OPTIMAL EXPERIMENTAL DESIGNS FOR SPATIALLY AND GENETICALLY CORRELATED DATA USING LINEAR MIXED MODELS

By

LAZARUS KATANA MRAMBA

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2016

Dedicated to :
My family: Dorothy, Emmanuel and Michelle
My Parents: Joseph and Margaret
My siblings: Mary, Jacob, Samson, Joshuah and Joyce

ACKNOWLEDGMENTS

I would like to thank God for having brought me this milestone. I am indebted to my supervisory committee members, Drs. S.A. Gezan, M. Kirst, G.F. Peter, D.R. Valle, and V.M. Whitaker for their incredible support, ideas, energy and time during my Ph.D. program and particularly thanking Dr. S.A. Gezan for giving me an opportunity to be his Ph.D. student.

I am very appreciative of the University of Florida early learning career and the Cooperative Forest Genetics Research Program (CFGRP) for funding my Ph.D. program, without which, it would not have been possible.

My deepest gratitude to my family members, parents and siblings for their abundant love, full support, continuous encouragement, understanding, and consistent prayers, that have kept me going and facilitated the progress of my studies to the climax.

Last but not least, I would like to thank everyone who has directly or indirectly contributed to the success of my Ph.D. program. This includes all my lecturers, friends, officemates and classmates.

TABLE OF CONTENTS

page

ACKNOWLEDGMENTS 4

LIST OF TABLES 8

LIST OF FIGURES 9

ABSTRACT 10

CHAPTER

1 INTRODUCTION 12

1.1 Background 12
1.2 Study Objectives 17

2 GENERATING EXPERIMENTAL DESIGNS FOR SPATIALLY AND GENETICALLY CORRELATED DATA USING MIXED MODELS 20

2.1 Introduction 20
2.1.1 Optimality Criteria 21
2.1.2 Study Objectives 22
2.2 Materials and Methods 23
2.2.1 Statistical Model 23
2.2.2 Pairwise Swap Algorithm 25
2.2.3 Algorithm Evaluation 26
2.2.4 Relative Design Efficiency 27
2.2.5 Data Simulation 28
2.2.6 Motivating Example 29
2.3 Results 32
2.3.1 Initial and Overall Design Efficiency 32
2.3.2 Analysis of Simulated Data 36
2.4 Discussion 37
2.5 Conclusion 43

3 EVALUATING ALGORITHMS EFFICIENCIES FOR EXPERIMENTAL DESIGNS WITH CORRELATED DATA 44

3.1 Introduction 44
3.2 Materials and Methods 46
3.2.1 Statistical Model 46
3.2.2 Algorithms 47
3.2.2.1 Simple pairwise algorithm 47
3.2.2.2 Greedy algorithm 48
3.2.2.3 Genetic neighborhood algorithm 48
3.2.2.4 Simulated annealing algorithm 49

3.2.3 Evaluation of Algorithms . . . 49
3.3 Results . . . 51
3.4 Discussion . . . 52
3.5 Conclusion . . . 60

4 IMPROVING NON-ORTHOGONAL EXPERIMENTAL DESIGNS WITH SPATIALLY AND GENETICALLY CORRELATED DATA . . . 61

4.1 Introduction . . . 61
4.2 Materials and Methods . . . 64
4.2.1 Statistical Models . . . 64
4.2.2 Optimization Procedure . . . 65
4.2.3 Evaluation of Experimental Conditions . . . 66
4.2.4 Relative Design Efficiency . . . 67
4.3 Results . . . 69
4.4 Discussion . . . 71

5 OPTIMALDESIGNMM: AN *R* PACKAGE FOR OPTIMIZING EXPERIMENTAL DESIGNS WITH CORRELATED DATA . . . 76

5.1 Introduction . . . 76
5.2 Statistical Models . . . 78
5.2.1 Case 1 . . . 79
5.2.2 Case 2 . . . 80
5.3 Example: RCB Designs with Regular-Grid Layouts . . . 81
5.4 Example: RCB Designs with Irregular-Grid Layouts . . . 87
5.5 Example: Designs with Genetic and Spatial Correlations . . . 87
5.6 Unequally Replicated Designs . . . 89
5.7 Generating Incomplete Block Designs . . . 95
5.8 Generating Augmented Designs . . . 97
5.9 Simulated Annealing Algorithms . . . 98
5.10 Extensions . . . 99
5.11 Discussion . . . 100

6 CONCLUSIONS . . . 102

APPENDIX

A OTHER OPTIMALITY CONDITIONS . . . 109

A.1 Completely Randomized Designs with Spatial Correlations . . . 109
A.1.1 Ordinary Least Squares Approach . . . 109
A.1.2 Matrix Approach . . . 110
A.2 Randomized Complete Block Designs with Fixed Blocks and Treatments Effects 110
A.3 Randomized Complete Block Designs with Random Blocks and Fixed Treatments Effects . . . 111

B EXTRA TABLES AND GRAPHS . . . 113

B.1 Overall Design Efficiency for Irregular-Grid $\Omega_A^{(30)}$ RCB Designs . . . 113
B.2 Initial and Overall Design Efficiency Table for $\Omega_A^{(196)}$ RCB Designs with 16 Blocks . . . 113
B.3 Initial and Overall Design Efficiency Graphs for $\Omega_A^{(196)}$ RCB Designs with 16 Blocks . . . 114
B.4 Boxplots of Overall Design Efficiency for $\Omega_A^{(30)}$ RCB Designs for Each Algorithm 115
B.5 Overall Design Efficiency Synergies for Non-Orthogonal Designs . . . 116
B.6 Pedigree Information for Full-Sib Families with 30 Offspring . . . 117
B.7 Pedigree Information for Half-Sib Families with 30 Offspring . . . 118
B.8 Pedigree Information for Full-Sib Families with 196 Offspring . . . 119
B.9 Pedigree Information for Half-Sib Families with 196 Offspring . . . 120

C R FUNCTIONS . . . 121

C.1 Simple Pairwise Algorithm . . . 121
C.2 Genetic Neighborhood Algorithm . . . 122
C.3 Simulated Annealing Algorithm . . . 124
C.4 Greedy Pairwise Algorithm . . . 126
C.5 Generate Matrices for RCB Designs . . . 128
C.6 Generate Matrices for Unequally Replicated Designs . . . 131
C.7 Generate Matrices for Augmented Designs . . . 134

REFERENCES . . . 137

BIOGRAPHICAL SKETCH . . . 141

LIST OF TABLES

Table page

2-1 Initial and overall design efficiency (IDE, ODE) summary statistics for RCB designs . 34

2-2 Prediction of genetic values for $\Omega_A^{(30)}$ and $\Omega_D^{(30)}$ experiments 39

2-3 Prediction of genetic values for $\Omega_A^{(196)}$ scenario with 16 blocks 40

3-1 Average of algorithms ODEs for $\Omega_A^{(30)}$ RCB experimental designs 53

3-2 Average of algorithms ODEs for $\Omega_A^{(196)}$ RCB experimental designs 57

3-3 Average of algorithms ODEs for $\Omega_D^{(30)}$ RCB experimental designs 57

4-1 Summary of ODEs for unequally replicated regular-grid designs 71

4-2 Summary of ODEs for incomplete block regular-grid designs 72

4-3 Summary of ODEs for augmented regular-grid designs 72

B-1 Overall design efficiency for irregular-grid $\Omega_A^{(30)}$ RCB designs 113

B-2 Initial and overall design efficiency for $\Omega_A^{(196)}$ RCB designs with 16 blocks 113

B-3 Pedigree information for full-sib families with 30 offspring 117

B-4 Pedigree information for half-sib families with 30 offspring 118

B-5 Pedigree information for full-sib families with 196 offspring 119

B-6 Pedigree information for half-sib families with 196 offspring 120

LIST OF FIGURES

Figure page

2-1 Motivating example showing rate of design improvement with time 31

2-2 Initial and overall design efficiency summary statistics based on RCB designs 35

2-3 Kernel densities for estimated heritabilities 38

3-1 A motivating $\Omega_A^{(30)}$ example showing successful swaps and ODE for simple pairwise,simulated annealing, greedy, and genetic neighborgood algorithms 53

3-2 A motivating $\Omega_A^{(30)}$ example with all trace values, showing rates of convergence for simple pairwise,simulated annealing, greedy, and genetic neighborgood algorithms . . 54

3-3 ODE % for $\Omega_A^{(30)}$, $\Omega_D^{(30)}$, and $\Omega_A^{(196)}$ for each algorithm 55

3-4 Average swaps for $\Omega_A^{(30)}$, $\Omega_D^{(30)}$, and $\Omega_A^{(196)}$ for each algorithm 56

4-1 Examples of regular and irregular-grid experimental layouts 68

4-2 Individual effective ODE % for unequally replicated designs 73

4-3 Individual effective ODE % for incomplete block designs 73

4-4 Individual ODE % for augmented designs 74

5-1 An illustration of the optimization process for RCB designs 85

B-1 Initial and overall design efficiency for $\Omega_A^{(196)}$ generated with 16 blocks 114

B-2 Boxplots of overall design efficiency for $\Omega_A^{(30)}$ RCB scenario for each search algorithm 115

B-3 Overall design efficiency synergies for incomplete block and unequally replicated designs 116

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

OPTIMAL EXPERIMENTAL DESIGNS FOR SPATIALLY AND GENETICALLY CORRELATED DATA USING LINEAR MIXED MODELS

By

Lazarus Katana Mramba

May 2016

Chair: Salvador A. Gezan
Major: Forest Resources and Conservation

Experimental designs with varying levels of spatial and genetic correlations require the use of appropriate statistical models, computational procedures and algorithms to optimally generate their layouts. Statistical models often ignore sources of variations to simplify the bottleneck of computational intensity, which, often results in imprecise estimation and poor prediction of parameters. This dissertation presents several procedures to generate improved experimental designs while accounting for both genetic and spatial correlations at the design stage. Appropriate linear mixed models were studied together with information based *A*- or *D*-optimality criteria. Illustrations were provided on a subset of experimental designs including randomized complete block designs, unequally replicated, incomplete block and unreplicated designs such as augmented block designs. Evaluation of relative design efficiency of experiments was done between initially randomly generated designs and improved designs generated after stochastic randomization procedure analyzed using a mixed model framework. Comparison of optimal design efficiencies were evaluated based on simple pairwise algorithm and its variants, simulated annealing and genetic neighborhood. An *R* package, $OptimalDesignMM$, has been developed that implements these procedure to improve designs of experiments. Results from randomized complete block designs had highest overall design efficiency achieved among genetically unrelated individuals at heritability $h^2 = 0.3$ and spatial correlation $\rho = 0.6$. Half-sib and full-sib families achieved highest improvements for relatively low $h^2 = 0.1$, with $\rho = 0.6$ or $\rho = 0.9$. Also, accuracy of prediction of genetic values increased with increase in

h^2 and ρ. In addition, better prediction accuracies were obtained when spatial variability was accounted for. From evaluation of efficiency of search algorithms, results indicated that simple pairwise and simulated annealing achieved highest design efficiency in all evaluated conditions based on A-optimality criterion. Results from non-orthogonal designs indicated that unequally replicated and incomplete block designs achieved highest mean reduction in average variance for experiments with genetically unrelated individuals whereas augmented designs recorded highest average variance reduction among full-sib families with lowest heritabilities. In conclusion, experimental designs have varied sources of variability and require appropriate statistical models and computational procedures to realize important design efficiencies.

CHAPTER 1
INTRODUCTION

1.1 Background

Designing an experiment is an essential stage in research settings that requires decisions to be made in order to choose the best out of a set of alternatives. An optimum decision procedure relies on choosing an experiment that is optimum in some sense. Thus, the problem of generation of optimal experimental designs is dependent on making maximum use of available information with a goal to be able to estimate parameters of interest accurately and with precision. Basic principles of generating experimental designs have been discussed in detail by researchers such as Yates (1939); Welham *et al.* (2015), namely: randomization, replication and blocking. Replications enable estimation of experimental error variance and, the more replications there are, the more precise inference can be made. Randomization makes sure that all experimental units are equally likely to receive any treatment thus minimizing systematic errors from the experimenter. Blocking controls for different sources of natural variation amongst experimental units. When applied appropriately, it can control for field variations and help to reduce background noise.

Field experiments are characterized by varied levels of environmental heterogeneity which influence the accuracy and precision of estimated parameters. Standard traditional designs assume, for simplicity, that residual errors are uncorrelated. However, when experimental units are measured or located in close proximities, their responses are likely to be more correlated than those spaced out, either in time or distance (Stroup, 2013; Gilmour *et al.*, 2009). In blocked designs, spatial correlation may arise due to the physical proximity among experimental units, and thus sharing of microsite variability, which influences the precision of the experiment. Spatially correlated errors are oftenly modeled using a 2-dimensional separable autoregressive spatial error structure of order 1 (Cullis *et al.*, 2006; Butler *et al.*, 2008; Gezan *et al.*, 2010; Gilmour *et al.*, 2009) but other spatial error structure are available (Littell *et al.*, 2006) that can be used where appropriate. Another form of correlation between experimental units that is

commonly present in plant studies is genetic relatedness, that needs to be accounted for (Littell *et al.*, 2006). Therefore, generating an optimal experiment requires a proper exploitation of all, or most of the above sources of variations.

Several computer routines and software such as CycDesigN (John and Williams, 1995) and other studies (Butler *et al.*, 2008; Cullis *et al.*, 2006) have implemented diverse approaches to generating experimental designs. However, most of the methods use approximations of some sort to optimize designs, for instance Cullis *et al.* (2006); Butler *et al.* (2008) used an approximation of A-optimality to improve experimental designs, and for some, they do not exploit the full correlated structure of the experimental units for possible spatial or genetic correlations. For instance, Filho and Gilmour (2003) explored genetic correlations but no spatial correlations between experimental units.

Randomized complete block (RCB) designs are one of the most frequently used designs, however, they often have some restrictions. First, RCB designs require that there is sufficient materials to replicate all treatments into several homogeneous blocks. Second, RCB designs are balanced designs, where treatment and block effects, and their contrasts, are orthogonal. In the event that the number of treatments is very large, RCB designs become unviable since the block structure will encompass heterogeneous conditions. In such situations, incomplete block (IB) designs which have smaller (incomplete) block sizes than the number of treatments, can be used to reduce environmental heterogeneity within blocks thus increasing precision of estimations. Unbalanced designs are such that, the contrasts for testing treatment effects are correlated due to unequal number of observations in different blocks. Balanced designs, such as RCB, have uncorrelated treatments effects with contrasts that are orthogonal. These designs have all treatments equally represented in every block and replicated as many times as possible. Thus, RCB designs are one of the most efficient layouts that yield accurate and precise estimated parameters when possible sources of variations have been adjusted for appropriately

Unbalanced experimental designs have been described in different forms (Federer, 1956; Federer and Raghavarao, 1975; Federer, 1998; Cullis *et al.*, 2006; William *et al.*, 2011) and

include extreme cases such as unreplicated trials of which augmented designs are examples of such. These are mainly used in early stages of breeding programs when replication of treatments is impossible due to lack of enough propagation materials (Moehring *et al.*, 2014). In particular, unbalaned and partially or fully unreplicated trials allow testing of several hundreds of treatments with little or no replication. Therefore, existence of unbalanced experimental designs is inevitable in many research settings, yet, adequate procedures to improve such experiments is lacking despite the advantages of using optimized designs which would yield more accurate and precise estimated parameters of interest.

Mixed models have become common in plant studies due to their potential to provide accurate estimation of variance components and unbiased predictions of treatment effects as they are flexible in modeling correlated errors, and, incorporation of heterogeneous structures inthe model. Mixed models contain both fixed and random effects, where, effects are said to be fixed if their levels were selected by nonrandom process or the specified levels included in the study consist of their entire population of possible levels. On the other hand, random effects are factors with levels that consist of a random sample of levels from a population of possible levels and inference is made to the whole population of levels. In a given experimental design, factors can be considered as fixed or random depending on the aim of the experiment and whether the observations are a random sample from a larger population from which an inference is to be made. Blocks, for instance, can be considered to be random effects if they are a random sample of all potential blocking units that can be selected such as plots. However, for unbalanced experimental designs, blocks are considered to be random effects since they are incomplete as not all treatments are equally represented in every block.

Mixed models are advantageous and highly applicable in designs of field experiments as they extend the linear models by allowing a more flexible specification of errors and other random factors (Stroup, 2013; Littell *et al.*, 2006; Mathew *et al.*, 2015; Brown and Prescott, 2015).

In plant breeding, it is of interest to improve and maximize the efficiency of field experimental designs to obtain better predictions of genetic or breeding values. Treating genotypes as random effects in the context of a mixed model allows for the incorporation of correlated information from relatives. Mixed models use restricted maximum likelihood to estimate variance components and they maximize correlations between the true and the predicted breeding values by minimizing the prediction error variance. It is important for statistical models to include all possible sources of variation to better correct the observed phenotypes to obtain estimated best linear unbiased predictions (BLUPs).

Prediction and estimation of breeding values and heritabilities may be inaccurate when phenotypic observations present with spatial and genetic correlations that the statistical models do not account for. Genetically, it is important to consider information about relatives since they share some alleles, and therefore their response is correlated. Statistically, random genetic effects are correlated and therefore a matrix of variance covariance ($\mathbf{A}$) between genotypes should be incorporated in a mixed model. Genetic relationships can be calculated using genetic theory (Falconer and Mackay, 1996) or molecular information such as SNPs (VanRaden, 2008). Incorporating genetic relationships into the mixed model is a more efficient use of the information about individuals, but also, genetic values of individuals not tested, but with relatives tested, can be predicted and selected. Often, pedigree information is used to define the genetic relationships among individual treatments, denoted as $\mathbf{G} = \sigma_g^2\mathbf{A}$, where σ_g^2 is variance of the treatments and $\mathbf{A}$ is a numerator relationship matrix.

As mentioned before, experimental units that are physically close together are strongly correlated than units farther apart as they share a common microsite environment. To account for spatial correlations, an error structure is incorporated into a mixed model, where one of the most common is the first order separable autoregressive correlation structure (AR1), that considers an spatial correlation among rows and a different correlation among columns (Gilmour *et al.*, 2009). Also, note that other suitable spatial error structures can be used (Stroup, 2013; Littell *et al.*, 2006; Cressie, 1993).

In order to generate an adequate experimental design, an optimal criterion has to be applied. The choice of which optimality criterion to use may depend on the objective of the experiment (Kuhfeld, 2010). Some of the optimality criteria are *A*, *D*, and *E* (Das, 2002). The most common optimality functions are *A*- and *D*-optimality (Wald, 1943; Chernoff, 1953; Cheng, 1983; Butler *et al.*, 2008; Kuhfeld, 2010), where *A*- minimizes the sum of the diagonal elements of a variance-covariance matrix of estimated treatment effects, denoted here as $\mathbf{M}(\Omega)$, that is equiavalent to minimizing the average variance of the treatment effects; and *D*- minimizes the determinant of $\mathbf{M}(\Omega)$, which in effect minimizes the generalized variance or the volume of an ellipsoid described by $\mathbf{M}(\Omega)$. Kuhfeld (2010) defines these criterion in terms of functions of eigenvalues, where *A*- minimizes the sum of eigenvalues and *D*- minimizes the product of eigenvalues. Other optimality procedures exist and include among others: *E*, *G*, *M*, *S*, and some combinations of these and more (John and Williams, 1995; Butler *et al.*, 2008; Kuhfeld, 2010; Wald, 1943; Cheng, 1983).

Although there is such a need to address generation of optimal experimental designs (Cheng, 1983; Butler *et al.*, 2008), focus has always been on the use of sophisticated analysis with little work done on development of efficient search algorithms to generate optimal designs mostly due to intensive computational demands; and therefore, most approaches tend to ignore practical conditions and assume that any flaws can be corrected at the analysis stage. Several computer algorithms exist that have been used to search for optimal designs. In most cases, they involve swapping pairs of treatments and re-evaluating the new layout. Approximation procedures have been used Butler *et al.* (2008) and other software such as CycDesignN make use of simulated annealing algorithms for simple experimental designs. To optimize experimental designs, intensive computational procedures are required and thus methods that are computationally efficient and that are based in the correct statistical structure of the data (with spatial and genetic correlations) need to be employed

The use of a a complex linear mixed model with the implementation of several search algorithms together with application of *A*- and *D*-optimality criteria are explored in detail for a

wide spectrum of genetic and environmental factors. Comparisons of their performances with respect to relative design efficiencies is undertaken. In addition, this study has endeavored to present an statistical methodology for improving experimental designs spanning from orthogonal to non-orthogonal experiments with incorporation of genetic and spatial correlations using a linear mixed models framework for both balanced and unbalanced data. Here, the models are formulated appropriately for an array of experimental designs including for RCB, unequally replicated, incomplete block and unreplicated designs (augmented) to illustrate different ways they can be used to improve experimental designs.

1.2 Study Objectives

An overall goal of this study is to evaluate the potential of a wide array of computational and statistical procedure to improve experimental designs. This present study evaluates the potential of the following search algorithms to improve experimental designs: (1) pairwise swap procedure, (2) greedy swap, which is a variant of the simple pairwise, where, a single or multiple pairs of treatments are swapped at a time; (3) simulated annealing where a cooling strategy is employed, and (4) genetic neighborhood algorithm where genetic relationships and physical proximity of pairs of treatments are evaluated.

The following are specific objectives for each of the Chapters:

1. In Chapter 2, the objective is to develop and evaluate statistical methods and algorithms in order to generate improved RCB designs while considering simultaneously genetic and spatial correlations at the design stage. Evaluations are done for several typical field conditions with varying levels of heritabilities, genetic relatedness, and spatial correlations that are incorporated into a mixed model framework. A simple pairwise swap algorithm is implemented and evaluated for both *A*- and *D*-optimality criteria. As a secondary objective, a simulated data is evaluated in order to assess prediction accuracies of genetic values, and estimation of heritabilities for initial (unimproved) and final (improved) designs under a combination of conditions.

2. In Chapter 3, the main objective is to evaluate the efficiency of diverse search algorithms to generate improved randomized complete block (RCB) designs, applying *A*- and *D*-optimality criteria, while accounting for both spatial and genetic correlations using linear mixed models with applications in plant breeding trials. Several varying field conditions that include a range of heritabilites, genetic relatedness structures and spatial correlations are evaluated.
3. In Chapter 4, the aim is to develop and evaluate statistical procedures to generate improved designs for unbalanced designs such as unequal replications, incomplete block and augmented designs for field experimental trials based on *A*-optimality criterion and incorporating spatial correlations and genetic relatedness.
4. In Chapter 5, the aim is to solidify the procedures presented in the previous Chapters and create an *R* package that can be used by other researcher with similar interests. Thus an *R* package, called *OptimalDesignMM* has been developed and continues to be improved by adding other useful functions as required to generate and improve block experimental designs.

The next chapters present specific objectives, detailed methods, and discussions with respect to optimization procedures. In all the chapters, spatial correlations were modeled using a 2-dimensional separable autoregressive 1st order variance structure (AR1) whereas genetic relatedness are modeled using a numerator relationship matrix under the linear mixed models framework.

Particularly, Chapter 2 describes a procedure to improve randomized complete block designs using a pairwise algorithm and both *A*- and *D*-optimality criterion. Also, this chapter extends the methods of optimal designs to simulation of a response variable for a continuous trait from which accuracy of predicted genetic values and estimated heritabilities are assessed for different genetic and environmental conditions.

In Chapter 3, relative design efficiencies for several algorithms including simple pairwise, simulated annealing, greedy pairwise and genetic neighborhood are evaluated and compared for

a wide range of experimental conditions, all based on a RCB designs with genetic and spatial correlations modeled using linear mixed models and based on *A*- and *D*-optimality criteria.

As mentioned before, the main goal in Chapter 4 is to develop and evaluate statistical procedures to generate improved designs for non-orthogonal experiments with blocking structure. They include, unequal replications, incomplete block and augmented designs, which were generated based on *A*-optimality criterion. These will be presented with appropriate derivation of linear mixed model equations. Here, both blocks and treatments are random effects, compared to the procedure presented in Chapter 2 where blocks were treated as fixed effects.

Finally, in Chapter 5, the goal is to bring together all *R* functions and document all procedures in order to develop a computational routine that is open source, user friendly, and available for researchers and practitioners. As such an *R* package, called **OptimalDesignMM** is presented. In this chapter, a brief description of the theory behind optimization is given and most important, several illustrations are provided with complete *R* code (syntax) and limited output. Extra tables and graphs are presented in Appendix B and *R* functions for the search algorithms are presented in Appendix C.

CHAPTER 2
GENERATING EXPERIMENTAL DESIGNS FOR SPATIALLY AND GENETICALLY CORRELATED DATA USING MIXED MODELS

2.1 Introduction

Plant and animal breeders often conduct and analyze large field trials with the aim of selecting the best genotypes for future breeding (Möhring, 2010). They do so by testing a large number of genotypes in single locations or multiple environments. Such field trials are characterized by varied levels of environmental heterogeneity such as spatial correlation, typically due to the physical proximity among experimental units, and microsite variability, which influence the precision of the experiment. The existence of correlated observations challenges the standard traditional design methods since they assume that residual errors are uncorrelated despite the fact that for breeding, treatments (*i.e*, genotypes), often have varied levels of genetic relatedness such as half-sib where genotypes share a common single parent, full-sib where they share both parents, and/or clonally propagated genotypes. Genetic relationships form the basis of parental or individual selection in animal and plant breeding programs. Prediction and estimation of parameters of interest such as breeding values and heritabilities may be imprecise when phenotypic observations contain other elements such as spatial and genetic correlations, yet most design of experiments, and in many cases statistical analysis, do not account for these elements. Implementation of appropriate statistical models at both the design and analysis stage is thus a vital component in this field. However, some authors have used extensions of Linear Mixed Models (LMM), at the analysis stage to model, to model this heterogeneity and improve the precision of treatment estimates (Stringer and Cullis, 2002; Gezan *et al.*, 2010; Cullis *et al.*, 1989), an approach that can be extended into the design stage.

Several field experimental designs are used in plant breeding programs including randomized complete block (RCB), incomplete block and row-column designs, where RCB is, at the present, the most common experimental layout used. Often, RCB designs are used with both blocks and treatments (genotypes) factors modeled as fixed effects in addition to assuming that residual errors are uncorrelated. Whether the parameters should be fixed or random effects

depend on the aim of the experiment and whether the observations are a random sample from a larger population from which an inference is to be made.

Therefore, it is of interest in plant breeding to improve the efficiency of experimental designs to obtain better predictions of genetic or breeding values. Thus, considering genotypes as random effects enables its prediction and allows for the incorporation of correlated information from relatives. This genetic information can be obtained by reading pedigree data to estimate expected relationships or by processing molecular data to estimate these relationships (Henderson, 1975; Mrode, 2014).

The generation of an improved experimental layout may require optimization routines, a process that often is computationally intensive, and it has been implemented in several software and studies with different approaches (Butler *et al.*, 2008; John and Williams, 1995; Resende *et al.*, 2005; Cullis *et al.*, 2006). In addition, the choice of which optimality criterion to use may depend on the objective of the experiment and whether the data is scale invariant on the response variable or not (Kuhfeld, 2010). Some of the optimality criteria include A, D, E, G, and Ms (Das, 2002).

2.1.1 Optimality Criteria

Statistically, a design is considered to be optimal if it maximizes the amount of information available, where the layout of its experimental units optimize a function of the variance-covariance matrix of treatment effects (Das, 2002). Information based A- and D-optimality criteria are perhaps the most frequently used procedures when choosing between designs (Butler *et al.*, 2008; Cullis *et al.*, 2006; Hooks *et al.*, 2009; Kuhfeld, 2010; Das, 2002).

A-optimality was first introduced by Chernoff (1953) using the Fisher's information matrix under the framework of a fixed effects model. The information matrix is used because standard errors of the mean (SEM) are calculated using the variance of treatment effects while standard error of differences (SED) are obtained from functions of variances and covariances. A-optimality criterion seeks to minimize the sum of the diagonal elements (*i.e*, trace) of the variance-covariance matrix of the treatment effects. Minimizing the trace implies, for a model

with fixed effects treatment factor, minimizing the average variance of the Best Linear Unbiased Estimator (BLUE) of the treatment effects, and, for a model with a random treatment factor, implies minimizing the average variance of the Best Linear Unbiased Predictor (BLUP) of the treatment effects (Das, 2002). *A*-optimality criterion is not scale invariant (Kuhfeld, 2010) on the response variable; however, this does not affect blocked designs since treatments are on the same scale. The objective function for *A*-optimality is expressed as:

$$A_{opt} = \text{argmin}\{trace[\mathbf{M}(\Omega)]\} \tag{2–1}$$

where $\mathbf{M}(\Omega)$ is the inverse of an information matrix (variance-covariance matrix) of the treatment effects calculated from the experimental layout Ω. More details about this matrix are presented later.

D-optimality is one of the most studied criterion, and it was first introduced by Wald (1943) with other researchers doing extensive work (Kiefer, 1959; Kiefer and Wolfowitz, 1959; Mandal, 2000; Yang, 2008). A design is *D*-optimal if it minimizes the determinant of $\mathbf{M}(\Omega)$, expressed as:

$$D_{opt} = \text{argmin}\{|\mathbf{M}(\Omega)|\} \text{ for } |\mathbf{M}(\Omega)| \neq 0. \tag{2–2}$$

Minimizing the determinant of an inverse of an information matrix is equivalent to minimizing the generalized variance of the treatment effects (Kuhfeld, 2010); hence, this criterion chooses an optimal design for which the volume of the joint confidence ellipsoid is minimized (Das, 2002). A desirable property of this criterion is that it uses both the diagonal and off-diagonal values in the variance-covariance matrix and it is scale invariant on the response variable and computationally efficient (Kuhfeld, 2010).

2.1.2 Study Objectives

This study evaluates the potential of the use of a linear mixed model framework to improve experimental designs. The main objective is to develop and evaluate statistical methods and algorithms in order to generate improved designs while considering simultaneously genetic and spatial correlations at the design stage. The field designs considered in this study correspond

to a RCB design tested and evaluated for several typical field conditions with varying levels of heritabilities, genetic relatedness, and spatial correlations that are incorporated into a LMM. A simple pairwise swap algorithm was implemented for both *A*- and *D*-optimality criteria and relative design efficiency measures were obtained.

As a secondary objective, simulated data that mimics real data in plant breeding trials was evaluated in order to assess prediction accuracies of genetic values, and estimation of heritabilities for initial (unimproved) and final (improved) designs under a combination of conditions.

2.2 Materials and Methods

2.2.1 Statistical Model

For this study, the procedure implemented is based on a LMM that considers blocks as fixed effects and genotypes as random effects. Treatments are considered random effects since they are a random sample from a much larger set of genotypes. This LMM can be expressed as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{g} + \mathbf{e} \tag{2–3}$$

where $\mathbf{y}_{nx1}$ is a vector of observations; $\mathbf{X}_{nxb}$ is a full column rank incidence matrix of fixed block effects; β_{bx1} is a vector of fixed effects (blocks); $\mathbf{Z}_{nxt}$ is a full column rank incidence matrix of random treatment effects; $\mathbf{g}_{tx1}$ is a vector of random effects (treatments); $\mathbf{e}_{nx1}$ is a vector of residual errors; n, b and t are the number of observations, blocks and treatments. The assumptions are:

$$\begin{bmatrix} \mathbf{g} \\ \mathbf{e} \end{bmatrix} \sim MVN \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \right)$$

with $\mathbf{V} = var(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}^{'} + \mathbf{R}$, where $\mathbf{G}$ and $\mathbf{R}$ are variance matrices for the genetic effects and residual errors, respectively. When residual errors are assumed to be independent and identically distributed (*iid*), $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$. Correlated errors were modeled using a 2-dimensional separable autoregressive spatial error structure of order 1 to model spatial variability along the rows and columns of the experimental layouts (Stringer and Cullis, 2002; Gezan *et al.*, 2010; Gilmour

et al., 2009), with $\mathbf{R} = \sigma_e^2 \Sigma_r(\rho_r) \otimes \Sigma_c(\rho_c)$ with

$$Var(e_{ij}) = \sigma_e^2 \quad \text{and} \quad Cov(e_{ij}, e_{i'j'}) = \sigma_e^2 \rho_r^{|dx|} \rho_c^{|dy|} \tag{2–4}$$

where $|dx| = |x_i - x_{i'}|$ and $|dy| = |y_j - y_{j'}|$ are the row and column absolute distances, respectively; $\otimes$ is a Kronecker product; and $\sum_r(\rho_r)$ and $\sum_c(\rho_c)$ are matrices with autocorrelation parameters ρ_r and ρ_c for rows and columns respectively, expressed as

$$\Sigma_r(\rho_r) = \begin{bmatrix} 1 & \rho_r & \rho_r^2 & \cdots & \rho_r^{r-1} \\ & 1 & \rho_r & \cdots & \rho_r^{r-2} \\ & & 1 & \cdots & \rho_r^{r-3} \\ & & & \ddots & \vdots \\ & & & & 1 \end{bmatrix} \quad \text{and} \quad \Sigma_c(\rho_c) = \begin{bmatrix} 1 & \rho_c & \rho_c^2 & \cdots & \rho_c^{c-1} \\ & 1 & \rho_c & \cdots & \rho_c^{c-2} \\ & & 1 & \cdots & \rho_c^{c-3} \\ & & & \ddots & \vdots \\ & & & & 1 \end{bmatrix}$$

When treatment effects are assumed to be genetically unrelated, $\mathbf{G} = \sigma_g^2 \mathbf{I}_{txt}$ where σ_g^2 is the treatments variance and $\mathbf{I}_{txt}$ is an identity matrix. For genetically related individuals, $\mathbf{G} = \sigma_g^2 \mathbf{A}_{txt}$ where $\mathbf{A}$ corresponds to the additive genetic numerator relationship matrix among individuals, often derived from pedigree (Henderson, 1975, 1984; Mrode, 2014; Gilmour *et al.*, 2009) or, more recently, with molecular information (VanRaden, 2008). Here, narrow-sense heritability h^2 is calculated as $h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2)$.

Estimation of β and $\mathbf{g}$ are done using mixed model equations as follows (Henderson, 1950):

$$\begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{Z} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z} + \hat{\mathbf{G}}^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ \mathbf{g} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{2–5}$$

$$\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{Z} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z} + \hat{\mathbf{G}}^{-1} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{y} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{y} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{y} \end{bmatrix} \tag{2–6}$$

$$= \begin{bmatrix} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{y} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{y} \end{bmatrix} = \begin{bmatrix} (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \\ \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y}-\mathbf{X}[\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}) \end{bmatrix}$$

$$= \begin{bmatrix} (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \\ \hat{\mathbf{G}}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\beta}) \end{bmatrix}$$

That is, $\hat{\beta} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$ which is the Empirical Best Linear Unbiased Estimator (EBLUE), and $\hat{\mathbf{g}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\beta})$ commonly referred to as Empirical Best Linear Unbiased Predictor (EBLUP). Variance components are often estimated using Restricted Maximum Likelihood (REML) assuming that both **g** and **e** have multivariate normal distributions (Patterson and Thompson, 1971).

Computing these LMM matrices with a goal to optimize field designs or at least improve the efficiency of an existing experimental design can be computationally slow especially for large experiments with thousands of entries. Butler *et al.* (2008) suggested an approximation to *A*-optimality by using the neighbor balance approach. However, an exact implementation requires the calculation of the variance-covariance matrix $\mathbf{C}^{22}$ that contains information about the random treatment effects from which the trace or log of determinant is calculated during the process of optimization. For the computation of this matrix, it has been shown by Harville (1997) and Hooks *et al.* (2009) that

$$\mathbf{C}^{22} = \mathbf{M}(\Omega) = Var(\hat{\mathbf{g}}-\mathbf{g}) = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}-\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z})^{-1} \qquad (2\text{–}7)$$
$$= (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}-\mathbf{Z}'\mathbf{K}_x\mathbf{Z})^{-1}$$

where $\mathbf{K}_x = \mathbf{R}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}$.

2.2.2 Pairwise Swap Algorithm

The procedure to improve an experimental design was implemented based on a simple pairwise swap (exchange) algorithm. Other search algorithms can be implemented; however, the focus in this study is to evaluate the potential of generating improved designs with spatial

and genetic correlations based on a swap procedure, and future studies will focus on probably faster and more efficient search algorithms. In the implemented search algorithm, random pairs of treatments belonging to the same block are swapped and evaluated using either an *A*- or *D*-optimality criterion, and in each iteration other pairs are evaluated.

The required inputs at different stages of generating the experimental designs are: 1) a choice of either *A*- or *D*- optimality criterion; 2) number of initial RCB designs (m) to be generated randomly; 3) number of "best" designs (s) to be selected from the m designs; and, 4) number of iterations (p) desired, which is the number of times the treatments will have to be swapped before a stopping rule is applied.

In brief, the swap procedure follows: (i) randomly generate m experimental layouts, Ω_i where $i = 1, 2, 3 \cdots, m$; (ii) for each layout, calculate a criterion value (that is, trace in case of *A*-optimality or determinant in case of *D*-optimality), and call it, say, τ_i; (iii) select s experimental layouts with the smallest τ_i; (iv) for each Ω_i of s_i, randomly interchange the position of a pair of treatments within a block to produce a new layout, say, Ω_j, and recalculate the new criterion value, τ_j; (v) if $\tau_i > \tau_j$, then accept τ_j and use Ω_j as the new layout, otherwise reject Ω_j; (vi) repeat steps (iv) to (v) for a total of p iterations.

The first output of the algorithm is obtained at the random generation of un-improved (initial) designs, that is, without swapping pairs of genotypes. This includes the traces and determinants (or preferably natural logarithms of the determinants for sparse matrices) of all the initial designs together with the actual layout of the s designs. The final output is a list with the final (improved) design, and other relevant matrices.

2.2.3 Algorithm Evaluation

Performance evaluation of the proposed swap algorithm was conducted based on a RCB designs that had either 30 or 196 genotypes with specified number of blocks and a combination of heritabilities (h^2 = 0.1, 0.3, and 0.6), spatial correlation levels (ρ = 0, 0.1, 0.3, 0.6 and 0.9), and genetic relationship structures that involved genetically unrelated individuals, half-sibling and full-sibling families. The evaluated conditions are only a subset of the typical field conditions

that usually occur in plant breeding and can be easily extended to other situations. A nugget effect estimate was also incorporated into the **R** matrix but fixed to zero arbitrarily to reduce the number of experimental conditions to be evaluated although Gezan *et al.* (2010) showed that incorporating nugget effects could also improve the design efficiency.

The following notations are used: $\Omega_A^{(30)}$ and $\Omega_D^{(30)}$ to identify an RCB design with $t = 30$ genotypes generated using A- and D-optimality criteria, respectively. In these scenarios, there are $b = 6$ blocks, $r_b = 5$ rows per block, $c_b = 6$ columns per block, with $R_T = 15$ total number of rows and $C_T = 12$ total columns. For half-sib families, pedigree files consisted of a structure based on five parents each with six individuals for offspring. Full-sib pedigree files consisted in a half-diallel with five parents for a total of 10 families each with three individuals. Similarly, $\Omega_A^{(196)}$ scenario identifies a RCB design with $t = 196$, with 4 blocks, $r_b = 14$, $c_b = 14$ generated based on A-optimality criterion. Pedigree for half-sib families consisted of 32 parents with approximately 6 individual offsprings each, and for full-sib families, this consisted in a several half-diallel with five parents each with eight additional crosses between diallels for a total 30 parents in 68 families each with approximately three offsprings.

Each of the evaluated conditions are shown in Table 2-1 where each combinations was replicated $\lambda = 10$ times. Each replicate had $m = 100$ initial RCB designs iterated and the best design was selected ($s = 1$) which was then optimized for $p = 5,000$ iterations to produce an improved experimental layout. Several parameters of interest were calculated including traces and determinants from the variance-covariance matrices, time taken for each run, initial and overall gains and number of successful swaps for every iteration. Time taken to run each replicate of a condition was also recorded. A 64-bit Desktop Windows Operating System Intel(R) Core(TM) i7-2600 CPU 3.40GHz, 8GB RAM was used for all evaluations. Programming was done in *R* (R Core Team, 2016) and computer code is available upon request to the authors.

2.2.4 Relative Design Efficiency

Measures of relative design efficiency were implemented as described below. Suppose that m initial designs are randomly generated for a given condition ξ_i, where $i = 1,2,3,\cdots,\xi$,

where ξ is the total number of conditions to be evaluated. Condition i is replicated j times, for $j = 1,2,3,\cdots,\lambda$. Consider a design that was generated under A-optimality criterion, let $\bar{A}_{ij} = \sum_{k=1}^{m} A_k/m$ be the average trace value from the m initial designs for condition i in replicate j. Let $A_{(min)ij}$ be the smallest trace from the m traces and $A_{(opt)ij}$ be the smallest trace achieved after p iterations of optimizing the best selected initial design that had the $A_{(min)ij}$ trace value. Hence, an Initial Design Efficiency (IDE) and an Overall Design Efficiency (ODE) for a condition i and replicate j are given by

$$\text{IDE:} \quad \alpha_{ij}^{A} = \frac{\bar{A}_{ij} - A_{(min)ij}}{\bar{A}_{ij}}; \qquad \alpha_{ij}^{D} = \frac{\bar{D}_{ij} - D_{(min)ij}}{\bar{D}_{ij}}; \tag{2–8}$$

$$\text{ODE:} \quad \gamma_{ij}^{A} = \frac{\bar{A}_{ij} - A_{(opt)ij}}{\bar{A}_{ij}}; \qquad \gamma_{ij}^{D} = \frac{\bar{D}_{ij} - D_{(opt)ij}}{\bar{D}_{ij}}; \tag{2–9}$$

where $i = 1,2,\cdots,\xi$ condition; $j = 1,2,\cdots,\lambda$ replicate. Note that when $m = 1$, the average trace value $\bar{A}_{ij}$ is simply replaced by the exact trace value of the particular design under consideration.

Summary statistics of IDE and ODE over the number of replicates per condition i were obtained.

2.2.5 Data Simulation

To evaluate the accuracy and precision of predicted random genetic effects and estimate narrow-sense heritabilities from a fitted LMM, a response variable $\mathbf{y}$ was simulated following the model

$$y_{ij} = \mu + g_{k(ij)} + E_{s(ij)} \tag{2–10}$$

where y_{ij} represents the observation on the ith row and jth column, μ is an overall mean that was arbitrarily fixed to 10 units, $g_{k(ij)}$ represents a k-th random genotype effect on the ith row and jth column and E_s is the structured residual error (Littell *et al.*, 2006; Gezan *et al.*, 2010). Similar experimental conditions as those described in Section 2.2.3 were considered. Correlated genetic and residual effects were obtained based on the Cholesky decomposition of multivariate

normal distributions with a zero expected value given by $\mathbf{G} = \sigma_g^2 \mathbf{A}$ and $\mathbf{R} = \sigma_e^2 \sum_r(\rho_r) \otimes \sum_c(\rho_c)$, respectively (Gilmour *et al.*, 2009; Mrode, 2014; Gezan *et al.*, 2010).

Three scenarios were generated: $\Omega_A^{(196)}$ with 16 blocks, $\Omega_A^{(30)}$ and $\Omega_D^{(30)}$ both with 6 blocks as described above. A total of $\xi = 12$ conditions for each scenario were evaluated, each with $\lambda = 50$ replicates, $m = s = 1$ and $p = 5,000$ iterations. For $\Omega_A^{(196)}$ scenario, the experimental layout had $t = 196$ individuals, $b = 16$, $r_b = 14$, $c_b = 14$, $R_T = 56$ and $C_T = 56$. Similarly, for both $\Omega_A^{(30)}$ and $\Omega_D^{(30)}$ scenarios, their layouts had $t = 30$ individuals, $b = 6$, $r_b = 5$, $c_b = 6$, $R_T = 15$ and $C_T = 12$. Pedigree structure for half-sib and full-sib families are identical to those described in Section 2.2.3. The spatial correlation levels evaluated under $\Omega_A^{(196)}$ scenario were $\rho = 0.3$ and 0.6 with narrow-sense heritabilities of 0.1, 0.3 and 0.6.

Analysis of data for the $\Omega_A^{(30)}$ and $\Omega_D^{(30)}$ scenarios was done by fitting two linear mixed models. Both of the models considered blocks as fixed effects, genotypes as random effects, but for the first model (Model 1) residual errors were modeled assuming no spatial correlations, whereas for the second model (Model 2) residual errors were modeled by fitting an $AR1 \otimes AR1$ correlation structure. For simplicity, under the $\Omega_A^{(196)}$ scenario, only the spatial model (Model 2) was fitted.

Pearson's product-moment correlation, r_g, between the predicted and true breeding values were computed, and estimation of heritabilities and their standard errors together with coefficient of variation (C.V. %) statistics were calculated. The software ASReml-R v. 3.0 (Gilmour *et al.*, 2009) and the R package *nadiv* (Wolak, 2012) were used to fit all models and to calculate approximated standard errors of the estimated heritabilities by using the delta method, respectively.

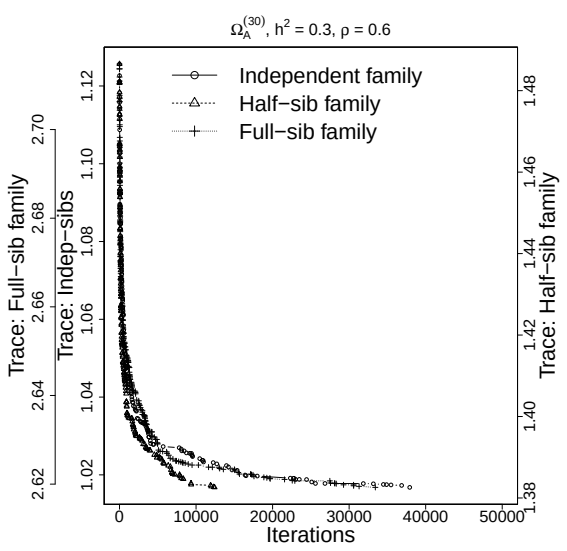
2.2.6 Motivating Example

The objective of this example is to illustrate how the rate of improvement of experimental designs varies for each condition. Details on number of successful swaps, relative design efficiencies and computational time taken for $p = 50,000$ iterations per single run are given.

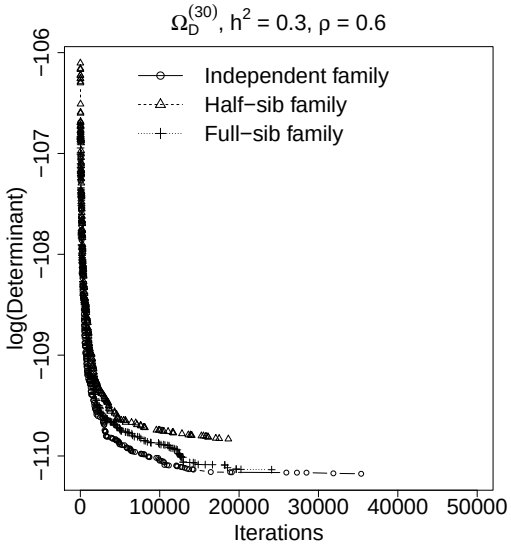
Figure 2-1 present findings based on $h^2 = 0.3$ and $\rho = 0.6$ for genetically unrelated individuals, half-sib and full-sib families for $\Omega_A^{(30)}$, $\Omega_D^{(30)}$ and $\Omega_A^{(196)}$ scenarios.

In this particular example, the highest average variance reduction of treatment effects for $\Omega_A^{(30)}$ scenario (ODE = 9.67 %) was obtained after 178 successful swaps achieved from the genetically unrelated individuals (Figure 2-1a). Half-sib individuals had ODE of 7.00 % with 172 successful swaps and full-sib had ODE of 3.52 % with 198 successful swaps. For $\Omega_D^{(30)}$ scenario, the largest reduction in the hyperplane volume (generalized variance) of the treatment effects was found to be an ODE = 3.51 % with 240 successful swaps from half-sib families. An ODE of 3.27 % with 205 successful swaps were achieved for the genetically unrelated individuals, and ODE of 3.21 % with 234 successful swaps from full-sib families (Figure 3-3b). Lower design efficiencies were obtained from the large experimental design, $\Omega_A^{(196)}$, since they require much more iterations than for small designs. The ODE was 4.52 % obtained after 2,589 successful swaps from the genetically unrelated individuals, ODE of 3.76 % with 2,685 successful swaps from half-sib families and ODE of 2.32 % with 2,616 successful swaps from full-sib families (see Figure 3-3c).

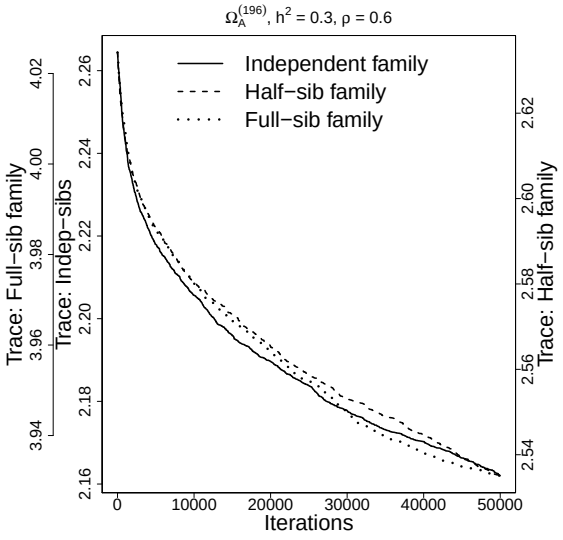
In both $\Omega_D^{(30)}$ and $\Omega_A^{(30)}$ scenarios, the rate of improvement is high in the first 5,000 to 10,000 iterations and almost flattens out thereafter. In contrast, scenario $\Omega_A^{(196)}$ (Figure 3-3c) shows that more than $50,000$ iterations might be required, and there is room for improvement of the experimental design as the slope is still steep which could explain the lower ODE values found. The optimal experimental layout from each of the above scenarios have the positions for most of the genotypes within a block changed in order to improve the design. Ideally, improved experiments tend to place siblings separately so as not to share microsites and not always be on the same side of the experimental layout as this would result in biased estimation of treatment effects due to confounding effects of spatial and genetic effects.



(a)



(b)



(c)

Figure 2-1. Design improvement trends for scenarios (a) $\Omega_A^{(30)}$, (b) $\Omega_D^{(30)}$, and (c) $\Omega_A^{(196)}$ based on $h^2 = 0.3$ and $\rho = 0.6$ for genetically unrelated individuals, half-sib and full-sib families, displaying successful traces and determinants during the optimization process based on $p = 50,000$ iterations. Scenarios $\Omega_A^{(30)}$ and $\Omega_D^{(30)}$ were generated with 6 blocks of dimensions 5 rows by 6 columns whereas $\Omega_A^{(196)}$ had 16 blocks of dimensions 14 rows by 14 columns.

2.3 Results

2.3.1 Initial and Overall Design Efficiency

From the evaluated conditions related to varying levels of heritability, genetic relatedness and spatial correlations where spatial correlation was greater than 0, Table 2-1 display summary statistics of the average IDE % and ODE % and their standard errors. The remaining conditions with zero spatial correlation ($\rho = 0$) resulted in null improvements hence, they are not shown in these tables but are presented in Figure 2-2. The results present here, show the percentage improvement in terms of reduction in average variance of the treatment effects when A-optimality criterion is used and in terms of reduction in volume of hyperplane when the D-optimality criterion is used.

As, expected, the average IDE values were all smaller than their respective ODE values. This would confirm the fact that randomly generating hundreds of experimental designs and simply choosing the best with respect to A- or D-optimality criteria results in designs with lower efficiencies compared to optimized designs. After applying the optimization procedure to improve the experimental layout, the average highest ODE = 8.739 % (S.E. = 0.065) from the optimal designs under $\Omega_A^{(30)}$ was obtained from the set of genetically unrelated individuals when $h^2 = 0.3$ and $\rho = 0.6$. Relatively lower gains were observed within half-sib and full-sib families compared to family with independent individuals. Specifically, the highest ODE of 7.262 % (0.031) among half-sib families occurred when $h^2 = 0.1$ and $\rho = 0.6$ which was also the case among full-sib families that recorded the highest ODE of 5.004 % (0.034) when $h^2 = 0.1$ and $\rho = 0.6$. Also experiments with either half-sib or full-sib families appear to achieve higher reduction of average variance of treatment effects when the heritabilities are very low (*i.e.* $h^2 = 0.1$). In addition, for any given heritability level, highest design improvement was always achieved when the spatial correlation level was 0.6.

From a relatively larger experimental design such as $\Omega_A^{(196)}$ scenario with four blocks (Table 2-1), on average, the overall highest ODE was obtained when the experiments consisted of genetically unrelated individuals with $h^2 = 0.1$ and $\rho = 0.9$ (ODE = 5.664 %, S.E. = 0.032).

Still among the genetically unrelated individuals, when $h^2 = 0.3$, large improvements occurred when $\rho = 0.9$ yielding and ODE of 4.559 % (S.E. = 0.032). However, for $h^2 = 0.6$, large design improvements (ODE = 3.213 %, S.E. = 0.036) occurred when $h^2 = 0.6$. Considering the half-sib families, overall highest reduction in average variance of treatment effects of 4.834 % (S.E. = 0.055) was observed when $h^2 = 0.1$ and $\rho = 0.9$. Similarly, among the full-sib families, highest ODE of 3.040 % (S.E. = 0.033) was obtained when $h^2 = 0.1$ and $\rho = 0.9$. These results indicate that an experiment with strong spatial correlations and with very low heritabilities may have considerable room for a design improvement of its design, over experiments with high heritabilities and low spatial correlations.

Experiments generated using *D*-optimality criterion (*i.e.* $\Omega_D^{(30)}$ scenario) had the highest reduction in volume of the hyperplane (ODE = 6.910 %, S.E. = 0.039) obtained when $h^2 = 0.1$ and $\rho = 0.9$ among the genetically unrelated individuals. Similarly, the highest ODE among half-sib family was 3.943 % (S.E. = 0.024) obtained when $h^2 = 0.1$ and $\rho = 0.9$. Experiments with full-sib families recorded highest design efficiencies of 3.114 % (S.E. = 0.023) when $h^2 = 0.3$ and $\rho = 0.6$. A Pearson's product-moment correlation of 0.98 between *A*- and *D*-optimality criteria for $\Omega_A^{(30)}$ and $\Omega_D^{(30)}$ was obtained reflecting a good agreement between these criteria. However, in practical field designs, only one of the optimality criteria is to be applied to help choose a relatively more efficient design than another based on the knowledge of existing experimental conditions.

The number of successful swaps for each condition and scenario were monitored for any possible trend. The mean number of successful swaps for independent, half-sib and full-sib families in the $\Omega_A^{(30)}$ scenario across all conditions were: 139 (min = 96, max = 208), 150 (93, 244) and 178 (97, 283). The average successful swaps under $\Omega_D^{(30)}$ scenario were 185 (144, 234), 190 (147, 257) and 199 (131, 260) for independent, half-sib and full-sib families, respectively, whereas under $\Omega_A^{(196)}$, the successful swaps for the same families were 894 (830, 959), 950 (828, 1,144) and 1,024 (844, 1,281). In general, it was noted that higher number of successful swaps were obtained when the treatments had lower heritabilities and spatial correlations.

Table 2-1. Summary statistics for Initial Design Efficiency (IDE) and Overall Design Efficiency (ODE) for RCB designs with 30 genotypes generated using A-optimality criterion $\left(\Omega_A^{(30)}\right)$ and D-optimality criterion $\left(\Omega_D^{(30)}\right)$ and for 196 genotypes generated using A-optimality criterion $\left(\Omega_A^{(196)}\right)$ with four blocks of dimensions 14 rows by 14 columns. All designs were evaluated with $\lambda = 10$ replicates per condition and iterated $p = 5,000$ times to improve the experimental layouts. ODE mean values that are starred ($\star$) are the overall largest improvements per set.

Condition			Design $\Omega_A^{(30)}$				Design $\Omega_A^{(196)}$				Design $\Omega_D^{(30)}$			
Pedigree	h^2	ρ	IDE %	S.E.	ODE %	S.E.	IDE %	S.E.	ODE %	S.E.	IDE %	S.E.	ODE %	S.E.
Indep	0.1	0.1	0.120	0.002	0.305	0.002	0.004	0.001	0.017	0.000	0.046	0.001	0.124	0.001
		0.3	0.490	0.015	2.159	0.020	0.017	0.002	0.203	0.005	0.171	0.006	0.845	0.007
		0.6	1.636	0.087	7.975	0.075	0.102	0.010	1.602	0.019	0.417	0.013	2.627	0.022
		0.9	1.566	0.078	6.948	0.062	0.382	0.062	5.664*	0.032	1.695	0.092	6.910*	0.039
	0.3	0.1	0.211	0.005	0.481	0.003	0.008	0.001	0.042	0.001	0.093	0.003	0.236	0.002
		0.3	0.792	0.038	3.038	0.016	0.046	0.008	0.507	0.007	0.320	0.007	1.418	0.012
		0.6	2.043	0.084	8.739*	0.065	0.140	0.016	2.767	0.031	0.541	0.019	3.272	0.017
		0.9	0.638	0.026	2.670	0.018	0.270	0.035	4.559	0.032	0.586	0.022	2.672	0.026
	0.6	0.1	0.227	0.006	0.483	0.005	0.013	0.002	0.063	0.003	0.098	0.001	0.249	0.002
		0.3	0.795	0.036	2.978	0.029	0.072	0.008	0.683	0.009	0.322	0.014	1.434	0.011
		0.6	1.435	0.051	6.025	0.071	0.253	0.027	3.213	0.036	0.524	0.021	3.041	0.024
		0.9	0.203	0.009	0.864	0.008	0.170	0.026	2.515	0.014	0.206	0.009	0.868	0.008
Half-sib	0.1	0.1	0.243	0.011	0.908	0.003	0.024	0.003	0.174	0.003	0.051	0.002	0.202	0.002
		0.3	0.689	0.016	3.310	0.027	0.036	0.006	0.805	0.006	0.166	0.006	0.969	0.007
		0.6	1.335	0.037	7.262*	0.031	0.148	0.026	2.015	0.014	0.433	0.024	2.601	0.021
		0.9	0.928	0.040	3.848	0.040	0.341	0.035	4.834*	0.055	0.934	0.024	3.943*	0.024
	0.3	0.1	0.218	0.006	0.686	0.003	0.016	0.002	0.144	0.002	0.092	0.003	0.283	0.001
		0.3	0.643	0.021	2.898	0.028	0.051	0.007	0.704	0.007	0.283	0.007	1.430	0.010
		0.6	1.427	0.056	6.409	0.048	0.138	0.016	2.646	0.041	0.549	0.019	3.277	0.016
		0.9	0.304	0.021	1.287	0.010	0.158	0.029	3.161	0.029	0.319	0.015	1.307	0.012
	0.6	0.1	0.168	0.005	0.416	0.004	0.010	0.001	0.094	0.002	0.101	0.003	0.265	0.003
		0.3	0.555	0.025	2.054	0.010	0.064	0.006	0.714	0.010	0.331	0.014	1.446	0.010
		0.6	0.841	0.050	3.530	0.031	0.159	0.013	2.846	0.033	0.573	0.018	3.081	0.017
		0.9	0.100	0.005	0.403	0.004	0.096	0.011	1.405	0.012	0.092	0.005	0.390	0.003
Full-sib	0.1	0.1	0.262	0.017	1.955	0.015	0.039	0.005	0.623	0.007	0.056	0.002	0.358	0.003
		0.3	0.675	0.019	4.485	0.039	0.115	0.018	1.944	0.015	0.190	0.011	1.173	0.007
		0.6	1.064	0.047	5.004*	0.034	0.116	0.018	2.829	0.032	0.362	0.010	2.419	0.016
		0.9	0.345	0.011	1.566	0.014	0.204	0.031	3.040*	0.033	0.371	0.013	1.529	0.010
	0.3	0.1	0.166	0.007	0.899	0.004	0.023	0.003	0.467	0.003	0.091	0.004	0.398	0.001
		0.3	0.457	0.019	2.153	0.023	0.087	0.012	1.259	0.007	0.257	0.010	1.458	0.013
		0.6	0.671	0.031	3.128	0.031	0.126	0.014	2.245	0.010	0.540	0.026	3.114*	0.023
		0.9	0.108	0.005	0.465	0.005	0.103	0.013	1.483	0.012	0.113	0.005	0.458	0.005
	0.6	0.1	0.087	0.003	0.271	0.002	0.014	0.002	0.204	0.001	0.094	0.002	0.284	0.004
		0.3	0.256	0.006	1.016	0.011	0.057	0.010	0.698	0.005	0.307	0.012	1.418	0.010
		0.6	0.346	0.017	1.421	0.013	0.145	0.023	1.896	0.014	0.553	0.026	3.029	0.018
		0.9	0.034	0.002	0.138	0.001	0.036	0.005	0.512	0.004	0.032	0.002	0.138	0.002

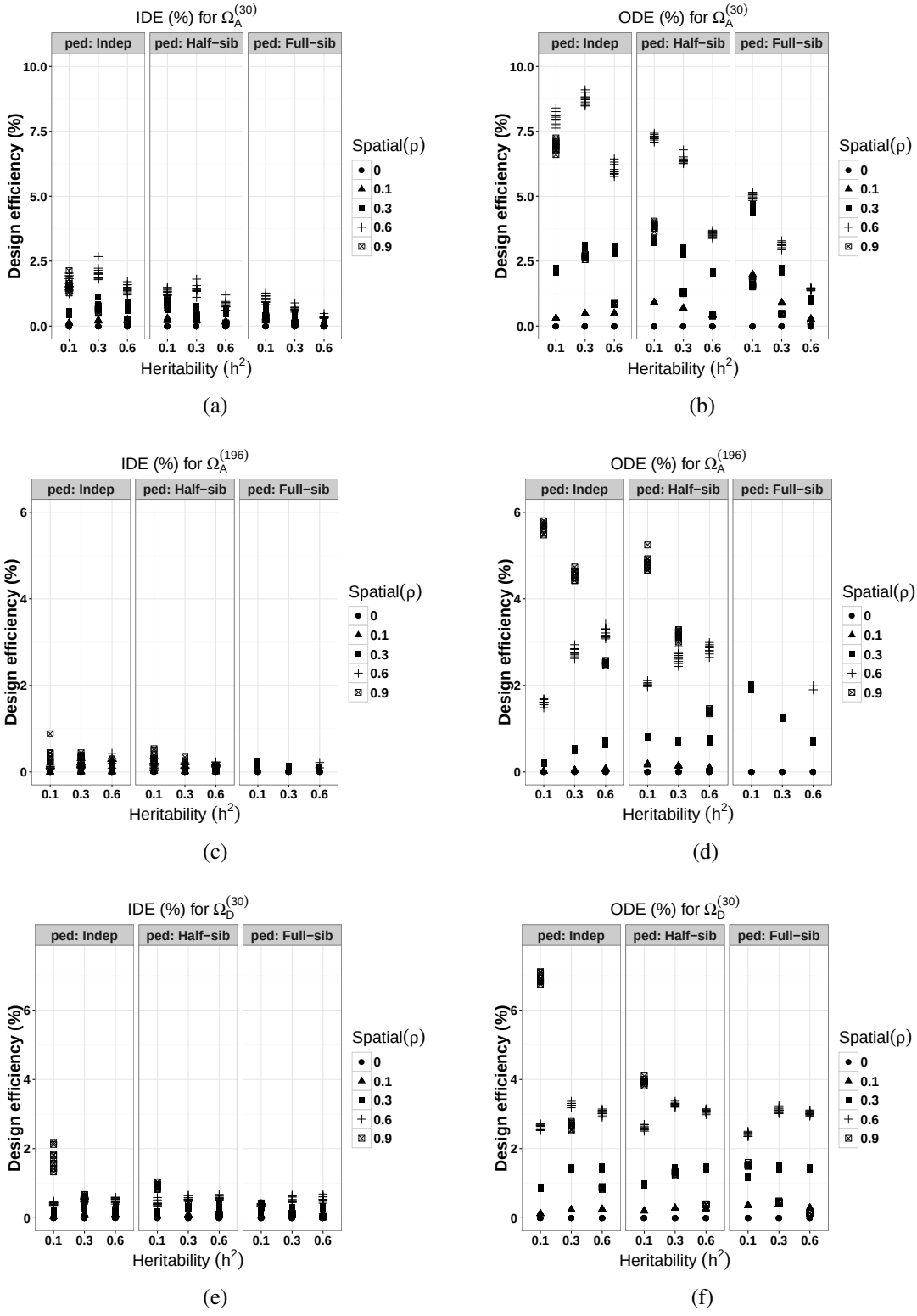


Figure 2-2. Initial design efficiency (IDE) and overall design efficiency (ODE) for varying scenarios, where $\Omega_A^{(196)}$ was generated with four blocks of dimensions 14 rows by 14 columns. Replicates were $\lambda = 10$, for $m = 100$ initial designs and $p = 5,000$ iterations.

2.3.2 Analysis of Simulated Data

A summary of the results obtained from the analysis of simulated data for assessing prediction accuracies for genetic values (BLUPs) and estimated heritabilities by fitting a LMM with and without a 2-dimensional separable spatial correlation structure denoted by Model 2 and 1, respectively are presented in Tables 2-2 and 2-3 each with 12 different conditions for scenarios $\Omega_A^{(30)}$ and $\Omega_D^{(30)}$. Descriptive statistics have been categorized based on initial (*i.e.* un-improved) and those obtained from final (*i.e.* improved) designs.

Considering results from the improved designs and for $\Omega_A^{(30)}$ scenario under Model 2, prediction accuracies of genetic values among half-sib and full-sib families were found to be very high, with a Pearson's correlation coefficient of $r_g = 0.984$ and $r_g = 0.981$, respectively, obtained when $h^2 = \rho = 0.6$. Similarly, $\Omega_D^{(30)}$ scenario under Model 2 resulted in strong prediction accuracies of genetic values among half-sib and full-sib families with Pearson's correlation coefficient of $r_g = 0.983$ and $r_g = 0.982$, respectively, also obtained when $h^2 = \rho = 0.6$. Pearson's correlation coefficients from the no-spatial model (Model 1) were relatively lower than those from Model 2 but still very strong. For instance, prediction accuracies for $\Omega_A^{(30)}$ based on Model 1 among half-sib and full-sib families were $r_g \approx 0.94$ and $r_g \approx 0.94$, respectively, obtained when $\rho = h^2 = 0.6$. As expected, the lowest predictive ability under each scenario was found when treatments had the lowest spatial and heritability values. The estimates of heritabilities were accurate with precision assessed using the coefficient of variation (C.V. %), which was larger for treatments with smaller h^2 values, decreasing with increasing h^2 in both half-sib and full-sib families under all evaluated conditions.

Results from $\Omega_A^{(196)}$ scenario (Table 2-3) had much stronger predictive ability of $r_g = 0.99$ for both half-sib and full-sib families when $\rho = h^2 = 0.6$ and $r_g \approx 0.83$ for the lowest prediction accuracy occurring when treatments had the lowest spatial ($\rho = 0.3$) and heritability levels ($h^2 = 0.1$). For each level of spatial correlation, r_g increased with increasing h^2 and similarly, for a given h^2 value r_g increased with ρ.

Precision of the estimated heritabilities as measured using C.V.% was found to be largest for smallest h^2 values, decreasing with increasing h^2 in both half-sib and full-sib families. Treatments with smaller heritability values were relatively more variable, presenting higher C.V. % than for treatments with large heritability values. In all scenarios, for a given spatial correlation level, the prediction accuracies increased with increasing heritabilities. For a given heritability value, prediction accuracies increased with increase in spatial correlations only for the spatial correlation model (*i.e.* Model 2). The C.V. % for $\Omega_A^{(196)}$ were notably smaller than for $\Omega_A^{(30)}$ and $\Omega_D^{(30)}$ scenarios, a result due to the large number of experimental units in $\Omega_A^{(196)}$.

Kernel densities of estimated heritabilities for Model 2 based on $\Omega_A^{(196)}$ scenario for four conditions are shown in Figure 2-3. Comparisons of precision of the estimated heritabilities between the initial and improved designs for full-sib and half-sib families are presented in this figure which clearly indicate that final designs are more precise than initial designs in estimating heritabilities under the evaluated conditions.

2.4 Discussion

Findings from this study indicate that both environmental and genetic factors influence the levels of design efficiency and accuracy in prediction of random genetic effects. A methodology to improve the generation of RCB designs by accounting for both spatial and genetic correlations using a mixed model approach was proposed and illustrated based on a simple pairwise algorithm and an information based optimality criterion. The linear mixed effect model was used in this study with blocks as fixed effects and treatments considered to be random effects. Note, however, that it is absolutely trivial to change these assumptions and consider treatment and blocks as fixed and random effects respectively depending on the study objective.

The illustration in Section 2.2.6 reveals that for small RCB designs, such as $\Omega_A^{(30)}$, $p =$ $10,000$ to $40,000$ iterations would substantially improve the efficiency of the design, and subsequently minimizing the average variances of the treatment effects to a great extent. On the other hand larger designs, such as $\Omega_A^{(196)}$, with many replicates would require at least $p = 50,000$ iterations. All simulations presented in this study, other than the motivating example, were based

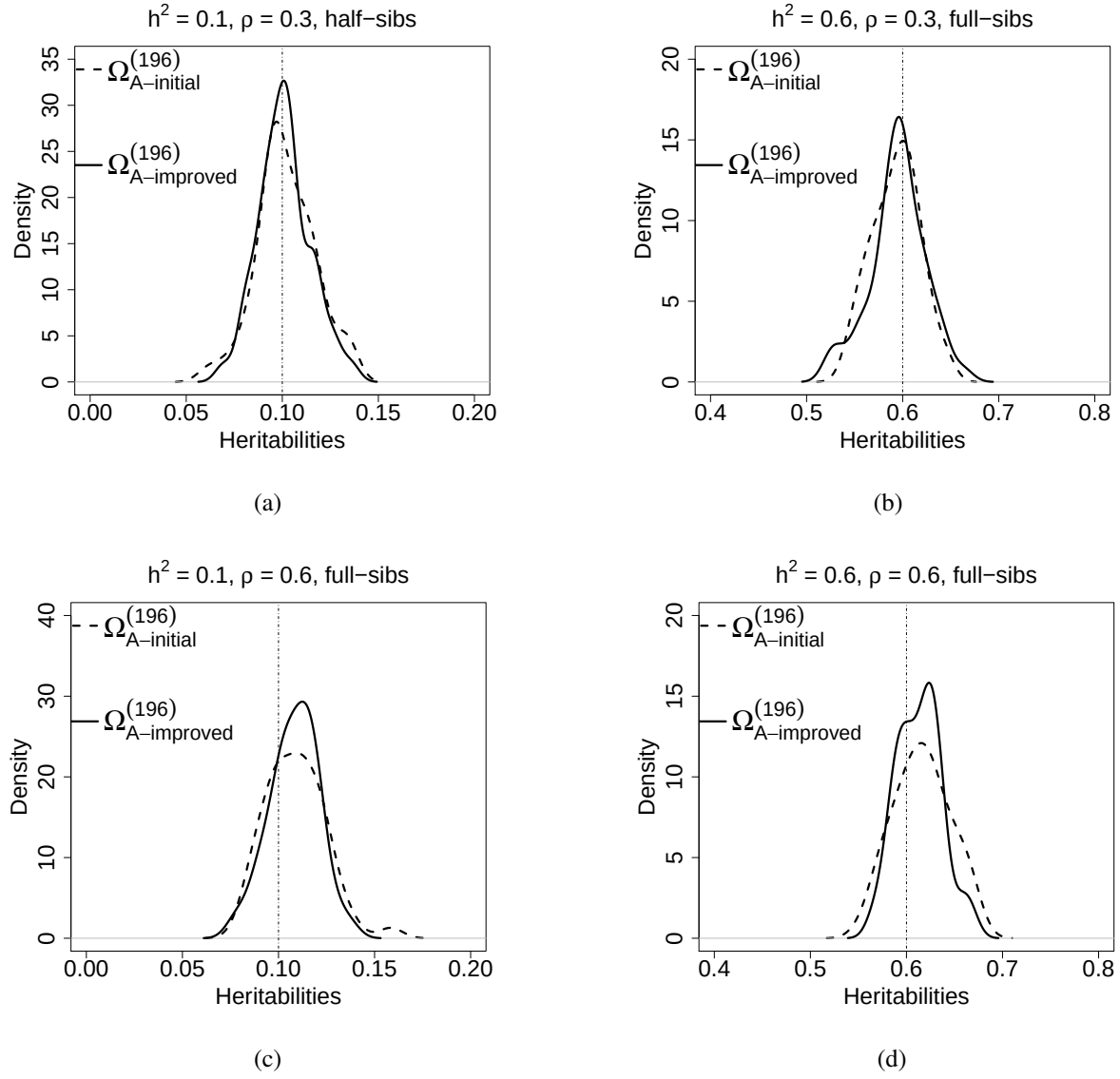


Figure 2-3. Kernel densities for estimated heritabilities, where (a) is from half-sib families and (b), (c) and (d) from full-sib families. The vertical line represents the true heritability. Both initial and improved designs for different heritabilities and spatial correlations are presented based on $\Omega_A^{(196)}$ with 16 blocks evaluated with $\lambda = 50$ replicates per condition. Each generated initial design was iterated $p = 5,000$ times.

Table 2-2. Summary statistics presented with Pearson's correlation coefficient (r_g) and estimated heritabilities ($\hat{h}^2$) together with their coefficient of variation (C.V. %) from $\Omega_A^{(30)}$ and $\Omega_D^{(30)}$ scenarios. Each condition had $\lambda = 50$ replicates each iterated $p = 5,000$ times. In Model 1, residuals were assumed to be uncorrelated and in Model 2, an $AR1 \otimes AR1$ spatial correlation structure was fitted.

				Estimates under $\Omega_A^{(30)}$						Estimates under $\Omega_D^{(30)}$					
	Conditions			Half-sib family			Full-sib family			Half-sib family			Full-sib family		
Model	ρ	h^2	Design	$\hat{h}^2$	C.V. %	r_g	$\hat{h}^2$	C.V. %	r_g	$\hat{h}^2$	C.V. %	r_g	$\hat{h}^2$	C.V. %	r_g
Model 1	0.3	0.1	Initial	0.091	56.932	0.626	0.105	57.551	0.647	0.107	59.429	0.638	0.122	56.807	0.675
			Improved	0.107	57.958	0.625	0.107	67.372	0.660	0.130	49.984	0.666	0.129	61.648	0.682
		0.3	Initial	0.303	25.573	0.845	0.288	29.828	0.830	0.309	31.097	0.851	0.307	31.874	0.833
			Improved	0.302	30.783	0.854	0.311	29.985	0.827	0.315	26.223	0.85	0.301	32.012	0.839
		0.6	Initial	0.609	10.661	0.946	0.592	14.29	0.939	0.581	13.029	0.942	0.578	14.226	0.932
			Improved	0.594	11.505	0.943	0.574	14.425	0.932	0.582	15.934	0.941	0.574	18.943	0.923
	0.6	0.1	Initial	0.119	48.321	0.644	0.117	61.06	0.618	0.126	46.39	0.659	0.114	61.02	0.682
			Improved	0.157	49.696	0.643	0.142	49.682	0.648	0.126	54.187	0.604	0.142	52.276	0.615
		0.3	Initial	0.291	30.602	0.852	0.317	26.024	0.844	0.315	26.474	0.859	0.308	32.348	0.832
			Improved	0.341	25.837	0.844	0.317	30.52	0.809	0.361	25.528	0.847	0.326	28.804	0.826
		0.6	Initial	0.602	12.046	0.950	0.613	11.383	0.940	0.601	13.728	0.945	0.608	15.518	0.939
			Improved	0.614	12.716	0.943	0.615	13.101	0.935	0.618	10.361	0.94	0.632	12.177	0.937
Model 2	0.3	0.1	Initial	0.088	55.335	0.685	0.1	58.724	0.665	0.088	61.306	0.686	0.106	56.899	0.694
			Improved	0.092	60.271	0.67	0.094	61.836	0.694	0.111	49.259	0.716	0.099	56.763	0.707
		0.3	Initial	0.284	25.788	0.874	0.282	29.295	0.857	0.307	27.117	0.876	0.302	29.865	0.857
			Improved	0.293	32.026	0.886	0.295	27.45	0.859	0.299	23.427	0.883	0.296	29.435	0.873
		0.6	Initial	0.594	12.037	0.954	0.582	14.984	0.949	0.577	11.451	0.955	0.560	14.18	0.944
			Improved	0.585	11.628	0.957	0.562	14.391	0.949	0.569	15.535	0.957	0.557	19.756	0.943
	0.6	0.1	Initial	0.091	38.893	0.83	0.087	51.023	0.797	0.095	35.528	0.824	0.088	45.976	0.802
			Improved	0.099	35.675	0.858	0.088	37.522	0.828	0.090	36.586	0.844	0.084	36.736	0.832
		0.3	Initial	0.268	22.697	0.943	0.284	23.124	0.938	0.278	22.175	0.944	0.267	25.216	0.933
			Improved	0.292	21.367	0.951	0.275	20.051	0.937	0.295	20.827	0.953	0.282	23.933	0.940
		0.6	Initial	0.565	13.205	0.982	0.571	10.841	0.979	0.565	13.732	0.981	0.569	15.987	0.977
			Improved	0.571	13.579	0.984	0.571	12.637	0.981	0.560	17.916	0.983	0.576	13.523	0.982

on $5,000$ iterations given that time was a limiting factor as the optimizing procedure could be computationally intensive. However, this number of iterations gave a sense of the magnitude of relative design efficiency and prediction accuracies expected to be achieved under the conditions evaluated, and do not limit the use of this algorithm operationally.

Our findings have shown that relative design efficiency varies according to existing experimental conditions including heritability and spatial correlation, genetic relationships, size of the experimental design, and optimality criterion of preference. This current study

Table 2-3. Summary statistics based on Model 2 with Pearson's correlation coefficient (r_g) and estimated heritabilties ($\hat{h}^2$) together with their coefficient of variation (C.V. %). A $\Omega_A^{(196)}$ scenario with 16 blocks of dimensions 14 rows by 14 columns is presented with $\lambda = 50$ replicates per condition each iterated $p = 5,000$ times. Residuals errors were modeled using $AR1 \otimes AR1$ spatial structure.

Conditions			Half-sib family			Full-sib family		
ρ	h^2	Design	$\hat{h}^2$	C.V. %	r_g	$\hat{h}^2$	C.V. %	r_g
0.3	0.1	Initial	0.100	14.279	0.855	0.102	15.152	0.839
		Improved	0.102	14.39	0.860	0.101	13.428	0.837
	0.3	Initial	0.301	8.125	0.948	0.295	9.237	0.946
		Improved	0.303	8.011	0.950	0.306	7.070	0.949
	0.6	Initial	0.593	4.172	0.983	0.601	4.727	0.984
		Improved	0.595	4.851	0.983	0.593	3.956	0.983
0.6	0.1	Initial	0.109	14.169	0.900	0.107	11.142	0.893
		Improved	0.108	11.813	0.902	0.106	13.072	0.895
	0.3	Initial	0.321	6.999	0.967	0.319	7.669	0.969
		Improved	0.314	7.649	0.969	0.311	7.458	0.969
	0.6	Initial	0.616	4.784	0.990	0.614	3.416	0.990
		Improved	0.613	3.746	0.990	0.609	4.206	0.990

has also shown that, based on A-optimality criterion, that the highest design efficiency can be achieved among genetically unrelated individuals with $h^2 = 0.3$ and $\rho = 0.6$. Both small and large experiments with half-sib and full-sib families can achieve greater improvements under low heritability levels of 0.1 and spatial correlations of 0.6. Filho and Gilmour (2003) also reported that larger improvements are found on those studies with genetically unrelated individuals where they accounted for genetic relatedness but residuals were assumed independent, thus not modeling spatial correlation. Relative design efficiencies based on D-optimality showed that the highest values of 6.910% (on average) can be achieved among genetically unrelated individuals when heritability is lowest at 0.1 and a strong spatial correlation of 0.9.

Unlike other studies that have discussed optimality procedures by fitting mainly fixed effects models (Das, 2002; John and Williams, 1995), the implemented procedure provides with results

that are practical for an array of scenarios for field experiments that present genetic relationships and/or spatial correlations. Further, this study has shown that, in RCB designs, under the absence of spatial correlations, $\rho = 0$, there are no gains in optimizing the designs using the implemented swap algorithm regardless of the level of h^2 and genetic correlations. However, in practice for any field trial there is always some level of spatial correlations due to physical proximity of the treatments, and all initial designs can be improved following the proposed procedure. Even more important, these results indicate that generating hundreds or thousands of unimproved initial experimental designs without using an optimizing procedure does not achieve significant improvements as when optimal procedures are implemented.

The strong Pearson's product-moment correlation of ≈ 0.98 between A- and D-optimality criteria is not unusual as both criteria are a convex function of the eigenvalues of the information matrix (Das, 2002; Kuhfeld, 2010). These results are in line with their mathematical definitions, since A-optimality is a function of the arithmetic mean of the eigenvalues and D-optimality is a function of the geometric mean of the eigenvalues (Kuhfeld, 2010). Hence, ODE increases as the average variances of the treatment effects decreases.

For the data simulated in this study under scenario $\Omega_A^{(196)}$ to evaluate prediction of genetic values and estimation of heritabilities, high prediction accuracies (≥ 0.90) were obtained from both initial and improved experimental layouts when $h^2 = 0.6$ and $\rho = 0.6$ from both Model 1 and Model 2. As expected, better prediction accuracies were found for Model 2 compared to Model 1, and more importantly predictions were more accurate from improved designs compared to initial designs under Model 2. This is likely the result of appropriately modeling spatially correlated errors which was not the case for Model 1. No clear trend of predictive ability of genetic values was found between initial and improved designs under Model 1.

Estimation of heritabilities was considerably more precise for both models under improved designs. The current study has found that genotypes with small heritability values will exhibit larger C.V. % compared to genotypes with large heritability values. Also, when the spatial correlations are low, the C.V. % for estimated heritability, as expected, is larger, and vice-versa.

Results from Tables 2-2 and 2-3 clearly show that for large experiments, smaller C.V. % values for heritabilities are obtained, compared to similar conditions, on smaller experiments. The prediction accuracies of genetic values from larger experiments is higher compared to that from small experiments but similar trends are observed as the accuracy increases with increasing levels of heritabilities.

In quantitative genetics it is considered that the phenotype of an individual is explained by a genetic and an environmental component using the expression: $P = G + E$ (Falconer and Mackay, 1996). However, the genetic component can be further partitioned into additive and non-additive (dominance, epistasis) components, with $G = A + D + I$. The algorithm implemented in this study focused in the estimation of additive effects. This is reasonable as some research have shown that in many plant and animal studies additive variance accounts for more than half, and in some cases about 100%, of the total genetic variance (Hill *et al.*, 2008). Also, these authors indicated that often presence of common environmental effects within full-sib families make it difficult to estimate dominance and epistatic components accurately due to potential confounding. Nevertheless, the current algorithm can be extended to other situations by combining more than one component. For example, if total genetic variance, and therefore broad-sense heritability, is known then this can be used in place of the narrow-sense heritability; however, additional assumptions will be required for the 'relatedness' between these effects, where independence can be used (*i.e.* $\mathbf{G} = \sigma_g^2\mathbf{I}$), or an approximation proportional to the relationship matrix ($\mathbf{G} = k\mathbf{A}$).

Pedigree-based genetic relationships have been used to estimate BLUP effects (Falconer and Mackay, 1996; Henderson, 1950; Patterson and Thompson, 1971). An alternative approach would be to use molecular markers to implement Genomic BLUP (GBLUP) commonly used in some genomic selection studies (Beaulieu *et al.*, 2014; Hill *et al.*, 2008; VanRaden, 2008). Here, an 'observed' relationship matrix is obtained based on molecular information, which replaces the original $\mathbf{A}$ matrix. Beaulieu *et al.* (2014) pointed out that a larger dataset with dense marker arrays and closely related individuals per family would be required in the case of marker-based models in order to achieve similar or higher prediction accuracies than those achievable by using

pedigree-based models. Also, Habier *et al.* (2007) stated that genomic prediction accuracies might yield superior results compared to pedigree-based if markers are in linkage disequilibrium with causal loci.

The proposed optimization procedure can be extended to other orthogonal and non-orthogonal complex experimental designs such as augmented designs, incomplete block, row-column and unequally replicated experiments. Other than using $AR1$ variance structure to model spatial correlations, other variance-covariance structures can be incorporated into the LMM framework to optimize designs (Stroup, 2013; Cressie, 1993; Gilmour *et al.*, 2009; Zuur *et al.*, 2009; Littell *et al.*, 2006). In addition, extensions to other stochastic search algorithms such as simulated annealing can also be implemented.

A limitation of this study was the long computation time required to generate improved designs for large experiments. The time taken for $\Omega_A^{(30)}$ scenario fro 5,000 iterations, on a 64-bit desktop computer was, on average, $\approx$ 3 minutes and for $\Omega_A^{(196)}$ scenario it took about 30 minutes. Further improvements are in progress with the implementation of more efficient computational routines and other object-oriented programming languages.

2.5 Conclusion

This study has demonstrated that simultaneous considerations for genetic and environmental correlations can be incorporated to generate better experimental designs with important improvements in relative design efficiency and prediction accuracies of random treatment effects. Also, for RCB designs, higher relative design efficiencies are achievable from genetically unrelated individuals compared to experiments with half-sib and full-sib families. For RCB designs with half-sib or full-sib families, optimization procedure may yield to important improvements under the presence of mild to strong spatial correlation levels and relatively low heritability values. As expected, accuracy of prediction of genetic values improves as levels of heritability and spatial correlations increase. Furthermore, improved designs present more precise estimates of heritabilities than their un-improved counterparts.

CHAPTER 3
EVALUATING ALGORITHMS EFFICIENCIES FOR EXPERIMENTAL DESIGNS WITH CORRELATED DATA

3.1 Introduction

Experimental designs in agricultural and forestry field trials are often conducted with an aim to evaluate and select best treatments (genotypes) with superior phenotypes for future breeding (Piepho *et al.*, 2008). Statistical principles that are useful in constructing experimental designs are replication, randomization and blocking (Welham *et al.*, 2015) together with an appropriate choice of physical layout and treatments such that results of an experiment can be inferred to a larger population. Replication ensures that estimates of treatment effects are reliable by repeating each treatment on many experimental units, and also replicated observations helps to control for background variations between experimental units and test for differences between treatments and their precision. Randomization ensures that allocation of treatments to experimental units is fair in order to reduce bias and mimic natural variation between units. On the other hand, blocking is done in such a way that treatments in similar blocks are more uniform (homozygous) than treatments across blocks to minimize background variation which increases precision and possibility to detect differences between treatments (Sarker and Singh, 2015; Welham *et al.*, 2015).

The process of improving experimental designs is often ignored due to intensive computational requirements especially for large experimental designs. However, John and Williams (1995) and Williams *et al.* (2006) have discussed efficient procedures to construct experimental designs for incomplete blocks, row-column and other cyclic designs, based on the assumption that treatments are fixed effects. Computational issues become more when large numbers of treatment levels are tested, and therefore variance-covariance matrices are difficult to compute, particularly if a model include random factors (such as blocks or genotypes). In addition, genetic relationships for plant breeding are important in design of experiments, which are often available as pedigree files or molecular markers; however, this information is often ignored. At the same time, it has been shown by many other studies such as Gezan *et al.*

(2010) that modeling spatial correlations (e.g. using autorregressive variance structures) in plant breeding results in more efficient designs than assuming residual errors are independent and identically distributed. Something that can be also incorporated into the generation of improved designs. The framework of linear mixed models is advantageous over traditional linear models since they allow specification of appropriate variance-covariance structures for both factors (e.g. genetic) and errors (e.g. environmental noise), providing greater flexibility.

To improve experimental designs, an optimality criterion is used that has to be maximized or minimized in the implemented search algorithm. *A*- and *D*- information based criterion are the most widely used procedures in field experiments to generate designs (Chernoff, 1953; Cullis *et al.*, 2006, 1989) and are also useful in the process of selecting the optimal design (Kuhfeld, 2010). *A*-optimality criterion minimizes the average variance of random treatment effects (see Chapter 2 for more details). It is expressed as: $A_{optim} = \text{argmin}\{trace[\mathbf{M}(\Omega)]\}$, where $\mathbf{M}(\Omega)$ is the inverse of an information matrix of the treatment (or genetic) effects from a given design layout Ω. *D*-optimality was introduced by Wald (1943) and minimizes the determinant of $\mathbf{M}(\Omega)$ which can be interpreted as minimizing the generalized variance of the treatment effects (Kuhfeld, 2010) by choosing designs which the volume of the joint confidence ellipsoid is minimized (Das, 2002), and is given by $D_{optim} = \text{argmin}\{|\mathbf{M}(\Omega)|\}$ for $|\mathbf{M}(\Omega)| \neq 0$. .

There are many search algorithms that can be used to find improved designs. Often, these involve interchanging pairs of treatments and re-evaluating the new layout. Some of the computer algorithms available include: 1) pairwise swap procedure, and its variants where a single or multiple pairs of treatments are swapped at a time (John and Williams, 1995), and simulated annealing (SA) where a cooling strategy is employed (Kirkpatrick *et al.*, 1983), among others. Most of the applications of these algorithms focus on the analysis of data and very little has been done on their applications to improve the designs of such experiments, yet, estimated parameters from improved designs are obtained with increased precision since variability of treatment effects is reduced.

The main objective of this study is to evaluate the efficiency of diverse search algorithms to generate improved randomized complete block (RCB) designs applying *A*- or *D*-optimality criteria, while accounting for both spatial and genetic correlations using linear mixed models with applications in plant breeding trials. This will be done by initially generating experimental layouts through a random process and later applying an array of proposed search algorithms to improve the initial experimental layouts. Several varying field conditions that include a range of heritabilites, genetic relatedness structures and spatial correlations are evaluated.

3.2 Materials and Methods

3.2.1 Statistical Model

The following linear mixed effect model (LMM) was used:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{g} + \mathbf{e}, \tag{3–1}$$

where $\mathbf{y}$ is a vector of observed phenotypes (responses); $\mathbf{X}$ is an incidence matrix of fixed block effects; β is a vector of fixed block effects; $\mathbf{Z}$ is an incidence matrix of random treatment effects; $\mathbf{g}$ is a vector of random treatment effects, with $\mathbf{g} \sim MVN(\mathbf{0}, \mathbf{G})$, where $\mathbf{G} = \sigma_g^2\mathbf{A}$ for genetically correlated observations with $\mathbf{A}$ being a numerator relationship matrix calculated from pedigree files to account for additive genetic relatedness between individuals, and $\mathbf{G} = \sigma_g^2\mathbf{I}$ for genetically unrelated individuals. Also, $\mathbf{e}$ is a vector of residual errors, with, $\mathbf{e} \sim MVN(\mathbf{0}, \mathbf{R})$ for spatially correlated data that is modeled with an autoregressive error structure of order 1 (Gilmour *et al.*, 2009) as

$$\mathbf{R} = \sigma_e^2 \sum_r (\rho_r) \otimes \sum_c (\rho_c); \tag{3–2}$$

with $Var(e_{ij}) = \sigma_e^2$ and $Cov(e_{ij}, e_{i'j'}) = \sigma_e^2 \rho_r^{|dx|} \rho_c^{|dy|}$, where $|dx| = |x_i - x_{i'}|$ and $|dy| = |y_j - y_{j'}|$ are the row and column absolute distances, respectively; $\otimes$ is a Kronecker product; and $\sum_r(\rho_r)$ and $\sum_c(\rho_c)$ are matrices with autocorrelation parameters ρ_r and ρ_c for rows and columns respectively. If residuals are assumed to be independent and identically distributed then $\mathbf{R} = \sigma_e^2\mathbf{I}_n$. To obtain the variance-covariance matrix of random treatment effects, linear mixed model normal equations

are solved as described by Henderson (1950) and their computations implemented as discussed by Hooks *et al.* (2009); Harville (1997) to give

$$\mathbf{M}(\Omega) = Var(\hat{\mathbf{g}} - \mathbf{g}) = (\mathbf{Z}^{'}\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} - \mathbf{Z}^{'}\mathbf{R}^{-1}\mathbf{X}(\mathbf{X}^{'}\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{R}^{-1}\mathbf{Z})^{-1} \tag{3–3}$$

from which the trace and determinant of the matrix $\mathbf{M}(\Omega)$ are calculated based on *A*- and *D*-optimality criteria, respectively.

3.2.2 Algorithms

The proposed algorithms comprise of 1) simple pairwise (SP), that only swaps a pair of treatments at a time; 2) variants of pairwise procedure that swaps a group of α treatments at a time, and is identified as greedy pairwise (GP); 3) a genetic neighborhood (GN), that takes into consideration the genetic relatedness of the direct neighboring experimental units; and 4) simulated annealing (SA), that also swaps a pair of treatments at a time, but also accepts poor designs with a given probability which diminishes with time.

For any of these algorithms, the procedure involves randomly generating m initial experimental layouts, denoted as Ω_i. For each layout, the variance-covariance matrix of the treatment effects $\mathbf{M}(\Omega)$ is obtained, and its criterion value is calculated (as trace or determinant). Next, from the m designs, the single “best" initial (non-improved) experimental layout with the best criterion value is selected. After this, a given optimization algorithm is applied for p iterations. For all implemented algorithms, the output is a list of objects including the improved experimental layout, a vector with criterion values and iterations of the sequentially accepted (successful) designs, and a vector of all criterion values from all iterations, whether the swap was successful or not. Following is a detailed description of the implemented algorithms.

3.2.2.1 Simple pairwise algorithm

For the SP algorithm, the following steps are undertaken after selecting the best initial Ω_i with criteria value τ_i: 1) randomly interchange a single pair of treatments within a randomly selected block to produce a new layout, Ω_j; 2) re-calculate a new criterion value τ_j; 3) if $\tau_i > \tau_j$,

accept Ω_j as the new layout; and 4) repeat steps 1 to 3 for a total of p iterations and produce the output.

3.2.2.2 Greedy algorithm

Greedy algorithms (GP) are more aggressive variants of the simple pairwise algorithm (SP) that allow multiple treatments to be randomly interchanged within a block. In order to evaluate a spectrum of alternative implementations, this algorithm was implemented by varying the number of treatments to be swapped simultaneously, denoted as $GP\alpha$, where α refers to the number of treatments swapped. The algorithm allows specification of any even number of treatments to be swapped at a time. Tested procedure were denoted as GP4, GP14 and GP98 for randomly swapping 4, 14 and 98 treatments simultaneously on each iteration within a randomly selected block, respectively. Numbers 14 and 98 were chosen as a proportion (or 50 %) of the treatments to be swapped at a time, in an experiment with 30 and 196 treatments, respectively, whereas 4 was chosen as a close value to 2 to detect any small changes in improvement of the design when a single pair or double pairs of treatments are swapped in each iteration. Steps 1 to 4 apply as described under the SP procedure.

3.2.2.3 Genetic neighborhood algorithm

The GN algorithm is defined as a method that makes use of genetic relatedness of the eight neighboring experimental units found in a 3×3 matrix using information provided by a numerator relationship matrix ($\mathbf{A}$) of the corresponding genotypes. Steps for this algorithm are: 1) randomly generate m initial designs and select the best (Ω_i) with the smallest trace, τ_i; 2) randomly select a treatment t_l from Ω_i; 3) identify the genetic correlation coefficients from the numerator relationship matrix for all experimental units within the nearest neighborhood of t_l; 4) if there exists a pairwise genetic relationship of 0.25 or higher between t_l and any other treatment t_k for $l \neq k$ within the neighboring matrix; 5) replace either one of the treatments with a another treatment that is at a distance of more than a unit (row or column) away; 6) if there are no treatments further than a unit away even though these neighbors are genetically correlated, randomly interchange t_l with t_k; 7) calculate the new criterion value, τ_j, based on the new design

layout Ω_j; 8) if $\tau_i > \tau_j$, accept Ω_j, otherwise reject Ω_j; and 9) repeat steps 2 to 8 for a total of p iterations. Note that if all the experimental units from a neighborhood are genetically unrelated, then the SP is applied.

3.2.2.4 Simulated annealing algorithm

SA is a probabilistic meta-heuristic and stochastic optimization procedure that prevents the search from getting trapped in a local optima by accepting some solutions with a set probability and lowering the temperature with time to make sure that poorer solutions are accepted with lower probabilities (Robert and Casella, 2010). The SA algorithm implemented in this study is described as follows: 1) randomly interchange a pair of treatments within a randomly selected block to produce a new layout, Ω_j and re-calculate a criterion value, τ_j; 2) if $\tau_i > \tau_j$, accept Ω_j as the new layout with probability 1.0; else do the following, 3) calculate $\triangle = \tau_j - \tau_i$ and set a cooling temperature $T_c[i] = 1/i$, for each i-th iteration, and calculate $v = \exp(-\triangle / T_c[i])$; 4) draw a random value u from a uniform distribution, and if $u < v$ accept Ω_j; and 5) repeat steps 1 to 4 for a total of p iterations.

3.2.3 Evaluation of Algorithms

The proposed algorithms were evaluated under varying experimental conditions to assess their effectiveness to improve a RCB design. For the greedy algorithm, a total of 4, 14 and 98 random pairs of swaps were performed simultaneously within a given random block. These are identified as GP4, GP14 and GP98. Conditions considered include narrow-sense heritabilities, h^2, of 0.1, 0.3, and 0.6, where $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$; unrelated individuals (Indep), half-sib and full-sib families; and a spatial correlation of $\rho = 0.6$. Every combination of conditions was repeated $\lambda = 10$ times for $p = 5,000$ iterations. All implementation and evaluation of algorithms was done using *R* (R Core Team, 2016).

Two main scenarios were considered. First, $\Omega_A^{(30)}$ and $\Omega_D^{(30)}$ that represent RCB designs with 30 genotypes generated using *A*- and *D*-optimality criteria, respectively. In these layouts, the designs had six blocks each of dimensions five rows by six columns. Here, pedigree from half-sib

families consisted of five male parents each with six individuals, and full-sib families consisted on a half-diallel with five parents for a total of 10 families each with three individuals.

The second scenario, $\Omega_A^{(196)}$ represents an RCB design with 196 genotypes generated using *A*-optimality criterion with four blocks of dimensions 14 rows by 14 columns per block. Pedigree files for half-sib families had 32 known parents each with six offspring, whereas full-sib families had 30 parents with several half-diallels for a total of 68 families each with approximately three offspring.

Note that, GP98 was implemented only for $\Omega_A^{(196)}$ scenario to swap 50 % of the total genotypes at every single iteration, and GP14 represents swapping about 50 % for $\Omega_A^{(30)}$ and $\Omega_D^{(30)}$ scenarios.

A motivating example was evaluated for all algorithms to investigate the level of design efficiencies and rates of convergence that can be obtained for a specified condition with all algorithms having to improve the same initial RCB experimental design. This was done using *A*-optimality criterion for an experiment with 30 genotypes, 6 blocks of sizes 5 rows and 6 columns, and comprised of half-sib families with five male parents each with six individuals, for $h^2 = 0.1$, $\rho = 0.6$, and a nugget effects of 0.1. Initially, $m = 1,000$ designs were randomly generated and the best one selected for optimization. All the proposed algorithms were made to improve this initial design by going through $p = 20,000$ iterations. Traces from both successful and unsuccessful swaps were observed together with the time taken for each algorithm. A swap was defined to be successful if the resulting design had a smaller trace than the previous as this translates to a reduction in average variance of the treatment effects (Das, 2002). The motivating example was run from a 64-bit windows operating system Intel(R) Core(TM) i7-4720HQ CPU@2.60GHz, RAM = 8.0GB using *R* (R Core Team, 2016).

In order to evaluate the improvement of a design, a relative overall design efficiency (ODE%), that quantifies how efficient the improved design is relative to an initially non-improved

design, and it was calculated as:

$$\text{Overall design efficiency (ODE):} \quad \gamma_{ij}^{A} = \frac{\bar{A}_{ij} - A_{(opt)ij}}{\bar{A}_{ij}}; \qquad \gamma_{ij}^{D} = \frac{\bar{D}_{ij} - D_{(opt)ij}}{\bar{D}_{ij}}; \tag{3–4}$$

$$i = 1, 2, \cdots, \xi \quad \text{condition}; \quad j = 1, 2, \cdots, \lambda \quad \text{replicate}$$

where $\bar{A}_{ij}$ and $\bar{D}_{ij}$ are averages of m initial traces and log(determinants), respectively, for i-th condition and j-th replicate, $A_{(opt)ij}$ and $D_{(opt)ij}$ are the smallest trace and log(determinant), respectively, obtained from an improved design. Estimates of ODE% over the $\lambda = 10$ replicates per condition were summarized.

3.3 Results

Results from the motivating example that was conducted for an RCB design with $h^2 = 0.1$ and $\rho = 0.6$ based on $\Omega_A^{(30)}$ with a nugget effect of 0.1 for an experiment with 6 blocks of 5 rows by 6 columns, are displayed in Figure 3-1 which plots traces obtained from successful swaps and their overall design efficiencies and Figure 3-2 showing the rate of convergence by plotting all the 20,000 traces obtained for each algorithm. From this illustration, the results indicate that simple pairwise (SP) algorithm had the highest design efficiency of 6.713 % with the highest number of successful swaps = 192 and took about 5.8 minutes for the 20,000 iterations. This was closely followed by the simulated annealing (SA) algorithm that had an ODE = 6.258 % with 139 successful swaps and took about 5.8 minutes. GP4 algorithm had an ODE = 5.552 % with 104 successful swaps and also took about 5.8 minutes and genetic neighborhood (GN) algorithm recorded the lowest ODE = 2.053 % with 12 successful swaps and took about 6.1 minutes.

Means and standard errors (S.E.) of overall design efficiency (ODE %) for the three scenarios, that is, $\Omega_A^{(30)}$, $\Omega_A^{(196)}$ and $\Omega_D^{(30)}$ for all algorithms are presented in Tables 3-1, 3-2 and 3-3, respectively. Figure 3-3 display visible trends of ODEs by genetic relatedness and heritability levels whereas Figure 3-4 shows the average number of successful swaps out of 5,000 (that is, swaps that were accepted due to the resulting design having a smaller criterion value than the previous) for each algorithm. Results indicate that for all experiments conducted based on $\Omega_A^{(30)}$ and $\Omega_A^{(196)}$ scenarios, simulated annealing (SA) and simple pairwise (SP) algorithms

achieved the highest ODE means in all evaluated conditions followed by GP4 (for $\Omega_A^{(30)}$) or GP98 (for $\Omega_A^{(196)}$) and lowest for genetic neighborhood (GN). Also, the overall highest ODEs were achieved when $h^2 = 0.3$ among genetically unrelated individuals for all algorithms. Among full-sib families, highest ODEs were achieved when $h^2 = 0.1$ and decreased with increasing heritability for all algorithms evaluated under $\Omega_A^{(30)}$ and $\Omega_A^{(196)}$ scenarios. SA recorded the highest average ODE = 7.403 % (S.E. = 0.063) followed by SP with average ODE = 7.398 % (S.E. = 0.066) all obtained when $h^2 = 0.3$ among genetically unrelated individuals. Algorithms SA, SP, GP4 and GP14 evaluated with half-sib families under $\Omega_A^{(30)}$ had highest ODEs obtained for treatments with lowest heritability of 0.1, whereas GN achieved its highest ODE when $h^2 = 0.3$ for the same family.

Based on $\Omega_D^{(30)}$ scenario, the best performing algorithm with highest average ODE among all conditions was SP, closely followed by GP4, GP14, GN and SA which recorded the lowest average ODE. Under this scenario, the overall highest ODEs were observed among genetically unrelated individuals for SP, GP4, and GP14 when $h^2 = 0.3$. Among half-sib families, highest ODEs occurred when $h^2 = 0.3$ but no clear trend among full-sib families.

Both $\Omega_A^{(30)}$ and $\Omega_D^{(30)}$ took, on average, about 2 to 3 minutes to improve a given initial experimental design for $p = 5,000$ iterations whereas $\Omega_A^{(196)}$ required about 25 to 40 minutes for the same number of iterations. Figure 3-4 indicate that the number of successful swaps decrease with increasing heritability especially for $\Omega_A^{(30)}$ and $\Omega_A^{(196)}$ scenarios with small difference in numbers between SA and SP algorithms but presents with large differences under $\Omega_D^{(30)}$ scenario. The number of successful swaps out of 5,000 appeared to be highest for SA and SP under *A*-optimality criterion. From $\Omega_D^{(30)}$ scenario, the number of successful swaps were highest for SA which recorded above 2,500 out of the 5,000 swaps but this was not realized in terms of improving the design efficiency under this criterion.

3.4 Discussion

Algorithms are used in research for instance, to optimize long term forest planning management using SA (Borges *et al.*, 2014), to estimate the optimum combination of stand

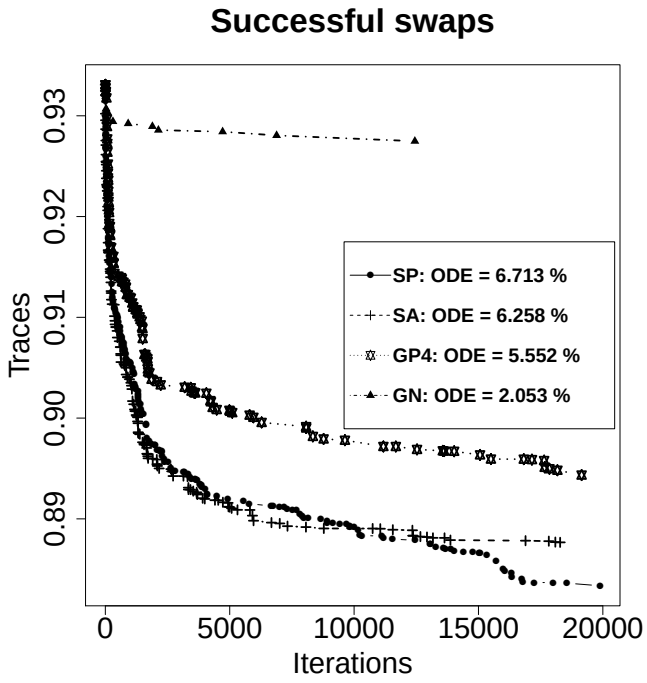
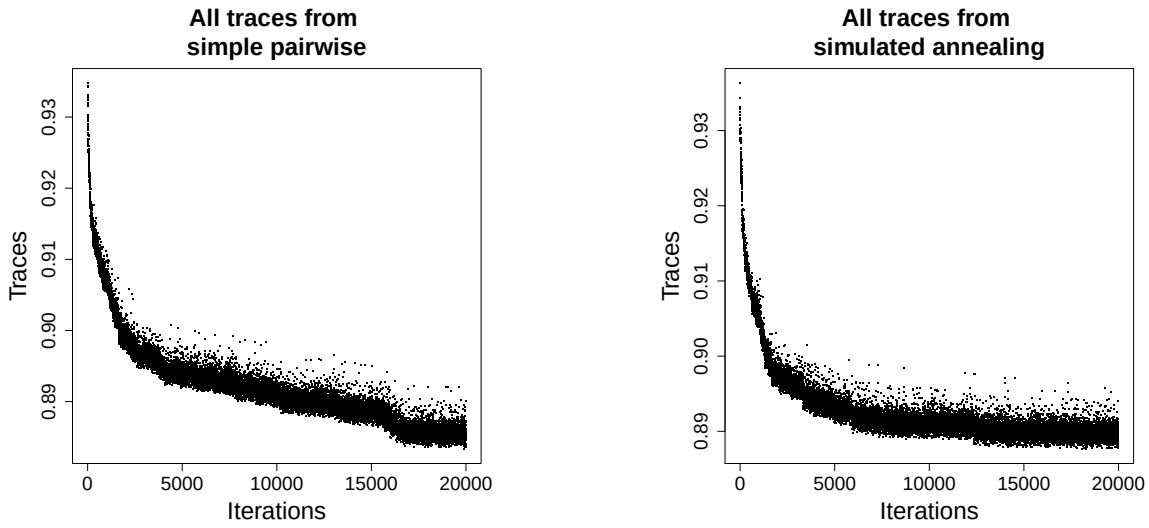


Figure 3-1. A motivating $\Omega_A^{(30)}$ example displaying the traces from successful swaps to convey the rate of convergence for simple pairwise (SP), simulated annealing (SA), greedy pairwise (GP4), and genetic neighborhood (GN) algorithms with their overall design efficiencies (ODE) evaluated for half-sib families with $h^2 = 0.1$, $\rho = 0.6$ with a nugget error of 0.1 iterated for 20,000.

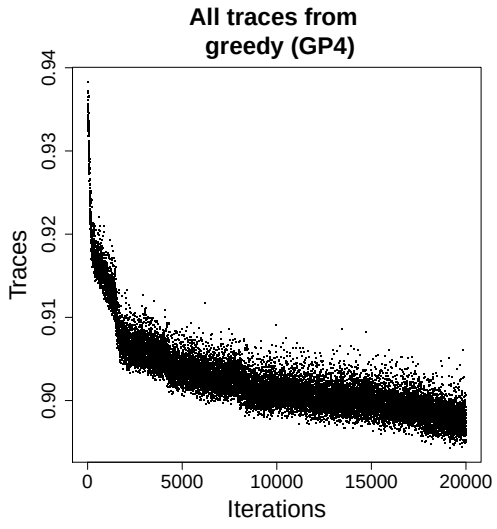
Table 3-1. Average (and standard errors) of algorithms overall design efficiencies (ODE%) for $\Omega_A^{(30)}$ RCB experimental designs at a spatial correlation of 0.6. Average ODEs from 10 replicates per condition are reported together with standard errors (S.E.) for simple pairwise (SP), greedy pairwise (GP4) and GP14, simulated annealing (SA) and genetic neighborhood (GN) procedures.

Condition		SP		GP4		GP14		SA		GN	
pedigree	h2	ODE %	S.E.	ODE %	S.E.	ODE %	S.E.	ODE %	S.E.	ODE %	S.E
	0.1	6.347	0.060	5.501	0.060	3.747	0.093	6.385	0.072	-	-
Indep	0.3	7.398	0.066	6.194	0.080	4.371	0.053	7.403	0.063	-	-
	0.6	5.109	0.044	4.414	0.057	3.110	0.054	5.222	0.064	-	-
	0.1	5.826	0.026	5.082	0.055	3.610	0.065	5.781	0.045	1.853	0.042
Half-sib	0.3	5.375	0.056	4.640	0.082	3.192	0.052	5.428	0.047	1.940	0.088
	0.6	3.066	0.028	2.663	0.023	1.858	0.033	3.131	0.028	1.064	0.033
	0.1	4.109	0.030	3.611	0.026	2.543	0.038	4.045	0.027	1.343	0.034
Full-sib	0.3	2.656	0.029	2.265	0.021	1.601	0.034	2.667	0.032	0.920	0.027
	0.6	1.247	0.006	1.065	0.009	0.755	0.012	1.247	0.013	0.460	0.011

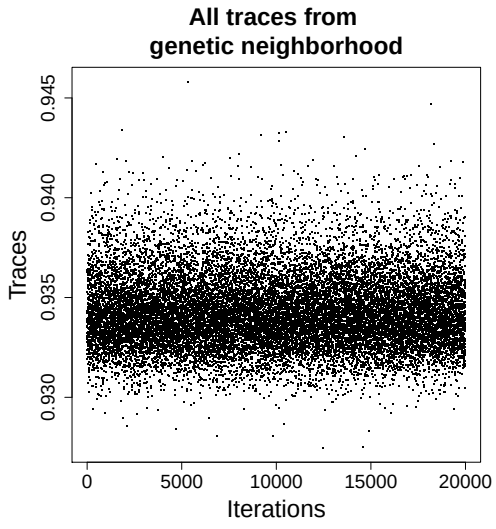


(a)

(b)



(c)



(d)

Figure 3-2. A motivating $\Omega_A^{(30)}$ example illustrating the rates of convergence of algorithms showing all traces obtained from these algorithms evaluated for half-sib families with $h^2 = 0.1$, $\rho = 0.6$ with a nugget error of 0.1 iterated for 20,000.

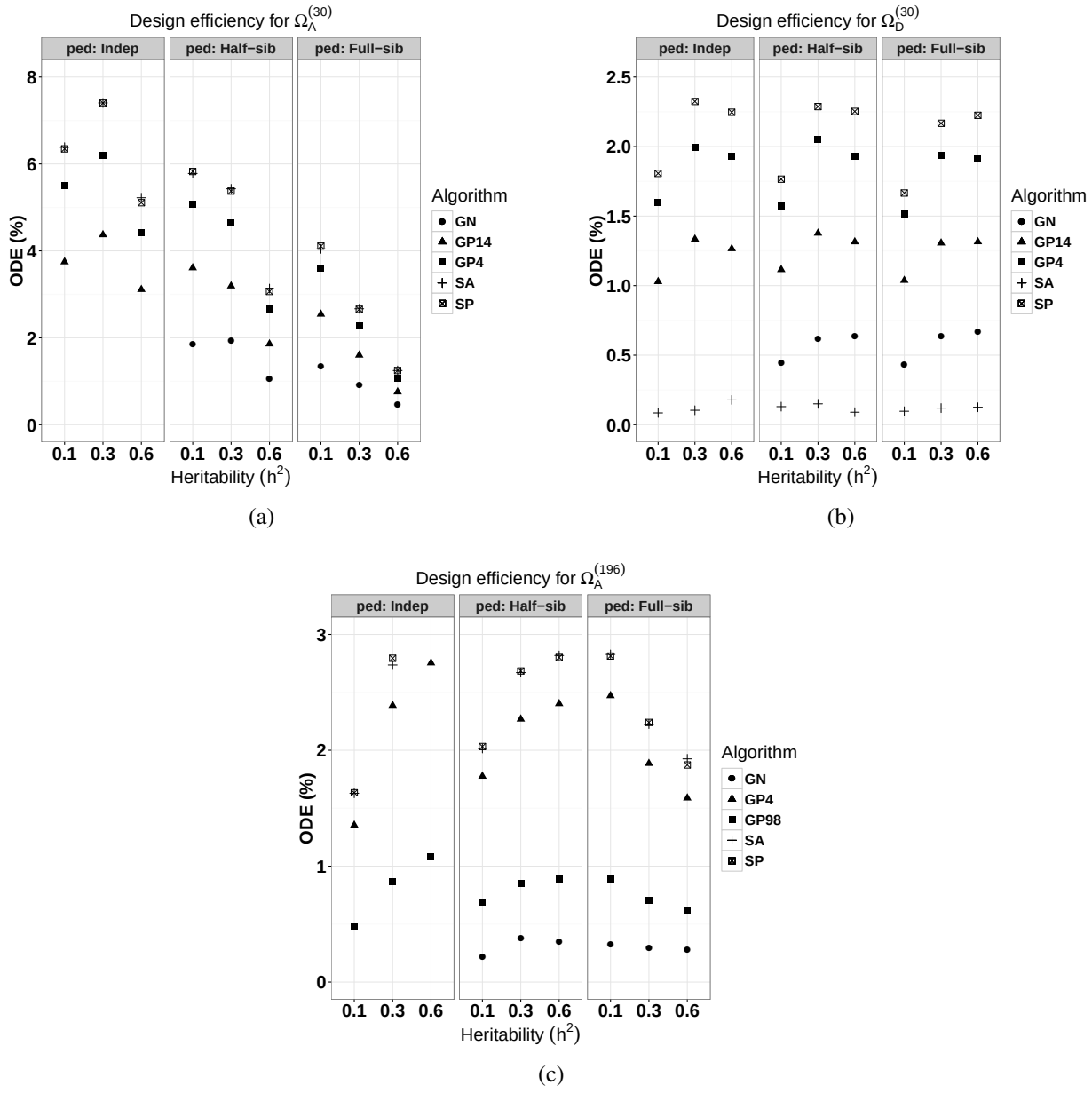


Figure 3-3. Overall design efficiency (ODE %) for (a) $\Omega_A^{(30)}$, (b) $\Omega_D^{(30)}$, and (c) $\Omega_A^{(196)}$ scenarios evaluated for simple pairwise (SP), greedy pairwise: GP4, GP14, and GP98, simulated annealing (SA) and genetic neighborhood (GN) algorithms iterated $p = 5,000$ times, with each condition replicated $\lambda = 10$ times, with $m = 100$ initially unimproved designs and $s = 1$ selected design.

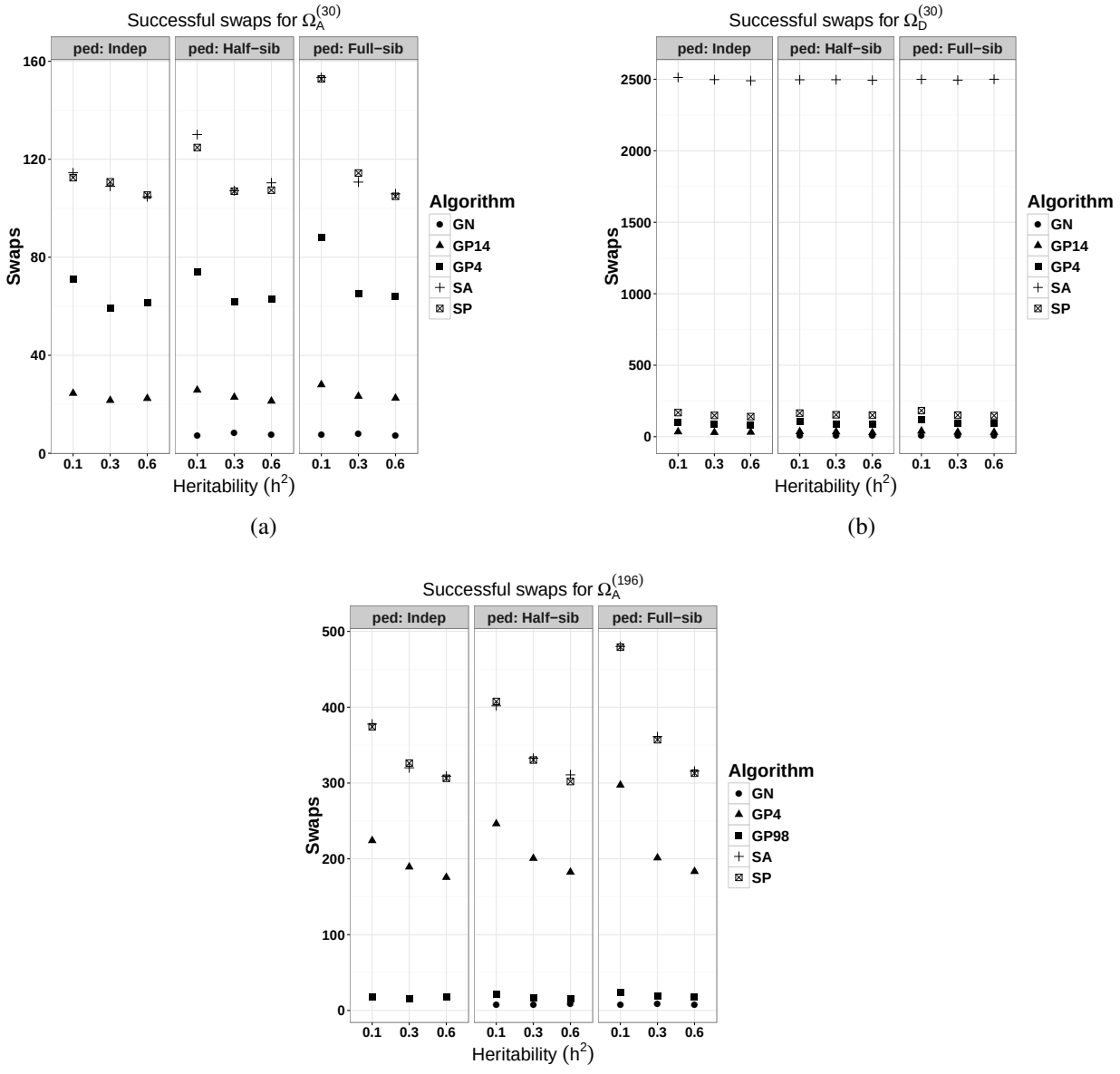


Figure 3-4. Average swaps for scenarios (a) $\Omega_A^{(30)}$, (b) $\Omega_D^{(30)}$, and (c) $\Omega_A^{(196)}$ based on simple pairwise (SP), greedy pairwise: GP4, GP14, and GP98, simulated annealing (SA) and genetic neighborhood (GN) algorithms.

Table 3-2. Average (and standard errors) of algorithms overall design efficiencies (ODE%) for $\Omega_A^{(196)}$ RCB experimental designs at a spatial correlation of 0.6. Average ODEs are reported together with standard errors (S.E.) for simple pairwise (SP), greedy pairwise: GP4 and GP98, simulated annealing (SA) and genetic neighborhood (GN) procedures.

Condition		SP		GP4		GP98		SA		GN	
pedigree	h2	ODE %	S.E.	ODE %	S.E.	ODE %	S.E.	ODE %	S.E.	ODE %	S.E
	0.1	1.633	0.013	1.354	0.018	0.481	0.008	1.629	0.015	-	-
Indep	0.3	2.794	0.020	2.387	0.017	0.864	0.024	2.736	0.034	-	-
	0.6	3.232	0.024	2.754	0.039	1.080	0.028	3.270	0.027	-	-
	0.1	2.032	0.023	1.776	0.019	0.690	0.018	2.016	0.019	0.216	0.014
Half-sib	0.3	2.684	0.018	2.269	0.009	0.851	0.024	2.670	0.019	0.381	0.013
	0.6	2.801	0.027	2.402	0.029	0.890	0.025	2.818	0.009	0.351	0.020
	0.1	2.813	0.018	2.471	0.022	0.888	0.025	2.827	0.014	0.324	0.015
Full-sib	0.3	2.240	0.016	1.886	0.023	0.702	0.020	2.226	0.021	0.297	0.011
	0.6	1.873	0.011	1.588	0.013	0.623	0.016	1.926	0.013	0.280	0.013

Table 3-3. Average (and standard errors) of algorithms overall design efficiencies (ODE%) for $\Omega_D^{(30)}$ RCB experimental designs at a spatial correlation of 0.6. Each condition was replicated 10 times with their average ODEs reported together with standard errors (S.E.) for simple pairwise (SP), greedy pairwise: GP4 and GP14, simulated annealing (SA) and genetic neighborhood (GN) procedures.

Condition		SP		GP4		GP14		SA		GN	
pedigree	h2	ODE %	S.E.	ODE %	S.E.	ODE %	S.E.	ODE %	S.E.	ODE %	S.E
	0.1	1.807	0.014	1.600	0.014	1.029	0.014	0.085	0.019	-	-
Indep	0.3	2.324	0.017	1.993	0.013	1.335	0.030	0.104	0.025	-	-
	0.6	2.247	0.021	1.930	0.021	1.265	0.032	0.178	0.027	-	-
	0.1	1.766	0.012	1.576	0.015	1.115	0.015	0.130	0.034	0.446	0.015
Half-sib	0.3	2.287	0.023	2.054	0.024	1.377	0.022	0.150	0.041	0.614	0.013
	0.6	2.253	0.024	1.933	0.020	1.315	0.019	0.090	0.023	0.637	0.023
	0.1	1.666	0.011	1.514	0.013	1.037	0.009	0.097	0.025	0.431	0.010
Full-sib	0.3	2.168	0.013	1.935	0.025	1.307	0.026	0.119	0.025	0.634	0.017
	0.6	2.225	0.027	1.913	0.023	1.316	0.025	0.125	0.019	0.669	0.022

paths for a given forest (Seo *et al.*, 2005) and Liu *et al.* (2006) to optimize spatially constrained harvest scheduling problems in forest planning and management. It has been applied in the current study to assess how well it can be used to improve the efficiency of experimental designs.

In this study, evaluation of algorithm efficiency to improve experimental designs has focused on the use of RCB designs in field trials with application in plant breeding. Presence of half-sib or full-sib families in experiments require appropriate modeling of their genetic correlations

and similarly, proximity of genotypes within rows and columns needs to be accounted for as those genotypes in close range may share microsite and thus, accounting for spatial correlations within rows and columns is necessary to minimize experimental error bias. Incorporating spatial correlations in a RCB design has been shown by Gezan *et al.* (2010) to produce designs that are nearly as efficient as those generated using a row-column designs with uncorrelated residual errors. The current study examines potential design efficiency levels that can be achieved when simple pairwise (SP), greedy pairwise algorithms denoted as GP4, GP14, and GP98, simulated annealing (SA) and genetic neighborhood (GN) algorithms in plant breeding programs.

From the motivating example that examined a specific condition where an RCB experiment ($\Omega_A^{(30)}$) with half-sib that had $h^2 = 0.1$ at $\rho = 0.6$ with a nugget error of 0.1 and iterated for 20,000 to improve a design, results indicate that the simple pairwise (SP) algorithm is the best as it managed to improve the initial experiment by reducing the average variance of treatment effects by 6.713 %. SA followed closed with an ODE of 6.258 %, not much different from that of SP. The more aggressive algorithms such as GP4, are unlikely to perform better than SP under the evaluated experimental conditions presented in this study with GN achieving the lowest design improvement levels. Also, GN had only 12 successful swaps which is a much smaller number than SP and SA who recorded 192 and 139 respectively. The rates of convergence as shown in Figure 3-2 is better for SP, SA and GP4 than for GN as its traces are randomly scattered.

Results from Tables 3-1, and 3-2, have shown that SP and SA algorithms achieves the highest relative design efficiencies under all experimental conditions for $\Omega_A^{(30)}$ and $\Omega_A^{(196)}$ scenarios with second best algorithm appearing to be GP4 followed by GP14 for $\Omega_A^{(30)}$ scenario or GP98 for $\Omega_A^{(196)}$ scenario and last by GN. These results could be attributable to the fact that SP procedure swaps a single pair of treatments per iteration, thus takes small steps in the search for optimality which makes it more likely to find an optimal condition than greedy algorithms that take large steps. Simulated annealing performed well under A-optimality criterion since it has the ability not to be trapped in a local minima by accepting a proportion of bad solutions using an exponential distribution and a cooling schedule. However, SA algorithm achieved lowest relative

design efficiencies for the same number of iterations of 5,000 under $\Omega_D^{(30)}$ scenario as shown in Table 3-3. It is not very clear why this is so, but it was observed that it accepted too many bad solutions as it tried not to get trapped in a local minima, that did not subsequently maximize the objective function.

Large design improvements have been observed among genetically unrelated individuals, which agrees with findings from Filho and Gilmour (2003) although they did not analyse varied levels of spatial correlations. Optimization based on *A*-criterion has revealed from this current study that a substantial decrease in average variance of treatment effects among full-sib families can be realized for treatments with very small narrow-sense heritabiliies ($h^2 = 0.1$). When full-sibs have strong narrow-sense heritabilities such as 0.6 at a spatial correlation of 0.6, little improvements on the design efficiencies can be achieved. For experimental designs that were evaluated under $\Omega_A^{(30)}$ scenario, the amount of design improvement was, for some conditions, about four times larger than that realized under $\Omega_A^{(196)}$ scenario. This is is because more iterations ($> 50,000$) are required for larger experiments to converge to an optimal solution than it would take a smaller experiment. The number of successful swaps displayed in Figure 3-4 indicate that they decrease with increasing heritability for all families for experiments evaluated under $\Omega_A^{(196)}$ and $\Omega_A^{(30)}$ scenarios for almost all algorithms.

The choice of *A*- or *D*-optimality criteria depends on the objective function to be maximized or minimized. Both criteria are a convex function of the eigenvalues of an information matrix (Das, 2002; Kuhfeld, 2010) since *A*-optimality is a function of the arithmetic mean of the eigenvalues whereas *D*-optimality is a function of the geometric mean of eigenvalues (Kuhfeld, 2010). Thus, an increase in overall design efficiency implies a decrease in the average variances of the random treatment effects. When matrices are sparse, it is not efficient to use the *D*-optimality criterion since the determinants are likely to be zeros and this may cause computational problems. The approach in this study used the natural logarithms of the determinants, although this did not solve the problem for large experiments such as the $\Omega_A^{(196)}$ scenarios. For these reasons, the authors would recommend *A*-optimality procedure. If

approximations to the procedure are required, then, a similar approach to that described by Butler *et al.* (2008) can be used.

The procedure presented in this study can be easily extended to other complex experimental designs such as non-orthogonal experiments that can be implemented with appropriate statistical models and an optimality criterion of choice. Other variants of the search algorithms can also be implemented. For instance, for genetic neighborhood procedure, a value different from 0.25 could be chosen to indicate which treatments to be swapped. In the implemented GN procedure, any two neighboring treatments that had a genetic relationship coefficient of 0.25 (for half-sib) or more were swapped. It is not known whether changing this value to a higher coefficient would increase the efficiency of GN algorithm.

3.5 Conclusion

The potential to improve experimental designs, particularly randomized complete block designs, has been shown, in this study, to be highest when simulated annealing and simple pairwise algorithms are used under *A*-optimality criteria, in which case, they also achieve the highest numbers of successful swaps. Similarly, under *D*-optimality criterion, simple pairwise records highest overall design efficiencies whereas simulated annealing performs poorest with the largest number of accepted swaps. In conclusion, the use of a simple pairwise algorithm based on *A*-optimality criterion under a linear mixed model framework to improve RCB experimental designs is desirable.

CHAPTER 4
IMPROVING NON-ORTHOGONAL EXPERIMENTAL DESIGNS WITH SPATIALLY AND GENETICALLY CORRELATED DATA

4.1 Introduction

In agricultural and forest field trials, experimental units may not necessarily occur with equal replications and may not be equally represented in each block as some treatments are more likely to be available in large numbers than others. Thus, the scarce treatments will be missing in some blocks while others will be represented in most or all blocks. Such is the case where treatments are available in different numbers due to differential rates in fecundity, greenhouse survival, loss of experimental units or insufficient subjects available. These experiments are said to be non-orthogonal since they are unbalanced and incomplete. Unequal replication of treatments in research studies is thus inevitable. Since such experimental designs are no longer balanced, the standard statistical methods for orthogonal designs cannot be used in these experiments and thus the need to incorporate appropriate statistical procedures such as the use of linear mixed models in the generation and analysis of non-orthogonal designs. Estimates of treatment means and variances are more variable and orthogonality is lost in unbalanced designs, implying that treatment effects might be inter-correlated. In addition, the sum of squares partitions for analysis of variance for a single effect will also convey some information about other effects (that is, non-orthogonal relationships). This might lead to contradictory results under the cell and marginal means models if the statistics used to test the hypothesis does not take care of non-orthogonality (Kuehl, 2000).

This study has considered evaluating a set of unequally replicated designs, incomplete block and augmented designs to lay down a procedure that could be extended later on to other designs of interest. In incomplete block (IB) designs, not all treatments are represented in every block. Blocking increases precision of estimates of interest as it enables comparison to be made under more homogeneous conditions. As blocks get larger, treatments within blocks become more heterogeneous and this reduces precision of the estimated parameters of interest. Thus, often smaller units (*i.e* incomplete blocks) are defined that contain only a portion of the treatments. In

addition, block size may be defined by some natural grouping of experimental units that could result into allocation of fewer units per block. IB designs can be classified into resolvable and nonresolvable designs, or balanced incomplete block (BIB) and partially balanced incomplete block (PBIB) designs. Resolvable IB designs have blocks that can be grouped together in a way to include all treatments replicated once in each of the groups. For BIB designs, treatment pairs occur in the same block an equal number of times whereas for PBIB designs, different treatment pairs occur in a block an unequal number of times implying that mean comparisons will have different levels of precision (John and Williams, 1995).

Augmented block designs are used most often in early stages of breeding programs for evaluation of a large number of new test treatments that are replicated once, planted with known controls that are replicated several times. (Federer, 1956; Federer and Raghavarao, 1975; Federer, 1998; Cullis *et al.*, 2006; William *et al.*, 2011). Replicated controls enable estimation of block effects, error variances and a connection of trials if conducted in multi-environments (Moehring *et al.*, 2014). These designs are useful when it is not possible to replicate the new test treatments and so, the available controls are replicated in large numbers within blocks to monitor experimental conditions, thus, acting as baseline. In the analysis, observations on test treatments are adjusted for field heterogeneity and scarce resources are utilized efficiently. However, augmented designs may have relatively few degrees of freedom for experimental error, which often results in reduced power to detect differences among treatments and they are inherently imprecise since treatments are unreplicated. Partially replicated augmented block designs (p-rep) replace controls with additional plots of replicated test treatments such that a proportion p of them is replicated, a procedure that avoids the need to have controls (Cullis *et al.*, 2006; William *et al.*, 2011). These studies use test treatments rather than controls to estimate experimental errors and make adjustments for field heterogeneity effects.

The search for a method to improve non-orthogonal designs becomes challenging when both treatments and blocks are random effects and genetic relationships as well as spatial correlations exist among observations. Blocks are considered to be random effects since they

are incomplete as not all treatments are equally represented in every block. Also, treatments are considered to be random effects in order to incorporate genetic relationships as genotypes may share one parent (half-sib) or both parents (full-sib). A practical experimental layout in plant breeding involves physically planting treatments in rows and columns of either regular or irregular-grid rectangular layouts. Experimental units that are physically close together are likely to be more spatially correlated than units farther apart as they share a common microsite environment. Spatial correlations and genetic relationships that exist in experiments have to be modeled appropriately using a linear mixed model which enables proper specification of genetic and spatial variance-covariance structures. In field trials, spatial and genetic correlations can be confounded if not properly modeled, which can mask differences in genotypic values of treatments, consequently reducing the precision of their estimates (See Chapter 2). Gonçalves *et al.* (2007) also reported that using spatial mixed models significantly resulted in a positive impact on selection decisions and increased the accuracy of genetic value prediction.

Generation of improved experimental designs requires the use of an optimality criterion. Many choices exist such as A, D, E, G and a suitable combination of these (John and Williams, 1995; Das, 2002). The method most commonly used is A-optimality criterion that seeks to minimize average variance of treatment effects and has been shown to be an effective method in plant breeding trials used to choose the best experiment designs (Chernoff, 1953; Cullis *et al.*, 2006, 1989). It can be expressed as $A_{optimality} = \text{argmin}\{trace[\mathbf{M}(\Omega)]\}$ where $\mathbf{M}(\Omega)$ is the inverse of an information matrix (*i.e* variance-covariance matrix) of the treatment effects from a given design Ω.

The present study aims to develop and evaluate statistical procedures to generate improved designs for unequal replications, incomplete block and augmented designs for field experimental trials based on A-optimality criterion. Spatial correlations are modeled using a 2-dimensional separable autoregressive 1st order variance structure (AR1) whereas genetic relatedness are modeled using a numerator relationship matrix. Relative design efficiencies between unimproved and improved experimental layouts are evaluated for an array of experimental conditions.

4.2 Materials and Methods

4.2.1 Statistical Models

A linear mixed effects model (LMM) was used with both blocks and treatments considered to be random effects and an overall mean as a fixed effect, which can be expressed as

$$\mathbf{y} = \mathbf{X}\mu + \mathbf{Z}_b\mathbf{b} + \mathbf{Z}_g\mathbf{g} + \mathbf{e} \quad \text{or equivalently as ,} \quad \mathbf{y} = \mathbf{X}\mu + \begin{bmatrix}\mathbf{Z}_b & \mathbf{Z}_g\end{bmatrix}\begin{bmatrix}\mathbf{b}\\ \mathbf{g}\end{bmatrix} + \mathbf{e} \tag{4–1}$$

$$= \mathbf{X}\mu + \mathbf{Z}\gamma + \mathbf{e}, \quad \text{where} \quad \mathbf{Z} = \begin{bmatrix}\mathbf{Z}_b & \mathbf{Z}_g\end{bmatrix} \quad \text{and} \quad \gamma = \begin{bmatrix}\mathbf{b}\\ \mathbf{g}\end{bmatrix} \quad \text{and} \quad \mathbf{G} = \begin{bmatrix}\mathbf{D}_b & \mathbf{0}\\ \mathbf{0} & \mathbf{G}_g\end{bmatrix}$$

where $\mathbf{X}$ is a design matrix with a column vector of n ones and μ is overall expected mean, $\mathbf{y}$: vector of response observations, $\mathbf{Z}_b$: is an incidence matrix of blocks effects and $\mathbf{Z}_g$ is an incidence matrix of treatment effects, $\mathbf{b}$ is a vector of random block effects such that $\mathbf{b} \sim MVN(\mathbf{0}, \mathbf{D}_b)$, $\mathbf{g}$ is a vector of random treatments effects such that $\mathbf{g} \sim MVN(\mathbf{0}, \mathbf{G}_g)$, $\mathbf{e}$ is a vector of random errors (residuals) such that $\mathbf{e} \sim MVN(\mathbf{0}, \mathbf{R})$ where $\mathbf{D}_b$, $\mathbf{G}_g$ and $\mathbf{R}$ are variance-covariance matrices for the blocks, treatments and residual errors, respectively.

Estimation of random effects is done by solving a set of linear mixed models equations (Henderson, 1975) yielding (Hooks *et al.*, 2009)

$$\mathbf{M}(\Omega) = Var(\hat{\mathbf{g}} - \mathbf{g}) = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z})^{-1} \tag{4–2}$$

$$= (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} - \mathbf{Z}'\mathbf{K}_x\mathbf{Z})^{-1}$$

where $\mathbf{K}_x = \mathbf{R}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}$. By expanding $\mathbf{Z}$ and $\mathbf{G}$, Equation 5–5 becomes

$$\mathbf{M}(\Omega) = \left\{ \begin{bmatrix}\mathbf{Z}_b & \mathbf{Z}_g\end{bmatrix}'\mathbf{R}^{-1}\begin{bmatrix}\mathbf{Z}_b & \mathbf{Z}_g\end{bmatrix} + \begin{bmatrix}\mathbf{D}_b^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}_g^{-1}\end{bmatrix} - \begin{bmatrix}\mathbf{Z}_b & \mathbf{Z}_g\end{bmatrix}'\mathbf{K}_x\begin{bmatrix}\mathbf{Z}_b & \mathbf{Z}_g\end{bmatrix} \right\}^{-1} \tag{4–3}$$

where $\mathbf{D}_b = \sigma_b^2\mathbf{I}_b$ and $\mathbf{G}_g = \sigma_g^2\mathbf{A}_g$, where $\mathbf{A}$ is a numerator relationship matrix calculated from a pedigree of genetic relationships, or $\mathbf{G}_g = \sigma_g^2\mathbf{I}_g$ for genetically unrelated individuals. When the residual errors are assumed to be independent and identically distributed, $\mathbf{R} = \sigma_e^2\mathbf{I}_n$. However, if

the residual errors are correlated, $\mathbf{R}$ could be modeled, for instance, using a spatial autoregressive covariance structure of order 1 (AR1) as $\mathbf{R} = \sigma_e^2 \sum_r(\rho_r) \otimes \sum_c(\rho_c)$ (Gezan *et al.*, 2010; Stroup, 2013), where $\sum_r(\rho_r)$ and $\sum_c(\rho_c)$ are autocorrelation matrices with spatial correlation parameters along the rows and columns of the experimental field design (Gilmour *et al.*, 2009), respectively. The matrix $\mathbf{M}(\Omega)$ can be expressed as:

$$\mathbf{M}(\Omega) = \begin{bmatrix} \sum_b(\Omega) & \sum_{bg}(\Omega) \\ \sum_{bg}(\Omega) & \sum_g(\Omega) \end{bmatrix} \tag{4–4}$$

where $\sum_g(\Omega)$ is the portion of the matrix that contains the variance-covariance of treatment effects, from which a trace is calculated. Thus,

$$A_{optimality} = \text{argmin} \left\{ trace \left[\sum_g(\Omega) \right] \right\}$$

4.2.2 Optimization Procedure

A simple pairwise (SP) algorithm was used in the optimization procedure as described in Chapter 2. The following steps are undertaken in order to improve a given experimental design:

1. generate several initial designs randomly and calculate for each one of them, a trace value, $\tau_i = trace[\sum_g(\Omega_i)]$ for $i = 1, 2, 3, \cdots, m$, where m is the number of initial designs,
2. select a design Ω_i with the smallest trace value τ_0 which minimizes the average variances of the treatment effects. Also, calculate an average trace value from the m traces,
3. obtain matrices $\mathbf{Z}$, $\mathbf{G}$, $\mathbf{R}$ and $\mathbf{K}$ as given in Equations 5–9, 5–5 and 5–7.
4. randomly select a pair of treatments within a randomly selected block and swap them to generate a new experimental layout Ω_j and use the LMM to evaluate its trace value $\tau_j = trace[\sum_g(\Omega_j)]$.
5. if $\tau_0 > \tau_j$, accept the new experimental design Ω_j and replace it with the previous Ω_i and denote its trace as τ_i to replace the previous trace value, otherwise, reject Ω_j, and
6. repeat steps 4 and 5 for p iterations and display the improved experimental design.

Generation of initial designs with unequal replications require a list of constraints provided by the breeder, indicating the smallest and largest number of replicates available for each treatment. On the other hand, incomplete block designs are generated with block sizes being smaller than the number of treatments whereas augmented designs are generated with un-replicated test treatments and replicated controls. Unequally replicated experiments were improved by randomly replacing treatments and swapping pairs of randomly selected treatments either within or across blocks. For IB designs, treatments were randomly swapped across and within blocks, without replacement. To improve augmented designs, treatments were swapped only within blocks.

4.2.3 Evaluation of Experimental Conditions

The following scenarios of experimental designs were evaluated. Unequally replicated experiments with 30 treatments were generated with 6 blocks of size $k = 30$ with dimensions 5 rows by 6 columns on a regular-grid and on an irregular-grid as shown in Figures 4-1a and 4-1c, respectively. Since treatments were not equally replicated, the number of replications r for treatment i ranged between 4 and 8 (that is, $4 \leq r_i \leq 8$). See Figure 4-1 for an example of the physical layout.

Incomplete block (IB) experiments were generated with a total of 30 treatments and block sizes, $k = 20$, with 6 blocks of dimensions 5 rows by 4 columns. Each treatment was replicated 4 times (that is, $r_i = r = 4$). The physical layout is displayed in Figure 4-1b. For augmented designs, the experiments were generated with a total of 492 un-replicated test treatments split into 3 incomplete blocks with each block having 164 un-replicated test treatments. Also, there were 3 controls that were replicated 12 times in each of these blocks. Thus, blocks were of size $k = 200$ of dimensions 10 rows by 20 columns.

Since the phenotype of a trait is determined by its genetic composition and environmental factors, evaluation of the algorithm was done for varying levels of genetic and environmental conditions. The degree of resemblance among relatives is determined by a narrow-sense

heritability which was calculated as

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_b^2 + \sigma_e^2}$$

where σ_g^2, σ_b^2, and σ_e^2 are the variance components for genotypes, blocks and residual errors, respectively, with narrow-sense heritabilities h^2 = 0.1, 0.3, and 0.6. The variance of the blocks (σ_b^2) was calculated as $0.2(1-h^2)$ and $\sigma_e^2 = (1-\sigma_{ms}^2)(1-h^2-\sigma_b^2)$, where σ_{ms}^2 is a nugget effect and, spatial correlations were set to 0, 0.3, 0.6 and 0.9.

Experimental conditions were evaluated among genetically unrelated individuals (Indep), half-sib and full-sib families. For the unequally replicated and IB designs, pedigrees for half-sib comprised of 5 male (or female) parents each with 6 individuals and full-siblings consisted in a half-diallel with 5 parents for a total of 10 families each with 3 individuals. Among the augmented designs, a pedigree for half-sib families comprised of 41 parents each with 12 individuals and full-sib consisted in a half-diallel with 12 parents for a total of 35 families each with about 14 individuals. The control treatments were genetically unrelated to the 492 test treatments.

Each evaluated condition was repeated $\lambda = 10$ times and each λ had $m = 100$ initial designs randomly generated and the best initial design chosen for optimization with $p = 5,000$ iterations as described in Section 4.2.2. As indicated earlier, Figure 4-1 shows examples of experimental layouts to be improved using the proposed algorithm. All computations were implemented in *R* (R Core Team, 2016) using high performance computers available at the University of Florida.

4.2.4 Relative Design Efficiency

A measure of relative overall design efficiency (ODE) was defined as a proportion of the differences between the average trace value from m initial designs and the trace value obtained from the improved design, expressed as

$$ODE = \frac{\bar{A}_{ij} - A_{(opt)ij}}{\bar{A}_{ij}} \tag{4–5}$$

19	2	21	16	12	24	27	23	15	2	5	21
2	1	16	11	21	12	9	1	6	9	13	30
7	11	9	17	28	4	26	6	17	20	19	18
7	3	30	2	4	11	25	23	8	30	12	3
3	14	30	17	16	23	13	20	27	1	29	5
20	7	14	23	23	13	17	14	13	29	15	29
6	24	29	4	23	16	19	6	19	27	21	26
7	21	18	17	21	10	16	6	25	12	18	2
22	4	10	30	22	26	1	7	23	23	4	27
16	22	18	27	15	10	22	19	8	26	20	22
24	25	20	5	27	11	2	1	20	8	29	8
9	22	5	26	3	24	2	25	28	7	7	5
25	29	30	3	26	18	6	10	1	13	13	3
8	11	28	3	15	14	8	28	15	5	12	15
7	14	14	19	24	9	8	10	17	24	28	9

(a)

24	30	11	2	7	4	29	28
21	3	12	27	19	23	30	1
6	20	21	23	2	19	5	8
15	17	20	14	25	16	8	11
17	25	12	7	14	13	16	16
26	22	2	4	1	25	21	2
29	23	18	12	4	4	8	20
27	17	24	28	6	13	9	13
27	24	3	11	29	22	14	10
10	23	5	25	11	9	26	30
18	6	9	5	7	29	1	13
6	3	19	28	9	17	12	8
5	15	19	26	28	16	15	18
27	7	30	18	10	26	21	22
14	24	4	22	3	10	1	15

(b)

28	22	1	17	29	14	8	17	27	20	13	1												
11	26	22	27	14	6	30	3	1	18	20	8												
2	19	26	5	17	7	14	2	18	22	19	3												
19	21	25	15	26	15	7	30	28	2	6	27												
15	4	29	21	13	21	15	10	6	4	20	30												
14	9	21	28	23	2	24	20	29	10	4	29	28	25	1	16	22	12	19	30	28	10	11	8
24	29	27	13	12	15	30	7	11	23	19	28	17	22	23	18	6	25	8	26	16	3	20	11
4	30	12	9	16	23	21	20	22	13	17	12	18	5	3	5	6	13	27	16	27	9	25	9
8	16	24	7	4	6	23	1	25	7	11	5	10	8	12	9	17	2	1	14	10	24	25	18
13	5	21	26	21	24	7	5	25	2	29	10	25	17	11	16	6	1	26	4	9	3	25	19

(c)

Figure 4-1. Motivating examples of non-orthogonal experimental designs with regular and irregular-grid layouts, where, (a) represents an unequally replicated experiment with regular-grid layout with 30 treatments and 6 blocks of sizes 5 rows by 6 columns with treatments replicated r_i times such that $4 \leq r_i \leq 8$, (b) is an incomplete block experimental layout with 30 treatments and 6 blocks of sizes $k = 20$, and (c) is an unequally replicated experiment with irregular-grid layout with 30 treatments and 6 blocks of sizes 5 rows by 6 columns with treatments replicated r_i times such that $4 \leq r_i \leq 8$.

where $i = 1,2,\cdots,\xi$ condition and $j = 1,2,\cdots,\lambda$ replicates per condition, where ξ is the number of conditions evaluated for that design, $\bar{A}_{ij}$ is the average trace value from $m = 100$ initially unimproved designs for condition i and replicate j and $A_{(opt)ij}$ is the trace value from the improved design of the ith condition and replicate j obtained after $p = 5,000$ iterations. Note that a single initial design ($m = 1$) may be generated and optimized using the described procedure. For that case, $ODE = \frac{\tau_{ij} - A_{(opt)ij}}{\tau_{ij}}$, where τ_{ij} is the trace of the initial design Ω for replicate j of condition i.

Among the unequally replicated designs, an effective ODE was obtained after the initial best experimental design was subjected to a process of replacing genotypes using a list of constraints to guide on the minimum and maximum number available for each genotype and then swapping pairs of genotypes within blocks. The effective ODE from incomplete block designs was obtained when the initial best experimental design was improved by swapping treatments across blocks and then swapping treatments either within or across blocks whereas ODE from augmented designs were obtained after swapping treatments within blocks.

4.3 Results

A summary of average percentage of overall design efficiencies (ODE %) with standard errors (S.E.) from each of the evaluated experimental conditions are given in Tables 4-1, 4-2 and 4-3, for unequally replicated designs, incomplete block and augmented designs, respectively, and details of their respective individual ODEs are presented in Figures 4-2, 4-3 and 4-4.

Unequally replicated experimental designs yielded, on average, a highest improvement level of 9.348 % (S.E. = 0.190) achieved when $h^2 = 0.3$ and $\rho = 0.6$ observed among genetically unrelated individuals (Indep). Also, at $h^2 = 0.1$, a mean highest level of design improvement of 9.283 % (S.E. = 0.184) was observed at a spatial correlation of 0.6. In addition, when $h^2 = 0.6$, a highest mean ODE of 6.207 % (S.E. = 0.110) was obtained at $\rho = 0.6$ among the genetically unrelated individuals. Results from the unequally replicated experiments also indicate that for a given heritability, average ODEs increase with increasing spatial correlations up to $\rho = 0.6$ and drop as spatial correlation increases to 0.9 among the Indep, half-sib and full-sib families.

Among half-sib families, a mean highest ODE of 7.650 % (S.E. = 0.085) was obtained when $h^2 = 0.1$ and $\rho = 0.6$, followed by an ODE of 6.607 % (S.E. = 0.085) obtained when $h^2 = 0.3$ and $\rho = 0.6$. When $h^2 = 0.6$, an ODE of 3.417 % (S.E. = 0.0.070) was obtained at spatial correlation of $\rho = 0.6$. Figure 4-2 shows all individual ODEs obtained for the evaluated conditions based on unequally replicated experiments. The same pattern, as that of half-sib and independent families, was observed among full-sib families, achieving a mean highest reduction in average variance of treatment effects with an ODE of 4.914 % (S.E. = 0.075) obtained when $h^2 = 0.1$ and $\rho = 0.6$. Still, among full-sib families, when $h^2 = 0.3$, a mean highest ODE of 3.246 % (S.E. = 0.063) was obtained at $\rho = 0.6$ and a mean highest ODE of 1.389 % (S.E. = 0.022) obtained when $h^2 = 0.6$ at $\rho = 0.6$. In general, among the unequally replicated experiments, for a specified heritability and spatial correlation level, individual ODEs (See Table 4-1) appear to decrease as genetic relationships increases. Also, for any given heritability, lowest ODEs were observed when spatial correlations were null ($\rho = 0.0$) and in some conditions, when $\rho = 0.9$ with $h^2 = 0.6$.

Incomplete block designs achieved a highest average ODE of 10.250 % (S.E. = 0.274) when $h^2 = 0.1$ and $\rho = 0.9$. For treatments with $h^2 = 0.3$ and $h^2 = 0.6$, mean highest reduction in average variance of treatment effects, with ODEs of 8.543% (S.E. = 0.139) and 6.854 % (S.E. = 0.122), respectively, were obtained at $\rho = 0.6$. For a fixed spatial correlation level, individual ODEs among full-sib families appear to decrease with increasing heritability as shown in Figure 4-3 with average highest ODEs obtained when the spatial correlation was 0.6 (Table 4-2). Among half-sib families, a highest ODE of 6.824 % (S.E. = 0.163) was obtained when $h^2 = 0.1$ at $\rho = 0.6$. For heritabilities of 0.3 and 0.6, mean highest ODEs of 6.413 % (S.E. = 0.180) and 4.337 % (S.E. = 0.087) were obtained when $\rho = 0.6$. Design improvements for incomplete block experiments with full-sib families were ODEs of 5.082 (S.E. = 0.089), 3.511 %(S.E. = 0.098) and 1.833 % (S.E. = 0.047) for $h^2 = 0.1$, 0.3 and 0.6, respectively, all of them obtained when $\rho = 0.6$.

Results from augmented experimental designs achieved a highest reduction of average variance of treatment effects among full sib families, with an average ODE of 3.752 % (S.E. = 0.093) obtained when $h^2 = 0.1$ and $\rho = 0.6$. Still among the full-sib families, treatments with

Table 4-1. Summary of average overall design efficiencies (ODE %) and standard errors (S.E) for unequally replicated regular-grid designs with varying genetic relatedness, spatial correlations (ρ) and narrow-sense heritability (h^2) based on $m = 100$ initial designs, $\lambda = 10$ replicates per condition and $p = 5,000$ iterations.

Condition		Indep		Half-sib		Full-sib	
h^2	ρ	ODE (%)	S.E	ODE (%)	S.E	ODE (%)	S.E
0.1	0.0	0.889	0.025	0.787	0.028	0.584	0.010
	0.3	2.968	0.057	3.533	0.087	3.846	0.056
	0.6	9.283	0.184	7.650	0.085	4.914	0.075
	0.9	7.362	0.152	3.985	0.053	1.569	0.028
0.3	0.0	1.918	0.039	1.488	0.037	0.777	0.028
	0.3	4.566	0.092	3.870	0.056	2.438	0.056
	0.6	9.348	0.190	6.607	0.085	3.246	0.063
	0.9	2.747	0.043	1.290	0.014	0.472	0.006
0.6	0.0	2.131	0.042	1.474	0.030	0.699	0.016
	0.3	4.562	0.115	2.920	0.070	1.367	0.038
	0.6	6.207	0.110	3.417	0.070	1.389	0.022
	0.9	0.879	0.017	0.390	0.007	0.138	0.001

heritabilities of 0.3 and 0.6 obtained their highest average ODEs of 2.939 % (S.E. = 0.092) and 2.387 % (S.E. = 0.083) at spatial correlation of 0.6. The ODEs increased with increasing spatial correlation, from $\rho = 0.0$ to 0.6 and decreased at a spatial correlation of 0.9 as shown in Table 4-3. For augmented experiments with genetically unrelated individuals, a trend was observed of highest design improvements that were achieved when $\rho = 0.9$ at all levels of heritabilities with the highest being an ODE of 1.881 % (S.E. = 0.120) obtained when $h^2 = 0.6$ and $\rho = 0.9$. Similarly, an ODE of 2.002 % (S.E. = 0.111) was the highest improvement achieved among half-sib families when $h^2 = 0.6$ and $\rho = 0.9$. Figure 4-4 shows individual ODEs with a tendency to increase with heritability and spatial correlation among genetically unrelated individuals, and decrease with increasing heritability among half-sib and full-sib families except when $h^2 = 0.6$ and $\rho = 0.9$ among half-sib families where large ODEs were obtained.

4.4 Discussion

Early generation field trials are used in plant breeding programs and play an important role in selection of promising genotypes for future breeding (Moehring *et al.*, 2014). They

Table 4-2. Summary of average overall design efficiencies (ODE %) and standard errors (S.E) for regular-grid incomplete block designs with varying genetic relatedness, spatial correlations (ρ) and narrow-sense heritability (h^2) based on $m = 100$ initial designs, $\lambda = 10$ replicates per condition and $p = 5,000$ iterations.

Condition		Indep		Half-sib		Full-sib	
h^2	ρ	ODE (%)	S.E	ODE (%)	S.E	ODE (%)	S.E
0.1	0.0	0.317	0.009	0.468	0.011	0.421	0.012
	0.3	1.870	0.052	2.657	0.060	3.481	0.106
	0.6	6.663	0.193	6.824	0.163	5.082	0.089
	0.9	10.250	0.274	5.919	0.138	2.641	0.053
0.3	0.0	0.711	0.038	0.733	0.032	0.476	0.014
	0.3	2.902	0.050	2.950	0.063	2.220	0.041
	0.6	8.543	0.139	6.413	0.180	3.511	0.098
	0.9	4.672	0.090	2.313	0.038	0.858	0.016
0.6	0.0	0.872	0.039	0.709	0.017	0.371	0.020
	0.3	3.332	0.105	2.470	0.084	1.243	0.026
	0.6	6.854	0.122	4.337	0.087	1.833	0.047
	0.9	1.604	0.036	0.743	0.010	0.257	0.005

Table 4-3. Summary of average overall design efficiencies (ODE %) and standard errors (S.E) for regular-grid augmented designs with varying genetic relatedness, spatial correlations (ρ) and heritability (h^2) levels with 492 un-replicated treatments and 3 controls, each replicated 12 times in each of the 3 incomplete blocks. Calculations were based on $m = 1$ initial design, $\lambda = 10$ replicates per condition and $p = 5,000$ iterations.

Condition		Indep		Half-sib		Full-sib	
h^2	ρ	ODE (%)	S.E	ODE (%)	S.E	ODE (%)	S.E
0.1	0.0	0.639	0.000	0.820	0.000	1.589	0.020
	0.3	0.733	0.001	1.235	0.009	3.274	0.049
	0.6	0.922	0.006	1.765	0.018	3.752	0.093
	0.9	1.090	0.028	1.477	0.037	2.002	0.061
0.3	0.0	1.070	0.000	0.922	0.000	1.740	0.101
	0.3	1.074	0.002	1.095	0.009	2.546	0.088
	0.6	1.025	0.005	1.433	0.015	2.939	0.092
	0.9	1.275	0.062	1.446	0.034	1.474	0.056
0.6	0.0	1.113	0.000	0.506	0.002	1.333	0.090
	0.3	1.086	0.002	0.646	0.006	1.718	0.101
	0.6	1.092	0.006	1.191	0.018	2.387	0.083
	0.9	1.881	0.120	2.002	0.111	1.106	0.039

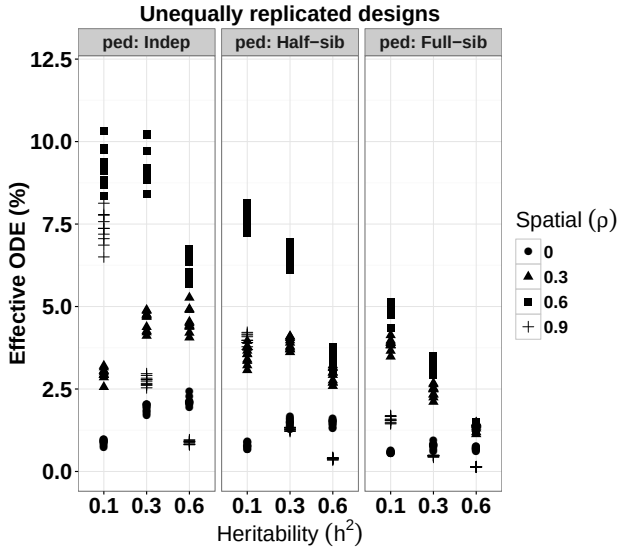


Figure 4-2. Individual effective overall design efficiencies (ODE %) for unequally replicated designs, evaluated with 30 treatments in 6 blocks of dimensions 5 rows by 6 columns based on $m = 100$ initial designs, $\lambda = 10$ replicates per condition and $p = 5,000$ iterations. The effective ODE was obtained after an initial design was subjected to first, random replacement of genotypes using a list of constraints and second, swapping of genotypes within blocks.

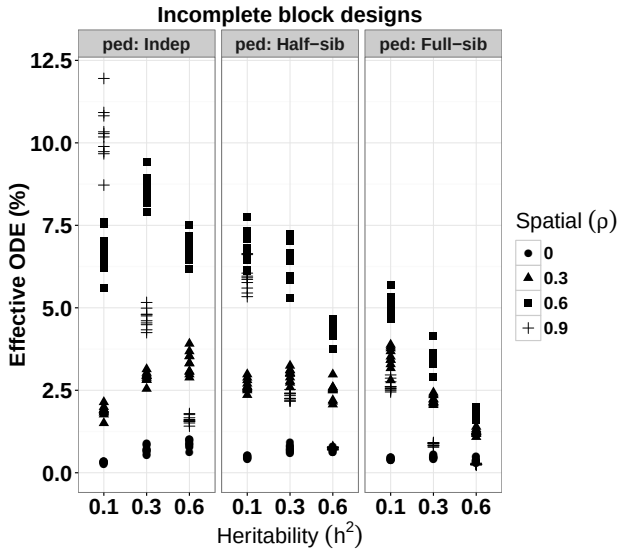


Figure 4-3. Individual effective overall design efficiencies (ODE %) for incomplete block designs, evaluated with 30 treatments in 6 blocks of dimensions 5 rows by 4 columns based on $m = 100$ initial designs, $\lambda = 10$ replicates per condition and $p = 5,000$ iterations. The effective ODE was obtained after an initial design was subjected to random swapping of genotypes, first within blocks, then either across or within blocks.

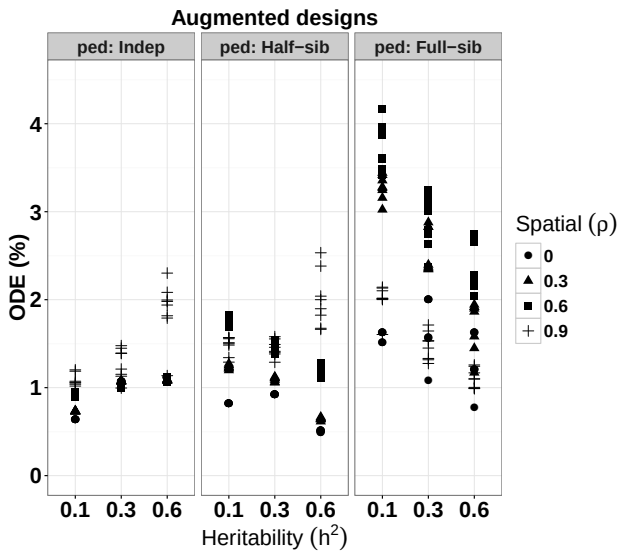


Figure 4-4. Individual overall design efficiencies (ODE %) for augmented designs with each condition replicated $\lambda = 10$ times based on $\Omega_A^{(495)}$ scenario with 3 blocks of sizes 10 rows by 20 columns. These experiments were evaluated with 492 un-replicated test treatments and 3 known controls replicated 12 times each in every block of dimensions 10 rows by 20 columns and iterated for $p = 5,000$ from $m = 1$ initial design.

involve evaluation of many new test treatments, that are often not feasible to be conducted in a randomized complete block (RCB) design, but require non-orthogonal structures. Augmented block designs are particularly important in early generation field trials as they allow a large number of un-replicated treatments to be screened (Federer, 1956; Federer and Raghavarao, 1975; Federer, 1998).

Results from the current study has shown that the highest average variance reduction of treatment effects of about 10 % for incomplete block (IB) designs and about 9 % for unequally replicated experiments can be achieved when conducted with genetically unrelated individuals with heritabilities of 0.1 and 0.3, respectively, at spatial correlations of 0.9 for IB and $\rho = 0.6$ for unequally replicated designs. This result agrees with Filho and Gilmour (2003) who also reported high levels of design improvement among genetically unrelated individuals. However, for augmented designs, the highest average variance reduction in treatment effects was observed among full-sib families at the lowest heritability of 0.1 with spatial correlation of 0.6.

Unlike RCB designs, it is still possible for non-orthogonal designs to improve their efficiencies even when spatial correlation is null (See Chapter 2) due to the lack of balance in the design. The reported results are all based on regular-grid experimental layouts since results from irregular-grid layouts showed a similar pattern of design efficiencies and the level of improvement relied on how far physically two blocks were apart from each other. Higher design efficiencies were obtained for irregular-grid experimental designs when blocks were more isolated than when they were placed side by side.

The proposed algorithm and procedure can be easily extended to include other complex experimental designs, genetic relationships, and spatial variance-covariance structures. Most importantly, statistical computation can be implemented with faster code and software to enable evaluation of larger numbers of genotypes. In addition, instead of using pedigree information to calculate a numerator relationship matrix (Falconer and Mackay, 1996; Patterson and Hunter, 1983), molecular markers can also be used to calculate a genomic relationship matrix (Hill *et al.*, 2008; VanRaden, 2008; Beaulieu *et al.*, 2014; Habier *et al.*, 2007). The process of improving IB and unequally replicated designs took about 5 minutes and about an hour for augmented designs using high performance computers with varied CPU processing speeds, hosted at the University of Florida.

In summary, non-orthogonal experimental designs have varied levels of design efficiencies under different experimental conditions. They can be generated and improved efficiently with the use of a linear mixed model framework to account for genetic relatedness, different levels of heritability and spatial correlations among experimental units by appropriately modeling their respective variance-covariance structures and incorporating an optimization algorithm that maximizes the information extracted from field trials.

CHAPTER 5
OPTIMALDESIGNMM: AN *R* PACKAGE FOR OPTIMIZING EXPERIMENTAL DESIGNS WITH CORRELATED DATA

5.1 Introduction

Experimental designs can have varied levels of environmental heterogeneity such as spatial correlation, which occur due to physical proximity among experimental units. Correlated observations, whether spatially or genetically, require appropriate modeling. These sources of variations affect prediction and estimation of parameters and may lead to imprecise estimates when not accounted for. At the analysis stage, it is a common practice to account for spatially correlated observations (Stringer and Cullis, 2002; Gezan *et al.*, 2010; Cullis *et al.*, 1989), an approach that can be extended into the design stage. Improving the efficiency of experimental designs is beneficial, as it results in reduced background variations. For instance, considering genotypes as random effects allows estimation of predictions (BLUPs) in the mixed models framework. Genetic information can be obtained by reading pedigree data to estimate expected relationships or by processing molecular data to estimate these relationships (Henderson, 1975; Mrode, 2014).

Many experimental designs exist, such as randomized complete block (RCB) designs and incomplete block (IB) designs, among others. The choice of an experimental design largely depends on the main objective of the experiment and availability of materials and technology to conduct the study. Most often, the objective is to compare the effect of treatments on a quantitative response. This is assessed by estimating parameters of interest such as treatment effects and its precision. These are accomplished by generating experimental designs considering replication of experimental units, blocking structure and randomization processes and most importantly, accounting for possible sources of variations in the experiment by using a appropriate statistical models and optimization procedure. Also, non-orthogonal experimental designs are prevalent in many research settings, and procedures to improve their generation is lacking. Non-orthogonal designs have been described by Federer (1956); Federer and Raghavarao (1975); Federer (1998); Cullis *et al.* (2006); William *et al.* (2011) among

others. They include unreplicated designs such as augmented designs, incomplete blocks and unequaly replicated designs. In particular, unreplicated trials allow testing of several hundreds of experimental units with little or no replications. In these settings, both blocks and treatments can be considered to be random effects.

Statistically, a design is optimal in some sense if it maximizes the amount of information available, by optimizing a function of a variance-covariance matrix of treatment effects (Das, 2002). The most common optimality criteria used in choosing study designs are *A*- and *D*- (Butler *et al.*, 2008; Cullis *et al.*, 2006; Hooks *et al.*, 2009; Kuhfeld, 2010; Das, 2002). *A*-optimality (Chernoff, 1953) minimizes the sum of diagonal elements (*i.e*, trace) of the variance-covariance matrix of treatment effects. *A*-optimality criterion is not scale invariant (Kuhfeld, 2010). It is expressed as:

$$A_{opt} = \text{argmin}\{trace[\mathbf{M}(\Omega)]\} \tag{5–1}$$

where $\mathbf{M}(\Omega)$ is the inverse of an information matrix (variance-covariance matrix) of the treatment effects obtained from a design, Ω. *D*-optimality is also a common optimization procedure (Wald, 1943; Kiefer, 1959; Kiefer and Wolfowitz, 1959; Mandal, 2000; Yang, 2008; Kuhfeld, 2010) as it seeks to minimize the determinant of $\mathbf{M}(\Omega)$, expressed as:

$$D_{opt} = \text{argmin}\{|\mathbf{M}(\Omega)|\} \text{ for } |\mathbf{M}(\Omega)| \neq 0. \tag{5–2}$$

Minimizing the determinant of an inverse of an information matrix is equivalent to minimizing the generalized variance of the treatment effects (Kuhfeld, 2010); There is a wide spectrum of search algorithms in literature that can be applied to find improved designs. These include: pairwise swap procedure (John and Williams, 1995), and simulated annealing (SA) (Kirkpatrick *et al.*, 1983), a method that applies a cooling strategy and known to avoid local optimal solutions. They have been used mostly in the analysis of data, with not much done in their applications to improve the efficiency of experimental designs.

Other *R* packages such as *agricolae* (Mendiburu, 2015), *algDesign* (Wheeler, 2014), *experiment* (Imai, 2013), *blockrand* (Snow, 2013), *crossdes* (Sailer, 2013), *OPDOE* (Simecek *et al.*, 2014) and *designGG* (Li *et al.*, 2013) can be useful to generate standard experimental designs. However, their approach of designing and optimizing/improving designs of experiments is different from that presented in this current package since a mixed model framework is used and both genetic and spatial correlations are accounted for at the design stage.

5.2 Statistical Models

Consider the following general linear mixed model:

$$\mathbf{y} = \mu + \mathbf{X}\beta + \mathbf{Z}_1\mathbf{b} + \mathbf{Z}_2\mathbf{g} + \mathbf{e} \tag{5–3}$$

where $\mathbf{y}$ is a vector of continuous phenotypic observations, such as height, yield; $\mathbf{X}$ is a full column rank incidence matrix of fixed block effects; β is a vector of fixed effects (blocks); $\mathbf{Z}_1$ is a full column rank incidence matrix of random effects; $\mathbf{b}$ is a vector of random effects; $\mathbf{Z}_2$ is a full column rank incidence matrix of another random effects; $\mathbf{g}$ is a vector of another random effects; $\mathbf{e}$ is a vector of residual errors;

The assumptions are that $\mathbf{b}$, $\mathbf{g}$ and $\mathbf{e}$ are uncorrelated and that, $\mathbf{b} \sim MVN(0, \mathbf{B})$, where $\mathbf{B} = \sigma_b^2\Phi$, where Φ is a suitable variance-covariance correlation structure. Similarly, $\mathbf{g} \sim MVN(0, \mathbf{G})$, where $\mathbf{G} = \sigma_g^2\mathbf{A}$, where $\mathbf{A}$ is suitable correlation matrix for $\mathbf{g}$, and $\mathbf{e} \sim MVN(0, \mathbf{R})$, where $\mathbf{R} = \sigma_e^2\Phi$, where Φ is a suitable spatial (residual) correlation structure. The correlation structure can take any form such as diagonal or AR1 or any other suitable for that experiment (See Littell *et al.* (2006); Cressie (1993) for more details).

It follows that $\mathbf{V} = var(\mathbf{y}) = var(\mathbf{Z}_1\mathbf{b}) + var(\mathbf{Z}_2\mathbf{g}) + var(\mathbf{e}) = \mathbf{Z}_1\mathbf{B}\mathbf{Z}_1^{'} + \mathbf{Z}_2\mathbf{G}\mathbf{Z}_2^{'} + \mathbf{R}$, where for instance, $\mathbf{G}$ and $\mathbf{R}$ could be variance matrices for genetic effects and residual errors, respectively. The mixed model Equation 5–4 can be solved using Henderson (1975) to obtain BLUPs and BLUEs. The variance-covariance matrix of random effects obtained from solving the mixed models equations can then be used in the optimality procedure from which the trace or determinant of that matrix is calculated with an aim of minimization.

Two specific cases are demonstrated in Subsections 5.2.1 and 5.2.2.

5.2.1 Case 1

This is the case where blocks are fixed effects and treatments and residual errors are random effects. This is typical for randomized complete block (RCB) designs. Treatments are considered random effects since they are a random sample from a much larger set of observations (population). This LMM can be expressed as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{g} + \mathbf{e} \tag{5–4}$$

where $\mathbf{y}$ is a vector of observations; $\mathbf{X}$ is a full column rank incidence matrix of fixed block effects; β is a vector of fixed effects (blocks); $\mathbf{Z}$ is a full column rank incidence matrix of random treatment effects; $\mathbf{g}$ is a vector of random effects (treatments); $\mathbf{e}$ is a vector of residual errors; The assumptions are:

$$\begin{bmatrix} \mathbf{g} \\ \mathbf{e} \end{bmatrix} \sim MVN \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \right)$$

with $\mathbf{V} = var(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}^{'} + \mathbf{R}$, where $\mathbf{G}$ and $\mathbf{R}$ are variance matrices for the genetic effects and residual errors, respectively. Correlated errors can be modeled using a 2-dimensional separable autoregressive spatial error structure of order 1 to model spatial variability along the rows and columns of the experimental layouts (Stringer and Cullis, 2002; Gezan *et al.*, 2010; Gilmour *et al.*, 2009), with $\mathbf{R} = \sigma_e^2 \sum_r(\rho_r) \otimes \sum_c(\rho_c)$, where $\sum_r(\rho_r)$ and $\sum_c(\rho_c)$ are matrices with autocorrelation parameters ρ_r and ρ_c for rows and columns respectively. For genetically related individuals, $\mathbf{G} = \sigma_g^2\mathbf{A}$ where $\mathbf{A}$ corresponds to the additive genetic numerator relationship matrix among individuals, often derived from pedigree (Henderson, 1975, 1984; Mrode, 2014; Gilmour *et al.*, 2009) or, more recently, with molecular information (VanRaden, 2008). Here, narrow-sense heritability h^2 is calculated as $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$.

From this model, $\hat{\beta} = (\mathbf{X}^{'}\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^{'}\hat{\mathbf{V}}^{-1}\mathbf{y}$ which is the Best Linear Unbiased Estimator (EBLUE), and $\hat{\mathbf{g}} = \hat{\mathbf{G}}\mathbf{Z}^{'}\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})$ referred to as Best Linear Unbiased Predictor (EBLUP).

For the computation of this matrix, it has been shown by Harville (1997) and Hooks *et al.* (2009)

$$\mathbf{M}(\Omega) = Var(\hat{\mathbf{g}} - \mathbf{g}) = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z})^{-1} \tag{5–5}$$

where,

$$A_{optimality} = \text{argmin}\,\{trace\,[\mathbf{M}(\Omega)]\}$$

5.2.2 Case 2

In this case, both blocks and treatments are considered to be random effects and an overall mean as a fixed effect. The statistical model can be expressed as

$$\mathbf{y} = \mathbf{X}\mu + \mathbf{Z}_b\mathbf{b} + \mathbf{Z}_g\mathbf{g} + \mathbf{e} \quad \text{or equivalently as ,} \quad \mathbf{y} = \quad \mathbf{X}\mu + \begin{bmatrix}\mathbf{Z}_b & \mathbf{Z}_g\end{bmatrix}\begin{bmatrix}\mathbf{b} \\ \mathbf{g}\end{bmatrix} + \mathbf{e} \tag{5–6}$$

$$= \mathbf{X}\mu + \mathbf{Z}\gamma + \mathbf{e}, \quad \text{where} \quad \mathbf{Z} = \begin{bmatrix}\mathbf{Z}_b & \mathbf{Z}_g\end{bmatrix} \quad \text{and} \quad \gamma = \begin{bmatrix}\mathbf{b} \\ \mathbf{g}\end{bmatrix} \quad \text{and} \quad \mathbf{G} = \begin{bmatrix}\mathbf{D}_b & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_g\end{bmatrix}$$

where $\mathbf{X}$ is a design matrix with a column vector of n ones and μ is overall expected mean, $\mathbf{y}$: vector of response observations, $\mathbf{Z}_b$: is an incidence matrix of blocks effects and $\mathbf{Z}_g$ is an incidence matrix of treatment effects, $\mathbf{b}$ is a vector of random block effects such that $\mathbf{b} \sim MVN(\mathbf{0}, \mathbf{D}_b)$, $\mathbf{g}$ is a vector of random treatments effects such that $\mathbf{g} \sim MVN(\mathbf{0}, \mathbf{G}_g)$, $\mathbf{e}$ is a vector of random errors (residuals) such that $\mathbf{e} \sim MVN(\mathbf{0}, \mathbf{R})$ where $\mathbf{D}_b$, $\mathbf{G}_g$ and $\mathbf{R}$ are variance-covariance matrices for the blocks, treatments and residual errors, respectively.

Estimation of random effects is done by solving a set of linear mixed models equations (Henderson, 1975) yielding

$$\mathbf{M}(\Omega) = \left\{ \begin{bmatrix}\mathbf{Z}_b & \mathbf{Z}_g\end{bmatrix}'\mathbf{R}^{-1}\begin{bmatrix}\mathbf{Z}_b & \mathbf{Z}_g\end{bmatrix} + \begin{bmatrix}\mathbf{D}_b^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_g^{-1}\end{bmatrix} - \begin{bmatrix}\mathbf{Z}_b & \mathbf{Z}_g\end{bmatrix}'\mathbf{K}_x\begin{bmatrix}\mathbf{Z}_b & \mathbf{Z}_g\end{bmatrix} \right\}^{-1} \tag{5–7}$$

where where $\mathbf{K}_x = \mathbf{R}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}$, $\mathbf{D}_b = \sigma_b^2\mathbf{I}_b$ and $\mathbf{G}_g = \sigma_g^2\mathbf{A}_g$, where $\mathbf{A}$ is a numerator relationship matrix calculated from a pedigree of genetic relationships, or $\mathbf{G}_g = \sigma_g^2\mathbf{I}_g$

for genetically unrelated individuals. The matrix $\mathbf{M}(\Omega)$ can be expressed as:

$$\mathbf{M}(\Omega) = \begin{bmatrix} \sum_b(\Omega) & \sum_{bg}(\Omega) \\ \sum_{bg}(\Omega) & \sum_g(\Omega) \end{bmatrix} \tag{5–8}$$

where $\sum_g(\Omega)$ is the portion of the matrix that contains the variance-covariance of treatment effects, from which a trace is calculated. Thus,

$$A_{optimality} = \text{argmin}\left\{ trace\left[\sum\nolimits_g(\Omega)\right]\right\}$$

OptimalDesignMM package aims to improve experimental designs by modeling simultaneously both the genetic relatedness and spatial correlations using a mixed models approach. Genetic relationships are incorporated by reading pedigree information. In addition residual errors are modeled either as independent or correlated such that $\mathbf{R} = \sigma^2{}_e\Phi$, where Φ is a suitable spatial variance-covariance structure. Initial designs are randomly generated and then improved following detailed procedures. The following sections illustrate how to use the functions build in this package to generate and improve experimental designs.

5.3 Example: RCB Designs with Regular-Grid Layouts

Suppose there are 30 treatments, each replicated once in 4 blocks of dimensions 5 rows per block by 6 columns per block, based on *A*-optimality criterion. As discussed above, three forms of genetic relatedness can be conducted: either the treatments being genetically unrelated, or half-siblings or full-siblings. Suppose the treatments have a heritability of $h^2 = 0.3$ and the treatments have a spatial correlation of 0.6 along the rows and columns. Nugget effects are also allowed, and they can be set to zero or to a value between 0 and 1.

To randomly generate an initially unimproved RCB regular-grid design, use the function $rcbd(blocks, genotypes, rb, cb, Tr, Tc, irregular = FALSE)$, where, *blocks* is a numerical value for the number of blocks, *genotypes* is a vector of treatments, *rb*, *cb*, Tr and Tc are integers for numbers of rows per block, columns per block, total rows and total columns, respectively, and *irregular* is a logical statement that is *FALSE* by default to indicate that the shape of the

experimental layout is a regular-grid and if set to $TRUE$ would indicate that it is *irregular* and as such, the *Col* and *Row* coordinates for all the treatments will have to be provided vectors by the user.

```
R> set.seed(100)
R> blocks = 4; genotypes = c(1:30); Tr = 10
        Tc = 12; rb = 5; cb = 6
R> matdf <- rcbd(blocks, genotypes, rb, cb, Tr, Tc)
R> head(matdf[order(matdf[,"Reps"]),])
          Row Col Reps Genotypes
    [1,]   1   1    1        10
    [2,]   1   2    1         8
    [3,]   1   3    1        16
  -------- truncated -----
```

The function $DesLayout(matdf, genotypes, cb, rb, blocks)$ prints the experimental layout, where $matdf$ is an output of the design shown above, and *genotypes*, *cb*, *rb* and *blocks* are as defined earlier.

```
> DesLayout(matdf, genotypes, cb, rb, blocks)
   , , 1
         [,1] [,2] [,3] [,4] [,5] [,6]
   [1,]   10    8   16    2   13   26
   [2,]   20    9   25    4   22   17
   [3,]    6    7   24   11    3    5
   [4,]   28   29   18   19   12   30
   [5,]   14    1   21   23   27   15
   , , 2
   ----- truncated -----
```

Next, the function

$$VarCov.rcbd(matdf, rhox, rhoy, h2, s20, Tr, Tc, criteria = \text{“}A\text{”}, Amat = FALSE, irregular = FALSE)$$

generate relevant matrices required in the mixed model equations. In this case, *matdf* is the design with columns sorted within rows, *rhox* and *rhoy* are spatial correlations along the rows and columns, respectively, $h2$ is narrow-sense heritability of the treatments, $s20$ is a nugget effect, Tr and Tc are as defined above, *criteria* sets the optimality criterion option for either *A*- or *D*- as desired, *Amat* is logical and by default set to *FALSE* implying that no pedigree is required assuming treatments are genetically unrelated, or, *Amat* can be given as a kinship matrix, that is, a numerator relationship matrix showing the pairwise genetic relationships of the treatments, and *irregular* for designs that have more than two sides such as L or T-shaped. The textit below is a simple illustration of the its usage.

```
R> rhox = 0.6; rhoy = 0.6; h2 = 0.3; s20 = 0
R> res = VarCov.rcbd(matdf, rhox, rhoy, h2, s20,
         Tr, Tc, criteria = "A")
R> attributes(res)
   names
    [1] "traceI" "Ginv"   "Rinv"   "K"
R> res$traceI # this is trace value
    [1] 1.582483
R> dim(res$Ginv)
    [1] 30 30
R> dim(res$Rinv)
    [1] 120 120
R> dim(res$K)
    [1] 120 120
```

The output from above are used as input in a function, $Optimize.rcbd(matdf, n, traceI, criteria, Rinv, Ginv, K)$, where $matdf$ and $criteria$ are as defined above, n is the number of iterations desired, that is, the number of times treatments have to be swapped. Some of the swaps will be accepted and therefore deemed successful and others rejected in a process to improve the design. Matrices *Rinv*, *Ginv* and K are inputs from the function *VarCov.rcbd*, where *Rinv* is an inverse of **R**, variance-covariance spatial correlations, *Ginv* is the inverse of **G** and **K** is required internally to calculate the variance-covariance of the random treatment effects. The output of this function includes a value that quantifies how efficient the improved design is compared to the initially unimproved, reporting the percentage overall design efficiency (ODE %). Note that when an A-optimality is used, this translates to the percentage average reduction in variance of the treatment effects.

```
R> # To improve a RCB design
R> traceI=res$traceI; criteria = "A"; n = 5000
R> Rinv = as.matrix(res$Rinv); Ginv = as.matrix(res$Ginv)
R> K = as.matrix(res$K)
R> ans <- Optimize.rcbd(matdf, n, traceI,
        criteria, Rinv, Ginv, K)
    [1] "Swapping within blocks: 3"
    [1] "Swapping within blocks: 4"
           --- truncated ---
    [1] "Swapping within blocks: 2966"
    [1] "ODE due to swapping pairs of treatments within blocks is: 7.16"
R> attributes(ans)
   $names
    [1] "TRACE"       "mat"         "Design_best"
```

Some basic graphics to display the rate of improvement during the optimization process are presented in Figure 5-1. An alternative procedure can be implemented by first, randomly

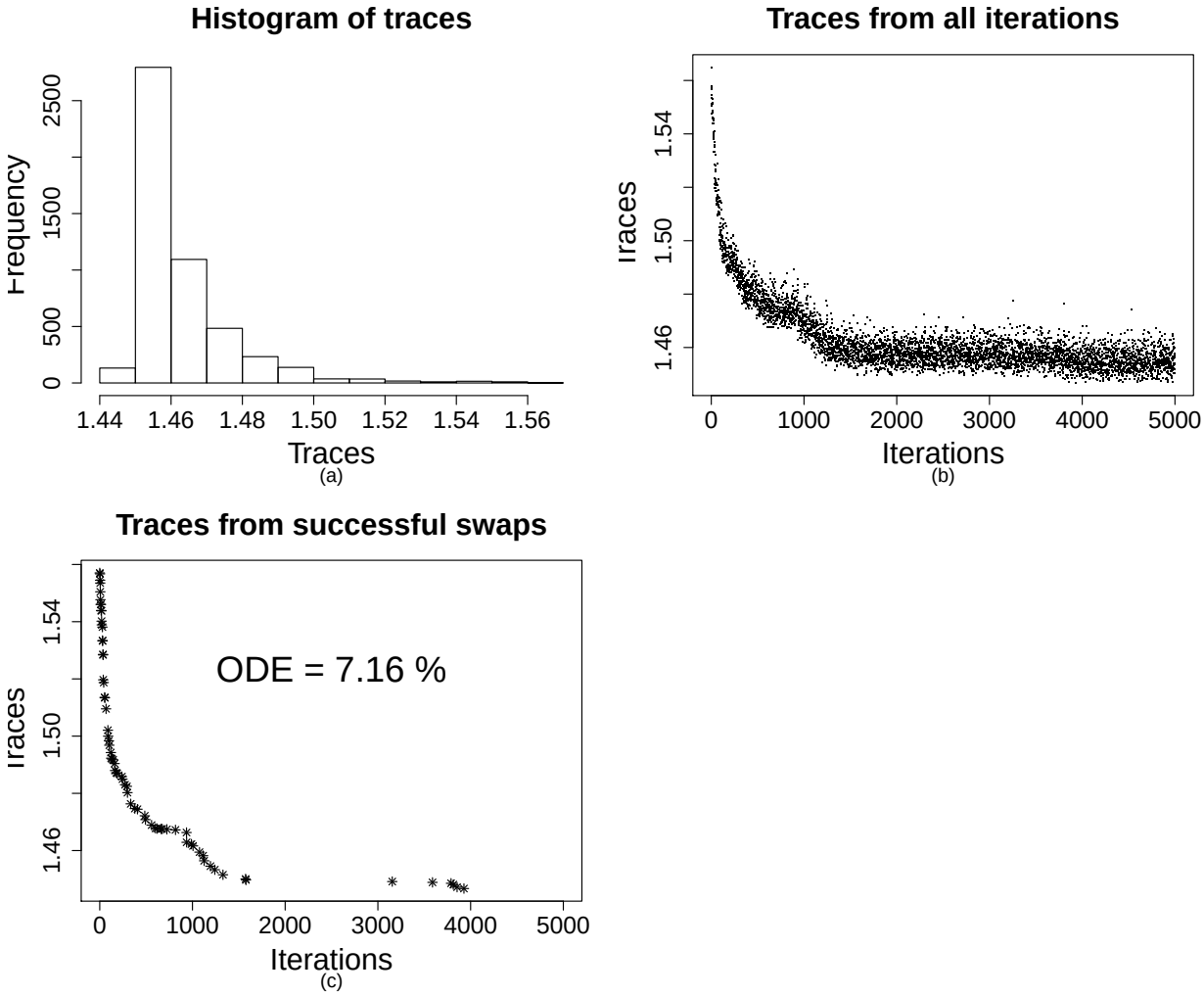


Figure 5-1. An illustration of the optimization process with a total of $p = 5,000$ iterations based on a simple pairwise algorithm for a regular-grid RCB example evaluated under A-optimality criterion with 30 treatments at heritability of 0.3, and spatial correlations of 0.6. The improvement in terms of overall design efficiency (ODE) was 7.16 % reduction in average variance of the treatment effects. (a) is a histogram of all 5,000 traces, (b) is a sctatter plot of all traces and their respective iterations and (c) displays only the successful traces that had lower traces compared to their previous designs.

generating, lets say, $m = 100$ initial unimproved designs, choose the best design ($s = 1$) and improve its layout by iterating it, lets say, $p = 5,000$ times using the textit below. The parameters p, m and s can take any value. However, $s = 1$ is currently implemented as the default due to computational bottlenecks in trying to improve many designs at the same time.

```
R> # Example: Improve designs after generating 100 initial designs
        set.seed(100)
        h2 = 0.3; rhox = 0.6; rhoy = 0.6; s20 = 0; criteria="A"
        blocks = 4; rb = 5; cb = 6; Tr = 10; Tc = 12
         genotypes = c(1:30)
R> res1 =         MultipleDesigns(DesN=100, blocks,
        genotypes, rb, cb, Tr, Tc, Amat = FALSE,
   criteria,h2,rhox,rhoy,s20,irregular=FALSE)
   [1] "generating initial design: 1"
   [1] "generating initial design: 2"
     --- truncated ---
   [1] "generating initial design: 99"
   [1] "generating initial design: 100"
R> attributes(res1)
   $names
    [1] "newmatdf" "initialValues1" "min_initialValues1" "initialValues2"
    [5] "min_initialValues2"
R> res2 <- VarCov.rcbd(matdf=res1$newmatdf,rhox,rhoy,
    h2,s20,Tr,Tc,criteria="A",Amat=FALSE,irregular=FALSE)
R> attributes(res2)
   $names
     [1] "traceI" "Ginv"   "Rinv"   "K"
```

```
R> Rinv = as.matrix(res2$Rinv); Ginv = as.matrix(res2$Ginv)
R> K = as.matrix(res2$K)
```

The spatial correlations along the rows and columns of the experimental design can vary since in most practical cases, there are gaps between rows to allow machinery services for weeding, pesticide control and harvesting or for other management practices.

5.4 Example: RCB Designs with Irregular-Grid Layouts

To generate experimental designs with irregular rectangular layouts, the input $irregular = FALSE$ has to be changed to $TRUE$ and the *Col* and *Row* textits provided as vectors. All other steps are similar to those for RCB designs with regular-grid layouts. Here is an illustration.

```
R> set.seed(100)
R> blocks = 3; genotypes = c(1:9); Tr=6; Tc=6; rb=3; cb=3
R> Row = c(rep(1:3,each=3), rep(4:6,each=3), rep(4:6,each=3))
R> Col = c(rep(1:3,3), rep(1:3,3), rep(4:6,3))
R> matdf <- rcbd(blocks, genotypes,rb,cb, Tr, Tc, irregular=TRUE)
R> res = VarCov.rcbd(matdf, rhox, rhoy, h2, s20, Tr, Tc,
   criteria = "A", irregular = TRUE)
R> traceI=res$traceI; criteria="A"; n=1000
R> Rinv = as.matrix(res$Rinv); Ginv = as.matrix(res$Ginv);
   K = as.matrix(res$K)
R> ans <- Optimize.rcbd(matdf,n,traceI,criteria,Rinv,Ginv,K)
```

5.5 Example: Designs with Genetic and Spatial Correlations

To exemplify the syntax for generating improved experimental designs with genetic relatedness and spatial correlations, suppose that an experiment involves full siblings (treatments that share both mother and father). The following textit could be used to generate and improve the designs.

```
R> h2 = 0.3; rhox = 0.6; rhoy = 0.6; s20 = 0; criteria="A"
R> blocks = 3; rb = 5; cb = 6; Tr = 15; Tc = 6; genotypes = c(1:30)
```

```
R> set.seed(100)
R> # generate an initial unimproved design
R> matdf <- rcbd(blocks, genotypes,rb,cb, Tr, Tc)
R> # improve the design
R> data("ped30fs")
R> Amat <- GenA(male = ped30fs[,"male"], female = ped30fs[,"female"])
R> # suppose we want only offsprings
R> Amat <- as.matrix(Amat[-c(1:5), -c(1:5)])
R> res <- VarCov.rcbd(matdf, rhox, rhoy, h2, s20, Tr, Tc, criteria="A",
    Amat, irregular=FALSE)
R> traceI=res$traceI; criteria="A"
R> Rinv = as.matrix(res$Rinv); Ginv = as.matrix(res$Ginv)
R> K = as.matrix(res$K)
R> ans <- Optimize.rcbd(matdf,n=5000,traceI,criteria,Rinv,Ginv,K)
   [1] "Swapping within blocks: 3"
   [1] "Swapping within blocks: 6"
    --- trucncated ---
   [1] "Swapping within blocks: 4987"
   [1] "Swapping within blocks: 4990"
   [1] "ODE due to swapping pairs of treatments within blocks is: 3.769"
R> attributes(ans)
   $names
   [1] "TRACE"   "mat"    "Design_best"
```

5.6 Generating Unequally Replicated Designs

A linear mixed effects model for unequally-replicated, incomplete blocks and augmented designs can be expressed as

$$y = \mu X + Z_b b + Z_g g + e \quad = \quad \mu X + \begin{bmatrix} Z_b & Z_g \end{bmatrix} \begin{bmatrix} b \\ g \end{bmatrix} + e \tag{5–9}$$

$$= \mu X + Z\gamma + e, \quad \text{where} \quad Z = \begin{bmatrix} Z_b & Z_g \end{bmatrix} \quad \text{and} \quad \gamma = \begin{bmatrix} b \\ g \end{bmatrix}$$

where X: is a column vector of ones, μ is overall expected mean, y: is vector of phenotypic observations, Z_b: is an incidence matrix of blocks effects and Z_g is an incidence matrix of treatment effects, b is a vector of random block effects, g is a vector of random treatments effects, e ia a vector of random errors (residuals). The random parameters are assumed to be a realization from a normal probability distribution with zero means and unknown variances which are estimated using restricted maximum likelihood in mixed models. There distributions are:

$$\begin{bmatrix} b \\ g \\ e \end{bmatrix} \sim MVN \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} D_b & 0 & 0 \\ 0 & G_g & 0 \\ 0 & 0 & R \end{bmatrix} \right) \tag{5–10}$$

Computations of the variance-covariance of the treatment effects are derived from the mixed model solutions(Henderson, 1950) and as described by (Hooks *et al.*, 2009) from where the A- and D-optimality criteria are calculated. To generate an initial unimproved unequally-replicated design, use the function *unequal.RBD* with the syntax shown below. A list has been generated with constraints of the number of treatments available for use in the experiment. The list has the minimum and maximum number allowed of each treatment, and has been textitd here as *min.u* and *max.u* respectively. Usually, this list is provided by the user and the column for frequency ($freq$) actually counts the number of treatments used in the field.

```
R> library(OptimalDesignMM)
R> set.seed(1000)
```

```
R> genotypes = 1:30;criteria="A"
R> blocks = 3; rb=5;cb=6;Tr=5;Tc=18;rhox=0.6
        rhoy=0.6;h2=0.3;s20=0
R> min.u = sample(1:3,length(genotypes),replace=TRUE)
R> max.u = sample(3:5,length(genotypes),replace=TRUE)
R> des1 = unequal.RBD(genotypes, blocks, max.u,
         min.u, rb, cb, Tr, Tc,
    irregular = FALSE)
R> attributes(des1)
    $names
   [1] "matdf"   "datam"   "sumFreq"
R> des1$sumFreq
   [1] 90
R> head(des1$datam)
         genot min.u max.u freq
   [1,]     1     1     4    4
   [2,]     2     3     3    3
   [3,]     3     1     4    3
       ---- truncated  ----
R> # To generate the physical layout
R> DesLayout(matdf=des1$matdf,genotypes, cb, rb, blocks)
```

Now, generate matrices required as input in the next optimization process and calculate a numerator relationship matrix.

```
R> data(ped30hs)
R> Amat <- GenA(male=ped30hs[,"male"], female = ped30hs[,"female"])
R> Amat <- as.matrix(Amat[-c(1:5), -c(1:5)])
R> ans1 <- unequal.VarCov(des1$matdf,rhox,rhoy,h2,s20,Tr,Tc,
```

```
  criteria="A",Amat,sigBl=FALSE,irregular=FALSE)
R> attributes(ans1)
   $names
   [1] "traceI" "Ginv"   "Rinv"   "K"
```

To improve the initial design generated above, optimization algorithms can be run based on swapping pairs within blocks, across blocks, replacing treatments, swapping randomly within and across blocks and by using a suitable combination of these pocedures as illustrated in the following textits.

```
R> traceI <- ans1$traceI
R> Rinv = as.matrix(ans1$Rinv); Ginv = as.matrix(ans1$Ginv)
         K = as.matrix(ans1$K)
 R> Results <- unequal.Optimize.SwapsWithin
         (matdf = des1$matdf, n = 5000,
   traceI, criteria, Rinv, Ginv, K)
   [1] "Swapping within blocks: 2"
   [1] "Swapping within blocks: 7"
   [1] "Swapping within blocks: 9"
    --- truncated ---
   [1] "Swapping within blocks: 4774"
   [1] "ODE due to swapping pairs of treatments within
    blocks is: 6.481182"
R> attributes(Results)
   $names
   [1] "ODE"   "TRACE"  "mat"   "Design_best"
```

To use the replacing treatments algorithm, the syntax to be used to improve the design is $unequal.Optimize.Rpl(matdf, n, traceI, criteria, Rinv, Ginv, K, genotypes, max.u, min.u)$ as shown here where the inputs are as previously defined.

```
R> # Optimize by replacing treatments using a list of constraints
R> Results <- unequal.Optimize.Rpl(matdf=des1$matdf,n=5000,
   traceI,criteria,Rinv,Ginv,K,genotypes,max.u,min.u)
    [1] "Replacing treatments: 3"
    [1] "Replacing treatments: 6"
     --- truncated ---
    [1] "Replacing treatments: 507"
    [1] "Replacing treatments: 737"
    [1] "ODE due to replacing treatments is: 5.483255"
```

Similarly, treatments may be swapped using the $unequal.Optimize.Anypair(matdf, n, traceI, criteria, Rinv, Ginv, K)$ which randomly swaps any two pairs of treatments that could be within or across blocks and accepts designs with smaller trace or determinants values than the previous design.

```
R> Results <- unequal.Optimize.Anypair(matdf = des1$matdf,
         n = 5000,
   traceI,criteria,Rinv,Ginv,K)
   [1] "Swapping treatments: 2"
   [1] "Swapping treatments: 4"
   [1] "Swapping treatments: 6"
      --- truncated ---
   [1] "Swapping treatments: 679"
   [1] "Swapping treatments: 770"
   [1] "ODE due to swapping any pairs of treatments is: 4.181371"
R> attributes(Results)
   $names
   [1] "ODE"   "TRACE"   "mat"   "Design_best"
```

Swapping across blocks is done using the syntax

```
R> Results <-unequal.Optimize.Across(matdf=des1$matdf,n=5000,
   traceI,criteria,Rinv,Ginv,K)
   [1] "Swapping treatments: 2"
   [1] "Swapping treatments: 5"
     ----- truncated  ----
   [1] "Swapping treatments: 1790"
   [1] "Swapping treatments: 2169"
   [1] "ODE due to swapping treatments across blocks: 4.931635"
```

If desired, improving the unequally-replicated designs can involve a combination of the above procedures, that is, replacement, swapping within, cross or by calling their respective functions. For instance

```
R> optimalDs <- function(matdf, n, criteria = "A", Amat = FALSE,
   sigBl = FALSE, irregular = FALSE)
  {
     res1 <- unequal.Optimize.Rpl(matdf = des1$matdf, n,
     traceI = ans1$traceI, criteria, Rinv = ans1$Rinv,
     Ginv = ans1$Ginv, K=ans1$K, genotypes, max.u, min.u)
     ans0 <- unequal.VarCov(matdf = res1$Design_best, rhox, rhoy,h2,
     s20, Tr, Tc, criteria, Amat, sigBl, irregular)
     res2 <- unequal.Optimize.SwapsWithin(matdf = res1$Design_best, n,
     traceI = ans0$traceI, criteria, Rinv = ans0$Rinv,
     Ginv = ans0$Ginv, K = ans0$K)
      ODE_Total = ((res1$mat[1,"value"] - res2$mat[nrow(res2$mat),
     "value"])/res1$mat[1,"value"])*100
     print(sprintf("Effective ODE is: %f", ODE_Total, "complete\n",
     sep = ""))
     list(ODE_effective = ODE_Total, Replacement = res1,
```

```
     WithinBlock = res2)
 }
```

Then, use it as shown below:

```
R> library(OptimalDesignMM)
R> genotypes = 1:30; criteria = "A"
R> blocks = 3; rb = 5; cb = 6; Tr = 15; Tc = 6
R> rhox = 0.3; rhoy = 0.9; h2 = 0.3; s20 = 0.1
R> min.u = sample(1:4,length(genotypes), replace=TRUE)
R> max.u = sample(4:8,length(genotypes), replace=TRUE)
R> des1 = unequal.RBD(genotypes, blocks, max.u, min.u,
    rb, cb, Tr, Tc, irregular = FALSE)
R> data("ped30fs")
R> data(ped30fs)
R> Amat <- GenA(male=ped30fs[,"male"], female = ped30fs[,"female"])
R> Amat <- as.matrix(Amat[-c(1:5), -c(1:5)])
R> ans1 <- unequal.VarCov(des1$matdf, rhox, rhoy, h2,
   s20, Tr, Tc, criteria = "A", Amat)
R> answer1 <- optimalDs(matdf = des1$matdf, n = 1000, criteria = "A", Amat)
   [1] "Replacing treatments: 3"
           --truncted --
   [1] "Replacing treatments: 588"
   [1] "ODE due to replacing treatments is: 3.792448"
        --truncated --
   [1] "Swapping within blocks: 853"
   [1] "ODE due to swapping treatments within blocks is: 1.412301"
   [1] "Effective ODE is: 5.151187"
```

5.7 Generating Incomplete Block Designs

For unbalanced incomplete block designs, it is required that $blocks * k = t * r$ (Kuehl, 2000), where k is block size, t is number of treatments and r is the number of times each treatment has been replicated in the whole experiment and in most cases, $k < t$. The textits illustrate ways of generating and improving such designs. First, we could right a function that wraps several functions of interest so that we can optimize an experiment by both swapping treatments across and within blocks: $optimal.ibd(matdf, n, criteria = \text{“}A\text{”}, Amat = FALSE, sigBl = FALSE, irregular = FALSE)$

```
R> optimal.ibd <- function(matdf, n, criteria = "A", Amat = FALSE,
   sigBl = FALSE, irregular = FALSE)
  {
     res0 <- unequal.Optimize.Across(matdf, n, traceI = ans1$traceI,
     criteria, Rinv = ans1$Rinv, Ginv = ans1$Ginv, K=ans1$K)
     ans01 <- unequal.VarCov(matdf = res0$Design_best, rhox,
     rhoy, h2, s20, Tr, Tc, criteria, Amat, sigBl, irregular)
     res2 <- unequal.Optimize.SwapsWithin(matdf = res0$Design_best,
     n, traceI=ans01$traceI, criteria, Rinv = ans01$Rinv,
     Ginv = ans01$Ginv,K = ans01$K)
     ODE_Total = ((res0$mat[1,"value"] - res2$mat[nrow(res2$mat),
     "value"])/res0$mat[1,"value"])*100
     print(sprintf("Effective ODE is: %f", ODE_Total, "complete\n",
     sep = ""))
     list(ODE_effective = ODE_Total, Swapping_AnyPair=res0,
     Swapping_within_blocks = res2)
 }
```

Then, generate an incomplete block design with blocks*blocksize = trt*replicates = $3*20 = 30*2$ as shown here

```
R> genotypes = 1:30; criteria = "A"; blocks = 3
R> rb = 5; cb = 4; Tr= 5; Tc = 12; rhox = 0.6
R> rhoy = 0.9; h2 = 0.3; s20 = 0
R> min.u = rep(2,length(genotypes))
R> max.u = rep(2,length(genotypes))
R> # generate the  initial design
R> des1 = unequal.RBD(genotypes, blocks, max.u, min.u, rb,
   cb, Tr, Tc, irregular = FALSE)
R> attributes(des1)
    $names
    [1] "matdf"   "datam"   "sumFreq"
R> des1$sumFreq
   [1] 60
R> DesLayout(des1$matdf, genotypes, cb, rb, blocks)
R> # generate matrices and trace values for input later
R> ans1 <- unequal.VarCov(des1$matdf,rhox,rhoy,h2,s20,Tr,Tc,criteria)
R> # Improve the above design
R> answer1 <- optimal.ibd(des1$matdf,n=1000,criteria = "A")
   [1] "Swapping treatments: 3"
        ---truncated ---
   [1] "Swapping treatments: 879"
   [1] "ODE due to swapping treatments across blocks: 4.450752"
   [1] "Swapping within blocks: 3"
          --truncated --
   [1] "Swapping within blocks: 625"
   [1] "ODE due to swapping treatments within blocks is: 0.640636"
   [1] "Effective ODE is: 5.062874"
```

```
R> attributes(answer1)
   $names
    [1] "ODE_effective"          "Swapping_Across"
R> ls(answer1$Swapping_Across)
      [1] "Design_best" "mat"         "ODE"         "TRACE"
R> ls(answer1$Swapping_within_blocks)
     [1] "Design_best" "mat"         "ODE"         "TRACE"
```

5.8 Generating Augmented Designs

Augmented designs with replicated controls and unreplicated new treatments can be generated as shown in this section where *CheckPlots* are the controls for each block that will be replicated, *Reps.Per.Block* is the number of times to replicate the controls in each block, *rhox* is spatial correlation along the rows, *rhoy* is spatial correlation along the columns of the experimental design.

```
R> CheckPlots = c(1:2); Treatments = c(3:92)
R> Reps.Per.Block = 5; rb = 8; cb = 5; blocks = 3
R> rhox=0.3; rhoy=0.9; h2=0.3; s20=0; criteria="A"
R> # generate design
R> matdf = rcbd.Augmented(blocks, Treatments, CheckPlots,
   Reps.Per.Block, rb, cb)
R> genotypes = c(CheckPlots, Treatments)
R> DesLayout(matdf, genotypes, cb, rb, blocks)
# calculate trace and other variance matrices
R> ans1 = unequal.Augmented.VarCov(matdf, rhox, rhoy, h2, s20,
   criteria = "A", Amat = FALSE, sigBl = FALSE)
R> attributes(ans1)
   $names
     [1] "traceI" "Ginv"   "Rinv"   "K"
```

```
R> # optimize the design by swapping treatments within blocks
R> answer1 <- unequal.Optimize.SwapsWithin(matdf,n=2000,
     traceI=ans1$traceI, criteria="A",Rinv=ans1$Rinv,
     Ginv=ans1$Ginv,K=ans1$K)
    [1] "Swapping within blocks: 2"
    [1] "Swapping within blocks: 3"
          --- truncated --
    [1] "Swapping within blocks: 1246"
    [1] "ODE due to swapping treatments within blocks is: 4.105"
R> attributes(answer1)
   $names
     [1] "ODE"  "TRACE"  "mat"  "Design_best"
```

5.9 Optimization Using Simulated Annealing

Other than improving a RCB experimental design using a simple pairwise swap algorithm, *OptimalDesignMM* has also implemented a simulated annealing procedure (Kirkpatrick *et al.*, 1983) which can be very efficient when used with an *A*-optimality criterion. Simulated annealing is a powerful optimization procedure that has the potential to prevent the search from getting trapped in a local minima or maxima (Robert and Casella, 2010). It has been used by other researchers with diverse applications such as to improve RCB designs, to optimize long term forest planning management (Borges *et al.*, 2014), to estimate an optimal combination level of stand paths for forests Seo *et al.* (2005) and Liu *et al.* (2006) to optimize harvest scheduling problems with spatial constraints in forest planning and management.

The steps to generate and improve RCB designs using simulated annealing is similar to the previous examples except that the function *Optimize_SimAnn_rcbd* has to be used to call the simulated annealing algorithm as shown in this example.

```
R> library(OptimalDesignMM)
     Loading required package: Matrix
```

```
     Loading required package: nadiv
R> h2 = 0.3; rhox = 0.6; rhoy = 0.6; s20 = 0; criteria="A"
R> blocks = 3; rb = 5; cb = 6; Tr = 15; Tc = 6
         genotypes = c(1:30)
R> set.seed(100)
R> matdf<- rcbd(blocks, genotypes,rb,cb, Tr, Tc)
R> res <- VarCov.rcbd(matdf,rhox,rhoy,h2,s20,Tr,Tc,criteria="A",
          Amat=FALSE,irregular=FALSE)
R> traceI=res$traceI; criteria="A"
R> Rinv = as.matrix(res$Rinv); Ginv = as.matrix(res$Ginv)
     K = as.matrix(res$K)
R> # Now go through 2000 iterations
R> ans <- Optimize_SimAnn_rcbd(matdf,n=2000,traceI,
         criteria,Rinv,Ginv,K)
   [1] "Swapping within blocks: 5"
   [1] "Swapping within blocks: 7"
       --- truncated ---
   [1] "Swapping within blocks: 1565"
   [1] "Swapping within blocks: 1892"
   [1] "ODE due to simulated annealing is: 6.409"
R> DesLayout(matdf=ans$Design_best, genotypes, cb, rb, blocks)
```

5.10 Extensions

The package *OptimalDesignMM* has also implemented more aggressive variants of the swap procedure, known as greedy algorithms that are similar to the simple pairwise swap procedure except that they allow more than 2 treatments to be swapped at the same time on every iteration. Greedy algorithms allow any even number of treatments to be randomly selected and swapped within blocks and the function *OptimizeGreedy.rcbd* requires the syntax shown below

where *gsize* is an even number of treatments to be swapped, with all other terms being as defined above.

$OptimizeGreedy.rcbd(matdf, n, traceI, criteria, gsize, Rinv, Ginv, K)$

An additional algorithm that has been implemented to improve experimental designs is known as genetic nearest neighbor which swaps treatments depending on how strong their genetic correlations are and how far they are positioned apart in the rectangular grid on the physical layout of an RCB experiment. Note that simulated annealing and simple pairwise swap algorithms have been shown to be superior to both the variants of greedy and genetic nearest neighbor algorithms (Chapter 3). The syntax for using the genetic nearest neighbor algorithm is:

$Optimize\_GNN\_rcbd(matdf, n, traceI, criteria, Amat, Rinv, Ginv, K)$

where *Amat* is a matrix of numerator relationship matrix (**A**) representing pairwise genetic relationships, also called a kinship matrix.

5.11 Discussion

The current version of this package *OptimalDesignMM* has demonstrated some stochastic procedures to improve experimental designs using linear mixed models approach, with illustrations provided for RCB designs and non-orthogonal experimental designs such as unequally-replicated designs, incomplete block and augmented block designs. Where an initial experimental design is already available, having been generated from another software tool, it possible to read it into this package, organize it and apply appropriate procedure to improve such designs. Extensions to allow user-defined matrix of correlations other than the autoregressive correlation of order 1 (AR1) are in progress. This will allow either other standard correlation and variance-covariance matrices such as uniform heterogeneous (CORUH), unstructured (US), uniform correlation (CORUV) also known as compound symmetry, diagonal (DIAG) (Littell *et al.*, 2006) and other user-defined variance-covariance matrices to be implemented. A numerator relationship matrix can be calculated either from outside this package using other standard packages and be read in as *Amat* or can be calculated from within this software. In the

later case, the pedigree file has to be ordered by generations with parents appearing on top of the list.

CHAPTER 6
CONCLUSIONS

Incorporation of spatial heterogeneity and genetic relatedness among experimental units is critical in plant breeding programs and can be successfully addressed both at the design and analysis stages as ignoring these elements results in less precise estimation and poor prediction of parameters. This research has focused on the design aspects, so that design layouts can be generated optimally, for different field conditions. Design aspects are often overlooked and ignored mainly due to computational difficulties, but rather more emphasis is given to analysis of data that has been already collected often from non-optimal designs Stroup (2013). It is also important to note that both balanced and unbalanced experimental designs are inevitable in many research settings, and the amount of improvement to be realized will depend on the type of experiment, field experimental conditions, search algorithms and a choice of optimality criterion.

In this research, several computational procedures and statistical models have been presented that can be used to generate improved experimental designs for balanced and unbalanced situations by considering simultaneously genetic and spatial correlations at the design stage. A linear mixed model framework has been used, due to its flexibility to incorporate a genetic relationship matrix and a spatial error structure to reduce background noise thus leading to more accurate and precise estimates of variance components and model effects. Also, the use of *A*- and *D*- information based optimality criteria to generate improved designs has shown to be fruitful in the current research and in other studies such as Cullis *et al.* (2006) and Butler *et al.* (2008).

Unlike many other studies that discussed optimality procedures by using fixed effects models where they assumed both blocks and treatments to be fixed effects (Das, 2002; John and Williams, 1995; Kuhfeld, 2010), or Filho and Gilmour (2003) who considered genetic relationships with no spatial error correlations, the implemented procedure presented in this research provides flexible options to account for genetic relationships and/or spatial correlations. Findings from this study have unfolded the potential variations in levels of design improvement

when experiments are conducted with varying field conditions, with results, as expected, indicating that relative design efficiency varies with levels of heritability, genetic and spatial correlations. It has been demonstrated throughout that simultaneous considerations for genetic and environmental correlations should be incorporated to generate better experimental designs with important improvements in relative design efficiency and prediction accuracies of random treatment effects.

Results from the evaluations performed in this study are presented for an array of conditions with heritability levels, h^2, of 0.1, 0.3, 0.6, spatial correlation levels, ρ, of 0.0, 0.1, 0.3, 0.6, 0.9, genetic relationships such as genetically unrelated individuals, half-sib and full-sib families and search algorithms including simple pairwise, simulated annealing, some variants of pairwise also known as greedy and genetic neighborhood procedures. The measure of relative design efficiency compares between initial (un-improved) randomly generated designs to that of improved, after several iterations, for all evaluated conditions. In this study, search algorithms have been applied to assess how well they can be used to improve the efficiency of experimental designs, with results indicating that a simple pairwise algorithm as well as simulated annealing can substantially improve the efficiency of experimental designs under *A*-optimality criterion, and also under *D*-optimality criterion when the simple pairwise procedure is used.

When simple pairwise algorithm is used, the observed criterion values which are traces for *A*- and determinants for *D*-optimality criterion have been found to be highly correlated. This is not unusual as both criteria are a convex function of the eigenvalues of information matrix (Das, 2002; Kuhfeld, 2010) and follows with their mathematical definitions, as *A*-optimality is a function of the arithmetic mean of the eigenvalues whereas *D*-optimality is a function of the geometric mean of the eigenvalues (Kuhfeld, 2010).

Specifically, results from Chapter 2 about improving randomized complete block designs, indicated that experiments with genetically unrelated individuals have a greater room of improvement as they had the highest overall design efficiencies of 8.739 % when evaluated with heritability of 0.3 and spatial correlation of 0.6. When RCB designs are generated with

half-sib or full-sib families, optimization procedure may yield to important improvements under the presence of mild ($\rho = 0.6$) to strong ($\rho = 0.9$) spatial correlation levels and relatively low heritability values ($h^2 = 0.1$). Also, as expected, results showed that accuracy of prediction of genetic values increases as the levels of heritability and spatial correlations increase and that, improved designs present, slightly more precise estimates of heritabilities than those from un-improved experiments. In addition, better prediction accuracies were also found for mixed models that accounted for spatial correlation using AR1 based compared to models that assumed that residual errors were identical and independently distributed (iid).

In Chapter 3, where focus was to search for computationally efficient algorithms and procedures to improve experimental designs, a process that has been deemed to be computationally challenging, with other researchers opting for approximations (Butler *et al.*, 2008), fundamental findings have been obtained. In particular, the evaluations, spanning several search algorithms, over a range of heritabilities and for a given spatial correlation of 0.6 gave promising improvements in design efficiencies. Results indicate that for the evaluated conditions, based on both *A*- and *D*-optimality criterion, the best performing algorithm is simple pairwise, which achieved the highest design efficiencies. Under *A*-criterion, both simple pairwise and simulated annealing procedures performed best with the lowest appearing to be the genetic neighborhood algorithm. Based on *D*-optimality criterion, results indicated that simulated annealing performs poorly than any other search algorithm. Relative design efficiencies observed from experiments with full-sib families show a decrease in their efficiencies as heritability increases under *A*-optimality criterion. In addition, the number of successful swaps in each of the search algorithm decrease with increasing heritability and are highest for both simulated annealing and simple pairwise procedure and lowest for genetic neighborhood algorithm. Findings from Chapter 3 indicate that there is a relevant potential to improve experimental designs, and that, the level of design improvement can be highest when simple pairwise or simulated annealing algorithms are used under *A*-optimality criterion. Most important is that, if

D-optimality criterion is used, then greater efficiency gains are achievable through the use of a simple pairwise search algorithm.

From Chapter 4, where non-orthogonal designs were evaluated for design efficiency under a span of conditions, results indicated that for the unequally replicated experiments, a high reduction in average variance of treatment effects (about ODE of 9 %) can be obtained among genetically unrelated individuals at $\rho = 0.6$ with $h^2 = 0.3$. When incomplete block designs are used, highest design improvements (about ODE of 10 %) are feasible among genetically unrelated individuals with $\rho = 0.9$ and $h^2 = 0.1$. Although these results agree with Filho and Gilmour (2003), who also reported high levels of design improvement among genetically unrelated individuals, the current research has revealed that it is not only about genetically relatedness, but also, levels of spatial correlations and genetic relationship play an important role in determining the magnitude of design efficiency.

Unreplicated trials, such as augmented designs, have the potential to achieve high design improvements among full-sib families when heritability is lowest, at, $h^2 = 0.1$, with $\rho = 0.6$. Unlike RCB designs that achieved no improvement at all when spatial correlation was zero, it is still possible to obtain some level of design improvements for unbalanced designs. In general, for unbalanced designs, there are varying levels of design efficiency that can be achieved for different experiments, given their levels of heritability, genetic relatedness and spatial correlations. Results obtained from designs with irregular-grid layouts showed a similar pattern of design efficiencies as that from regular-grid, with the level of improvement being dependent on how far physically two blocks are set apart from each other. Higher design efficiencies can be obtained from irregular-grid experimental designs when blocks are more isolated than when they are set close together, reflecting the need to find appropriate experimental designs for irregular-grids.

Time required to improve an experimental design varies. For instance, RCB designs with 30 genotypes and six blocks, on average, takes about 2 to 3 min for 5,000 iterations. Incomplete block and unequally replicated designs with 30 genotypes and six blocks, require $\approx$ 5 min, and takes $\approx$ 1 hr for augmented design that has 492 unreplicated test treatments and three replicated

controls arranged in a total of three blocks of dimensions 10 rows by 20 columns. These times were obtained using high performance computers with varied CPU processing speeds. It takes about the same time, or a little bit shorter, to run similar jobs, one at a time, from a 64-bit windows operating system Intel(R) Core(TM) i7-4720HQ CPU@2.60GHz, RAM = 8.0GB using *R* (R Core Team, 2016). However, the advantage of using high performance computing is parallelization, where multiple jobs run at the same time.

In summary, this study has shown that experimental designs have varied levels of design efficiencies under different experimental conditions. They can be generated and improved efficiently using a mixed model framework by including sources of genetic and non-genetic variations in the model and by implementing an optimization algorithm that maximizes the information extracted from field trials together with a choice of optimality criteria. From this research, the recommended optimality criterion is *A*- due to its consistency in results based on simple pairwise and simulated annealing algorithms, its low computational demands, as it only calculates a trace not the determinant of a variance-covariance matrix of the random treatment effects, and most importantly it can easily handle highly sparse variance-covariance matrices of random treatment effects.

An *R* package, called, *OptimalDesignMM*, that implements the procedures described in this research, using mixed models amidst other search algorithms and optimal criteria to improve the efficiency of experiment designs, has been developed. Although other *R* packages such as *agricolae* (Mendiburu, 2015), *algDesign* (Wheeler, 2014), *experiment* (Imai, 2013), *blockrand* (Snow, 2013), *crossdes* (Sailer, 2013), *OPDOE* (Simecek *et al.*, 2014) and *designGG* (Li *et al.*, 2013) exist, their approach of designing and optimizing/improving designs of experiments is different from that presented by *OptimalDesignMM* package since a mixed model framework is used here and both genetic and spatial correlations are simultaneously accounted for at the design stage, whereas most of the aforementioned packages base their designs on fixed effects models, particularly, where treatments are fixed effects.

It is also envisaged that this *R* library will expand its functionalities to include many other variance-covariance matrices and more experimental designs that have not yet been implemented. Also, it will be more computationally efficient with the advent of faster programming languages and techniques.

The proposed procedures can be easily extended to include other complex experimental designs and variance-covariance error structures (Stroup, 2013; Cressie, 1993; Gilmour *et al.*, 2009; Zuur *et al.*, 2009; Littell *et al.*, 2006). Computational efficiency of the presented algorithms can be improved, for instance, by limiting the use of loops in the functions and adopting more vectorization. In addition, other variants of search algorithms can be implemented. For the search algorithms that did not do well, such as genetic neighborhood procedure, a value different from 0.25 could be chosen to indicate which treatments to be swapped. It is not known whether changing this value to a higher coefficient would increase the efficiency of the genetic neighborhood algorithm.

Most importantly, computational efficiency can be improved with more efficient programming languages such C++, Fortran or Python. It is highly recommended to write computer code in a faster programming environment and interface that with the free and open source *R* (R Core Team, 2016), which will solve computational challenges usually encountered in generation of large optimal field trials with thousands of treatments. Also, for the simulated annealing, different variants of cooling schedules can be developed and tested for efficiency. Other than using pedigree information to calculate a numerator relationship matrix (Falconer and Mackay, 1996), molecular markers such as SNPs can be used to calculate a genomic relationship matrix (Beaulieu *et al.*, 2014; Hill *et al.*, 2008; VanRaden, 2008) which can be easily incorporated under the developed algorithms, in the linear mixed model to account for genetic sources of variation, as pointed out by Habier *et al.* (2007), who stated that genomic prediction accuracies might yield superior results compared to pedigree-based if markers are in linkage disequilibrium with causal loci.

In conclusion, this research has demonstrated that due to the existence of correlated observations, as evidenced by varying genetic and non-genetic experimental conditions, the use of a linear mixed model framework to account for possible sources of variations and incorporation of a simple pairwise or simulated annealing search algorithm, together with a suitable optimality criterion, such as *A*-, have a great potential to improve the efficiency of balanced and unbalanced experimental designs.

APPENDIX A
OTHER OPTIMALITY CONDITIONS

A.1 Completely Randomized Designs with Spatial Correlations

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\tau + \mathbf{e} = \mathbf{W}\gamma + \mathbf{e}, \quad \text{where,} \quad \mathbf{W} = \begin{bmatrix} \mathbf{1} & \mathbf{X} \end{bmatrix} \quad \text{and} \quad \gamma = \begin{bmatrix} \mu \\ \tau \end{bmatrix}, \tag{A–1}$$

where $\mathbf{y}$ is a vector of observations; $\mathbf{1}$ is a column of ones; μ is an overall mean; $\mathbf{X}$ is an incidence matrix of fixed treatment effects; τ is a vector of fixed treatment effects; $\mathbf{e}$ is a vector of residual errors such that $\mathbf{e} \sim N(0, \mathbf{R})$, where $\mathbf{R} = \sigma_e^2\theta$, θ is a spatial correlation matrix such as $\mathbf{R} = \sigma_e^2\Sigma_r(\rho_r) \otimes \Sigma_c(\rho_c)$ (Gilmour *et al.*, 2009); $\mathbf{W}$ is a partitioned incidence matrix of all fixed effects and γ is a partitioned vector of all fixed treatment effects.

A.1.1 Ordinary Least Squares Approach

$$\mathbf{e} = \mathbf{y} - \mathbf{W}\gamma \tag{A–2}$$

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{W}\gamma)'(\mathbf{y} - \mathbf{W}\gamma) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{W}\gamma - \mathbf{W}\gamma'\mathbf{y} + \mathbf{W}\gamma'\mathbf{W}\gamma$$

$$= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{W}\gamma + \gamma'\mathbf{W}'\mathbf{W}\gamma$$

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \gamma} = 0 - 2\mathbf{y}'\mathbf{W} + 2\mathbf{W}'\mathbf{W}\gamma \quad \text{and set it equal to zero, gives:}$$

$$2\mathbf{y}'\mathbf{W} = 2\mathbf{W}'\mathbf{W}\hat{\gamma} \Rightarrow \mathbf{W}'\mathbf{W}\hat{\gamma} = \mathbf{y}'\mathbf{W} \Rightarrow \hat{\gamma} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{y}'\mathbf{W}$$

$$var(\hat{\gamma}) = (\mathbf{W}'\mathbf{W})^{-1}\hat{\mathbf{R}}\mathbf{W}[(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}]' = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\hat{\mathbf{R}} = (\mathbf{W}'\mathbf{W})^{-1}\hat{\mathbf{R}}$$

$$var(\hat{\gamma}) = (\mathbf{W}'\mathbf{W})^{-1}\hat{\mathbf{R}} = (\mathbf{W}'\hat{\mathbf{R}}^{-1}\mathbf{W})^{-1} = \mathbf{M},$$

where $\mathbf{M}$ is the matrix to be minimized for a given design.

Thus, $A_{opt} = \text{argmin}\{trace(\mathbf{M})\}$ and $D_{opt} = \text{argmin}\{determinant(\mathbf{M})\}$

A.1.2 Matrix Approach

$$\text{From} \quad \mathbf{y} = \mathbf{W}\gamma + \mathbf{e} \tag{A–3}$$

$$\mathbf{W}'\mathbf{y} = \mathbf{W}'\mathbf{W}\gamma + \mathbf{W}'\mathbf{e} = \mathbf{W}'\mathbf{W}\gamma, \quad \text{since} \quad \mathbf{W}'\mathbf{e} = 0, \quad \text{thus,} \quad \gamma = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y}$$

$$var(\hat{\gamma}) = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\hat{\mathbf{R}}[(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}']' = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\hat{\mathbf{R}} = (\mathbf{W}'\mathbf{W})^{-1}\hat{\mathbf{R}}$$

$$var(\hat{\gamma}) = (\mathbf{W}'\mathbf{W})^{-1}\hat{\mathbf{R}} = (\mathbf{W}'\hat{\mathbf{R}}^{-1}\mathbf{W})^{-1} = \mathbf{M},$$

where $\mathbf{M}$ is the matrix to be minimized for a given design.

$A_{opt} = \text{argmin}\{trace(\mathbf{M})\}$ and $D_{opt} = \text{argmin}\{determinant(\mathbf{M})\}$ where $\mathbf{M}$ is as given in Equation A–3.

A.2 Randomized Complete Block Designs with Fixed Blocks and Treatments Effects

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{W}_\alpha\alpha + \mathbf{W}\tau + \mathbf{e} = \begin{bmatrix}\mathbf{1} & \mathbf{W}_\alpha\end{bmatrix}\begin{bmatrix}\mu \\ \alpha\end{bmatrix} + \mathbf{W}\tau + \mathbf{e} = \mathbf{X}\beta + \mathbf{W}\tau + \mathbf{e} \tag{A–4}$$

$$\text{denoting,} \quad \mathbf{X} = \begin{bmatrix}\mathbf{1} & \mathbf{W}_\alpha\end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix}\mu \\ \alpha\end{bmatrix}$$

$$\text{This can be written as} \quad \mathbf{y} = \mathbf{P}\nu + \mathbf{e}, \text{where} \quad \mathbf{P} = \begin{bmatrix}\mathbf{X} & \mathbf{W}\end{bmatrix} \quad \text{and} \quad \nu = \begin{bmatrix}\beta \\ \tau\end{bmatrix} \tag{A–5}$$

where $\mathbf{y}$ is a vector of observations; $\mathbf{1}$ is a column of ones; μ is an overall mean; $\mathbf{W}_\alpha$ is an incidence matrix of fixed block effects; α is a vector of fixed block effects; $\mathbf{W}$ is an incidence matrix of fixed treatment effects; τ is a vector of fixed treatment effects; $\mathbf{e}$ is a vector of residual errors such that $\mathbf{e} \sim N(0, \mathbf{R})$, where $\mathbf{R} = \sigma_e^2\theta$, θ is a spatial correlation matrix; $\mathbf{X}$ and $\mathbf{P}$ are partitioned incidence matrices of fixed effects and β and ν are partitioned vectors of fixed effects.

$$From \quad \mathbf{y} = \mathbf{P}\nu, \tag{A–6}$$

$$\mathbf{P}'\mathbf{y} = \mathbf{P}'\mathbf{P}\nu + \mathbf{P}'\mathbf{e} \Rightarrow \nu = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{y}$$

$$var(\hat{\nu}) = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\hat{\mathbf{R}}[(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}']' = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{P}(\mathbf{P}'\mathbf{P})^{-1}\hat{\mathbf{R}} = (\mathbf{P}'\mathbf{P})^{-1}\hat{\mathbf{R}}$$

$$var(\hat{\nu}) = (\mathbf{P}'\mathbf{P})^{-1}\hat{\mathbf{R}} = (\mathbf{P}'\hat{\mathbf{R}}^{-1}\mathbf{P})^{-1}$$

$$\text{Note that} \quad \mathbf{P}'\hat{\mathbf{R}}^{-1}\mathbf{P} = \begin{bmatrix}\mathbf{X} & \mathbf{W}\end{bmatrix}' \hat{\mathbf{R}}^{-1} \begin{bmatrix}\mathbf{X} & \mathbf{W}\end{bmatrix} = \begin{bmatrix}\mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{W} \\ \mathbf{W}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{W}'\hat{\mathbf{R}}^{-1}\mathbf{W}\end{bmatrix} \tag{A–7}$$

To find $(\mathbf{P}'\hat{\mathbf{R}}^{-1}\mathbf{P})^{-1}$, we can use therem 8.5.11 from Harville (1997) to find the inverse of a partitioned matrix. That is,

$$(\mathbf{P}'\hat{\mathbf{R}}^{-1}\mathbf{P})^{-1} = \begin{bmatrix}\mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{W} \\ \mathbf{W}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{W}'\hat{\mathbf{R}}^{-1}\mathbf{W}\end{bmatrix}^{-} = \begin{bmatrix}\mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22}\end{bmatrix}^{-} = \begin{bmatrix}\mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22}\end{bmatrix} \tag{A–8}$$

$$\text{It follows that} \quad \mathbf{M} = \mathbf{C}^{22} = (\mathbf{W}'\hat{\mathbf{R}}^{-1}\mathbf{W} - \mathbf{W}'\hat{\mathbf{R}}^{-1}\mathbf{X}(\mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{W})^{-1} \tag{A–9}$$

$A_{opt} = \text{argmin}\{trace(\mathbf{M})\}$ and $D_{opt} = \text{argmin}\{determinant(\mathbf{M})\}$ where $\mathbf{M}$ is as given in Equation A–9.

A.3 Randomized Complete Block Designs with Random Blocks and Fixed Treatments Effects

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{W}\gamma + \mathbf{Z}\mathbf{b} + \mathbf{e} = \begin{bmatrix}\mathbf{1} & \mathbf{W}\end{bmatrix} \begin{bmatrix}\mu \\ \gamma\end{bmatrix} + \mathbf{Z}\mathbf{b} + \mathbf{e} = \mathbf{X}\tau + \mathbf{Z}\mathbf{b} + \mathbf{e} \tag{A–10}$$

$$\text{where} \quad \mathbf{X} = \begin{bmatrix}\mathbf{1} & \mathbf{W}\end{bmatrix} \quad \text{and} \quad \tau = \begin{bmatrix}\mu \\ \gamma\end{bmatrix}, \quad \text{that is,} \quad \mathbf{y} = \mathbf{X}\tau + \mathbf{Z}\mathbf{b} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of observations; $\mathbf{1}$ is a column of ones; μ is an overall mean; $\mathbf{W}$ is an incidence matrix of fixed treatment effects; γ is a vector of fixed treatment effects; $\mathbf{Z}$ is an

incidence matrix of random block effects; $\mathbf{b}$ is a vector of random block effects such that $\mathbf{b} \sim N(0, \mathbf{D})$, where $\mathbf{D}$ is the variance-covariance matrix of block effects, say, for instance, $\mathbf{D} = \sigma_b^2 \mathbf{I}$; $\mathbf{e}$ is a vector of residual errors such that $\mathbf{e} \sim N(0, \mathbf{R})$, where $\mathbf{R} = \sigma_e^2 \theta$, θ is a spatial correlation matrix; $\mathbf{X}$ is a partitioned incidence matrix of fixed effects and τ is a partitioned vector of fixed effects. It is assumed that $\mathbf{b}$ and $\mathbf{e}$ are uncorrelated.

$$var(y) = \hat{V} = \mathbf{Z}\hat{D}\mathbf{Z}^{'} + \hat{R} \tag{A–11}$$

Solving the linear mixed model equations for best linear unbiased estimates (BLUEs) and best linear unbiased predictors (BLUPs) using Henderson (1950) procedure yields

$$\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{Z} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z} + \hat{\mathbf{D}}^{-1} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{y} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{y} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{y} \end{bmatrix} \tag{A–12}$$

$$= \begin{bmatrix} \mathbf{C}^{11} & \mathbf{C}^{12} \\ \mathbf{C}^{21} & \mathbf{C}^{22} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{y} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{y} \end{bmatrix} = \begin{bmatrix} (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \\ \hat{\mathbf{D}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}[\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}) \end{bmatrix}$$

$$= \begin{bmatrix} (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \\ \hat{\mathbf{D}}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}) \end{bmatrix}$$

The variance-covariance matrix to be minimized can be obtained using theorem 8.5.11 from Harville (1997), giving,

$$\mathbf{M} = var(\hat{\tau} - \tau) = (\mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{X} - \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{Z}[\mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z} + \hat{\mathbf{D}}^{-1}]^{-1}\mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{X})^{-1} \tag{A–13}$$

$A_{opt} = \text{argmin}\{trace(\mathbf{M})\}$ and $D_{opt} = \text{argmin}\{determinant(\mathbf{M})\}$ where $\mathbf{M}$ is as given in Equation A–13.

APPENDIX B
EXTRA TABLES AND GRAPHS

B.1 Overall Design Efficiency for Irregular-Grid $\Omega_A^{(30)}$ RCB Designs

Table B-1. Summary statistics for overall design efficiency (ODE) for irregular-grid $\Omega_A^{(30)}$ RCB designs, each condition replicated $\lambda = 10$ times for $p = 5,000$ iterations.

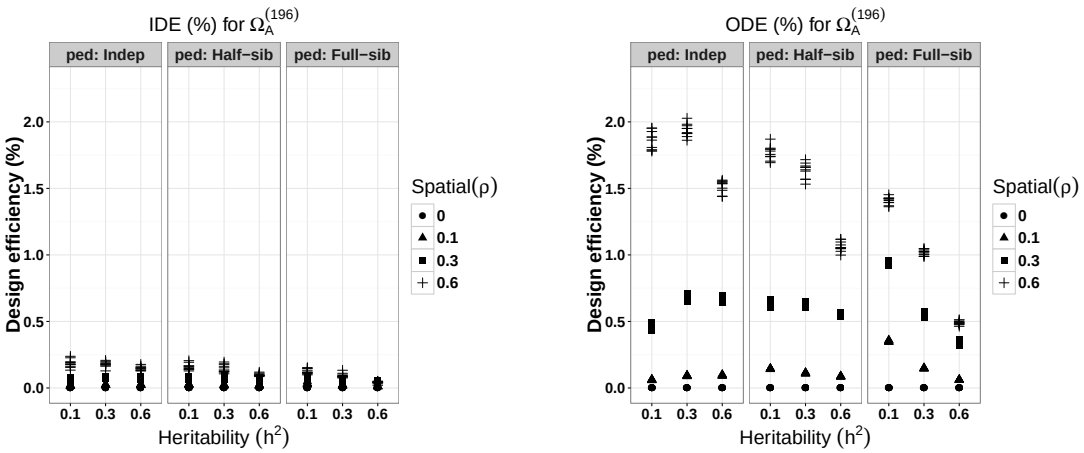
Pedigree	h^2	ρ	mean ODE %	S.E.
Indep	0.3	0.6	8.540	0.337
Half-sib	0.3	0.6	5.515	0.257
Full-sib	0.3	0.6	2.728	0.081

B.2 Initial and Overall Design Efficiency Table for $\Omega_A^{(196)}$ RCB Designs with 16 Blocks

Table B-2. Summary statistics for initial design efficiency (IDE) and overall design efficiency (ODE) for $\Omega_A^{(196)}$ RCB designs with 16 blocks of dimensions 14 rows by 14 columns, each condition replicated $\lambda = 10$ times for $p = 5,000$ iterations. ODE % mean values that are starred ($\star$) are the overall largest improvements per family.

	Condition		Indep		Half-sib		Full-sib	
Efficiency	h^2	ρ	mean (%)	S.E.	mean (%)	S.E.	mean (%)	S.E.
		0.1	0.017	0.001	0.025	0.001	0.032	0.002
	0.1	0.3	0.064	0.002	0.073	0.002	0.082	0.003
		0.6	0.184	0.010	0.160	0.008	0.124	0.006
IDE		0.1	0.023	0.001	0.021	0.001	0.020	0.001
	0.3	0.3	0.081	0.002	0.080	0.003	0.060	0.002
		0.6	0.177	0.007	0.148	0.010	0.093	0.005
		0.1	0.023	0.001	0.020	0.001	0.013	0.001
	0.6	0.3	0.073	0.002	0.067	0.003	0.040	0.002
		0.6	0.147	0.005	0.098	0.003	0.040	0.003
		0.1	0.061	0.001	0.144	0.001	0.351	0.002
	0.1	0.3	0.469	0.006	0.638	0.006	0.941	0.005
		0.6	1.863	0.022	1.768*	0.017	1.408*	0.009
ODE		0.1	0.092	0.001	0.111	0.001	0.148	0.001
	0.3	0.3	0.670	0.006	0.634	0.005	0.550	0.004
		0.6	1.938*	0.015	1.633	0.019	1.018	0.007
		0.1	0.095	0.001	0.087	0.001	0.062	0.001
	0.6	0.3	0.661	0.005	0.549	0.003	0.340	0.004
		0.6	1.516	0.015	1.064	0.012	0.489	0.005

B.3 Initial and Overall Design Efficiency Graphs for $\Omega_A^{(196)}$ RCB Designs with 16 Blocks



(a) (b)

Figure B-1. (a) Displays initial design efficiency (IDE) and (b) overall design efficiency (ODE) for $\Omega_A^{(196)}$ generated with 16 blocks of dimensions 14 rows by 14 columns. A total of 36 field conditions each with $\lambda = 10$ replicates are presented. Initial $m = 100$ designs were generated and the best one selected and iterated for $p = 5,000$ times.

B.4 Boxplots of Overall Design Efficiency for $\Omega_A^{(30)}$ RCB Designs for Each Algorithm

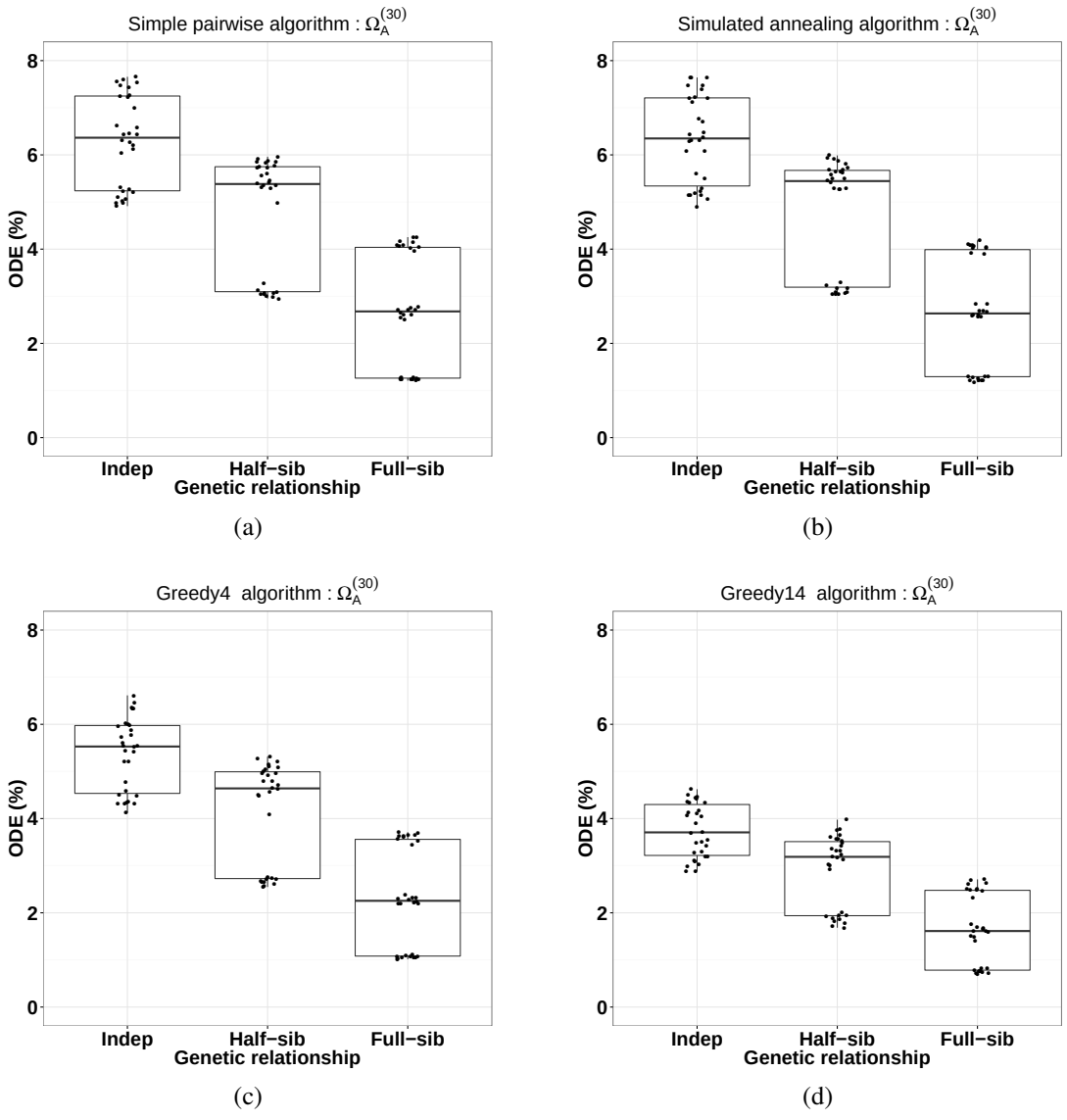
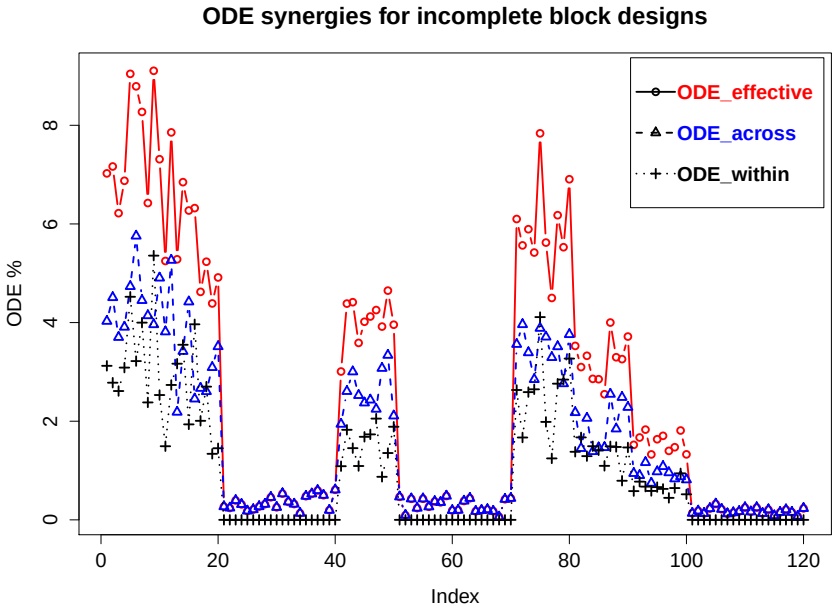
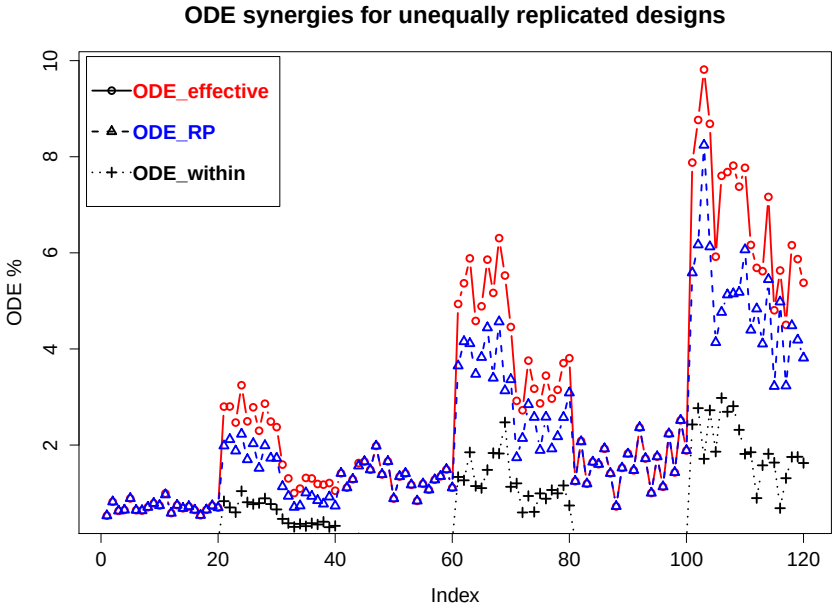


Figure B-2. Boxplots of overall design efficiency for experiments that were evaluated based on $\Omega_A^{(30)}$ scenario with 6 blocks of dimensions five rows by six columns. with $m = 100$ initial designs and $p = 5,000$ iterations.

B.5 Overall Design Efficiency Synergies for Non-Orthogonal Designs



(a)



(b)

Figure B-3. Overall design efficiency (ODE %) synergies for (a) incomplete block and (b) unequally replicated designs, based on $\Omega_A^{(30)}$ scenario for all families, evaluated at h^2 = 0.1, 0.3, and 0.6 and ρ = 0.0 and 0.6, with $m = 1$ initial design, $\lambda = 10$ and $p = 5,000$ iterations.

B.6 Pedigree Information for Full-Sib Families with 30 Offspring

Table B-3. Pedigree information for experiments that consisted of full-sib families with 30 offspring and 5 parents.

Id	sire	dam
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	1	2
7	1	2
8	1	2
9	1	3
10	1	3
11	1	3
12	1	4
13	1	4
14	1	4
15	1	5
16	1	5
17	1	5
18	2	3

Id	sire	dam
19	2	3
20	2	3
21	2	4
22	2	4
23	2	4
24	2	5
25	2	5
26	2	5
27	3	4
28	3	4
29	3	4
30	3	5
31	3	5
32	3	5
33	4	5
34	4	5
35	4	5

B.7 Pedigree Information for Half-Sib Families with 30 Offspring

Table B-4. Pedigree information for experiments that consisted of half-sib families with 30 offspring and 5 parents.

Id	sire	dam
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	1	0
7	1	0
8	1	0
9	1	0
10	1	0
11	1	0
12	2	0
13	2	0
14	2	0
15	2	0
16	2	0
17	2	0
18	3	0

Id	sire	dam
19	3	0
20	3	0
21	3	0
22	3	0
23	3	0
24	4	0
25	4	0
26	4	0
27	4	0
28	4	0
29	4	0
30	5	0
31	5	0
32	5	0
33	5	0
34	5	0
35	5	0

B.8 Pedigree Information for Full-Sib Families with 196 Offspring

Table B-5. Pedigree information for experimental designs that consisted of full-sib families with 196 offspring and 30 parents.

Id	sire	dam
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0
9	0	0
10	0	0
11	0	0
12	0	0
13	0	0
14	0	0
15	0	0
16	0	0
17	0	0
18	0	0
19	0	0
20	0	0
21	0	0
22	0	0
23	0	0
24	0	0
25	0	0
26	0	0
27	0	0
28	0	0
29	0	0
30	0	0
31	1	2
32	1	2
33	1	2
34	1	3
35	1	3
36	1	3
37	1	4
38	1	4
39	1	4
40	1	5
41	1	5
42	1	5
43	2	3
44	2	3
45	2	3
46	2	4
47	2	4
48	2	4
49	2	5
50	2	5
51	2	5
52	3	4
53	3	4
54	3	4
55	3	5
56	3	5
57	3	5
58	4	5
59	4	5
60	4	5
61	6	7
62	6	7
63	6	7
64	6	8
65	6	8
66	6	8
67	6	9
68	6	9
69	6	9
70	6	10
71	6	10
72	6	10
73	7	8
74	7	8
75	7	8
76	7	9
77	7	9
78	7	9
79	7	10
80	7	10
81	7	10
82	8	9
83	8	9
84	8	9
85	8	10
86	8	10
87	8	10
88	9	10
89	9	10
90	9	10
91	11	12
92	11	12
93	11	12
94	11	13
95	11	13
96	11	13
97	11	14
98	11	14
99	11	14
100	11	15
101	11	15
102	11	15
103	12	13
104	12	13
105	12	13
106	12	14
107	12	14
108	12	14
109	12	15
110	12	15
111	12	15
112	13	14
113	13	14
114	13	14
115	13	15
116	13	15
117	13	15
118	14	15
119	14	15
120	14	15
121	16	17
122	16	17
123	16	17
124	16	18
125	16	18
126	16	18
127	16	19
128	16	19
129	16	19
130	16	20
131	16	20
132	16	20
133	17	18
134	17	18
135	17	18
136	17	19
137	17	19
138	17	19
139	17	20
140	17	20
141	17	20
142	18	19
143	18	19
144	18	19
145	18	20
146	18	20
147	18	20
148	19	20
149	19	20
150	19	20
151	21	22
152	21	22
153	21	22
154	21	23
155	21	23
156	21	23
157	21	24
158	21	24
159	21	24
160	21	25
161	21	25
162	21	25
163	22	23
164	22	23
165	22	23
166	22	24
167	22	24
168	22	24
169	22	25
170	22	25
171	22	25
172	23	24
173	23	24
174	23	24
175	23	25
176	23	25
177	23	25
178	24	25
179	24	25
180	24	25
181	26	27
182	26	27
183	26	27
184	26	28
185	26	28
186	26	28
187	26	29
188	26	29
189	26	29
190	26	30
191	26	30
192	26	30
193	27	28
194	27	28
195	27	28
196	27	29
197	27	29
198	27	29
199	27	30
200	27	30
201	27	30
202	28	29
203	28	29
204	28	29
205	28	30
206	28	30
207	28	30
208	29	30
209	29	30
210	29	30
211	1	6
212	1	6
213	4	19
214	4	19
215	7	11
216	7	11
217	8	22
218	8	22
219	12	16
220	12	16
221	13	30
222	13	30
223	17	25
224	17	25
225	24	29
226	24	29

B.9 Pedigree Information for Half-Sib Families with 196 Offspring

Table B-6. Pedigree information for experiments that consisted of half-sib families with 196 offspring and 32 parents.

Id	sire	dam	Id	sire	dam	Id	sire	dam	Id	sire	dam	Id	sire	dam
1	0	0	47	3	0	93	11	0	139	18	0	185	26	0
2	0	0	48	3	0	94	11	0	140	18	0	186	26	0
3	0	0	49	3	0	95	11	0	141	19	0	187	26	0
4	0	0	50	3	0	96	11	0	142	19	0	188	26	0
5	0	0	51	4	0	97	11	0	143	19	0	189	27	0
6	0	0	52	4	0	98	11	0	144	19	0	190	27	0
7	0	0	53	4	0	99	12	0	145	19	0	191	27	0
8	0	0	54	4	0	100	12	0	146	19	0	192	27	0
9	0	0	55	4	0	101	12	0	147	20	0	193	27	0
10	0	0	56	4	0	102	12	0	148	20	0	194	27	0
11	0	0	57	5	0	103	12	0	149	20	0	195	28	0
12	0	0	58	5	0	104	12	0	150	20	0	196	28	0
13	0	0	59	5	0	105	13	0	151	20	0	197	28	0
14	0	0	60	5	0	106	13	0	152	20	0	198	28	0
15	0	0	61	5	0	107	13	0	153	21	0	199	28	0
16	0	0	62	5	0	108	13	0	154	21	0	200	28	0
17	0	0	63	6	0	109	13	0	155	21	0	201	29	0
18	0	0	64	6	0	110	13	0	156	21	0	202	29	0
19	0	0	65	6	0	111	14	0	157	21	0	203	29	0
20	0	0	66	6	0	112	14	0	158	21	0	204	29	0
21	0	0	67	6	0	113	14	0	159	22	0	205	29	0
22	0	0	68	6	0	114	14	0	160	22	0	206	29	0
23	0	0	69	7	0	115	14	0	161	22	0	207	29	0
24	0	0	70	7	0	116	14	0	162	22	0	208	30	0
25	0	0	71	7	0	117	15	0	163	22	0	209	30	0
26	0	0	72	7	0	118	15	0	164	22	0	210	30	0
27	0	0	73	7	0	119	15	0	165	23	0	211	30	0
28	0	0	74	7	0	120	15	0	166	23	0	212	30	0
29	0	0	75	8	0	121	15	0	167	23	0	213	30	0
30	0	0	76	8	0	122	15	0	168	23	0	214	30	0
31	0	0	77	8	0	123	16	0	169	23	0	215	31	0
32	0	0	78	8	0	124	16	0	170	23	0	216	31	0
33	1	0	79	8	0	125	16	0	171	24	0	217	31	0
34	1	0	80	8	0	126	16	0	172	24	0	218	31	0
35	1	0	81	9	0	127	16	0	173	24	0	219	31	0
36	1	0	82	9	0	128	16	0	174	24	0	220	31	0
37	1	0	83	9	0	129	17	0	175	24	0	221	31	0
38	1	0	84	9	0	130	17	0	176	24	0	222	32	0
39	2	0	85	9	0	131	17	0	177	25	0	223	32	0
40	2	0	86	9	0	132	17	0	178	25	0	224	32	0
41	2	0	87	10	0	133	17	0	179	25	0	225	32	0
42	2	0	88	10	0	134	17	0	180	25	0	226	32	0
43	2	0	89	10	0	135	18	0	181	25	0	227	32	0
44	2	0	90	10	0	136	18	0	182	25	0	228	32	0
45	3	0	91	10	0	137	18	0	183	26	0			
46	3	0	92	10	0	138	18	0	184	26	0			

APPENDIX C
R FUNCTIONS

C.1 Simple Pairwise Algorithm

```
Optimize.rcbd<- function(matdf,n,traceI,criteria,
Rinv,Ginv,K) {
        newmatdf <- matdf
        trace <- traceI
        mat <- NULL
        mat <- rbind(mat, c(value = trace, iterations = 0))
        Design_best <- newmatdf
        Des <- list()
        TRACE <- c()
        newmatdf <- SwapPair(matdf = matdf)
        for (i in 2:n) {
        newmatdf <- SwapPair(matdf = newmatdf)
        TRACE[i] <- NewValue.rcbd(matdf=newmatdf,
        criteria, Rinv, Ginv, K)
        Des[[i]] <- newmatdf
        if (NewValue.rcbd(matdf=newmatdf, criteria,
        Rinv, Ginv, K) < trace) {
                print(sprintf("Swapping within blocks: %d", i,
                "complete\n",sep = ""))
                Design_best <- Des[[i]] <- newmatdf
                Design_best <- newmatdf
                trace <- NewValue.rcbd(matdf=newmatdf,
        criteria, Rinv, Ginv, K)
                mat <- rbind(mat, c(trace = trace, iterations = i))
```

```
        }
        if (NewValue.rcbd(matdf=newmatdf, criteria, Rinv, Ginv,
        K) > trace & nrow(mat) <= 1) {
                newmatdf <- matdf
                Des[[i]] <- matdf
                Design_best <- matdf
        }
        if (NewValue.rcbd(matdf=newmatdf, criteria, Rinv, Ginv,
        K) > trace & nrow(mat) > 1) {
                newmatdf <- Des[[length(Des) - 1]]
                Des[[i]] <- newmatdf
                Design_best <- newmatdf
        }
        }
        ODE = (((mat[1,"value"]) - (mat[nrow(mat),"value"]))/
        (mat[1,"value"]))*100
        print(sprintf("ODE due to swapping pairs of treatments within
        blocks is: %f", ODE, "complete\n", sep = ""))
        list(TRACE = c(as.vector(mat[1, "value"]), TRACE[!is.na(TRACE)]),
                        mat = mat, Design_best = Design_best)
        }
```

C.2 Genetic Neighborhood Algorithm

```
Optimize_GNN_rcbd<- function(matdf, n, traceI, criteria, Amat,
Rinv, Ginv, K){
        newmatdf <- matdf
        trace <- traceI
        mat <- NULL
```

```
mat <- rbind(mat, c(value = trace, iterations = 0))
Design_best <- newmatdf
Des <- list()
TRACE <- c()
newmatdf <- Neighbor_rcbd(matdf,Amat)
for (i in 2:n) {
newmatdf <- Neighbor_rcbd(matdf,Amat)
TRACE[i] <- NewValue.rcbd(matdf=newmatdf, criteria,
Rinv, Ginv, K)
Des[[i]] <- newmatdf
if (NewValue.rcbd(matdf=newmatdf, criteria, Rinv,
Ginv, K) < trace) {
        print(sprintf("Swapping treatments: %d",
i, "complete\n",         sep = ""))
        Design_best <- Des[[i]] <- newmatdf
        Design_best <- newmatdf
        trace <- NewValue.rcbd(matdf=newmatdf,
        criteria, Rinv, Ginv, K)
        mat <- rbind(mat, c(trace = trace, iterations = i))
}
if(NewValue.rcbd(matdf=newmatdf, criteria, Rinv, Ginv,
        K) > trace & nrow(mat)<=1){
        newmatdf <- matdf
        Des[[i]] <- matdf
        Design_best <- matdf
}
if(NewValue.rcbd(matdf=newmatdf, criteria, Rinv, Ginv,
```

```
        K) > trace & nrow(mat)>1){
                newmatdf <- Des[[length(Des) - 1]]
                Des[[i]] <- newmatdf
                Design_best <- newmatdf
        }
        }
        ODE = (((mat[1,"value"]) -
                (mat[nrow(mat),"value"]))/(mat[1,"value"]))*100
        print(sprintf("ODE due to applying GNN procedure:
        %f", ODE, "complete\n", sep = ""))
        list(TRACE = c(as.vector(mat[1, "value"]), TRACE[!is.na(TRACE)]),
                        mat = mat, Design_best = Design_best)
}
```

C.3 Simulated Annealing Algorithm

```
Optimize_SimAnn_rcbd<- function(matdf,n,traceI,
criteria,Rinv,Ginv,K) {
newmatdf <- matdf
trace <- traceI
mat <- NULL
mat <- rbind(mat, c(value = trace, iterations = 0))
Design_best <- newmatdf
Des <- list()
TRACE <- c()
newmatdf <- SwapPair(matdf = matdf)
for (i in 2:n) {
        newmatdf <- SwapPair(matdf = newmatdf)
        TRACE[i] <- NewValue.rcbd(matdf=newmatdf,
```

```
        criteria, Rinv, Ginv, K)
        Des[[i]] <- newmatdf
        if (NewValue.rcbd(matdf=newmatdf, criteria, Rinv,
 Ginv, K) < trace) {
                print(sprintf("Swapping within blocks: %d",
                i, "complete\n",
                sep = ""))
                Design_best <- Des[[i]] <- newmatdf
                Design_best <- newmatdf
                trace <- NewValue.rcbd(matdf=newmatdf, criteria,
                                Rinv, Ginv, K)
                mat <- rbind(mat, c(trace = trace, iterations = i))
        }
Temp<-c()
if (NewValue.rcbd(matdf=newmatdf, criteria,
Rinv, Ginv, K) > trace)
{
dif <- setdiff(NewValue.rcbd(matdf=newmatdf,
criteria, Rinv, Ginv, K),trace)
Temp[i] <- 1/i
accept  = exp(-dif/Temp[i])
u = runif(1)
if (u < accept){
Design_best <- Des[[i]] <- newmatdf
Design_best <- newmatdf
trace <- NewValue.rcbd(matdf=newmatdf, criteria, Rinv, Ginv, K)
}
```

```
if (u > accept & nrow(mat) <= 1) {
        newmatdf <- matdf
        Des[[i]] <- matdf
        Design_best <- matdf
}
if (u > accept & nrow(mat) > 1) {
newmatdf <- Des[[length(Des) - 1]]
Des[[i]] <- newmatdf
Design_best <- newmatdf
}
}
}
ODE = (((mat[1,"value"]) - (mat[nrow(mat),"value"]))/
(mat[1,"value"]))*100
print(sprintf("ODE due to simulated annealing is:
%f", ODE, "complete\n",
        sep = ""))
list(TRACE = c(as.vector(mat[1, "value"]),TRACE[!is.na(TRACE)]),
        mat = mat, Design_best = Design_best)
}
```

C.4 Greedy Pairwise Algorithm

```
     OptimizeGreedy.rcbd<- function(matdf,n,traceI,criteria,
     gsize,Rinv,Ginv,K) {
     newmatdf <- matdf
     trace <- traceI
     mat <- NULL
     mat <- rbind(mat, c(value = trace, iterations = 0))
```

```
Design_best <- newmatdf
Des <- list()
TRACE <- c()
newmatdf <- SwapGreedy(matdf = matdf,gsize = gsize)
for (i in 2:n) {
        newmatdf <- SwapGreedy(matdf = newmatdf,gsize = gsize)
        TRACE[i] <- NewValue.rcbd(matdf=newmatdf, criteria,
        Rinv, Ginv, K)
        Des[[i]] <- newmatdf
        if (NewValue.rcbd(matdf=newmatdf, criteria, Rinv,
        Ginv, K) < trace) {
                print(sprintf("Swapping greedly within blocks:
         %d", i, "complete\n",        sep = ""))
                Design_best <- Des[[i]] <- newmatdf
                Design_best <- newmatdf
                trace <- NewValue.rcbd(matdf=newmatdf, criteria,
         Rinv, Ginv, K)
                mat <- rbind(mat, c(trace = trace, iterations = i))
        }
        if (NewValue.rcbd(matdf=newmatdf, criteria, Rinv, Ginv,
         K) > trace & nrow(mat) <= 1) {
                newmatdf <- matdf
                Des[[i]] <- matdf
                Design_best <- matdf
        }
        if (NewValue.rcbd(matdf=newmatdf, criteria, Rinv, Ginv,
         K) > trace & nrow(mat) > 1) {
```

```
                        newmatdf <- Des[[length(Des) - 1]]
                        Des[[i]] <- newmatdf
                        Design_best <- newmatdf
        }
}
ODE = (((mat[1,"value"]) -
         (mat[nrow(mat),"value"]))/(mat[1,"value"]))*100
print(sprintf("ODE due to greedly swapping pairs of treatments
 within blocks is: %f", ODE, "complete\n", sep = ""))
list(TRACE = c(as.vector(mat[1, "value"]), TRACE[!is.na(TRACE)]),
 mat = mat, Design_best = Design_best)
}
```

C.5 Generate Matrices for RCB Designs

```
VarCov.rcbd <- function(matdf, rhox, rhoy, h2, s20, Tr, Tc,
criteria="A", Amat=FALSE,irregular=FALSE) {
if(nrow(matdf)==length(unique(matdf[,"Genotypes"]))){
        X <- as.matrix(matdf[, "Reps"])
}
if(nrow(matdf) > length(unique(matdf[,"Genotypes"]))){
        X <- Matrix::sparse.model.matrix(~as.factor(matdf[,
        "Reps"])-1)
}
s2e <- (1 - s20) * (1 - h2)
stopifnot(s2e > 0)
m = length(unique(matdf[,"Genotypes"]))
if(is.matrix(Amat)){
G <- h2 * as.matrix(Amat)
```

```
Ginv <- round(chol2inv(chol(as.matrix(G))),7)
Ginv <- as(Ginv, "sparseMatrix")
}
else{
        Ginv <- round((1/h2) * Matrix::Diagonal(m),7)
        Ginv <- as(Ginv, "sparseMatrix")
}
Z<- Matrix::sparse.model.matrix(~as.factor(matdf[,
"Genotypes"]) - 1)
# calculating R and its inverse for spatial analysis
bb <- length(unique(matdf[,"Reps"]))
matdf <- matdf[order(matdf[,"Row"],matdf[,"Col"]),]
if(irregular==TRUE){
        R <- Matrix::Diagonal(nrow(matdf))
        for(i in 1:(nrow(matdf)-1)) {
        x1  <- matdf[,"Col"][i]
        y1  <- matdf[,"Row"][i]
                for (j in (i+1):nrow(matdf)){
                x2 <- matdf[,"Col"][j]
                y2  <- matdf[,"Row"][j]
                R[i,j]<-(rhox^abs(x2 -x1))*(rhoy^abs(y2 -y1))
        }
}
R = as.matrix(round(s2e*R,7))
R[lower.tri(R)] <- t(R)[lower.tri(R)]
R <- as(R, "sparseMatrix")
Rinv <- round(chol2inv(chol(R)),7)
```

```
Rinv <- as(Rinv, "sparseMatrix")
}
if(irregular==FALSE){
        sigx <- Matrix::Diagonal(Tc)
        sigx <- rhox^abs(row(sigx) - col(sigx))
        sigy <- Matrix::Diagonal(Tr)
        sigy <- rhoy^abs(row(sigy) - col(sigy))
        R <- round(s2e * kronecker(sigy, sigx),7)
        R <- as(R, "sparseMatrix")
        Rinv <- round(chol2inv(chol(R)),7)
        Rinv <- as(Rinv, "sparseMatrix")
}
C11 <- Matrix::crossprod(as.matrix(X),
as.matrix(Rinv)) %*% as.matrix(X)
C11inv <- solve(C11)
k1 <- Rinv %*% as.matrix(X)
k2 <- Matrix::tcrossprod(as.matrix(C11inv), as.matrix(X))
k3 <- k2 %*% Rinv
K <- k1 %*% k3
K <- as(K, "sparseMatrix")
temp0 <- Matrix::crossprod(Z, Rinv) %*% Z + Ginv -
 Matrix::crossprod(Z, K) %*% Z
C22 <- solve(temp0)
C22  <- as(C22, "sparseMatrix")
Ginv = round(Ginv,7)
Rinv = Matrix::drop0(round(Rinv,7))
K = round(K,7)
```

```
C22 = round(C22,7)
if (criteria == "A") {
        return(c(traceI = sum(Matrix::diag(C22)), Ginv = Ginv,
                Rinv = Rinv,   K=K))
}
if (criteria == "D") {
        deTm = Matrix::det(C22)
        return(c(doptimI = log(deTm), Ginv = Ginv, Rinv = Rinv, K=K))
}
}
```

C.6 Generate Matrices for Unequally Replicated Designs

```
unequal.VarCov <- function(matdf, rhox, rhoy, h2,s20, Tr, Tc,
criteria = "A", Amat = FALSE, sigBl = FALSE, irregular = FALSE)
{
if(irregular ==FALSE & Tr*Tc != nrow(matdf))
 stop("check Tr by Tc dimensions")
X <- matrix(1,nrow = nrow(matdf))
# determine number of blocks
bb <- length(unique(matdf[,"Reps"]))
if(is.numeric(sigBl))
{
        Binv <- (1/sigBl)*Matrix::Diagonal(bb)
}else{
        sigBl <- 0.2*(1 - h2)
        Binv <- (1/sigBl)*Matrix::Diagonal(bb)
        Binv <- as(Binv, "sparseMatrix")
}
```

```
s2e <- (1 - s20) * (1 - h2 - sigBl)
stopifnot(s2e > 0)
m = length(unique(matdf[,"Genotypes"]))
if(is.matrix(Amat)) {
        Gg <- h2 * as.matrix(Amat)
        Gg <- round(solve(Gg),7)
        Gg <- as(Gg, "sparseMatrix")
}else{
        Gg <- (1/h2) * Matrix::Diagonal(m)
        Gg <- as(Gg, "sparseMatrix")
}
Ginv <- Matrix::bdiag(Binv,Gg)
Zg <- Matrix::sparse.model.matrix(~as.factor(matdf[,
"Genotypes"]) - 1)
Zb <- Matrix::sparse.model.matrix(~as.factor(matdf[,
"Reps"]) - 1)
Z <- Matrix::cBind(Zb,Zg)
if(irregular==TRUE){
        R <- Matrix::Diagonal(nrow(matdf))
        for(i in 1:(nrow(matdf)-1)) {
                x1  <- matdf[,"Col"][i]
                y1  <- matdf[,"Row"][i]
                for (j in (i+1):nrow(matdf)){
                x2 <- matdf[,"Col"][j]
                y2  <- matdf[,"Row"][j]
                R[i,j]<-(rhox^abs(x2 -x1))*(rhoy^abs(y2 -y1))
        }
```

```
}
R = as.matrix(round(s2e*R,7))
R[lower.tri(R)] <- t(R)[lower.tri(R)]
R <- as(R, "sparseMatrix")
Rinv <- round(chol2inv(chol(R)),7)
Rinv <- as(Rinv, "sparseMatrix")
}
if(irregular==FALSE){
        sigx <- Matrix::Diagonal(Tc)
        sigx <- rhox^abs(row(sigx) - col(sigx))
        sigy <- Matrix::Diagonal(Tr)
        sigy <- rhoy^abs(row(sigy) - col(sigy))
        R <- round(s2e * kronecker(sigy, sigx),7)
        R <- as(R, "sparseMatrix")
        Rinv <- round(chol2inv(chol(R)),7)
        Rinv <- as(Rinv, "sparseMatrix")
}
C11 <- t(X) %*% Rinv %*% X
C11inv <- 1/C11
K <- round(Rinv %*% X %*% C11inv %*% t(X) %*% Rinv ,7)
K <- as(K, "sparseMatrix")
Z <- as.matrix(Z)
temp0 <- t(Z) %*% Rinv %*% Z + Ginv - t(Z) %*% K %*% Z
C22 <- solve(temp0)
C22 <- round(C22[-(1:bb), -(1:bb)],7)
C22 <- as(C22, "sparseMatrix")
if (criteria == "A") {
```

```
        return(c(traceI = sum(Matrix::diag(C22)), Ginv = Ginv,
                Rinv = Rinv, K = K))
}
if (criteria == "D") {
        deTm = Matrix::det(C22)
        return(c(doptimI = log(deTm), Ginv = Ginv, Rinv = Rinv, K=K))
}
}
```

C.7 Generate Matrices for Augmented Designs

```
unequal.Augmented.VarCov <- function(matdf, rhox, rhoy, h2, s20,
        rb, cb, criteria = "A", Amat = FALSE, sigBl = FALSE){
X <- matrix(1,nrow = nrow(matdf))
X<-Matrix(X)
bb <- length(unique(matdf[,"Reps"]))
if(is.numeric(sigBl))
{
        Binv <- (1/sigBl)*Matrix::Diagonal(bb)
        s2e <- (1 - s20) * (1 - h2 - sigBl)
}else{
        sigBl <- 0.2*(1 - h2)
        s2e <- (1 - s20) * (1 - h2 - sigBl)
        Binv <- (1/sigBl)*Matrix::Diagonal(bb)
}
stopifnot(s2e > 0)
m = length(unique(matdf[,"Genotypes"]))
if(is.matrix(Amat)){
        Gg <- h2 * as.matrix(Amat)
```

```
        Gg <- Matrix::drop0(Gg)
        Gg <- round(chol2inv(chol(Gg)),7)
}else{
        Gg <- (1/h2) * Matrix::Diagonal(m)
        Gg <- as(Gg, "sparseMatrix")
}
Ginv <- Matrix::bdiag(Binv,Gg)
Zg <- Matrix::sparse.model.matrix(~as.factor(matdf[,
"Genotypes"]) - 1)
Zb <- Matrix::sparse.model.matrix(~as.factor(matdf[,
"Reps"]) - 1)
Z <- Matrix::cBind(Zb,Zg)
if(rhox==0 & rhoy==0){
h = nrow(matdf)/bb
rinv = (1/s2e) * Matrix::Diagonal(h)
Rinv <- do.call(bdiag, replicate(bb, rinv, simplify=FALSE))
}
else{
        matX <- subset(matdf,matdf[,"Reps"]==1)
        sigx <- Matrix::Diagonal(cb)
        sigx <- rhox^abs(row(sigx) - col(sigx))
        sigy <- Matrix::Diagonal(rb)
        sigy <- rhoy^abs(row(sigy) - col(sigy))
        R <- round(s2e * kronecker(sigy, sigx),7)
        R <- as(R, "sparseMatrix")
        Rinv <- round(chol2inv(chol(R)),7)
        Rinv <- as(Rinv, "sparseMatrix")
```

```
        Rinv <- do.call(Matrix::bdiag, replicate(bb, Rinv,
        simplify = FALSE))
}
X <- as.matrix(X)
C11 <- t(X) %*% Rinv %*% X
C11inv <- 1/C11
K <- round(Rinv %*% X %*% C11inv %*% t(X) %*% Rinv ,7)
K <- as(K, "sparseMatrix")
Z <- as.matrix(Z)
temp0 <- t(Z) %*% Rinv %*% Z + Ginv - t(Z) %*% K %*% Z
C22 <- solve(temp0)
C22 <- round(C22[-(1:bb), -(1:bb)],7)
C22 <- as(C22, "sparseMatrix")
if (criteria == "A") {
return(c(traceI = sum(Matrix::diag(C22)), Ginv = Ginv,
         Rinv = Rinv, K = K))
}
if (criteria == "D") {
deTm = Matrix::det(C22)
return(c(doptimI = log(deTm), Ginv = Ginv, Rinv = Rinv, K=K))
}
}
```

REFERENCES

Beaulieu J, Mackay J, Rainville A, Bousquet J (2014). "Genomic selection accuracies within and between environments and small breeding groups in white spruce." *BMC Genomics*, **15**, 1048.

Borges P, Eid T, Bergseng E (2014). "Applying simulated annealing using different methods for the neighborhood search in forest planning problems." *European Journal of Operational Research*, **233**, 700–710.

Brown H, Prescott R (2015). *Applied Mixed Models in Medicine*. Third edition. John Wiley and Sons Ltd, UK.

Butler DG, Eccleston JA, Cullis BR (2008). "On an approximate optimality criterion for the design of field experiments under spatial dependence." *Australian and New Zealand Journal of Statistics*, **50**, 295–307.

Cheng CS (1983). "Construction of optimal balanced incomplete block designs for correlated observations." *The Annals of Statistics*, **11**, 240–246.

Chernoff H (1953). "Locally optimal designs for estimating parameters." *The Annals of Mathematical Statistics*, **24**, 586–602.

Cressie NAC (1993). *Statistics for Spatial Data*. Revised edition. John Wiley & Sons, Inc, New York, USA.

Cullis BR, Lill W, Fisher J, Read B, Gleeson A (1989). "A new procedure for the analysis of early generation variety trials." *Journal of Applied Statistics*, **38**, 361–375.

Cullis BR, Smith AB, Coombes NE (2006). "On the design of early generation variety trials with correlated data." *Journal of Agricultural, Biological and Environmental Statistics*, **11**, 381–393.

Das A (2002). "An introduction to optimality criteria and some results on optimal block design." In *Design Workshop Lecture Notes*, pp. 1–21. Indian Statistical Institute, Kolkata, Theoretical Statistics and Mathematics Unit, New Delhi, India.

Falconer DS, Mackay TFC (1996). *Introduction to Quantitative Genetics*. Fourth edition. Longman Group Ltd, England, Pearson Prentice Hall.

Federer WT (1956). "Augmented (or hoonuiaku) designs." *Hawaiian Planters' Record*, **55**, 191–208.

Federer WT (1998). "Recovery of interblock, intergradient, and intervarietal information in incomplete block and lattice rectangle designed experiments." *Biometrics*, **54**, 471–481.

Federer WT, Raghavarao D (1975). "On augmented designs." *Biometrics*, **31**(1), 29–35.

Filho JSB, Gilmour SG (2003). "Planning incomplete block experiments when treatments are genetically related." *Biometrics*, **59**, 375–381.

Gezan SA, White TL, Huber DA (2010). "Accounting for spatial variability in breeding trials: A simulation study." *Agronomy*, **102**, 1562–1571.

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009). *ASReml User Guide Release 3.0*. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK. URL www.vsni.co.uk.

Gonçalves E, St Aubyn A, Martins A (2007). "Mixed spatial models for data analysis of yield on large grapevine selection field trials." *Theoretical and Applied Genetics*, **115**(5), 653–663.

Habier D, Fernando RL, Dekkers JCM (2007). "The impact of genetic relationship information on genome assisted breeding values." *Genetics*, **177**, 2389–2397.

Harville DA (1997). *Matrix Algebra from a Statistician's Perspective*. New York, Inc, Springer-Verlag, USA.

Henderson CR (1950). "The estimation of genetic parameters." *The Annals of Mathematical Statistics*, **21**, 309–310.

Henderson CR (1975). "Use of all relatives in intraherd prediction of breeding values and producing abilities." *Dairy Science*, **58**, 1910–1916.

Henderson CR (1984). *Applications of Linear Models in Animal Breeding*. University of Guelph, Ontario, Canada.

Hill WG, Goddard ME, Visscher PM (2008). "Data and theory point to mainly additive genetic variance for complex traits." *PLoS Genetics*, **4**(2).

Hooks T, Marx D, Kachman S, Pedersen J (2009). "Optimality criteria for models with random effects." *Revista Colombiana de Estadística*, **32**, 17–31.

Imai K (2013). *experiment: R package for designing and analyzing randomized experiments*. R package version 1.1-1.

John JA, Williams ER (1995). *Cyclic and Computer Generated Designs*. Second edition. Chapman and Hall, UK.

Kiefer J (1959). "Optimum experimental design." *Journal of the Royal Statistical Society*, **21**, 272–319.

Kiefer J, Wolfowitz J (1959). "Optimum designs in regression problems." *Annals of Mathematical Statistics*, **30**, 271–294.

Kirkpatrick S, Gelatt CD, Vecchi MP (1983). "Optimization by simulated annealing." *Science*, **220**, 671–680.

Kuehl RO (2000). *Design of Experiments: Statistical Principles of Research Design and Analysis*. Second edition. Brooks/Cole, Cengage Learning, CA, USA.

Kuhfeld WF (2010). "Experimental design: Effiency, Coding, and Choice Designs." *Technical report*, SAS®.

Li Y, Swertz M, Vera G, Breitling R, Jansen R (2013). *designGG: Computational tool for designing genetical genomics experiments.* R package version 1.1.

Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Schabenberger O (2006). *SAS® for Mixed Models*. Second edition. Cary, NC, USA.

Liu G, Han S, Zhao X, Nelson JD, Wang H, Wang W (2006). "Optimisation algorithms for spatially constrained forest planning." *Ecological Modelling*, **194**(4), 421 – 428.

Mandal S (2000). *Construction of optimizing distributions with applications in estimation and optimal design*. Ph.d. dissertation, University of Glasgow, UK.

Mathew B, Holand A, Koistinen P, Léon J, Sillanpää M (2015). "Reparametrization-based estimation of genetic parameters in multi-trait animal model using integrated nested laplace approximation." *Theoretical and Applied Genetics*, pp. 1–11.

Mendiburu F (2015). *agricolae: Statistical procedures for agricultural research*. R package version 1.2-3.

Moehring J, Williams ER, Piepho HP (2014). "Efficiency of augmented p-rep designs in multi-environmental trials." *Theoretical and Applied Genetics*, **127**(5), 1049–1060.

Möhring J (2010). *Mixed modelling for phenotypic data from plant breeding*. Ph.D. thesis, Institut für Kulturpflanzenwissenschaften, Universität Hohenheim.

Mrode RA (2014). *Linear Models For the Prediction of Animal Breeding Values*. Third edition. CABI Publishing, USA.

Patterson HD, Hunter EA (1983). "The efficiency of incomplete block designs in national list and recommended list cereal trials." *Agricultural Science*, **101**, 427–433.

Patterson HD, Thompson R (1971). "Recovery of inter-block information when block sizes are unequal." *Biometrika*, **58**(3), 545–554.

Piepho HP, Möhring J, Melchinger AE, Büchse A (2008). "BLUP for phenotypic selection in plant breeding and variety testing." *Euphytica*, **161**, 209–228.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Resende M, Stringer J, Cullis B, Thompson R (2005). "Joint modelling of competition and spatial variability in forest field trials." *Revista Matemática e Estatica*, **23**, 7–22.

Robert CP, Casella G (2010). *Introducing Monte Carlo Methods with R*. Use R! Springer, New York, London, USA.

Sailer MO (2013). *crossdes: Construction of crossover designs*. R package version 1.1-1.

Sarker A, Singh M (2015). "Improving breeding efficiency through application of appropriate experimental designs and analysis models: A Case of Lentil(Lens culinaris Medikus subsp. culinaris) yield trials." *Field Crops Research*, **179**, 26–34.

Seo JH, Vilčko F, Orois SS, Kunth S, Son YM, Gadow KV (2005). "A case study of forest management planning using a new heuristic algorithm." *Tree Physiology*, **25**, 929–938.

Simecek P, Pilz J, Wang M, Gebhardt A (2014). *OPDOE: Optimal design of experiments*. R package version 1.0-9.

Snow G (2013). *blockrand: Randomization for block random clinical trials*. R package version 1.3.

Stringer JK, Cullis BR (2002). "Application of spatial analysis techniques to adjust for fertility trends and identify interplot competition in early sugarcane selection trials." *Australian Journal of Agricultural Research*, **53**, 911–918.

Stroup WW (2013). *Generalized Linear Mixed Models, Modern Concepts, Methods and Applications*. Chapman & Hall, New York, USA.

VanRaden PM (2008). "Efficient methods to compute genomic predictions." *Journal of Dairy Science*, **91**(11), 4414–4423.

Wald A (1943). "On the efficient design of statistical investigations." *The Annals of Mathematical Statistics*, **14**, 134–140.

Welham SJ, Gezan SA, Clark SJ, Mead A (2015). *Statistical Methods in Biology*. Design and Analysis of Experiments and Regression. Chapman & Hall, Boca Raton, USA.

Wheeler B (2014). *AlgDesign: Algorithmic experimental design*. R package version 1.1-7.3.

William E, Piepho HP, Whitaker D (2011). "Augmented p-rep designs." *Biometrical Journal*, **53**(1), 19–27.

Williams ER, John JA, Whitaker D (2006). "Construction of Resolvable Spatial Row-Column Designs." *Biometrics*, **62**, 103 – 108.

Wolak ME (2012). "nadiv: an R package to create relatedness matrices for estimating non-additive genetic variances in animal models." *Methods in Ecology and Evolution*, **3**(5), 792–796.

Yang M (2008). "A-optimal designs for generalized linear models with two parameters." *Journal of Statistical Planning and Inference*, **138**, 624–641.

Yates F (1939). "The recovery of inter-block information in varietal trials arranged in three dimensional lattices." *The Annals of Eugenics*, **9**(2), 136–156.

Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer, New York, USA.

BIOGRAPHICAL SKETCH

Lazarus Mramba was born in Kizurini, Kaloleni division, Kilifi county, Kenya. He attended Kizurini primary school and sat for the Kenya Certificate of Primary Education before joining Katana Ngala secondary school at Matsangoni where he attended form one and form two. He later joined St. Georges High school in Kaloleni, Giryiama where he was enrolled in form three and form four and sat for the Kenya Certificate of Secondary Education in November 1995. In April 1997, he joined Jomo Kenyatta University of Agriculture and Technology (JKUAT) for an undergraduate degree and graduated four years, later with a Bachelor of Science degree (BSc) majoring in physics with all core mathematics courses. Lazarus was in October 2001 employed by Oshwal academy in Mombasa county as a Physics assistant teacher where he worked until April 2006 before joining Kenya Medical Research Institute (KEMRI)-Wellcome Trust Research Programme as an intern statistician, then a junior statistician and later as a research statistician. He was admitted to the University of Leeds in the United Kingdom for a Master of Statistics degree (MSc. Statistics) in 2009-2010 before returning to work for the KEMRI-Wellcome Trust Research Programme. Later in January 2012, Lazarus was awarded a scholarship by the University of Florida, in Gainesville, United States of America, where he pursued a Ph.D. program with a focus in quantitative genetics and graduate statistical courses for a period of about 4 years and graduated in May, 2016.