Plant Breeding Know About the Statistical Aspects of Genotype × Environment Interactions? *Crop Science*, *56*(5), 2119-2140.

Venables, B., & Ripley, B. (2002). Modern Applied Statistics With S. In (4 ed.).

Wu, P.-Y., Ou, J.-H., & Liao, C.-T. (2023). Sample size determination for training set optimization in genomic prediction. *Theoretical and Applied Genetics*, *136*(3), 57.

Wu, P. Y., Tung, C. W., Lee, C. Y., & Liao, C. T. (2019). Genomic Prediction of Pumpkin Hybrid Performance. *Plant Genome*, *12*(2).

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

Yan, W., Kang, M. S., Ma, B., Woods, S., & Cornelius, P. L. (2007). GGE biplot vs. AMMI analysis of genotype-by-environment data. *Crop Science*, *47*(2), 643-653.

doi:10.6342/NTU202501399

Appendix A – $Var(\hat{g}_c)$ and $Cov(g_c, \hat{g}_c)$ are equivalent mathematically

Consider the MGE model in Eq. (5), described as:

$$\boldsymbol{y}_t = \boldsymbol{\mu}_t + \boldsymbol{g}_t + \boldsymbol{e}_t.$$

The total genetic effect $\boldsymbol{g}_t$ and residual effect $\boldsymbol{e}_t$ are assumed to be independent and follow multivariate normal distributions respectively:

$$\boldsymbol{g}_t \sim MVN(0, \boldsymbol{G}_t), \boldsymbol{G}_t = \begin{bmatrix} \sigma^2_{G\times 1}\boldsymbol{K}_1 & \cdots & \sigma^2_G \boldsymbol{K}_{1T} \\ \vdots & \ddots & \vdots \\ \sigma^2_G \boldsymbol{K}_{T1} & \cdots & \sigma^2_{G\times T}\boldsymbol{K}_T \end{bmatrix};$$

$$\boldsymbol{e}_t \sim MVN(0, \boldsymbol{R}_E), \boldsymbol{R}_E = \begin{bmatrix} \sigma^2_{E1}\boldsymbol{I}_{n_1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \sigma^2_{ET}\boldsymbol{I}_{n_T} \end{bmatrix}.$$

Therefore, the variance of the phenotype $\boldsymbol{y}_t$ of the training set could be obtained by:

$$Var(\boldsymbol{y}_t) = Var(\boldsymbol{\mu}_t + \boldsymbol{g}_t + \boldsymbol{e}_t) = Var(\boldsymbol{g}_t) + Var(\boldsymbol{e}_t) = \boldsymbol{G}_t + \boldsymbol{R}_E.$$

A block diagonal matrix $\boldsymbol{M}_t$ is given by:

$$\boldsymbol{M}_t = \begin{bmatrix} (\sigma^2_{E1})^{-1}(\boldsymbol{I}_{n_1} - \bar{\boldsymbol{J}}_{n_1}) & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & (\sigma^2_{ET})^{-1}(\boldsymbol{I}_{n_T} - \bar{\boldsymbol{J}}_{n_T}) \end{bmatrix},$$

where its j^{th} diagonal sub-matrix equal to $(\sigma^2_{Ej})^{-1}(\boldsymbol{I}_{n_j} - \bar{\boldsymbol{J}}_{n_j})$, $j = 1, 2, \ldots\ldots, T$. Each sub-matrix is orthogonal to the vector space spanned by the corresponding j^{th} population mean sub-vector $\mu_j \boldsymbol{1}_{n_j}$ in $\boldsymbol{\mu}_t$. $\bar{\boldsymbol{J}}_{n_j}$ is a matrix of order n_j with every element equal to $1/n_j$. Here, $\boldsymbol{M}_t$ has properties that $\boldsymbol{M}_t \boldsymbol{1}_{n_t} = \boldsymbol{0}$, and that $(\sigma^2_{Ej})^{-1}(\boldsymbol{I}_{n_j} - \bar{\boldsymbol{J}}_{n_j})\ \bar{\boldsymbol{J}}_{n_j} = \boldsymbol{0}$.

doi:10.6342/NTU202501399

By Henderson's mixed model equations (Henderson, 1975), the BLUP for the total genetic effects $\boldsymbol{g}_t$ could be expressed as:

$$\hat{\boldsymbol{g}}_t = (\boldsymbol{M}_t + \boldsymbol{G}_t^{-1})^{-1}\boldsymbol{M}_t\boldsymbol{y}_t,$$

and from Henderson (1977), the BLUP for $\boldsymbol{g}_c$ could be obtained by:

$$\hat{\boldsymbol{g}}_c = \boldsymbol{G}_{ct}\boldsymbol{G}_t^{-1}\hat{\boldsymbol{g}}_t \;\; = \boldsymbol{G}_{ct}(\boldsymbol{M}_t\boldsymbol{G}_t + \boldsymbol{I}_{n_t})^{-1}\boldsymbol{M}_t\boldsymbol{y}_t.$$

Hence, $Var(\hat{\boldsymbol{g}}_c)$ and $Cov(\boldsymbol{g}_c, \hat{\boldsymbol{g}}_c)$ could be proven equivalent as follows:

$$\begin{aligned}
Var(\hat{\boldsymbol{g}}_c) &= \boldsymbol{G}_{ct}\left(\boldsymbol{M}_t\boldsymbol{G}_t + \boldsymbol{I}_{n_t}\right)^{-1}\boldsymbol{M}_t[Var(\boldsymbol{y}_t)]\boldsymbol{M}_t\left(\boldsymbol{G}_t\boldsymbol{M}_t + \boldsymbol{I}_{n_t}\right)^{-1}\boldsymbol{G}_{ct}^T \\
&= \boldsymbol{G}_{ct}\left(\boldsymbol{M}_t\boldsymbol{G}_t + \boldsymbol{I}_{n_t}\right)^{-1}\boldsymbol{M}_t[\boldsymbol{G}_t + \boldsymbol{R}_E]\;\boldsymbol{M}_t\left(\boldsymbol{G}_t\boldsymbol{M}_t + \boldsymbol{I}_{n_t}\right)^{-1}\boldsymbol{G}_{ct}^T \\
&= \boldsymbol{G}_{ct}\left(\boldsymbol{M}_t\boldsymbol{G}_t + \boldsymbol{I}_{n_t}\right)^{-1}\boldsymbol{M}_t(\boldsymbol{G}_t\boldsymbol{M}_t + \boldsymbol{R}_E\boldsymbol{M}_t)\left(\boldsymbol{G}_t\boldsymbol{M}_t + \boldsymbol{I}_{n_t}\right)^{-1}\boldsymbol{G}_{ct}^T \\
&= \boldsymbol{G}_{ct}\left(\boldsymbol{M}_t\boldsymbol{G}_t + \boldsymbol{I}_{n_t}\right)^{-1}\boldsymbol{M}_t[\left(\boldsymbol{G}_t\boldsymbol{M}_t + \boldsymbol{I}_{n_t}\right) - \begin{bmatrix} \bar{\boldsymbol{J}}_{n_1} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \bar{\boldsymbol{J}}_{n_T} \end{bmatrix}]\left(\boldsymbol{G}_t\boldsymbol{M}_t + \boldsymbol{I}_{n_t}\right)^{-1}\boldsymbol{G}_{ct}^T \\
&= \boldsymbol{G}_{ct}\left(\boldsymbol{M}_t\boldsymbol{G}_t + \boldsymbol{I}_{n_t}\right)^{-1}\boldsymbol{M}_t[\left(\boldsymbol{G}_t\boldsymbol{M}_t + \boldsymbol{I}_{n_t}\right)]\left(\boldsymbol{G}_t\boldsymbol{M}_t + \boldsymbol{I}_{n_t}\right)^{-1}\boldsymbol{G}_{ct}^T \\
&= \boldsymbol{G}_{ct}\left(\boldsymbol{M}_t\boldsymbol{G}_t + \boldsymbol{I}_{n_t}\right)^{-1}\boldsymbol{M}_t\boldsymbol{G}_{ct}^T \\
Cov(\boldsymbol{g}_c, \hat{\boldsymbol{g}}_c) &= Cov[\boldsymbol{g}_c, \boldsymbol{G}_{ct}\left(\boldsymbol{M}_t\boldsymbol{G}_t + \boldsymbol{I}_{n_t}\right)^{-1}\boldsymbol{M}_t\boldsymbol{y}_t] \\
&= \boldsymbol{G}_{ct}\left(\boldsymbol{M}_t\boldsymbol{G}_t + \boldsymbol{I}_{n_t}\right)^{-1}\boldsymbol{M}_t Cov(\boldsymbol{g}_c, \boldsymbol{y}_t) \\
&= \boldsymbol{G}_{ct}\left(\boldsymbol{M}_t\boldsymbol{G}_t + \boldsymbol{I}_{n_t}\right)^{-1}\boldsymbol{M}_t Cov[\boldsymbol{g}_c, (\boldsymbol{\mu}_t + \boldsymbol{g}_t + \boldsymbol{e}_t)] \\
&= \boldsymbol{G}_{ct}\left(\boldsymbol{M}_t\boldsymbol{G}_t + \boldsymbol{I}_{n_t}\right)^{-1}\boldsymbol{M}_t Cov(\boldsymbol{g}_c, \boldsymbol{g}_t) \\
&= \boldsymbol{G}_{ct}\left(\boldsymbol{M}_t\boldsymbol{G}_t + \boldsymbol{I}_{n_t}\right)^{-1}\boldsymbol{M}_t\boldsymbol{G}_{ct}^T
\end{aligned}$$

This completes the proof that $Var(\hat{\boldsymbol{g}}_c)$ and $Cov(\boldsymbol{g}_c, \hat{\boldsymbol{g}}_c)$ are equivalent.

doi:10.6342/NTU202501399

Appendix B – Supplementary Materials

The supplementary figures and tables of the all three datasets analyzed in this study are freely accessible. The supplementary files can be downloaded from the Figshare webpages: (https://doi.org/10.6084/m9.figshare.29432471.v1) for those of the simulation studies and (https://doi.org/10.6084/m9.figshare.29432477.v1) for those of the real data analyses, respectively.

The original data includes the phenotype data and the kinship matrix obtained by Eq. (2). The original data of the three datasets could be retrieved from the Figshare webpage: (https://doi.org/10.6084/m9.figshare.29453807.v1).

All of the methods in this study are conducted in R language. The R code implemented in this research could be retrieved from the GitHub webpage: (https://github.com/simonb08601003/Training-set-optimization-in-genomic-selection-for-multi-environment-trials).

doi:10.6342/NTU202501399