Linköping University | Department of Computer and Information Science
Master's thesis, 30 ECTS | Statistics and Machine Learning
2022 | LIU-IDA/LITH-EX-A--22/001--SE

Optimal designs for sub-regions' effects in multi-environment crop variety testing

Biswas Kumar

Supervisor : Maryna Prus
Examiner : Frank Miller

LINKÖPING UNIVERSITY

Linköpings universitet
SE–581 83 Linköping
+46 13 28 10 00 , www.liu.se

Abstract

Growing population, limited cultivable land size, and climate change are the major challenges that the world is facing at present. To counter these issues, the United Nations has advocated the need for genetic diversity in agriculture. In order to provide a credible recommendation to the farmers, the crops must be thoroughly tested in multi-environment trials to obtain trustworthy empirical support. Since the environment is a large, complex and diverse set-up, it could be logically divided into smaller sub-regions based on their homogeneity between them. The allocation of crop trials in different sub-regions could be further determined by optimal design criterion, which estimates optimal designs to ensure estimation of parameters without bias and with minimum variance.

A hierarchical linear mixed model which involves sub-region trials was considered. The best linear unbiased estimator and its covariance matrix that measure the changes of variables together, for the sub-regions' effects were obtained. The standard and weighted A-design criterion were formulated using approximate and exact methods. The real data example of the Indian maize zone was considered and its variances were used for the calculation of the optimal design. The OptimalDesign package in R was used for the implementation of optimal design, which illustrated that the standard A-criterion gives equal weightage to the sub-region for the allocation of trials. The weighted A-Criterion resulted in the allocation of trials to sub-region based on the value of their coefficients. The constraints-based examples were also performed, which highlights the effective consideration of limitations that may arise due to any practical issue or requirements. It was also observed that higher variance in the sub-region leads to lower allocation of trials which can assist in decision-making.

Acknowledgments

I would like to express my sincere gratitude to my thesis supervisor Maryna Prus for providing me with the opportunity to contribute towards the research in Optimal designs for sub-regions' effects in multi-environment crop variety testing, with this thesis. Her patience, guidance, enthusiasm, and suggestions helped me in all the time of research and writing this thesis.

Besides my supervisor, I would also like to thank my examiner Professor Frank Miller for his comments and suggestions during seminars and review meeting. His valuable feedback has improved the quality of the thesis. I would also like to thank my course leader, Associate Professor Oleg Sysoev, for his comments and the suggestions he provided all along and especially during our seminars, to meet the goals of the thesis. I cannot imagine better mentorship for my thesis.

I am also thankful to my opponent Hoda Fakharzadehjahromy for her constructive feedback during the meetings, especially during the review meeting.

Finally, I would like to thank my parents, my wife, my brother, and the rest of my family for their unconditional love and support, without which I could not have enrolled in the course and started my thesis work.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

Population growth is a critical challenge that is putting pressure on the cultivable land, which is limited in nature. Moreover, climate change negatively impacts the productivity and nutritional quality of crops. The world bank has highlighted that food security will be a severe challenge as the world needs to produce about 70% more food by 2050 to feed an estimated 9 billion people [1]. The United Nations agency has stressed the need for genetic diversity in agriculture to combat climate change. They have backed the enhancement of agricultural and food systems that can improve the livelihoods and health of people, thereby building healthier ecosystems. Farmers around the world constantly need the recommendation of new crop varieties that can perform and provide higher crop yields and increased resilience to climate-related risks.[2]

To provide a credible recommendation to the farmers, it is important that the crops are thoroughly tested in multi-environment trials to obtain trustworthy empirical support. To reduce the complexity of a large environment that is diverse, it seems logical to divide such an environment into sub-regions such that they have a similar environment within the sub-region but differences between them. Sub-regions could be identified as geographical areas within the environment wherein the yield of crop genotype (or crop varieties), and climate are more homogeneous than the overall target environment [3]. For instance, a country could be considered as an environment, and its area could be divided into smaller distinct parts (sub-regions). Locations are smaller parts within sub-regions and are homogeneous.

Once the sub-region is divided, the next question arises about the allocation of trials in the sub-regions. This calls for an optimal number of trials to produce a robust empirical basis for the recommendation of crop varieties to farmers.

1.2 Problem

Effects are the impacts or differences. There are two significant effects (random and fixed) that influence the outcome of the design of allocation trials. The random effects are majorly attributed to the crop varieties, whereas the fixed effects aspects the sub-regions' effects. The

optimal allocation of trials has been recently discussed in Prus and Piepho [3] concerning the random genotype effects. However, the fixed sub-regions' effect and its optimal number of trials still need to be researched extensively as the related designs may differ from those for the genotype effects.

1.3 Goal

This study is focused on the sub-regions' effects, and therefore, the goal is to research about three specific questions:

1. Determine the best linear unbiased estimator and its covariance matrix for the sub-regions' effects in the underlined linear mixed model.
2. Formulate and analyze the related design criteria (standard and weighted A-criteria, which minimize the trace of the (adjusted) covariance matrix of the BLUE).
3. Compute optimal designs – optimal numbers of allocations to sub-regions – using the OptimalDesign package in R or analytically (in simple cases) for the proposed real data example.

1.4 Motivation

The solutions highlighted by the UN [2] was the primary motivation for conducting this thesis work. It provides me an opportunity to contribute to the recommendation of optimum crop variety using refined statistical models, thereby helping farmers and betterment of societies. Moreover, studying fixed and random effects gives me insight into the practical constraints while designing a real-life model.

2 Literature Review

Linear mixed models (LMM) are considered an extension of simple linear models (LM) to allow both fixed (repeatable, model population-average effects) and random effects. It is particularly used when there is non-independence in the data, such as arising from a hierarchical structure. For example, locations are nested within sub-regions. The solutions associated with the fixed effects are called Best Linear Unbiased Estimate (BLUE), whereas the solutions for random effects are known as Best Linear Unbiased Prediction (BLUP) [4].

In 2017, Isik et al.[5] compared the traditional analysis of variance (ANOVA) based on ordinary least squares (OLS) methods to mixed models. They found that under certain conditions, results from ANOVA and mixed models analysis are largely equivalent. Still, the mixed model is usually the best fit when data is limited or has missing values. They further stated that randomized field trials are regularly conducted to evaluate the performance of new crop varieties.

In 1975, C. R. Henderson[6] developed mixed model equations for random effects to obtain BLUPs of animals' breeding values. He was well known for the development of methods for the estimation of variance components in unbalanced data settings of mixed models. He was regarded as a pioneer in animal breeding for his contributions and had effectively utilized his mixed model equations.

In 2005, Piepho and Möhring[7] stated that if sufficient genotype and sub-region interactions are available wherein the sub-region is part of testing environments, then neighboring sub-regions can be exploited for more precise estimates. Moreover, the genotypic mean estimate for the whole target region can be used for the balanced data, which predicts better than simple means per sub-region. They also proposed a method based on BLUP that allows a weighted combination of data from several sub-regions and compared that method to other estimators.

In 2013, Kleinknecht et al [8] worked on finding a solution for selecting the best genotype per zone and borrowing information across zones to improve accuracy. They analyzed data using mixed models for the correlation of genetic effects between zones and data per zones. They concluded that covariance could enhance estimation accuracy, and hence mean perfor-

mance of the sub-region should be considered.

In 2021, Prus and Piepho[3] proposed a solution to the allocation of trials to the different sub-region. They used a LMM and illustrated that the optimal allocation depends on the variance–covariance structure for genotypic effects nested within sub-regions. Furthermore, they proposed an analytical approach for the computation of optimal designs for BLUP of genotype effects which is random, and illustrate the obtained results by a real data example from an Indian nationwide maize variety.

The proposed thesis study shall utilize the LMM mentioned in Prus and Piepho's 2021 paper and will try to estimate covariance of BLUE for sub-regions effects of fixed nature for the same data example from the Indian nation-wide maize variety. In addition, the computation of optimal designs for BLUE shall also be conducted.

3 Model Specification

3.1 Linear Mixed Models

LMM contains both fixed and random effects. Fixed effects are the parameters that stay constant and do not vary with observations or individuals. In simple words, it means that any change they cause to an individual is the same across the group, such as age and sub-region effects [9]. On the other hand, random effects are the parameters that change with each sampling. This highlights that during sampling such mixed models, fixed effects will return consistent results while random effects will return variable results.

The generic expression for the LMM can be represented as follows [6]:

$$y = X\beta + Zu + e \tag{3.1}$$

where, y is a $n \times 1$ observational vector, X is a known $n \times p$ matrix, β is a $p \times 1$ unknown fixed vector, Z is a known $n \times q$ matrix, u is a $q \times 1$ non-observable random vector with null means and, e is a $n \times 1$ observational error vector with null means.

$\text{Cov} \begin{pmatrix} u \\ e \end{pmatrix} = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix}$

where,
G and R are both nonsingular.
As u and e are two uncorrelated random vectors, the expression can be simplified to

$$Cov(u) = G, Cov(e) = R. \tag{3.2}$$

3.2 Linear Mixed Model - The Specific Model

In our study, sub-regions could be treated as fixed effect parameter as it is homogeneous in nature. Therefore, the conditions for the sub-region are considered to be constant for all the observations. On the contrary, locations could be considered random effect parameters as it is nested within sub-regions and can be sampled [3].

LMM is particularly used when there is non-independence in the data, such as arises from a hierarchical structure. In addition to fixed and random effects, the nested location data used

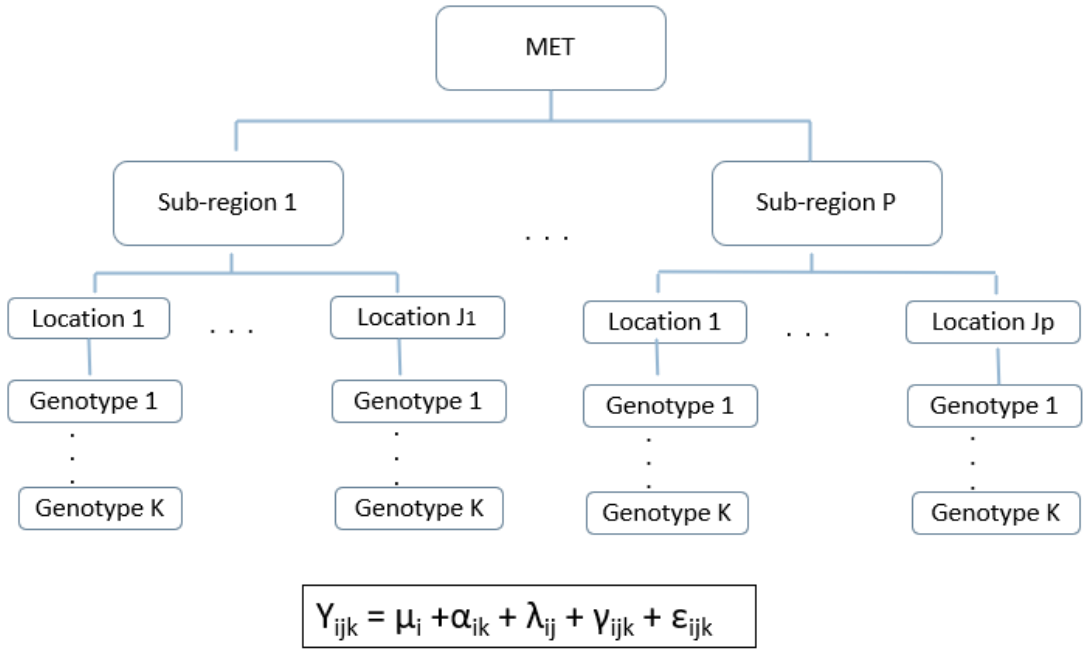
in our study is hierarchical, which also advocates for the usage of LMM [4].

Prus and Piepho [3] has highlighted an expression for the LMM, which involves sub-region trials that includes multiple replications. For simplicity, the proposed model for this thesis is the adjusted model with single replication, wherein we consider P sub-regions with nested locations such that $J = \sum_{i=1}^{P} J_i$, then the observation Y for K crop varieties in each location can be expressed as :

$$Y_{ijk} = \mu_i + \alpha_{ik} + \lambda_{ij} + \gamma_{ijk} + \epsilon_{ijk} \tag{3.3}$$

where, i = 1,...,P (number of sub-regions), j = 1,...,Ji (number of locations in i-th sub-region), k = 1,...,K (number of genotypes in each location). The variables such as μ_i is the fixed effect of i-th sub-region, α_{ik} highlights interaction effect of genotype k in sub-region i, λ_{ij} describes the effect of j-th location within i-th sub-region,γ_{ijk} highlights the effect of the k-th genotype in the j-th location within the i-th sub-region, and ϵ_{ijk} is observational error. The variance for location effect is given by $\text{var}(\lambda_{ij}) = \sigma^2_\lambda$, the variance for genotype effect is given by $\text{var}(\gamma_{ijk}) = \sigma^2_\gamma$ and for observational error $\text{var}(\epsilon_{ijk}) = \sigma^2$. The $Cov(\alpha_k) = \sigma^2 D$, where D is positive definite and $\alpha_k = (\alpha_{1k}, ..., \alpha_{Pk})^T$. It may be noted that random effects and experimental errors have zero mean and are uncorrelated [3].

Figure 3.1: The schematic diagram of the proposed model (highlighted by equation no.3.3)



The proposed model for equation 3.3 can be visualized in Figure 1. It represents the hierarchical structure and the relevant effects.

The original model of Prus and Piepho [3] also included replication l per genotypes in

each location and was given by :

$$Y_{ijkl} = \mu_i + \alpha_{ik} + \lambda_{ij} + \gamma_{ijk} + b_{ijl} + \epsilon_{ijkl} \tag{3.4}$$

which also includes b_{ijl} that highlights the effect of the l-th replication in location j in sub-region i.

4 Estimators

4.1 Best Linear Unbiased Estimator- BLUE

Isik et. al.[5] has noted that the fixed effects are known constants that will remain the same over repeated sampling. Fixed effects do not have a covariance structure, and examples of such fixed effects from a practical perspective include effects of different soil types, fertilizer treatments and environmental effects[10]. Isik et. al has further asserted that Best Linear Unbiased Estimator(BLUE) is the least square mean for the fixed effects. The Gauss-Markov theorem states that the ordinary least squares estimators have the lowest sampling variance within the class of linear unbiased estimators [11], and Christensen[12] proved that these Gauss-Markov theorem-based least squares estimates are the best linear unbiased estimates.

It is important to note that BLUE is not a method in itself but represents statistical properties of methods or solutions, which stands for[4]:
Best: minimum sampling variance what is being estimated
Linear: the estimates are linear functions of the observations
Unbiased: the expected value of the estimates is equal to their true values
Estimates: it refers to the algorithms that generate the estimated values

Henderson [6] gave a mixed model equation based on equation(3.1) and equation(3.2), which can generate both BLUE(β) and BLUP (u) as below:

$$\begin{bmatrix} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z + G^{-1} \end{bmatrix} \times \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^TR^{-1}y \\ Z^TR^{-1}y \end{bmatrix} \tag{4.1}$$

Equation(4.1) equates to :

$$X^TR^{-1}X\hat{\beta} + X^TR^{-1}Z\hat{u} = X^TR^{-1}y \tag{4.2}$$

$$Z^TR^{-1}X\hat{\beta} + (Z^TR^{-1}Z + G^{-1})\hat{u} = Z^TR^{-1}y \tag{4.3}$$

On solving equation (4.3) for $\hat{\mu}$, we get:

$$\hat{u} = (Z^TR^{-1}y - Z^TR^{-1}X\hat{\beta})(Z^TR^{-1}Z + G^{-1})^{-1} \tag{4.4}$$

Now, Christensen[12] substituted the value of $\hat{\mu}$ from equation(4.4) in equation(4.2), to get simplified equation as:

$$X^T V^{-1} X\hat{\beta} = X^T V^{-1} y \tag{4.5}$$

where V is the covariance matrix of the mixed model equation(3.1) such that $Cov(y) = V$. Further, upon solving for $\hat{\beta}$ in equation(4.5),we get:

$$BLUE(\beta) = \hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y \tag{4.6}$$

The covariance is given by:

$$COV(\hat{\beta}) = (X^T V^{-1} X)^{-1} \tag{4.7}$$

Henderson et al.[13] had proved that $\hat{\beta}$ of equation (4.6) is a solution for the GLS equations (4.5) by showing $VV^{-1} = I$.

4.2 Computation of Covariance Matrix of BLUE

In 2021, Prus and Piepho [3] introduced the design matrices and have re-written the LMM (equation 3.3) in the vector form as:

$$Y = X\mu + Z\alpha + \tilde{\epsilon} \tag{4.8}$$

$$Y = (1_k \otimes F)\mu + (I_k \otimes F)\alpha + \tilde{\epsilon} \tag{4.9}$$

where, $F = block - diag(1_{LJ_1}, \ldots, 1_{LJ_P})$. $X = (1_k \otimes F)$ and $Z = (I_k \otimes F)$ are the design matrices for fixed (μ) and random (α) respectively. $\tilde{\epsilon}$ includes λ, γ, and ϵ components. 1_k is a vector of length k with all elements equal to 1, and I_k is the identity matrices(k x k). $\otimes$ represents the Kronecker product which is an operation that transforms two matrices into a larger matrix that contains all possible products of the entries of the two matrices [14].

The covariance matrix of the interaction effect is given by $Cov(\alpha) = G = \sigma^2 I_K \otimes D$. The expected value $E(Y) = X\mu$ and its covariance matrix is $Cov(Y) = V = ZGZ' + R$.

Prus and Piepho have further derived the covariance of $\tilde{\epsilon}$ in their paper, which is highlighted below:

$$R = Cov(\tilde{\epsilon}) = \sigma^2((v_1 1_K 1_K^T + v_2 I_K) \otimes I_J \otimes (1_L 1_L^T) + v_3(1_K 1_K^T) \otimes I_{LJ} + I_{LJK})$$

Since this thesis has used single replication of genotype in location, so upon customizing the above equation for L=1 and ignoring associated replication effects in location $v_3 = 0$, we get:

$$R = Cov(\tilde{\epsilon}) = \sigma^2(v_1 1_K 1_K^T + (v_2 + 1) I_K) \otimes I_J \tag{4.10}$$

Substituting these values of Z, G, and R in equation (4.8), we get :

$$V = ZGZ' + R$$

$$V = (I_k \otimes F)(\sigma^2 I_K \otimes D)(I_k \otimes F)^T + \sigma^2(v_1 1_K 1_K^T + (v_2 + 1) I_K) \otimes I_J$$

Using the Kronecker product rule mentioned in equation D, we can write:

$$V = I_k \otimes (\sigma^2 FDF^T) + I_k \otimes (\sigma^2(v_2 + 1) I_J) + \frac{1}{K} 1_K 1_K^T \otimes (\sigma^2 v_1 K I_J)$$

$$V = I_k \otimes (\sigma^2 FDF^T + (v_2 + 1) I_J) + \frac{1}{K} 1_K 1_K^T \otimes (\sigma^2 v_1 K I_J)$$

Using the formula from equation A,

$$V^{-1} = \frac{1}{K}1_K 1_K^T \otimes \left[\sigma^2(FDF^T + (v_2 + 1 + Kv_1)I_J)\right]^{-1} + (I_k \frac{1}{K}1_K 1_K^T \otimes \left[\sigma^2 FDF^T + (v_2 + 1)I_J)\right]^{-1}$$

Let, $V_1 = \left[\sigma^2(FDF^T + (v_2 + 1 + Kv_1)I_J)\right]^{-1}$ and $V_2 = \left[\sigma^2 FDF^T + (v_2 + 1)I_J)\right]^{-1}$, then we can write :

$$X^T V^{-1} X = (1_k^T \otimes F^T)(\frac{1}{K}1_K 1_K^T \otimes V_1 + I_k \frac{1}{K}1_K 1_K^T \otimes V_2)(1_k \otimes F)$$

Using the Kronecker product rule mentioned in equation D, we can write:

$$X^T V^{-1} X = \frac{1}{K}1_K^T 1_K 1_K^T 1_K \otimes (F^T V_1 F) + 1_K^T (I_K - \frac{1}{K}1_K 1_K^T)(1_k \otimes (F^T V_2 F))$$

$$X^T V^{-1} X = KF^T V_1 F = \frac{K}{\sigma^2} F^T (FDF^T + aI_J)^{-1} F$$

$$X^T V^{-1} X = \frac{K}{\sigma^2 a} F^T (F(\frac{1}{a}D)F^T + I_J)^{-1} F$$

where, $a = (v_2 + 1 + Kv_1)$

Using the formula from equation B, the equation could be written as:

$$X^T V^{-1} X = \frac{K}{\sigma^2 a}(F^T F - F^T F(aD^{-1} + F^T F)^{-1} F^T F)$$

Using the formula from equation C, the equation could be expressed as

$$X^T V^{-1} X = \frac{K}{\sigma^2 a}(\frac{1}{a}D + (F^T F)^{-1})^{-1} \tag{4.11}$$

Now, using equation 4.7 and equation 4.11, the covariance could be written as:

$$Cov(\hat{\mu}) = (X' V^{-1} X)^{-1} = \frac{\sigma^2 a}{K}(\frac{1}{a}D + (F^T F)^{-1}) \tag{4.12}$$

5 Optimal Design

Design of experiments (DOE) is a systematic process to determine the relationship between the input factors and the results of a process [15]. Optimal design is part of the DOE which are experimental in nature with an objective to estimate the parameters without bias and with minimum variance. Optimal designs are valued because it has three significant advantages in DOE, which are outlined below [16]:

1. It minimizes the cost of experiments as it allows the estimation of parameters with few experimental runs.
2. It is flexible and can accommodate multiple factors, such as processes and mixtures.
3. It works with constraints or limitations and helps optimize the design.

Optimal design deals with minimizing the variance of estimators; therefore, in this paper, we need to reduce the covariance matrix of BLUE with related design criteria. Minimization of the variance which also corresponds to maximization of the information, is one of the primary objective of the optimal design.

In general, there are two types of optimal designs, i.e., the exact and the approximate. The information or the moment matrices have been discussed, and the same has been extended for the covariance of BLUE.

5.1 Exact and Approximate Design

Exact Design

$$\xi := \begin{pmatrix} x_1 & \dots & x_P \\ J_1 & \dots & J_P \end{pmatrix} \tag{5.1}$$

where $x_1 \dots x_P$ represents the sub-regions and $J_1 \dots J_P$ denotes number of locations in each sub-region.

Approximate Design

$$\xi := \begin{pmatrix} x_1 & \dots & x_P \\ w_1 & \dots & w_P \end{pmatrix} \tag{5.2}$$

where $w_i = J_i/J$ is the weight of locations within the sub-region i. The approximate design should fulfill the below conditions:
i). All weights sum up to unity

$$\sum_{n=1}^{P} w_i = 1$$

ii). Each weight should be positive and range between 0 and 1.

$$0 \leqslant w_i \leqslant 1$$

Information Matrix

The Information Matrix (IM) is the inverse matrix of the variance matrix. IM is important because it has been used as an input in the computation of the A-optimality criterion (or A-criterion) later in this paper. Prus and Piepho[3] have expressed information matrix for approximate design as below:

$$M(\xi) = diag(w_1 \dots w_P) \tag{5.3}$$

and for exact design as follows:

$$M(\xi) = \frac{1}{LJ} F^T F$$

For L=1, it simplifies to:

$$M(\xi) = \frac{1}{J} F^T F \tag{5.4}$$

Extending the definition of the covariance of BLUE from equation (4.12), we get

$$Cov(\xi) = \frac{\sigma^2 a}{K} (\frac{1}{a} D + M(\xi)^{-1}) \tag{5.5}$$

5.2 Design Criteria

A specific optimality criterion is optimized to formulate an optimal design[17]. There are different design criteria (such as A, C, D, E, S and T) that correspond to the explicit goals that the design must achieve to be optimal. In this thesis, the A-criterion has been used. The reason for selecting A-criterion is that it is one of the most commonly used criteria for designing of MET. It is based on minimizing the trace of the covariance matrix of BLUE.

A-Criterion

A-criteria minimizes the average variance of the parameter estimates [18]. A design (ξ_{opt}) could be termed as A-optimal if it meets below condition [19]:

$$tr(Cov(\hat{\mu}_{\xi_{opt}})) \leqslant tr(Cov(\hat{\mu}_{\xi}))$$

for all other design (ξ) , and $\hat{\mu}$ is BLUE. In simple words, it means that the sum of diagonal elements (or trace) of covariance matrix of BLUE, has to be less than or equal to any other design to be termed as A-optimal design.

Standard A-criterion

$$\Phi_A(\xi) = tr(Cov(\xi))$$

Using equation (5.5), we can re-write the expression as:

$$\Phi_A(\xi) = tr(\frac{\sigma^2 a}{K}(\frac{1}{a}D + M(\xi)^{-1}))$$

Neglecting constant factor $\frac{\sigma^2 a}{K}$, we get

$$\Phi_A(\xi) = tr(M(\xi)^{-1} + \Delta))$$

$$\Phi_A(\xi) = tr(M(\xi)^{-1}) + tr(\Delta))$$

where $\Delta = \frac{\sigma^2 D}{K}$ and $tr(A + B) = tr(A) + tr(B)$ is one of the property of trace. Again, Δ and D are independent of designs, so ignoring it we get:

$$\Phi_A(\xi) = tr(M(\xi)^{-1}) \tag{5.6}$$

The minimum solution could be represented as $w_i = \frac{1}{P}$, where P is the number of sub-regions and $i = 1, ..., P$, and the rounding of the values gives the exact designs [3]. This can be proved analytically but in this thesis, it has been illustrated by using the OptimalDesign (OD) package in section 7 of this thesis.

Weighted A-criterion

The weighted A-criterion is given by:

$$\Phi_{Aw}(\xi) = tr(L \cdot (M(\xi)^{-1} + \Delta)) \tag{5.7}$$

where, $L = diag(l_1, \ldots, l_p)$ denoted coefficient of sub-regions with the condition that $l_1, \ldots, l_p > 0$. Δ has no influence on design and can be neglected.

It may be noted that the OptimalDesign package introduced in next section works with classical A-criterion $\Phi_A(\xi)$ only. In order to work with weighted A-criterion $\Phi_{Aw}(\xi)$, it should be re-written in below form:

$$\Phi_A(\xi) : \Phi_{Aw}(M(\xi)^{-1}) = \Phi_A(\tilde{M}(\xi)^{-1})$$

Then,

$$tr(L \cdot M(\xi)^{-1}) = tr(\tilde{M}(\xi)^{-1})$$

Also, $\tilde{M} = \tilde{F}^T\tilde{F}$. Now, in order to find out $\tilde{F}$ required in the package, L should be written in the form $\tilde{L} \cdot \tilde{L}^T$ to work with the package. If L is a positive definite, as in the case of this thesis, the $\tilde{L}$ is also positive definite [20].

Then,

$$\begin{aligned} tr(L \cdot M(\xi)^{-1}) &= tr(\tilde{L} \cdot \tilde{L}^T \cdot M(\xi)^{-1}) \\ &= tr(\tilde{L}^T \cdot M(\xi)^{-1} \cdot \tilde{L}) \\ &= tr(\tilde{L}^T \cdot (F^T F)^{-1} \cdot \tilde{L}) \\ &= tr(\tilde{L}^{-1} \cdot (F^T F) \cdot (\tilde{L}^T)^{-1})^{-1} \\ &= tr(\tilde{L}^{-1} F^T \cdot F(\tilde{L}^T)^{-1})^{-1} \\ &= tr(F \cdot (\tilde{L}^{-1})^T F \cdot (\tilde{L}^{-1})^T)^{-1} \\ &= tr(\tilde{F}^T \cdot \tilde{F})^{-1} \end{aligned} \tag{5.8}$$

6 OptimalDesign(OD) package in R

The OptimalDesign (OD) package in R is a toolbox for computing optimal and efficient Designs of Experiments (DOE). The package has been created by Harman and Filova (2016) [21]. Some of the functions in this package require the 'gurobi' software and its accompanying R package. The academic license of the software is available to download for free. The exact and approximate designs can be effectively calculated by functions such as od.MISOCP and od.SOCP, respectively. The objective of both functions is to minimize the value of the optimality criterion over all possible designs and choose the optimal one. It may be noted that both functions for the calculation of exact and approximate designs have similar arguments, but the output may differ because of the algorithms as they are not the same.

Some of the important arguments of OD package in R, are described below (also see [21]):

- F: It represents the m-dimensional regressors corresponding to n design points in a matrix (n x m) form, where n is the number of the design points and m is the number of parameters (or feature variables). It is assumed that $n \geqslant m \geqslant 2$.
- b, A: These represent linear constraints on a set of permissible designs w, wherein "b" is a real vector of length k and "A" is the $k \times n$ matrix of reals numbers. The linear constraints $A \times w \geqslant b$, $w_0 \leqslant w$ define the set of permissible designs w (where w_0 is described below).
- w_0: The non-negative vector of length n representing the design to be augmented. The default w_0=NULL, which is set to the vector of zeros.
- crit: It represents the design criterion such as "A" that will be used in the function.

An efficient exact design is computed by means of integer quadratic programming, which solves mathematical optimization problems involving quadratic functions. The idea is to use a quadratic criterion, which approximates the target criterion in the neighborhood of the information matrix of the optimal approximate design.

The best permissible design can be found using the w.best method. The design is formally a non-negative vector $w = (w_1, \ldots, w_n)^T$. It may be noted that "w" is called an "approximate"

design when w_i is the approximate (possibly non-integer) whereas, it is called an “exact" design if it exhibits an exact (integer) number of replications of independent trials.

7 Optimal design - example

This thesis plan to use variance data that comes from the zoned Indian maize crop study published in the paper authored by Kleinknecht et al.[8]. They had worked on finding a solution for selecting the best genotype per zone. The focus of this thesis is to use the variances from the paper to calculate optimal designs and not on the data itself.

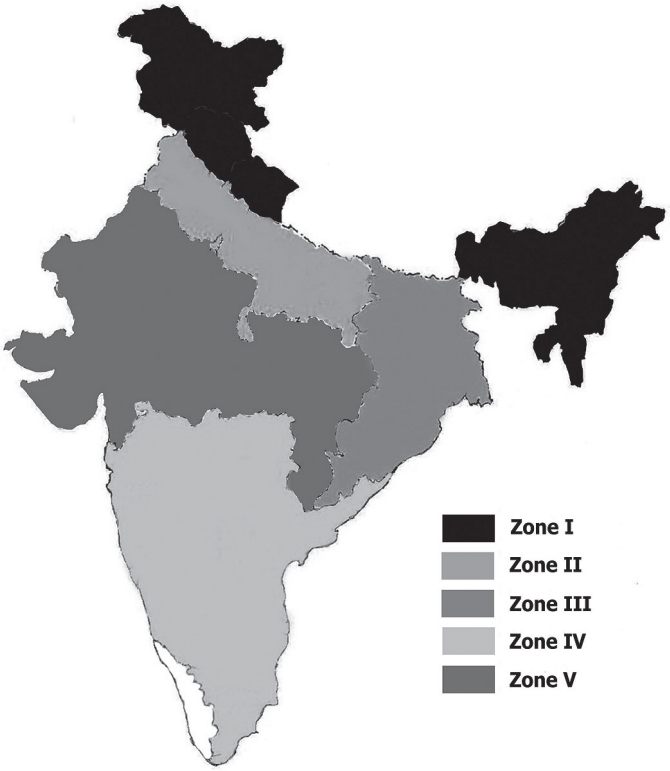
7.1 Description

Maize is traditionally grown during the monsoon (rainy) season in many parts of India across a wide range of environments, extending from extreme semiarid to subhumid and humid regions. It is therefore, the diverse environment has been classified into five zones on account of similar climatic conditions. The map showing different zones is shown in Figure 7.1. These zones are the Northeast and Northwest Himalayas (Zone I), Northeast Indo-Gangetic plains (Zone II), Northwest Indo-Gangetic plains (Zone III), Peninsular India (Zone IV), and central and western India (Zone V).

The data was collected by the All India Coordinated Maize Improvement Program (AICMIP) coordinated by the Directorate of Maize Research, New Delhi. Further, each zone comprises of four to six trial locations, and the data was collected for four maturity groups: extra early, early, medium, and late maturity. Ten series for each maturity group were composed for the data available for the years 1995 - 2006. Kleinknecht et al.[8] computed variance components of the original series, which was presented in the paper.

This thesis has considered the first-order factor-analytic (FA), which is a statistical model, mentioned in the Kleinknecht et al. (2013) [8] and has further used the variance components from the extra-early maturity group, which corresponds to Tables 6 and 7, reported in the reference paper[8]. The covariance matrix V for the genotype × zone effects for the extra early maturity presented in the reference paper [8] could be represented by equation 7.1.

Figure 7.1: Indian maize breeding zones [8]



$$V = \begin{pmatrix} 567 & 254 & 239 & 485 & 328 \\ 254 & 155 & 118 & 240 & 162 \\ 239 & 118 & 155 & 226 & 153 \\ 485 & 240 & 226 & 488 & 310 \\ 328 & 162 & 153 & 310 & 215 \end{pmatrix} \tag{7.1}$$

The variance components has been derived in this thesis for its design problems, mirroring the reported variance components in the paper[8] and reference derivation in Prus and Piepho's (2021)[3] paper. The summary has been reported in Table 7.1.

Table 7.1: Variance (Var) components used in this thesis (Model-T) and in paper of Kleinknecht et al. (Model-K) [8]

Effect	Model-T	Model-K	Var: Model-T	Var: Model-K
Zone + mean	μ_i	$\mu + z_h + za_{hk} + a_k$	fixed	426 + 107 + 153
Genotype × zone	α_{ik}	$g_{i(h)} + gza_{ihk} + ga_{ik}$	$\sigma^2 D$ for α_k	V + $31I_5 1_5^T + 18I_5$
Location × zone	λ_{ij}	$l_{jh} + la_{jhk}$	σ^2_λ	1129 + 1000
Gen × loc × zone + Error	$\gamma_{ijk} + \epsilon_{ijkl}$	$gh_{ijh} + e_{hijk}$	$\sigma^2_\gamma + \sigma^2$	160 + 333

The variances for the effects were obtained in Table 7.1. while comparing the model and its components.

7.2 Methods - Exact and Approximate Designs for Fixed effects

The zones were divided into 5 sub-regions. The exact and approximate design was calculated for various combinations of locations within the sub-regions (p=5).

Standard A-Criterion

The working of the algorithm can be represented by pseudocode, which is defined in the documentation of OD package in R and is reproduced below [21]:

Algorithm 1. Optimal Design

Function: “optimal" - created a function to calculate the optimal design which process data and uses od.MISOCP and od.SOCP of OptimalDesign package.
Input : number of sub-regions as “subregions", matrix L as “weights", number of locations as “locations".
Output: exact design and approximate design.

Step 1 : Choleski decomposition of matrix L was computed using chol function and its transpose was obtained.

Step 2 : Diagonal matrix of dimension P was built for the total number of sub-regions.

Step 3 : Regressors corresponding to design points were computed by applying matrix multiplication to solution of linear equation with diagonal matrix obtained in step 2.

Step 4 : b and A that represents linear constraints on a set of permissible designs w, were assigned wherein “b” was a real vector of length k (locations) and “A” is the k (locations) × p (subregions) matrix of 1.

Step 5 : The exact and approximate designs were calculated by using functions such as od.MISOCP and od.SOCP of OptimalDesign package respectively. The regressors computed in step 3 along-with b, A obtained in step 4 were used as input and crit="A" was used in these functions to obtain A-optimal designs.

Step 6 : The w.best component of the output of od.MISOCP function, from step 5, represents exact design. The w.best component of od.SOCP function output was further divided by locations to present approximate design.

Weighted A-Criterion

Two different weight matrices (L_1 and L_2) have been used to showcase the results of the weighted A-Criterion and provide inferences. The details of the weight matrices are highlighted ahead.

L_1 (Weight matrix - based on the coefficients that correspond to the area of the sub-region)

Prus and Piepho’s (2021)[3] have used the coefficients that correspond to the area of the sub-region as l_1 = 813685, l_2 = 432716, l_3 = 477365, l_4 = 995298, l_5 = 1174818. These coefficients constitute diagonal matrix L_1, to be used as L for step 2, which was further transformed into $L = \tilde{L} \cdot \tilde{L^T}$ as per equation (5.8) to work with the package. Using these inputs, the algorithm

from Standard A-criterion was further followed to obtain exact and approximate weighted A-criterion. L_1 could be represented as:

$$L_1 = \begin{pmatrix} 813685 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 432716 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 477365 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 995298 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1174818 \end{pmatrix} \tag{7.2}$$

L_2 (Weight matrix - based on the variance of the genotype effects in the sub-region)

This thesis also considered an alternative weight matrix L_2 for the weighted A-criterion. It was based on the variance of the genotype effects in the sub-regions. The condition for L_2 was set and the diagonal elements of the L_2 are the inverse variances where $(v_i = \frac{1}{Var(\alpha_{ik})})$. The $Var(\alpha_{ik})$ are the diagonal elements of the covariance matrix of random effects $\sigma^2 D$ as mentioned in Table 7.1. The elements of the resultant square (5 x 5) matrix L_2 could be represented as:

$$L_2 = \begin{pmatrix} 0.0016 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0049 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0049 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0018 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.004 \end{pmatrix} \tag{7.3}$$

A-Criterion with Constraints

Constraints are the limitations in the design-space which need to be optimized to meet the minimal quality that an investigator wishes to maintain. These minimum design objectives can be imposed due to practical issues or to meet certain requirements.

This thesis will try to implement the constraint in the existing design as below:

- less than or equal to 10% of all locations in 1st sub-region

This constraint results in the below conditions:

For $w_1 <= 10$ (For example, in the case of $J = 100$ ($J \geqslant 100$ for R package, but in reality it is 100), $J_1 \leqslant 10$), the parameters that regulate constraints in the algorithm were set as below:
A = $\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -1 & 0 & 0 & 0 & 0 \end{pmatrix}$ and, b = c(100,-10)

- more than or equal to 25% of all locations in 1st sub-region

This constraint results in the below conditions:

For $w_1 >= 25\%$ (in the case of J=100, it is $J_1 \geqslant 25$), the parameters that regulate constraints in the algorithm were set as below:
A = $\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$ and, b = c(100,25)

8 Results

The implementation of the above methods for exact and approximate designs was carried out, and the results have been represented under their respective subsections. It may be noted that results presented in the tables below have been rounded to two decimal points for approximate design for presentation purposes.

8.1 Standard A-Criterion

The standard A-Criterion with and without constraints, for three different numbers of locations (100,200 and 400) was carried out. The result for standard A-criterion without constraints is represented by Table 8.1. The result for the Standard A - criterion with constraints- less than or equal to 10% in 1st location for similar location trials is depicted in Table 8.2. whereas Table 8.3. contains the result of the Standard A - criterion with constraints- at-least 25% in 1st location.

Computing OD for standard A-Criterion without constraints

Table 8.1: Standard A - criterion without constraints

Locations	Approximate Design					Exact Design				
J	w_1	w_2	w_3	w_4	w_5	J_1	J_2	J_3	J_4	J_5
100	0.20	0.20	0.20	0.20	0.20	20	20	20	20	20
200	0.20	0.20	0.20	0.20	0.20	40	40	40	40	40
400	0.20	0.20	0.20	0.20	0.20	80	80	80	80	80

Computing OD for standard A-Criterion with constraints- less than or equal to 10% in 1st location

Table 8.2: Standard A - criterion with constraints- less than or equal to 10% in 1st location

Locations	Approximate Design					Exact Design				
J	w_1	w_2	w_3	w_4	w_5	J_1	J_2	J_3	J_4	J_5
100	0.10	0.22	0.22	0.22	0.22	10	22	23	22	23
200	0.10	0.22	0.22	0.22	0.22	20	45	45	45	45
400	0.10	0.22	0.22	0.22	0.22	40	89	91	90	90

Computing OD for standard A-Criterion with constraints- at-least 25% in 1st location

Table 8.3: Standard A - criterion with constraints- at-least 25% in 1st location

Locations	Approximate Design					Exact Design				
J	w_1	w_2	w_3	w_4	w_5	J_1	J_2	J_3	J_4	J_5
100	0.25	0.19	0.19	0.19	0.19	25	19	19	18	19
200	0.25	0.19	0.19	0.19	0.19	50	38	38	37	37
400	0.25	0.19	0.19	0.19	0.19	100	75	75	75	75

8.2 Weighted A-criterion

The weighted A-Criterion for three different numbers of locations (100,200, and 400) using two different constraints (less than or equal to 10% in 1st location and at-least 25% in 1st location was carried out. The constrained results related to coefficients of sub-regions are depicted in Table 8.4 and Table 8.5, respectively. In addition, these constraints were also implemented for weight matrices L_1 and L_2, and the results are highlighted in Table 8.6 and Table 8.7, respectively. It may be noted that exact design and approximate design uses different algorithms for their estimations. This means that exact designs are not rounded approximate designs.

Weight matrix- based on the coefficient of sub-regions

Computing OD for Weighted A-criterion without constraints

Table 8.4: Weighted A-criterion (L_1 without constraints)

Locations	Approximate Design					Exact Design				
J	w_1	w_2	w_3	w_4	w_5	J_1	J_2	J_3	J_4	J_5
100	0.21	0.15	0.16	0.23	0.25	21	15	16	23	25
200	0.21	0.15	0.16	0.23	0.25	42	30	32	46	50
400	0.21	0.15	0.16	0.23	0.25	83	61	64	92	100

Computing OD for Weighted A-criterion with constraints- less than or equal to 10% in 1st location

Table 8.5: Weighted A-criterion (L_1 with Constraints)- less than or equal to 10% in 1st location

Locations	Approximate Design					Exact Design				
J	w_1	w_2	w_3	w_4	w_5	J_1	J_2	J_3	J_4	J_5
100	0.10	0.17	0.18	0.26	0.28	10	17	18	26	29
200	0.10	0.17	0.18	0.26	0.28	20	35	36	52	57
400	0.10	0.17	0.18	0.26	0.28	40	69	72	105	114

Computing OD for Weighted A-criterion with constraints-at-least 25% in 1st location

Table 8.6: Weighted A-criterion (L_1 with Constraints)- at-least 25% in 1st location

Locations	Approximate Design					Exact Design				
J	w_1	w_2	w_3	w_4	w_5	J_1	J_2	J_3	J_4	J_5
100	0.25	0.14	0.15	0.22	0.24	25	14	15	22	24
200	0.25	0.14	0.15	0.22	0.24	50	29	30	44	47
400	0.25	0.14	0.15	0.22	0.24	100	58	60	87	95

Weight matrix- based on the variance of the genotype effects in the sub-regions

Computing OD for Weighted A-criterion without constraints

Table 8.7: Weighted A-criterion (L_2 without Constraints)

Locations	Approximate Design					Exact Design				
J	w_1	w_2	w_3	w_4	w_5	J_1	J_2	J_3	J_4	J_5
100	0.14	0.25	0.25	0.15	0.21	14	25	24	15	22
200	0.14	0.24	0.24	0.16	0.22	29	49	49	30	43
400	0.14	0.25	0.25	0.15	0.22	57	101	95	60	87

It may be noted that approximate designs are optimal, but the exact design algorithm follows a heuristic approach for the solution; the output may produce bit-varying results during different runs. However, all these designs are highly efficient.

Computing OD for Weighted A-criterion with constraints- less than or equal to 10% in 1st location

Table 8.8: Weighted A-criterion (L_2 with Constraints)- less than or equal to 10% in 1st location

Locations	Approximate Design					Exact Design				
J	w_1	w_2	w_3	w_4	w_5	J_1	J_2	J_3	J_4	J_5
100	0.10	0.26	0.26	0.16	0.22	10	25	26	15	24
200	0.10	0.26	0.26	0.16	0.22	20	52	49	32	47
400	0.10	0.26	0.26	0.16	0.22	40	124	95	57	84

Computing OD for Weighted A-criterion with constraints- at-least 25% in 1st location

Table 8.9: Weighted A-criterion criterion (L_2 with Constraints)- at-least 25% in 1st location

Locations	Approximate Design					Exact Design				
J	w_1	w_2	w_3	w_4	w_5	J_1	J_2	J_3	J_4	J_5
100	0.25	0.21	0.22	0.13	0.19	25	21	22	13	19
200	0.25	0.22	0.22	0.13	0.18	50	41	40	29	40
400	0.25	0.22	0.21	0.13	0.19	100	86	86	57	71

9 Discussion

The optimal design for the sub-regions' effects in multi-environment crop variety testing was studied, and the model was considered (equation 3.3). The BLUE and its covariance matrix were derived. The standard and weighted A-Criterion were formulated, which minimize the trace of the based on the variance of the genotype effects in the sub-region covariance matrix of BLUE. Moreover, the exact and approximate design methods were studied, and the results of all the examples exhibit them. This thesis considered the FA model mentioned in the Kleinknecht et al. (2013) [8] as the input matrix and has also used the variances for the effects as highlighted in Table 7.1. The OptimalDesign (OD) package in R was used to find out the optimal number of locations for all the examples.

The standard A-Criterion was implemented for five sub-regions (p=5), three different values of total locations J. The result is highlighted in Table 8.1. It shows that the weights of locations are the same in the approximate design for the different numbers of locations, which further relates to the equal distribution of locations in the exact design. This was expected as Standard A-Criterion gives equal weightage to each sub-region. This illustrates that the number of locations is equally divided into the number of sub-regions. So, in general, the solution could be represented as $w_i = \frac{1}{P}$, where P is the number of sub-regions and $i = 1, ..., P$, and the rounding of the values gives the exact designs [3]. The standard A-Criterion was also implemented for two different constraints. The first constraint - less than or equal to 10% in 1st location was implemented, and the result was depicted in Table 8.2. It shows that the first location has weight for the approximate design was 0.10 (or 10%), and the same proportion was for the exact design. Similarly, the second constraint- at-least 25% in 1st location was also implemented, and the weight proportion was limited to 0.25 (or 25%) for the 1st (first) sub-region, as highlighted in Table 8.3. It may also be noted that the balance numbers of locations were equally divided among four weights between sub-regions 2 (two) to 5 (five). This highlights that the standard A-Criterion algorithm worked well with and without constraints, and the results were on the expected lines.

The weighted A-Criterion was also implemented, and the sub-region coefficients were considered. The weighted coefficients L_1 were taken from Prus and Piepho's (2021) [3]. The result is highlighted in Table 8.4. It was observed that the weights of different sub-regions in approximate design impacted the distribution of number of locations. Large coefficients

could attract a larger allocation of weights. A similar trend was observed in the exact design where allocation of higher number of locations was assigned to larger sub-region. In the case of implementation of the weighted A-design criterion with the alternate weight matrix L_2, a higher number of locations was assigned to sub-regions with lower variances (or higher weights). The results are highlighted in Table 8.7.

Two different constraints were also chosen and implemented in a weighted A-Criterion design. The first constraint was to limit the allocation of the locations in the 1st location to less than or equal to 10%. The second constraint was to include at-least 25% of all allocations in 1st location. The first case considered the coefficient of sub-regions as the input matrix highlighted as weight matrix L_1. The result for this case, and for the first constraint, is highlighted in Table 8.5. It can be observed that weight (w_1) was limited to 0.10 (or 10%) for different trials of locations. The exact design also echoed the same as 10% of the locations were obtained for the first location as optimum design. Similarly, for the weight matrix L_2, the result highlighted in Table 8.8. also depicts the same narrative. The weight corresponding to the 1st location w_1 for approximate design was limited to 0.1, and the exact design was just 10% of the total locations in the first location (J_1). A similar trend was observed for the second constraint (at least 25% in the first location) as well, where the weight (w_1) in the approximate design was limited to 0.25, and 25% of the allocation of locations was made in the exact design. The result is highlighted in Table 8.6 for the L_1. Table 8.9 depicts the result obtained when the input matrix was considered as L_2. The constraint for (w_1) in approximate design was maintained at 0.25, and J_1 was limited to 25% of the allocation of locations in the first sub-region of the exact design.

It was interesting to note that the alternate weight matrix L_2, which includes a representation of the variance of the FA model (equation 7.1), can be related to the result with an observation that the higher value of variance (in the FA model), leads to a lower weight in the weighted A-criterion that results in a lower number of allocation of trials in the sub-region.

10 Conclusion

This section summarizes the findings based on the results in light of the thesis objectives. The goal of this thesis was to study about three research questions.

- Determine the best linear unbiased estimator and its covariance matrix for the sub-regions' effects in the underlined linear mixed model.

The BLUE was determined (equation 4.6), and its covariance matrix for the sub-regions' effects was derived (equation 4.12). It was observed that the derived covariance matrix was inversely proportional to the number of genotypes K and directly proportional to the variance of observational errors σ^2. It also depends on the design matrix F. So, the higher the number of genotypes will lead to a lower value of the covariance matrix of BLUE.

- Formulate and analyze the related design criteria (standard and weighted A-criteria, which minimize the trace of the (adjusted) covariance matrix of the BLUE).

The design criteria for the Standard and Weighted A-criterion were formulated and analyzed, which minimized the trace of the (adjusted) covariance matrix of BLUE. The same was represented by equation 5.6 and equation 5.8. The exact and approximate designs were also discussed. Interestingly, it was noted that the criterion was independent of the covariance matrix of random effects.

- Compute optimal designs – optimal numbers of allocations to sub-regions – using the OptimalDesign package in R or analytically (in simple cases) for the proposed real data example.

The real data example of the Indian maize zone was considered, and its variances from the paper[8] were used for the calculation of the optimal design. These calculations were done by the OptimalDesign package in R. Additionally, the constraints-based design was also studied, and the result of effective implementation of constraints was demonstrated. It was highlighted that the Standard A - criterion gives equal weightage to the sub-regions. Using the weighted A-criterion with sub-region coefficients, it was demonstrated that large area gets higher allocation of locations in a sub-region. Moreover, constraints-based examples were followed, showcasing that the number of optimal allocations to sub-regions can easily follow any practical constraint.

There was also some evidence in the study of the weighted A-criterion, which highlights that a higher variance in a sub-region leads to a lower allocation of locations in a sub-region. These were supported by Tables 8.4-8.9, which included the FA model. This illustration can also be backed by a practical standpoint, as regions with lower information or higher variance are a bit risky for allocating a higher number of locations as the confidence in crop productivity would be low. This is also important from the financial point of view as it can avoid exposure of crops to lesser unknown areas and avoid losses.

A practical example could be the distribution of rice plants for cultivation in a given heterogeneous environment. The weights of sub-regions may be influenced by water availability or rainfall forecast in the sub-region. Since water is an essential factor in the growth of rice plants, a higher weightage of the sub-region will see an increased allocation of locations for that sub-region. This further indicates a lower allocation of the location to sub-region where the water scarcity is high. In this way, the wastage of rice plants could be avoided, and the optimum utilization of resources could be achieved, thereby saving cost and time. Thus, sub-region coefficients based weighted A-criterion, or the optimal design algorithms in general, can help in huge savings and aligns with the objectives of the thesis, in line with recommendations of the UN [2] for the need for genetic diversity in agriculture to combat climate change.

11 Limitations, Future Work and Ethical Consideration

The thesis had its own limitations, wherein a few design criteria were performed. The raw data was not available, and the variance of an Indian maize study [8] was only used with a focus on the implementation of optimal design.

Future work could include the implementation of different design criteria and applications in other real-life situations.

Ethical consideration: The important ethical concerns related to data access and its usage are highlighted below:

- Privacy and distribution: This thesis has used variance data that comes from the zoned Indian maize crop study published in the paper authored by Kleinknecht et al.[8]. The thesis has further used data that correspond to the coefficients of the area of the subregion highlighted in Prus and Piepho's (2021)[3]. These data are not private and are available in the academic domain. Proper citations have been made in this thesis.
- Cost: The data is available for free.
- Usage: This thesis has used data for academic research purposes with no commercial ambitions.

Bibliography

[1] WorldBank. "Climate-Smart Agriculture". In: *Article at The World Bank website* (2021, Retrieved 2022-09-15).

[2] J.G.Silva. "Feeding the World Sustainably". In: *Article at UN website* (2012, Retrieved 2022-09-15).

[3] M.Prus and H.P.Piepho. "Optimizing the Allocation of Trials to Sub-regions in Multi-environment Crop Variety Testing". In: *International Biometric Society* (2021).

[4] Dr.S.A.Gezan. "BLUEs, BLUPs and Breeding Values, what is the difference?" In: *Article at VSNi website* (2021, Retrieved 2022-09-15).

[5] J.Holland F.Isik and C.Maltecca. "Genetic Data Analysis for Plant and Animal Breeding". In: *Springer Nature* (2017).

[6] C.R.Henderson. "Best Linear Unbiased Estimation and Prediction under a Selection Model". In: *International Biometric Society* (1975).

[7] H.P.Piepho and J.Möhring. "Best Linear Unbiased Prediction of Cultivar Effects for Subdivided Target Regions". In: *Crop Sci. 45:1151–1159* (2005).

[8] Möhring Kleinknecht and H.P.Piepho. "Comparison of the performance of best linear unbiased estimation and best linear unbiased prediction of genotype effects from zoned Indian maize data". In: *Crop Science* 53 (2013).

[9] S. Glen. "Fixed Effects / Random Effects / Mixed Models and Omitted Variable Bias". In: *StatisticsHow* (2020, Retrieved 2022-12-25).

[10] W.D.Beavis. "Quantitative Genetics- Multi Environment Trials: Mixed Effects Models". In: *Plant Breeding E-learning in Africa* (2021.Retrieved 2022-10-09).

[11] J.M.Wooldridge. "Introductory Econometrics: A Modern Approach". In: *5th ed South-Western Cengage Learning* (2013).

[12] R. Christensen. "Plane answers to complex questions : the theory of linear models". In: *Springer, New York* (2011).

[13] S.R.Searle C.R.Henderson O.Kempthorne and C.M.v.Krosigk. "The Estimation of Environmental and Genetic Trends from Records Subject to Culling". In: *International Biometric Society* (1959).

[14] M. Taboga. "Kronecker product". In: *Lectures on matrix algebra, StatLect* (2021, Retrieved 2022-12-25).

[15] K. Sundararajan. “Kronecker product”. In: *isixsigma* (2021, Retrieved 2022-12-25).

[16] A.N.Donev A.C.Atkinson and R.D.Tobias. “Optimum Experimental Designs”. In: *SAS Oxford University Press, Oxford* (2007).

[17] J.J.Borkowski W.Limmun and B.Chomtee. “Weighted a-optimality criterion for generating robust mixture designs”. In: *Computers and Industrial Engineering, Volume 125, Issue C* (2018).

[18] B. Jones, K. Allen-Moyer, and P. Goos. “A-optimal versus D-optimal design of screening experiments”. In: *Journal of Quality Technology, Volume 53, number 4* (2021).

[19] A.Das. “An introduction to optimality criteria and some results on optimal block design. In Design Workshop Lecture Notes”. In: *Citeseer* (2002).

[20] D.A.Harville. “Matrix Algebra From a Statistician’s Perspective”. In: *Chapter 18. Sums (and Differences) of Matrices, Springer* (2001).

[21] L.Filova R.Harman. “Package ‘OptimalDesign’”. In: *R package, version 0.2* (2016).

12 Annexure

12.1 Equation A

$$(I_k \otimes A + \frac{1}{K}1_k 1_k^T \otimes B)^{-1} = \frac{1}{K}1_k 1_k^T \otimes (A+B)^{-1} + (I_K \frac{1}{K}1_k 1_k^T) \otimes A^{-1} \qquad A, B \in R^{P \times P}$$

Proof:

$$(I_k \otimes A + \frac{1}{K}1_k 1_k^T \otimes B)(\frac{1}{K}1_k 1_k^T \otimes (A+B)^{-1} + (I_K \frac{1}{K}1_k 1_k^T) \otimes A^{-1})$$

$$= \frac{1}{K}1_k 1_k^T \otimes (A(A+B)^{-1}) + \frac{1}{K}1_k 1_k^T \frac{1}{K}1_k 1_k^T \otimes (B(A+B)^{-1}) + (I_K \frac{1}{K}1_k 1_k^T) \otimes AA^{-1}) + \frac{1}{K}1_k 1_k^T (I_K \frac{1}{K}1_k 1_k^T) \otimes (BA^{-1})$$

Since, $\frac{1}{K}1_k 1_k^T \frac{1}{K}1_k 1_k^T = \frac{1}{K}1_k 1_k^T$, $\frac{1}{K}1_k 1_k^T (I_K \frac{1}{K}1_k 1_k^T) = 0$ and $AA^{-1} = I_P$, we get:
$= \frac{1}{K}1_k 1_k^T \otimes ((A+B)(A+B)^{-1}) + (I_K \frac{1}{K}1_k 1_k^T) \otimes I_P$
$= I_K \otimes I_P$
$= I_{KP}$

12.2 Equation B

$(R + STU)^{-1} = R^{-1} - R^{-1}S(T^{-1} + UR^{-1}S)^{-1}UR^{-1}$; [20]

12.3 Equation C

$(A^{-1} + B^{-1})^{-1} = A - A(A+B)^{-1}A$; [20]

12.4 Equation D

Kronecker- product rules
$(A \otimes B)(A \otimes C) = A \otimes (B + C)$
$(A \otimes B)(C \otimes D) = AC \otimes BD$