Data-driven optimization decision support for plant breeding

by

Samira Karimzadeh

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Industrial Engineering

Program of Study Committee:
Sigurdur Olafsson, Major Professor
Stephen B Vardeman
Heike Hofmann
Guiping Hu
Qing Li

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation/thesis. The Graduate College will ensure this dissertation/thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2021

DEDICATION

To my beloved family and my best friend Athena.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this dissertation. First and foremost, Dr. Sigurdur Olafsson for his guidance, patience and support throughout this research and the writing of this dissertation. His insights and words of encouragement have often inspired me and renewed my hopes for completing my graduate education. I would also like to thank my committee members for their efforts and contributions to this work: Dr. Stephen B. Vardeman, Dr Heike Hofmann, Dr. Guping Hu and Dr. Qing Li. I would additionally like to thank Syngenta Seeds Inc for granting this research, and many insightful discussions that informed our understanding of the problem. I also want to thank our Iowa State University colleagues who have worked with us on this project and provided thoughtful feedback, especially Reyhaneh Bijari and Hanisha Vemireddy.

ABSTRACT

A commercial plant breeding program is a complex operation. A typical operation starts with several tens of thousands of experimental genotypes (e.g., soybean varieties or corn hybrids) that are planted in what is termed an early-stage experiment. At the end of the growing season, the most promising genotypes are selected for advancement. That is, they will be planted the next year again with the expectation that the best will eventually become commercialized. Thus, advanced genotypes are planted over multiple years, starting with early-stage experiments of tens of thousands of genotypes planted in a few locations, extending to late-stage experiments where a few of the best performers are planted in a larger number of locations.

Decisions regarding the advancement of a specific plant genotype is challenging due to a very limited number of observations. But identifying subsets of varieties that are performing similarly to the current commercial seeds in variety of the environments can help to identify potential commercial plant seeds. To aid the decision making, machine learning can be used to predict the yield of a genotype based on past observations of multiple genotypes. In this dissertation we propose predictive models using observation of commercialized varieties that are able to identify varieties those perform similarly to the different environments in laboratory stage when there is no field observation available.

Also, we must deal with a complex assignment problem. Practically speaking, experimental genotypes belong to groups based on their similarities, such as relative maturity (RM), and stage of the experiment. Trials that are similar in terms of stage and RM should also be placed together as much as possible to keep the environmental condition consistent to make the comparison fair among those. Furthermore, within a plant breeding program, there are typically multiple breeding groups, each responsible for making decisions regarding a portfolio of experimental genotypes would also like their trials to be positioned close to each other for convenience. We,

therefore, develop an optimization model that focuses on desirable properties, that is, placing trials from the same breeding groups and with the same RM and stage together while indirectly reducing wasted space in the field. The model is scalable and provides advantage of reaching the optimal solution over the entire breeding program in a fully automated two-phase process.

CHAPTER 1. INTRODUCTION

1.1 Overview

A commercial plant breeding program is a complex operation. A typical operation starts with several tens of thousands of experimental genotypes (e.g., soybean varieties or corn hybrids) that are planted in what is termed an early-stage experiment. At the end of the growing season, the most promising genotypes are selected for advancement. That is, they will be planted the next year again with the expectation that the best will eventually become commercialized. Thus, advanced genotypes are planted over multiple years, starting with early-stage experiments of tens of thousands of genotypes planted in a few locations, extending to late-stage experiments where a few of the best performers are planted in a larger number of locations.

To have a reliable estimation on the performance of a plant seed for decision making, breeders need to observe the seed varieties across the years and locations (environments) which is not possible in early stages due to having several thousands of varieties. But identifying subsets of varieties that are performing similarly to the current commercial seeds in variety of the environments can help to identify potential commercial plant seeds. In practice this is it is only economically feasible to plant each seed variety in a limited number of locations are selected by plant breeders which is a challenging assignment problem itself for various reasons. In fact in a commercial plant breeding trials from different breeding groups those are responsible for making decisions regarding a portfolio of experimental genotypes, in different stages which is the year in the breeding program the genotype is planting and with varying relative maturity (RM) that is, the number of growing days needed for the plant to mature must all be planted in the same field. Meeting all these requirement at the same time is not possible due to limited capacity of the fields. In the first part of this dissertation we propose predictive models using observation of commercialized varieties that are able to identify varieties those perform similarly to the different

environments in laboratory stage when there is no field observation available. Then in the second part of the dissertation we provide an optimization model to optimize the trials field assignment process.

1.2 Research Questions

We assume that observations of genotype varieties those have similar genetic-by-environment (GxE) interaction effect can be used to predict phenotypic performance of a target variety when the target observations is not sufficient to make advancement decision. Thus in Chapter 2 of this dissertation we are trying to answer a practical question: *How can we find clusters of genotypes with similar GxE effects using phenotype data?* To define subsets of genotype varieties those have similar GxE interaction effect, data clustering can be used which requires a complete matrix of GxE similarity/dissimilarity of each and every pair of varieties. It is not possible to define this GxE similarity for all variety pairs, as it requires the varieties having been planted in the same environments. Therefore, the proximity matrix obtained by phenotype data has huge percentage of missing observations and no standard clustering method is directly applicable. Thus we have to address a technical difficulty of *How can we do data clustering when we only have proximity matrix with mostly missing values?* A proximity matrix completion method is developed to solve this issue.

Although PMC algorithm proposed in Chapter 2 can help with identifying subsets of GxE similar varieties, it requires variety pairs to be observed in same environment repeatedly which only happen in late stages where breeders have enough information about varieties. In other words, breeders are interested to identify subsets of GxE similar varieties for early stage varieties that number of phenotypic observations is very limited to make advancement decision. Thus, breeders are interested to answer this question: *How to predict the GxE similarity of genotypes using genetic information?* to make sure are they able to make a reliable decision regarding advancement of a trial genotype in advance without observing its phenotypic performance across the various environments using observations of similar GxE varieties in the past. We have

developed a framework in Chapter 3 of this dissertation which defines an empirical measure of GxE dissimilarity using phenotypic observation so that machine learning can be used to predict GxE similarity using genetic variables.

Finally from operational point of view, breeders are interested to know: *How should field trials be planted to aid advancement decisions?* This can be achieved knowing how to model field trials mathematically in a way that accounts for all of the plant breeder concerns, while still having the ability to solve the problem under tight time constraints. Chapter 4-5 of this dissertations provide optimization modelling approach to aim this goal.

Therefore this dissertation facilitates the breeding process in two main approaches of *Predictive models to aid advancement decisions* and *Optimization of field trials*.

1.3 Predictive models to aid advancement decisions

Decisions regarding the advancement of a specific plant genotype is challenging due to a very limited number of observations. To aid the decision making, machine learning can be used to predict the yield of a genotype based on past observations of multiple genotypes, but this approach is challenged by the fact that yield is a function of genotype, environment, and the genotype-by-environment (GxE) interaction effects. As a result, the sufficiency of past observations in target environments is crucial in building the prediction model. When the observations of the target genotype are not enough to train an accurate model, using observation of related genotypes those have the same GxE behavior is a possible solution, but identifying those related genotypes is challenging as well.

A data clustering approach could be applied to identify such sets of related genotypes that require a proximity matrix of GxE similarity of all genotype pairs. Furthermore, it is not practically possible to define this GxE similarity for all genotype pairs, as it requires the genotypes having been planted in the same environments repeatedly. This will not happen for most pairs for various reasons. For some pairs, their environmental requirements are simply too dissimilar, which makes them unlikely to be planted in the same location. For some others, it is

only economically feasible to plant each genotype in a limited number of locations. Thus, it naturally gives rise to a proximity matrix where most of the values are missing for either of these two fundamentally different reasons. Therefore, if the proximity matrix has missing values, no standard clustering method is directly applicable. Imputation can be done to replace missing values, but none of the current imputation methods such as imputing the mean or median of the observations could address the reason for values are missing. As a solution, we propose the Proximity Matrix Completion (PMC) algorithm in Chapter 1 that can impute missing values that addresses reasons values are missing. To determine which case applies, the data is modeled as a graph, and a set of maximum cliques in the graph is found. The overlap between cliques then determines the case and hence the method of imputation for each missing data point.

Besides the clustering, we are able to determine pairs of genotypes that have the same preference to the environment (low dissimilarity in GxE effect) using historical data where we have a wide observation of genotypes in common environments. However, that only happens in the late stages of a breeding program when the breeders have enough information to make a decision. Since the GxE effect is a response of genetic interaction to the environment, the similarity in genetic markers in a pair of genotypes can be used to predict similarity/dissimilarity in the GxE effect. In Chapter 2 we deploy a supervised learning model on commonality in genetic markers can predict a pair of varieties related or non-related as a binary response of a classification problem. Using related genotypes versus non-related genotypes affects yield prediction accuracy and benefits of the advancement decision-making process. This early prediction of related genotypes in the early stages can save time and money in different ways. Having advanced knowledge of related genotypes, breeders can plan to plant experiments in a way that maximizes information gain in the early stage. Knowing that related genotypes have the same GxE preference to environment, breeders can avoid planting related genotypes in the same environments and expand the experiment to more environments through related genotypes. We also validate the capability of the proposed GxE similarity measure in estimating phenotypic

stability by comparing the results with some of the existing stability/adaptability models and discuss its contribution to the field.

1.4 Optimization of field trials

Having all information about related genotypes and target environments, we must deal with a complex assignment problem. Practically speaking, experimental genotypes belong to groups based on their similarities, such as relative maturity (RM), and stage of the experiment. Trials that are similar in terms of stage and RM should also be placed together as much as possible to keep the environmental condition consistent to make the comparison fair among those. Furthermore, within a plant breeding program, there are typically multiple breeding groups, each responsible for making decisions regarding a portfolio of experimental genotypes would also like their trials to be positioned close to each other for convenience. We, therefore, develop an optimization formulation in Chapter 3 that focuses on desirable properties, that is, placing trials from the same breeding groups and with the same RM and stage together while indirectly reducing wasted space in the field. The core idea of our formulation is to split each field into blocks and then create homogenous blocks. By favoring placing similar trials together in blocks, the optimization formulation indirectly favors using fewer blocks, which reduces wasted space in the field.

However, the proposed optimization model can minimizes the wasted space and provide the most favorable arrangement of trials within in a field it has limitations for scaling that to the whole breeding program. First, a commercial breeding program includes multiple fields which trials should be placed and splitting trials among those fields requires same considerations as arranging trials within a field. Also, there are other practical considerations that make a field suitable for a specific group of trials or force the program to avoid a particular field for some trials. So the current formulation is not applicable to the entire breeding program. To address this issue we propose a two phase solution in Chapter 4 for optimal assignment of trials to the fields and arrangement within each. Another limitation with the baseline optimization model is

that the formulation is limited to three mentioned objectives and these criteria may change from one year to the next as practices within the breeding program evolve. For instance there could be decision groups that requires trials of same group to be placed nearby for dicision purpose. So the model requires some level of flexibility in terms of objectives to adjust to the program. The new formulation not only provides an advancement on the existing model to help with scalability and applicability in practice but also makes the process fully automated through the two phase optimization.

1.5 Contributions

One cycle of breeding program takes several years and requires vast testing resources. Such predictive models can optimize the breeding program and accelerate it by providing knowledge on performance of varieties in a breeding program years in advance using machine learning techniques. Knowing that similar GxE varieties have same GxE preference to environment breeders can avoid planting similar GxE varieties in same environments and expand the experiment to more environments through similar GxE varieties in a way that maximizes information gain. A precise prediction of the yield of a target variety in early stages will provide a valuable source of information for breeders to make a better decision on future of a variety in breeding program. In other words, determining a set of similar GxE varieties breeders are able to predict yield of a target variety in some unseen environments. A better evaluation of the target performance saves considerable amount of time and cost. Also, a better evaluation of a target performance using prediction can refuse an unwanted decision of keeping a future failure variety or discarding a potential winner variety based on few unreliable observations in early stages.

Along with all the predictive advancement we improve the breeding process from practical point of view as well. We consider optimization of field trial experiments from the perspective of the plant breeder. Namely, given a set of trials from multiple breeding groups, how should the trials be assigned to specific locations within a large field? The overall goal is improved utilization of each field, but each breeding group would also prefer their trials to be positioned close to each

other for convenience, and trials that are similar in terms of stage and RM should also be placed together as much as possible as this makes them easier to compare and evaluate for possible advancement.

In the end, a correct advancement decision is paramount and anything that can be done to facilitate the comparison between competing experimental genotypes, and would thus aid the advancement process, would be of great value. In fact, the benefit of good trial placement is likely higher than the monetary benefit of reduced waste, although the former is hard to quantify since it depends on ultimately selecting the best varieties and hybrids for commercialization. Needless to say that an optimization tool can solve the assignment problem in order of seconds and help the breeders to save weeks by solving the problem manually. Also the solution provided by the model is the optimal solution which could not be easily obtained by human. A comparison of the model solution to a commercial breeding man-made solution shows considerable improvement. Finally, in practice there are last minute changes to the program like adding new trials or changing the objectives in terms of adding a new consideration or changing the importance of one. In such situation the model still would provide the optimal solution quick regardless of the changes but in manual approach it will require a lot more time and resources to solve the new problem along with more complexity for human brain to handle.

CHAPTER 2. DATA CLUSTERING USING PROXIMITY MATRICES WITH MISSING VALUES

Samira Karimzadeh and Sigurdur Olafsson
Department of Industrial and Manufacturing Systems Engineering, Iowa State University
published in *Expert Systems With Applications*

Abstract

In most applications of data clustering the input data includes vectors describing the location of each data point, from which distances between data points can be calculated and a proximity matrix constructed. In some applications, however, the only available input is the proximity matrix, that is, the distances between each pair of data point. Several clustering algorithms can still be applied, but if the proximity matrix has missing values no standard method is directly applicable. Imputation can be done to replace missing values, but most imputation methods do not apply when only the proximity matrix is available. As a partial solution to fill this gap, we propose the Proximity Matrix Completion (PMC) algorithm. This algorithm assumes that data is missing due to one of two reasons: complete dissimilarity or incomplete observations; and imputes values accordingly. To determine which case applies the data is modeled as a graph and a set of maximum cliques in the graph is found. Overlap between cliques then determines the case and hence the method of imputation for each missing data point. This approach is motivated by an application in plant breeding, where what is needed is to cluster new experimental seed varieties into sets of varieties that interact similarly to the environment, and this application is presented as a case study in the paper. The applicability, limitations and performance of the new algorithm versus other methods of imputation are further studied by applying it to datasets derived from three well-known test datasets.

2.1 Introduction

Data clustering is a well−studied field and many clustering algorithms have been proposed for finding clusters in data. This includes classic but still widely used methods such as k−means and hierarchical clustering Xu and Tian (2015), as well as more recent variants, such as kernel k−means Das et al. (2008), gaussian kernel clustering Güngör and Özmen (2017), Bayesian clustering CHEN et al. (2007) and quantum clustering Shuiping et al. (2013). The applicability of these algorithms, however, depends on the available input data.

Data clustering aims to find groups of points that are similar, which implies that all clustering methods require a distance between any two points to assess their similarity. In most applications those points are characterized by some vectors, which we can think of as the primary input variables, and distances between points are calculated based on the observed values of these vectors. But sometimes those input variables are not available and all we know is a distance or proximity matrix, that is, the distance of one point to another, not how it was or could be obtained. Well−known examples of only having a proximity matrix available occur when we are interested in similarity between documents Jian-Ping and Lihui (2014). While starting from the proximity matrix may appear to simplify the clustering calculations – we need those distances anyway − it is, in fact, limiting since many clustering methods assume the availability of the original vectors. For example, k−means iteratively calculates a centroid for a cluster and then assigns each data point to the closest centroid. Without the original vectors we cannot calculate the centroid and cannot apply k−means or similar methods. Fortunately, some clustering methods are proximity based and only require a distance between points as inputs. This includes hierarchical clustering Ellen (1986), partitioning around medioids Ng and Han (1994), fractal clustering Dan (2000), quantum clustering Shuiping et al. (2013) and certain graph−theory based clustering methods Amir et al. (1999). Such methods can thus be applied in cases where only the proximity matrix is available.

Missing values pose another difficulty in data clustering. Imputation methods are commonly used to deal with missing values and here we propose a new method of this type. While other

solutions exist, using an imputation method has the advantage that it can be applied once to modify the data; and then any clustering algorithm can subsequently be used to cluster the modified data. Many imputation methods are applicable to deal with missing data when a vector characterization of the original data is available; but we are not aware of many imputation methods that can be applied for data clustering where the proximity matrix is partially observed and is the only available information about the data points. In particular, what may be considered advanced methods for imputation, including hot/cold deck imputation, regression imputation, interpolation, and extrapolation all use vectors characterizing each data point to estimate missing values and are therefore not applicable when only the proximity matrix is available Daniel (2003). The only option that we are aware of for such cases is therefore to use a summary statistic such as the mean or median for the imputation. It is possible for such methods to perform well on a specific dataset, but they are also known to be biased and do not account for the fundamental reason for *why* the data is missing Baraldi and Enders (2010). Furthermore, such statistical summary imputations consider all missing values to be the same, which is often not true in practice. Other approaches, such as maximum likelihood and expected maximization can provide an unbiased estimate of missing values; however, as noted above, these algorithms assume that that the missing values are missing values in the vectors characterizing each data point Daniel (2003), as opposed to missing values in the proximity matrix. As far as we know, no previous work has therefore systematically addressed the issue of missing values in the proximity matrix when only the proximity matrix is observed, reducing the users options to simple imputation of mean or median, which may not be effective.

Based on the reasons outlined above, we contend that the lack of effective imputation methods for the scenario where only the proximity matrix is available is significant because simply imputing a mean or a median does not address the reason why a value is missing, which is recognized as a critical issue Garciarena and Santana (2017); Jaemun et al. (2016). In other words, imputing all missing values with the same estimation may end up clustering objects that have no reasonable relation in same clusters. As an illustrative example of an application area

where this would be important, consider a recommendation system where clustering is used to identify groups of raters have the same taste. Intuitively there are likely specific reasons of missing proximity values between different groups of raters; namely they deliberately choose not to rate the same items. Imputing uniform proximity for every missing value may lead some clustering algorithms to place raters that belong to disjoints groups together. There is therefore a need for an imputation method that considers the reason a value is missing.

In order to partially address the shortcoming in the current state−of−the−art, we develop a new method for imputation of missing values in a proximity matrix where the missing values are not missing at random. We do not address all possible reasons for why an observation may be missing, but specifically address scenarios where missing values can be assumed to fall in one of two categories: missing due to complete dissimilarity and missing due to lack of observations. These categories are motivated by a plant science case study to be discussed in detail later and we believe them to be applicable in other areas as well. To get a better quality of data clustering, we therefore need to distinguish these two missing value categories and estimate those missing values appropriately. This is a novel contribution since, as far as we know, no existing imputation methods consider the importance of these two different types of missing values.

The goal of this paper is thus to provide a new method for dealing with specific types of missing data when only the proximity matrix is available. To achieve this goal, in Section 2 we present a graph reformulation of the proximity matrix to identify the reason a value is missing and estimate each missing value as a unique case using a maximum clique algorithm. In Section 3 we evaluate the generality and performance of our imputation method compared to the limited existing benchmark imputation methods using the well−known Iris data. This evaluation suggests a generalization of our method that is presented and evaluated in Section 4. In Section 5 we evaluate the applicability of our method to data with less favorable structure via two more publicly available dataset, and we conclude the paper in Section 6 with a case study based on a real application in plant breeding that motivated this work.

2.2 Proximity Matrix Completion Algorithm

In general, we assume that we have n data points $\{x_i : i = 1, 2, \ldots, n\}$ and we have a distance metric that gives us a the distance $d_{ij} = distance\,(x_i, x_j)$ between pairs of those data points, resulting in what we will refer to as a *proximity matrix*

$$\mathrm{D} = \begin{pmatrix} d_{11} & \ldots & d_{1n} \\ \vdots & \ddots & \ldots \\ d_{n1} & \cdots & d_{nn} \end{pmatrix}.$$

As noted in the introduction, when this matrix is completely observed there is a limited set of existing clustering methods can be applied directly to generate clusters of these data points. This includes classic hierarchical clustering such as the single−line and complete−link algorithms, the latter of which will be used for illustration later in this paper. However, the case where this matrix is only partially observer, either at−random or not−at−random, has not been sufficiently addressed.

Our imputation method utilizes a graph theoretic formulation of the data points, primarily to distinguish between the two categories of missing not−at−random values in the proximity matrix (missing due to complete dissimilarity, and missing due to lack of observations). Specifically, we let $V = \{1, 2, \ldots, n\}$denote the set of data points (rows and columns in D), which we will now interpret as vertices in a graph $G = (V, E)$, where an edge $e = (i, j) \in E$ represent an observed similarity between the corresponding data points (and a lack of an edge implies a missing value in D). A clique $C \subseteq V$ in G is a subset of vertices such that each pair of distinct vertices is connected by an edge. A clique thus corresponds to a subset of data points with complete distance information. A maximum clique is a clique with the property that if one more vertex is added, that subset of vertices is no longer a clique.

Example 2.1 Suppose we are given a proximity matrix with many missing values as follows:

$$\mathrm{D} = \begin{pmatrix} 0 & 0.35 & 0.24 & NA & NA & NA & NA & NA & NA \\ 0.35 & 0 & NA & 0.25 & 0.31 & NA & NA & NA & NA \\ 0.24 & NA & 0 & 0.45 & NA & 0.19 & NA & 0.53 & 0.40 \\ NA & 0.25 & 0.45 & 0 & 0.11 & 0.13 & NA & 0.45 & NA \\ NA & 0.31 & NA & 0.11 & 0 & NA & NA & NA & NA \\ NA & NA & 0.19 & 0.13 & NA & 0 & 0.12 & 0.32 & NA \\ NA & NA & NA & NA & NA & 0.12 & 0 & NA & NA \\ NA & NA & 0.53 & 0.45 & NA & 0.32 & NA & 0 & 0.12 \\ NA & NA & 0.40 & NA & NA & NA & NA & 0.12 & 0 \end{pmatrix}.$$

Note that D is symmetric. This matrix corresponds to a graph $G = (V, E)$ with nine vertices $V = \{1, 2, \ldots, 9\}$ and edges

$E =$

$\{(1,2), (1,3), (2,4), (2,5), (3,4), (3,6), (3,8), (3,9), (4,5), (4,6), (4,8), (6,7), (6,8), (8,9)\}$,

And the distances can be viewed as the weights of the edges. This graph is further visualized in Figure 1, and we observe that we can form a partition $V = \bigcup_{i=1}^{6} C_i$ of the vertices using maximum cliques

$C_1 = \{3,4,6,8\}, C_2 = \{3,8,9\}, C_3 = \{2,4,5\}, C_4 = \{1,2\}, C_5 = \{6,7\}, C_6 = \{1,3\}$.

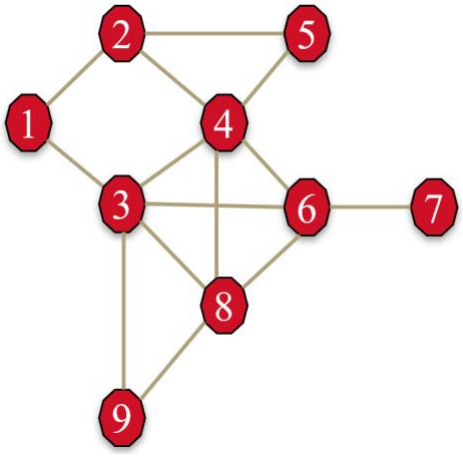

Figure 2.1 Graph corresponding to example proximity matrix

The key idea of our imputation method is, in fact, to utilize the concept of a maximum clique as the basis for dealing with missing data in the proximity matrix D, and specifically distinguishing two types of missing values. Missing values in the proximity matrix imply that the corresponding graph is also incomplete, that is, the graph has missing edges. We use a set of maximum cliques in the graph to identify how to construct a complete graph, corresponding to a new proximity matrix $\hat{D}$ that does not contain any missing values. The algorithm for this is detailed below.

Algorithm *Proximity Matrix Completion*

- Let $G = (V, E)$ be the graph corresponding to a proximity matrix D, where each vertex in V corresponds to a row/column and an edge in E corresponds to a value in D (and $(v, u) \notin E$ implies that $d_{vu}$is missing).

- Any non−missing value in D will be the same in $\hat{D}$, that is,

$$\hat{d}_{vu} = d_{vu}, \forall (v, u) \in E \tag{2.1}$$

- To determine how missing values are imputed, identify a subset of maximum cliques $C_1, \ldots, C_m$ such that

$$V = \bigcup_{i=1}^{m} C_i. \tag{2.2}$$

  Finding maximum cliques is a hard but well−studied problem, and any standard maximum clique algorithm may be used for this step. In particular, we have used the classic Bron–Kerbosch algorithm Coen and Joep (1973). Note that these cliques are not unique and may overlap.

- (**Imputation of missing−due−to−complete−dissimilarity values**.) Let $v, u \in V$ be two vertices such that $\forall C_1, C_2 : v \in C_1, u \in C_2 \Rightarrow C_1 \cap C_2 = \varnothing$, that is, these vertices do not belong to any maximum cliques that intersect. Then we assume that disconnected nodes in

different partitions should in a sense be disconnected because they are completely dissimilar, that is, the corresponding distance values are not missing at random but missing because they are essentially cannot be compared due to complete dissimilarity. We hence introduce an edge with a maximum distance for any such pair of vertices:

$$\hat{d}_{uv} = d_{\max}, \forall (v,u) \notin E : \forall C_1, C_2 : v \in C_1, u \in C_2 \Rightarrow C_1 \cap C_2 = \varnothing. \tag{2.3}$$

- (**Imputation of missing−due−to−incomplete−observations values**.) Let $v, u \in V$ be vertices such that $\exists C_1, C_2 : v \in C_1, u \in C_2, C_1 \cap C_2 \neq \varnothing$. Even if those vertices are not connected, that is, $(v,u) \notin E$, there must exist at least one other vertex with edge to both oan f those vertices, and we can therefore triangulate a value for $\hat{d}_{ij}$ using one or more of those vertices in common. First define the set o,f all vertices that have edges to both u and v

$$V_0(u,v) = \{t \in V : (u,t) \in E, (t,v) \in E\}. \tag{2.4}$$

We now use $V_0(u,v)$ to triangulate an estimate of the distance between u and v and we add an edge $(v,u) \in E$ with value

$$\hat{d}_{uv} = \frac{1}{|V_0(u,v)|} \sum_{t \in V_0(u,v)} |d_{ut} - d_{tv}|. \tag{2.5}$$

With the non−missing values address in Step 1 of the algorithm above, and every missing−value case falling into either the scenario described in Step 3 or Step 4, the algorithm has constructed a complete graph and a corresponding proximity matrix $\hat{D} = \left(\hat{d}_{ij}\right)$ that has no missing values. Figure (2) shows a flowchart of the complete PMC algorithm.

Example 2.2: Applying the PMC algorithm to the proximity matrix D in Example 2.1 gives a complete proximity matrix $\hat{D}$ where the missing edges $\{(2,7),(2,9),(5,7),(5,9),(1,7),(7,9)\}$ are estimated in step 4 as values are missing due to incomplete dissimilarity and the rest of missing edges are estimated in step 3 by triangulation. As a case in point for missing edge d_{14} in step 4 $V_0(u,v) = \{2,3\}$ which V_2 comes from intersection of cliques $C_3 \cap C_4$ and V_3 from

intersection of cliques $C_1 \cap C_6$. So, triangulated estimate of distance $\hat{d}_{14} = \frac{1}{2}(|0.35 - 0.25| + |0.24 - 0.45|)$. The completed proximity matrix is

$$\hat{D} = \begin{pmatrix} 0 & 0.35 & 0.24 & 0.16 & 0.04 & 0.05 & 1.00 & 0.29 & 0.16 \\ 0.35 & 0 & 0.16 & 0.25 & 0.31 & 0.12 & 1.00 & 0.20 & 1.00 \\ 0.24 & 0.16 & 0 & 0.45 & 0.34 & 0.19 & 0.07 & 0.53 & 0.40 \\ 0.16 & 0.25 & 0.45 & 0 & 0.11 & 0.13 & 0.01 & 0.45 & 0.19 \\ 0.04 & 0.31 & 0.34 & 0.11 & 0 & 0.02 & 1.00 & 0.34 & 1.00 \\ 0.05 & 0.12 & 0.19 & 0.13 & 0.02 & 0 & 0.12 & 0.32 & 0.21 \\ 1,00 & 1.00 & 0.07 & 0.01 & 1.00 & 0.12 & 0 & 0.20 & 1.00 \\ 0.29 & 0.20 & 0.53 & 0.45 & 0.34 & 0.32 & 0.20 & 0 & 0.12 \\ 0.16 & 1.00 & 0.40 & 0.19 & 1.00 & 0.21 & 1.00 & 0.12 & 0 \end{pmatrix}.$$

2.3 Numerical Example: Iris Data

We first evaluate the PMC algorithm using a well−known dataset that is often used for illustrating data clustering, namely the iris data first introduced by FISHER (1936). The iris data describes three types of iris flowers (iris setosa, iris versicolor and iris virginica) based on four variables (petal length and width, sepal length and width). There are 50 examples of each type of iris for a total of 150 data points. Since the correct classes are known this is a useful dataset to evaluate the effectiveness of the proposed approach, as well as its potential limitations. The structure of the iris data, with two class values that appear more similar (iris versicolor and iris virginica) and one that is easier to separate (iris setosa), allows us to simulate both scenarios that fit closely with the PMC algorithm assumptions and scenarios that fit less well.

We design an experiment based on two different structures of missing data and use three clustering methods: single−link and complete−link hierarchical clustering, and partitioning around mediod (PAM) partitional clustering. The different structures allow us to evaluate how deviations from the assumptions of the PMC algorithm affect its performance. The three algorithms chosen allow us to evaluate how the clustering algorithm fit for the specific data

impacts the performance of the new PMC algorithm. Since we must use a clustering methods that requires only the proximity matrix the choice of methods is limited and these methods demonstrate a range of fit with the data. For the iris data single−link clustering completely misses the mark by combining the two similar class values into one cluster, complete link hierarchical clustering performs better, but PAM has the best performance on this data. We first calculate a complete proximity matrix and from that point onward assume that this is the only input data (that is, we do not have access to the four primary variables again). We first assume data missing−at−random and remove various percentage of the data. We then evaluate how well complete−link, single−link and PAM k−mediods clustering algorithms recovers the correct clustering results after the missing values are imputed using the PMC algorithm. Here there are no missing values due to complete dissimilarity so no values should be calculated according to Step 3 of the PMC algorithm. This may be considered a non−ideal case for the PMC algorithm, but it should still be able to impute values that lead to a clustering algorithm recovering the correct clusters a reasonable fraction of time. For the second setting we remove all distance comparisons between certain types of iris flowers, namely, iris setosa is assumed completely dissimilar, before removing a certain percentage of the remaining distance values at random (incomplete information). We then apply the PMC algorithm to impute values. Here some values are missing due to complete dissimilarity and should be imputed according to Step 3, while others are missing due to incomplete information and should be imputed according to Step 4 of the PMC algorithm. This setting is therefore a better fit with the assumptions of the PMC algorithm, and we would expect it to perform well.

For the first experiment, when values are missing at random, we tested the PMC algorithm for different percentage of missing values ranging from 10% to 95%, cluster using a complete link algorithm into three clusters, and determine how well we identify the true iris classes. As noted above, imputation methods that use observation of the original vectors cannot be compared to the PMC algorithm because we assume that we only have access to the partially observed proximity matrix. The benchmarks that we use are therefore a simple imputation of the mean

and imputation of the median. Table 1 compares average accuracy of the PMC algorithm to the benchmarks with 1000 replications of the imputation and clustering process.

Table 2.1 PMC vs.benchmark imputations (values missing-at-random)

Measure	Sparsity level	Complete-link			Single-link			PAM k-medoids		
		PMC Algorithm	Impute Mean	Impute Median	PMC Algorithm	Impute Mean	Impute Median	PMC Algorithm	Impute Mean	Impute Median
Accuracy	None missing	84.0%	84.0%	84.0%	65.3%	65.3%	65.3%	94.7%	94.7%	94.7%
	10% missing	**59.8%**	57.3%	53.4%	**66.2%**	65.4%	65.4%	**73.9%**	74.6%	76.7%
	30% missing	**61.8%**	59.4%	55.4%	**65.9%**	65.6%	65.6%	**80.2%**	71.0%	74.9%
	50% missing	**62.3%**	57.1%	53.5%	**67.1%**	65.8%	65.8%	**86.3%**	75.7%	76.6%
	70% missing	**64.0%**	54.4%	53.2%	**67.7%**	66.1%	66.1%	**85.5%**	54.0%	54.6%
	90% missing	36.6%	**55.4%**	55.0%	33.5%	**57.4%**	57.5%	**64.8%**	37.2%	35.7%
	95% missing	34.9%	**54.9%**	54.7%	33.5%	**33.7%**	34.9%	**36.4%**	32.8%	32.7%
Silhouette	None missing	0.51	0.51	0.51	0.51	0.51	0.51	0.52	0.52	0.52
	10% missing	**0.51**	0.41	0.44	**0.47**	**0.47**	**0.47**	**0.54**	0.48	0.48
	30% missing	**0.52**	0.33	0.35	**0.5**	0.37	0.37	**0.54**	0.37	0.38
	50% missing	**0.53**	0.25	0.26	**0.51**	0.26	0.26	**0.55**	0.26	0.27
	70% missing	**0.53**	0.15	0.16	**0.51**	0.15	0.15	**0.56**	0.11	0.12
	90% missing	−0.13	**0.06**	**0.06**	−0.35	**0.04**	0.03	**0.28**	−0.01	−0.02
	95% missing	−0.11	**0.03**	**0.03**	−0.08	**0**	−0.01	**0.07**	−0.02	−0.03

The first row of Table 1 in each section shows how well each algorithm can cluster the iris data with a completely observed proximity matrix. When the data has missing−at−random values, the PMC algorithm results in higher accuracy than the benchmark imputation methods for up to 80% sparsity in the proximity matrix for agglomerative hierarchical algorithms, and consistently better accuracy for PAM k−mediod clustering method. Although the performance of clustering methods might seem similar for some levels of sparsity, especially for the hierarchical methods, the silhouette width shows that the quality of final clusters found from the data imputed using the PMC method could be considered better.

The accuracy of the method decreases by increasing the sparsity rate for every imputation method. However, for an extremely sparse distance matrix in this case 90% sparse, simply imputing the mean or median outperforms the PMC algorithm for this missing data structure based on the clustering method. For the benchmark methods the accuracy decreases slowly but steadily, while for the PMC algorithm the accuracy first increases slightly and then drops of very quickly. These observations are in fact intuitive. The PMC algorithm works based on triangulation when there is a maximum clique in common and for an extremely sparse data the

algorithm is often unable to find any vertices in common to estimate the missing edge. In such cases the algorithm assumes complete dissimilarity and estimates the missing edge accordingly, which for this experiment should never be done (that is, here all missing values are missing−at−random). For moderate to large percentage of missing data this does not happen, but for extremely high percentage of missing data this suddenly starts occurring frequently, which explains why a simple imputation method performs better when we have high percentage of missing data. The reason why, using the PMC algorithm, the accuracy first increases before decreasing rapidly has to do with a property of the iris data. It is well known that for this dataset one type of iris is easy to separate from the others while two have some overlap in the explanatory variables. What we observe is that for high percentage of missing data many of the overlapping pairs in the proximity matrix are removed and replaced by triangulated values based on non−overlapping pairs, making it easier for clustering algorithm to correctly separate these two iris types. While interesting, this is thus due to an idiosyncrasy of this test data, not a pattern we can expect to generalize to many other datasets.

Table 2.2 PMC vs. benchmark imputations (values missing due to two reason)

Measure	Sparsity level	Complete- link			Single-link			PAM k-medoids		
		PMC Algorithm	Impute Mean	Impute Median	PMC Algorithm	Impute Mean	Impute Median	PMC Algorithm	Impute Mean	Impute Median
Accuracy	None missing	84.0%	84.0%	84.0%	65.3%	65.3%	65.3%	94.7%	94.7%	94.7%
	50% missing	**61.4%**	34.7%	32.5%	**66.7%**	65.4%	65.4%	**75.7%**	60.1%	61.9%
	60% missing	**59.4%**	31.9%	31.3%	**67.6%**	65.6%	65.6%	**66.5%**	63.3%	61.2%
	70% missing	**61.6%**	32.3%	31.3%	**68.7%**	65.8%	65.8%	61.1%	**61.4%**	49.3%
	80% missing	**63.3%**	34.2%	32.1%	**69.2%**	65.8%	64.3%	**62.4%**	52.1%	43.1%
	90% missing	**39.5%**	38.3%	36.9%	**66.0%**	53.5%	42.4%	**67.2%**	35.5%	35.4%
	95% missing	35.4%	**39.6%**	38.9%	**65.8%**	34.1%	34.1%	**57.5%**	34.5%	34.9%
Silhouette	None missing	0.51	**0.51**	0.51	**0.51**	0.51	0.51	**0.52**	0.52	0.52
	50% missing	**0.53**	0.18	0.18	**0.53**	0.11	0.01	**0.57**	0.26	0.17
	60% missing	**0.54**	0.14	0.15	**0.56**	0.08	0	**0.58**	0.21	0.13
	70% missing	**0.54**	0.11	0.11	**0.56**	0.05	−0.01	**0.58**	0.14	0.09
	80% missing	**0.54**	0.08	0.07	**0.56**	0.03	−0.01	**0.59**	0.04	0.03
	90% missing	−0.16	**0.04**	0.03	**0.45**	0.01	−0.01	**0.51**	0	−0.01
	95% missing	−0.17	**0.02**	**0.02**	**0.36**	0	−0.01	**0.28**	−0.01	−0.01

Table 2 shows same comparison when values are missing both at random and due to complete dissimilarity. This means that the data is missing for the two underlying reasons assumed by the PMC algorithm. As expected, PMC performs significantly better than the benchmark imputation

methods because of ability to distinguish between types of missing values. However, for extremely high percentage of missing data, a simple imputation of mean or median may still be better. In particular, when the PMC algorithm is used in conjunction with the complete−link algorithm and more than 80% of the data is missing, the clusters obtained may not be sensible as evidenced by the negative Silhouette width. However, when applied in conjunction with either single−link or PAM, imputing data using PMC results in sensible clusters and better clusters than the benchmark methods even for the highest percentage of missing values. It therefore appears to be a combination of the clustering method and the data that may cause PMC to perform poorly for high percentage of missing data. While our results indicate that this rarely happens when the PMC assumptions regarding why data is missing are satisfied (Table 2), this potential limitation will be addressed in the next section.

2.4 Extension of PMC for High-Percentage Missing Data

As is noted before, the PMC algorithm assumes that every missing data point in the proximity matrix is missing for one of two reasons: 1) missing due to complete dissimilarity of the objects being compared, or 2) missing due to lack of observations (random or not−at−random). As explained in Section 2, we furthermore assume that we can identify each case via the existence of maximal cliques. In the case of overlapping maximal cliques we assume the second case, and if there are no overlapped maximal cliques we assume the first case. As we have seen for the iris data in Section 3 this may work well in practice, but it is also possible that very high percentage of missing−at−random data may result in missing observations with no overlapping maximal cliques. In other words, large percentage of missing data may result in missing values in the proximity matrix being incorrectly identified as missing due to complete dissimilarity. For such cases it is possible to extend the PMC algorithm while still utilizing the maximal clique concept.

We specifically suggest that the following extension may work well for datasets with very high percentage of missing data, where most of the data is missing at random. Instead of imputing a maximal distance for all values with no overlapping cliques, we find the number of cliques needed

to connect two vertices and if it meets a minimum number then we use another imputation. In other words, we apply Step 4 unchanged but essentially split Step 3 cases depending on how close the vertices are in terms of overlapping cliques. Vertices that are, say, only two cliques apart could be assigned a mean or a median value, whereas all vertices that are more than two cliques apart would be assigned the maximum distance as in the standard Step 3 procedure.

Table 2.3 Extended PMC vs. benchmark imputations (values missing-at-random)

Clustering method	Sparsity level	Accuracy					Silhouette				
		PMC/No extension	PMC/ Mean	PMC/ Median	Imputing Mean	Imputing Median	PMC/ No extension	PMC/ Mean	PMC/ Median	Imputing Mean	Imputing Median
Complete-link	None missing	84.0%	84.0%	84.0%	84.0%	84.0%	0.51	0.51	0.51	0.51	0.51
	50% missing	62.3%	**67.3%**	66.7%	57.1%	53.5%	**0.53**	**0.53**	**0.53**	0.25	0.26
	60% missing	63.6%	**69.2%**	68.6%	56.0%	53.4%	0.53	**0.54**	**0.54**	0.2	0.21
	70% missing	64.0%	**70.7%**	70.1%	54.4%	53.2%	0.53	**0.55**	**0.55**	0.15	0.16
	80% missing	48.5%	**61.8%**	62.7%	54.7%	53.6%	0.18	**0.55**	**0.55**	0.1	0.11
	90% missing	36.6%	**58.8%**	58.1%	55.4%	55.0%	−0.13	0.37	**0.39**	0.06	0.06
	95% missing	34.9%	**59.6%**	59.0%	54.9%	54.7%	−0.11	0.14	**0.15**	0.03	0.03
Single-link	None missing	65.3%	65.3%	65.3%	65.3%	65.3%	0.51	0.51	0.51	0.51	0.51
	50% missing	**67.1%**	**67.1%**	**67.1%**	65.8%	65.8%	**0.51**	**0.51**	**0.51**	0.26	0.26
	60% missing	**67.5%**	**67.5%**	**67.5%**	65.9%	65.9%	**0.51**	**0.51**	**0.51**	0.2	0.21
	70% missing	**67.7%**	**67.7%**	**67.7%**	66.1%	66.1%	**0.51**	**0.51**	**0.51**	0.15	0.15
	80% missing	49.4%	49.4%	49.4%	**66.1%**	**66.1%**	**0.16**	**0.16**	**0.16**	0.1	0.1
	90% missing	33.5%	33.5%	33.5%	57.4%	**57.5%**	−0.35	−0.21	−0.21	**0.04**	**0.03**
	95% missing	33.5%	33.5%	33.5%	33.7%	**34.9%**	−0.08	−0.05	−0.07	**0**	**−0.01**
PAM k-medoids	None missing	94.7%	94.7%	94.7%	94.7%	94.7%	0.52	0.52	0.52	0.52	0.52
	50% missing	86.3%	**86.6%**	**86.6%**	75.7%	76.6%	**0.55**	**0.55**	**0.55**	0.26	0.27
	60% missing	85.9%	**86.2%**	**86.2%**	53.1%	57.0%	**0.56**	**0.56**	**0.56**	0.18	0.18
	70% missing	85.5%	**85.7%**	**85.7%**	54.0%	54.6%	**0.56**	**0.56**	**0.56**	0.11	0.12
	80% missing	84.3%	**85.1%**	**85.1%**	46.2%	45.0%	0.56	**0.57**	**0.57**	0.04	0.04
	90% missing	64.8%	73.3%	**74.4%**	37.2%	35.7%	0.28	0.42	**0.43**	−0.01	−0.02
	95% missing	36.4%	**46.8%**	44.2%	32.8%	32.7%	0.07	**0.08**	0.07	−0.02	−0.03

We repeat the two experiments from Section 4 and the results are shown in Tables 3−4. The results indicate that extending the algorithm to impute mean or median of the observed distances for missing values when there is no overlap among cliques helps to improve the accuracy as intended, especially when the percentage of missing values is high and when the values are exclusively missing−at−random. When the assumptions of the PMC algorithm are better satisfied (second experiment), then the original PMC algorithm performs best except for extremely high percentage of missing values, where the extended PMC algorithm outperforms all other approaches.

Overall, the experiments with the iris data indicate that the PMC algorithm is useful for a variety of missing data structures. Furthermore, its major limitation is in cases where the

Table 2.4 Extended PMC vs. benchmark imputations (values missing due to two reason)

Clustering method	Sparsity level	Accuracy					Silhouette				
		PMC/No extension	PMC/ Mean	PMC/ Median	Imputing Mean	Imputing Median	PMC/No extension	PMC/ Mean	PMC/ Median	Imputing Mean	Imputing Median
Complete-link	None missing	84.0%	84.0%	84.0%	84.0%	84.0%	0.51	0.51	0.51	0.51	0.51
	50% missing	**61.4%**	34.3%	34.1%	34.7%	32.5%	**0.53**	0.2	0.2	0.18	0.18
	60% missing	**59.4%**	32.5%	32.5%	31.9%	31.3%	**0.54**	0.2	0.2	0.14	0.15
	70% missing	**61.6%**	31.0%	30.9%	32.3%	31.3%	**0.54**	0.2	0.2	0.11	0.11
	80% missing	**63.3%**	30.1%	30.0%	34.2%	32.1%	**0.54**	0.2	0.21	0.08	0.07
	90% missing	**39.5%**	35.4%	34.2%	38.3%	36.9%	−0.16	**0.16**	**0.16**	0.04	0.03
	95% missing	35.4%	**38.9%**	38.1%	39.6%	38.9%	−0.17	**0.07**	**0.07**	0.02	0.02
Single-link	None missing	65.3%	65.3%	65.3%	65.3%	65.3%	0.51	0.51	0.51	0.51	0.51
	50% missing	66.7%	**66.8%**	**66.8%**	65.4%	65.4%	**0.53**	0.14	0.04	0.11	0.01
	60% missing	67.6%	**67.7%**	**67.7%**	65.6%	65.6%	**0.56**	0.21	0.11	0.08	0
	70% missing	**68.7%**	**68.7%**	**68.7%**	65.8%	65.8%	**0.56**	0.28	0.19	0.05	−0.01
	80% missing	**69.2%**	69.1%	69.1%	65.8%	64.3%	**0.56**	0.34	0.27	0.03	−0.01
	90% missing	**66.0%**	**66.0%**	**66.0%**	53.5%	42.4%	**0.45**	0.27	0.21	0.01	−0.01
	95% missing	**65.8%**	**65.8%**	**65.8%**	34.1%	34.1%	**0.36**	0.16	0.11	0	−0.01
PAM k-medoids	None missing	94.7%	94.7%	94.7%	94.7%	94.7%	0.52	0.52	0.52	0.52	0.52
	50% missing	**75.7%**	63.0%	63.2%	60.1%	61.9%	**0.57**	0.32	0.22	0.26	0.17
	60% missing	**66.5%**	63.9%	58.2%	63.3%	61.2%	**0.58**	0.38	0.29	0.21	0.13
	70% missing	61.1%	**64.9%**	61.3%	61.4%	49.3%	**0.58**	0.41	0.36	0.14	0.09
	80% missing	62.4%	**65.5%**	65.2%	52.1%	43.1%	**0.59**	0.44	0.41	0.04	0.03
	90% missing	**67.2%**	65.9%	65.5%	35.5%	35.4%	**0.51**	0.41	0.38	0	−0.01
	95% missing	**57.5%**	44.8%	43.0%	34.5%	34.9%	**0.28**	0.07	0.04	−0.01	−0.01

percentage of missing data becomes so large that many pairs are misidentified as missing due to complete dissimilarity. However, even in such cases the extended PMC algorithm can be applied effectively in its place. Of course, as expected, the PMC algorithm performs best when the missing data structure is the closest to its assumptions and results for both experiments illustrate that selecting an effective clustering algorithm is also important to the effectiveness of the PMC algorithm.

The results reported in tables 1−4 are average performance of PMC algorithm versus benchmark imputation methods over 1000 replication of the experiments. A t−test with 95% confidence interval has been done on accuracy difference of PMC to benchmark imputation methods to evaluate the significance of the improvement that PMC algorithm can achieve. Table 5 reports the results of these tests. As is shown in the table, the PMC algorithm can result in significant improvements in quality of clustering output even in extremely sparse proximity matrices.

Table 2.5 T-test of 95% CI on PMC accuracy improvement over benchmark imputations

Clustering method	Sparsity level	Missing only at random				Missing at random/complete dissimilarity			
		PMC over imputing Mean	PMC over imputing Median	PMC extended to Mean over imputing Mean	PMC extended to Median over imputing Median	PMC over imputing Mean	PMC over imputing Median	PMC extended to Mean over imputing Mean	PMC extended to Median over imputing Median
Complete-link	50% missing	5.2	8.8	10.23	13.22	26.71	28.88	−0.35	1.57
	60% missing	7.62	10.21	13.17	15.21	27.5	28.09	0.69	1.27
	70% missing	9.6	10.78	16.36	16.93	29.29	30.25	−1.3	−0.45
	80% missing	−6.24	−5.11	7.03	9.14	29.06	31.19	−4.1	−2.11
	90% missing	−18.86	−18.37	3.34	3.13	1.18	2.57	−2.9	−2.68
	95% missing	−20	−19.74	4.61	4.3	−4.23	−3.47	−0.73	−0.81
Single-link	50% missing	1.25	1.25	1.25	1.25	1.31	1.31	1.39	1.39
	60% missing	1.58	1.58	1.58	1.58	2.09	2.09	2.11	2.11
	70% missing	1.66	1.66	1.66	1.66	2.94	2.94	2.94	2.94
	80% missing	−16.69	−16.69	−16.69	−16.69	3.34	4.89	3.3	4.85
	90% missing	−23.97	−24.01	−23.97	−24.01	12.5	23.6	12.47	23.57
	95% missing	−0.23	−1.39	−0.23	−1.39	31.72	31.79	31.67	31.75
PAM k-medoids	50% missing	10.64	9.77	10.92	10.09	15.68	13.83	2.98	1.29
	60% missing	32.77	28.9	33.08	29.21	3.22	5.35	0.61	−2.92
	70% missing	31.51	30.95	31.65	31.09	−0.28	11.81	3.5	11.96
	80% missing	38.14	39.29	38.94	40.06	10.29	19.25	13.44	22.07
	90% missing	27.6	29.16	36.07	38.75	31.72	31.81	30.43	30.13
	95% missing	3.61	3.71	14.05	11.5	23	22.58	10.33	8.03

2.5 Evaluation on Different Datasets

The iris data experiment reported in the previous two sections provided insights into the effectiveness of the PMC algorithm and specifically demonstrated that while it is generally applicable, it is more effective when the structure of the missing data is close to it's assumptions (that is, both missing at random and due to complete dissimilarity) and when the chosen clustering algorithm is effective for the data (e.g., PAM or complete−link versus single−link). What remains to be investigated is how its effectiveness depends on the nature of the data itself. In this section we explore this via two very different datasets, both available through the UCI repository of machine learning datasets. The first dataset is the wine data, which includes 178 observation of chemical analysis of 13 quantities of three types of wine grown in the same region of Italy Aeberhard et al. (1992). The second data set is the glass data, which includes 214 observations of six types of glass and is motivated by criminal investigations at the scene of a crime Evett and Spiehler (1987). Table 6 provides cluster quality results for the three candidate clustering methods for these two datasets with no missing values. We observe that PAM performs best for the wine data, whereas complete−link performs best for the glass data. Since we now

want to investigate the effect of the data itself, we ran 1000 replications of the experiment described in Section 3 and 4 above using the best clustering method on each dataset (PAM for wine data and complete−link for glass data). Tables 7−8 show results of these experiments.

Table 2.6 Candidate clustering methods performance on new datasets

Dataset/clustering method	Wine		Glass	
	Accuracy	Silhouette	Accuracy	Silhouette
Complete-link	67.4	0.54	50.5	0.59
Single-link	29.8	0.49	21.0	0.37
PAM k-mediods	70.8	0.57	33.2	0.31

Table 2.7 PMC performance over new datasets (values missing-at-random)

Clustering method/dataset	Sparsity level	Accuracy					Silhouette				
		PMC/No extension	PMC/ Mean	PMC/ Median	Imputing Mean	Imputing Median	PMC/No extension	PMC/ Mean	PMC/ Median	Imputing Mean	Imputing Median
Complete-link/glass	None missing	50.5%	50.5%	50.5%	50.5%	50.5%	0.59	0.59	0.59	0.59	0.59
	50% missing	37.5%	39.1%	39.1%	**39.3%**	38.2%	0.36	**0.47**	**0.47**	0.25	0.23
	60% missing	37.2%	39.1%	39.1%	**39.2%**	38.6%	0.35	**0.5**	**0.5**	0.2	0.17
	70% missing	37.3%	39.2%	39.2%	**39.4%**	39.1%	0.36	**0.53**	**0.53**	0.14	0.12
	80% missing	37.8%	39.1%	39%	**39.7%**	39.4%	0.24	0.55	**0.56**	0.09	0.07
	90% missing	35.9%	37.7%	37.4%	**40.5%**	39.9%	0.09	0.42	**0.44**	0.05	0.03
	95% missing	34%	37.7%	37.8%	**41.5%**	41.3%	−0.17	0.12	**0.13**	0.02	0.02
PAM k-medoids/wine	None missing	70.8%	70.8%	70.8%	70.8%	70.8%	0.57	0.57	0.57	0.57	0.57
	50% missing	**70.1%**	**70.1%**	**70.1%**	66.8%	65.3%	**0.55**	**0.55**	**0.55**	0.23	0.22
	60% missing	**70.1%**	**70.1%**	**70.1%**	60.4%	60.0%	**0.55**	**0.55**	**0.55**	0.14	0.12
	70% missing	**70.2%**	**70.2%**	**70.2%**	55.7%	55.9%	**0.54**	**0.54**	**0.54**	0.07	0.06
	80% missing	**70.2%**	**70.2%**	**70.2%**	46.3%	46.0%	**0.53**	**0.53**	**0.53**	0.02	−0.01
	90% missing	55.0%	**68.3%**	68.1%	36.2%	36.7%	0.21	0.41	**0.42**	−0.02	−0.04
	95% missing	34.3%	**49.2%**	45.9%	34.7%	35.4%	0.05	**0.06**	0.04	−0.02	−0.03

These results illustrate how the conclusions of Section 3 and Section 4 might change for data with different structure, and this structure is insightful to interpret the results. The glass data has seven class values and the ideal clustering result would identify seven clusters corresponding to these values. However, there are two majority classes, including one that represents 35.5% of the data. From a classification perspective we could therefore simply predict the majority class, ignore the other six class values and achieve 35.5% classification accuracy on the data. We notice that while the best clustering algorithm (complete−link) achieves a better performance in separating the seven class values when given the whole data, when at least 50% of the data is

Table 2.8 PMC performance over new datasets (values missing due to two reason)

Clustering method/dataset	Sparsity level	Accuracy					Silhouette				
		PMC/No extension	PMC/ Mean	PMC/ Median	Imputing Mean	Imputing Median	PMC/No extension	PMC/ Mean	PMC/ Median	Imputing Mean	Imputing Median
Complete-link/glass	None missing	50.5%	50.5%	50.5%	50.5%	50.5%	0.59	0.59	0.59	0.59	0.59
	50% missing	37.2%	37.2%	37.2%	**39.8%**	38.9%	**0.37**	**0.37**	**0.37**	0.25	0.21
	60% missing	37.3%	37.3%	37.3%	**39.9%**	38.8%	**0.37**	**0.37**	**0.37**	0.2	0.16
	70% missing	36.9%	36.9%	36.9%	**39.4%**	38.9%	**0.37**	**0.37**	**0.37**	0.14	0.12
	80% missing	31.3%	31.3%	31.3%	**39.9%**	39.6%	0.09	0.09	0.09	**0.1**	0.07
	90% missing	36.7%	36.7%	36.7%	40.2%	**40.3%**	0.03	0.03	0.03	**0.05**	0.03
	95% missing	34.1%	34.1%	34.1%	41.7%	**41.8%**	−0.2	−0.2	−0.2	**0.02**	**0.02**
PAM k-medoids/wine	None missing	70.8%	70.8%	70.8%	70.8%	70.8%	0.57	0.57	0.57	0.57	0.57
	50% missing	**70.8%**	**70.8%**	**70.8%**	65.5%	66.0%	**0.55**	**0.55**	**0.55**	0.21	0.14
	60% missing	**70.4%**	**70.4%**	**70.4%**	59.3%	55.4%	**0.54**	**0.54**	**0.54**	0.08	0.05
	70% missing	**70.0%**	**70.0%**	**70.0%**	50.4%	45.5%	**0.53**	**0.53**	**0.53**	0	−0.03
	80% missing	**68.9%**	**68.9%**	**68.9%**	38.7%	38.3%	**0.48**	**0.48**	**0.48**	−0.01	−0.07
	90% missing	**54.9%**	**54.9%**	**54.9%**	34.1%	35.9%	**0.32**	**0.32**	**0.32**	−0.01	−0.05
	95% missing	**40.9%**	**40.9%**	**40.9%**	34.7%	35.9%	**0.02**	**0.02**	**0.02**	−0.02	−0.04

missing (at random or otherwise), the performance becomes similar to simply predicting that all the data belongs to the majority class. It is therefore not surprising that imputing the mean or median is very competitive relative to the PMC algorithm, and it is likely not worthwhile to use the PMC algorithm for such datasets.

The wine data has a very different structure. Here removing 50% of the data at random appears to have minimal effect on the performance. For such data, the PMC algorithm dominates the performance of imputing the mean or the median, but the patterns observed in Section 3 and Section 4 for how relative performance changes as the percentage of missing data increases are much less pronounced. For such data the effectiveness of the PMC algorithm appears clear, but on the other hand, there is little motivation to use the extended PMC over the basic PMC algorithm. The results from these two datasets thus provide some insights into the type of data where the PMC algorithm and its extension can be expected to be most effective.

2.6 Case Study: Plant Breeding

We now turn to the application that motivated the PMC algorithm. While we have developed a general method to treat missing values in a proximity matrix, and we have developed some insights into when it is most effective, our original motivation comes from an application in plant

breeding. In this application our data points correspond to different experimental plant varieties that will potentially become commercial crop varieties. In this context it is valuable to know the similarity between those varieties as it pertains to how they react to different environments, what is generally referred to as genotype−by−environment (GxE) effects Des Marais et al. (2013); Li et al. (2018). Thus, a distance measure can be defined where zero difference implies two varieties respond in exactly the same way to different growing environment, and the larger the value the more dissimilar the varieties. A proximity matrix is thus defined, but only the proximity matrix is available. Furthermore, it is not possible to define this GxE similarity for all variety pairs, as it requires the varieties having been planted in the same environments. This will not happen for most pairs for various reasons. For some pairs, their environmental requirements are simply too dissimilar, which makes them unlikely to be planted in the same location. For such cases, there are no environments in common for explainable reasons and the corresponding values in the proximity matrix will be *missing due to complete dissimilarity*. Other elements of the matrix are missing for unknown but potentially not−at−random reasons. The main reason for this is it is only economically feasible to plant each variety in a limited number of locations (decisions made by plant breeders). The corresponding values in the proximity matrix will be *missing due to incomplete observations*. This application thus naturally gives rise to a proximity matrix where most of the values are missing for two fundamentally different reasons (environmental dissimilarity versus breeder decisions) and we are interested in finding clusters based on this incomplete proximity matrix.

For this case study we start from a two−way data matrix where rows correspond to experimental soybean varieties and columns correspond to test environments. The response in the matrix is the mean yield observed for a variety in that environment. What is of interest is to determine what we call GxE similarity, that is, how similarly two varieties interact with the same environment, similarity of GxE interactions.

Example 5.1: To illustrate what we mean by GxE similarity, let's consider a small synthetic example of 5 varieties in 9 environments. Suppose the following yield observations in table 9 are made:

Table 2.9 Sample field observations

	E1	E2	E3	E4	E5	E6	E7	E9
V1	68.0	60.4	72.2	57.6	59.9	81.1	52.2	90.4
V2	59.8	51.1	62.4	49.8	37.6	60.2	31.9	68.7
V3	48.9	39.2	52.7	37.5	38.1	62.9	33.8	70.6
V4	64.5	57.4	67.3	55.7	42.6	65.2	37.4	74.0
V5	63.3	55.7	67.2	52.2	53.7	77.3	48.3	84.5

It may, for example, seem that V_2 and V_3 are similar as they have lower yield and the other three have higher yield. However, the similarity we are interested in is not similarity of yield, but rather similarity of the GxE interactions. Considering the sum of squared difference in estimated interaction effects in this two−way table and then normalizing the numbers to be between zero and one, these yield observations are converted into proximity matrix:

$$D = \begin{pmatrix} 0 & 0.88 & 0.03 & 0.93 & 0.01 \\ 0.88 & 0 & 0.95 & 0.01 & 0.88 \\ 0.03 & 0.95 & 0 & 1 & 0.01 \\ 0.93 & 0.01 & 1 & 0 & 0.93 \\ 0.01 & 0.88 & 0.01 & 0.93 & 0 \end{pmatrix}.$$

We note that V_1, V_3 and V_5 are very similar, not because they have similar yield (in fact, V_1 has very high yield but V_3 has very low yield), but because they have similar preference for each environment. All three varieties have higher than expected yield in E_6, E_7 and E_9, and lower than expected performance in the other four environments. The opposite is true for V_2 and V_4, which are hence similar to each other but dissimilar to the other three varieties.

Understanding GxE similarity is important when it comes to predicting plant phenotype, such as yield, and hence to decision making in commercial plant breeding. Specifically, if the two varieties have exactly the same interaction effects between genotype and environment, yield can

be modeled with a non−interaction model that predicts yield simply as a function of genotype main effects and environment main effects, which makes predicting yield in new environments simpler. Having the ability to predict yield of experimental soybean varieties in environments where they have not been tested aids decision makers when breeders need to decide if to keep a specific variety in breeding program and plant the variety again in following year, or discard it from the program.

Plant phenotype is often modeled as function of genotype effect x_v, environmental effect x_e, and the interaction effects x_{ve} between genotype and environment (GxE). We will use the following model of the yield μ_{ve} of variety v in environment e

$$\mu_{ve} = \mu + x_v + x_e + x_{ve} + \varepsilon_{ve} \tag{2.6}$$

where μ denotes the overall average yield and ε_{ve} models the variability. We are interested in similarity of the interaction effects, that is, how similar x_{ve} and x_{ue} are across all environments for a pair of varieties (v, u). If variety v is tested (planted) in a set of environments $E(v)$ and u is similarly tested in $E(u)$, where the set of common testing environments is non−empty, that is, $E(v, u) \equiv E(v) \cap E(u) \neq \varnothing$, then we can estimate this similarity (across common testing environments) as follows. Dropping ε_{ve}, since we will be using empirically observed averages, we note that $x_{ve} = \mu_{ve} - (\mu + x_v) - x_e$, and x_e is fixed in a common testing environment $e \in E(v, u)$. Furthermore, $\mu_v \equiv \mu + x_v$ represents the average yield of variety v, and we can write the difference in interaction effects between the two varieties in a fixed environments as $x_{ve} - x_{ve} = (\mu_{ve} - \mu_v) - (\mu_{ue} - \mu_u)$. We can therefore calculate the estimated dissimilarity or distance between the two varieties as

$$d\,(v, u) = \frac{1}{K} \frac{1}{|E(v, u)|} \sum_{e \in E(v,u)} \left| (\overline{\mu}_{ve} - \overline{\mu}_v) - (\overline{\mu}_{ue} - \overline{\mu}_u) \right|. \tag{2.7}$$

Here $\overline{\mu}_{ve}$ is the observed average yield of variety v in environment e, and $\overline{\mu}_v$ is the observed overall average yield for variety v. We finally normalize the values to be between zero and one, by dividing by $K = \max_{v,u} \frac{\sum_{e \in E(v,u)} |(\overline{\mu}_{ve} - \overline{\mu}_v) - (\overline{\mu}_{ue} - \overline{\mu}_u)|}{|E(v,u)|}$, resulting in a proximity matrix D.

This proximity matrix will have a large percentage of missing values. Having a complete proximity matrix D requires that we have planted each pair of varieties in at least one common test environments. As noted in the introduction this is not possible for two reasons. First, it is economically infeasible to plant every experimental variety in every location every year. Breeders must make choices as to which locations to plant each experimental variety and there are many GxE distances missing in the proximity matrix due to simple lack of observations. Second, there are other values missing due to complete dissimilarity because they are not intended for the same location. Soybeans, like other crops, have what is called relative maturity (RM) that we can think of as indicating the number of days needed to mature before first frost. Varieties with small RM values are appropriate for locations with short growing seasons and varieties with large RM values are appropriate for locations with longer growing seasons. While there is significant overlap in where different varieties are tested, certain varieties will simply never be tested against each other, that is, they have complete dissimilarity. Therefore, distinguishing different types of missing values in a sparse proximity matrix is crucial for this application.

To demonstrate the PMC algorithm in practice we apply it to observation of 1033 soybean varieties that are part of a commercial breeding program. We start by building a proximity matrix using only field observations. This set of 1033 varieties was specifically selected as varieties that have been planted the most widely, but even in this case the proximity matrix is just 40% filled. (Depending on where they are in the breeding program, other sets of varieties will results in proximity matrices with much higher percentage of missing values, many over 90% missing.) The PMC algorithm can now be applied to fill in the similarity matrix completely, either by imputation through triangulation or by assigning the maximum value of one (complete dissimilarity). For this case study of 1033 soybean varieties, Figure 3 graphically represents the sparsity of the proximity matrix. For this graph, the varieties have been ordered by the size of the maximum cliques to which they belong. Color indicates the number of common planting environments, that is, at least 20 (yellow), 1-19 (green), or zero (white). Yellow dots correspond to pairs of varieties with substantial number of testing environment in common

($|E(v,u)| \geq 20$)and green dots correspond to measures of GxE based on a few environments in common ($|E(v,u)| \in [1,19]$). Finally, white dots represent pairs of varieties with no common testing environments.

As is mentioned before, in this case study values are missing because of lack of observation in common environment or due to complete dissimilarity of varieties based upon RM. So sorting the proximity matrix based on RM given more insights into the nature of the missing values. Figure 4 represents the same proximity matrix as in Figure 2 ordered by RM from low rm to the left/high, high rm to the right/low). Color indicates at least 20 common testing environments (yellow), 1-19 common testing environments (green), overlapping maximum cliques (white), and no overlapping maximum cliques (blue).

Reordering the proximity matrix makes it clear that pair of varieties where $|E(v,u)| \geq 20$ (yellow) have mostly similar RM values (close to diagonal), and pairs with no common testing environments (neither yellow nor green) tend to have very dissimilar RM values (far from diagonal). Figure 4 further identifies those variety pairs that belong to at least one pair of overlapping maximum cliques (white) versus those with no overlapping maximum cliques (blue). The former values we assume are missing due to incomplete observations and are imputed according to Step 4 in the PMC algorithms, and the latter values we assume are missing due to complete dissimilarity and are imputed according to Step 3 in the PMC algorithm. Values imputed as maximum dissimilarity are thus identified in blue in Figure 4, which makes intuitive sense given the large difference in RM values (furthest from diagonal), whereas most of the imputed values are calculated according to the triangulation approach.

After applying the PMC algorithm we now have a complete proximity matrix and can apply standard clustering algorithms to find clusters, that is, sets of soybean varieties with similar GxE interactions. As an illustration, we consider one such cluster obtained by again applying complete−link algorithm to the completed proximity matrix. This cluster includes seven varieties and Figure 5 shows a graphical representation of the yield of those seven varieties in three environments. The columns show yield above or below environmental average in three

environments (e1, e2, e3). Varieties in the same cluster are expected to have the same preference for environments (that is, the same GxE interactions). Here we have plotted the yield versus the environmental average, so values above zero indicate that the variety does better than average in this environment and vice versa for values below zero. With a few exceptions, these varieties that were clustered together do in fact have preference for the same environments (same GxE interactions). For example, varieties V1 and V3 yield above environment average in environment E1 and below average in environment E2.

This case study serves to illustrate how clustering with incomplete proximity matrices may arise in real applications. Furthermore, it serves to illustrate how data may be missing in the proximity matrix for different underlying reasons, including the two assumed by the PMC algorithm proposed in this paper. The purpose of this section is thus to demonstrate the need for the PMC algorithm to analyze real data, but a complete discussion of the value of identifying soybean varieties with same GxE interactions to decision making in a commercial breeding program is not possible within the scope of this paper. However, as noted before, within a subset of varieties in the same cluster the GxE interactions can now be ignored and a simple non−interaction model of yield will be appropriate. This allows us to better predict and compare yield, ultimately contributing to the decision as to which experimental varieties should be advanced.

2.7 Conclusions

We have presented the Proximity Matrix Completion (PMC) algorithm for imputing missing values before clustering when only the proximity matrix is available. The algorithm assumes that values are missing in the proximity matrix due to one of two underlying reason: incomplete observations or complete dissimilarity. This assumption is motivated by an application in plant breeding where the goal is to find clusters of experimental soybean varieties with the same genotype−by−environment (GxE) interactions. However, an extended PMC algorithm we present is general and could be applied to any scenario where we need to perform clustering and only an

incomplete proximity matrix is available. To further evaluate and provide insights into the effectiveness of the PMC algorithm, we also presented numerical results using the well−known iris data, where we simulated different types of missing−value structures. These results indicate that the PMC algorithm is in fact effective for a wide−range of missing value structures, and that it can be effectively extended to applications with very high percentage of missing data which common imputation methods are unable to deal with. One notable benefit of the PMC algorithm over benchmarks is treating every missing value as a unique case while benchmarks consider all missing values uniform, which affects the results specially in highly sparse datasets. In practice there are reasons why values are missing and PMC algorithm tries to identify those reasons by taking advantage of a graph formulation of the clustering. Specifically, identifying maximum cliques in the graph and doing triangulation on overlaps among these maximum cliques form the basis for identifying and treating difference cases of missing values in our approach. Being an imputation method the PMC algorithm does is independent to clustering method subsequently applied. As is shown is numerical results the PMC algorithm can be used with any clustering method that only requires the proximity matrix as input. However, as might be expected, its effectiveness is higher if the clustering method fits well with the structure of the data to be clustered.

2.8 References

(2000). *Using the Fractal Dimension to Cluster Datasets*. ACM.

Aeberhard, S., Coomans, D., and Olivierde, V. (1992). Comparative analysis of statistical pattern recognition methods in high dimensional settings. 27:1065−−1077.

Amir, B.-D., Ron, S., and Zohar, Y. (1999). Clustering gene expression patterns. 6:281−−297.

Baraldi, A. N. and Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1):5–37.

CHEN, C., DURAND, E., FORBES, F., and FRANÇOIS, O. (2007). Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes*, 7(5):747–756.

Coen, B. and Joep, K. (1973). finding all cliques of an undirected graph. 16:575−−577.

Daniel, N. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. pages 328−−362.

Das, S., Abraham, A., and Konar, A. (2008). Automatic kernel clustering with a multi-elitist particle swarm optimization algorithm. *Pattern Recognition Letters*, 29(5):688–699.

Des Marais, D. L., Hernandez, K. M., and Juenger, T. E. (2013). Genotype-by-environment interaction and plasticity: Exploring genomic responses of plants to the abiotic environment. *Annual Review of Ecology, Evolution, and Systematics*, 44(1):5–29.

Ellen, V. (1986). Implementing agglomerative hierarchic clustering algorithms for use in document retrieval.

Evett, I. W. and Spiehler, E. J. (1987). Rule induction in forensic science. pages 107−−118.

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

Garciarena, U. and Santana, R. (2017). An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Systems with Applications*, 89:52–65.

Güngör, E. and Özmen, A. (2017). Distance and density based clustering algorithm using gaussian kernel. *Expert Systems with Applications*, 69:10–20.

Jaemun, S., Ohbyung, K., and Kun, C. (2016). Adaptive pairing of classifier and imputation methods based on the characteristics of missing values in data sets. 46:485−−493.

Jian-Ping, M. and Lihui, C. (2014). Proximity−based k−partitions clustering with ranking. 41:7095−−7105.

Li, X., Guo, T., Mu, Q., Li, X., and Yu, J. (2018). Genomic and environmental determinants and their interplay underlying phenotypic plasticity. *Proceedings of the National Academy of Sciences*, 115(26):6679–6684.

Ng, R. T. and Han, J. (1994). Efficient and effective clustering methods for spatial data mining. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, page 144–155, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Shuiping, G., Xiong, Z., Yangyang, L., Cong, X., and Licheng, J. (2013). Multi−elitist immune clonal quantum clustering algorithm. 101:275−−289.

Xu, D. and Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193.

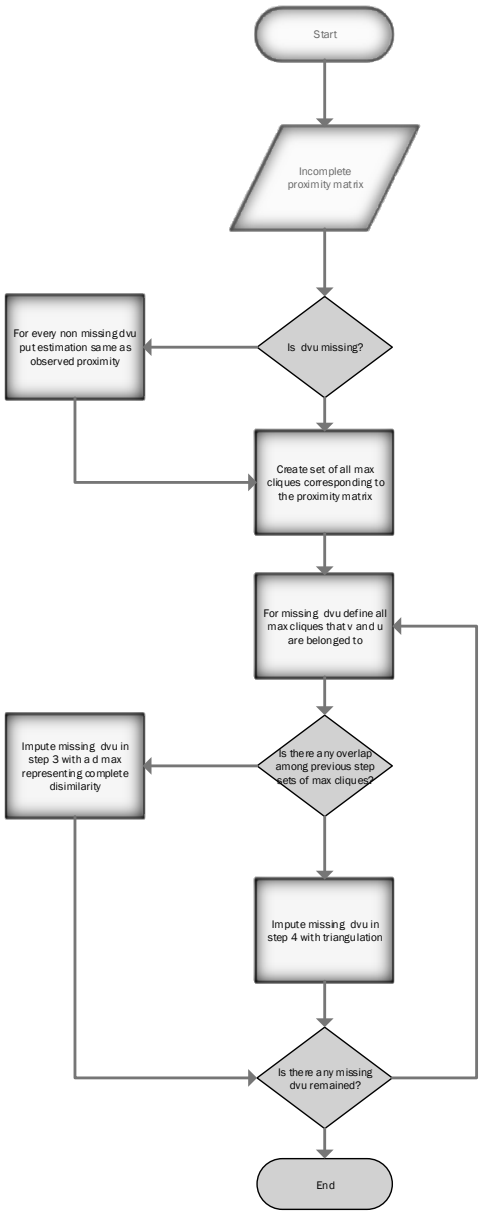

Figure 2.2 PMC flowchart

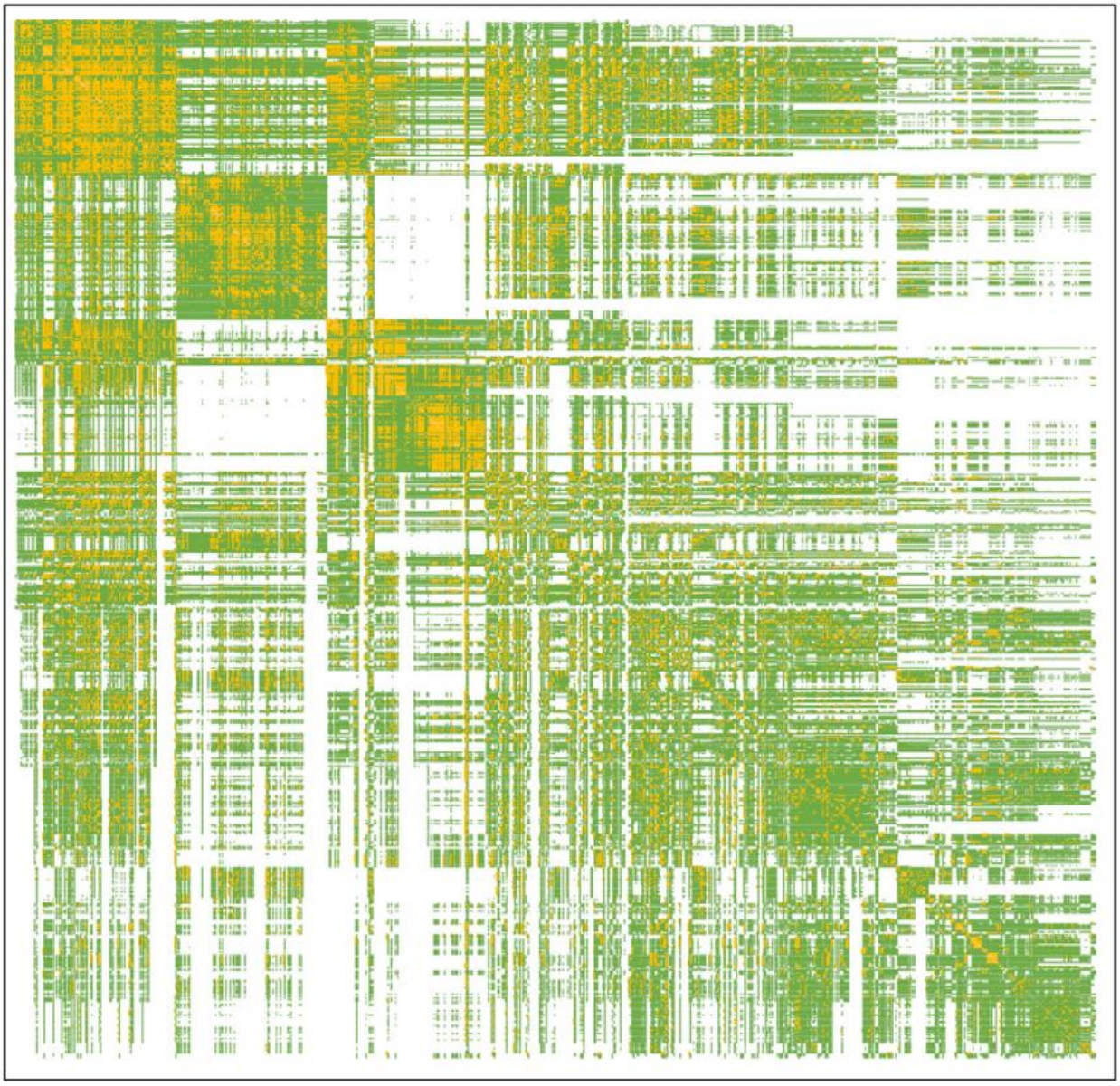

Figure 2.3 Sparsity of the initial proximity matrix for 1033 soybean varieties

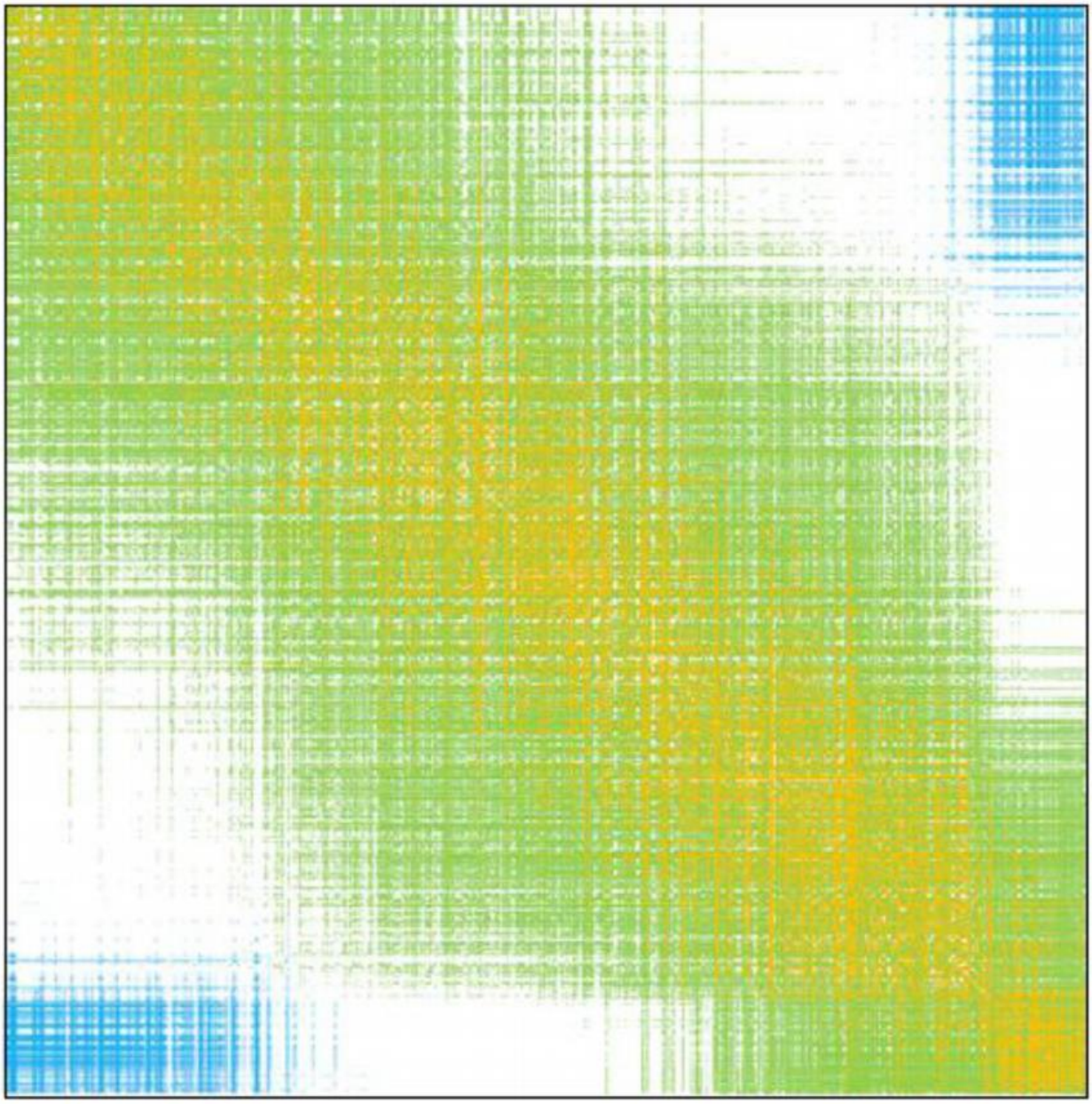

Figure 2.4 Graphical representation of sparsity with varieties reordered by rm value

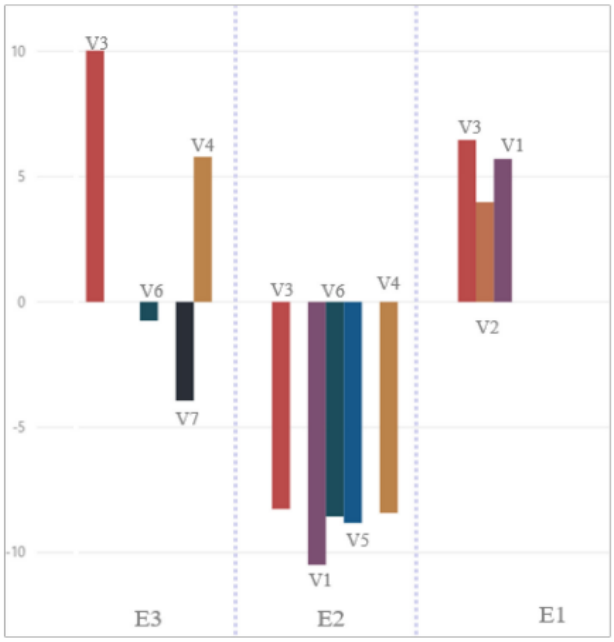

Figure 2.5 Yield values for an example cluster of seven varieties (v1 – v7)

CHAPTER 3. PREDICTING THE SIMILARITY OF GxE INTERACTIONS OF SOYBEAN VARIETIES USING GENETIC DATA

Samira Karimzadeh and Sigurdur Olafsson

Department of Industrial and Manufacturing Systems Engineering, Iowa State University

Abstract

In commercial plant breeding decisions regarding the advancement of a specific plant variety, that is, if the variety should continue within the program and be planted again the following year, is challenging due to very limited amounts of observations. To aid the decision making, machine learning can be used to predict yield of a seed variety based on past observations of multiple varieties, but this approached is challenged by the fact that yield is a function of genotype, environment and the genotype-by-environment (GxE) interaction effects. As a result, sufficiency of the past observations in target environments is crucial in building the prediction model. Where the observations of the target variety is not enough to train an accurate model, using observation of similar GxE varieties, that is those have same GxE behavior is a possible solution but identifying those similar GxE varieties is challenging as well. In this study, we have developed a supervised prediction model to identify similar GxE varieties using genetic data when there is no yield observation available to identify similar GxE varieties directly based on the observed GxE effects. The study shows how using similar GxE varieties versus non-similar GxE varieties affects yield prediction accuracy and benefits advancement decision making process in a case study.

3.1 Introduction

GxE interactions are important to plant breeding. Most work has focused on phenotypic stability and/or adaptabilitiy. Our work focuses on another aspect of GxE interactions, namely

GxE similarity, which is different but somewhat related to stability. The importance of selecting genotype based on their phenotypic stability has long been well understood in plant breeding. The analysis of phenotypic stability based on phenotype observations was originated by Finlay and Wilkinson (1963) and followed by Ribaut et al. (1996); Via et al. (1995). More recently several studies have investigated the genomic underpinnings of phenotypic stability, and methods for predicting phenotypic stability based on genotype data such as Kusmec et al. (2018) and Arnold et al. (2019). Such methods potentially allow for selecting genotypes based on their phenotypic stability at various stages of the breeding and experimental process.

To make the idea of phenotypic stability more concrete, consider a traditional linear model involving genetic effect, environmental effect, and genetic-by-environment (GxE) interaction effect Becker and Léon (1988).

$$y_{ij} = \mu + g_i + h_j + b_{ij} + \varepsilon_{ij} \tag{3.1}$$

In this equation the b_i factor represents the sensitivity of each genotype to the environmental effect, that is, to the quality of the environment in which it is grown. This model can also be rewritten in terms of the normalized phenotype, that is the observed phenotype minus the environmental mean, which eliminates the environmental effect from the equation above. The response thus only has two components: the genetic effect (G) and the genetic-by-environment (GxE) interaction effects.

$$\tilde{y}_{ij} = y_{ij} - (\mu + h_j) = g_i + b_i h_j + \varepsilon_{ij} \tag{3.2}$$

The objective of Finlay-Wilkinson analysis of phenotype stability is to estimate the b_i factor as genotype sensitivity to the environment quality. The lower the b_i is the more stable (less plastic)

genotype comes in terms of responding to environmental inputs. In other words, stability is a measure helps plant breeders to identify varieties which consistently over perform in target environments with low sensitivity to the environment quality Happ et al. (2021).

In this paper we address a related but slightly different objective. We investigate predicting groups of genotypes that behave similarly with respect to phenotypic stability, that is, groups of genotypes that have similar genetics-by-environment (GxE) effects. Thus, we consider a full-interaction linear model. The goal of this paper is to identify subsets of genotypes $I = I_1 \cup I_2 \cup ... \cup I_m$ such that $b_{i_1} \approx b_{i_2}, \forall i_1, i_2 \in I_k, \forall k$. The advantages of being able to make such predictions are considerable. Such information could be utilized as an input to creating trial groups for planting, and the analysis of observed field data is simplified by the knowledge that some genotypes have the same GxE effects, as it effectively implies that for the subsets of those genotypes there are no GxE interaction effects, only a main environmental effect. The ability to predict such subsets could thus be put into effective use by plant breeders.

Our motivation is thus that understanding GxE similarity helps understanding of phenotype such as yield. knowing that the yield of a genotype is a function of the genetics of the plant (G), the environment in which the plant is grown (E) and the way genetic of the variety as internal characteristics of the plant reactions to the external factors in a target environment (GxE effect), breeders need to observe a variety in a wide range of locations across the years to cover target environments precisely and make a reliable prediction of the yield. Collecting such data requires resources in terms of land and time that are simply not feasible in practice. In such situation, observations of the other varieties in past might become useful in training a prediction model on the yield. The key here is how to make observations of other varieties usable to predict yield of a target variety?

As mentioned before, yield of an individual is a function of genotype, environment and GxE interaction. Genotype of a variety determine efficiency of a plant to capture environmental inputs and convert it into biomass output. Also, environment represent quality of supplies such as nutrients, water, radiation etc. that a plant need to flourish. These two elements, genetic and

environment are measurable through historical data for every type of genetic and environment. That means yield of a target variety can be measured by expected yield of the variety across all environments, corrected by the quality of the environment and preference of the variety for a target environment. Thus, knowing expected yield of every variety we are able to measure difference between the quality of genetics interaction to the environment for each pair of two varieties those have been observed in common environments. Considering the fact that factors such as weather, soil quality, precipitation, etc. are fixed within a location in a particular year called environment, expected quality of an environment is measurable and the same for a pair of varieties in common environment. So finding any pairs of varieties that have same preference to the environment (GxE) in commonly observed environment can gives a variety interact similarly to the environment than its observation be used to expand knowledge of a target variety in unseen environments those the similar variety is planted. This preference to the commonly observed environments can be estimated as GxE dissimilarity having the two variety observed in enough number of environment in common.

3.2 Data

In this study, we use a commercial soybean breeding program data to build our prediction model. As we model the GxE interaction effect by genetic data we need genotypic information of varieties to use as predictor variables along with the phenotypic information to build an estimation on GxE similarity of variety pairs. So, two main datasets are used to build the training dataset. The phenotypic dataset includes over 430K phenotypic observations of nearly 2300 soybean varieties in 1700 environments from 2009 to 2017. This dataset contains three columns corresponding to genotype, environment and phenotype that means each observation in this dataset represents a genotype observed phenotype in a particular environment. The main purpose of this dataset is to build a source of GxE interaction effect to use as response variable in the prediction model. As mentioned before, GxE is the result of internal factors (Genetic) interaction to the external factors (Environment). So, to have an unbiased estimation of GxE

interaction effect, the sufficiency of phenotypic observation across the environments is essential. Figure 1 shows a histogram on number of environments our soybean varieties have been observed.

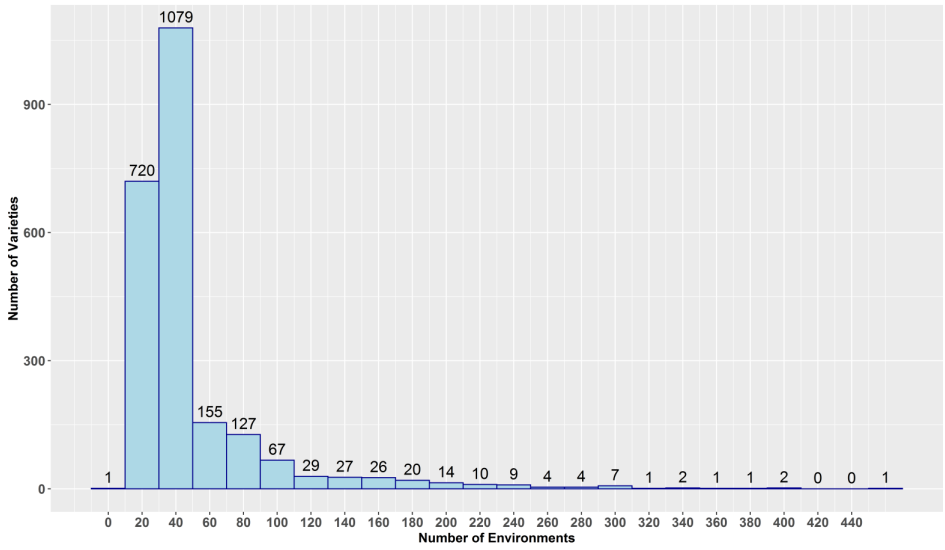

Figure 3.1 Histogram for number of planting environments

As is shown majority of varieties are observed in 20-40 environments which is statically a reasonable sample size for our estimation. Simply speaking by GxE interaction effect we are interested in having an estimation on how a genotype over perform the environment average using a random sample of environments. Thus, number of varieties are observed in each environment would help to get a better estimation on the quality of the environment. Figure 2 show a histogram on number of varieties are planted in each environment.

As is shown majority of environments cover more than 20 varieties that means we can get a fair estimation of environment average performance.

For genetic data we have used a SNP [1] genotyping array of 5602 genetic variables. For each of the soybean varieties used in this study a subset of genetic variables information is available. Decision on number and genetic variables to explore has made by breeders based on the available resources for genotyping and prior knowledge on the genotype. So the number and subset of

[1]Single Nucleotide Polymorphism

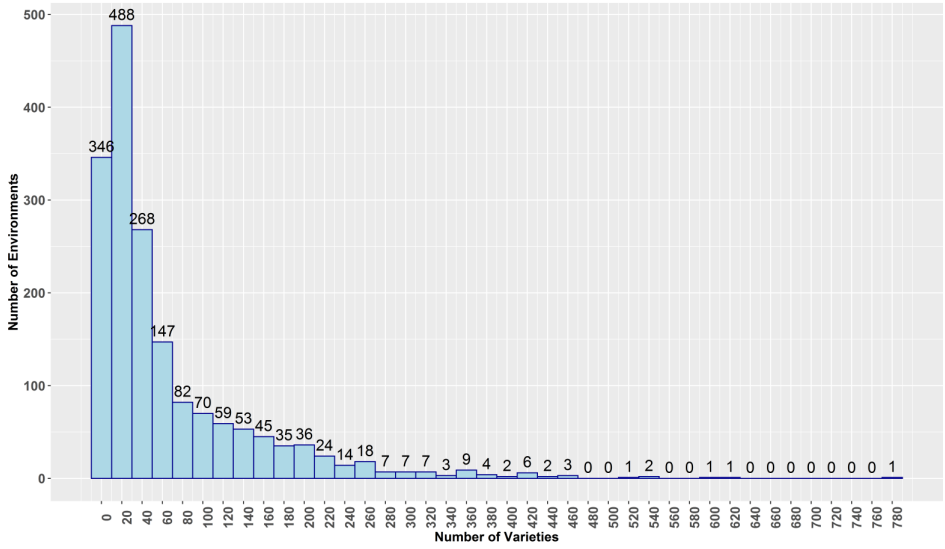

Figure 3.2 Histogram for number of planted varieties

genotyped variables could vary for each genotype. Study shows that a few thousands of genetic variables are sufficient in genome pool for crop plants to identify important adaptive genes Yun-Gyeong et al. (2015). Figure 3 shows histogram for number of genetic variables are genotyped in our varieties' set.

As is shown majority of varieties have around 2000 genetic variables genotyped and there are only few genotypes that are massively genotyped in recent years due to advancement in genotyping technologies.

3.3 Methods

In commercial plant breeding the main source of information for breeders in advancement process is the phenotypic observations of the varieties in experimental environments. Basically varieties those can outperform other varieties in the decision group specially previously commercialized varieties will be chosen to be advanced and planted for at least one more year. So the sufficiency of the phenotypic observation in terms of planting the target variety in wide range

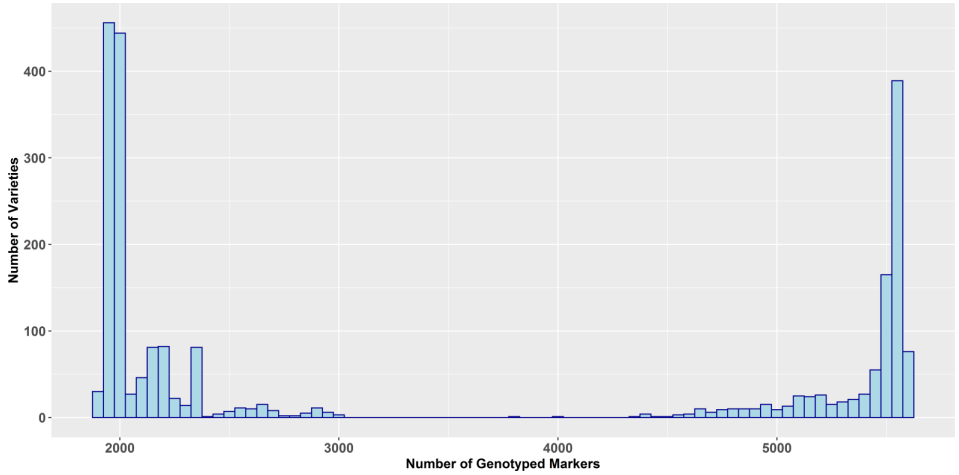

Figure 3.3 Histogram for number of genotyped genetic variables

of environments along with other experimental and commercial varieties is crucial and need huge amount of resources. Therefore, advancement decisions in early stages are based on limited number of observation that makes the decisions weak. Considering the fact that GxE is a function of genetic interaction to environment, identifying similar GxE varieties those have same GxE interaction effect to our target variety, we can use observation of the similar GxE varieties to estimate the target variety performance in unseen environment where the similar GxE variety has is planted. But in order to define such similar GxE varieties based on phenotypic observations we need the two variety have been observed in common environments repeatedly. We use the framework presented in Figure 4 to build an empirical source of GxE dissimilarity of two varieties using phenotypic observation of late stage varieties to put into prediction model with genetic data as known responses in order to predict GxE dissimilarity of two varieties in early stage that the phenotypic observations of a target variety is not sufficient to make advancement decision thus using observations of similar GxE varieties are essential to estimate phenotypic performance of the target variety.

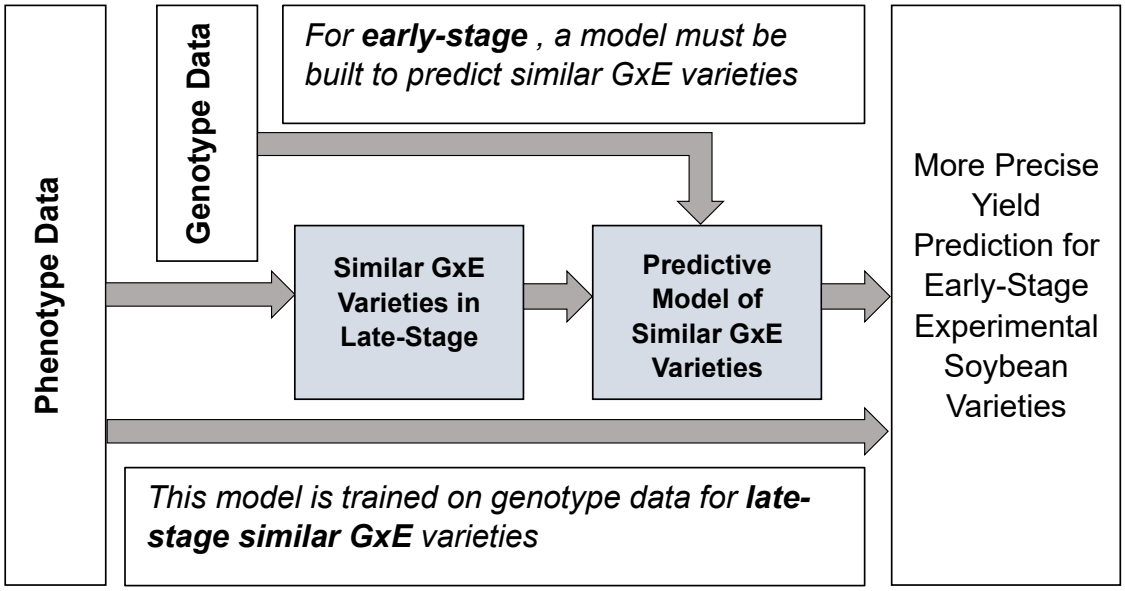

Figure 3.4 Prediction model framework

3.3.1 Data Labeling

In order to learn from the data described above, we first come up with appropriate labeling. In the labeling process understanding GxE similarity is important and a critical step. As noted before, plant phenotype is often modeled as function of genotypic effect g_i , environmental effect h_j , and the interaction effects between genotype and environment b_{ij} provided in equation (3.1) where μ denotes the overall average yield and ε_{ij} models the variability. We are interested in similarity of the interaction effects, that is, how similar $b_{i_1 j}$ and $b_{i_2 j}$ are across all environments for a pair of varieties (i_1, i_2) . If variety i_1 is tested (planted) in a set of environments $E(i_1)$ and i_2 is similarly tested in $E(i_2)$, where the set of common testing environments is non-empty, that is, $E(i_1, i_2) \equiv E(i_1) \cap E(i_2) \neq \varnothing$, then we can estimate this similarity (across common testing environments) as follow. Dropping the ε_{ij} , since we will be using empirically observed averages, we note that $y_{i_1 j} = y_{i_1 j} - (\mu + g_{i_1}) - h_j$, and h_j is fixed in a common testing environment $e \in E(i_1, i_2)$. Furthermore, $y_{i_1} \equiv \mu + g_{i_1}$ represents the average yield of the variety i_1 , and we

can write the difference in interaction effect between the two varieties in a fixed environment as $b_{i_1 j} - b_{i_2 j} = (y_{i_1 j} - g_{i1}) - (y_{i_2 j} - g_{i2})$. We can therefore calculate the estimated GxE dissimilarity as:

$$d(i_1, i_2) = \frac{1}{K} \frac{1}{|E(i_1, i_2)|} \sum_{e \in E(i_1, i_2)} |\left(\overline{y}_{i_1 j} - \overline{g}_{i_1}\right) - \left(\overline{y}_{i_2 j} - \overline{g}_{i_2}\right)|$$
$$K = \max_{i_1, i_2} \frac{\sum_{e \in E(i_1, i_2)} |\left(\overline{y}_{i_1 j} - \overline{g}_{i_1}\right) - \left(\overline{y}_{i_2 j} - \overline{g}_{i_2}\right)|}{|E(i_1, i_2)|} \tag{3.3}$$

Here $\overline{y}_{ij}$ is the observed average yield of variety i in environment j , and $\overline{g}_i$ is the observed overall average yield for variety i. We finally normalize the values to be between zero and one, by dividing by K that is resulting a proximity matrix

$$D = \begin{pmatrix} 0 & \dots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \dots & 0 \end{pmatrix}$$

where all the varieties have a distance to some other varieties in the same scale from zero to one that represents the most similar to the most dissimilar.

Since GxE effect is a response of genetic interaction to the environment, the similarity in genetic in a pair of varieties can be used to predict similarity/dissimilarity in GxE effect. So, by determining pairs of similar GxE varieties using the empirical measure discussed, we're able to train a predictive model on genetic data in common to predict similarity/dissimilarity of GxE interaction effect between two varieties. Simply speaking, a supervised learning model on commonality in genetic data can predict a pair of varieties similar GxE or dissimilar GxE as a binary response of the classification problem.

In order to build such genetic variables in a common training source, we make a three-class predictor model of genetic variables in common in which class (+1) correspond to have genetic variable m_j in common, class (-1) represents different values for variable m_j between the two varieties of a pair and class (0) is set for missing variable m_j in either variety of the pair. Finally, to make the response variable, we then use a cutoff of .25 to map the normalized dissimilarity

measured above to a binary response of relatedness. Any pair with dissimilarity estimation of less than or equal to 0.25 are labeled as similar GxE, and all pairs with dissimilarity measure above 0.25 are called dissimilar GxE.

3.3.2 Variable Selection

Building the genetic data, we have a training dataset, includes thousands of predictor variables to train the prediction model. As discussed in the data section for each genotype there is a subset of genetic variables available and that subset could vary. Also, from the genotypic information we know that most of the varieties are missing majority of genetic variables. Thus, to avoid over-fitting, a variable selection model could be used not only for increasing the quality of the prediction model but also for managing feasibility of the training process in terms of computation power and time. To identify the predictors, the model relies on more to make a prediction model; we have used variable importance methods of all types, from regression filter methods to wrapper methods and practical method such as information gain. Specifically, LASSO variable importance measure is used to cover filter methods, Relief is used from all wrapper methods and completeness of genetic information is used as a practical method for discarding predictor variables class (0) those are missing or have not been genotyped often for a practical reason. Using any of these three variable selection methods reduces the number of predictors considerably.

LASSO validated around 1400 variables with a non-zero score to use. Relief weight also gives a similar number of variables to use according to figure (4) which presents relief weights sorted descending. Approximately after 1500 variables the weights' plot gets almost flat.

But taking the information availability for variety pairs shown in figure (5) as the base for predictor variables to use, we'll get slightly more variables to use in the prediction model which is expected considering the common range for number of genotyped genetic variables.

Another challenge that should be addressed here is the class-imbalanced issue. In other words, in such prediction problems in which the positive class (here similar GxE variety pairs) has extremely fewer observation compared to the negative class (in our case 27% positive), there is a

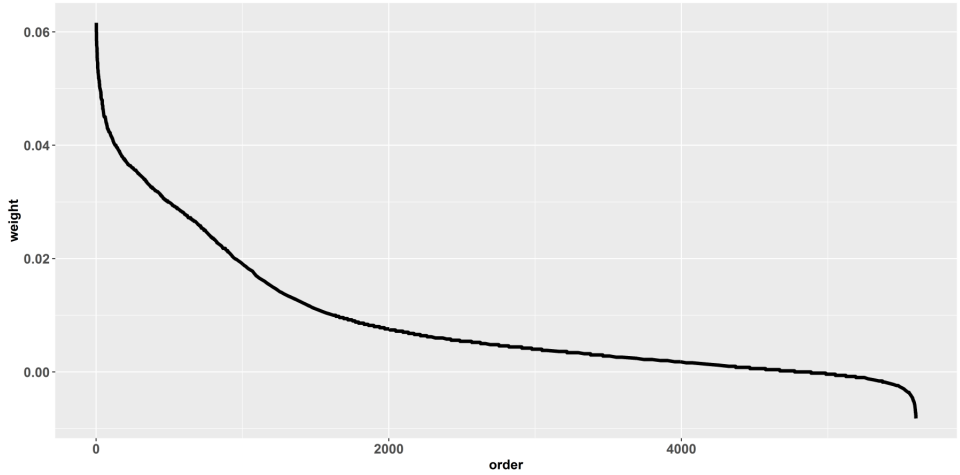

Figure 3.5 Ordered relief weights

natural tendency in prediction models to ignore the minority positive class to reach a better accuracy. While the objective of the prediction is to catch the minority class. We have used the synthetic minority class oversampling technique (SMOTE) to address the class-imbalance issue Chawla et al. (2002).

3.3.3 Predictive Models

The approach we have taken in this study is predictive modeling. Among all predictive models available in toolbox we have chosen some advanced models those are well known for predicting nonlinear response as we have some clue from the literature that GxE interaction effect is a complex nonlinear response of genetic interaction with the environment. LASSO regression, random forest and xg-boosted trees are the chosen models. Considering the scale of predictors we use, ensemble models such as random forest gives us the advantage of taking majority votes of several uncorrelated weak learners as a professional committee of predictions. Also, boosting methods such as xgboost could help in getting a better balance in bias-variance of the prediction where there is a high risk of over-fitting in weak learners. LASSO regression is also used to add the power of regularization and shrinkage as it is a popular model in analysing correlation and

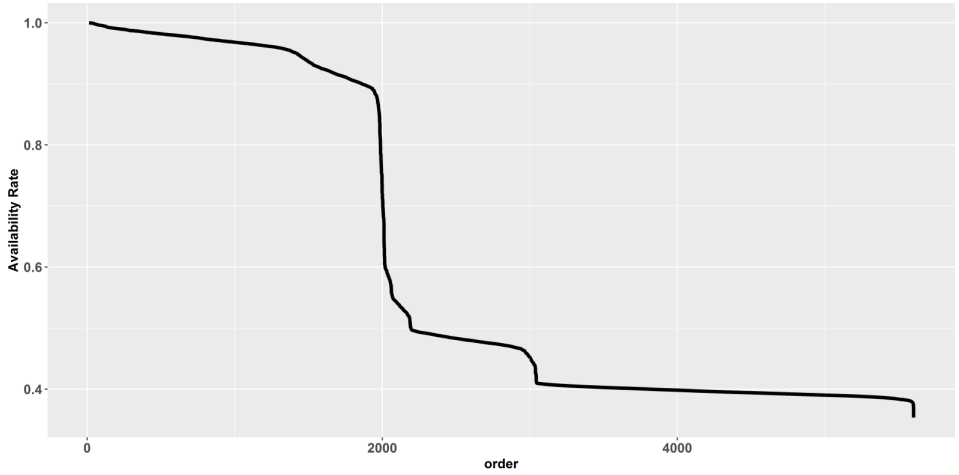

Figure 3.6 genome availability ordered by rate

linear dependencies in high dimensional genomic dataWang et al. (2018). Although there are lots of other advanced method could be used, computational power resource and tuning simplicity led us to these three models. The prediction model goal is to identify similar GxE varieties using genetic variables commonality information as predictor variables. Although relatedness is a binary response but we believe predicting probability of the posterior which means probability that the two variety being similar GxE varieties gives us some advantages over predicting the binary class. Using the probabilities let us to add an extra level to the tuning process by taking a probability cutoff that gives the best performance.

3.3.4 Model Validation

To build the prediction model, we need two components in the dataset. First, the genetic information of varieties which is provided as a set of 5602 genetic variables we use as predictors and GxE dissimilarity for every pair of the varieties as the response. For each pair of soybean varieties in our dataset, we can make an estimation of how similarly they interact with the environment using phenotypic observation of those in the environments that they have in common. That is called estimated GxE dissimilarity. For minimizing the randomness effect and

having a reliable estimation of the GxE we restrict the pairs to those that have at least 20 environments in common as that is a statically reasonable sample size for our estimation and is satisfied by most of the varieties of the study.

Furthermore, to have an unbiased estimation of error, we split the data into two sets for training and evaluation so that every target variety has no observation in the train set in neither side of a pair. Using this approach, we are able to evaluate the model performance for a new target variety that has never been observed before. We keep 20% of varieties for evaluation, which gives 190 over 18K variety pairs, and the rest of the 80% containing around 2100 target varieties and 104K pairs, for the train purpose.

3.4 Numerical Results

Addressing all the challenges in this prediction model, we evaluate our prediction model based on the model, applied feature selection method and re-sampling technique. Table 1 shows the classic supervised learning performance metrics for all the possible 24 setups.

Table 3.1 Prediction models' performance metrics

		Imbalanced data				SMOTE balanced			
Prediction Model	Feature Selection	Accuracy	Specificity	recall	rrecision	Accuracy	Specificity	Recall	Precision
LASSO Regression	None	0.7	0.79	0.31	0.24	0.6	0.61	0.52	0.24
	Completeness	0.75	0.86	0.22	0.27	0.57	0.58	0.5	0.26
	Relief	0.75	0.87	0.21	0.25	0.37	0.28	0.79	0.25
	LASSO	0.76	0.88	0.19	0.26	0.58	0.6	0.49	0.22
Random Forest	None	0.78	0.92	0.15	0.27	0.78	0.92	0.16	0.28
	Completeness	0.79	0.93	0.15	0.3	0.79	0.93	0.15	0.3
	Relief	0.78	0.92	0.16	0.29	0.78	0.92	0.16	0.29
	LASSO	0.79	0.93	0.15	0.31	0.79	0.93	0.16	0.32
Boosted Trees	None	0.78	0.92	0.15	0.27	0.78	0.92	0.15	0.28
	Completeness	0.75	0.87	0.19	0.24	0.78	0.92	0.16	0.29
	Relief	0.75	0.87	0.22	0.26	0.78	0.91	0.17	0.27
	LASSO	0.75	0.87	0.2	0.26	0.79	0.93	0.14	0.28

Considering only the overall accuracy of the models' performance are considerable. But bringing the problem complexity and main goal of the prediction, the models performance are reasonable. A desired model here is the one that have a high specificity and recall at the same

time which is very difficult to achieve. By high specificity we mean the model can identify dissimilar GxE varieties correctly as specificity is a metric represents the ration of truly caught dissimilar GxE variety pairs to all are predicted as dissimilar GxE. Thus, the more mistakes the model make in predicting a similar GxE varieties as dissimilar GxE the lower is the specificity. Checking the specificity numbers, most of the models are performing well in catching dissimilar GxE varieties. The best performance models got here is catching 93% of the dissimilar GxE varieties correctly. But the main purpose of developing these prediction models is to identify the similar GxE varieties which could be represented by recall and precision of the predictions. The higher the recall is means the less similar GxE varieties are missed as the recall is ratio of truly similar GxE variety pairs to all are labeled as similar GxE. As is shown the best recall performance is 79% obtained by Lasso Regression using SMOTE balanced data over the relief selected variables. But a high recall is not the only objective. Basically a model can improve the recall by deteriorating the specificity and predicting more cases as similar GxE. So the desired model is the one can precisely predict the similar GxE varieties. In other words the more truly similar GxE varieties we get over the all pairs predicted as similar GxE the higher the precision is. It concludes the winner here is random forest model that using LASSO variable selection and SMOTE balanced data as the base by catching 1/3 of the similar GxE variety pairs. However all the models performances are competing closely which is the absolute advantage of tuning the probability cutoff. In fact the random forest model can achieve almost the same performance without re-sampling that means using less computation power and time.

Getting back to the main motivation of the prediction model, We are looking for a model can identify varieties with similar GxE effects as a target variety to use observation of those varieties in estimating the performance of the target. To make sure the sets of predicted varieties with similar GxE effects are the best to use for estimating target variety performance, we evaluate all the prediction models using average MSPE of the prediction over 190 target varieties. To have a fair comparison of the MSPE numbers we pick all the varieties predicted as similar GxE using the tuned cutoff as the similar GxE set performance and all the varieties have probability of being

similar GxE greater than $(1 - c)$ as the dissimilar GxE set performance. Table 2 shows the average MSPE of predicting target phenotype using observation of the predicted similar GxE/dissimilar GxE set for 190 target soybeans.

Table 3.2 Average MSPE of yield prediction over 190 target variety

		Imbalanced data		SMOTE balanced	
Prediction Model	Feature Selection	Similar GxE	Dissimilar GxE	Similar GxE	Dissimilar GxE
LASSO Regression	None	8.07	11.04	8.32	9.48
	Completeness	8.19	10.59	8.02	9.79
	Relief	8.46	11.41	8.64	9.42
	LASSO	8.21	10.76	8.25	9.96
Random Forest	None	8.14	11.32	7.84	11.83
	Completeness	8.27	12.04	8.03	11.63
	Relief	8.17	12.24	8.21	10.14
	LASSO	8.11	12.09	8.05	11.96
Boosted Trees	None	8.42	10.93	NA	NA
	Completeness	8.49	9.1	8.39	9.07
	Relief	8.36	9.12	8.35	9.1
	LASSO	8.47	9.15	8.41	9.05

As is shown using observation of predicted similar GxE varieties has a considerable impact on predicting the yield in either scenarios. So, no matter how many similar GxE varieties we can catch using the prediction model, we need a model that is able to exclude the dissimilar GxE and avoid the randomness they can add to the yield prediction.

In other words, the similarity in GxE interaction effect for similar GxE varieties let to have a precise approximate of the target variety performance in unseen environment where the similar GxE is observed. So even having few number of similar GxE varieties takes advantage over averaging variance added by dissimilar GxE observations. Figure 7 represent a sample of GxE dissimilarity of a similar GxE and dissimilar GxE varieties to a target variety over the common environments.

As is shown for similar GxE varieties the GxE difference is low and stable across the common environments, while dissimilar GxE varieties have a bigger GxE difference with high variability. So using observation of similar GxE varieties will conclude a lower MSPE.

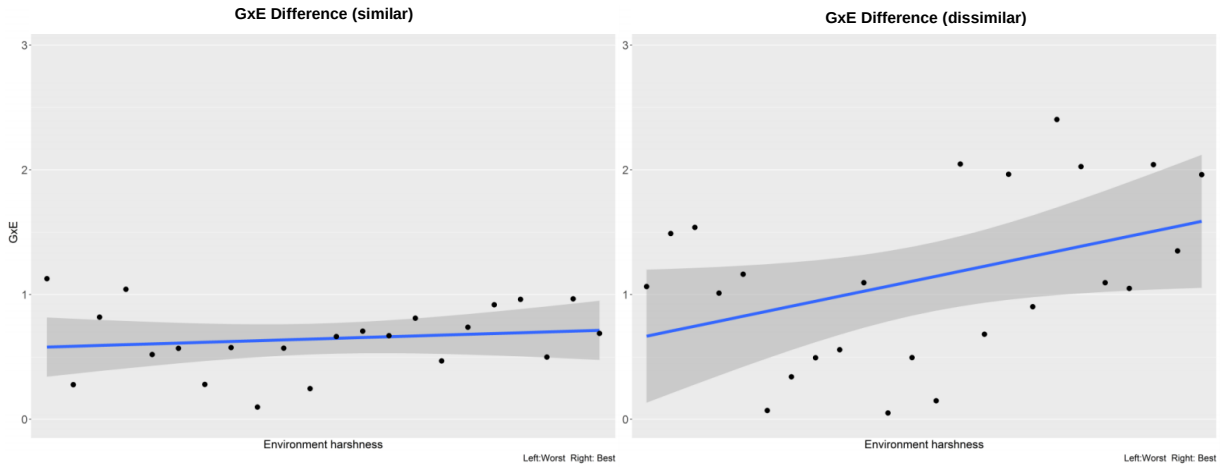

Figure 3.7 Similar/dissimilar GxE varieties absolute GxE difference to a target variety

3.5 Discussion

In this section we compare our empirical GxE dissimilarity model with some of the existing stability/adaptability models and discuss its contribution to the field. To start, we compare the classic model of phenotypic stability analysis by Finlay and Wilkinson (1963) with our estimation of a pair of varieties GxE dissimilarity. Following the Finlay-Wilkinson (FW) approach, we can estimate the stability of a phenotype in two steps: First estimate the environmental effect from a simple main-effect model $y_{ij} = \mu + g_i + h_j + \varepsilon_{ij}$ and then Substituting the estimate h_j into the main model of $y_{ij} = \mu + g_i + h_j + b_{ij} + \varepsilon_{ij}$ to estimate slopes b_i for each genotype. Figure 8 shows normalized absolute difference of a target variety estimated b_i (stability index) to the varieties stability that the target has been observed repeatedly compared to our empirical measure of normalized GxE dissimilarity.

Considering a cut-off of 0.25 for our normalized GxE dissimilarity measure to call two varieties similar GxE it is shown there is no strong correlation between the GxE dissimilarity and stability similarity. A correlation test gives sample correlation of 0.060 with 95% intervals $[0.45, 0.74]$ and

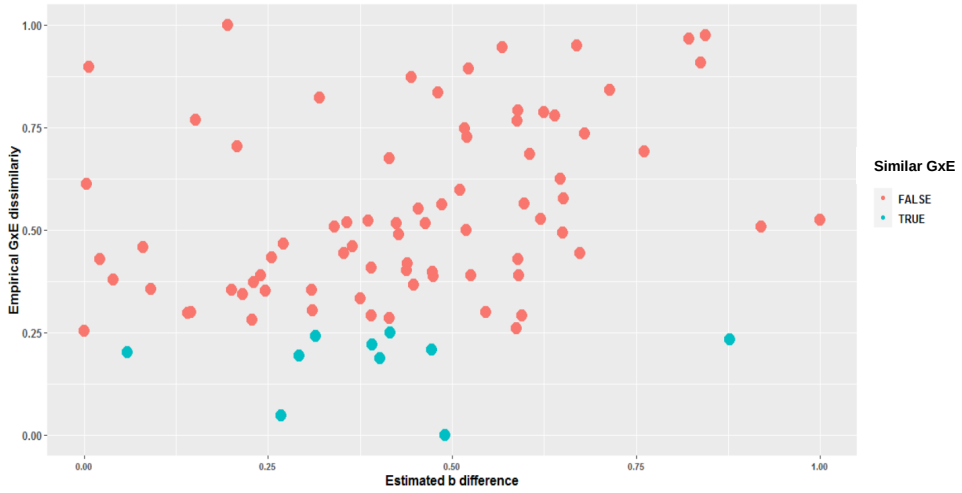

Figure 3.8 Correlation between GxE dissimilarity and stability

p-value of $5.049e-16$ for our sample of 18K variety pairs. That means two variety with different stability to the environment may show similar GxE interaction in variety of environments.

Next we compare the normalized GxE dissimilarity measure to genotype main effects and genotype × environment interaction effects (GGE) introduced by Yan and Tinker (2006).Figure 9 shows, GGE biplot for the same target variety discussed above besides the normalized GxE dissimilarity matrix introduced in this study.

The GxE matrix shows the GxE dissimilarities normalized by terget. Thus, the matrix is not symmetric. It means for a pair of varieties GxE dissimilarities might be slightly different and that is because the set of the other varieties each of the varieties if the pair have been observed frequently are different. For example, similar GxE variety 'RG5' is the most similar GxE variety to the target 'TG'. However, for the variety 'RG5', target 'TG' is the third most similar GxE after varieties 'RG2' and 'RG8'. It also can be observed that the most similar GxE variety to 'RG5' is not in this set of the varieties of the discussion.

Based on the biplot target variety 'TG' and similar GxE variety 'RG7' are not similar. But the GxE matrix says 'RG7' is the 6th similar GxE variety to the target variety 'TG' from the 10

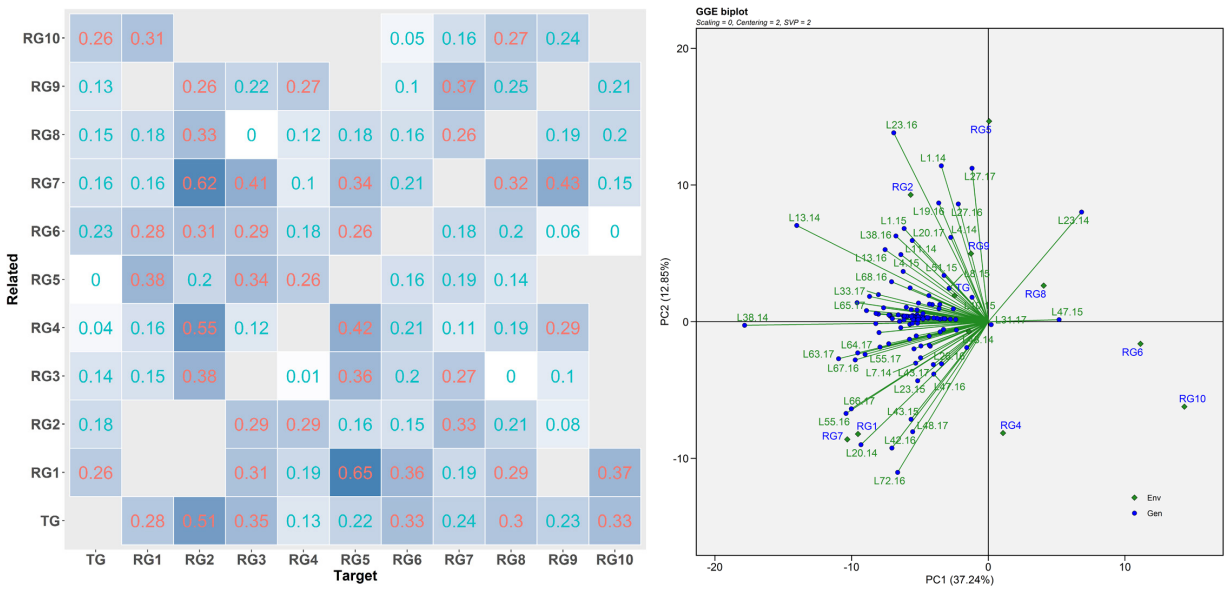

Figure 3.9 GGE biplot vs GxE dissimilarity matrix

empirically similar GxE varieties. Also, 'RG7' and 'RG1' are the most similar varieties based on the biplot which the dissimilarity matrix defines those as similar GxE too. This agreement is hold for most of the pair-wised comparisons of the biplot and the normalized GxE dissimilarity. Figure 10 shows detailed GxE interaction for target variety 'TG' and empirically similar GxE variety 'RG7' in common environments to address the reason for disagreement between the bipot and the empirical GxE measure. According to this detailed GxE interaction plot, the two varieties have same interaction to the environment as they both perform above or below the environment average phenotypic performance with a close magnitude. However they might perform a bit differently in poor environments but as it goes to more rich environments their interaction to the environment becomes identical.

To have a better understanding of this pair-wised comparisons of GxE a heatmap of GxE interaction effect is provided in Figure 11. The top heatmap shows the observed phenotype of the varieties in environments and the bottom heatmap shows the phenotypic difference to the environment average. In these heatmaps the x-axis of environments is sorted by the average

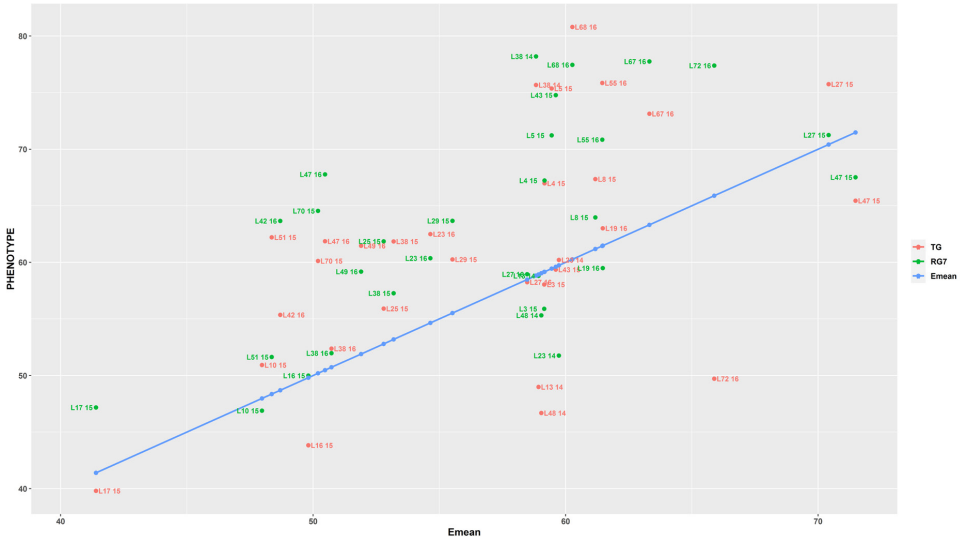

Figure 3.10 TG and RG7 detailed GxE

environment phenotype increasing from right(worst) to the left(best). It can be concluded from the heatmaps that this set of empirically similar GxE varieties are interacting similarly to the environments.

Figure 3.11 GxE dissimilarity Heat-map

Considering the advantage that similar GxE varieties can provide in predicting the yield (phenotype) of the target variety and having the prediction model, we can identify groups of similar GxE varieties. Identifying such groups let the breeders to plan where to plant each variety to maximize the information gain. One way of identifying such groups is to make the prediction

for similar GxE varieties using all the information is available and then cluster the varieties to similar GxE groups using a common similar GxE variety. Figure 12 show heat-map of available normalized GxE dissimilarity for a sample group that are considered similar GxE through a common predicted similar GxE variety.

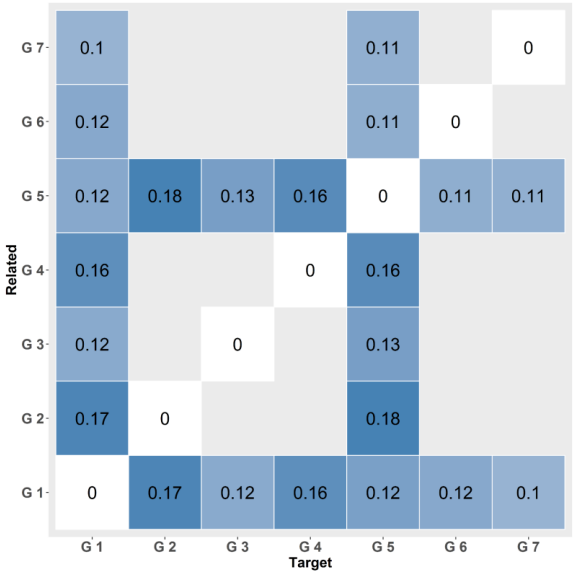

Figure 3.12 Heat-map of GxE dissimilarity: scale[0,1]

As is shown, most of the empirical GxE dissimilarities are missing due to practical limitations. Basically, clustering the varieties into groups of similar GxE varieties having only the phenotypic observations requires a complete matrix of empirical GxE dissimilarity which is not practical. In order to have a complete GxE dissimilarity matrix, we need to have each pair of varieties observed in common environments repeatedly. In practice it only happens for late stage experimental varieties to have been planted with other varieties repeatedly. Also those commonly observed varieties are usually commercialized varieties (e.g. G1,G5 in the heatmap) so breeders have a baseline that if an experimental variety is good enough to be advanced. Thus, measuring the GxE dissimilarity of experimental varieties empirically is not possible for most of the cases and imputation is required. But there are issues with regular imputation methods make them

ineffective for breeding context. In a recent study Karimzadeh and Olafsson (2019) address those issue and provide an imputation method for breeding context can estimate the reason a value is missing in the proximity matrix using graph theory and impute that based on the reason the value is missing.

3.6 Conclusion

This advance prediction of GxE similarity specially, in early stages can save time and money in different ways. Having knowledge of similar GxE varieties in advance, breeder can plan planting experiments in a way that maximizes information gain. Knowing that similar GxE varieties have same GxE preference to environment breeders can avoid planting similar GxE varieties in same environments and expand the experiment to more environments through similar GxE varieties. If the two varieties have exactly the same interaction effects between genotype and environment, yield can be modeled with a non-interaction model that predicts yield simply as a function of genotype main effects and environment main effects, which makes predicting yield in new environments simpler and provide information on variety performance in more target environments which is a key in making decision about advancing a variety or discarding it from the breeding process.

Also, a precise prediction of the yield of a target variety in early stages will provide a valuable source of information for breeders to make a better decision on future of a variety in breeding program. In other words, determining a set of similar GxE varieties those have a similar GxE interaction as the target and using observations of those we are able to predict yield of a target variety in some unseen environments and have a better evaluation of the target performance saves considerable amount of time and cost. Also, a better evaluation of a target performance using prediction can refuse an unwanted decision of keeping a future failure variety or discarding a potential winner variety based on few unreliable observations in early stages.

3.7 References

Arnold, P. A., Kruuk, L. E. B., and Nicotra, A. B. (2019). How to analyse plant phenotypic placticity in response to a changing climate. pages 1235–1241.

Becker, H. and Léon, J. (1988). Stability analysis in plant breeding. 101:1–23.

Chawla, N., Bowyer, K., Hall, L., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Finlay, K. and Wilkinson, G. (1963). The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research*, 14:742–754.

Happ, M. M., Graef, G. L., Wang, H., Howard, R., Posadas, L., and Hyten, D. L. (2021). Comparing a mixed model approach to traditional stability estimators for mapping genotype by environment interactions and yield stability in soybean. *Frontiers in Plant Science*, 12:542.

Karimzadeh, S. and Olafsson, S. (2019). Data clustering using proximity matrices with missing values. *Expert Systems with Applications*, 126:265 – 276.

Kusmec, A., de Leon, N., and Schnable, P. S. (2018). Harshnessing phenotypic placticity to improve maize yield.

Ribaut, J., Hoisington, D., Deutsch, J., and de Leon, C. D. J. G. (1996). Identification of quantitative trait loci under drought conditions in tropical maize. 1. flowering parameters and the anthesis-silking interval. *Theoretical and Applied Genetics volume*, pages 905–914.

Via, S., Gomulkiewicz, R., De Jong, G., Scheiner, S. M., Schlichting, C. D., and Van Tienderen, P. H. (1995). Adaptive phenotypic plasticity: consensus and controversy. *Trends in Ecology & Evolution*, 10(5):212–217.

Wang, H., Lengerich, B. J., Aragam, B., and Xing, E. P. (2018). Precision Lasso: accounting for correlations and linear dependencies in high-dimensional genomic data. *Bioinformatics*, 35(7):1181–1187.

Yan, W. and Tinker, N. (2006). Biplot analysis of multi-environment trial data: Principles and applications. *Canadian Journal of Plant Science*, 83:623–645.

Yun-Gyeong, L., Namhee, J., Ji Hong, K., Kwanghee, L., Kil Hyun, K., Ali, P., Bo-Keun, H., Sung-Taeg, K., Beom-Seok, P., Jung-Kyung, M., Namshin, K., and Soon-Chun, J. (2015). Development, validation and genetic analysis of a large. *the plant journal*, pages 625–636.

CHAPTER 4. OPTIMIZATION OF FIELD TRIALS FOR PLANT BREEDING

Samira Karimzadeh and Sigurdur Olafsson

Department of Industrial and Manufacturing Systems Engineering, Iowa State University

Under review in *Computer & Industrial Engineering*

Abstract

We describe a mathematical programming formulation and solution for a problem that occurs in commercial plant breeding operations. In such operations, a large number of experimental plant varieties are planted each year in different field trials, and assigning these trials to specific locations within each field is an important problem. What is of importance is not just an efficient solution that minimizes wasted space, but a solution that is favored by the plant breeders who oversee a set of such trials. This involves multiple considerations. First, it is advantageous to place trials involving plants that mature at the same time together. Second, each breeding group within the program prefers their own trials to be placed close together, and finally, trials that are in the same stage of the breeding process should ideally also be placed together. Specifically, early-stage trials where a very large number of experimental genotypes are planted in small plots should preferably be placed together, and late-stage trials where much fewer experimental genotypes planted in larger plots should be planted together. A mathematical programming formulation to successfully optimize trial locations should account for all three dimensions as well as make good use of space. The core idea of the formulation we present here is to split each field into blocks and then create blocks where the homogeneity is enforced via an objective function that penalizes deviations from a perfectly homogenous block. By favoring placing similar trials together in blocks, the model indirectly favors using fewer blocks, which in turn reduces wasted

space in the field. This optimization model was found to result in solutions that are a significant improvement of existing practices, both in terms of space utilization and in terms of having properties that are found desirable by plant breeders.

4.1 Introduction

Mathematical programming has been used in a wide range of agricultural applications over the years Mérel and Howitt (2014)Wang and Jiang (2012)Sarker and Quaddus (2002)Moeinizade et al. (2019), but there are still many areas where there are unexplored opportunities for taking advantage of mathematical programming and other industrial engineering methods to improve agricultural operations. Commercial plant breeding, in particular, involves operations ranging from the scheduling of planting equipment to the packing and warehousing of seed after harvest, where traditional industrial engineering methods could play a significant role but have not been heavily utilized. In this paper, we show how a mathematical programming approach can not just automate but significantly improve the solution quality of a difficult process within a commercial plant breeding operation, which until now had been performed without any such support, namely the planning of plant breeding trials in the field. This problem shares many features with other assignment problems that have been addressed extensively by the industrial engineering community, but also has some unique elements that derive from the application context.

4.1.1 Field Trial Placement

A commercial plant breeding program is a complex operation that stretches over multiple years. It typically starts with several tens of thousands of experimental genotypes (e.g., soybean varieties or corn hybrids) that are planted in what is termed an early-stage experiment and involves a small plot for each trial. At the end of the growing season, the most promising genotypes are selected for advancement; that is, they will be planted again the next year with the expectation that the best will eventually become commercialized. Thus, advanced genotypes are planted over multiple years, starting with early-stage experiments of tens of thousands of

genotypes planted in a few locations and small plots, extending to late-stage experiments where a few of the best performers are planted in a larger number of locations and bigger plots. Each experimental genotype is compared with others that will grow under similar conditions, primarily based on relative maturity (RM), that is, the number of growing days needed for the plant to mature. Furthermore, within a plant breeding program, there are typically multiple breeding groups, each responsible for making decisions regarding a portfolio of experimental genotypes. Trials from different breeding groups, in different stages and with varying RM must all be planted in the same field.

We consider optimization of field trial experiments from the perspective of the plant breeder. Namely, given a set of trials from multiple breeding groups, how should the trials be assigned to specific locations within a large field? The overall goal is improved utilization of each field, but each breeding group would also prefer their trials to be positioned close to each other for convenience, and trials that are similar in terms of stage and RM should also be placed together as much as possible as this makes them easier to compare and evaluate for possible advancement. In the end, a correct advancement decision is paramount and anything that can be done to facilitate the comparison between competing experimental genotypes, and would thus aid the advancement process, would be of great value. In fact, the benefit of good trial placement is likely higher than the monetary benefit of reduced waste, although the former is hard to quantify since it depends on ultimately selecting the best varieties and hybrids for commercialization.

4.1.2 Related Work

Field trial optimization has been addressed from the perspective of design of experiments Sarandon and Sarandon (1995) Zhang et al. (2019) Groot et al. (2012) but as far as we know not from the breeder perspective, that is, it has not been previously addressed from an operational perspective of assigning trials in a way that meets breeder preferences. However, other somewhat similar operational problems in agriculture have been addressed using mathematical programming and here we briefly mention the literature for some of those areas.

One of the critical operational problems is the planting and/or harvesting schedule. The time window a seed is planted and the resulting crop is harvested has a great impact on the amount and quality of the final product. To address this problem, Florentino, et al. implemented a mathematical programming model for selecting sugarcane varieties to be planted and an optimal schedule for planting and harvesting to maximize production in the sugarcane industry H.d.O et al. (2020). The routing of the relevant machinery is another area where optimization can improve agricultural operations. As mentioned above, both the planting and harvesting time is a key in maximizing plant yield. Needless to say, these two processes are mechanical and the machinery required for these operations is limited. Thus machinery routing is a critical issue in operation efficiency. To address this issue, Neungmatcha and Sethanan proposed a mixed integer model to optimize the transportation operation in sugarcane field Neungmatcha and Sethanan (2015). Jensen, et al. also have proposed a graph-search based path planning method for In-field and inter-field path planning transport of agricultural units to minimize the traveling distance and improve productivity of the whole system Jensen et al. (2012). The work of Lamsal, et al. also focuses on harvesting logistic systems and compares a variety of economically significant systems using two practical conditions: a multiple, independent producers and no on-farm storage Lamsal et al. (2016). There are many other relevant operational problems. For example, in one study Zhang and Gu attempt to maximize the economic profit through water resource allocation Zhang and Guo (2016). In another study, Zhang, et al. discuss the importance of water reservoir operation optimization Zhang et al. (2013), and finally, Solano, et al. modeled sustainability of the operation through minimizing the waste in agricultural operations Caicedo Solano et al. (2020).

4.2 Optimization Model

In this section we describe the optimization model we designed to improve field planting operations. While reducing wasted space is certainly an objective of this improvement effort, simply taking the utilization or wasted space as the objective function of an optimization formulation would likely lead to impractical solutions. Packing trials into fields too tightly may

lead to trials from the same breeding group being placed far apart, and dissimilar trials being placed together. This would lead to such an optimized trial layout being viewed unfavorably by the breeders, that is, the end-users of the planting layout, and hence be unlikely to be implemented. We therefore, develop an optimization formulation that focuses on desirable properties, that is, placing trials from the same breeding groups and with the same RM and stage together; while indirectly reducing wasted space in the field. The core idea of our formulation is to split each field into blocks, and then create homogenous blocks. By favoring placing similar trials together in blocks, the optimization formulation indirectly favors using fewer blocks, which in turn reduces wasted space in the field. Thus, the utilization of fields is improved while generating field layouts that are more desirable to the end-users (that is, the plant breeders) than existing field layouts.

4.2.1 Notation

The notation we are use for the formulation of this field layout problem is as follows. We let F denote the set of fields. There is a set B of breeders who need to plant trials, each trial belonging to one of a set S of stages and a set R of relative maturity groups. We let T denote this set of trials, and partition the trials into subsets T_{bsr} according to breeding group, stages and relative maturity. That is,

$$T = \bigcup_{b \in B} \bigcup_{s \in S} \bigcup_{r \in R} T_{bsr}.$$

A breeder $b \in B$ might have trials in different stages $s \in S$ or relative maturity groups $r \in R$. Each trial has some size c_i for $i \in T$; and each block has a capacity b_j that cannot be exceeded. There is a distance matrix D that defines the distance between blocks:

$$D = \begin{pmatrix} 0 & \dots & d_{1j} \\ \vdots & \ddots & \vdots \\ d_{j1} & \dots & 0 \end{pmatrix}, j = |F|.$$

Adjacent blocks will have distance of one, blocks separated by a single block will have distance of two and so forth. Distance between trials is taken as the distance between the block to

which they are assigned, which means that the distance between trials that are assigned to the same block is taken as zero.

The decision variables simply determines the block assignment for each of the trials, that is,

$$x_{ij} = \begin{cases} 1, & \text{if trial } \ i \text{ is assigned to block } \ j \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

We then define several additional binary variables to keep track of different breeding groups, stages and RM values as they are assigned to blocks.

$$\begin{aligned} y_{bj} &= 1 \text{ if breeder } b \text{ has trial}s \text{ assigned to block } j \\ z_{bij} &= 1 \text{ if breeder } b \text{ has trial}s \text{ assigned to blocks } i \text{ and } j \\ g_{sj} &= 1 \text{ if stage group } s \text{ has trial}s \text{ assigned to block } j \\ q_{sij} &= 1 \text{ if stage group } s \text{ has trial}s \text{ assigned to block } i \text{ and } j \\ u_{rj} &= 1 \text{ if RM group } r \text{ has trial}s \text{ assigned to block } j \\ v_{rij} &= 1 \text{ if RM group } r \text{ has trial}s \text{ assigned to blocks } i \text{ and } j \\ w_j &= 1 \text{ if block } j \text{ is used} \end{aligned}$$

This problem is essentially an assignment task. Each trial $i \in T$ will require certain amount of space c_i in the field $j \in F$ where it planted, and each field has some maximum capacity of $b_j, \forall j \in F$. The binary decision variables x_{ij} are defined to determined these paired assignments in a very straightforward manner. Specifically, $x_{ij} = 1$ if and only if trial t_i is assigned to field f_j. The remaining variables, namely y, z, g, q, u, v, are auxiliary variables to keep track of what may be considered the key component of the formulation, namely the penalty costs in the objective function. Those will be discussed in more detailed next.

4.2.2 Objective Function

The key to our formulation is to define penalties for assigning trials by the same breeding group, experimental stage, relative maturity (RM) far apart. We will also assign a small penalty for using each block, which will favor solutions to use as few blocks as possible. The motivation behind this approach is that it is not possible to obtain the ideal assignment where there is no

distance between any trials from the same breeding group, stage and RM. Thus, hard constraints would result in an infeasible formulation and assigning appropriately weighted penalties is a sensible approach that results in practical solutions. We use a single penalty weight for each type, defined as follows.

$$\begin{aligned} P_1 &= \text{penalty weight due to breeder,} \\ P_2 &= \text{penalty weight due to stage group,} \\ P_3 &= \text{penalty weight due to RM group,} \\ P_4 &= \text{penalty weight due to activating a field block.} \end{aligned}$$

These penalty weights will apply to both assigning trials to different blocks and to the distance between blocks when trials of the same type are assigned to different blocks.

Using the above penalty weights, the objective function measures the total cost of assigning trials in set T to available fields in set F using linear mathematical model as follow:

$$\begin{aligned} \min \quad & P_1 \left(\sum_{b=1}^{|B|} \sum_{i=1}^{|F|-1} \sum_{j=i+1}^{|F|} d_{ij} z_{bij} + M \sum_{b=1}^{|B|} \sum_{j=1}^{|F|} y_{bj} \right) \\ & + P_2 \left(\sum_{s=1}^{|S|} \sum_{i=1}^{|F|-1} \sum_{j=i+1}^{|F|} d_{ij} q_{sij} + M \sum_{s=1}^{|S|} \sum_{j=1}^{|F|} g_{sj} \right) \\ & + P_3 \left(\sum_{r=1}^{|R|} \sum_{i=1}^{|F|-1} \sum_{j=i+1}^{|F|} d_{ij} v_{rij} + M \sum_{r=1}^{|R|} \sum_{j=1}^{|F|} u_{rj} \right) \\ & + P_4 \sum_{j=1}^{|F|} w_j \end{aligned} \tag{4.2}$$

We observe that the objective function consists of four parts, corresponding to the penalties due to breeder, stage, RM and number of blocks. The first three parts have two components each, the first one corresponding to the distance between blocks if similar trials are assigned to different blocks, which we refer to as distance cost; and the second corresponding to not assigning such similar trials to the same block, which we refer to as occupation cost. Needless to say, the magnitude of these three penalty weights determines how this objective function prioritize the similarity of assignments for a set of trials. For each of the three groups let say breeders, there exist a binary variable z_{bij} that indicates if the breeder $b \in B$ has trial(s) in fields F_i and F_j to

consider in the total cost. Also, for every breeder $b \in B$ there exist a dummy cost of MP_1 where $M>>max(D)$ on scale of M for breeder b and field j using binary variable y_{bj} to avoid ghost assignments. Simply speaking, based on the magnitude of the penalty costs, the objective function might enable a ghost y_{bj} that fits the constraints with lower penalty where there is no actual trial(s) from group b assigned to the field j. The same structure exist for the stage and relative maturity groups.

Although the distance penalty constraints force the model to put related trials as close as possible, in some cases there could be unrelated trials fit into the same field block which the model put them in different fields because there is no direct penalty for using maximum capacity of each field. The last term in the objective function is intended to take care of maximizing each field capacity usage and avoid activating a field when there are some capacity in other field to assign trials without violating connections' criteria.

4.2.3 Constraints

There are only two physical constraint needed. The first such constraint ensures that capacity needed for the trials assigned to each field does not exceed the field capacity:

$$\sum_{i=1}^{|T|} c_i x_{ij} \leq b_j \quad \forall j \in \{1, ..., |F|\}. \tag{4.3}$$

The second set of physical constraints ensures that each trial is assigned to exactly one field:

$$\sum_{j=1}^{|F|} x_{ij} = 1 \quad \forall i \in \{1, ..., |T|\}. \tag{4.4}$$

In addition, there are technical constraints that are needed to ensure the correct penalties in the objective function:

$$\sum_{j=1}^{|F|} x_{ij} \leq |T_b| y_{bj} \qquad \forall b \in \{1, ..., |B|\} \quad \forall i \in T_b, \forall T_b \subseteq T \quad \forall j \in \{1, ..., |F|\} \tag{4.5}$$

$$y_{bi} + y_{bj} \leq 1 + Z_{bij} \qquad \forall b \in \{1, ..., |B|\} \quad \forall i \in \{1, ..., |F| - 1\} \quad \forall j \in \{i + 1, ..., |F|\} \tag{4.6}$$

$$\sum_{j=1}^{|F|} x_{ij} \leq |T_s| g_{sj} \qquad \forall s \in \{1, ..., |S|\} \quad \forall i \in T_s, \forall T_s \subseteq T \quad \forall j \in \{1, ..., |F|\} \tag{4.7}$$

$$g_{si} + g_{sj} \leq 1 + q_{sij} \qquad \forall s \in \{1, ..., |S|\} \quad \forall i \in \{1, ..., |F| - 1\} \quad \forall j \in \{i + 1, ..., |F|\} \tag{4.8}$$

$$\sum_{j=1}^{|F|} x_{ij} \leq |T_r| u_{rj} \qquad \forall r \in \{1, ..., |R|\} \quad \forall i \in T_r, \forall T_r \subseteq T \quad \forall j \in \{1, ..., |F|\} \tag{4.9}$$

$$u_{ri} + u_{rj} \leq 1 + v_{rij} \qquad \forall r \in \{1, ..., |R|\} \quad \forall i \in \{1, ..., |F| - 1\} \quad \forall j \in \{i + 1, ..., |F|\}. \tag{4.10}$$

The first set of constraints (4.5) forces the model to enable auxiliary variable y_{bj} if there exist trial(s) from breeder b assigned to the field j in order to penalize the objective function. This second set of constraints (4.6) forces the model to enable auxiliary variable z_{bij} if there exist trial(s) from breeder b assigned to the field i and j in order to penalize the objective function in

scale of the distance between the two fields. The third set of constraints (4.7) forces the model to enable auxiliary variable g_{sj} if there exist trial(s) from stage s assigned to the field j in order to penalize the objective function. The fourth set of constraints (4.8) forces the model to enable auxiliary variable q_{sij} if there exist trial(s) from stage s assigned to the field i and j in order to penalize the objective function in scale of the distance between the two fields. The fifth set of constraints (4.9) forces the model to enable auxiliary variable u_{rj} if there exist trial(s) from relative maturity group r assigned to the field j in order to penalize the objective function. The final set of constraints (4.10) forces the model to enable auxiliary variable v_{rij} if there exist trial(s) from relative maturity group r assigned to the field i and j in order to penalize the objective function in scale of the distance between the two fields.

With our approach, what might be thought of as constraints by the breeders (that is, placing trials in the same stage in the same blocks) are not formulated as constraints but rather incorporated into the objective function as penalties. However, in practice there are also some hard constraints to be considered. Such constraints can be of two types, either trials must be planted together because they need to be compared directly, or they cannot be planted together, for example if some trials should be sprayed with a chemical and others should not. Thus, the model must also include both inclusive and exclusive hard constraints. In the exclusive case that trials should not be planted in the same field we simply use the following constraint:

$$x_{ai} + x_{bi} \leq 1. \tag{4.11}$$

And for the inclusive case where trials must be planted in the same field, these two constraint must be added:

$$\begin{aligned} x_{bi} - x_{ai} &\leq 0 \\ x_{ai} - x_{bi} &\leq 0. \end{aligned} \tag{4.12}$$

Finally, we add a constraint with the purpose of eliminating solutions that use extra blocks. The main purpose of the model is to minimize the distance of related trials, resulting in a planting that is viewed favorably by breeders, an indirect goal of the optimization model is to assign trials

into the experimental fields in a way that minimizes wasted space. While creating homogeneous blocks tends to lead to fewer blocks, there are also some scenarios where equivalent solutions can be generated (in terms of the penalty values) by using an extra block that is not necessary. We therefore add the following constraint:

$$\sum_{b=1}^{|B|} y_{bj} \leq |B| w_j \quad \forall j \in \{1, ..., |F|\}. \tag{4.13}$$

This constraint makes sure that the penalty variable of using each field is activated in the objective function, and hence helps minimize the number of blocks used and reduce wasted space.

4.3 Case Study

The mathematical program formulated in the previous section was directly developed to address a scenario faced every year at multiple sites by a major commercial plant breeding program. To illustrate the use of the model, we present a case study involving a single field from this plant breeder. The starting point is a field assignment that was planted in a prior year and our results compare this actual implementation to the solution that would have been obtained using our formulation. The result reported here were obtained using academic license Gurobi solver for linear programming and the results were implemented by the company created by expert schedulers manually. We ran the model on a 64-bit windows 10 desktop computer with two 3.40GHz Intel(R) Core(TM) i7-4770 CPU, 8 processing core and 16.0GB RAM.

4.3.1 Data Description

The naming and plant location details of the trials are confidential but can be summarized as follows. A planting plan was implemented in a prior year that involved a total of 71 trials including 2 exclusive trials assigned to 10 blocks within a field. The trials are spread among 11 breeding groups, 7 stage groups and 3 RM groups. Figure 1 shows distribution of the trials in each group and we observe that one breeder supervised almost half of the trials, with the other spread among the remaining 10 groups. Most of the trials are in early-stage (stage 1-3), with

fewer larger late-stage trials (stage 4-7). Finally, most of the trials are the same RM groups with a few trials in further two groups. This will be typical for most planting locations since there is a natural match between the RM of the genotype being evaluated and the geographical location (primarily latitude) of the site of the experiment.

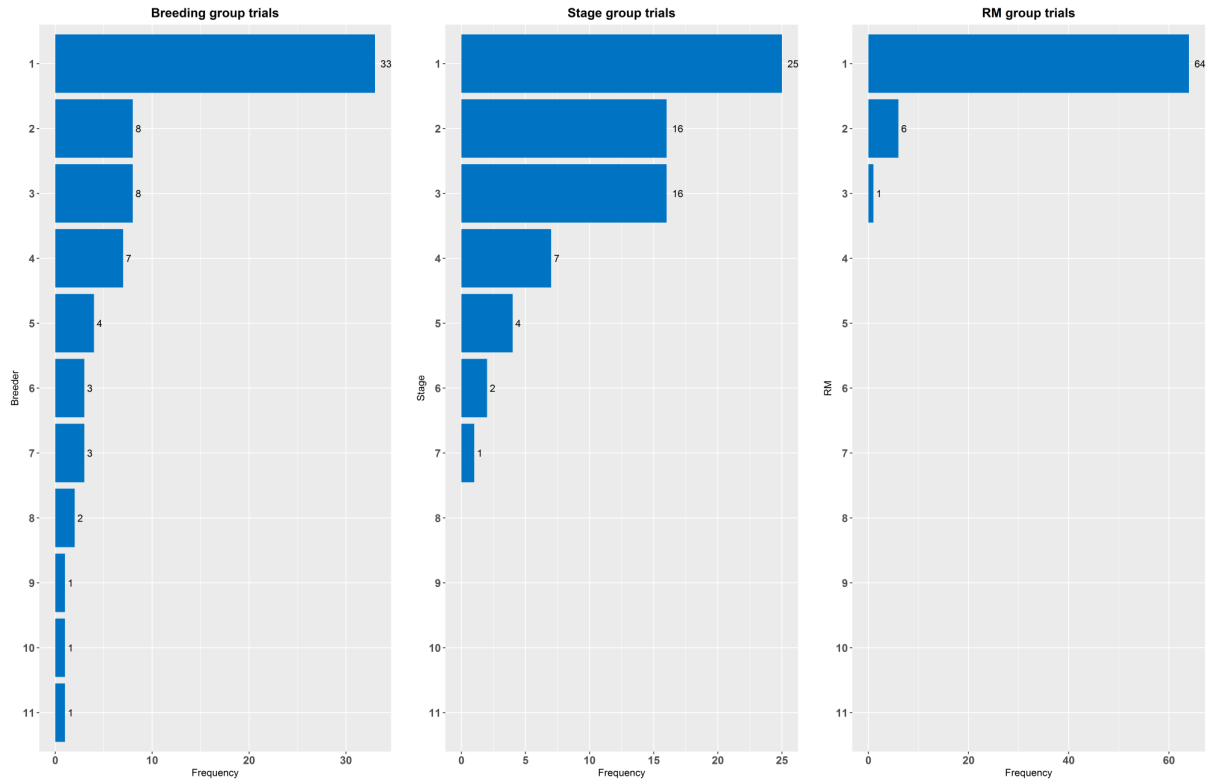

Figure 4.1 Breeder, stage and RM groups distribution of 71 trials

4.3.2 Model Parameters

As described in detail in Section 2 above, the goal is to place as many similar trials as possible together, while also improving field utilization. The solution will depend on the relative size of the penalty weights for different types of groups. For the implemented solution, those weights were set in consultation with experts as 30,50,20,1 for breeder, stage, RM and field activation,

respectively. Thus, placing similar stage trials together has the highest priority, followed by breeding group and RM value. The weight for the block activation is very small and thus essentially only makes a difference where there are multiple solutions that are identical except that one uses fewer blocks than the others. The importance order of stage, breeding group and RM, where given and the specific values were adjusted somewhat by looking at the desirability of solutions obtained via different weights. To determine the distances among the 10 blocks we use horizontal distance between the blocks. Since all the blocks in this case study are the same width we use the absolute difference of the IDs based on fields' geographical order which conclude the following distance matrix:

$$D = \begin{pmatrix} 0 & 1 & 2 & ... & 9 \\ 1 & 0 & 1 & ... & 8 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 9 & 8 & 7 & ... & 0 \end{pmatrix}.$$

When solving the optimization problem, a "big M" value of $M = 100$ is used in the formulation to avoid activating any variable y_j, g_j or u_j by the solver to reach a lower cost when the corresponding group does not have any trial(s) in block j. Mathematically speaking these three groups of binary variables are the dominant part of the cost function. When comparing optimized solutions obtained using different penalty values and the original solution that was constructed manually we substitute the $M = 100$ with $M = 1$ for a fair cost comparison between the solutions. We are thus able to measure the cost for any given solution x and objective coefficients C. In this approach, the number of fields each group has occupied would not dominate the total cost. In fact, the focus in this approach is on the distance cost of groups' assignment.

4.3.3 Comparison of Optimal and Original Solution

Table 1 shows both the total cost and a detailed costs' break down for the original and optimal solution. As is shown in this table, the optimal solution can considerably improve the total cost, with the objective function value of the manual solution originally implemented being

Table 4.1 Cost breakdown for original and optimal layout

	Original Layout	Optimal Layout	Original/Optimal
Breeder Cost	**2220**	**570**	**3.89**
Occupation	660	420	1.57
Distance	1560	150	10.40
Stage Cost	**2100**	**550**	**3.82**
Occupation	750	450	1.67
Distance	1350	100	13.50
RM Cost	**3520**	**540**	**6.52**
Occupation	140	140	1.00
Distance	3380	400	8.45
Field Cost	**10**	**5**	**2.00**
Occupation	10	5	2.00
Total Cost	**7850**	**1665**	**4.71**

nearly four times that of the optimal solution. The largest improvement is with respect to the RM penalty. However, improvements in all areas are considerable. Using this optimization model, breeders are able to put the trials in the fields in a way that is more favorable from all three aspects, and with the optimal solution using only five blocks versus the ten used in the original solution, wasted space within the field is significantly reduced. To provide some insights into how the mathematical programming approach improves the original layout, we show the details of the two layouts in Figure 2.

As indicated in this figure the original layout providing a reasonable assignment for each of the three objectives. Trials of each group are assigned to adjacent blocks as much as possible and there are numerous similarities between the two solutions. For example, both solutions assign all 16 trials in stage group S2 to a single block, and all trials belonging to breeding group B6 to a single block. However, there are also important differences, such as stage group S3 being spread across five blocks in the original solution but is combined in a single block in the optimal solution. Similarly, breeding group B11 has trials assigned in four blocks in the original solution but has all its trials in a single block in the optimal solution. Furthermore, the optimal layout obtained by the mathematical model provides a more homogeneous assignment in a way that no group has trials in more than two adjacent block unless a hard constraint or capacity limit force it. The

Figure 4.2 Original vs. optimal layout trial assignments to blocks

		Original Layout										Optimal Layout									
Field Blocks		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Stage Group	**S1**	22							3											17	8
	S2								16												16
	S3			4	3	4	2	3											16		
	S4		6		1														7		
	S5		1																1		
	S6									1	1						1	1			
	S7		2	2															4		
Breeding Group	**B1**				1														1		
	B2		2	1															3		
	B3			2	1														3		
	B4		1																1		
	B5				2														2		
	B6								8												8
	B7					2	1	1											4		
	B8		1																1		
	B9		5	1						1	1						1	1	6		
	B10	22							11											17	16
	B11			2		2	1	2											7		
RM Group	**R1**		2	1	3														6		
	R2		1																1		
	R3	22	6	5	1	4	2	3	19	1	1						1	1	21	17	24

intuitive reason for how such major improvement between two good enough solutions is possible is that considering number of trials and all the pair-wised relationships among the groups it is too difficult for the expert schedulers to manually try all possible solutions to get the minimum cost. Checking the details of the assignment, nearly all the gaps in original layout are caused by stage since the stage homogeneity is the priority for the company.

4.3.4 Sensitivity Analysis

As previously stated the penalty weights used in the above solution were obtained with input for experts. If the solution is extremely sensitive to small changes in those penalties it would limit

its practical use due to the potential need to tune the penalties for each field, which would not be practical. In order to evaluate sensitivity of the model to penalty weights we run the model with different set of penalties. In the first three experiment we put the focus on stage, breeder and RM respectively by giving a higher weight to that group. Experiment 1 is the main experiment that is discussed above. Then we solve the model for equal weights in experiment four and for the next three experiments, we run the model for optimizing only one of the three groups breeder/stage/RM at the time. Table 2 shows the cost breakdowns and gained improvements for these experiments.

Table 4.2 Cost breakdown for different penalty weights

	***Experiment 1** Stage Focus	**Experiment 2** Breeder Focus	**Experiment 3** RM Focus	**Experiment 4** Equal Focus	**Experiment 5** Breeder Only	**Experiment 6** Stage Only	**Experiment 7** RM Only
Penalty Costs							
P1 (Breeder)	30	**50**	30	30	**100**	0	0
P2 (Stage)	**50**	30	20	30	0	**100**	0
P3 (RM)	20	20	**50**	30	0	0	**100**
P4 (Field)	1	1	1	1	1	1	1
Original / Optimal Cost Ratio							
Breeder Cost	**3.89**	**3.22**	**3.89**	**2.74**	**3.89**		
Occupation	1.57	1.57	1.57	3.14	1.57		
Distance	10.40	5.78	10.40	2.60	10.40		
Stage Cost	**3.82**	**3.82**	**3.82**	**3.82**		**3.82**	
Occupation	1.67	1.67	1.67	1.67		1.67	
Distance	13.50	13.50	13.50	13.50		13.50	
RM Cost	**6.52**	**6.78**	**6.78**	**6.78**			**6.78**
Occupation	1.00	2.00	2.00	2.00			2.00
Distance	8.45	8.45	8.45	8.45			8.45
Field Cost	**2.00**	**2.00**	**2.00**	**2.00**	**2.00**	**2.00**	**2.00**
Occupation	2.00	2.00	2.00	2.00	2.00	2.00	2.00
Total Cost	**4.71**	**4.26**	**5.70**	**4.59**	**3.89**	**3.81**	**6.77**
Time to Solve (Seconds)							
	59	22	64	47	4	5	6

The results show that while some of the solutions differ, the solution is not very sensitive to how the penalties are set. All the solutions use five blocks and all the solutions have the same stage cost relative to the original solution. The other penalties components differ for some of the solutions but all are similar. Optimizing only one of the objectives at the time, the model makes the most improvement in RM cost followed by breeder and stage cost respectively. Also, comparing the improvements obtained by the optimization model in first four experiment shows RM is the main area to improve.

4.4 Conclusion

We have modeled the assignment of experimental planting trials within a field as a mathematical program. This formulation aims not just to reduce wasted space in the field, but to create planting plans that have properties that are viewed favorably by the breeders. In particular, the assignment aims to place trials from the same breeding group, trial stage and relative maturing (RM) close together. We demonstrate the application of this model via a case study that optimizes an actual field trial assignment that was planted in a previous year. The results show that the optimized solution reduces waste, has more similar trials placed together, and is robust with respect to exact values of the parameters needed by the mathematical program. The model can be solved reasonably fast and thus multiple fields can be solved within practical time constraints. This is important since breeders will have large number of fields. For example, the example field described in Section 3 is one of three fields at this site and the breeder has tens of similar sites in multiple countries. Future work will focus on optimizing multiple fields simultaneously by both assigning trials to fields and within each field.

4.5 References

Caicedo Solano, N. E., García Llinás, G. A., and Montoya-Torres, J. R. (2020). Towards the integration of lean principles and optimization for agricultural production systems: a conceptual review proposition. *Journal of the Science of Food and Agriculture*, 100(2):453–464.

Groot, J. C., Oomen, G. J., and Rossing, W. A. (2012). Multi-objective optimization and design of farming systems. *Agricultural Systems*, 110:63 – 77.

H.d.O, F., Jones, D., AdeIrawan, C., DjamilaOuelhadj, BanafeshKhosravi, and R.Cantane, D. (2020). An optimization model for combined selecting, planting and harvesting sugarcane varieties. *Annals of Operations Research*.

Jensen, M. A. F., Bochtis, D., Sørensen, C. G., Blas, M. R., and Lykkegaard, K. L. (2012). In-field and inter-field path planning for agricultural transport units. *Computers & Industrial Engineering*, 63(4):1054 – 1061.

Lamsal, K., Jones, P. C., and Thomas, B. W. (2016). Harvest logistics in agricultural systems with multiple, independent producers and no on-farm storage. *Computers & Industrial Engineering*, 91:129 – 138.

Moeinizade, S., Hu, G., Wang, L., and Schnable, P. S. (2019). Optimizing selection and mating in genomic selection with a look-ahead approach: An operations research framework. *G3 (Bethesda, Md.)*, 9(7):2123–2133.

Mérel, P. and Howitt, R. (2014). Theory and application of positive mathematical programming in agriculture and the environment. *Annual Review of Resource Economics*, 6(1):451–470.

Neungmatcha, W. and Sethanan, K. (2015). Optimal mechanical harvester route planning for sugarcane field operations using particle swarm optimization. *KKU Engineering Journal*, 42.

Sarandon, S. J. and Sarandon, R. (1995). Mixture of cultivars: Pilot field trial of an ecological alternative to improve production or quality of wheat (triticum aestivum). *Journal of Applied Ecology*, 32(2):288–294.

Sarker, R. and Quaddus, M. (2002). Modelling a nationwide crop planning problem using a multiple criteria decision making tool. *Computers & Industrial Engineering*, 42(2):541 – 553.

Wang, Y.-M. and Jiang, P. (2012). Alternative mixed integer linear programming models for identifying the most efficient decision making unit in data envelopment analysis. *Computers & Industrial Engineering*, 62(2):546 – 553.

Zhang, D. and Guo, P. (2016). Integrated agriculture water management optimization model for water saving potential analysis. *Agricultural Water Management*, 170:5 – 19.

Zhang, Y., Wang, X., Xu, F., Song, T., Du, H., Gui, Y., Xu, M., Cao, Y., Dang, X., Rensing, C., Zhang, J., and Xu, W. (2019). Combining irrigation scheme and phosphorous application levels for grain yield and their impacts on rhizosphere microbial communities of two rice varieties in a field trial. *Journal of Agricultural and Food Chemistry*, 67(38):10577–10586. PMID: 31490682.

Zhang, Z., Zhang, S., Wang, Y., Jiang, Y., and Wang, H. (2013). Use of parallel deterministic dynamic programming and hierarchical adaptive genetic algorithm for reservoir operation optimization. *Computers & Industrial Engineering*, 65(2):310 – 321.

CHAPTER 5. SIMULTANEOUS FIELD ASSIGNMENT AND FIELD OPTIMIZATION OF PLANT TRIALS

Samira Karimzadeh and Sigurdur Olafsson

Department of Industrial and Manufacturing Systems Engineering, Iowa State University

Abstract

Typical operation starts with several tens of thousands of experimental genotypes (e.g., soybean varieties or corn hybrids) that are planted in what is termed an early-stage experiment. At the end of the growing season, the most promising genotypes are selected for advancement. That is, they will be planted the next year again with the expectation that the best will eventually become commercialized. Each experimental genotype is compared with others that will grow under similar conditions, primarily based on relative maturity (RM), that is, the number of growing days needed for the plant to mature. Furthermore, within a plant breeding program, there are typically multiple breeding groups, each responsible for making decisions regarding a portfolio of experimental genotypes. Thus, a correct advancement decision is paramount and anything that can be done to facilitate the comparison between competing experimental genotypes, and would thus aid the advancement process, would be of great value.

We deploy a two-phased model to consider optimization of trials assignment from perspective of the breeding program. Namely, given a set of trials from multiple breeding groups, how should the trials be assigned to specific locations (Fields) and arranged within each field? The overall goal is improved utilization of each field and keep similar trials as close as possible for operational convenience. However, the benefit of good trial placement is likely higher than the monetary benefit of reduced waste and operational costs, although the former is hard to quantify since it depends on ultimately selecting the best varieties and hybrids for commercialization.

5.1 Introduction

We aim to advance mathematical formulation of an operation in plant breeding from scalability point of view for practical purpose. In this operation a large number of experimental plant varieties are planted each year in different field trials through a breeding program. The baseline solution provides a mathematical formulation that minimizes the waste space along with arranging trials within a field in a way that is the most favorable for the breeding program from three aspects. First, it places trials involving plants that mature at the same time together. Second, each breeding group within the program prefers their own trials to be placed close together, and finally, trials that are in the same stage of the breeding process should ideally also be placed together. There are limitations with the baseline solution from practical point of view. First, a commercial breeding program includes multiple fields which trials should be placed, and splitting trials among those fields requires same considerations as arranging trials within a field. Also, there are other practical considerations that make a field suitable for a specific group of trials or force the program to avoid a particular field for some trials. Another limitation with the baseline formulation is that the formulation is limited to three mentioned objectives. and these criteria may change from one year to the next as practices within the breeding program evolve. For instance there could be management practices like spraying method that requires trials of same spraying system to be placed nearby for operational purpose. So the model requires some level of flexibility in terms of objectives to adjust to the program.

So the baseline formulation is not applicable to the entire breeding program. To address these issues we propose a two phase solution illustrated in figure 1 for assigning trials to the field and arrange those within a field. The two-phase approach provides an advancement on the existing solution to help with scalability and applicability in practice.

In the two-phase approach we consider each field a big block of a master field and solve the optimization model with 90% capacity of fields. This proportional capacity usage is a way to guarantees phase two feasibility when we arrange trials within a field considering the field blocks. Also if there are trials need to be place alone in a field block, we break each of fields into $n + 1$

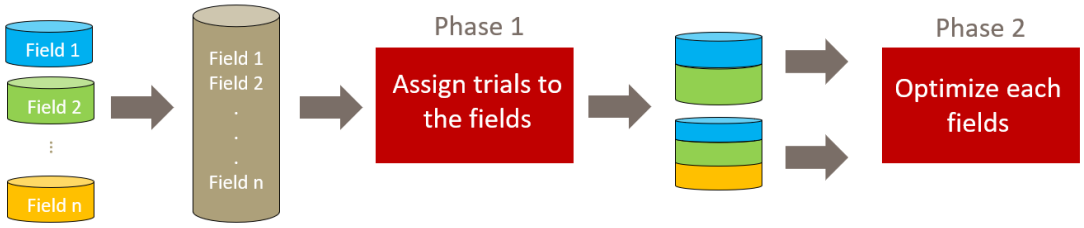

Figure 5.1 Two-phase approach for optimization.

blocks based on the field blocks capacity which n is number of standalone trials, and use a large distance for distance between blocks of a field to other fields to assure optimality of the two-phase solution. Needless to say that practical constraints of usability of each field for different group of trials in phase one can be assured by hard constraints. After obtaining phase one solution we run the model for each of the fields with full capacity based on the assigned trials of the phase one.

5.2 Optimization Model

In this section we describe the scalable optimization model we designed to improve field planting operations. We use the model described in Chapter (4) of this dissertation as the baseline formulation and try to modify that to reach scalabilty along with couple of improvement in hard constraints to speed up the model.

5.2.1 Notation

The notation we use for the formulation of this trial assignment problem is as follows. We let F denote the set of fields. Each trial belongs to different sets G_k consideration groups need to plant trials as close as possible, preferably in same field. We let T denote this set of trials, and partition the trials into subsets T_{g_k} according to breeding group, stages, relative maturity and any other consideration group. That is,

$$T = \bigcup_{s \in G_k} T_s \quad \forall k \in \{1, 2, ..., k\}.$$

There is a distance matrix D that defines the distance between field's blocks:

$$D = \begin{pmatrix} 0 & ... & d_{1j} \\ \vdots & \ddots & \vdots \\ d_{j1} & ... & 0 \end{pmatrix}, j = |F|.$$

Adjacent blocks will have distance of one, blocks separated by a single block will have distance of two and so forth. Distance between trials is taken as the distance between the block to which they are assigned, which means that the distance between trials that are assigned to the same block is taken as zero.

The decision variables simply determines the block assignment for each of the trials, that is,

$$x_{ij} = \begin{cases} 1, & \text{if trial } \ i \text{ is assigned to block } \ j \\ 0, & \text{otherwise.} \end{cases} \tag{5.1}$$

We then define two types of additional binary variables to keep track of different consideration groups are assigned to blocks.

$$\begin{aligned} y_{sj} &= 1 \text{ if group } s \in G_k \text{ has trials assigned to block } j \\ z_{sij} &= 1 \text{ if group } s \in G_k \text{ has trials assigned to blocks } i \text{ and } j \end{aligned}$$

We also consider a binary variable to keep track of used field blocks to penalize in objective function for minimizing space usage.

$$w_j = 1 \text{ if block } j \text{ is used}$$

This problem is essentially an assignment task. Each trial $i \in T$ will require certain amount of space c_i in the field $j \in F$ where it planted, and each field has some maximum capacity of $b_j, \forall j \in F$. The binary decision variables x_{ij} are defined to determined these paired assignments in a very straightforward manner. Specifically, $x_{ij} = 1$ if and only if trial t_i is assigned to field f_j. The remaining variables are auxiliary variables to keep track of what may be considered the key component of the formulation, namely the penalty costs in the objective function. Those will be discussed in more detailed next.

5.2.2 Objective Function

The key to our formulation is to define penalties for assigning trials by the same consideration group far apart. We will also assign a small penalty for using each block, which will favor solutions to use as few blocks as possible. The motivation behind this approach is that it is not possible to obtain the ideal assignment where there is no distance between any trials from the same consideration group. Thus, hard constraints would result in an infeasible formulation and assigning appropriately weighted penalties is a sensible approach that results in practical solutions. We use a single penalty weight for each type, defined as follows.

$$\begin{aligned} P_g &= \text{penalty weight due to consideration group } G_k, \\ P_f &= \text{penalty weight due to activating a field block.} \end{aligned}$$

These penalty weights will apply to both assigning trials to different blocks and to the distance between blocks when trials of the same type are assigned to different blocks.

Using the above penalty weights, the objective function measures the total cost of assigning trials in set T to available fields in set F using linear mathematical model as follow:

$$\begin{aligned} \min \quad & \sum_{g=1}^{k} P_g \left(\sum_{s=1}^{|G_k|} \sum_{i=1}^{|F|-1} \sum_{j=i+1}^{|F|} d_{ij} z_{sij} + M \sum_{s=1}^{|G_k|} \sum_{j=1}^{|F|} y_{sj} \right) \\ & + P_f \sum_{j=1}^{|F|} w_j \end{aligned} \tag{5.2}$$

We observe that the objective function consists of $k+1$ parts, corresponding to the penalties due to consideration groups $\{1, 2, ..., k\}$ and number of blocks. The first k parts have two components each, the first one corresponding to the distance between blocks if similar trials are assigned to different blocks, which we refer to as distance cost; and the second corresponding to not assigning such similar trials to the same block, which we refer to as occupation cost. Needless to say, the magnitude of these penalty weights determines how this objective function prioritize the similarity of assignments for a set of trials. For each of the groups let say breeders, there exist a binary variable z_{sij} that indicates if the breeder $s \in G_k$ has trial(s) in fields F_i and F_j to consider in the total cost. Also, for every breeder $s \in G_k$ there exist a dummy cost of MP_g where

M>>$max(D)$ on scale of M for breeder s and field j using binary variable y_{sj} to avoid ghost assignments. Simply speaking, based on the magnitude of the penalty costs, the objective function might enable a ghost y_{sj} that fits the constraints with lower penalty where there is no actual trial(s) from subgroup s in consideration group G_k assigned to the field j.

Although the distance penalty constraints force the model to put related trials as close as possible, in some cases there could be unrelated trials fit into the same field block which the model put them in different fields because there is no direct penalty for using maximum capacity of each field. The last term in the objective function is intended to take care of maximizing each field capacity usage and avoid activating a field when there are some capacity in other field to assign trials without violating connections' criteria.

5.2.3 Constraints

There are only two physical constraint needed. The first such constraint ensures that capacity needed for the trials assigned to each field does not exceed the field capacity:

$$\sum_{i=1}^{|T|} c_i x_{ij} \leq b_j \quad \forall j \in \{1, ..., |F|\}. \tag{5.3}$$

The second set of physical constraints ensures that each trial is assigned to exactly one field:

$$\sum_{j=1}^{|F|} x_{ij} = 1 \quad \forall i \in \{1, ..., |T|\}. \tag{5.4}$$

In addition, there are technical constraints that are needed to ensure the correct penalties in the objective function:

$$\sum_{j=1}^{|F|} x_{ij} \leq |T_s| y_{sj} \qquad \begin{array}{l} \forall s \in \{1, ..., |G_k|\} \\ \forall i \in T_s, \forall T_s \subseteq T \\ \forall j \in \{1, ..., |F|\} \end{array} \tag{5.5}$$

$$y_{si} + y_{sj} \leq 1 + Z_{sij} \qquad \begin{array}{l} \forall s \in \{1, ..., |G_k|\} \\ \forall i \in \{1, ..., |F| - 1\} \\ \forall j \in \{i + 1, ..., |F|\} \end{array} \tag{5.6}$$

The first set of constraints (5.5) forces the model to enable auxiliary variable y_{sj} if there exist trial(s) from subgroup s of consideration group G_k assigned to the field j in order to penalize the objective function. This second set of constraints (5.6) forces the model to enable auxiliary variable z_{sij} if there exist trial(s) from subgroup s of consideration group G_k assigned to the field i and j in order to penalize the objective function in scale of the distance between the two fields.

Finally, we add a constraint with the purpose of eliminating solutions that use extra blocks. The main purpose of the model is to minimize the distance of related trials, resulting in a planting that is viewed favorably by breeding program, an indirect goal of the optimization model is to assign trials into the experimental fields in a way that minimizes wasted space. While creating homogeneous blocks tends to lead to fewer blocks, there are also some scenarios where equivalent solutions can be generated (in terms of the penalty values) by using an extra block that is not necessary. We therefore add the following constraint:

$$\sum_{i=1}^{|T|} x_{ij} \leq |T| w_j \quad \forall j \in \{1, ..., |F|\}. \tag{5.7}$$

This constraint makes sure that the penalty variable of using each field is activated in the objective function, and hence helps minimize the number of blocks used and reduce wasted space.

With our approach, what might be thought of as constraints by the consideration groups (that is, placing trials belong to the same breeders, stage, RM group , etc. in the same blocks) are not formulated as constraints but rather incorporated into the objective function as penalties. In other words, these technical constraint are responsible to enable auxiliary variables corresponding to penalty costs in objective function. However, in practice there are also some hard constraints to be considered. Such constraints can be of two types, either trials must be planted together because they need to be compared directly, or they cannot be planted together, for example if some trials should be sprayed with a chemical and others should not. Thus, the model must also include both inclusive and exclusive hard constraints.

For exclusive case that trials of two groups T_1 and T_2 such $T_1, T_2 \subset T$ and $T_1 \cap T_2 = \emptyset$ cannot be planted together we use an auxiliary variable h_j :

$$h_j = \begin{cases} 1, & \text{if trial(s) from } T_1 \text{ is assigned to block } j \\ 0, & \text{otherwise.} \end{cases} \tag{5.8}$$

and two hard constraints:

$$\sum_{i \in T_1} x_{ij} \leq |T_1| h_j \quad \forall j \in \{1, ..., |F|\} \tag{5.9}$$

$$\sum_{i \in T_2} x_{ij} \leq |T_2|(1 - h_j) \quad \forall j \in \{1, ..., |F|\} \tag{5.10}$$

needless to say that decision variables h_j are only for controlling exclusive trials hard constraint and there should be no penalty for those in objective function.

And for the inclusive case that trials of two groups T_1 and T_2 such $T_1, T_2 \subset T$ and $T_1 \cap T_2 = \emptyset$ must be planted in the same field, no extra auxiliary variable is needed and these two constraints must be added:

$$x_{bj} - x_{aj} \leq 0 \quad \forall a \in T_1, \forall b \in T_2, \forall j \in \{1, ..., |F|\} \tag{5.11}$$

$$x_{aj} - x_{bj} \leq 0 \quad \forall a \in T_1, \forall b \in T_2, \forall j \in \{1, ..., |F|\} \tag{5.12}$$

5.3 Case Study

The mathematical program formulated in the previous section was directly developed to address a scenario faced every year at multiple sites by a major commercial plant breeding program. To illustrate the use of the model, we present a case study involving four fields from this breeding program. The starting point is a field assignment that was planted in a prior year and our results compare this actual implementation to the solution that would have been obtained using our approach. The result reported here were obtained using academic license Gurobi solver for linear programming and the results were implemented by the company created by expert schedulers manually. We ran the model on a 64-bit windows 10 desktop computer with two 3.40GHz Intel(R) Core(TM) i7-4770 CPU, 8 processing core and 16.0GB RAM.

5.3.1 Data Description

The naming and plant location details of the trials are confidential but can be summarized as follows. The planting plan was implemented in a prior year that involved a total of 547 trials assigned to 4 fields and 39 blocks within the fields. The trials are spread among 25 breeding groups, 8 stage groups and 3 RM groups. Figure 2 shows the distribution of trials within each of the three consideration groups. As is shown most of the trials are the same RM groups with a few trials in further two groups. This will be typical for most planting locations since there is a natural match between the RM of the genotype being evaluated and the geographical location (primarily latitude) of the site of the experiment.

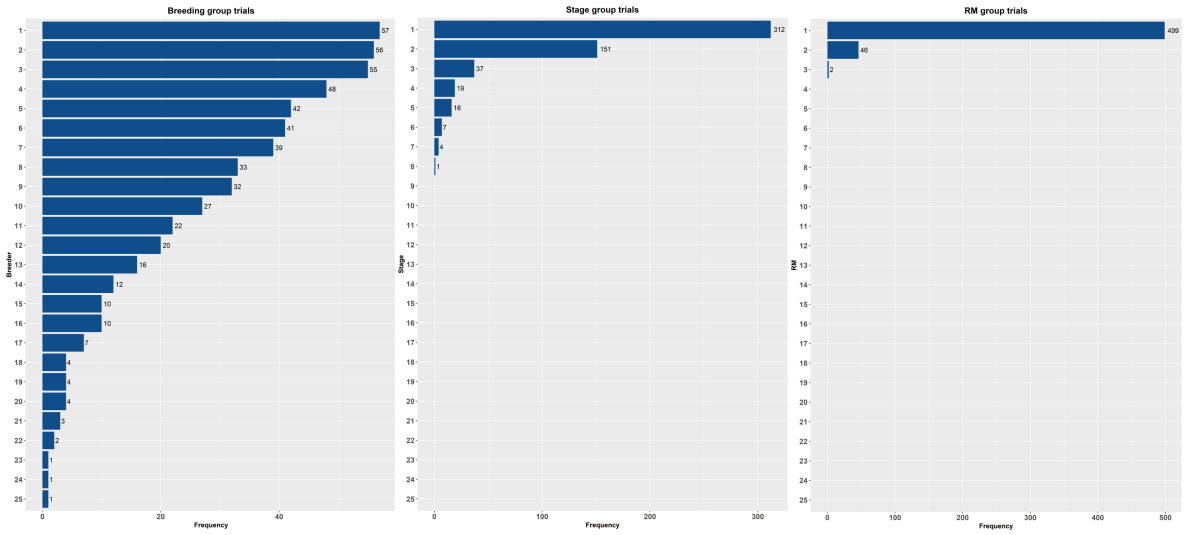

Figure 5.2 Distribution of the trials for breeder, stage and RM groups

5.3.2 Model Parameters

As described in detail in optimization model section, the goal is to place as many similar trials as possible together, while also improving fields utilization. The solution will depend on the relative size of the penalty weights for different types of groups. For the implemented solution, those weights were set in consultation with experts as 30,50,20,1 for breeder, stage, RM and field

activation, respectively. Thus, placing similar stage trials together has the highest priority, followed by breeding group and RM value. The weight for the block activation is very small and thus essentially only makes a difference where there are multiple solutions that are identical except that one uses fewer blocks than the others. The importance order of stage, breeding group and RM, where given and the specific values were adjusted somewhat by looking at the desirability of solutions obtained via different weights.

When solving the optimization problem, a "big M" value of $M = 1000$ is used in the formulation to avoid activating any occupation variable y_{sj}, by the solver to reach a lower cost when the corresponding group does not have any trial(s) in block j. Mathematically speaking these occupation binary variables are the dominant part of the cost function. When comparing optimized solutions obtained using different penalty values and the original solution that was constructed manually we substitute the $M = 1000$ with $M = 1$ for a fair cost comparison between the solutions. We are thus able to measure the cost for any given solution x and objective coefficients C. In this approach, the number of fields each group has occupied would not dominate the total cost. In fact, the focus in this approach is on the distance cost of groups' assignment.

5.3.3 Comparison of optimal and original solution

Although optimizing the arrangement of the trials within a field can provide considerable advantage over manual arrangement, the two-phase optimization approach can minimize the distance cost for all of the consideration groups for the entire breeding program. Table 1 shows the improvement for breeding groups:

As is shown, using the two-phase approach 6 out of 9 breeding group those have trials in multiple fields are able to move all trials in a single field. Although for some breeder such as breeder B2 having roughly 30% of trials in one field and 70% in could be a reasonable assignment but breeders B14 and B18 having all trials in one field and a single trial in another field is not favorable at all. So this sort of re-arrangement between the fields can provide a more favorable

Table 5.1 Two-phase optimization improvements for breeders

Breeding	Original Layout				Two-phase Layout			
group	Field 1	Field 2	Field 3	Field 4	Field 1	Field 2	Field 3	Field 4
B1	35			22	35		22	
B2	**40**			**16**	**56**			
B3		33		22			38	17
B4	**31**			**17**	**48**			
B5	**42**				**41**		**1**	
B6			41					41
B7		**8**		**31**				**39**
B8				33				33
B9			32				32	
B10				27				27
B11				22				22
B12	20				20			
B13	16				16			
B14		**1**		**11**				**12**
B15	10				10			
B16		**8**	**2**			**10**		
B17		7				7		
B18		**3**	**1**			**4**		
B19				4			4	
B20		4				4		
B21		3				3		
B22		2				2		
B23		1				1		
B24				1				1
B25		1				1		

assignment and reduce the operational costs considerably. This improvement also holds for stage groups as well. Table 2 shows the two-phase approach improvement for stage groups;

Table 5.2 Two-phase optimization improvements for stage groups

Stage	Original Layout				Two-phase Layout			
	Field 1	Field 2	Field 3	Field 4	Field 1	Field 2	Field 3	Field 4
S1	**43**	**25**	**41**	**203**	**75**		**65**	**172**
S2	151				151			
S3		2	35			5	32	
S4		**16**		**3**				**19**
S5		16				16		
S6		7				7		
S7		4				4		
S8		1						1

similar to the breeding groups, there are improvements for stage groups using the two-phase optimization approach as well. With this approach stage group S1 is able to reduce fields from four fields to three fields and stage group S4 can have all its trials in a single field.

Table 3 shows the total cost and cost breakdown for the two optimization approaches. Based on the costs breakdown, the two phase approach can save more field space compared to within field optimal arrangement of the trials. The two-phase approach saves two more blocks by using 33 out of 39 field blocks. However this improvement in field utilization is not the main purpose of neither two-phase optimization nor within field optimization.

Table 5.3 Cost breakdown for original and optimal layout

Cost Category	Original Layout	Within Field Optimization	Two-phase Optimization	Original/ Within Field	Original/ Two-phase
Breeder Cost	**8250**	**4110**	**5700**	**2.01**	**1.45**
Occupation	2190	1620	1710	1.35	1.28
Distance	6060	2490	3990	2.43	1.52
Stage Cost	**26750**	**17900**	**14750**	**1.49**	**1.81**
Occupation	2400	2000	1850	1.20	1.30
Distance	24350	15900	12900	1.53	1.89
RM Cost	**24100**	**8580**	**7940**	**2.81**	**3.04**
Occupation	1420	980	920	1.45	1.54
Distance	22680	7600	7020	2.98	3.23
Field Cost	**39**	**35**	**33**	**1.11**	**1.18**
Occupation	39	35	33	1.11	1.18
Total Cost	**59139**	**30625**	**28423**	**1.93**	**2.08**

Considering the main objective which is putting trials from same breeding group as the first priority followed by trials from same stage and RM groups, the two phase approach can provide slightly better cost compared to the baseline optimization which is assigning trials to the field by a breeding expert and optimal arrangement of trials within each field using the baseline formulation. Comparing the cost ratios for the breeding group shows that the two phase optimization is able to make better assignment of trials to the fields by considering all cost priorities for the entire breeding program while manual assignment of the trials to the field focuses on the most important objective of keeping trials of the same breeding group as close as possible.This ensures the importance of automating the entire trial assignment process in a two-phase optimization program.

5.4 Conclusion

We have used a mathematical formulation aimed to improve utilization of each field, as it assures each breeding group would also prefer their trials to be positioned close to each other for convenience, and trials that are similar in terms of stage and RM should also be placed together as much as possible as this makes them easier to compare and evaluate for possible advancement to provide a scalable optimization model can achieve the same goal for entire breeding program. The results show that the two-phase optimized solution reduces waste space and cost of putting similar trials far apart more than within field optimization.

Needless to say that the optimization tool can solve the assignment problem in order of seconds and help the breeders to save weeks by solving the problem manually. The two-phase model is a way to make the process automated. Also, the solution provided by the model is the optimal solution which could not be easily obtained by human. A comparison of the model solution to a commercial breeding man-made solution shows considerable improvement. Finally, in practice there are last minute changes to the program like adding new trials or changing the objectives in terms of adding a new consideration or changing the importance of one. In such situation the model still would provide the optimal solution quick regardless of the changes but in manual approach it will require a lot more time and resources to solve the new problem along with more complexity for human brain to handle.

CHAPTER 6. CONCLUSION

6.1 Summary

This dissertation was aimed to support plant breeding process in two ways. First helping in making advancement decisions in a more reliable approach by providing a data-driven method to identify groups of similar GxE varieties that enhances the quality of phenotypic performance estimation and a prediction method which enables breeders to make advancement decision ahead of the time by identifying similar GxE varieties in laboratory stage using genetic data. Second, it helps with enhancing the breeding process through optimizing trails assignment between and within fields.

A data clustering approach could be applied to identify sets of related genotypes those have similar GxE interaction effect that require a proximity matrix of GxE similarity of all genotype pairs. Furthermore, it is not practically possible to define this GxE similarity for all genotype pairs, as it requires the genotypes having been planted in the same environments repeatedly. Thus, it naturally gives rise to a proximity matrix where most of the values are missing for either of these two fundamentally different reasons.For some pairs, their environmental requirements are simply too dissimilar, which makes them unlikely to be planted in the same location and for others this is it is only economically feasible to plant each seed variety in a limited number of locations are selected by plant breeders. Therefore, if the proximity matrix has missing values, no standard clustering method is directly applicable. Imputation can be done to replace missing values but considering the reason of missing values in imputation process has important effect in final output of clustering. The results show increasing number of missing values, statistical summary imputation methods loose ability to estimation and in very high sparse proximity matrices the PMC algorithm may fall into a wrong step of estimation.

Besides the clustering, we are able to determine pairs of genotypes that have the same preference to the environment (low dissimilarity in GxE effect) using the prediction model trained by historical data where we have a wide observation of genotypes in common environments in laboratory stage when there are no phenotypic observations available. Knowing that similar GxE varieties have same GxE preference to environment breeders can avoid planting similar GxE varieties in same environments and expand the experiment to more environments through similar GxE varieties in a way that maximizes information gain. A precise prediction of the yield of a target variety in early stages will provide a valuable source of information for breeders to make a better decision on future of a variety in breeding program. In other words, determining a set of similar GxE varieties breeders are able to predict yield of a target variety in some unseen environments in advanced that saves considerable amount of time and cost. Also, a better evaluation of a target performance using prediction can refuse an unwanted decision of keeping a future failure variety or discarding a potential winner variety based on few unreliable observations in early stages.

Along with all the predictive advancement we improve the breeding process from practical point of view as well. Modeling all the preferences and constraints in linear space, we're able to get a layout that fits the preferences the best. Given a set of trials from multiple breeding groups, we are able to split trials between the fields and arrange those within each field using the two-phase optimization model in a fully automatic and scalable process. Needless to say that an optimization tool can solve the assignment problem in order of seconds and help the breeders to save weeks by solving the problem manually. Also the solution provided by the model is the optimal solution which could not be easily obtained by human. A comparison of the model solution to a commercial breeding man-made solution shows considerable improvement. Also, in practice there are last minute changes to the program like adding new trials or changing the objectives in terms of adding a new consideration or changing the importance of one. In such situation the model still would provide the optimal solution quick regardless of the changes but in manual approach it will require a lot more time and resources to solve the new problem along

with more complexity for human brain to handle. These sort of resource optimization can provide field space to put more trials in practice that adds more knowledge to the program.

6.2 Future work

A correct advancement decision is paramount and anything that can be done to facilitate the comparison between competing experimental genotypes, and would thus aid the advancement process, would be of great value. In fact, the benefit of good trial placement is likely higher than the monetary benefit of reduced waste, although the former is hard to quantify since it depends on ultimately selecting the best varieties and hybrids for commercialization. As is mentioned, one of the advantages of identifying similar GxE varieties is that we can model yield of the similar GxE varieties with a non-interaction model that predicts yield simply as a function of genotype main effects and environment main effects, which makes predicting yield in new environments simpler. Having the ability to plan trials placement in a way that enables breeders to predict yield of experimental soybean varieties in environments where they have not been tested through similar GxE varieties aids decision makers when breeders need to decide if to keep a specific variety in breeding program and plant the variety again in following year, or discard it from the program with higher precision. So incorporating GxE similarities in trials placement optimization process can evolve the breeding process in a way that maximizes the information gain along with all discussed benefits. Needless to say that the two-phase optimization framework can simply enable this evolution.