Optimizing the design of planting experiments for agricultural crops

By

Ashley Kate Swift

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Industrial Engineering

Program of Study Committee:
Sigurdur Olafsson, Major Professor
Stephen Vardeman
Daniel Nordman

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

2019

TABLE OF CONTENTS

ABSTRACT

In commercial breeding, new genotypes are constantly being created and need to be tested to understand how a specific seed will perform in its target locations. A major constraint is that a genotype needs to go through multiple years of testing before it can be commercialized. With the volume of new genotypes that are constantly being enhanced, it is unrealistic to test every genotype in every target environment. Here, a methodology has been created that considers the fact that there are limited resources, whether it be limited space or a limited number of each genotype available in a single planting season. This new approach works by using the observations of genotypes that were planted and then inferring the performance of specific genotypes in certain environments. For agricultural crops, not all genotypes respond in the same way when planted in a certain environment. This phenomenon is describing genotype by environment (GxE) interaction. Numerous methods exist that aim to predict plant performance and specifically quantify and understand the GxE interaction. Here, five models are first evaluated on four different crop datasets. The Biclustering model is one model considered and it is effective at determining which genotypes have no GxE interaction in a subset of environments. This model works well with sparse data which is what exists in practice. Therefore, the Biclustering model is used to find subsets of genotypes and environments that have little to no GxE interaction.

In a subset of genotypes and environments with no interaction, genotypes can be planted in a strategic, methodical pattern so that the phenotype of unplanted genotypes can be inferred. Depending on the amount of physical resources available, two approaches can be utilized to gain information about unplanted genotypes. Given a set number of genotypes that can be planted, the first approach aims to maximize the number of known genotype/

environment pairs. The term genotype/environment pair refers to the phenotype that exists for a single genotype in a single environment. The second approach determines how many observations are required to infer every genotype/environment pair within a dataset. Additional constraints can be introduced to create a more realistic model.

The effectiveness of these two approaches can be illustrated using small-scale experimental designs that can be translated to full-scale commercial cases. In order to evaluate the effectiveness of the experimental designs created, both optimized and random models are compared to the original phenotypic responses. Validation indicates that optimizing the location of genotypes allows more inferences to be made, implying that creating an optimized planting plan can improve the understanding of genotypes. If this approach is applied in practice, it can facilitate further research as additional information can be gained from existing resources.

CHAPTER 1. INTRODUCTION

In the context of commercial plant breeding, new genotypes are continuously being created, tested, and modified to ultimately identify genotypes that successfully generate high yields or another desired phenotype. For breeders to determine how a new genotype will perform in different environments, it ultimately must be planted in each desired environment. However, due to limited resources of seeds and space paired with long growing seasons, it is impractical and cost prohibitive to plant each genotype in every target environment. This is especially true because a genotype requires multiple years of testing to understand how an environment affects its phenotypic performance.

To maximize the resources and time available, strategic models can be applied to gain the most information about how genotypes will perform, even in unplanted environments. In the context of crops, a phenotype is a set of observable characteristics that differentiate individual plants. Yield and flowering time are two phenotypic responses that can be measured when analyzing the performance of a genotype. The resulting phenotype of any crop is dependent upon how a genotype and environment interact.

Determining the phenotype of an untested genotype is not as simple as inferring the response based on how a genotype performed in a nearby environment. When two genotypes are planted in a perceived good environment and a perceived difficult environment, there is no guarantee that the genotypes will respond with the same magnitude or have a direct relationship. This phenomenon is describing the genotype by environment (GxE) interaction. The GxE interaction is illustrated in Figure 1. Each line represents the performance of a different genotype in nine different environments. If no interaction existed, every genotype would respond in the same way between environments.

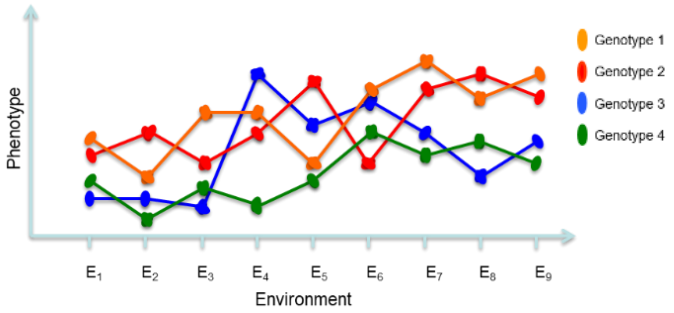

Figure 1: Example of how GxE interaction could result in a field

Clearly, this figure includes GxE interaction. The genotype (G) main effect explains that the orange genotype outperforms the green genotype in every environment. The environment (E) main effect explains that each genotype performs better in E_9 compared to E_2. When the lines are crossing, there is a GxE interaction that is causing genotypes to respond differently between different environments.

The genotype by environment (GxE) interaction is so fickle and difficult to quantify that there are numerous methods that have been proposed that aim to quantify and understand this interaction. These methods work similarly in most cases but may outperform other methods in certain instances. To determine the best approach to use for the evaluation of optimized planting designs, five different methods are evaluated. Each model is tested on four different datasets involving varying complexity of the phenotype and the crop data analyzed. In the order discussed, the methods include the following: Additive Model, All Interactions model, Regression on the Mean model, Additive Main Effects and Multiplicative Interactions model, and Biclustering model.

CHAPTER 2. BENCHMARKING EXISTING AND PROPOSED MODELS

2.1 Quantifying GxE – Existing Models

In the first four models, which were collectively analyzed by Malosetti et al. (2013), there is a heavy reliance on statistics to assist in describing and understanding the GxE interaction. The Additive Model is a simple approach that models the phenotype (μ_{ij}) as a sum of two main factors, genotype (G_i) and environment (E_i). This model focuses on genotypes $i \in I$ and environments $j \in J$. This simple no-interaction model can be modeled as followed.

$$\mu_{ij} = \mu + G_i + E_j + \epsilon_{ij}$$

In each equation, μ is a mean value that exists in the absence of (G_i) and (E_i) while ϵ_{ij} encompasses the normal error. In the Additive Model, the interaction of G_i and E_i do not help predict the phenotype (μ_{ij}). As aforementioned, it is not realistic to assume that the phenotype can be modeled as only a combination of genotype and environment main effects. Therefore, the next model, All Interaction, aims to incorporate the GxE interaction.

$$\mu_{ij} = \mu + G_i + E_j + GEI_{ij} + \epsilon_{ij}$$

The All Interaction model has the same structure as before with the addition of the term representing the GxE interaction (GEI_{ij}). A third model included in the comparison, that was introduced by Finlay et al. (1963), is the Regression on the Mean model. This approach aims to explain more of the GEI_{ij} by including a different variable in the model. GEI_{ij} is treated as a regression line on the quality of an environment. The quality is quantified by analyzing the average phenotype of all genotypes planted in that environment.

$$\mu_{ij} = \mu + G_i + E_j + b_i E_j + \epsilon_{ij}$$

In the Regression on the Mean Model, the term $b_i E_j$ aims to encompass the GxE interaction effect and represents the regression on the environment (E) main effect. The fourth model discussed by Gollob (1968) is the Additive Main Effects and Multiplicative Interactions (AMMI) model. It is less constrained and allows more than one environmental quality variable to be used where there are K multiplicative terms.

$$\mu_{ij} = \mu + E_j + \sum_{k=1}^{K} b_{ik} z_{jk} + \epsilon_{ij}$$

The variable b_{ik} represents the sensitivity of a genotype and z_{jk} represents the quality of the environment. This model can be thought of as using the principal components to quantify GxE interaction.

2.2 Biclustering Model

Knowing that the interaction between genotypes (G) and environments (E) can complicate the understanding of how a genotype will respond in certain environments, the Biclustering model aims to strategically group genotypes and environments into subsets, referred to as cells, where each single cell has no GxE interaction. To model the phenotypic response of a dataset, a two-way table (grid) with genotype as the rows (m) and environment as the columns (n) can be constructed. If each genotype in every environment is treated as its own no-interaction cell, then mxn no-interaction models exist. The overall GxE interaction for a dataset is determined by analyzing the differences that exist between all of the no-interaction models. The key to success for the Biclustering model is its ability to methodically group the genotypes and environments into homogeneous cells where the phenotype can be strictly attributed to the genotype and environment. This homogeneous cell is referred to as a no-interaction cell because the phenotypes can be expected to respond in

the same way if a genotype is planted in different environments within the cell. The goal is that the individual cells can be grouped together and still have no interaction so that the GxE interaction can be described with less complexity.

A no-interaction cell implies that the genotypes and environments within the cell can be modeled with the Additive Model where genotype and environment alone determine the resulting phenotype. Figure 2 illustrates how a complex model with each genotype/ environment pair acting as a single no-interaction cell can be shuffled to find sets of perfectly linear cells where the interaction among certain genotypes in set environments is negligible.

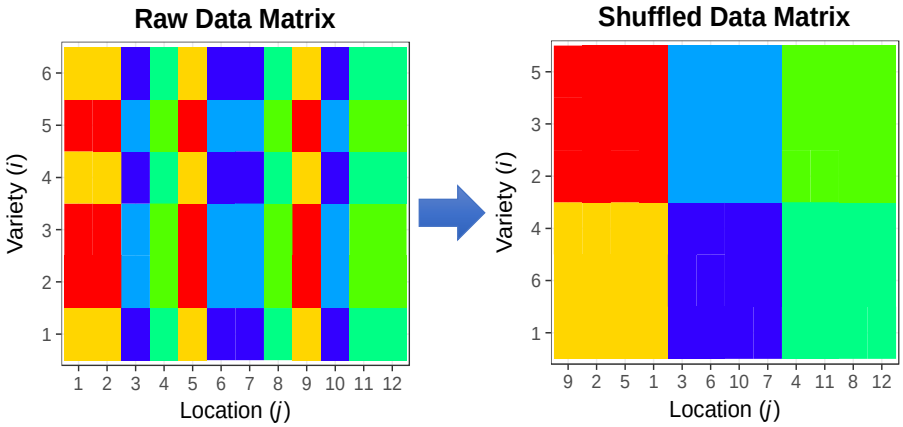

Figure 2: Illustration of how biclustering adds order to a perceived complex model

Graphically, a model with no interaction would have genotypes represented by parallel lines that respond to the set of environments in the exact same manner. This is shown in Figure 3(A). Figure 3(B) – Figure 3(D) show forms of interaction that ultimately complicate the understanding of phenotypic performance.

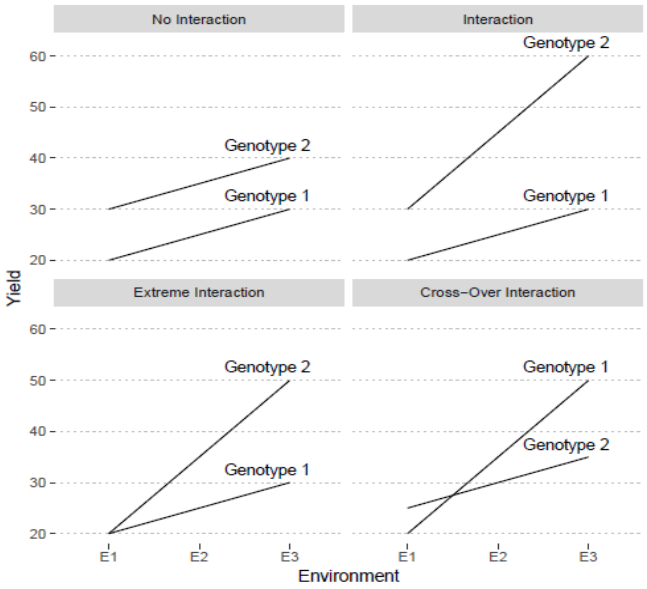

Figure 3: Four varying examples of how GxE interaction can occur

2.3 Benchmarking

The five models just described are summarized in the table below. These five models were evaluated on four datasets of differing crops with varying complexity. ANOVA tables were constructed to quantify how much each term was contributing to the sum of squares and to the complexity of the model. Understanding the residual error that exists in each model is the key focus.

Table 1: Varying models that are used to quantify the GxE interaction

	Model	Formulation
(1)	Additive Model (No-interaction)	$\mu_{ij} = \mu + G_i + E_j + \epsilon_{ij}$
(2)	All Interactions	$\mu_{ij} = \mu + G_i + E_j + GEI_{ij} + \epsilon_{ij}$
(3)	Regression on the Mean	$\mu_{ij} = \mu + G_i + E_j + b_i E_j + \epsilon_{ij}$
(4)	Additive Main Effects and Multiplicative Interactions (AMMI)	$\mu_{ij} = \mu + E_j + \sum_{k=1}^{K} b_{ik} z_{jk} + \epsilon_{ij}$
(5)	Biclustering	*Combination of Additive Models for each cell*

The Additive Model and All Interactions model are the two extremes of quantifying GxE interaction. When a dataset only has one observation of a genotype in each environment, all of the GxE interaction in the no-interaction model will be in the error term.

On the contrary, all of the GxE interaction in the All Interactions model will be included in the interaction term and no residual error will result. Therefore, Models (3), (4), and (5) will be the focus when comparing the effectiveness of each method. Table 2 presents the four crops that will be compared along with the phenotype that is measured. In terms of ability to quantify the GxE interaction, crops can be categorized as a simple or complex crop. The phenotype analyzed can also affect the results. Flowering time is easier to predict than yield. The data that is used for this comparison has varying amounts of missing data. With the more traditional approaches, incomplete data can negatively affect the ability of the model to capture the GxE interaction.

Table 2: Data information of the crops that are used to compare models

Crop	Phenotype	Genotypes	Environments	Missing Data
Sorghum	Flowering Time	236	7	3%
Rice	Flowering Time	175	9	3%
Maize	Yield	210	7	0%
Soybean	Yield	131	73	73%

2.3.1 Sorghum

In the following ANOVA tables, the sum of squared error (SSE) and degrees of freedom are the measures compared between models. The goal of each model is to most accurately predict the phenotype. All else equal, a model with a higher SSE and lower residual degrees of freedom is preferred because the model is capturing more of the GxE interaction with less complexity. In Model 3 – Model 5, the residual degrees of freedom can be analyzed as a measure of the complexity of the model. Table 3 – Table 6 illustrate the results of applying the first four models to the sorghum dataset obtained from Li et al. (2018). This dataset has 97% complete data for 236 genotypes evaluated in seven locations. The phenotypic response is flowering time. In the Additive Model shown in Table 3, the resulting SSE is more than the sum of squares that is captured by the G main effect. In the All

Interactions model, all of the error is captured in the GxE term; however, the model is as complex as it can be with no residual degrees of freedom. The Regression on the Mean and AMMI model in Table 5 and Table 6, respectively, both meaningfully manipulate the data to capture the GxE interaction. Therefore, these two models will be compared directly to the Biclustering model to determine its effectiveness.

Table 3: No-interaction Model of Sorghum

Model	Main Effect	D.F.	S.S.
No Interaction	G	236	53,020,738
	E	6	199,593,392
	Error	1,367	67,635,610

Table 4: All Interaction Model of Sorghum

Model	Main Effect	D.F.	S.S.
All Interaction	G	236	53,020,738
	E	6	199,593,392
	GxE	1367	67,635,610
	Error	0	0

Table 5: Regression on Mean Model of Sorghum

Model	Main Effect	D.F.	S.S.
Regression on Mean	G	236	53,020,738
	E	6	199,593,392
	Reg of E	236	52,422,217
	Error	1131	15,213,393

Table 6: AMMI Model of Sorghum

Model	Main Effect	D.F.	S.S.
AMMI	G	236	52,973,148
	E	6	199,640,982
	PC1	241	53,483,854
	PC2	239	5,581,472
	Error	887	7,570,284

Introducing the Regression of E term to Table 5 allows the model to explain as much about the data as the G main effect. The first principal component in the AMMI model explains more than the G main effect and, compared to the Regression on Mean model, the complexity is comparable. For the sorghum data, the AMMI model has the lowest SSE so its results will be compared to the Biclustering model.

The Biclustering model is very flexible. Depending on the number of row and column clusters created, the model can be as simple or as complex as the desired application. In order to compare the Biclustering model to the others just described, the number of row and column clusters that are constructed in the model are selected so that the residual degrees of freedom of the AMMI model with two principal components and the Biclustering model

match. If the SSE of the Biclustering model is lower than the AMMI model, it indicates that the Biclustering model is just as effective or better than the AMMI model.

Because the Biclustering model is dynamic, the model was explored to determine if there was an optimal set of row and column clusters that captured the sum of squares without adding the same degree of complexity. Figure 4 illustrates how the SSE changes based on varying the number of row and column clusters and degrees of freedom. The change is linear, so the final row and column clusters utilized can be determined based on the user's goals. This general pattern was observed for each dataset.

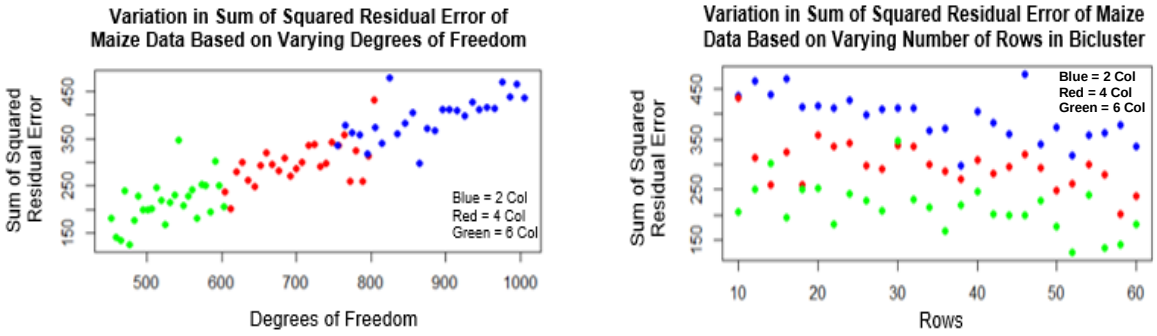

Figure 4: Variation in SSE resulting from differing row and column clusters and degrees of freedom

In order to get the Biclustering model to have a residual error around 887 like the AMMI model, the sorghum data was strategically split into two rows and three columns, as shown in Figure 5. For the biclustering algorithm to determine which cell the environments and genotypes should be put into, the differences of the phenotype from the environment average was used as the measure. This measure was used for every dataset for consistency purposes.

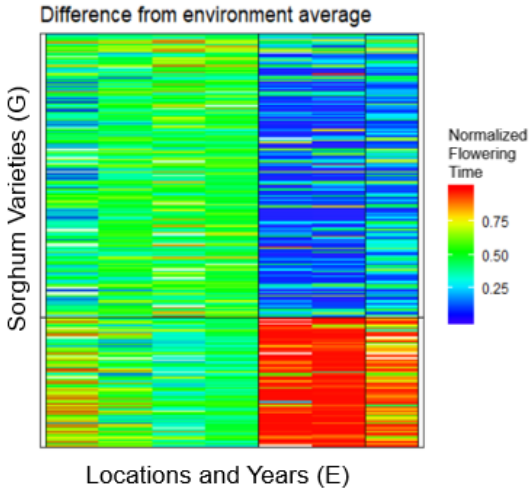

Figure 5: Heat Map of Shuffled Sorghum Data (2x3)

Splitting the data into two rows and three columns results in a residual degree of freedom of 896, as seen in Table 7. The SSE is slightly less than the SSE of the AMMI model, but the residual degrees of freedom is higher for the Biclustering model. Therefore, the Biclustering model is just as effective as the other models evaluated when applied to the sorghum data.

Table 7: Biclustering Model of Sorghum

Model	Main Effect	D.F.	S.S.
Biclustering	G	236	53,020,738
	E	6	199,593,392
	No Int Cells	471	60,230,522
	Error	896	7,405,088

When the biclustering algorithm is shuffling the data, the exact same genotypes and environments are not always put into the same cluster as the previous run. Therefore, the biclustering algorithm was run multiple times to find the minimum SSE possible which is recorded in Table 8. Running the algorithm multiple times also indicates how much variability existed in the model.

Table 8: Variation in Biclustering Model

Variability in Residual Sum of Squares (Sorghum)	
Minimum	7,405,088
Maximum	11,437,121
Average	10,223,972
Standard Deviation	1,741,186

2.3.2 Rice

The rice dataset is the next case that is compared. The phenotypic response here is also flowering time. The number of row and column clusters that were selected in this example for biclustering were again selected so that the Biclustering model and the AMMI model both had comparable residual degrees of freedom. Table 9 – Table 12 show the residual error for each model. In this dataset, the G and E main effects were able to explain much more of the model compared to the sorghum data. The AMMI model was further able to explain more than 90% of the original error with only two principal components.

Table 9: No-interaction Model of Rice

Model	Main Effect	D.F.	S.S.
No Interaction	G	175	134,189
	E	8	485,135
	Error	1,355	58,071

Table 10: All Interaction Model of Rice

Model	Main Effect	D.F.	S.S.
All Interaction	G	175	134,189
	E	8	485,135
	GxE	1,355	58,071
	Error	0	0

Table 11: Regression on Mean Model of Rice

Model	Main Effect	D.F.	S.S.
Regression on Mean	G	175	134,189
	E	8	485,135
	Reg of E	175	41,697
	Error	1,180	16,374

Table 12: AMMI Model of Rice

Model	Main Effect	D.F.	S.S.
AMMI	G	175	134,189
	E	8	485,135
	PC1	182	47,971
	PC2	180	4,272
	Error	993	5,828

The Biclustering model with similar complexity to the AMMI model splits the data into three row (genotype) clusters and three column (environment) clusters, as seen in Figure 6.

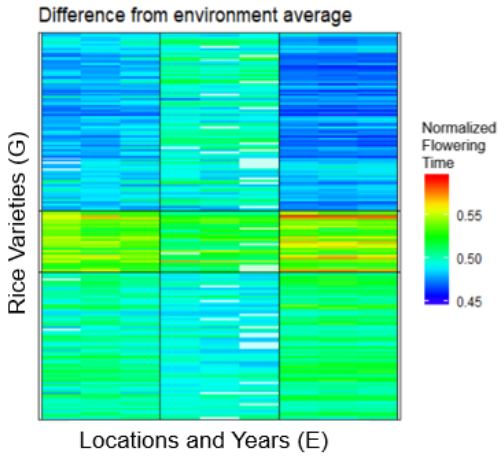

Figure 6: Heat Map of Shuffled Rice Data (3x3)

The residual degrees of freedom and sum of squares of the Biclustering model is compiled in Table 13. In this case, the sum of squared error for the Biclustering model is more than that of the AMMI model, but the two are comparable.

Table 13: Biclustering Model of Rice

Model	Main Effect	D.F.	S.S.
Biclustering	G	175	134,189
	E	8	485,135
	No Int Cells	362	51,586
	Error	993	6,485

Because the biclustering algorithm can have varying results depending on how well the algorithm is able to correctly group the genotypes and environments, multiple iterations were run to determine the minimum error that can result with a 3x3 cell. The results are compiled in Table 14.

Table 14: Variation in Biclustering Model

Variability in Residual Sum of Squares (Rice)	
Minimum	6,485
Maximum	11,747
Average	7,345
Standard Deviation	1,284

2.3.3 Maize

The third dataset that was used to compare the linear models to the Biclustering model was the maize data. Introduced by Ribaut et al. (1996, 1997) and used by Malosetti et al. (2013), in this dataset, 210 genotypes were planted in seven different environments. The sum of squared error in the no-interaction model is more than the sum of squares that can be described by the genotype (G) main effect. When applying each model, the lowest SSE again results from the AMMI model as seen in Table 15 – Table 18.

Table 15: No-interaction Model of Maize

Model	Main Effect	D.F.	S.S.
No Interaction	G	210	614
	E	7	5,679
	Error	1,470	813

Table 16: All Interaction Model of Maize

Model	Main Effect	D.F.	S.S.
All Interaction	G	210	614
	E	7	5,679
	GxE	1,470	813
	Error	0	0

Table 17: Regression on Mean Model of Maize

Model	Main Effect	D.F.	S.S.
Regression on Mean	G	210	614
	E	7	5,679
	Reg of E	210	230
	Error	1,260	583

Table 18: AMMI Model of Maize

Model	Main Effect	D.F.	S.S.
AMMI	G	210	614
	E	7	5,679
	PC1	216	242
	PC2	214	173
	Error	1,040	398

The Biclustering model uses a 3x3 cell to estimate the GxE interaction. The resulting Biclustering model can be seen in Figure 7.

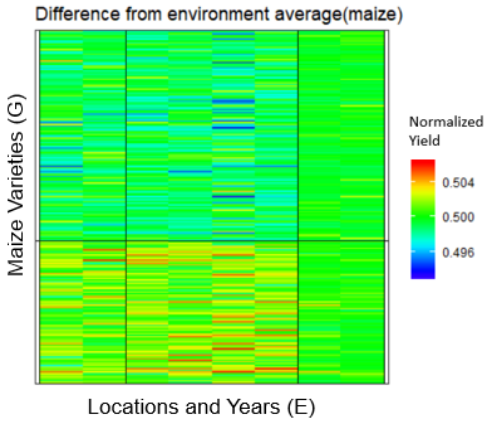

Figure 7: Heat Map of Shuffled Maize Data (3x3)

Like the rice dataset, the SSE that results from the biclustering algorithm is slightly higher than what was achieved by the AMMI model. This is shown in Table 19. The methods are still comparable; the linear models just outperform the Biclustering model in this case. Like the other models, the variation in the Biclustering model was analyzed to understand the stability of the Biclustering model on the Maize data. This is illustrated in Table 20.

Table 19: Biclustering Model of Maize

Model	Main Effect	D.F.	S.S.
Biclustering	G	210	614
	E	7	5,679
	No Int Cells	425	349
	Error	1,045	464

Table 20: Variation in Biclustering Model

Variability in Residual Sum of Squares (Maize)	
Minimum	464
Maximum	558
Average	492
Standard Deviation	40

2.3.4 Soybean

The most important comparison made in the context of this paper is the benchmarking done on the soybean dataset. This dataset was collected commercially by Syngenta Seeds and contains a large amount of missing data. The first three datasets evaluated were relatively complete and the Biclustering model was a contender in the

effectiveness of the model. However, in commercial practice, it is more common to have missing data in the sense that not every genotype is planted in every location so the biclustering cell (m rows and n columns) becomes sparse. Prior to the introduction of the Biclustering model, there was not a model that was effective at quantifying the GxE interaction when the data was mostly incomplete. Table 21 – Table 24 show the error that the linear models attained. For the no-interaction model, the SSE was double what the genotype (G) term was able to quantify, and the AMMI model was only able to reduce the SSE by less than 20%.

Table 21: No-interaction Model of Soybean

Model	Main Effect	D.F.	S.S.
No Interaction	G	131	26,098
	E	72	187,595
	Error	2,407	51,706

Table 22: All Interaction Model of Soybean

Model	Main Effect	D.F.	S.S.
All Interaction	G	131	26,098
	E	72	187,595
	GxE	2,407	51,706
	Error	0	0

Table 23: Regression on Mean Model of Soybean

Model	Main Effect	D.F.	S.S.
Regression on Mean	G	131	26,098
	E	72	187,595
	Reg of E	107	4,028
	Error	2,300	47,679

Table 24: AMMI Model of Soybean

Model	Main Effect	D.F.	S.S.
AMMI	G	131	14,975
	E	72	198,719
	PC1	202	4,590
	PC2	200	3,971
	Error	2,005	43,144

The results of splitting the 131 genotypes and 72 environments into three rows and three columns can be seen in Figure 8.

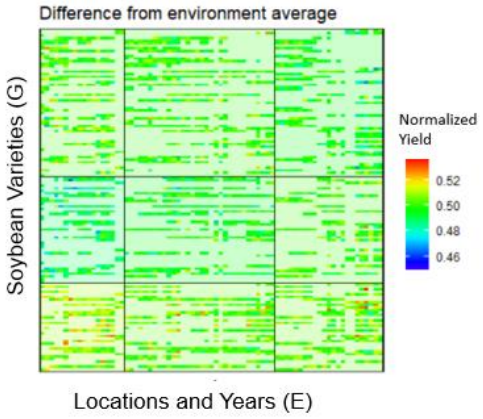

Figure 8: Heat Map of Shuffled Soybean Data (3x3)

With a 3x3 cell, the complexity of the model is less than that of the AMMI model and the SSE is lower. The results of the Biclustering model is depicted in Table 25. The results show that the Biclustering model was able to outperform the best linear model, AMMI, in terms of complexity and residual error.

Table 25: Biclustering Model of Soybean

Model	Main Effect	D.F.	S.S.
Biclustering	G	131	26,098
	E	72	187,595
	No Int Cells	333	17,947
	Error	2,074	33,759

It is even more reassuring in the effectiveness of the Biclustering model because even with variation, the maximum error calculated for the Biclustering model still captures more of the GxE interaction than any other linear model. The range of variability in the Biclustering model is illustrated in Table 26.

Table 26: Variation in Biclustering Model

Variability in Residual Sum of Squares (Soybean)	
Minimum	33,759
Maximum	39,911
Average	37,057
Standard Deviation	925
Range	6,152

When there is no interaction between the genotype and environments within a cell, one can predict how a genotype will perform in an unplanted environment within a cell. This concept is crucial for the models that have been constructed. Within a no-interaction cell, the phenotype of one genotype in an unplanted environment can be inferred based on how the phenotypes differed between two genotypes planted in the same environment. In other words, inferences can be made for genotypes in untested environments based directly from an observation of another genotype planted in that environment.

The Biclustering model outperforms the other models when using sparse data. The data used to demonstrate the effectiveness of the proposed optimization model is sparse data, which is representative of what happens commercially in the agricultural industry. Therefore, the Biclustering model is used extensively for the remainder of the discussion. Finding sets of genotypes and environments that interact in the same way is crucial in implementing the methodology utilized in this research. Each specific set of genotypes and environments can be referred to as a cell. There are a variety of ways that the genotypes can be planted to gain enhanced information. Several experimental designs are created and explored to examine the effectiveness of differing planting patterns.

Each experimental design has the goal of inferring the most genotype/environment pairs by strategically planting the limited resources. Two complementary methodologies can be utilized depending on the number of genotypes and other limited resources available. The approaches differ based on whether it is feasible to infer every genotype/environment pair. The first approach represents a common case. If a limited number of genotypes can be planted across all environments, the approach determines how genotypes can be arranged in order to maximize the number of genotype/environment pairs that can be inferred. The

second approach is the ideal case where the objective is to determine how many genotype/environment pairs need to be planted, so that every pair can be inferred within a dataset. A variety of small-scale experimental designs are constructed with logical constraints based on the two methodologies discussed. The intention is for these small-scale experimental designs to be duplicated in a larger scale for the use by commercial breeders.

CHAPTER 3. MATERIALS AND METHODS

Optimization formulations were developed to solve the two methodologies. For both, genotype/environment pairs are either tested (planted) or inferred. Minimally, one connection is required to make an inference for a genotype/environment pair. In order to create a connection, the desired genotype (g) needs to be planted in the same environment (e') as another known genotype (g'). The known genotype (g') also needs to be planted in the environment (e) of the desired genotype (g). This situation is illustrated in Figure 9. This connection can come from any genotype/environment pair within a no-interaction cell. The connection becomes void across cells.

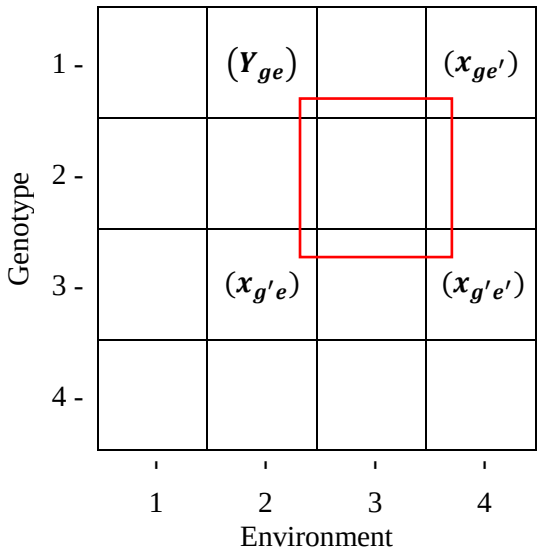

Figure 9: Illustration of how an inference can be made to predict an unobserved genotype/environment pair Y_{ge} based on the phenotypes of $x_{g'e}, x_{ge'}$ and $x_{g'e'}$.

To solve the first approach, a maximization formulation has been constructed. The objective of this approach is to maximize the number of genotype/environment pairs that can be inferred. The constraint is the set number of genotypes that can be planted across the environments. The premise is that more genotype/environment pairs can be inferred if the limited resources are strategically planted. Each variable is binary.

3.1 Approach 1

$Maximize$:

$$\sum_{g}\sum_{e} Y_{ge} \qquad 1(a)$$

$Subject\ to$:

$$\sum_{g}\sum_{e} x_{ge} \leq P \qquad 1(b)$$

$$\sum_{g\prime}\sum_{e\prime} x_{ge\prime} * x_{g\prime e} * x_{g\prime e\prime} + x_{ge} \geq Y_{ge} \qquad \forall\ ge \in Z_i \qquad 1(c)$$

where:

$x_{ge} = \begin{cases} 1 & \text{genotype (g) planted in environment (e)} \\ 0 & \text{genotype (g) not planted in environment (e)} \end{cases}$

$Y_{ge} = \begin{cases} 1 & \text{genotype (g) planted or reached in environment (e)} \\ 0 & \text{genotype (g) not planted nor reached in environment (e)} \end{cases}$

$P =$ maximum number of genotype/environment pairs to be planted

$Z_i =$ cell of genotypes and environments determined to be related

Equation 1(c) is used to determine whether a genotype/environment pair can be inferred. It involves multiplying all three variables together. If the product is zero, it implies that no connections were made for the specific pair. Although useful, this formulation makes this optimization model non-linear. In order to ensure convergence to an optimal solution, Equation 1(c) was linearized in the following manner. Where multiplication of variables exists in the problem formulations, the following linearization is applied.

Linearization of Variable Multiplication:

$$A_{geg'e'} \leq x_{g'e} \qquad \forall\ geg'e' \in Z_i \qquad 2(a)$$

$$A_{geg'e'} \leq x_{ge'} \qquad \forall\ geg'e' \in Z_i \qquad 2(b)$$

$$A_{geg'e'} \geq x_{g'e} + x_{ge'} - 1 \qquad \forall\ geg'e' \in Z_i \qquad 2(c)$$

$$B_{geg'e'} \leq A_{geg'e'} \qquad \forall\ geg'e' \in Z_i \qquad 2(d)$$

$$B_{geg'e'} \leq x_{g'e'} \qquad \forall\ geg'e' \in Z_i \qquad 2(e)$$

$$B_{geg'e'} \geq A_{geg'e'} + x_{g'e'} - 1 \qquad \forall\ geg'e' \in Z_i \qquad 2(f)$$

$$Y_{ge} \leq \sum_{g'} \sum_{e'} B_{geg'e'} + x_{ge} \qquad \forall\ ge \in Z_i \qquad 2(g)$$

where:

$x_{ge} = \begin{cases} 1 & \text{genotype (g) planted in environment (e)} \\ 0 & \text{genotype (g) not planted in environment (e)} \end{cases}$

$Y_{ge} = \begin{cases} 1 & \text{genotype (g) planted or reached in environment (e)} \\ 0 & \text{genotype (g) not planted nor reached in environment (e)} \end{cases}$

$A_{geg'e'} = \begin{cases} 1 & x_{g'e} \text{ and } x_{ge'} \text{ were both planted} \\ 0 & \text{at least one of } x_{g'e} \text{ and } x_{ge'} \text{ were not planted} \end{cases}$

$B_{geg'e'} = \begin{cases} 1 & x_{g'e}, x_{ge'}, \text{and } x_{g'e'} \text{ were both planted} \\ 0 & \text{at least one of } x_{g'e}, x_{ge'}, \text{and } x_{g'e'} \text{ were not planted} \end{cases}$

$Z_i =$ cell of genotypes and environments determined to be related

For the second approach, a minimization formulation has been constructed. The objective of this approach is to minimize the number of genotype/environment pairs that needs to be planted. In this approach, a specified number of genotype/environment pairs must be observed or inferred.

3.2 Approach 2

$Minimize$:

$$\sum_{g}\sum_{e} x_{ge} \tag{3(a)}$$

$Subject\ to$:

$$\sum_{g}\sum_{e} Y_{ge} \geq D \tag{3(b)}$$

$$\sum_{g'}\sum_{e'} x_{ge'} * x_{g'e} * x_{g'e'} + x_{ge} \geq Y_{ge} \qquad \forall\ ge \in Z_i \tag{3(c)}$$

where:

$x_{ge} = \begin{cases} 1 & \text{genotype (g) planted in environment (e)} \\ 0 & \text{genotype (g) not planted in environment (e)} \end{cases}$

$Y_{ge} = \begin{cases} 1 & \text{genotype (g) planted or reached in environment (e)} \\ 0 & \text{genotype (g) not planted nor reached in environment (e)} \end{cases}$

$D =$ desired amount of genotype/environment pairs to be planted and inferred

$Z_i =$ cell of genotypes and environments determined to be related

With the simple constraints of the approaches above, the required minimum number of genotypes planted is the sum of genotypes (m) and environments (n) minus one (m+n-1). In the simplest of cases, this optimal solution can be achieved by planting an “L”, “T”, or some variation of the shape. This implies that all pairs can be inferred by planting every genotype in one environment and planting one genotype in every environment. Three variations of this case are illustrated in Figure 10.

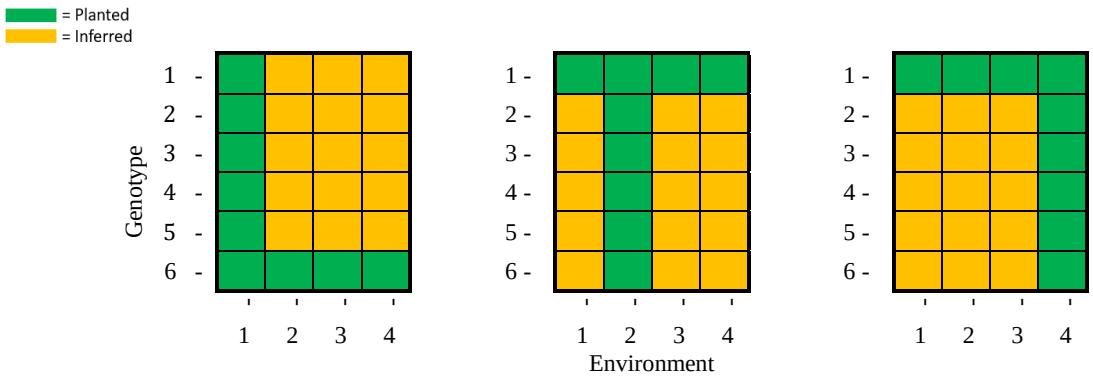

Figure 10: Three variations of the minimum required planting locations to infer an entire cell.

The situations illustrated achieve the goal of reaching every location; however, in practice, this is not a realistic or a commonly applied approach. Therefore, additional constraints are introduced to better represent current practices while making an improvement to the current system. The following constraints can be added in any combination to the model for either approach to create a more realistic set of experimental designs.

3.3 Additional Constraints

The following constraint singlehandedly ensures that the algorithm does not plant every genotype in only one environment. This constraint limits the number of genotypes that can be planted in any one environment. S_i is the maximum number of genotypes that can be planted in an environment and it can be modified to resemble the goal of the experiment.

Constraint 1:

$$\sum_{g} x_{ge} \leq S_i \qquad \forall\ e \in Z_i \qquad (4)$$

where:

$x_{ge} = \begin{cases} 1 & \text{genotype (g) planted in environment (e)} \\ 0 & \text{genotype (g) not planted in environment (e)} \end{cases}$

$S_i =$ maximum number of genotypes to be planted in any one environment

$Z_i =$ cell of genotypes and environments determined to be related

The next constraint aims to ensure that a genotype is planted in more than one environment. In order to learn more about how the genotype performs in a variety of environments, it should be planted in at least a few different environments.

Constraint 2:

$$\sum_{e} x_{ge} \leq W_i \, (C_i + 1) \qquad \forall \; g \in Z_i \qquad 5(a)$$

$$\sum_{e} x_{ge} \geq L_i * W_i \qquad \forall \; g \in Z_i \qquad 5(b)$$

where:

$x_{ge} = \begin{cases} 1 & \text{genotype (g) planted in environment (e)} \\ 0 & \text{genotype (g) not planted in environment (e)} \end{cases}$

$W_i = \begin{cases} 1 & \text{indicator that requirement was met} \\ 0 & \text{indicator that requirement was not met} \end{cases}$

$C_i =$ number of environments in cell i

$L_i =$ minimum number of environments a specific genotype has to be planted in'

$Z_i =$ cell of genotypes and environments determined to be related

The third constraint is particularly beneficial when a Biclustering model is not completely linear. This constraint sets a minimum number of connections that needs to be formed before a genotype/environment pair can be inferred. With this constraint, the breeder can specify the number of connections desired before an inference can be made. The more phenotypic responses averaged together, the more realistic the inference will be.

Constraint 3:

$$\sum_{g'} \sum_{e'} x_{ge'} * x_{g'e} * x_{g'e'} + M_i * x_{ge} - M_i + 1 \geq Y_{ge} \qquad \forall \; ge \in Z_i \qquad (6)$$

where:

$x_{ge} = \begin{cases} 1 & \text{genotype (g) planted in environment (e)} \\ 0 & \text{genotype (g) not planted in environment (e)} \end{cases}$

$Y_{ge} = \begin{cases} 1 & \text{genotype (g) planted or reached in environment (e)} \\ 0 & \text{genotype (g) not planted nor reached in environment (e)} \end{cases}$

$M_i =$ minimum number of connections required to infer a genotype/environment

$Z_i =$ cell of genotypes and environments determined to be related

Adding constraints to Approach 1 and Approach 2 can cause the resulting planting plan to change. Figure 11 illustrates one example of how a planting plan for a 6x4 cell varies with additional constraints when a maximum of nine genotype/environment pairs can be

planted. Figure 11(a) has no additional constraints. Figure 11(b) – Figure 11(d) reflect Constraint 1 – Constraint 3, respectively.

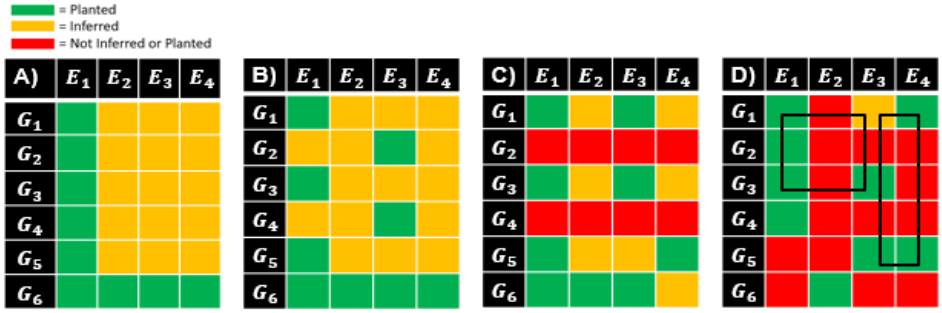

Figure 11: Examples of how additional constraints influence the planting design. Plot A – Set number of genotype/environment pairs planted, Plot B – Maximum number of genotypes in a certain environment, Plot C – Minimum number of environments planted if a genotype is planted, Plot D – Specified number of connections required for an unplanted environment to be inferred

In order to measure the success of the optimized models, each time that the model is evaluated, identical constraints are applied to a random model. With randomness, there will be the same or fewer pairs that can be inferred because no thought went into the placement of genotype/environment pairs.

This research problem is a two-step problem. Optimizing a planting plan is the first step. The goal of each experimental design is to infer as many pairs/cells as possible. It is understood that there are some planting designs with numerous optimal solutions in terms of how a grid can be structured regarding what cells are planted and which are inferred. For the evaluation below, each example uses only one of the optimal solutions produced. The next steps in the process are built upon the chosen optimal solution. After an optimal planting design has been determined, the effectiveness of the model is tested.

The biclustering algorithm is applied to three different datasets to identify genotypes and environments that have little to no interaction in each. For this research, the biclustering algorithm is serving as a way to determine genotypes and environments that have minimal interaction, implying that the phenotypes respond in the same manner. To compare the

optimized planting plan to a random plan, the phenotypic responses of the complete cells, that were determined by the biclustering algorithm, are going to be treated as the ground truth. The optimized and random plans will specify which cells will contain the true, observed phenotypic data. Only the cells deemed planted by the optimization model will have the phenotypic data inserted; the remainder will be blank initially. The phenotypes of those blank (unplanted) cells will be inferred, if possible, based on the connections that were formed. This inference can be made by taking the difference of phenotypes in an environment where both were planted and applying that difference in an unknown environment. The cells that are not inferred from the planted locations use row and column averages if possible or are left blank.

Once the experimental designs have the planted and inferred phenotypic responses input, the next step is to compare the amount of error that exists between the optimized or random models and the original data.

CHAPTER 4. RESULTS AND DISCUSSION

To evaluate the effectiveness of the different experimental designs, the phenotypic values from the original bicluster cell are compared to estimates using models generated from the optimized and random experimental designs. First, complete experimental planting designs are compared to designs where the cell cannot be fully inferred. By reducing the number of planting locations that can be observed, the experimental design becomes scarcer.

4.1 Complete vs. Incomplete Design

To illustrate the effect of limiting the number of genotype/environment pairs, a series of plots was created and evaluated using three different datasets. A dataset containing six genotypes and four environments was constructed to illustrate how strategically placing pairs in a no-interaction cell leads to more information gain and less error than if the identical number of observations was planted at random. Figure 12 shows how many pairs can be inferred as the number of observations are reduced from twelve to eight using the optimization models described above. For this case, the number of pairs planted is the main constraint and only one connection is required to make an inference for the phenotype.

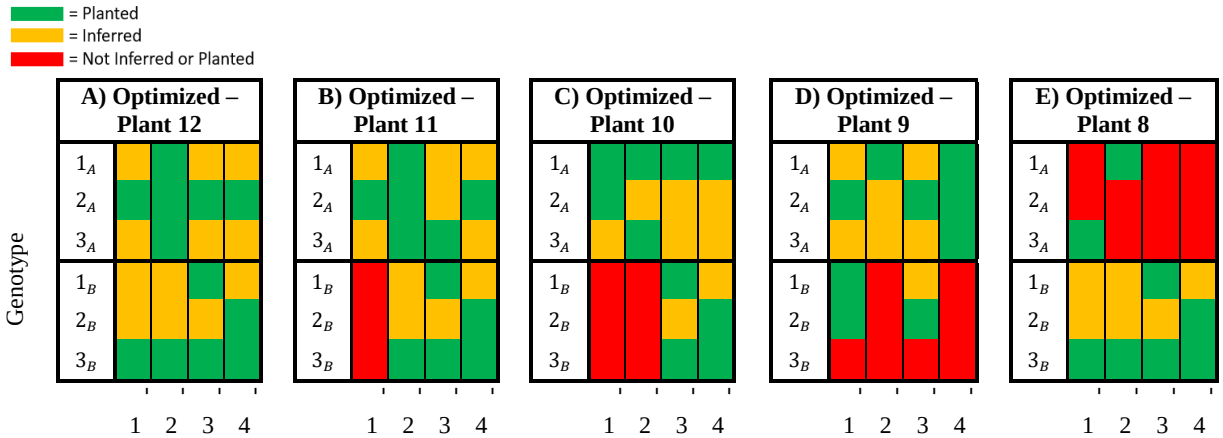

Figure 12: Results of using optimization to create a planting design where planted pairs are reduced from 12 to 8.

Figure 13 illustrates the same situation as Figure 12 except in this case; the pairs are randomly selected. When randomly selecting the experimental design, there is not a case where all of the pairs can be inferred.

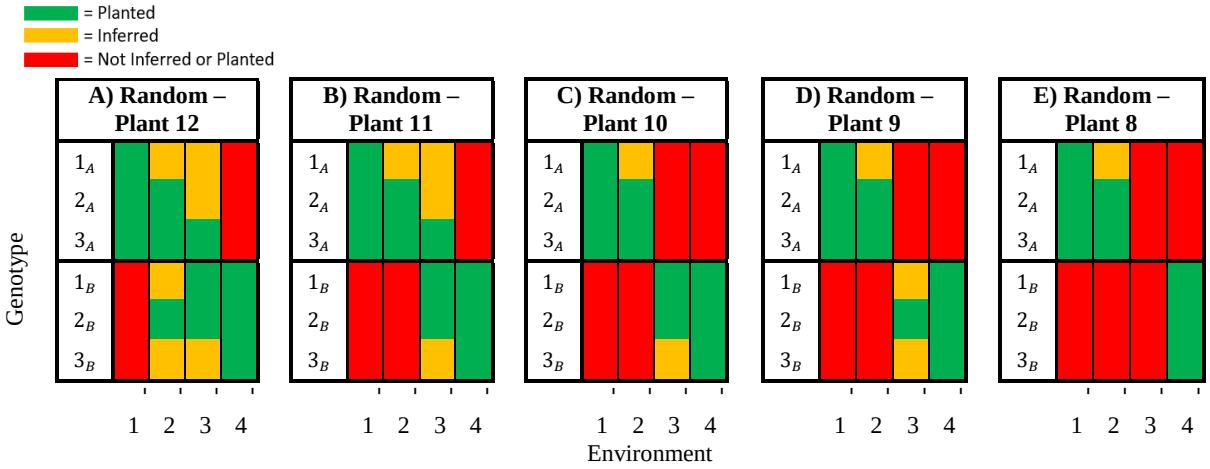

Figure 13: Results of randomly placing observations to create a planting design where planted pairs are reduced from 12 to 8.

In each case in this section, two no-interaction cells were constructed using the biclustering algorithm. Figure 14 – Figure 16 illustrate the original data that was reshuffled into two no-interaction cells. This data is the ground truth that every design is attempting to achieve. In Figure 14, each biclustering cell is perfectly linear and therefore has no interactions. This implies that the Additive Model has a residual sum of squares of zero.

The model that these results were obtained from required one connection. When the model is completely linear, having one or multiple connections will lead to equivalent results for inferred phenotypes. When the cell is not perfect, adding more connections is a way to gain additional understanding of what the phenotype for an unplanted pair would be. Figure 14 – Figure 16 illustrate the two cells that the biclustering algorithm created along with the phenotypic values of each pair. Figure 15 and Figure 16 use the sorghum and rice data for the evaluation. These two datasets were planted in an academic study. With academically

collected data, more factors can be monitored compared to commercial seeds. Therefore, this data is the next best alternative to a constructed, perfectly linear model.

	E_1	E_2	E_3	E_4
G_1	3	6	5	3
G_2	9	5	6	9
G_3	1	4	3	1
G_4	8	4	5	8
G_5	10	6	7	10
G_6	2	5	4	2

	E_1	E_2	E_3	E_4
G_1	3	6	5	3
G_3	1	4	3	1
G_6	2	5	4	2
G_2	9	5	6	9
G_4	10	6	7	10
G_5	8	4	5	8

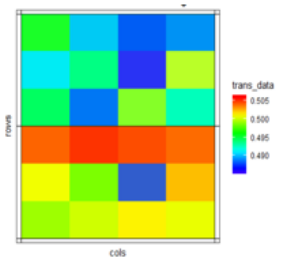

Figure 14: In order from left to right, the original data of a constructed dataset is shown followed by the data separated into two no-interaction cells. Last is a heat map illustrating the difference each pair is from the environmental average.

	E_1	E_2	E_3	E_4
G_1	48	2	89	-32
G_2	30	73	61	22
G_3	-149	-195	-122	-115
G_4	89	61	35	-61
G_5	-199	-217	-136	-61
G_6	-95	-217	-122	-102

	E_1	E_2	E_3	E_4
G_5	-199	-217	-136	-61
G_6	-95	-217	-122	-102
G_3	-149	-195	-122	-115
G_4	89	61	35	-61
G_1	48	2	89	-32
G_2	30	73	61	22

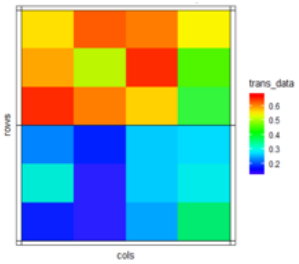

Figure 15: In order from left to right, the original data of Sorghum phenotypes is shown followed by the data separated into two no-interaction cells. Last is a heat map illustrating the difference each pair is from the environmental average.

	E_1	E_2	E_3	E_4
G_1	-4.6	-3.0	-7.3	0.4
G_2	-1.6	-4.0	-6.3	-5.7
G_3	-0.6	0.02	0.7	0.4
G_4	0.4	-1.0	-6.3	1.4
G_5	2.4	3.0	2.7	2.4
G_6	-2.6	-6.0	-0.3	-3.7

	E_1	E_2	E_3	E_4
G_3	-0.6	0.02	0.7	0.4
G_4	0.4	-1.0	-6.3	1.4
G_5	2.4	3.0	2.7	2.4
G_6	-2.6	-6.0	-0.3	-3.7
G_1	-4.6	-3.0	-7.3	0.4
G_2	-1.6	-4.0	-6.3	-5.7

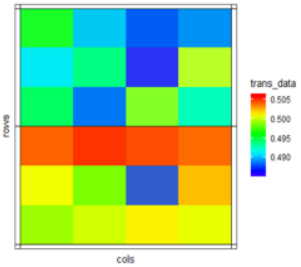

Figure 16: In order from left to right, the original data of Rice phenotypes is shown followed by the data separated into two no-interaction cells. Last is a heat map illustrating the difference each pair is from the environmental average.

In order to evaluate the effectiveness of each model with the different datasets, the information was combined as follows. For each unique experimental design using both the optimized and random designs as seen in Figure 17 and Figure 18 respectively, the model was tested using the constructed data, the sorghum data, and the rice data. From there, the data that was planted in each experimental design was directly applied in the new model.

Using the data that was planted, all of the pairs within a cell that could be were inferred. If no inference could be made, the environment and genotype average was placed within the cell.

Once the cells were completed as much as possible, a model was created and used to predict the original phenotypes in each dataset. The differences between the predicted phenotypes and original phenotypes were summarized by calculating the sum of squared error. As seen in Figure 17 and Figure 18, for the first two datasets in all but one experiment, the optimized model had a lower SSE compared to the random model. This indicates that when the Biclustering model can effectively find no-interaction cells, the results are favorable. The results of the Rice dataset indicate that when not every genotype/environment pair can be planted, like in the last case where only eight observations were used, the optimization model performs better than randomly planting genotypes in environments.

	Constructed Model - Optimized					**Sorghum Model – Optimized**					**Rice Model – Optimized**				
Pairs Planted	**12**	**11**	**10**	**9**	**8**	**12**	**11**	**10**	**9**	**8**	**12**	**11**	**10**	**9**	**8**
SSE - Combined 6x4 Cell	57	63	74	92	82	55,287	55,974	97,743	82617	613,26	259	150	179	186	223
SSE – 3x4 Cell 1	0	0	0	0	21	16,502	18,216	33,852	32,585	27,459	82	33	58	113	61
SSE – 3x4 Cell 2	0	8	29	38	0	20,205	18,677	20,211	24,968	20,205	140	124	120	71	140
SSE – Sum of Cell 1 & Cell 2	**0**	**8**	**29**	**38**	**21**	**36,707**	**36,893**	**54,063**	**57,553**	**47,664**	**221**	**157**	**178**	**184**	**201**

Figure 17: Summary of sum of squared error for each optimized experimental design and dataset.

	Constructed Model – Random					**Sorghum Model – Random**					**Rice Model – Random**				
Pairs Planted	**12**	**11**	**10**	**9**	**8**	**12**	**11**	**10**	**9**	**8**	**12**	**11**	**10**	**9**	**8**
SSE - Combined 6x4 Cell	64	72	78	82	90	66,623	64,233	60,425	66,396	62,708	122	109	109	150	166
SSE – 3x4 Cell 1	3	3	8	8	8	24,732	24,732	23,234	23,234	23,234	44	44	44	44	44
SSE – 3x4 Cell 2	4	29	29	32	38	26,587	22,903	22,903	37,883	32,026	83	72	73	152	121
SSE – Sum of Cell 1 & Cell 2	**7**	**32**	**37**	**40**	**46**	**51,319**	**47,635**	**46,137**	**61,117**	**55,260**	**127**	**116**	**117**	**196**	**165**

Figure 18: Summary of sum of squared error for each random experimental design and dataset.

4.2 Genotypic and Environment Constraints

This next section evaluates how planting designs can be altered to incorporate constraints that are common in commercial practice. A dataset containing ten genotypes and

four environments was constructed to illustrate how the following constraints affected the design of a planting pattern for optimized and random models. Figure 19 and Figure 20 include a series of experimental designs. Figure 19(a) and Figure 20(a) include the minimum number of observations to infer every cell. No other constraints are added. In Figure 19(b) and Figure 20(b), a constraint is added to limit the number of genotypes that can be planted in any one environment. This constraint is valuable in two ways. In practice, every location has a set amount of area to be used for planting. Second, multiple locations need to be planted to determine how genotypes perform across different environments. As shown in Figure 19(b), no new observations need to be added in order to infer every genotype/ environment pair. The observations are simply shifted so that no more than three genotypes are in an environment within a cell. Figure 19(c) and Figure 20(c) illustrate how the experimental planting design would change if a genotype needed to be planted in more than one environment if it was planted at all. In practice, a genotype is planted in multiple environments to evaluate its performance. In the optimized series of plots, when a constraint is added that requires a genotype to be planted in more than two environments, an experimental design cannot be created where every pair can be inferred if only sixteen pairs are planted. Therefore, Figure 19(d) and Figure 20(d) were created to show that more observations were required to infer every pair. With additional constraints, the random models do not do as well in terms of the cells that can be inferred.

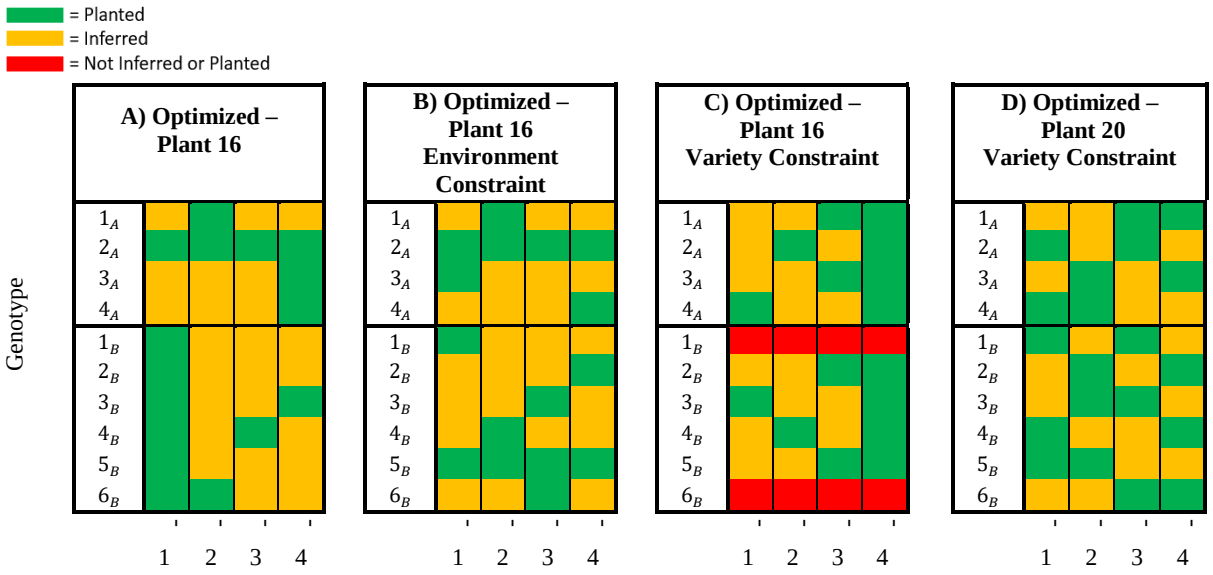

Figure 19: Results of using optimization to create a planting design. This sequence of images explores how the planting pattern changes with additional constraints.

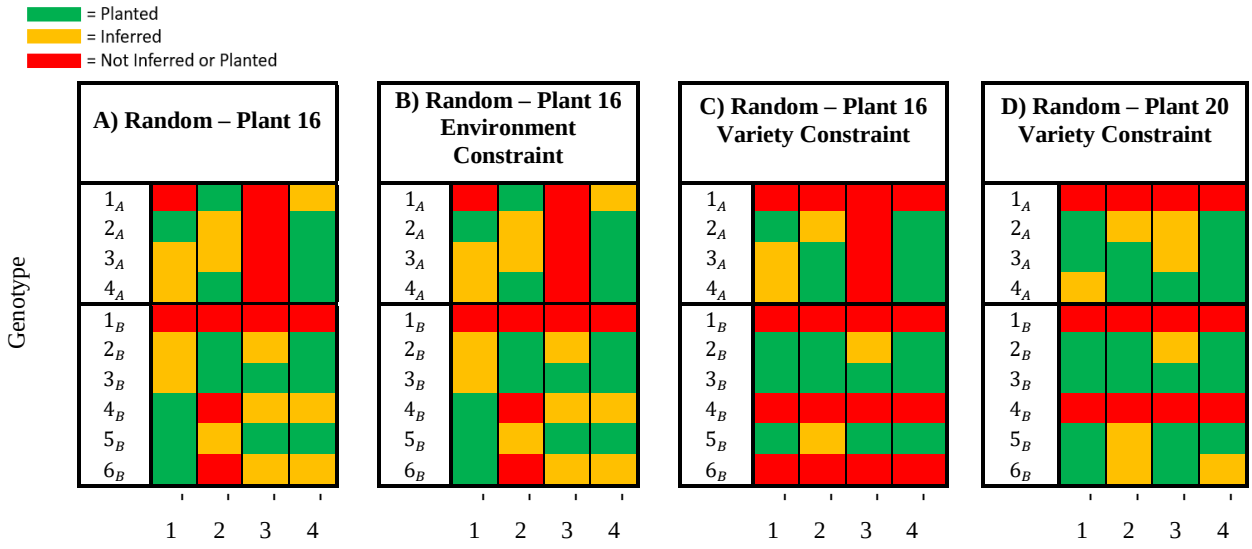

Figure 20: Results of randomly planting observations to create a planting design. This sequence of images explores how the planting pattern changes with additional constraints.

Figure 21 and Figure 22 show the exact data that was analyzed using each of the different experimental designs that were created. The biclustering algorithm again split the data into two cells. The first was a 4x4 cell and the second was a 6x4 cell.

	E_1	E_2	E_3	E_4
G_1	-278	-103	-5	-5
G_2	-296	-54	8	89
G_3	-197	-78	7	-60
G_4	-229	-65	-5	21
G_5	-181	-186	-73	-61
G_6	-166	-79	-73	-61
G_7	-136	-78	-110	-47
G_8	-136	-78	-46	22
G_9	-250	-79	-73	-19
G_{10}	-197	-12	-19	-46

	E_1	E_2	E_3	E_4
G_8	-136	-78	-46	22
G_{10}	-197	-12	-19	-46
G_2	-296	-54	8	89
G_4	-229	-65	-5	21
G_5	-181	-186	-73	-61
G_6	-166	-79	-73	-61
G_7	-136	-78	-110	-47
G_1	-278	-103	-5	-5
G_9	-250	-79	-73	-19
G_3	-197	-78	7	-60

Figure 21: In order from left to right, the original data of Sorghum phenotypes is shown followed by the data separated into two no-interaction cells.

	E_1	E_2	E_3	E_4
G_1	-0.6	2.0	3.77	1.34
G_2	-13.6	-11.0	-13.3	-12.7
G_3	-13.6	-10.0	-13.3	-13.7
G_4	-0.6	0.02	2.7	1.4
G_5	2.4	-1.0	0.7	1.4
G_6	-10.6	-9.0	-11.3	-13.7
G_7	-13.6	-12.0	-14.3	-11.7
G_8	-10.6	-8.0	-13.3	-12.7
G_9	-0.6	-1.0	1.7	0.4
G_{10}	-7.6	-7.0	-10.3	-9.7

	E_1	E_2	E_3	E_4
G_9	-0.6	-1.0	1.7	0.4
G_1	-0.6	2.0	3.77	1.34
G_4	-0.6	0.02	2.7	1.4
G_5	2.4	-1.0	0.7	1.4
G_{10}	-7.6	-7.0	-10.3	-9.7
G_2	-13.6	-11.0	-13.3	-12.7
G_3	-13.6	-10.0	-13.3	-13.7
G_6	-10.6	-9.0	-11.3	-13.7
G_7	-13.6	-12.0	-14.3	-11.7
G_8	-10.6	-8.0	-13.3	-12.7

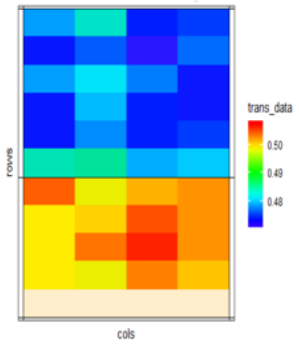

Figure 22: In order from left to right, the original data of Rice phenotypes is shown followed by the data separated into two no-interaction cells. Last is a heat map illustrating the difference each pair is from the environmental average.

Figure 23 summarizes the error that resulted for each of the four variations of the optimized models. The goal of this comparison was to see how each constraint affected the error of the model.

	Sorghum Model – Optimized				**Rice Model – Optimized**			
Pairs Planted	**Std-16**	**Env-16**	**Geno-16**	**Geno-20**	**Std-16**	**Env-16**	**Geno-16**	**Geno-20**
SSE - Combined 10x4 Cell	353,287	132,936	138,111	154,075	112	132	219	93
SSE – 4x4 Cell 1	127,998	78,609	54,814	55,491	46	46	36	31
SSE – 6x4 Cell 2	193,118	118,588	63,759	88,213	36	81	123	28
SSE – Sum of Cell 1 & Cell 2	321,116	197,197	118,573	143,704	82	127	159	59

Figure 23: Summary of sum of squared error for each optimized experimental design and dataset.

Adding additional constraints to the first model for the sorghum data has a considerable impact on the reduction of SSE that results. Requiring the models to rely on multiple

environments and genotypes provides enhanced information to describe the cell. The rice data did not perform as consistently between models; however, it is evident that the model with the lowest SSE was the last model with the additional constraint of the requirement that a genotype needed to be planted in multiple environments. The highest SSE did result when the model was unable to infer every genotype/environment pair. This indicates that when a planting design can be constructed to reach every genotype/environment pair, the inferences are positively affecting the model and resulting in less residual error.

CHAPTER 5. CONCLUSIONS

There are several methods that exist to quantify GxE interaction. Depending on the goal in mind and the dataset available, differing methods have more or less success. The primary goal of this research was to find a way to get an increased understanding about phenotypic performance in the presence of limited resources. Therefore, it was validated that using the Biclustering model is an effective method to determine what genotypes and environments have no interaction. By shuffling a dataset to find no-interaction cells, the Biclustering model is quantifying the GxE interaction. This model is not only effective for sparse data, but it has also been shown to be as effective as other methods when used to classify the GxE interaction for complete datasets involving both simple and complex crops.

To evaluate the effectiveness of the optimization model, the Biclustering model was utilized to determine genotypes and environments with minimal interaction. Using the subsets of no-interaction cells found by the Biclustering model, the information was used to illustrate that information can be gained when crops are planted according to a strategically designed planting plan. The information gained is an inference of the performance of other genotypes in unplanted environments. Known phenotypic data from complete planting plans was compared to new models to illustrate the effectiveness of the optimized and random models.

In practice, planting plans need to be constructed before crops are planted in order to save resources. Optimizing planting plans can be extended to be applied before crops are planted in the fields. For this to be successful, breeders and researchers need to continue to improve their understanding of what genotypes and environments are related so that the inferences made about unplanted genotype/environment pairs reflect the truth. Using the

Biclustering model is just one way to understand the relationship between genotypes and their environments. As the understanding of how genotypes perform in certain environments increases, this research can become an effective approach to save resources because more information can be obtained from the genotypes and environments that are already being planted.

REFERENCES CITED

Bondari, K., Statistical analysis of genotype X environment interaction in agricultural research. Experimental Statistics, Coastal Plain Station, University of Georgia

Finlay, K. W., Wilkinson, G. N., 1963. The analysis of adaptation in a plant-breeding programme. *AustralianJournal of Agricultural Research* 14, 742–754.

Gollob, H. F., 1968. A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika* 33, 73–115.

Li, J., Reisner, J., Pham, H., Olafsson, S., Vardeman, S., 2019. Biclustering for missing data. Information Sciences Submitted.

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

Malosetti, M., Ribaut, J. M., van Eeuwijk, F. A., 2013. The statistical analysis of multi-environment data: Modeling genotype-by-environment interaction and its genetic basis. *Frontiers in Physiology.* 4(44), 1-8

Reisner, J., Pham, H., Olafsson, S., Vardeman, S., Li, J., 2019. biclustermd: An R package for biclustering with missing data. The R Journal Submitted.

Ribaut, J. M., Hoisington, D. A., Deutsch, J. A., Jiang, C., Gonzalez-De-Leon, D., 1996. Identification of quantitative trait loci under drought conditions in tropical maize. 1. Flowering parameters and the anthesis-silking interval. *Theoretical and Applied Genetics* 92, 905–914.

Ribaut, J. M., Jiang, C., Gonzalez-de Leon, D., Edmeades, G. O., Hoisington, D. A., 1997. Identification of quantitative trait loci under drought conditions in tropical maize. 2. Yield components and marker-assisted selection strategies. *Theoretical and Applied Genetics* 94, 887–896.