Accounting for rank uncertainty in decision making for plant breeding

by

Reyhaneh Bijari

A dissertation 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
Qing Li
Kyung Min
Daniel Nordman
Stephen Vardeman

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. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2022

DEDICATION

To my family who always love, encourage and believe in me.

And, to the loving memory of my dear nephew, Foad.

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 this research and writing this dissertation.

First and foremost, I would like to express the deepest appreciation to my advisor Dr. Sigurdur Olafsson, an incredible leader, brilliant mentor, and supportive advisor, for his detailed guidance and encouragement that have often inspired me throughout my research.

I would also like to thank Dr. Vardeman for his insightful ideas, guidance and feedbacks throughout my journey.

I would also like to thank the rest of my dissertation committee: Dr. Min, Dr. Nordman and Dr. Li for being in my committee to provide me guidance and suggestions.

I also want to thank my master's advisor, Dr. Maryam Esmaeili for always being an inspirational person to me.

ABSTRACT

This dissertation is devoted to helping solve real-world plant breeding problems using innovative data science. There have been lots of efforts in the area of plant breeding to improve the quality of decisions made in such programs. While the use of new techniques has increased in this area, there exist lots of limitations in these programs that tie to unavoidable uncertainties that need to be taken into account for proper analysis. This work addresses a plant breeding decision-making challenge that stems from having a very limited number of environments observed for each plant breeding trial. We propose new methods that plant breeders can utilize when analyzing the genotypes' performance. Specifically, to capture the inherent uncertainty due to the specific set of environments observed, we propose a bootstrapping approach to estimating the distribution of rank and constructing confidence intervals around it. We also a new approach to compare genotypes probabilistically and offer a new ranking method based on pairwise probabilistic comparisons of genotypes. The methods are provided in an R package for analysis of plant breeding experiments for all users. We believe plant breeding would benefit from the body of this work as it tries to fill the gap in the analysis of multi-environment trails' data.

CHAPTER 1. GENERAL INTRODUCTION

1.1 Overview

The world's population is dramatically increasing, and it will increase the food demand substantially. Based to the Food and Agriculture Organization of the United Nations (FAO), agriculture needs to produce almost 50 percent more than it did in 2012 as the population will reach almost 10 billion in 2050 (FAO, 2017). Considering the limited natural resources, we need to employ the growing technology in all aspects, from mechanical to AI, to address this challenge and meet the food demand for such a population. We believe that using the new-born techniques will lead to a higher quality food supply, both with respect to health and amount while keeping the human footprint as small as possible on the environment.

Plant breeding programs are designed to select the experimental genotypes with the best genetic properties which perform relatively consistently across potential environments (Happ et al., 2021). Due to the inherent complexity of the decision-making in plant breeding programs, each program might take multiple years from the initial stages to commercialization. Commercialized genotypes are the ones that have been planted and outperformed the other experimental genotypes with respect to some phenotypic traits, e.g., yield, in different stages over multiple years.

Phenotypic traits such as the yield of genotypes depend heavily on the environmental conditions they face. They are affected differently by environmental conditions, which is commonly referred to as genotype by environment (G×E) interactions (Comstock and Moll, 1963). These genotype-by-environment interactions cause unavoidable uncertainties because each genotype is only planted in a limited set of locations and it is hard to detect them with such limited data. However, it is substantially important to account for these uncertainties as the G×E effects' contribution to different traits such as yield dramatically varies (Saltz et al., 2018) and

affects the genotypes' relative performance, i.e., their ranks. This makes the decision-making very challenging for breeders as G×E may be the primary contributor to rank uncertainty. This dissertation accounts for the uncertainty due to the limited set of planting locations to help breeders have reliable advancement decisions.

A great deal of research has been into advancement decision-making from different perspectives. In this regard, ranking is an essential component where selecting the best is of interest. Some researches predict the experimental genotypes' performance in unseen environments and do the ranking based on their predicted phenotypic traits (e.g., yield). Some rank them based on the information gained from their observed performance because they consider the fact that prediction might not be reliable while it is not easy to predict the weather and production conditions in advance. There are also studies that consider mixed models that take into account stability measures. In this dissertation, we develop a new methodology to capture and quantify the ranking uncertainty due to the G×E effect variability which is hard to be detected with a limited number of locations in each plant breeding experiment. We propose new tools to help breeders improve their decision-making.

1.2 Problem Statement

In a typical scenario, a commercial plant breeder may have thousands of experimental genotypes (e.g., soybean varieties or corn hybrids) to consider each year; and must decide as to which genotypes should be advanced and planted for at least one more year and stay in contention for becoming commercial genotypes. Such decisions are made based on many phenotypic properties and may, to a large extent, come down to ranking these genotypes based on one or more phenotypic traits.

Experimental genotypes are frequently ranked according to mean phenotypic response, and such a rank is then used as the basis of further decision-making, for example, to determine which experimental genotypes (e.g., soybean varieties or corn hybrids) should be advanced within a breeding program. However, this rank is problematic because while there often exists a large

amount of data available, early in a breeding program, relatively few observations may be available for each experimental variety to help breeders in decision making when determining if a crop will be advanced to the next stage of testing (grown for another year).

Breeding programs have practical limitations and capacity deficiencies to plant several thousands of genotypes in all potential locations. Therefore, it is not possible to make sure that sufficient information is earned about each experimental genotype's performance. This issue may greatly impact the relative performance of an experimental genotype to other competing genotypes, which impacts the ranking. The observed environments may result in one experimental genotype appearing better than its true genetic potential, whereas another appears worse, and it reverses their true relative performance. That is why rankings based on mean phenotypic response (e.g., yield) do not account for the effect of uncertainty that stems from only observing limited multi-environment trials (METs) for testing. Sometimes, it even becomes more challenging as such decisions are usually made under tight time constraints as the turnaround between harvesting and decision-making is very short.

There have been lots of methodologies in the literature that tries to address this issue by estimating the performance of the genotypes in unseen locations and ranking them based on their estimated values. We will show while they consider genetic-by-environment (G×E) interaction effects, they fail to capture the uncertainty of rank due to the limited number of locations in trials.

This research is dedicated to ranking, which is an essential step in advancement decision-making in real world where we face ranking under uncertainty. We develop a new methodology to capture the overall ranking uncertainty and provide rank confidence intervals to quantify the variability of rank. We also define a new ranking structure for experimental plant genotypes that differs significantly from what is assumed in the existing literature.

1.3 Dissertation Structure

The studies in this dissertation account for the uncertainty due to the set of planting locations, which are usually very limited, with the goal of providing valuable insights for plant breeders to

make better advancement decisions. The dissertation thus aims to apply data science techniques to fill the gap in capturing (G×E) interaction and improve the decision-making in plant breeding by investigating new models and methods that can be applied as decision support tools.

In the first research topic presented in chapter 2, we address the uncertainty inherent in making decisions based on a severely limited number of environments. In such trials, the relative rank of each genotype in each environment can be estimated and used to make advancement decisions. However, the uncertainty of the rank due to the the limited number of observed environments is generally unknown and this uncertainty is typically large due to the magnitude of the interactions (G×E effects) relative to the main effects that we are trying to discriminate. To address this shortcoming, in this chapter we propose using bootstrap resampling to estimate the probabilities of each rank and propose a procedure for converting those probabilities into a confidence interval of ranks. We evaluate the empirical coverage of those confidence intervals using simulated data, and present a case study that demonstrates its application and compares the new method to one of the widely-used existing approaches. Ultimately, this research aims to describe a new method for quantifying rank uncertainty that will assist plant breeders with advancement decisions.

In chapter 3, we investigate a new method for comparing genotypes. Namely, we propose to replace mean-based comparison with a probabilistic comparison that defines the best genotype as the one that is more likely to be the best across its target planting environments, versus the existing approach of defining the best genotype as the one that has the best mean. To estimate these probabilities, we again use resampling of environments or a bootstrap approach. Another similarity to chapter 2 is that the G×E effects are the key here, and we show that due to different G×E effects the probabilistic comparison is sometimes different than a simple mean comparison. We further evaluate the underlying reasons for these differences and show that the probabilistic comparison accounts for the uncertainty caused by observing limited environments.

Chapter 4 describes how the methods from chapter 2 and 3 can be applied for decision support in plant breeding. Specifically, we propose a ranking method based on the probabilistic

comparison of chapter 3 and describe an R package that we have developed to implement the ranking as well as the rank confidence intervals of chapter 2. As noted above, the probabilistic comparison that is the basis of the new ranking method, accounts for G×E interactions in addition to the mean performance. It needs to be mentioned that these interactions has been widely recognized before, and we thus compare the new probabilistic ranking with two popular ranking approaches that have been used with the same aim in plant breeding practice for decades.

Finally, the last section, chapter 5, briefly explains the conclusions that the research has derived along with future work suggestions.

1.4 Summary of Contributions

The contribution of this dissertation centers on accounting for the uncertainty that stems from the large and complex interactions between genotypes and environments, that is G×E effects, which are difficult to quantify due to limited environments in each breeding experiment. Even though in the context of agriculture, there is a great deal of statistical methods available to aid decision making, such methods usually provide a point estimate of the phenotypic response, such as yield, typically in combination with some other summary statistics aimed to capture stability. These summary measures are then combined and used as the basis for ranking experimental genotypes.

The main contributions of this thesis are three-fold and all revolve around the same central theme:

1. A method for using resampling to quantify the uncertainty of rank in plant breeding experiments is proposed and evaluated; and specifically a novel procedure that uses resampling estimates of rank probability to construct approximate confidence intervals of rank is developed.

2. In order to capture the full complexity of G×E interactions, a new approach is suggested for comparisons of genotypes, namely to prefer the genotype that is more likely to perform better across a set of environments. The relevant probabilities are again estimated using a

resampling procedure, and the differences between the proposed probabilistic and the standard means-based approaches are investigated and explained.

3. The methods proposed are implemented in an R package that can be used for decision support by providing probabilistic rank and rank confidence intervals. The new ranking is compared to existing approaches that have similar aims.

This dissertation thus provides novel insights into how plant breeding decisions should be made and results in methods that can be applied to assist plant breeders with advancement decisions.

1.5 References

Comstock, R. E. and Moll, R. H. (1963). Genotype x environment interactions. *Symposium on Statistical Genetics and Plant Breeding*, pages 164–196.

FAO (2017). *The future of food and agriculture – Trends and challenges*. Rome.

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 [Glycine max (L.) Merr.]. *Frontiers in Plant Science*, 12.

Saltz, J. B., Bell, A. M., Flint, J., Gomulkiewicz, R., Hughes, K. A., and Keagy, J. (2018). Why does the magnitude of genotype-by-environment interaction vary? *Ecol. Evol.*, 8(12):6342–6353.

CHAPTER 2. QUANTIFYING UNCERTAINTY OF RANK IN PLANT BREEDING EXPERIMENTS

Authors: Reyhaneh Bijari [1], Hanisha Vemireddy [1], and Sigurdur Olafsson [1]

[1] Department of Industrial and Manufacturing Systems Engineering, Iowa State University

2.1 Abstract

Plant breeders aim to select the experimental genotypes with the best genetic properties. In a typical scenario, a commercial plant breeder may have thousands of experimental corn hybrids or soybean varieties to consider each year; and must make a decision as to which hybrids or varieties should be advanced and planted for at least one more year. This decision, may to a large extent, come down to ranking these genotypes based on one or more phenotypic traits such as yield. However, in most cases, the uncertainty of the ranking makes this decision challenging because it is only possible to observe each experimental genotype in a relatively small number of environments. This may greatly impact the performance relative to other competing genotypes, which may significantly impact the ranking. The fundamental reason why this is true is that each experimental genotype has significant genetics-by-environment interaction (G×E) that will vary across the set of observed environments. Thus, the observed environments may result in one experimental genotype appearing better than its true genetic potential, whereas another appears worse, reversing the true relative performance. This paper proposes a new tool that plant breeders can utilize when analyzing the genotypes' performance. Specifically, to capture the inherent uncertainty due to the specific set of environments observed, we propose a bootstrapping approach to estimating the distribution of rank and constructing confidence intervals around it. We show through synthetic data experiments, constructed to mimic real observations of soybean yields, that the proposed approach is effective in the sense that the empirical coverage of the

confidence interval closely matches the theoretical coverage. These experiments demonstrate that this approach can be reliably used when evaluating plant breeding data. We also present a case study based on commercial soybean breeding data, demonstrating how these rank confidence intervals may be useful in practice.

2.2 Introduction

Plant breeders are routinely faced with the issue of experimental genotypes evaluation. After each growing season, experimental plant varieties (e.g., new soybean varieties or corn hybrids) are compared based on their observed yields and other phenotypic traits. Those that perform the best are advanced to the next year's field experiments, with the ultimate goal is to select those that have the potential to become commercial products. These decisions are usually made under tight time constraints as the turnaround between harvesting and decision-making is very short.

Breeders are typically interested in understanding both genetic effects (e.g., select the genotype with the highest yield) and genotype-by-environment (G×E) interactions to evaluate the stability and adaptability of experimental genotypes (Happ et al., 2021). There is a great deal of statistical methods available to aid in this analysis (Olivoto et al., 2019; Yang, 2007), with additive main effects and multiplicative interaction (AMMI) and best linear unbiased prediction (BLUP) being the most frequently used modeling approaches (Piepho, 1994; Piepho et al., 2008; Gauch, 2013; van Eeuwijk et al., 2016). Such methods provide a point estimate of the phenotypic response, such as yield, and while such can be used as the basis of ranking experimental genotypes, there appears to be much less attention within the plant breeding literature on direct estimates of rank and in particular, the uncertainty of rank. This paper aims to describe a new method for quantifying rank uncertainty that will assist plant breeders with advancement decisions. Specifically, we propose the use of bootstrapping to construct confidence intervals of ranks.

Ranking is an essential component of numerous decision-making processes where selecting the best is of interest, and there is, therefore a great deal of related literature in other application

domains. Gol (1996) defined a model-based uncertainty measure to make comparisons with the ranking concept in the education context. They suggested that rank intervals can be considered as a better estimator of schools' performances. There have been other methods that tried to provide uncertainty measures for individual ranks but continue failing to provide any measures of uncertainty of the overall ranking of populations (Aitkin and Longford, 1986; Laird and Louis, 1989).

Bootstrap sampling for confidence intervals has also been investigated before (DiCiccio and Efron, 1996; Efron, 1987). Seol (2016) used bootstrapped CIs to detect Rasch model fit statistics' misfitting under various simulation scenarios. It has received extensive attention in ranking as well. Hall and Miller (2010) studied both the theoretical and the numerical properties of bootstrap estimators of the data features' distributions on rankings. Wright et al. (2011, 2014) presented different uncertainty measures and proposed new uncertainty criteria to define estimates of ranks, both parametric and nonparametric, based on bootstrapping data. They introduced the concept of *the most probable ranking* and concluded that bootstrapping gives little better results of all presented methods. Wright et al. (2019) suggested alternate visualizations to overlapping/non-overlapping confidence intervals to improve the way of analyzing the rankings.

In the context of agriculture, there have been studies on ranking of phenotypic data. Simko and Linacre (2010) addressed the issue of partial rankings in plant breeding using the Rasch model and estimated final ranks. Simko et al. (2012) compared two methods of ranking, the rank aggregation approach and the projected values approach. They found these rankings and the integrated rating significantly correlated. A recent study introduced a mixed model for yield and compared the ranking results with the results of traditional yield stability measurements used for ranking (Happ et al., 2021). Literature has investigated bootstrapping for rank confidence intervals. In fact, previous studies primarily analyzed bootstrap CIs for rank given independent observations (DiCiccio and Efron, 1996; Efron, 1987). Such studies are therefore not directly applicable to the case of plant breeding, where the underlying structure involves subsets of observations with significant correlation among the observations within each subset. This is

important in practice because the success of bootstrapping methods depends heavily on the underlying structure. In this paper, we show how the proposed approach works well for both real and simulated plant breeding data and thus provides a useful method for plant breeders. While literature has tried to estimate point estimates of quantitative phenotypes (e.g. yield) and rank the genotypes based on their relative performance, in this paper, we try to fill the gap in the literature and capture the uncertainty of ranking due to limited multi-environment trials (METs) for testing.

2.3 Bootstrap Rank Confidence Interval Construction Methodology

2.3.1 Structure of Plant Breeding Experiments

When comparing different genotypes, breeders are interested in genotypes with the most favorable genetic gain which their high performances are stable facing different environmental conditions. Even though breeders tend to rank according to the main G effects, the genetic-by-environment interaction (G×E) effects contribute to the uncertainty of rank. In fact, they may be the primary contributor to the rank uncertainty. G×E effects are of interest with respect to stability and adaptability. If all the observations were directly representative of this genetic effect in the same manner, that is, they were simply observations of the genetic effect, comparing genotypes would be relatively straightforward. However, each genotype is planted in a fixed number of m environments, and for each of those environments, the genotype main effect is constant, whereas the G×E interaction effect varies. The observations of each genotype are thus highly correlated due to the main G effect, but yet drawn from different distributions, due to the varying G×E effects. It is thus the G×E effects that complicate the comparison and ranking of genotypes; as for the fixed planted locations, the average G×E effects may be very different for any two genotypes. Furthermore, this suggests that to capture the uncertainty of the ranking, our goal should be to capture the uncertainty due to only observing a limited number of environments. The following section suggests a procedure that aims to achieve this goal.

2.3.2 Bootstrapping to Capture Environment Sampling Variation

Let A denote a random subset of m_0 environments out of a total of m environments ($m_0 \leqslant m$). We assume that each environment j has an equal chance of being selected (observed) or $P(j \in A) = \frac{m_0}{m}, j = 1, 2, \ldots, m$. For a fixed set $A = a$, we assume that the rank $r_i(a)$ of each genotype can be calculated deterministically based on observations; that is, there is no uncertainty associated with the observations in each environment, only the set of environments that are observed. Neither of those assumptions is true in real plant breeding experiments, but we will demonstrate numerically that under confidence intervals constructed still provide useful coverage. Reflecting on the validity of the assumptions, we note that in real plant breeding experiments, planting locations and years are not selected at random, as most observations will be from a small number of years and locations may be planned, for example to evaluate each genotype in a variety of environments. Similarly, there is always uncertainty associated with observations in each environment, but we argue that due to the size of G×E effects relative to the main G effects, this is relatively small relative to the uncertainty due to the set of selected environments. The effect of deviations from both assumptions will be evaluated using simulated data in Section 2.4.

Let R_i denote the rank of genotype i in a random set A of m_0 environments. This is a random variable since the set of observed environments is random, but for a fixed $A = a$, its value is given by $r_i(a)$. We are interested in estimating the parameter $\mu_i = E[R_i]$, namely the expected rank of genotype i, $i = 1, 2, \ldots, n$. By conditioning on the observed subset of m_0 environments and letting S denote the space of all possible subsets, we observe that $\mu_i = E[R_i] = \sum_{a \in S} E[R_i | A = a] \cdot P(A = a) = \frac{1}{|S|} \sum_{a \in S} r_i(a)$, where we have simply used the assumption that each subset is equally likely. We can estimate the parameter of interest as $\hat{\mu}_i = r_i(A)$. We observe that this is an unbiased estimator:

$$E[\hat{\mu}_i] = E[r_i(A)] = \sum_{a \in S} E[r_i(A) | A = a] \cdot P(A = a) = \frac{1}{|S|} \cdot \sum_{a \in S} r_i(a) = \mu_i. \tag{2.1}$$

We have an estimate of the expected rank μ_i for each genotype $i = 1, 2, \ldots, n$; but this depends on the random set A of environments. Our main purpose in this paper is to capture the

uncertainty the selection of observed environments has on the estimate of this parameter, namely we are interested in making a statement regarding the distribution of μ_i. Specifically, we want to construct an interval I_i such that $P(\mu_i \in I_i) \geqslant 1-\alpha$, which we can think of as an approximate confidence interval.

Our approach uses bootstrap sampling to capture the variability in the set A of environments. Given B bootstrap samples $A_1, A_2, \ldots, A_B$, instead of a single rank, we now have B ranks $r_i(A_b)$ for each genotype and B estimates $\hat{\mu}_{ib} = r_i(A_b)$ of the parameter of interest. These estimates provide the basis for constructing a confidence interval for μ_i in addition to the point estimate. The specific procedure we propose is described below.

Procedure Rank CI

Step 0: Generate B bootstrap samples $A_1, A_2, \ldots, A_B$ of the set A of environments and select a confidence level, α.

Step 1: Estimate an empirical probability distribution $p_i(r) = P(\mu_i = r), i = 1, 2, \ldots, n$ for the expected rank of genotype i (the parameter of interest), based on the bootstrap samples as the number of occurrences of that rank in the B bootstrap samples of locations. Specifically, first calculate the rank $r_i(A_b)$ for each genotype $i = 1, 2, \ldots, n$; and each bootstrap sample $b = 1, 2, \ldots, B$. Then define an indicator variable

$$\chi_{irb} = \begin{cases} 1, & \text{if } r_i(A_b) = r \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

$i = 1, 2, \ldots, n; r = 1, 2, \ldots, n; b = 1, 2, \ldots, B$. Here χ_{irb} indicates if the rank of genotype i in the b^{th} bootstrap sample equals r or not. The estimate of the probability of having each rank is the proportion of bootstrap samples where this rank is observed, namely

$$\hat{p_i}(r) = \frac{1}{B} \sum_{b=1}^{B} \chi_{irb}. \tag{2.3}$$

Step 2: As the starting point of the confidence interval for the parameter μ_i, calculate the most likely rank of this genotype $i = 1, 2, \ldots, n$; namely the rank with the highest estimated probability:

$$r_i^{max} = \text{argmax}_{r \in 1,2,...,n} \hat{p}_i(r) \tag{2.4}$$

Set the initial upper and lower bounds of the confidence interval as this starting point

$$l_i = r_i^{max} \tag{2.5}$$

$$u_i = r_i^{max}. \tag{2.6}$$

If $\hat{p}_i(r_i^{max}) \geqslant 1 - \alpha$, then a confidence interval of the desired level has been constructed, stop and return $I_i = [l_i, u_i]$ as the final confidence interval for the rank of genotype i. Otherwise, continue to Step 3.

Step 3: Since $\sum_{r=l_i}^{u_i} \hat{p}_i(r) < 1 - \alpha$, the confidence interval must be enlarged. Thus, extend the interval by one rank position in each direction; that is, update the current upper and lower bounds of the rank interval as follows:

$$l_i = \min(l_i - 1, 1) \tag{2.7}$$

$$u_i = \max(u_i + 1, n) \tag{2.8}$$

If $\sum_{r=l_i}^{u_i} \hat{p}_i(r) \geqslant 1 - \alpha$, continue to Step 4. Otherwise repeat Step 3.

Step 4: Since the confidence intervals are extended both up and down simultaneously, the interval may have been constructed unnecessarily large, so consider removing one position. If $\hat{p}_i(l_i) \geqslant \hat{p}_i(u_i)$ and $\sum_{r=l_i}^{u_i-1} \hat{p}_i(r) \geqslant 1 - \alpha$; that is, the confidence interval still has the desired coverage after removing the upper bound position u_i from the interval, then make this change permanent:

$$u_i = u_i - 1.$$

If $\hat{p}_i(l_i) < \hat{p}_i(u_i)$ and $\sum_{r=l_i+1}^{u_i} \hat{p}_i(r) \geqslant 1 - \alpha$ that is, that is, the confidence interval still has the desired coverage after removing the lower bound position l_i from the interval, then make this change permanent:

$$l_i = l_i + 1. \tag{2.9}$$

Stop and return $[l_i, u_i]$ as the final confidence interval for the rank of genotype i.

Using this algorithm, first, we estimate the distribution of ranks for each genotype. We then calculate the confidence interval of ranks from this distribution.

NOTES

1. Given the assumptions made, the approximate confidence interval constructed is conservative, and sometimes we will systematically observe $P(r_i \in I_i) \geq 1 - \alpha$. In the extreme case, if $m_0 = m$, then the rank of each genotype can be inferred exactly from the observed environments, $\hat{p}_i(r_i^{max}) = 1$ and the procedure will construct intervals $I_i = r_i^{max}$, that is, terminate in Step 2 with $P(r_i \in I_i) = 1$. Since all the probability will be assigned to a single rank, the excess coverage will be α. Observing all, or even a large percentage of all environments is unrealistic in practice; but in general, we expect this excess coverage to increase as a function of the proportions of environments observed. Thus, under certain conditions this excess coverage may be large. This will be demonstrated via simulated data in Section 2.4 below.

2. On the other hand, if the assumptions do not hold true, the empirical coverage of the interval may not be the predicted coverage; and it could be lower than predicted even if the confidence interval is conservative. In particular, the following assumptions are made.

   (a) First, the construction assumes no noise given a specific set of environments, namely that $r_i(a)$ can be calculated given $A = a$. If the observations are noisy, the empirical coverage may be smaller than predicted. The effect of noise on the coverage of the confidence intervals will be evaluated via simulated data.

   (b) Second, the construction assumes that each environment is equally likely to be observed. This will never be true in practice, but the effect on the confidence interval coverage is unclear and will also be evaluated via simulated data.

2.4 Data

We argued above that the ability to obtain meaningful rankings of experimental genotypes is essential to informed advancement decisions in plant breeding. Furthermore, we argued that a meaningful ranking is not just limited to a single point estimate of ranks but also an evaluation of rank uncertainty and confidence intervals for the ranks. In this section, we analyze data from a commercial soybean breeding program and show how the proposed methods can provide valuable insights into advancement decisions. Capturing rank uncertainty can assist the breeders in individual advancement decisions as when they consider each genotype, they will have a confidence interval of rank that provides insights into the possible ranks of the genotype relative to others.

As noted above, our data comes from a commercial soybean breeding program. It contains a combination of phenotypic information of soybean varieties, namely yield, and managerial data such as planting date and harvesting date, with some field data such as location, latitude, and longitude. The whole dataset contains information of approximately 313 thousand genotypes in more than 700 locations and more than ten thousand experiments over ten years. The experimental locations are scattered throughout the United States and Canada. To illustrate the new methods proposed here, we focus on a small part of this data, namely a set of 38 experiments that are relatively uniform and planted in 2017. This results in the analysis of 730 genotypes (soybean varieties) over 166 test locations. A visualization of planting locations for all experiments is given in Figure 2.1.

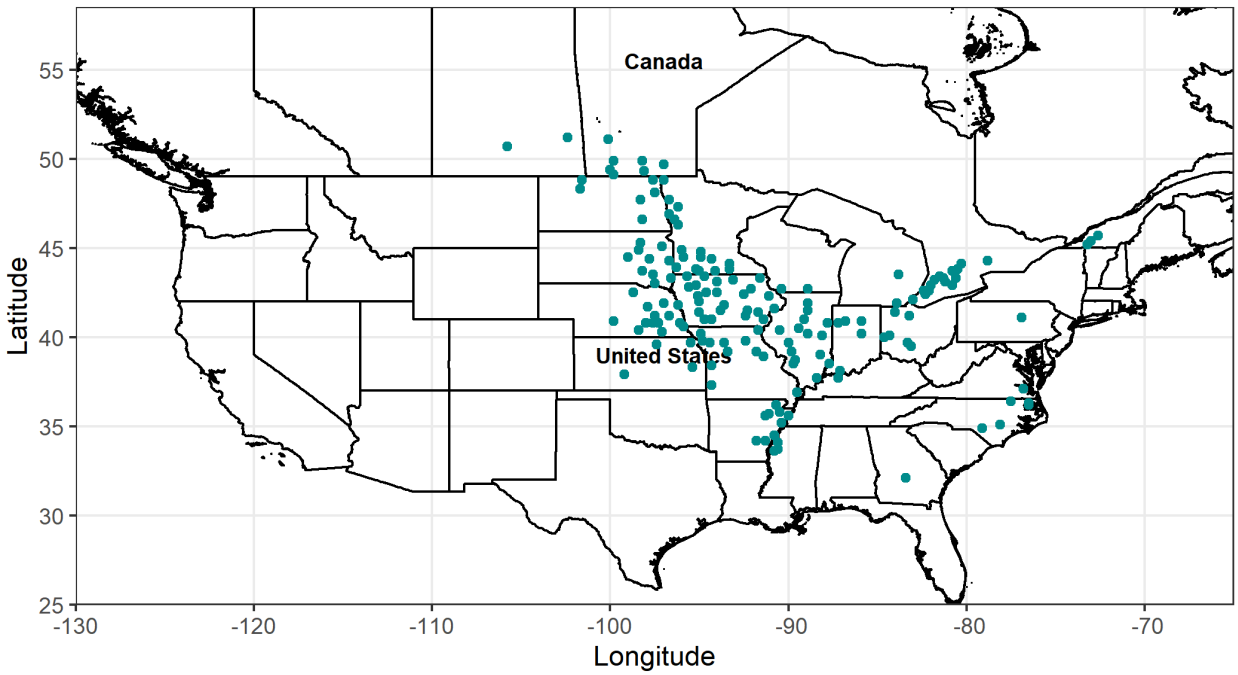

Figure 2.1: Planting locations for the experimental data.

2.5 Results

2.5.1 Correctness of Rank CIs

As noted in the introduction, bootstrapping procedures for rank confidence intervals are not new, but the success of such procedures will depend on the underlying data and its application, as well as the specifics of how the confidence intervals are constructed. The question is thus, how well this procedure work for phenotype data obtained from multi-environment trials (METs) of commercial crops? In other words, if we construct a $1 - \alpha\%$ confidence interval according to our procedure, will the empirical coverage be close to this predicted confidence level? To answer, ideally, we would like to experiment on real data derived from actual METs, and we will do that in the case study reported in the next section, but the drawback to using real data is that the ground truth is unknown. That is, there is no way for us to know the true rank of the genotypes.

To systematically evaluate the empirical rank of our confidence intervals, we simulate data that mimics real data but has a known ground truth; that is, the correct rank is known. To generate this data, we assume that the phenotype of interest is plant yield and assume the following linear model involving genetic (g_i), environmental (h_j), and genetic-by-environment interaction effects ($b_i h_j$) (Becker and Léon, 1988; van Eeuwijk et al., 2016).

$$\tilde{y}_{ij} = \mu + g_i + h_j + b_i h_j + \epsilon_{ij}. \tag{2.10}$$

To simplify the presentation, we assume that the response of interest is yield minus the environmental mean, eliminating the environmental effect from the above equation. The response thus only has two components: the genetic effect (G) and the genetic-by-environment (G×E) effects.

$$\begin{aligned} y_{ij} &= \tilde{y}_{ij} - (\mu + h_j) \\ &= g_i + b_i h_j + \epsilon_{ij}. \end{aligned} \tag{2.11}$$

We take Eq.2.11 as our starting point; namely, we assume a linear model $y_{ij} = g_i + b_i h_j + \epsilon_{ij}$ for the normalized yield of the i^{th} variety in the j^{th} environment. We then assume a distribution for $g_i \sim F_G$, $b_i \sim F_{GI}$, and $h_j \sim F_L$, and assume that $\epsilon_{ij} \sim N(0, \sigma)$ follows a normal distribution with zero mean.

For this study, distributions F_G, F_{GI} and F_L are determined by comparing simulated data to real yield data observed from a commercial soybean breeding program. Specifically, we attempt to mimic early-stage soybean experiments. For each of these experiments, we have 38 experimental varieties usually planted in a small number of locations within a single year in the Midwest, U.S. The normalized yield of a representative sample of six such experiments is plotted in Figure 2.2. Note that the yield is normalized by subtracting the environmental average. As expected, these distributions do not look perfectly identical; still, based on these observations, we conclude that a realistic simulated yield data would have a symmetric peak around zero with an approximate range of $[-25, 25]$. We experiment with three shapes and distributions of finite range distributions for both the genetic main effect (g_i) and the G×E interaction effects, contributed from the

environment (h_j), namely a uniform distribution, a beta distribution, and a normal distribution, along with three different shapes of G×E interaction effects, contributed from the genotype (b_i) following a beta distribution. In all cases, the sum of the two effects ($G + G \times E$) provides a reasonable fit to the empirical data. We also experiment with different ranges and shape parameters for each of the distributions considered. For the distributions considered for the g_i term, the uniform distributions have ranges $[-15, 15]$, $[-20, 20]$ and $[-30, 30]$, the normal distributions have mean of zero and standard deviations of $\frac{15}{3}$, $\frac{20}{3}$, and $\frac{30}{3}$, and the beta distributions used have scale parameters of 15, 20 and 30, and shape parameters of 5, 15 and 25. As the environmental effect is empirically twice as big as genetic main effect, we consider twice as big as the aforementioned distributions for h_j. The beta distributions considered for b_i term are $beta(8, 4)$, $beta(8, 8)$, and $beta(4, 8)$.

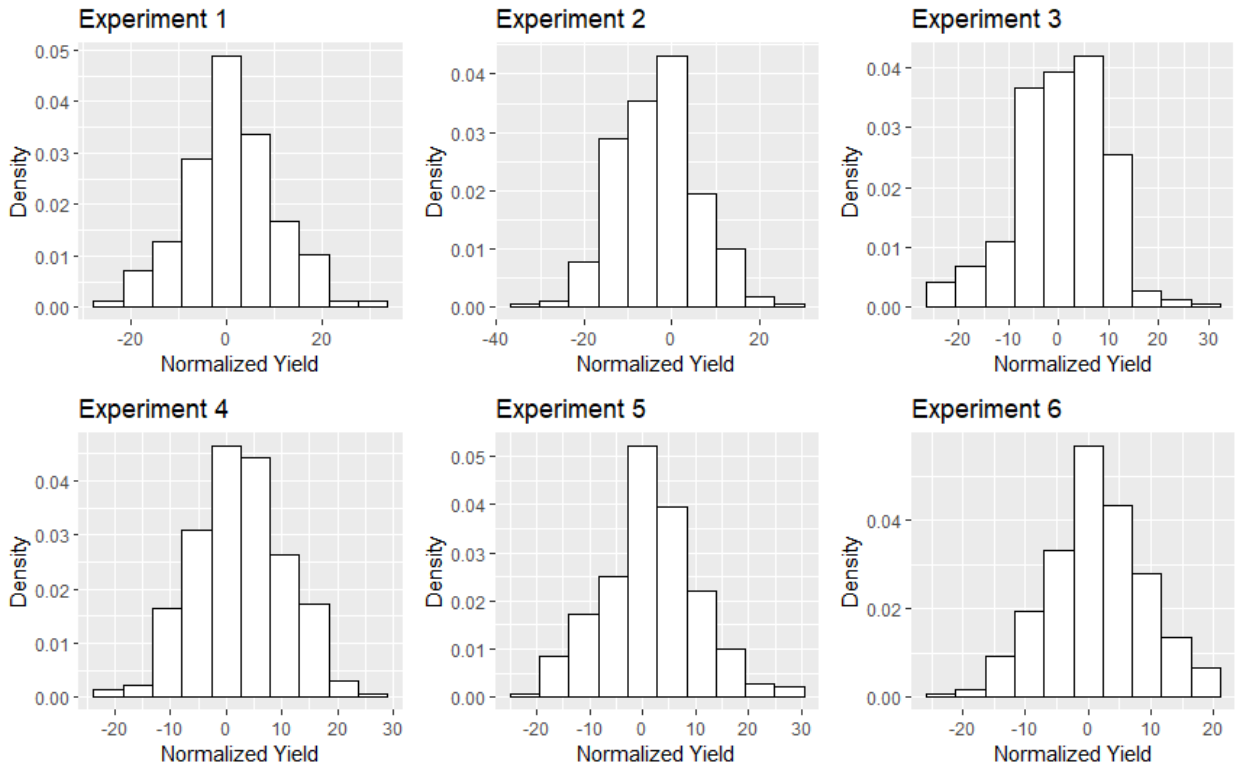

Figure 2.2: Empirically observed distributions for the normalized yields of commercial soybean experiments.

Using the distributions described above, we simulated the observed yield for 200 locations. The rank based on the average across these 200 locations is the ground truth rank. With the true rank being known, we can then evaluate how well the empirical coverage of the rank CIs matches with theoretical coverage. For example, if we construct rank CIs with $\alpha = 0.05$, we expect that those intervals include the ground truth rank 95% of the time, and for our procedure to work, we should observe empirical coverage close to this value.

To evaluate the empirical coverage, we mimic a MET where in practice, we can only observe a fraction of all possible locations. Specifically, we evaluate the empirical coverage for 10%, 50%, and 90% of the locations being observed. We also vary the variance of the normally distributed noise ($\epsilon = 1$ versus $\epsilon = 15$). These two parameters jointly determine the uncertainty, with more uncertainty when a small percentage of locations is observed (say, 10%) and the noise is high at each location ($\alpha = 15$). Finally, we repeat the experiment using different confidence levels ($\alpha = 0.05$, 0.10, 0.20), and the results for each of confidence levels are presented in Table 2.1, 2.2, and 2.3, respectively. Each empirical coverage number is based on ten replications for those specific parameter settings.

Note that each row in Table 2.1 corresponds to a specific shape of the main G effect and the G×E interaction effects but multiple parameters for those distributions for a fixed b_i and confidence level of $\alpha = 0.05$. For example, for a row with a left-skewed genotype contribution to the G×E effect (b_i), and a uniform main G effect (g_i), and a normal h_j (environment contribution to G×E effect), all three uniform ranges and all three standard deviation values of the normal distributions are used, each replicated 10 times, so each number in the row is an average of 90 empirical coverage values.

Table 2.1: Summary of Empirical Coverage at $\alpha = 0.05$

Fraction of Environments Observed			**10%**		**50%**		**90%**	
Noise (Std Dev of Normal)			1	15	1	15	1	15
Genotype Main Effect (G)	Environment Contribution to Interaction Effect (E in G×E)	Genotype Contribution to Interaction Effect (G in G×E)	**Empirical Coverage %**					
skewed beta	skewed beta	left skewed	93.72	87.04	98.89	92.98	99.97	94.13
skewed beta	beta	left skewed	92.33	84.91	98.65	91.90	99.97	93.79
skewed beta	normal	left skewed	91.88	85.07	98.64	92.24	99.98	93.75
skewed beta	uniform	left skewed	86.23	83.04	97.84	93.59	100.00	97.41
beta	skewed beta	left skewed	94.16	88.21	98.96	93.32	99.97	94.60
beta	beta	left skewed	93.01	86.68	98.74	92.91	99.98	94.19
beta	normal	left skewed	92.73	86.57	98.88	92.41	99.99	94.88
beta	uniform	left skewed	88.05	85.11	98.03	94.55	99.99	97.74
normal	skewed beta	left skewed	95.09	90.65	99.20	94.39	99.97	95.33
normal	beta	left skewed	94.55	89.95	99.05	94.66	99.97	95.51
normal	normal	left skewed	93.91	89.71	99.04	94.07	99.98	95.49
normal	uniform	left skewed	91.19	88.34	98.72	96.25	99.99	98.54
uniform	skewed beta	left skewed	95.97	91.91	99.41	95.25	99.99	96.17
uniform	beta	left skewed	94.89	91.50	99.30	95.19	100.00	96.17
uniform	normal	left skewed	94.64	91.26	99.33	95.18	99.99	96.75
uniform	uniform	left skewed	92.69	90.96	98.74	96.22	100.00	98.38
skewed beta	skewed beta	right skewed	94.84	87.18	98.78	92.49	99.75	93.44
skewed beta	beta	right skewed	93.80	85.12	98.60	91.42	99.78	92.58
skewed beta	normal	right skewed	93.66	84.96	98.70	91.36	99.90	92.43
skewed beta	uniform	right skewed	89.97	84.47	98.33	92.34	99.98	94.18
beta	skewed beta	right skewed	95.29	88.30	98.90	93.09	99.75	93.88
beta	beta	right skewed	94.48	86.72	98.78	92.12	99.83	93.40
beta	normal	right skewed	94.02	86.86	98.87	92.41	99.94	92.77
beta	uniform	right skewed	91.20	86.13	98.60	92.70	99.98	94.22
normal	skewed beta	right skewed	96.09	90.55	99.18	94.40	99.87	95.16
normal	beta	right skewed	95.76	90.18	99.17	94.14	99.91	94.69
normal	normal	right skewed	95.20	90.19	99.10	94.11	99.95	94.16
normal	uniform	right skewed	93.16	89.48	99.00	93.75	99.99	93.95
uniform	skewed beta	right skewed	97.13	91.92	99.39	95.00	99.83	95.74
uniform	beta	right skewed	96.44	91.42	99.50	94.50	99.90	95.89
uniform	normal	right skewed	96.17	91.51	99.39	94.92	99.96	95.96
uniform	uniform	right skewed	94.54	91.57	99.35	94.38	100.00	96.52
skewed beta	skewed beta	symmetric	94.21	87.08	98.85	92.72	99.93	93.72
skewed beta	beta	symmetric	93.18	84.85	98.72	91.61	99.95	93.39
skewed beta	normal	symmetric	92.82	84.91	98.74	91.60	99.96	93.26
skewed beta	uniform	symmetric	88.28	83.72	98.16	93.34	99.99	96.01
beta	skewed beta	symmetric	94.60	88.12	98.98	93.23	99.94	94.15
beta	beta	symmetric	93.73	86.57	98.83	92.49	99.96	93.83
beta	normal	symmetric	93.63	86.96	98.88	92.98	99.97	93.71
beta	uniform	symmetric	89.24	85.64	98.33	93.99	99.99	96.29
normal	skewed beta	symmetric	95.50	90.58	99.19	94.40	99.95	95.07
normal	beta	symmetric	95.09	90.15	99.18	94.08	99.95	95.16
normal	normal	symmetric	94.51	89.77	99.10	94.40	99.99	94.47
normal	uniform	symmetric	92.24	89.78	98.58	95.37	99.99	97.31
uniform	skewed beta	symmetric	96.49	92.02	99.46	95.29	99.93	95.98
uniform	beta	symmetric	95.85	91.75	99.35	94.89	99.97	96.10
uniform	normal	symmetric	95.57	91.32	99.44	95.35	99.99	95.23
uniform	uniform	symmetric	93.58	91.21	99.08	95.53	100.00	96.94

Table 2.2: Summary of Empirical Coverage at $\alpha = 0.10$

Fraction of Environments Observed			**10%**		**50%**		**90%**	
Noise (Std Dev of Normal)			1	15	1	15	1	15
Genotype Main Effect (G)	Environment Contribution to Interaction Effect (E in G×E)	Genotype Contribution to Interaction Effect (G in G×E)	**Empirical Coverage %**					
skewed beta	skewed beta	left skewed	89.42	80.76	97.33	88.06	99.86	89.62
skewed beta	beta	left skewed	87.40	78.15	96.73	86.58	99.87	89.19
skewed beta	normal	left skewed	86.77	78.39	96.76	87.13	99.91	89.10
skewed beta	uniform	left skewed	79.51	76.03	95.07	88.68	99.95	94.71
beta	skewed beta	left skewed	90.08	82.14	97.51	88.55	99.87	90.33
beta	beta	left skewed	88.33	80.37	97.01	87.87	99.89	89.96
beta	normal	left skewed	87.97	80.25	97.18	87.33	99.91	90.80
beta	uniform	left skewed	81.86	78.55	95.58	90.17	99.95	94.98
normal	skewed beta	left skewed	91.51	85.14	98.03	90.12	99.89	91.53
normal	beta	left skewed	90.63	84.38	97.67	90.47	99.88	92.01
normal	normal	left skewed	89.66	84.17	97.61	89.74	99.92	92.07
normal	uniform	left skewed	85.73	82.05	96.66	92.70	99.97	96.54
uniform	skewed beta	left skewed	92.96	86.93	98.52	91.52	99.91	92.89
uniform	beta	left skewed	91.19	86.35	98.26	91.34	99.92	92.76
uniform	normal	left skewed	90.83	86.08	98.33	91.41	99.94	93.51
uniform	uniform	left skewed	87.71	85.46	97.04	92.61	99.95	96.35
skewed beta	skewed beta	right skewed	91.18	80.94	97.28	87.36	99.33	88.78
skewed beta	beta	right skewed	89.61	78.52	96.90	85.98	99.39	87.57
skewed beta	normal	right skewed	89.38	78.30	97.00	85.95	99.62	87.34
skewed beta	uniform	right skewed	84.21	77.61	96.18	87.09	99.90	89.74
beta	skewed beta	right skewed	91.90	82.29	97.56	88.24	99.36	89.49
beta	beta	right skewed	90.56	80.39	97.23	86.88	99.51	88.66
beta	normal	right skewed	90.00	80.54	97.31	87.19	99.77	87.69
beta	uniform	right skewed	85.73	79.69	96.71	87.57	99.91	89.62
normal	skewed beta	right skewed	93.19	85.00	98.16	90.11	99.59	91.28
normal	beta	right skewed	92.62	84.72	98.02	89.85	99.66	90.68
normal	normal	right skewed	91.87	84.59	97.97	89.92	99.85	89.91
normal	uniform	right skewed	88.64	83.73	97.50	89.31	99.96	89.33
uniform	skewed beta	right skewed	94.76	86.95	98.66	91.14	99.58	92.27
uniform	beta	right skewed	93.66	86.28	98.66	90.49	99.73	92.46
uniform	normal	right skewed	93.29	86.35	98.57	91.16	99.86	92.61
uniform	uniform	right skewed	90.66	86.41	98.15	90.05	99.96	93.11
skewed beta	skewed beta	symmetric	90.16	80.76	97.32	87.66	99.74	89.11
skewed beta	beta	symmetric	88.58	78.20	96.92	86.19	99.78	88.51
skewed beta	normal	symmetric	88.09	78.29	97.00	86.18	99.85	88.35
skewed beta	uniform	symmetric	82.12	76.83	95.64	88.40	99.95	92.41
beta	skewed beta	symmetric	90.78	82.14	97.58	88.41	99.76	89.80
beta	beta	symmetric	89.42	80.22	97.31	87.29	99.81	89.23
beta	normal	symmetric	89.14	80.64	97.23	88.09	99.86	89.30
beta	uniform	symmetric	83.35	78.97	96.02	89.34	99.95	92.99
normal	skewed beta	symmetric	92.24	85.10	98.02	90.10	99.82	91.16
normal	beta	symmetric	91.58	84.47	97.93	89.54	99.80	91.24
normal	normal	symmetric	90.73	83.89	97.97	90.01	99.89	89.98
normal	uniform	symmetric	87.18	83.92	96.83	91.64	99.93	94.34
uniform	skewed beta	symmetric	93.81	86.91	98.67	91.58	99.79	92.58
uniform	beta	symmetric	92.63	86.48	98.42	90.87	99.87	92.77
uniform	normal	symmetric	92.28	85.89	98.45	91.69	99.94	91.66
uniform	uniform	symmetric	89.02	85.88	97.67	91.82	99.96	93.87

Table 2.3: Summary of Empirical Coverage at $\alpha = 0.20$

Fraction of Environments Observed			**10%**		**50%**		**90%**	
Noise (Std Dev of Normal)			1	15	1	15	1	15
Genotype Main Effect (G)	Environment Contribution to Interaction Effect (E in G×E)	Genotype Contribution to Interaction Effect (G in G×E)	**Empirical Coverage %**					
skewed beta	skewed beta	left skewed	80.91	69.96	92.88	78.54	99.03	80.76
skewed beta	beta	left skewed	77.90	67.04	91.57	76.59	99.09	79.93
skewed beta	normal	left skewed	76.98	67.22	91.47	77.25	99.30	79.98
skewed beta	uniform	left skewed	68.18	64.72	87.95	78.96	99.44	87.89
beta	skewed beta	left skewed	81.91	71.53	93.29	79.27	99.11	81.58
beta	beta	left skewed	79.32	69.53	92.22	78.36	99.20	81.42
beta	normal	left skewed	78.67	69.39	92.32	77.89	99.34	82.34
beta	uniform	left skewed	71.16	67.39	89.11	81.20	99.43	87.87
normal	skewed beta	left skewed	84.14	75.09	94.49	81.57	99.30	83.68
normal	beta	left skewed	82.46	74.09	93.65	82.04	99.30	84.53
normal	normal	left skewed	81.19	73.75	93.68	81.24	99.33	84.50
normal	uniform	left skewed	75.54	71.64	91.27	84.88	99.66	91.34
uniform	skewed beta	left skewed	86.29	77.39	95.65	83.77	99.40	85.82
uniform	beta	left skewed	83.64	76.36	95.11	83.45	99.43	85.29
uniform	normal	left skewed	82.95	76.66	94.86	83.39	99.45	86.26
uniform	uniform	left skewed	78.04	75.33	92.13	85.01	99.59	91.07
skewed beta	skewed beta	right skewed	83.66	70.22	93.16	77.68	97.66	79.73
skewed beta	beta	right skewed	81.20	67.44	92.16	75.85	97.77	77.95
skewed beta	normal	right skewed	80.76	67.21	92.29	75.82	98.27	77.63
skewed beta	uniform	right skewed	73.68	66.44	90.32	77.01	99.25	80.81
beta	skewed beta	right skewed	84.75	71.74	93.81	78.89	97.86	80.71
beta	beta	right skewed	82.72	69.57	92.96	77.19	98.15	79.52
beta	normal	right skewed	81.88	69.82	93.08	77.49	98.66	78.38
beta	uniform	right skewed	75.78	68.71	91.34	77.58	99.33	80.59
normal	skewed beta	right skewed	86.98	75.20	95.25	81.49	98.37	83.28
normal	beta	right skewed	85.67	74.33	94.80	81.02	98.55	82.86
normal	normal	right skewed	84.46	74.26	94.66	81.24	99.04	81.73
normal	uniform	right skewed	79.69	73.16	92.89	80.22	99.52	79.54
uniform	skewed beta	right skewed	89.65	77.42	96.42	83.11	98.65	84.90
uniform	beta	right skewed	87.42	76.57	96.29	82.01	98.99	85.28
uniform	normal	right skewed	86.75	76.79	95.87	83.06	99.27	85.52
uniform	uniform	right skewed	82.50	76.77	94.59	81.66	99.42	85.36
skewed beta	skewed beta	symmetric	81.97	69.92	93.02	78.06	98.66	80.04
skewed beta	beta	symmetric	79.56	67.19	92.01	76.05	98.80	79.02
skewed beta	normal	symmetric	78.87	67.14	92.21	76.14	99.02	78.90
skewed beta	uniform	symmetric	71.31	65.44	89.25	78.79	99.40	84.26
beta	skewed beta	symmetric	82.89	71.48	93.59	79.14	98.75	80.96
beta	beta	symmetric	80.90	69.35	92.86	77.71	98.93	80.24
beta	normal	symmetric	80.30	69.74	92.60	78.52	99.17	80.45
beta	uniform	symmetric	72.77	67.90	90.04	80.09	99.46	85.34
normal	skewed beta	symmetric	85.40	75.01	94.70	81.53	99.06	83.25
normal	beta	symmetric	84.08	74.16	94.19	80.65	99.04	82.97
normal	normal	symmetric	82.57	73.91	94.34	81.23	99.27	81.62
normal	uniform	symmetric	77.50	73.80	91.60	83.24	99.51	88.01
uniform	skewed beta	symmetric	87.73	77.30	96.16	83.76	99.08	85.26
uniform	beta	symmetric	85.86	76.68	95.49	82.75	99.31	85.37
uniform	normal	symmetric	85.09	76.14	95.58	83.80	99.41	84.59
uniform	uniform	symmetric	80.00	75.67	93.25	83.92	99.45	87.01

From the empirical coverage results reported in Table 2.1, we can make several observations. First, the empirical coverage is good overall and for each of the three confidence levels, the empirical coverage matches or exceeds the theoretical coverage for most parameter settings. Second, the shape of genotype contribution to the G×E effect (b_i), seems to have not a significant relationship with the empirical coverage, and this can be seen in Figure 2.3.

Third, there is a relationship between the shape of the main G effect distribution and the shape of the G×E interaction effects, especially E's contribution in G×E interaction effects (h_j) with the empirical coverage. The coverage appears to be lower for the skewed beta distribution and highest for the normal and uniform distributions regarding the main effect. This is further illustrated in Figure 2.4, from which we can also note that this relationship is consistent regardless of the level of uncertainty (that is, fractions of locations observed and the random noise at each location).

On the other hand, uniform distribution of E's contribution in G×E interaction effects (h_j) has the lowest empirical coverage in the scenario where both a small number of locations is observed and the noise is low (see the top-left plot in Figure 2.5). In fact, in this scenario, the relationship is exactly the opposite as for the main effect. We note that for the main effect, the distribution is across genotypes, whereas for the interaction effects, the distribution is across locations, which may explain the differences in behaviors. If the distribution of the main G effect is such that many genotypes have very similar effects, i.e. skewed beta, then the coverage of the CIs is observed to be lower, whereas if the main effects of each genotype are uniformly distributed (essentially spaced somewhat evenly apart), they are easier to rank, and the empirical coverage is observed to be higher.

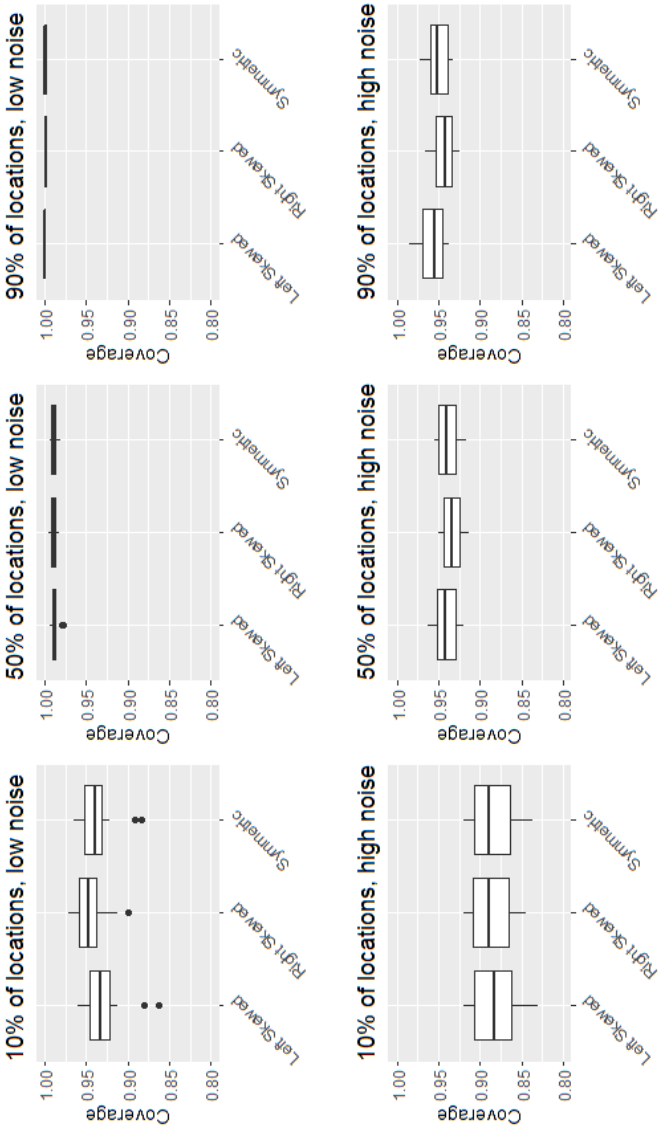

Figure 2.3: Empirical coverage for different shapes of b_i with 95% theoretical coverage.

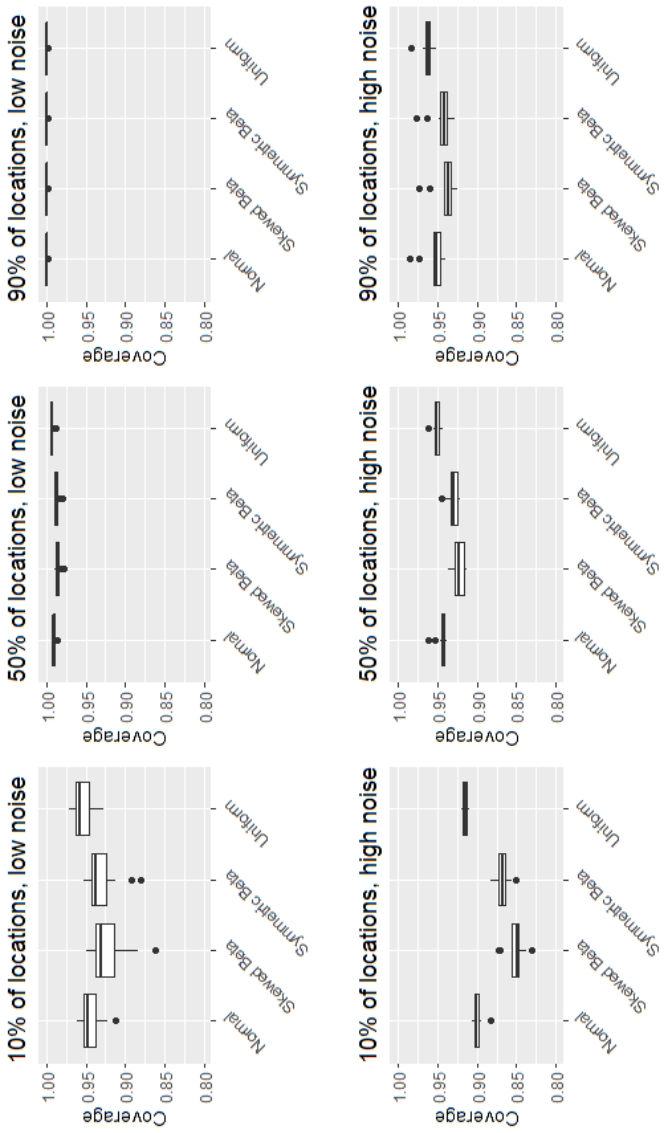

Figure 2.4: Empirical coverage for different shapes of main G effect (g_i) with 95% theoretical coverage.

On the other hand, the uniform distribution of G×E interaction effects across locations may exacerbate the difficulties due to only observing a fraction of all possible locations. This observation is further supported by the results in Figure 2.5, as we note that as the percentage of locations observed becomes large (90%), the lower empirical coverage for uniform h_j effects completely disappear. In practice, the shape parameters of these distributions could be estimated from the observed METs data to determine if we expect empirical coverage to be slightly higher or slightly lower than the theoretical coverage.

The third and final observation that is evident from all three figures is that less uncertainty implies higher empirical coverage and vice versa. This is true for both uncertainties due to the limited number of locations being observed and for the noise observed at each location. This is not an unexpected observation, but for analysis of non-simulated data, the amount of uncertainty estimate in the experiment would give us an idea of the empirical coverage of the rank CIs relative to other experiments.

The proposed rank confidence intervals work as expected; that is, the observed empirical coverage is close to the predicted theoretical coverage. If most of the possible environments are observed, then the empirical coverage will tend to be higher than the theoretical coverage, and vice versa when only a small percentage of locations is observed. Higher observation noise will also lead to worse empirical coverage, especially when the main G effect has a peaked and skewed distribution, in which many genotypes will have a similar main G effect. If only a few locations are observed but with little uncertainty, then the uniform distribution of the G×E effects with respect to environments would also lead to reduced empirical coverage. However, even in the worst scenarios, the difference in empirical and theoretical coverage is sufficiently small that we believe the rank CIs will still be useful. Finally, we note that even though we have focused on illustrating the patterns at $\alpha = 0.05$ significance level, the results reported in Table 2.1 illustrate the same patterns held at other significance levels.

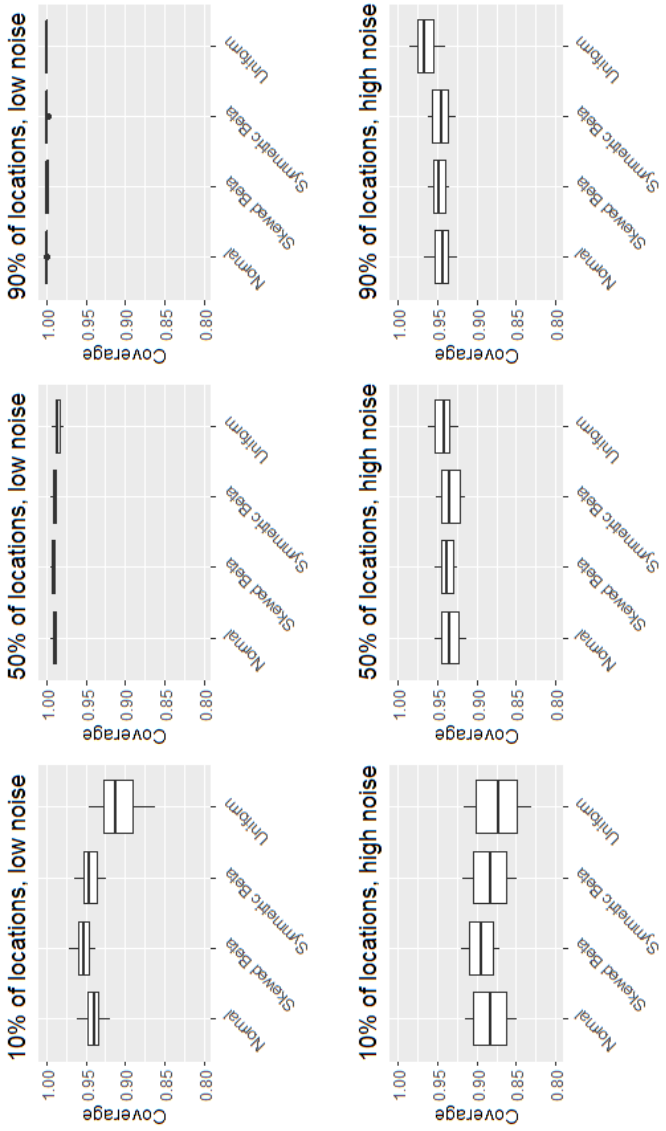

Figure 2.5: Empirical coverage for different shapes of h_j with 95% theoretical coverage.

2.5.2 Application of Rank Confidence Intervals

This section illustrates a practical use of rank confidence intervals and compares the information gained from the rank confidence intervals with the information obtained from a popular method for ranking in agriculture context, namely, BLUP analysis as it is a popular method for ranking in agriculture. For these results, the BLUP analysis was performed using the *metan* package in R (Olivoto and L'ucio, 2020) and the rank CIs are also obtained using R. The order in which the rank confidence intervals for genotypes are displayed is independent of the construction of the confidence intervals. Here, they are displayed in an order the same as in the predicted response. We performed the analysis for two soybean experiments. The first one is an example of an experiment with relatively low uncertainty and a clear winner. The second one has much higher uncertainty and more ambiguity regarding which experimental soybean variety is the best.

The BLUP analysis and the rank CIs for the first experiment are shown in Figures 2.6 and 2.7. In this experiment, there is one soybean variety, G4, that clearly outperforms all others. This is indicated by both the BLUP output and the rank CIs. In such transparent cases, the breeder will likely recognize this without the assistance of any statistical method. The more interesting comparison happens for cases that are less transparent. For example, varieties G11 and G9 appear virtually identical when comparing the BLUP confidence intervals, whereas the rank confidence intervals, while still largely overlapping, show that G11 has a significantly higher probability of ranking in the top 5 spots while G9 have just 4% probability of being in top 5 ranks within 80% confidence interval. A more interesting contrast comes from the width of the confidence intervals.

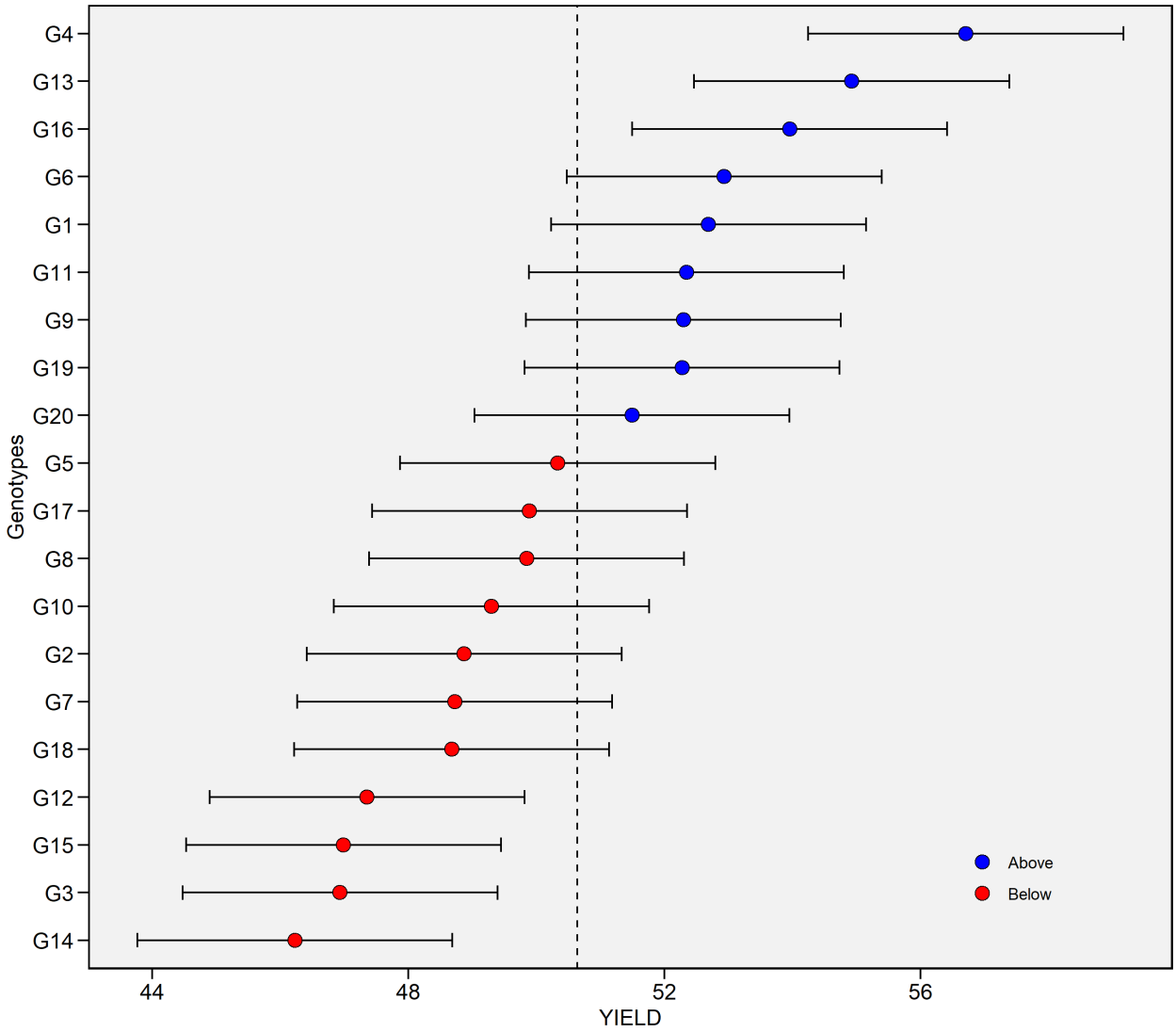

Figure 2.6: BLUP confidence intervals of yielD for a certain experiment.

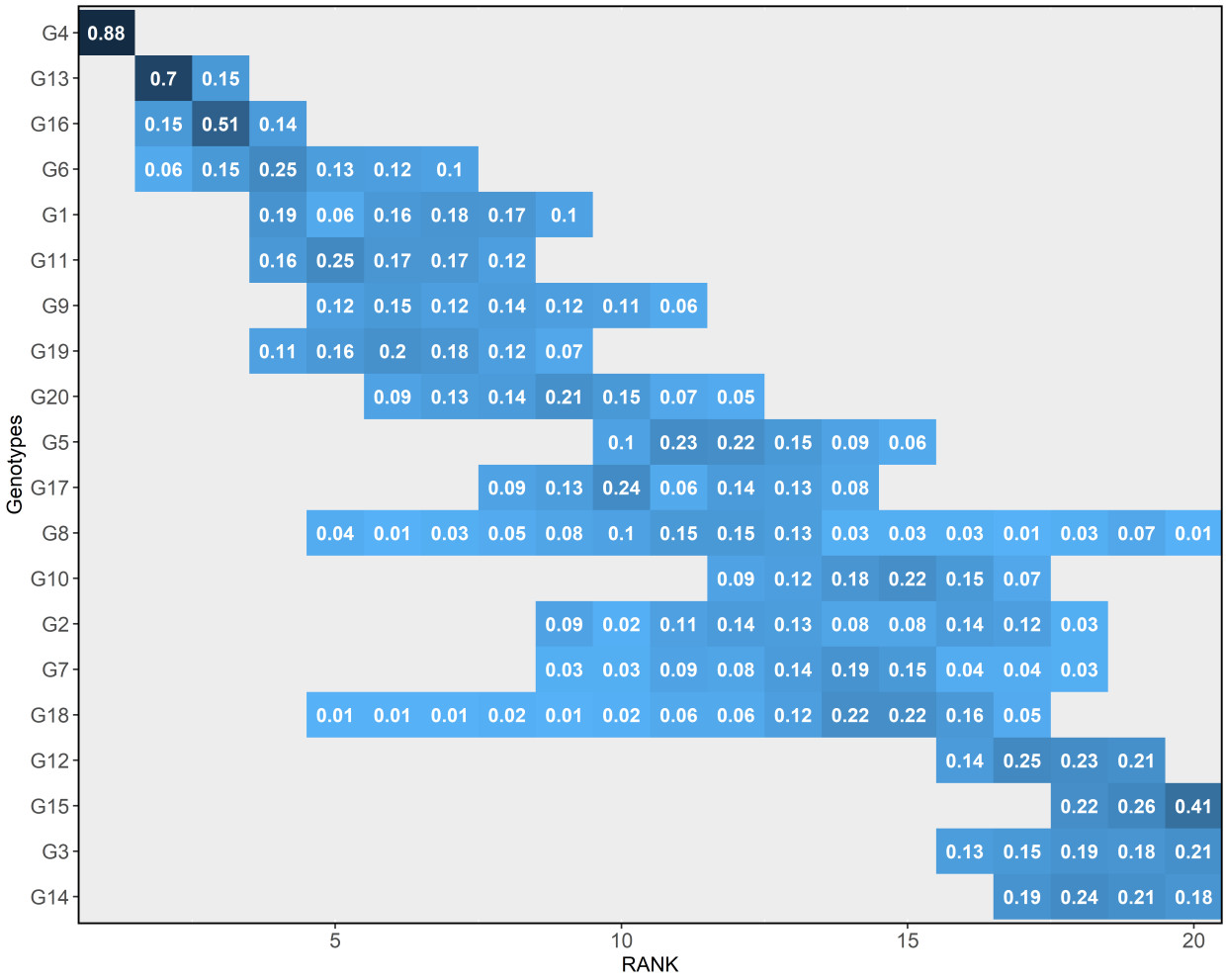

Figure 2.7: Rank confidence intervals of yield for a certain experiment.

The width of the confidence intervals for the BLUP predictions, as shown in Figure 2.6, does not depend on the individual genotype, whereas the width of the rank CIs, Figure 2.7, varies significantly from one variety to the next. For example, since G4 ranks 1^{st} with estimated probability of $0.83 > 0.80$, the width is only one spot. Similarly, the probability of G13 being in the top 3 exceeds 0.8, so the width is only three spots. The reason for these tight confidence intervals is the relative difference in main effects, namely, relative to the other varieties in this experiment. The yields of varieties G4 and G13 are easily seen as best and second best, respectively.

As noted above, statistical tools such as BLUP or rank CIs may be viewed as more beneficial for breeders when decisions are not as clear. The BLUP predictions indicate a very similar yield for G11 and G9. However, while the rank CI for G11 is tight (rank 4–7), the confidence interval for G11 spans 16 rank positions (rank 1–19). The intuitive explanation for this is in the relative G×E effects, that is, the G×E effects relative to other varieties in the experiment. Variety G11 performs more predictable compared to G9 while G9 has a highly variable rank. It should be noted that the BLUP analysis also estimates the G×E effects, and it would be possible to infer similar conclusions about G×E from those as what we obtain from the rank confidence interval width. Another key difference that should be emphasized is that while the BLUP analysis provides an estimation of the main genotype effects (and the G×E effects, although they are not shown here), the rank CIs provide information on the relative main genotype effect (and indirectly on the relative G×E effects via the CI width).

The analyses for the second experiment are shown in Figures 2.8 and 2.9. For this experiment, it is much less clear which variety is the top variety. The BLUP analysis predicts G21 as having the highest yield, but the prediction interval overlaps significantly with all of the above-average varieties.

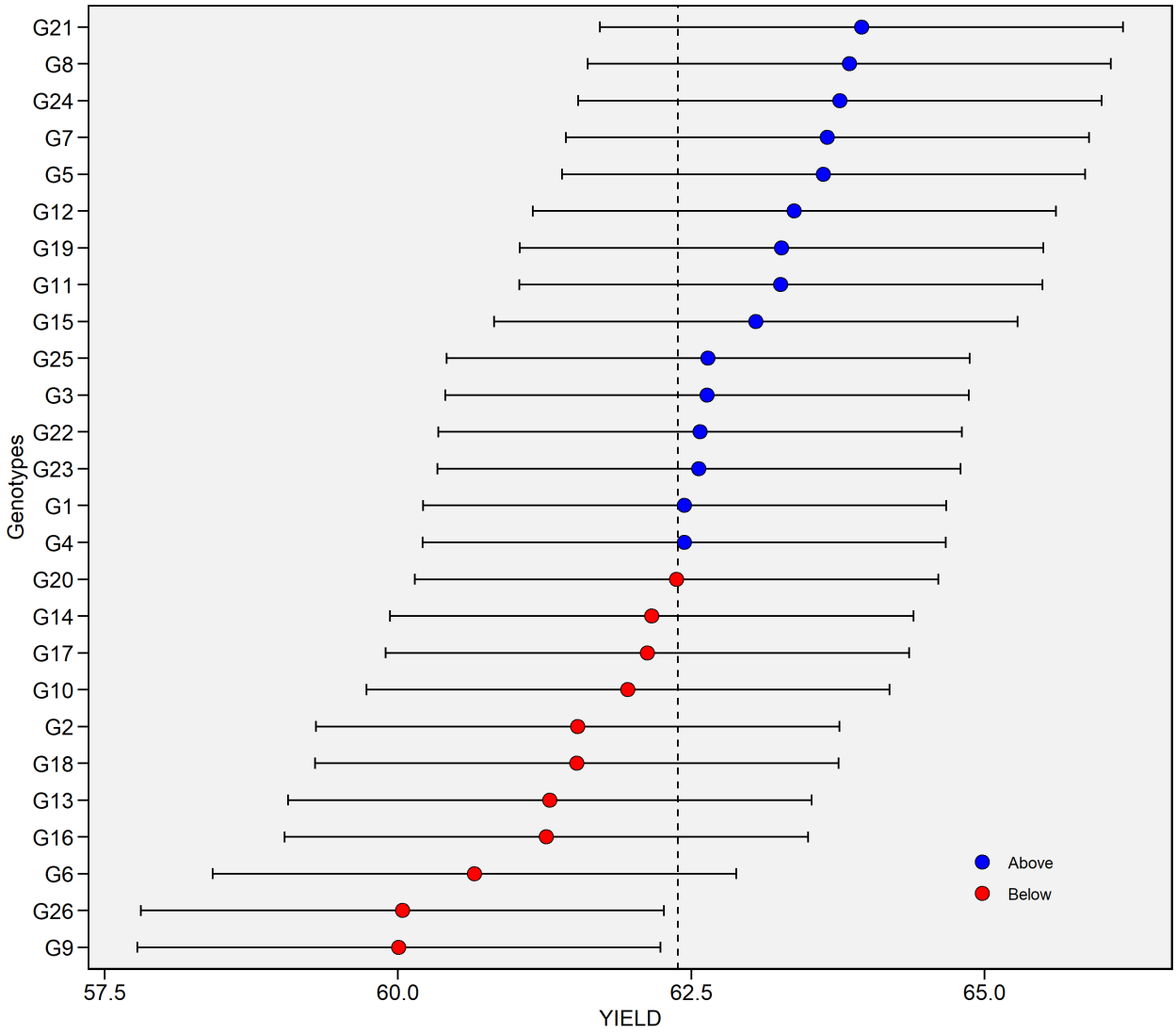

Figure 2.8: BLUP confidence intervals of yielD for a certain experiment.

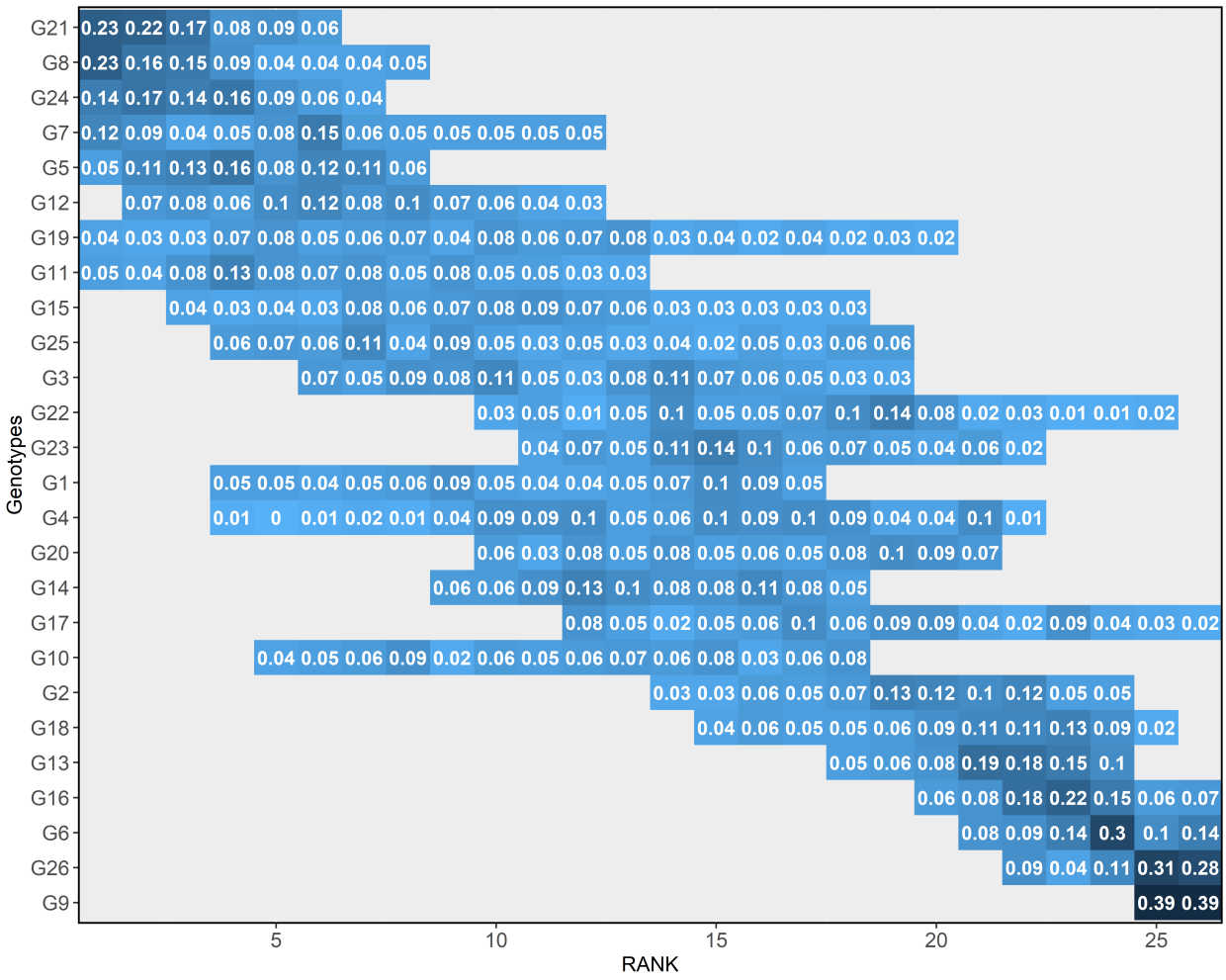

Figure 2.9: Rank confidence intervals of yield for a certain experiment.

The BLUP analysis also indicates that G21, G8, and G24, G7, and G5 are almost the same. On the other hand, the rank CIs provide more insights into this set of varieties and show that G21 has the highest probability of top rank among these five and it has the tightest confidence interval (ranks 1–6), indicating good performance in many locations, whereas the CI for G8 includes rank 1-8, and for G7 includes 1–12 spots with the probability of top rank being much smaller. Several varieties have very wide CIs, including G19 (rank 1–20) and G11 (rank 1–13). These varieties rank 7^{th}, and 8^{th} respectively, in the BLUP analysis. Their rank CI widths relative to other top varieties have an additional piece of information that breeders can consider for advancement decisions. While these genotypes ranked 7^{th} and 8^{th} in the BLUP output, they have the probability of 4% and 5% for the being the top variety. Rank CIs help the breeder to identify such genotypes in the case BLUP analysis cannot detect between them. Furthermore, the rank CIs would help separate these genotypes as G11 has a CI width much smaller than G19, perhaps indicating higher stability, as well as a higher probability of top rank.

2.6 Conclusions

We have proposed novel rank confidence intervals for plant breeding experiments. Such rank CIs do not provide a prediction of the response, and they are not intended to replace existing models such as BLUP or AMMI but rather provide a complementary analysis. This has been demonstrated via a comparison with a BLUP analysis. While the BLUP analysis provides an accurate prediction of the response (yield), the rank CIs focus explicitly on rank and hence the relative performance. In practice, the primary purpose of breeders is often comparison and selection among two or more genotypes. In such cases, where relative performance is most important, the new method may prove particularly helpful as it tries to capture the uncertainty due to the limited environments in which genotypes have been planted.

For further insights into the rank CIs, we argue that the width of the rank CIs captures two aspects of relative performance: separation in response for the genotypes (essentially the genotype main effects) and the relative G×E interaction effects as tighter CIs imply more stable genotypes

if the main effects are similar. This provides further information that complements traditional stability analysis.

2.7 Acknowledgments

Syngenta Seeds Inc Grant supports this work. We thank them for their support. Special thanks to Jerad Benson, Antoine Botrel, Greg Doonan, Andy Kuhl, Hieu Pham, Marcia Almeida De Macedo, and Ye Han for their insightful discussions and recommendations. We would also like to express our sincere thanks to our Iowa State University colleagues who have worked with us on this project and provided thoughtful feedback, especially Samira Karimzadeh and Stephen Vardeman.

2.8 Conflict of Interest

The authors declare that they have no conflict of interest.

2.9 References

(1996). League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159(3):385–443.

Aitkin, M. and Longford, N. (1986). Statistical Modelling Issues in School Effectiveness Studies. *Journal of the Royal Statistical Society: Series A (General)*, 149(1):1–26.

Becker, H. C. and Léon, J. (1988). Stability Analysis in Plant Breeding. *Plant Breeding*, 101(1):1–23.

DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3).

Efron, B. (1987). Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, 82(397):171–185.

Gauch, H. G. (2013). A Simple Protocol for AMMI Analysis of Yield Trials. *Crop Science*, 53(5):1860–1869.

Hall, P. and Miller, H. (2010). MODELING THE VARIABILITY OF RANKINGS. *The Annals of Statistics*, 38(5):2652–2677.

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 [Glycine max (L.) Merr.]. *Frontiers in Plant Science*, 12.

Laird, N. M. and Louis, T. A. (1989). Empirical Bayes Ranking Methods. *Journal of Educational Statistics*, 14(1):29–46.

Olivoto, T. and L'ucio, A. D. (2020). metan: an r package for multi-environment trial analysis. *Methods in Ecology and Evolution*, 11(6):783–789.

Olivoto, T., Lúcio, A. D. C., da Silva, J. A. G., Marchioro, V. S., de Souza, V. Q., and Jost, E. (2019). Mean Performance and Stability in Multi-Environment Trials I: Combining Features of AMMI and BLUP Techniques. *Agronomy Journal*, 111(6):2949–2960.

Piepho, H. P. (1994). Best Linear Unbiased Prediction (BLUP) for regional yield trials: a comparison to additive main effects and multiplicative interaction (AMMI) analysis. *Theoretical and Applied Genetics*, 89(5).

Piepho, H. P., Möhring, J., Melchinger, A. E., and Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*, 161(1-2).

Seol, H. (2016). Using the Bootstrap Method to Evaluate the Critical Range of Misfit for Polytomous Rasch Fit Statistics. *Psychological Reports*, 118(3):937–956.

Simko, I., Hayes, R. J., and Kramer, M. (2012). Computing integrated ratings from heterogeneous phenotypic assessments: A case study of lettuce postharvest quality and downy mildew resistance.

Simko, I. and Linacre, J. (2010). Combining partially ranked data in plant breeding and biology: II. Analysis with Rasch model. *Communications in Biometry and Crop Science*.

van Eeuwijk, F. A., Bustos-Korts, D. V., and Malosetti, M. (2016). What should students in plant breeding know about the statistical aspects of genotype × environment interactions? *Crop Science*, 56:2119–2140.

Wright, T., Klein, M., and Wieczorek, J. (2011). An Overview of Some Concepts for Potential Use in Ranking Populations Based on Sample Survey Data.

Wright, T., Klein, M., and Wieczorek, J. (2014). *Ranking Populations Based on Sample Survey Data*.

Wright, T., Klein, M., and Wieczorek, J. (2019). A Primer on Visualizations for Comparing Populations, Including the Issue of Overlapping Confidence Intervals. *The American Statistician*, 73(2):165–178.

Yang, R.-C. (2007). Mixed-Model Analysis of Crossover Genotype-Environment Interactions. *Crop Science*, 47(3).

CHAPTER 3. ACCOUNTING FOR G×E INTERACTIONS WHEN COMPARING PHENOTYPIC RESPONSE: A PROBABILISTIC APPROACH

Authors: Reyhaneh Bijari [1], Sigurdur Olafsson [1]

[1] Department of Industrial and Manufacturing Systems Engineering, Iowa State University

3.1 Abstract

In the plant breeding multi-environment trials, the mean phenotypic responses of two experimental genotypes are often close and depend on the specific target environments in which they are observed. This makes a simple comparison of the mean response problematic in practice. We propose a new approach to comparing genotypes that selects the genotype that is more likely to perform better rather than the one that has the better mean. Our implementation uses bootstrap resampling to estimate the probability that one genotype outperforms another in a set of observed environments. This not only provides more information, that is, the probability of one genotype being preferable versus a simple mean comparison, but we also show that due to the different G×E effects, the probabilistic comparison is sometimes different than a simple mean comparison. We further evaluate the underlying reasons for these differences and show that the probabilistic comparison accounts for the uncertainty caused by observing limited environments.

3.2 Introduction

Plant breeding is costly and time-consuming and it takes years to go from the initial experimental stages to commercialization (Luckett and Halloran, 2017). Each year, plant breeders want to select the best genotype for the potential target environments by assessing their relative performance to the other experimental genotypes planted in a subset of target environments

(Abberton, 2012; Vargas et al., 2013; Cooper and DeLacy, 1994). They are faced with a difficult selection decision. Given a large set of experimental plant varieties or hybrids, which should be selected as having the highest potential, that is, which genotypes should be advanced and planted for at least one more year. This selection is always imperfect due to the amount of randomness in the observed performance of these genotypes. There are practical limitations on the size of the subset to be selected and advanced. Still, in the end, there are only a few genotypes that become a success – for a commercial plant breeder, become commercial varieties or hybrids. The ultimate concern is not to miss out on a potential success, even if it happened to perform relatively poorly in a single year of experiments. The cost of a miss is much higher than the cost of including a genotype that ultimately does not succeed. Indeed, it is expected that most of the genotypes that are selected and advanced will ultimately not succeed.

At its core, this complex selection process comes down to making pair-wise comparisons between genotypes (Vargas et al., 2013). A pair-wise comparison may incorporate many factors. For example, a set of experimental soybean varieties with a similar relative maturity (RM) will be compared according to their normalized yield, where the normalization might take into account such factors as the planting locations and the RM values. Typically, when it comes to finding the best genotypes, comparison is made based on some mean comparison of one or more phenotype, while also considering some measures of stability or adaptability. The mean may be observed directly or predicted using some standard statistical models. While it may have corrective factors and normalization, but it inherently remains a comparison of means. In this paper, we suggest that mean comparison may not be the best approach for advancement selection, but instead the comparison should be based on comparing probabilities of performing better across a set of planting environments. In other words, we suggest that the selected genotype should be the genotype that is more likely to perform better rather than the genotype with the better mean. Of course, selecting based on mean and selecting based on such probability will often agree. For example, if the mean difference is large. However, it is likely to be different in important cases, such as for two genotypes with difficult to discriminate means and when we only observe a small

portion of possible environments. Additionally, we will show that the differences depend on the G×E structure and magnitude.

3.3 Motivating Example

As noted in the introduction, it is a standard practice in plant breeding to compare experimental genotypes according to their observed or predicted mean phenotype (most commonly yield) and select the genotype with the better mean for advancement (Schrama et al., 2018). While this is a very intuitive approach, it has some limitations that will be explored in this paper. To obtain some insights into those limitations, we start by presenting a simplified motivating example.

Suppose we have two soybean varieties, Variety 1 and Variety 2, that could be planted in five environments, E_1, E_2, E_3, E_4, E_5, and the phenotype of interest is yield measured in bushels/acre. Note that these five environments represent the entire universe of possible environments (locations and years), so in practice this would be a very large number of environments. The yield of Variety 1 is 73.0, 51.0, 61.0, 48.0, and 55.0 in the five environments, respectively; and the yield of Variety 2 is 63.5, 53.5, 59.5, 50.5, and 56.5 in the five environments, respectively. By calculating the average we can observe that Variety 1 has the better mean across all possible environments, or 57.6 versus 57.1. Variety 1 also has the better average rank (1.4 versus 1.6) since it is has higher yield in more environments. However, a crop will never be planted in all possible environments, not even within a single year, so another quantity of interest would be which variety is *more likely* to have higher yield when planted in some fixed number of environments? For example, we select two out of the five environments at random. There are exactly ten such pairs and it is easy to verify that Variety 1 performs better in only four out of those ten pairs. Specifically, Variety 1 does better in every pair that includes the highest-yielding environment E_1 and Variety 2 does better in every pair that does not include E_1. Thus, we can conclude that even though Variety 1 has a higher mean yield it will only have higher yield 40% of the time if the two varieties are planted in two randomly selected environments.

We argue that this small-scale example mirrors what happens in real situations, where there is a large set of potential target environments, and each year a crop is only planted in a very small subset of all possible environments. This example therefore motivates the main idea of this paper, which is to have probabilistic comparison along with mean-based comparison.

A couple of important observations need to be made about the motivating example. First, this scenario would not happen if the varieties had the same G×E structure. The yield in this example is calculated according to $y_{ij} = 50 + g_i + e_j + h_{ij}$, where $g_i \in \{3, 2.5\}$ is the genetic effect of the two varieties, $e_j \in \{10, -2, 4, -5, 1\}$ is the environmental effect and the G×E interaction (h_{ij}) is given in Table 3.1.

Table 3.1: G×E effects values for the motivating example.

G×E effects	E_1	E_2	E_3	E_4	E_5
Variety 1	10	0	4	0	1
Variety 2	3	3	3	3	3

We note that $\sum_{j=1}^{5} h_{1j} = \sum_{j=1}^{5} h_{2j} = 15$ (row sums in the table), and on the average the interaction is therefore the same; but whereas Variety 2 is very stable, Variety 1 is able to double the environmental effect of good environments but is neutral in poor environments. Variety 1 is precisely the type of variety that we expect to appear better with respect to mean performance versus a probabilistic approach.

Second, if the difference in main genetic effect is sufficiently large then the mean-based and probabilistic approach will always reach the same conclusion. For example, if $g_1 = 3$ and $g_2 = 2.4$ but everything else stays the same, that is, the difference in mean increases from 0.5 to 0.6 bushels/acre, it is easy to verify that Variety 1 will be selected 50% of the time based on two random environments. And if $g_2 = 1.9$ then Variety 1 will be selected 70% of the time. Thus, the proposed probabilistic approach is primarily relevant for comparisons where the difference in genotype effects is relatively small; but we argue that those are also the comparisons that are the most important to plant breeders in practice.

The observation made above regarding differences in G×E structure suggests that there is a relationship between the proposed approach and stability measures. This interesting connection will not be explored in detail in this paper, but it should be pointed out that no stability measure could completely replace the probability-based approach. Just like the mean phenotype is a summary statistic of the probability distribution of phenotype across environments, stability measures provide another complementary summary statistic. While considering two or more such summary statistics is certainly preferable to a single statistic, no summary statistics can completely replace considering the whole probability distribution. In fact, in some sense *the probabilistic approach accounts for the entire G×E structure* in whatever is selected as the target environments, rather thus the mean and some measure(s) of stability.

3.4 Probabilistic Pairwise-Comparison Methodology

Quantitative analysis of phenotype data for advancement decisions is heavily based on what may be considered as the analysis of mean performance (Reckling et al., 2021; Schrama et al., 2018). The genotype effect of experimental genotypes is estimated using some model that combines all available input data (e.g., phenotype observations and genetic markers) and when two experimental genotypes are compared according to their mean yield (or other phenotypic response of interest). As argued through the motivational example above, such mean analysis may be misleading, especially as extreme responses of genotypes in different environments pull the mean up or down and do not reflect the absolute superiority of one genotype over the other. This problem is unavoidable and causes uncertainty because of the limited sampling of locations. We therefore propose a new statistic to compare genotypes that, instead of predicting the mean, estimates the probability distribution that one genotype performs better than another. If the distribution of yield (or any other phenotype) is symmetric in the planted locations, then this is not probabilistic and mean comparison would produce the same result. However, if they have asymmetric distributions with respect to locations planted, which we believe to most often be the case in practice, then we argue that a probabilistic comparison is more sensible. This is because

the probabilistic comparison combines genetic gain, that is, the main effects, with stability/adaptability, that is, the interaction effects, as well as the distribution of the environmental effects in the observed environments. This happens without estimating these effects directly but is reflected in the estimated probabilities.

The goal is to understand the performance of n genotypes in m target environments (locations and year). Let y_{ij} denote the phenotype of genotype i in environment j, where $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$. What is traditionally of interest is the mean of each genotype across all environments, denoted $g_i = \frac{1}{m}\sum_{j=1}^{m} y_{ij}$, and may be thought of as representing the genotype effect. As noted above, what is specifically of interest is comparing two genotypes i_1, i_2, which is usually done based on the genotype effect and one approach would be to try to estimate the mean difference $g_{i_1} - g_{i_2}$ with as much precision as possible and use this as the basis of a decision. However, the analysis here is based on the indicator function

$$I(i_1, i_2) = \begin{cases} 1 & g_{i_1} > g_{i_2}, \\ 0 & g_{i_1} \leq g_{i_2}. \end{cases} \tag{3.1}$$

It is impossible to observe every possible environments. A plant breeder observes some small sample $A \subset \{1, 2, \ldots, m\}$ of m_0 environments, where $m_0 << m$, and uses the observed values $\{y_{ij}\}, j \in A; i = 1, \ldots, n$, to obtain estimates $\hat{g}_i = \frac{1}{m_0}\sum_{j \in A} y_{ij}$. From the observed environments A it is straightforward to obtain a point estimate of the indicator of interest

$$\hat{I}_A(i_1, i_2) = I_A(i_1, i_2). \tag{3.2}$$

This will simply equal either zero or one, depending on the relative performance of the two genotypes in this set A of environments. Based on this estimate, a plant breeder might conclude that genotype i_1 is better than genotype i_2 if $\hat{I}(i_1, i_2) = 1$. This is equivalent to directly estimating the mean difference $\hat{g}_{i_1} - \hat{g}_{i_2}$ and making a decision based on this mean. However, as we argued above, focusing on the mean can be misleading if the two genotypes have different G×E structures, in which case it would be more informative to know the distribution of the indicator function in addition to the mean.

To characterize the *distribution* of $I(i_1, i_2)$, we start by generating B bootstrap samples A_1, A_2, ..., A_B by resampling the set A of environments with replacement. This results in B estimates of the indicator

$$\hat{I}_{A_1}(i_1, i_2), \hat{I}_{A_2}(i_1, i_2), ..., \hat{I}_{A_B}(i_1, i_2). \tag{3.3}$$

Thus, instead of a single estimate, we now have B estimates, capturing the uncertainty due to the set of environments that were observed. Using the estimates calculated on the bootstrap samples, the probability of genotype i_1 being better than genotype i_2 in a randomly selected set of environments can now be estimated as

$$\hat{P}\left(g_{i_1} > g_{i_2}\right) = \frac{1}{B} \sum_{b=1}^{B} \hat{I}_{A_b}(i_1, i_2) \tag{3.4}$$

Given these probability estimates, a plant breeder could now conclude that genotype i_1 is better than genotype i_2 if $\hat{P}\left(g_{i_1} > g_{i_2}\right) > \frac{1}{2}$.

In this paper we refer to decisions based on equation (3.2) as mean-based selection and decisions based on (3.4) as probabilistic selection. Thus, decisions are determined by either the following sets of genotype pairs, which completely describe which genotypes should be selected over others:

$$S_{Mean} = \left\{(i_1, i_2) \in \{1, 2, \ldots, n\}^2 : I_A(i_1, i_2) = 1\right\}, \tag{3.5}$$

$$S_{Prob} = \left\{(i_1, i_2) \in \{1, 2, ..., n\}^2 : \hat{P}(g_{i_1} > g_{i_2}) > \frac{1}{2}\right\}. \tag{3.6}$$

As far as we know, this type of probabilistic selection has not been proposed before in the plant breeding domain, whereas incorporating the mean-based selection is standard practice.

An example of the mean-based and probabilistic selection approaches is illustrated in Figure 3.1. In this example the two approaches would reach different conclusions because while one genotype has a better mean across the observed locations, this is not true for the majority of the resampled subsets of locations. This paper explores when these two approaches result in different conclusions, that is, $S_{Mean} \neq S_{Prob}$, and the explanations behind those differences.

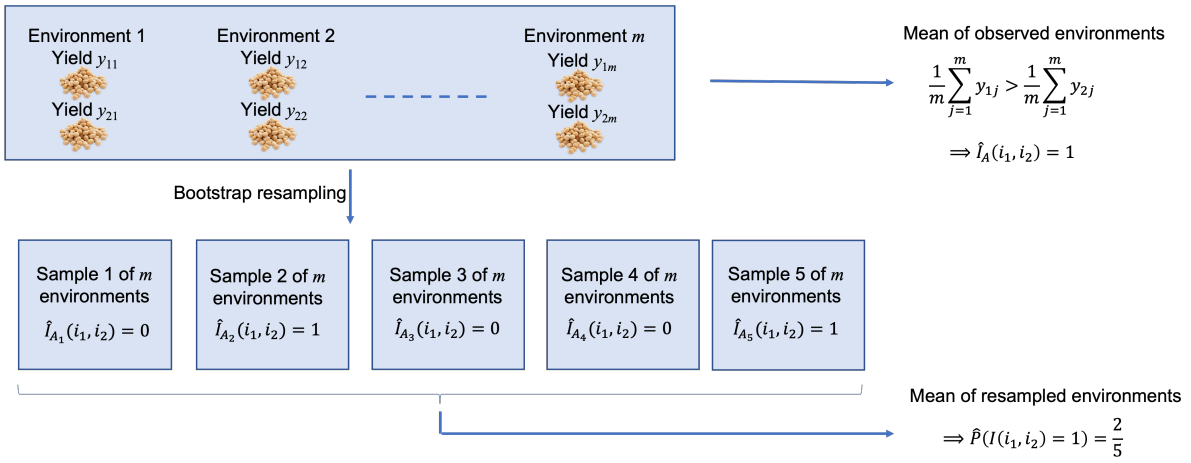

Figure 3.1: Mean versus probabilistic comparison of two genotypes. Based on direct observations the first genotype is better, but as it is only better in two out of five resampled environments the probabilistic comparison favors the second genotype.

3.4.1 Simulated Data

To provide insights into if and when probabilistic pair-wise comparison differs from mean-based comparison, we generate simulated data that can be considered as a generalized version of the motivational example introduced in a previous section. To generate this data, we assume that the phenotype of interest is plant yield, although any other phenotype could be used, and that yield follows what might be considered a standard linear model involving genetic (G_i), environmental (E_j), and genetic-by-environment interaction effects, which we refer to as G×E effects (Becker and Léon, 1988; van Eeuwijk et al., 2016). As the noise does not provide any insight with respect to the purpose in this research, for simplicity and clarity, we ignore the noise from the following equation.

$$\tilde{y}_{ij} \quad = \quad \mu + G_i + E_j + G_i \times E_j + \epsilon_{ij}. \tag{3.7}$$

As will be further described below, the simulation then generates values for each effect accroding to specific distribution $G_i \sim F_G$, $E_j \sim F_E$, and assumes different structures for the $G \times E$ interactions. This experiment includes four factors that will be set as follows:

1. Difference in main effect. Will be determined by the generated G effects for each genotype.

2. Similarity in interactions. Will generate three types: a) same in all environments, b) very good in good environments, neutral in others, c) very good in good environments and very poor in poor environments. The sum of the G×E effects will be kept as constant.

3. Magnitude of interactions. We consider a set the average magnitude of the G×E effects as half or double to the average magnitude of the main G effect.

4. Fraction of environments observed. We consider the scenarios with 5%, and 10% and 100% of the environments observed.

Following the above description, after generating a mean genotype effect, we generate three genotypes with three levels of G×E structures. The first is completely stable. i.e., has it has the same G×E effects over all environments. The second might be considered adaptive, i.e., takes advantage of good environments and performs very well in those environments. Finally, the third one is highly variable, with both very good and very poor performance based on the E effect. All three types have a exact same mean phenotype if we observe all of m environments, while their distributions are different.Therefore, with any subset of the m environments, the mean phenotypic response would also be different. Furthermore, for each set of genotypes, we consider uniform distribution of $[-1.5, 1.5]$ for F_G and uniform distribution of $[-10, 10]$ for F_E. As explained, G×E distributions derive from G and E effects based on their structures. To cover every aspect, we consider different magnitude of G×E with different fractions of environments to explore how different are the comparisons based on different contributing factors.

The synthetic data considered in this section consists 30 different genotypes with 3 different G×E interaction effects on 100 environments. For each three set of genotypes, genetic main effect identical, i.e., we have 10 distinct genetic effects. Probabilistic comparisons give us information on how certain we can be with the comparison we are making. The description of the simulated genotypes is shown in Table 3.2.

Table 3.2: Thirty simulated genotypes. A set of three genotypes has identical main genetic effect (G), but each of those three has a different $G \times E$ structure.

Name	Main Effect (G)	G×E Structure
GS1.44	1.44	Stable
GA1.44	1.44	Adaptive
GV1.44	1.44	Variable
GS1.39	1.39	Stable
GA1.39	1.39	Adaptive
GV1.39	1.39	Variable
GS1.13	1.13	Stable
GA1.13	1.13	Adaptive
GV1.13	1.13	Variable
GS0.81	0.81	Stable
GA0.81	0.81	Adaptive
GV0.81	0.81	Variable
GS0.09	0.09	Stable
GA0.09	0.09	Adaptive
GV0.09	0.09	Variable
GS0.39	-0.39	Stable
GA0.39	-0.39	Adaptive
GV0.39	-0.39	Variable
GS0.52	-0.52	Stable
GA0.52	-0.52	Adaptive
GV0.52	-0.52	Variable
GS0.66	-0.66	Stable
GA0.66	-0.66	Adaptive
GV0.66	-0.66	Variable
GS1.23	-1.23	Stable
GA1.23	-1.23	Adaptive
GV1.23	-1.23	Variable
GS1.29	-1.29	Stable
GA1.29	-1.29	Adaptive
GV1.29	-1.29	Variable

3.4.2 Estimated probabilities for pair-wise comparison

We start by looking at the estimated probabilities of one genotype being superior to another genotype. The results are displayed in Figure 3.2 and demonstrate how the probabilities of one genotype being better than another depend on the genetic main effect differences and the differences in G×E structures. All plots in Figure 3.2 are showing the comparison among the genotype with better main G main effect with respect to the other genotype in the pair. Furthermore, the genotypes are ordered according to their main effects with the one with the best mean yield being furthest to the right. Several observations can be made.

- If the G×E structure does not differ, then the probabilities are all either zero or one, that is there is no ambiguity. The three plots on the top in Figure 3.2 compares pairs with identical G×E structure. In those cases, the probability that the genotype with the higher mean is better is always one, even if the difference in the means is very small.

- One the other hand, when comparing genotypes that differ with respect to their interactions to environments, the likelihood of one genotype being better than the other genotype gets smaller as their genetic effects (yield similarity) get apart from each other. This trend is consistent in all three plots shown on the bottom row of the Figure 3.2 and the bigger the genetic difference, the more certain the comparison gets probabilistically.

- Finally, a more subtle observation is that absolute certainty (probability equal one) is observed when comparing adaptive genotypes to stable and variable genotypes, while it is not observed in comparison of two other structures. The reason may be that adaptive genotypes never exhibit very bad performance as they take advantage of good environments; therefore, when the difference is large enough, the comparison becomes certain at a fast rate.

As noted above, values over 0.5 indicate the win probability of genotype with higher genetic main effect over the other genotype and vice versa. It can be further observed from figure 3.2 that stable genotype is always selected over the adaptive and variable genotypes when the G main effect is bigger, even if they are very close; that is, for the plots in the bottom-right and

bottom-left, the probabilities are always greater than 0.5 for all pair-wise comparisons. On the other hand, the comparison among highly variable genotypes and adaptive ones is more complicated. Even though adaptive genotypes take advantage of good environments, when comparing with highly variable genotypes with a same mean, it depends on bad environments as well. Accordingly, for genotypes with close G main effect, we might select the genotype with lower G main effect. This statement is true for the case of variable genotype comparison with stable genotypes as well. This illustrates the point that when comparing genotypes, it is very important to consider the distribution of environments, because, for detecting the adaptive genotype, there should be high enough proportion of good environments so that it reflects in them taking advantage of good environments. Otherwise, since the highly variable G×E structure goes to both good and bad extreme directions, it is possible that under certain situations, this structure might be preferable and should be called over the other structures. This shows the complexity of the the process and as it can be seen, the probabilistic comparison incorporates the distribution of environments and takes into account some of the underlying components and constraints when making the comparison.

3.4.3 Comparison of probabilistic selection and mean selection

In the previous section, we explored the probabilistic comparisons and noted that the genotype with the better mean is not always the genotype that is more likely to perform better, given the different G×E structure of the pair of genotypes, as well as the distribution of the environmental effects in the target environments. In this section, we further explore whether decisions made using the new probabilistic approach characterized by Equation (3.4) are different than decisions made using the traditional mean approach characterized by Equation (3.2), that is, is $S_{Prob} = S_{Mean}$? And if yes, what circumstances lead to such differences? Thus, we look at the fraction of pair-wise comparisons that are different as a function of both the mean yield difference and difference in $G \times E$ structure.

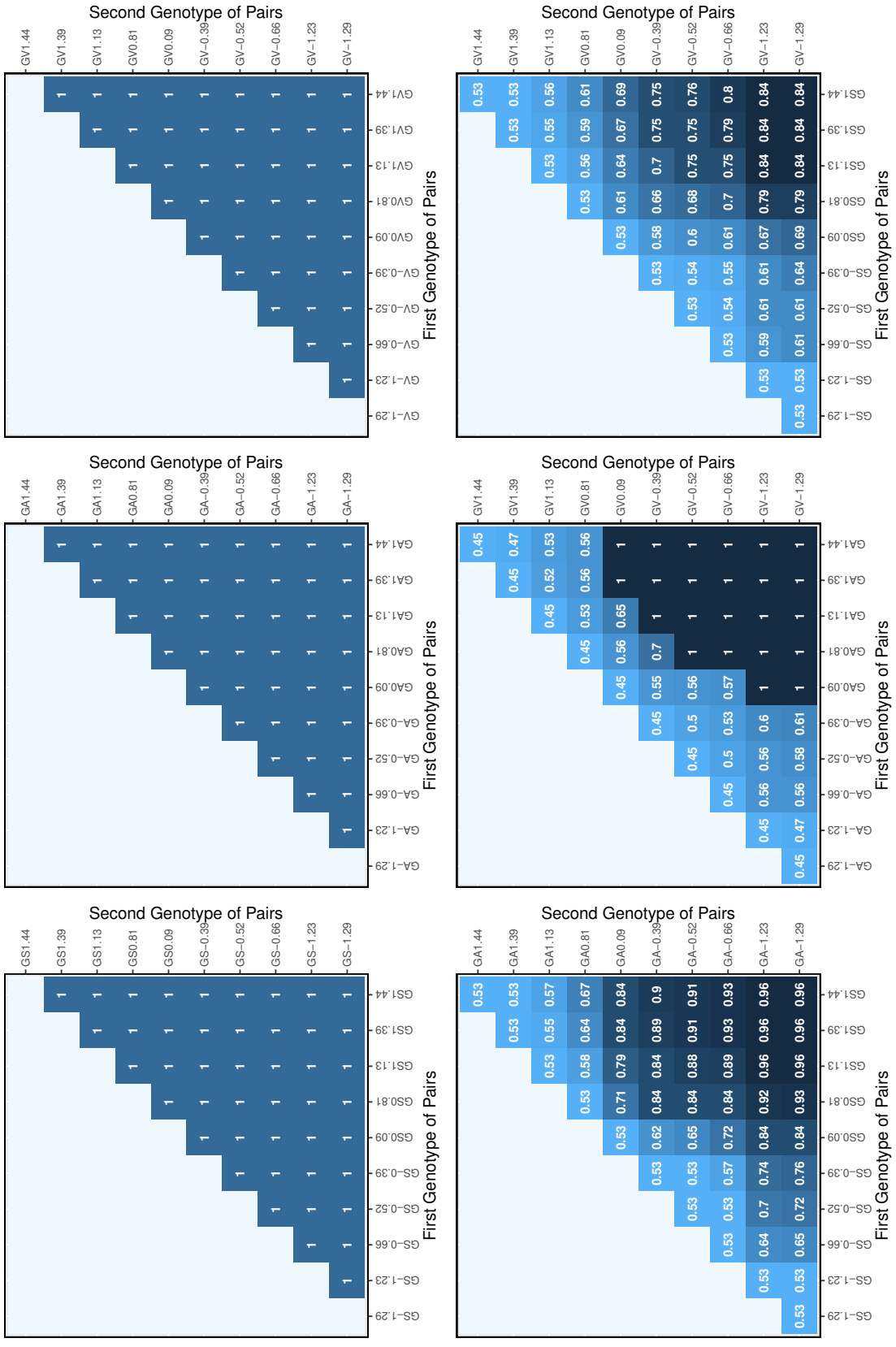

Figure 3.2: Heat maps showing probabilistic comparison of genotypes pairs where the G×E structure is stable vs stable (top-left), stable vs adaptive (bottom-left), adaptive vs adaptive (top-middle), adaptive vs variable (bottom-middle), variable vs variable (top-right), and stable vs variable (bottom-right). Genotypes are ordered according to main effects.

For this purpose, we calculate the probabilistic and mean comparisons for the synthetic data defined in Table 3.2. For further insights, we calculate the results of probabilistic comparisons when the average G×E magnitude is either reduced by half or doubled. Figure 3.3 illustrates the cases where the results of two methods differ or are the same with respect to their difference in yield for different G×E structures and magnitudes.

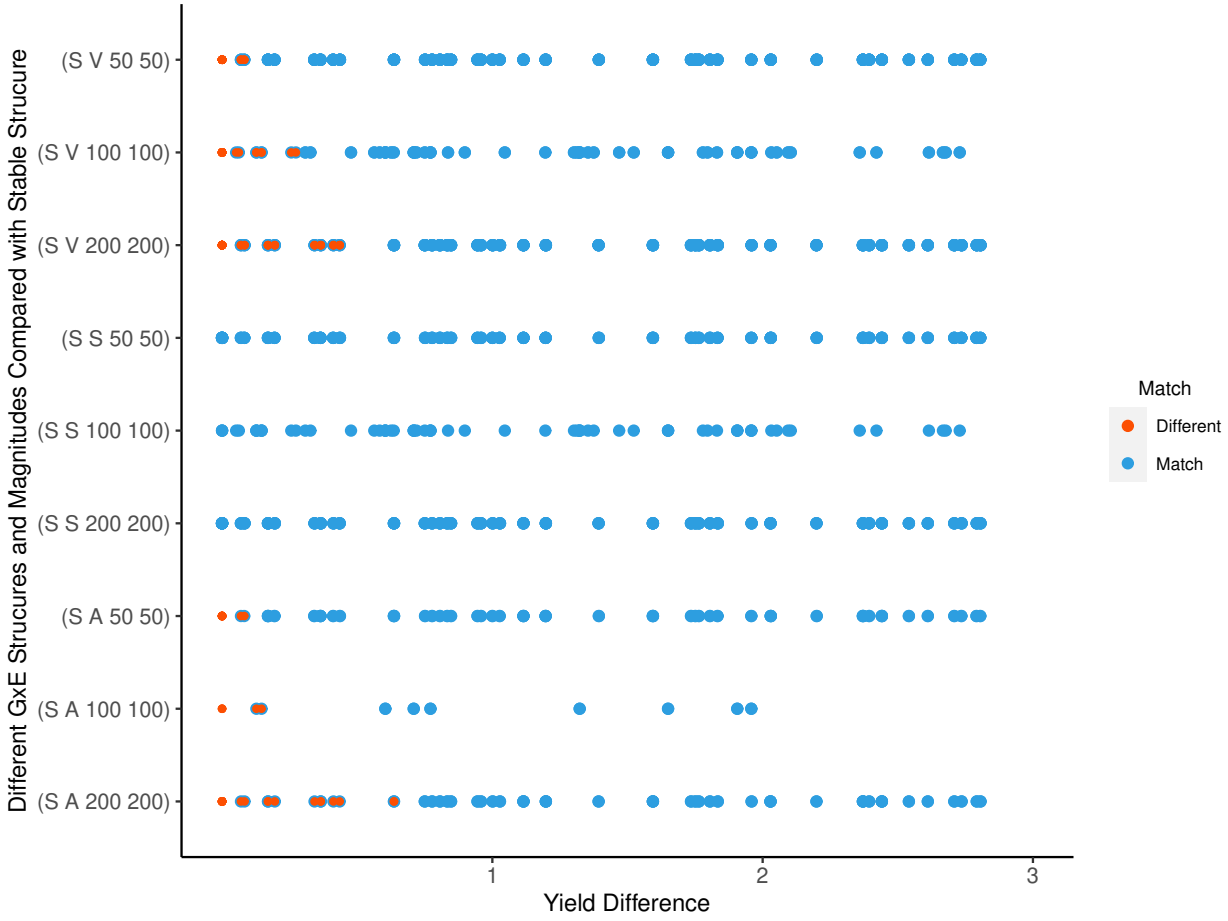

Figure 3.3: Mean and probabilistic comparison match/mismatch for stable genotypes and three levels of the magnitude of G×E interactions (50, 100 and 200). The red dots indicate pairs where the genotype with the better mean is *not* the genotype that is *more likely* to perform better.

As it can be seen, when comparing two perfectly stable genotypes, the comparison between two stable genotypes would always be the same no matter how big the G×E magnitude is. The comparison gets challenging when the structures are different and that makes the decision on the

desirability of genotypes complicated. When the interaction structures differ, the two methods give different results when the genotypes are close. Additionally, the mismatch cases increase when the G×E magnitude increases. This is consistent for both adaptive and variant genotypes comparisons to stable genotypes.

Similar plots are shown in Figure 3.4 for the adaptive and highly variable genotypes. The results of same structures' comparisons are consistent with what have been detected in Figure 3.3, if the G×E structure is the same then the genotype with the better mean is always the genotype that is more likely to perform better. It also indicates, as expected that if the magnitude of the G×E effect is larger, then it is more frequent that the genotype with the better mean is not the genotype that is more likely to perform better.

The results above show that the frequency of when the genotype with the better mean is not the genotype that is more likely to perform better depends on the differences of G×E structures and the magnitude of the G×E effects, as well as the difference in the main genotype effects. However, this frequency may also depend on the set of environments, and in particular the fraction of environments where the genotypes have been planted. Figure 3.5 shows the results for all comparisons for the case when 10% of the locations are observed.

For additional insights into when the genotype with the better mean is not the genotype that is more likely to perform better, a summary mean and probabilistic pair-wise comparison between different genotypes with different G×E magnitudes has been presented in Table 3.3.

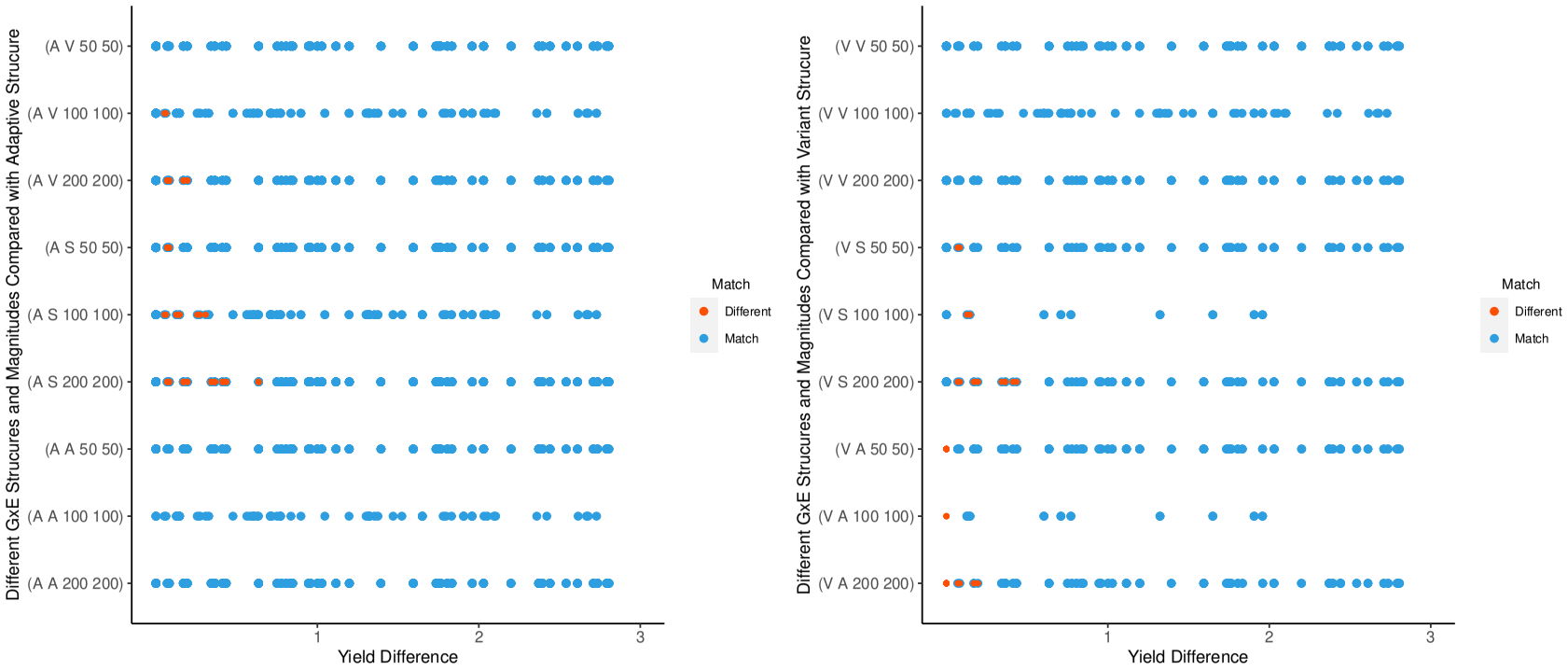

Figure 3.4: Mean and probabilistic comparison match/mismatch for adaptive and highly variable genotypes, and three levels of the magnitude of G×E interactions (50, 100 and 200). The red dots indicate pairs where the genotype with the better mean is *not* the genotype that is *more likely* to perform better.

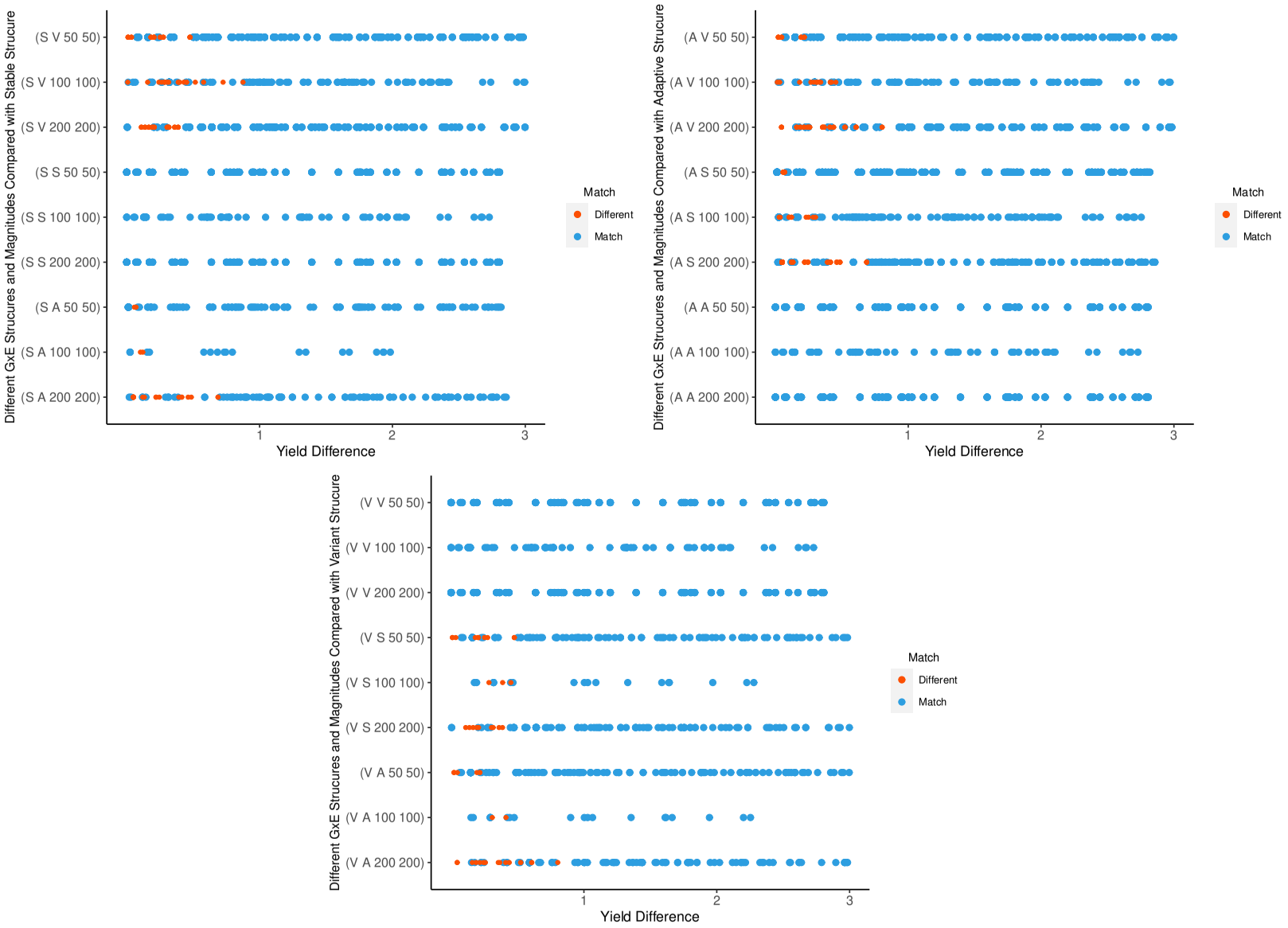

Figure 3.5: Mean and probabilistic comparison match/mismatch of different genotypes pairs when observing 10% of the locations.

It reports the fraction of time that the two definitions of what constitutes the best genotype disagree for all of the scenarios considered above as well as three cases for the fraction of locations observed: 5%, 10%, and 100%. Rather than a continuous scale of mean yield difference we have used the previously reported results to create three buckets as follows. The first bucket has a yield difference of less than 0.5, the second between 0.5 and 1, and finally above one. These buckets can be thought of as genotypes with yields that are very similar, somewhat similar, and significantly different, respectively. In two scenarios the genotype with the better mean is always the genotype that is more likely to perform better, namely when the yield difference is larger than one and if the G× E structure is identical for the two genotypes. This further supports the observations made above. In other cases, the fraction of pairs where the one with the higher mean is not more likely to be better can be very high. For example, when comparing stable and adaptive genotypes with a large magnitude of G× E effects and similar mean yield, the fraction is 43.8%, 68.8% and 68.8% when observing 5%, 10% and 100% of the environments, respectively. In general, as expected, the fraction is higher if the yield difference is smaller and the magnitude of the G× E effects is larger, but it is worth noting that the pattern is more complex when comparing the highly variable genotypes. As noted above, this is explained by the fact that the probabilistic approach accounts for the distribution of environmental effects for the target environments, as well as the main G effects and the G× E interaction effects.

Table 3.3: Fraction of Pairs where Genotype with Better Mean is Less Likely to Perform Better

Pairs Compared	Fraction of locations observed								
	5%			10%			100%		
	Yield Difference								
	[0,0.5)	[0.5,1)	[1,∞)	[0,0.5)	[0.5,1)	[1,∞)	[0,0.5)	[0.5,1)	[1,∞)
(S S 50 50)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(S S 100 100)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(S S 200 200)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(S A 50 50)	10.3	0.0	0.0	6.3	0.0	0.0	43.8	0.0	0.0
(S A 100 100)	0.0	0.0	0.0	28.6	0.0	0.0	71.4	0.0	0.0
(S A 200 200)	43.8	0.0	0.0	68.8	10.5	0.0	68.8	11.1	0.0
(S V 50 50)	11.5	0.0	0.0	31.6	0.0	0.0	45.9	0.0	0.0
(S V 100 100)	11.8	0.0	0.0	31.8	23.5	0.0	62.2	0.0	0.0
(S V 200 200)	0.0	0.0	0.0	49.1	0.0	0.0	69.4	0.0	0.0
(A A 50 50)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(A A 100 100)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(A A 200 200)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(A V 50 50)	0.0	0.0	0.0	17.7	0.0	0.0	0.0	0.0	0.0
(A V 100 100)	8.7	0.0	0.0	29.5	0.0	0.0	5.4	0.0	0.0
(A V 200 200)	38.9	15.6	0.0	60.8	15.7	0.0	15.3	0.0	0.0
(V V 50 50)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(V V 100 100)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
(V V 200 200)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

The analysis shown in this section demonstrates that comparing genotypes can get challenging under certain circumstances. When the difference in main effects is large or the two genotypes have the same G×E structure then the comparison is straightforward as the genotype with the better mean is also the one that is more likely to perform better across the target environments. However, in cases where the means are closer and the G×E structure differs, this may not be the case and indeed in such cases the simulated data experiments show that in some cases the majority of pair-wise comparisons differ in these two criteria. These differences also depend on the magnitude of the G×E effects and the specific subsets of environments that are observed. The proposed probabilistic comparison incorporates all of those factors in order to identify the genotype that is more likely to perform better across the environments.

3.5 Conclusions

Experimental genotypes are frequently compared according to mean phenotypic response, and such comparison is then used as the basis of further decision-making, for example, to determine

which experimental soybean varieties or corn hybrids should be advanced within a breeding program. This paper introduces a new way to compare genotypes, namely to estimate and select genotypes based on which genotype is most likely to perform better across a set of environments. We further evaluate how often this differs from a simple means-based selection, and evaluate the underlying reasons for why the genotype with the better mean is not always most likely to perform best. The probabilistic approach accounts for both mean and G×E interactions and thus incorporates both main effects and in some sense the stability and adaptability of the genotypes. However, the probability estimates account for not only the uncertainty that stems from the selection of planting locations but also the distribution of locations in which they have been planted.

Results on simulated data demonstrate that when the difference in main effects is large or the two genotypes have the same G×E structure then the genotype with the better mean is also the one that is more likely to perform better across the target environments. However, these cases are likely to be considered straightforward in practice as any selection approach is likely to result in the same decision. It is in the more difficult cases where the simulation results show that the probabilistic and mean-based approaches differ. Specifically, when the means are close, the G×E structure differs, and the magnitude of the G×E effects is large relative to the main effects, then the majority of pair-wise comparisons may differ. By incorporating the genotype's main effects, the G×E effects, and the distribution of the environmental effects for the target environments into a single probability, the proposed approach provides a new way to identify the genotype that is more likely to perform better across the environments.

The next step in this research will be to incorporate the proposed approach into a decision support tool that can provide a probabilistic ranking of genotypes, and compare this novel approach to ranking with existing approaches. Further research will explore how the proposed approach relates to existing stability measures as well as existing methods that combine mean and stability into a single metric.

3.6 References

Abberton, M. (2012). Molecular plant breeding. by y. xu. wallingford, uk: Cabi (2012), pp. 752, £59.95. isbn 9781845939823. *Experimental Agriculture*, 48(3):461–461.

Becker, H. C. and Léon, J. (1988). Stability Analysis in Plant Breeding. *Plant Breeding*, 101(1):1–23.

Cooper, M. and DeLacy, I. H. (1994). Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theoretical and Applied Genetics*, 88:561–572.

Luckett, D. and Halloran, G. . (2017). CHAPTER 4 PLANT BREEDING. In Pratley, J., editor, *Principles of Field Crop Production*. Graham Centre for Agricultural Innovation, Charles Sturt University: Wagga Wagga Australia.

Reckling, M., Ahrends, H., Chen, T.-W., Eugster, W., Hadasch, S., Knapp, S., Laidig, F., Linstädter, A., Macholdt, J., Piepho, H.-P., Schiffers, K., and Döring, T. F. (2021). Methods of yield stability analysis in long-term field experiments. a review. *Agronomy for Sustainable Development*, 41.

Schrama, M., de Haan, J., Kroonen, M., Verstegen, H., and Van der Putten, W. (2018). Crop yield gap and stability in organic and conventional farming systems. *Agriculture, Ecosystems & Environment*, 256:123–130.

van Eeuwijk, F. A., Bustos-Korts, D. V., and Malosetti, M. (2016). What should students in plant breeding know about the statistical aspects of genotype × environment interactions? *Crop Science*, 56:2119–2140.

Vargas, M., Combs, E., Alvarado, G., Atlin, G., Mathews, K., and Crossa, J. (2013). Meta: A suite of sas programs to analyze multienvironment breeding trials. *Agronomy Journal*, 105:11–19.

CHAPTER 4. METANALYZER: AN R PACKAGE FOR PROBABILISTIC RANKING AND RANK CONFIDENCE INTERVALS

Authors: Reyhaneh Bijari [1], Sigurdur Olafsson [1]

[1] Department of Industrial and Manufacturing Systems Engineering, Iowa State University

4.1 Introduction

Analyzing experimental genotypes is one of the meticulous steps in plant breeding programs. Breeders try to identify the best genotypes with respect to one or some of the phenotypic traits in multi-environment trials (METs), where they are planted in a subset of the target population of environments (Abberton, 2012; Vargas et al., 2013). Due to complicated genotype-by-environment interactions (G×E) effects, and the uncertainty that stems from observing genotypes' performance in a very limited number of environments in plant breeding programs, there has been extensive attention to study and capture their contribution to phenotypic traits, e.g., yield (Cooper and DeLacy, 1994; Pour-Aboughadareh et al., 2022). With respect to analytical tools, over decades, there have been many tools being developed that employ traditional measures from Shukla (Shukla, 1972) to regression-based methods (Eberhart and Russell, 1966) and up-to-date ones such as AMMI models (Sa'diyah and Hadi, 2016) to detect such interactions and have an inclusive analysis of the multi-environment trials. One of the recent tools is the `metan` package in `R` (Olivoto and L'ucio, 2020) which has collected a set of functions to analyze MET datasets, whether with a comprehensive list of stability measures to different models for performance prediction of genotypes. Another tool for MET analysis is `META-R` that provides performance predictions and data features' correlation for the set of locations in which genotypes are planted. `Spatial META-R` is a similar application that uses `R` and `ASReml v1.0` and provides statistical analyses to account of G×E interaction effects and estimates the genotypes'

performance. Many useful packages are also available for this purpose like `baystability` which gives Bayesian stability analysis of G×E interactions. Another useful packages are `GEInfo`, `GEint`, `stability`, `statgenGxE`, and `MPGE`, to name a few.

In this chapter, we develop a package called METanalyzeR that provides a framework for implementing the proposed methodologies in the previous chapters. Furthermore, as data in real applications is not usually like the standard datasets and needs to be processed before any analysis, we define a set of proposed functions that we have found practical in real applications.

We also propose a new ranking approach for comparing genotypes in MET datasets, which have been derived from the concepts of previous chapters. The new method differs significantly from what is assumed in the existing literature as it considers the probabilistic comparison of genotypes when ranking them and tries to give complementary information to the decision-maker.

In the following sections of this chapter, first, we explain two new methodologies that this dissertation offers to the field of plant breeding for the evaluation of corresponding experimental genotypes. The new proposed ranking approach is described in detail. Following that, we briefly go through the rank confidence intervals concept proposed in Chapter 2. In the third section, we go through the METanalyzeR package and explain its functions and implementations. This tool tries to provide users the two new proposed methods for MET datasets' analysis. We will continue analyzing different ranks, followed by some visualization. Finally, we will use the package for some standard datasets and report the outputs for the use cases.

4.2 Methods

Ranking is an essential component of numerous decision-making processes where the selection of the best is of interest and there is therefore a great deal of related literature in different application domains. Statistical methods for subset selection and ranking-and-selection have been received extensive attention for decades (Gibbons et al., 1979; Seong-Hee Kim and Nelson, 2007), with many of those focus on determining how many samples are needed to make a selection (Dudewicz, 1980). In the type of plant breeding applications we address, however, the number of

samples is fixed, and we need to address the uncertainty of the ranking considering this limitation. Govindarajulu and Harvey (1974) presented the ranking and subset selection problem from a Bayesian view. They suggested that ranking populations just by the posterior probabilities might not be an ideal approach unless one understands the selection process's purpose. In another study, Laird and Louis (1989) developed ranking methods based on the conditional distribution of ranks instead of the conditional distribution of parameters by treating the ranks of the prior parameters as the parameters of interest. Previous studies have primarily analyzed ranks considering observations independent (DiCiccio and Efron, 1996; Efron, 1987) and therefore, they are not directly applicable to the case of plant breeding, where there exists significant correlation among the observations.

4.2.1 Probabilistic Ranking of Experimental Crops

In this study, we define a new ranking structure for experimental plant genotypes that differs significantly from what is assumed in the existing literature and use bootstrapping to comparing genotypes probabilistically and deriving probabilistic ranking. We show how the proposed approach works well for real plant breeding data and thus provides a useful tool for plant breeders. In the following part, we explain the new ranking method that is based on the probabilistic comparison defined in chapter 3 which accounts for the uncertainty caused by observing limited environments in real world plant breeding experiments.

4.2.1.1 Algorithm

The procedure is about the act of swapping. This algorithm starts with a table with some ranking as the starting point. Starting from the top genotype, it is compared with the next genotype in the table. It then swaps if the second genotype is better based on the probabilistic approach. If the second genotype is not better, it moves down and compares the second genotype with the third one. The question here is: "Is the third genotype better than the second one?" If yes, another swap occurs; otherwise, it moves down. The algorithm continues down and then

starts from the top again. The first time it makes down without swapping, the algorithm will stop, and it has converged.

For completeness, a description of the new probabilistic ranking is provided in algorithm 1.

Algorithm 1: Probabilistic Rank

Data: A table with some ranking as the starting point
Result: Sorted genotypes' table based on probabilistic rank
$swaps$: swapping counter;
r : rank of genotype;
n: number of genotypes;
g_r: genotype with rank r;
$swaps = 1$;
while $swap \neq 0$ **do**
    $swap = 0$;
    $r = 1$;
    while $r \leq n - 1$ **do**
        Compare the pairwise comparison probability between g_r, and g_{r+1};
        if $P(g_r < g_{r+1})$ **then**
            rank $g_r \leftarrow r$;
            rank $g_{r+1} \leftarrow r + 1$
        else
            rank $g_r \leftarrow r + 1$;
            rank $g_{r+1} \leftarrow r$;
            $swap \leftarrow swap + 1$
        end
        $r \leftarrow r + 1$
    end
end

4.2.2 Rank Confidence Intervals

Current advancement starts with a ranked list of genotypes, usually ranked according to corrected mean yield (yield minus environmental average) and do not account for the variability from a relatively small subset of planting locations. This inherent uncertainty makes the decision-making challenging because it dramatically impacts the relative performance of genotypes and might not reflect their true ranks, mainly because of significant and differing genotype-by-environment interactions (G×E) that genotypes might have across the set of

observed environments. Rank Confidence Intervals try to illustrate this uncertainty by estimating the distribution of rank and helping decision-makers (e.g., breeders). It shows the separation of genotypes and also the relative G×E interaction effects through the widths of CIs (the tighter, the more stable). This can be insightful information, especially when the genotypes' main effects are similar.

4.3 METanalyzeR

In this section, we describe the R package we have developed to implement the new multi-environment trials analysis methods from rank confidence intervals proposed in Chapter 2 to new ranking method from Chapter 3 as new decision support tools in plant breeding. The package also collects some of the existing popular methods for METs analysis. METanalyzeR has used some popular functions (Olivoto and L'ucio, 2020) for the sake of comprehensiveness and to illustrate comparison among different methods for METs analysis. We compare the new probabilistic ranking with three popular ranking approaches that have been used with the same aim in plant breeding practice for decades. It also includes functions we have found useful for data pre-processing in real-world MET data if the information is provided in the dataset. To illustrate the main features of the package, three example datasets (oat data (Olivoto and L'ucio, 2020), rapeseed data (Wright, 2021), maize data (Malosetti et al., 2013)) are embedded to METanalyzeR for further exploration.

4.3.1 Package Overview

In this section, we explain the functions created in the package. For this purpose, we start with a flowchart of the main functions in the package in Figure 4.1.

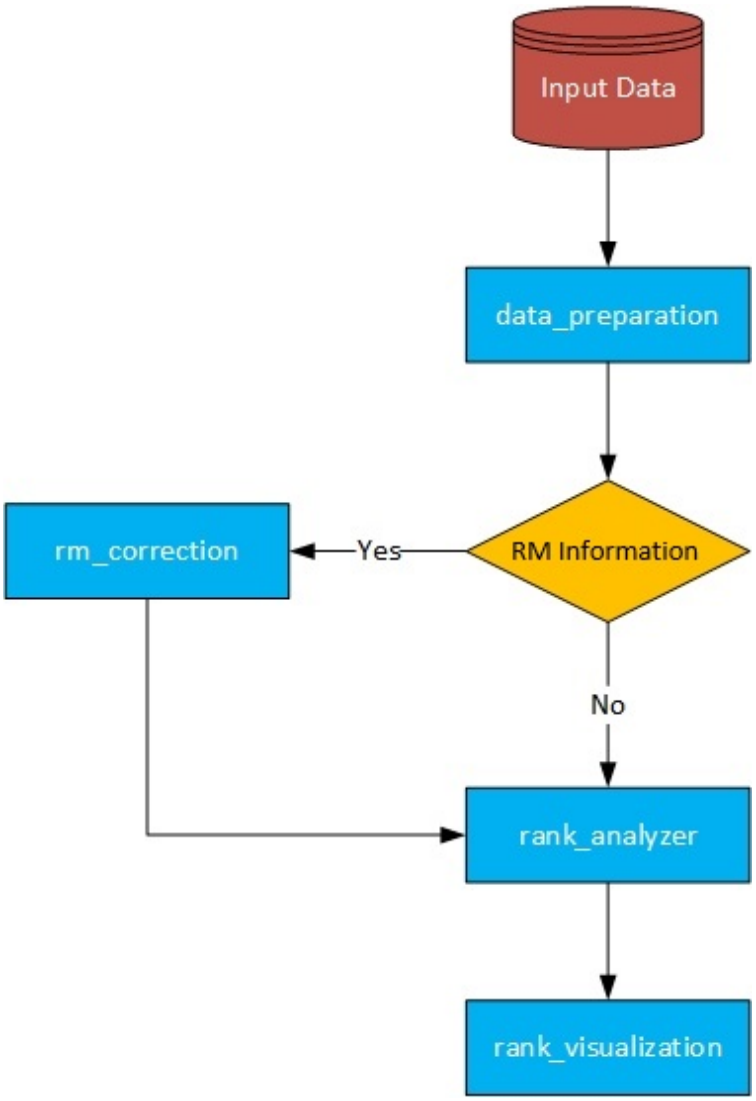

Figure 4.1: Flowchart of the main functions in METanalyzeR package.

As shown in the flowchart, there are four functions in the high level of this package:

- `data_preparation()`
- `rm_correction()`
- `rank_analyzer()`
- `rank_visualization()`

4.3.2 Analysis of Rank and Rank Uncertainty Using METanalyzeR

In this section, we show how different functions are connected and how we can use the provided workflow to explore plant breeding multi-environments trials.

With this regard, the input of the package would be datasets consist of genotypes information with at least their phenotypic responses in the locations and years (environments) they were observed in. We assume that the data we are analyzing is a complete dataset of genotypes in the intended experiment, i.e., all genotypes have been observed in all environments of the experiment.

To start the analysis, the input dataset should have the following necessary columns:

- **GENOTYPE:** A character list of genotypes' names to be analyzed in the experiment.
- **LOCATION:** A character list of locations that genotypes were planted in.
- **YEAR:** A numeric list of the years the experiment was conducted in.
- **REPNO:** A numeric list of replications that each genotype has been observed in a specific location-year.
- **PT:** A continuous list of phenotypic response observed in each set of specific location-year.

It worth mentioning that the user can use direct observations or estimated phenotypic responses, e.g., BLUP or AMMI estimations of yield as an input of the workflow. As explained in Chapter 2, these methods are popular for estimating genotypes' mean phenotype in a multi-environment trials (Sa'diyah and Hadi, 2016) and can be incorporated in any of the comparison/selection methods.

The input data first goes to `data_preparation` function. Since the environment is representative of location-year, which is the feature that genotypes are compared with each other in, the function creates this feature if it is not available in the original dataset. It returns the corrected phenotype (phenotype minus environmental mean) named `Corrected_PT` along with `M_Corrected_PT` as its average:

Since environment is representative of location-year, we keep both `ENVIRONMENT` and `LOCATION` for further consideration on `LOCATION` effects in future studies.

- **Inputs:** A dataframe including at least GENOTYPE, LOCATION, YEAR, REPNO, PT.
- **Ouputs:** A dataframe including GENOTYPE, LOCATION, YEAR, ENVIRONMENT, REPNO, PT, Corrected_PT, M_Corrected_PT, ACTUAL_PT.

After this step, depending on the need for removing the locations' effect with respect to relative maturity bands, i.e., their physiological maturity information, `RM_correction` function can be used. The term *Relative Maturity* is especiallay used for soybean and corn hybrids. It serves as a criteria to compare genotypes with similar maturity bands. This set of data pre-prosessing can be used as of the user's preference. This function takes and gives the following arguments:

- **Inputs:** A dataframe including GENOTYPE, LOCATION, YEAR, ENVIRONMENT, REPNO, PT, Corrected_PT, M_Corrected_PT.
- **Ouputs:** A dataframe including GENOTYPE, LOCATION, YEAR, ENVIRONMENT, REPNO, PT, Corrected_PT, M_Corrected_PT, CORRECTED_PT_RM.

Following the data pre-processing functions, as shown in the flowchart, the next function that is implemented, is `rank_analyzer`. This function is one of the most important functions in this package and is responsible for the analysis derived from bootstrapping approach.

The pre-processed output as `experiment_data` input argument along with other input data such as`genotypes_set`, `boots_matrix`, and `method` will be fed into `rank_analyzer` This function has the following input arguments:

- **experiment_data**: A dataset (e.g., output dataset of data pre-processing) including all the experiment (trial) information.
- **output_name**: A character value indicating the name of the file to be saved on the directory /Outputs/. If this directory doesn't exist, it will be generated.
- **genotypes_set**: A vector of genotypes' subset to be analyzed. The default genotype set includes all genotypes in the experiment.
- **boots_matrix**: A matrix of bootstrap samples of the locations. By default, the value for this argument is `NULL`, which means a matrix of bootstrap resamples should be generated with size 1000. Otherwise, the user should provide the matrix.
- **method**: The method to do the analysis based on it. It can be whether `mean_phenotype`, `probabilistic` or `both`. The default method is `both`.

It worth mentioning that `boots_matrix` is set as an input for the sake of traceability. However, to make the package applicable for all potential users, such as breeders who might not be interested in feeding a matrix of locations and seek only the output of the function, the function sets a set of bootstrap samples if the user would not provide it. In this regard, the output of this function is automatically saved for further analyses.

As mentioned above, `rank_analyzer` is one of the package's important functions, which utilizes different helper functions to generate the outputs of interest. The helper functions are:

- `RankPhenotype_extractor()`
- `gen_comparison()`
- `genotype_list_probs()`
- `CI_calculator()`
- `pairwise_probs()`

- `Probabilistic_ranks()`

Figure 4.2 shows the flowchart of the helper functions within `rank_analyzer` function.

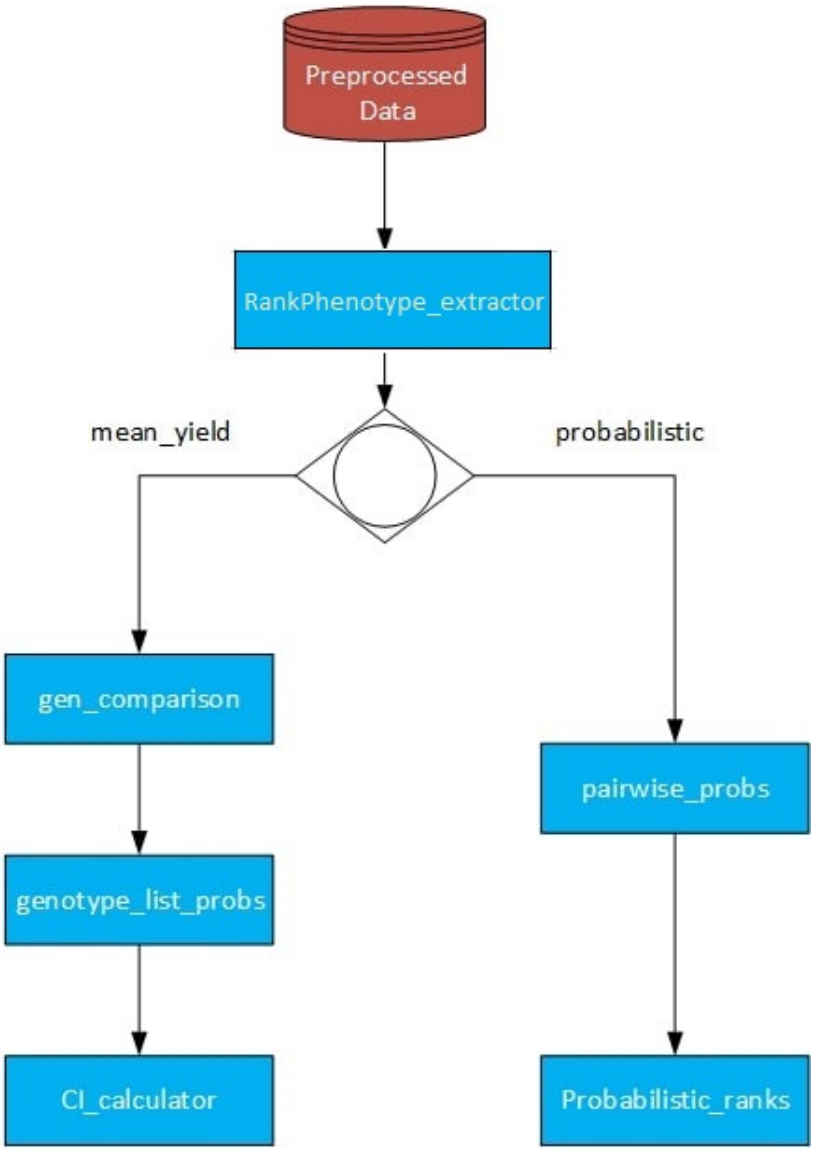

Figure 4.2: Flowchart of the helper functions within `rank_analyzer` function.

In the first step, the pre-processed data will be fed into the `RankPhenotype_extractor`. This function creates rank table and phenotypic response table for the experiment data. In the next step, if the defined method is `mean_phenotype`, `gen_comparison` function will be executed. Rank table information is the main input of this function.It creates the the frequency distribution of locations for each potential rank for all genotypes.

The path leads toward `genotype_list_probs` and the distribution of the rank probabilities for each genotype will be produced. This output along with the `p` (indicating the percentage corresponding to the confidence interval) will then feed into the `CI_calculator` which computes the CI for each genotype.

If the defined method is `probabilistic`, the path goes to `pairwise_probs` function. This function creates pairwise probability comparisons. This output is then used for in `Probabilistic_ranks` to generate probabilistic ranks. Here is the `rank_analyzer` function with its default arguments' values:

```
rank_analyzer = function(experiment_data,
                         output_name  = 'output',
                         genotype_set = unique(experiment_data$GENOTYEP),
                         boots_matrix = NULL,
                         method = 'both')
```

As stated before, this function uses bootstrapping for generating its following objects, which can be generated based on the method defined in the function input argument. The function provides the following outputs:

- **experiment_data**: The dataset analyzed.
- **CI_data**: A dataframe of all ranks probabilities for each genotype.
- **CI_Info**: A dataframe of genotypes with their most probable rank.
- **boots_ranks**: Frequency matrix of genotypes' ranks in all locations throughout all bootstrap resamples.

- **boots_PTs**: A dataframe of genotypes' mean phenotype in all bootstrap resamples.
- **pairwise_probs**: A dataframe of pairwise comparison probabilities.
- **probabilistic_ranks**: A dataframe including probabilistic ranks of genotypes along with their mean phenotype ranks.

With regard to the first output, we keep the input dataset which has been analyzed through `rank_visualization` for the sake of traceability and further exploration.

4.3.3 Visualization of Rank Uncertainty

Once all the outputs are generated, the users can visualize the information through the heatmap of the genotypes' ranks (whether traditional observed mean phenotype or probabilistic) confidence intervals. If the generated rank is `mean_phenotype`, rank CIs generated with this function can be an insightful tool for breeders with respect to genotypes' ranks because it tries to show the uncertainty roots in absolute ranks through their confidence intervals. The other method would be `probabilistic` which delivers the probabilistic ranks heatmap. As explained in section 4.2.1, we propose a new ranking approach which takes advantage of probabilistic comparison of genotypes and ranks them accordingly (See algorithm 1). The function sorts the genotypes based on their most probable ranks for visualization.

- **Inputs:**
  - **input_name**: A character value indicating the name of the saved file from `rank_analyzer` function which is stored.
  - **p**: A float in $[0, 1]$ used for constructing the confidence interval. The default value is set for 0.8.
  - **n_top**: A numeric value specifying the number of n-top genotypes to visualize. By default, it shows the top two genotypes based on the method ranks are calculated.
- **Ouputs:** A ggplot object of the heatmap for the CI of the ranks.

If the `input_name` does not exist, it prints out to the screen that *"The file doesn't exist. First, run the `rank_analyzer` function to generate the input for this function."*

Here is the `rank_visualization` function with its default arguments' values:

```
rank_visualization = function(input_name,
                              p = 0.8,
                              n_top = 2,
                              method = 'mean_phenotype')
```

4.4 Use Cases

This package has been developed for the purpose of multi-environment trails' analysis using new proposed probabilistic approaches. For the rest of this chapter, we bring two datasets of *oat* and *rapeseed* to demonstrate the concepts and applicability of the package for real world datasets. We have used these datasets as a demonstration guide and one can use much bigger datasets for further explorations of the package. It is noteworthy to mention that while we use yield as the phenotypic response in the following sections, the package can handle any continuous phenotypic response.

4.4.1 Oat Dataset

In this part, we will utilize the multi-environment trial of oat yield from `metan` package (Olivoto and L'ucio, 2020). The dataset has 420 observations, 10 genotypes in 14 environments with 3 replications in each environment for all genotypes. It is noteworthy to mention that columns' names are not compatible with the package requirements and one should rename them if using this dataset.

4.4.1.1 Main Functions and Their Outputs

First, as shown in Figure 4.1, the data goes to `data_preparation` function. Below is a summary of the output dataframe columns with some of their data and their structures. As

mentioned before, REPNO is indicating the number of replications each genotype has in each environment and PT is representing the phenotype, i.e., yield. Also, since ENVIRONMENT is not available in the dataset, the data_preparation function creates it by default. Yet, it keeps both ENVIRONMENT and LOCATION for further consideration on LOCATION effects in potential future explorations.

```
    oat_data = data_preparation(data)
    glimpse(oat_data)
```

```
## Rows: 420
## Columns: 9
## Groups: GENOTYPE, ENVIRONMENT [140]
## $ ENVIRONMENT    <chr> "E1 2020", "E1 2020", "E1 2020", "E1 2020", "E1 2020", ~
## $ GENOTYPE       <fct> G1, G1, G1, G2, G2, G2, G3, G3, G3, G4, G4, G4, G5, G5,~
## $ REPNo          <fct> 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1~
## $ PT             <dbl> 2.16700, 2.50304, 2.42732, 3.20750, 2.93290, 2.56484, 2~
## $ YEAR           <dbl> 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2~
## $ LOCATION       <fct> E1, E1, E1, E1, E1, E1, E1, E1, E1, E1, E1, E1, E1, E1,~
## $ P_Env          <dbl> 2.520683, 2.520683, 2.520683, 2.520683, 2.520683, 2.520~
## $ Corrected_PT   <dbl> -0.353683333, -0.017643333, -0.093363333, 0.686816667, ~
## $ M_Corrected_PT <dbl> -0.15489667, -0.15489667, -0.15489667, 0.38106333, 0.38~
```

As we have mentioned in section 4.3.2, rank_analyzer provides a set of outputs based on the provided bootstrap matrix. Here we bring the outputs for 50 iterations to serve its purpose as a demonstration guide. The user would use this function for much larger bootstrap samples for more accurate results. In the following, the summary of three outputs along with some of their data and columns' structures for oat experiment data are shown.

```
glimpse(rank_analysis_info[[2]])
```

This is showing a summary of the second output of the dataframe `rank_analyzer` function consisting of all ranks probabilities for each of 10 genotypes in oat dataset.

```
## Rows: 10
## Columns: 13
## $ GENOTYPE       <fct> G8, G3, G2, G7, G4, G1, G9, G6, G5, G10
## $ `1`            <dbl> 0.67, 0.32, 0.09, 0.01, 0.01, 0.00, 0.00, 0.00, 0.00, 0~
## $ `2`            <dbl> 0.26, 0.60, 0.04, 0.04, 0.01, 0.01, 0.00, 0.00, 0.00, 0~
## $ `3`            <dbl> 0.04, 0.06, 0.45, 0.27, 0.04, 0.04, 0.01, 0.02, 0.02, 0~
## $ `4`            <dbl> 0.01, 0.02, 0.22, 0.39, 0.14, 0.05, 0.06, 0.04, 0.01, 0~
## $ `5`            <dbl> 0.01, 0.00, 0.11, 0.11, 0.39, 0.11, 0.09, 0.06, 0.03, 0~
## $ `6`            <dbl> 0.00, 0.00, 0.05, 0.10, 0.17, 0.21, 0.17, 0.14, 0.08, 0~
## $ `7`            <dbl> 0.00, 0.00, 0.02, 0.03, 0.14, 0.21, 0.19, 0.19, 0.12, 0~
## $ `8`            <dbl> 0.01, 0.00, 0.00, 0.02, 0.08, 0.15, 0.16, 0.27, 0.15, 0~
## $ `9`            <dbl> 0.00, 0.00, 0.00, 0.01, 0.03, 0.14, 0.17, 0.17, 0.32, 0~
## $ `10`           <dbl> 0.00, 0.00, 0.01, 0.03, 0.01, 0.08, 0.15, 0.11, 0.28, 0~
## $ M_Corrected_PT <dbl> 0.33, 0.28, 0.07, 0.07, -0.03, -0.07, -0.16, -0.14, -0.~
## $ Rank           <dbl> 1, 2, 3, 3, 5, 6, 9, 7, 7, 10
```

Another output that `rank_analyzer` gives us is the pairwise comparison probabilities. This dataframe is reported as the sixth output of this function.

```
glimpse(rank_analysis_info[[6]])
```

```
## Rows: 90
## Columns: 3
## Groups: GENOTYPE [10]
## $ GENOTYPE    <chr> "G1", "G1", "G1", "G1", "G1", "G1", "G1", "G1", "G1", "G10~
## $ Compared_to <chr> "G10", "G2", "G3", "G4", "G5", "G6", "G7", "G8", "G9", "G1~
## $ Prob        <dbl> 0.464, 0.329, 0.071, 0.386, 0.550, 0.500, 0.286, 0.000, 0.~
```

If the user considers `method` in its input parameters as `probabilistic` or `both`, one of the main outputs of the `rank_analyzer` function would be the new proposed probabilistic ranks. It is derived from the embedded helper function, `Probabilistic_ranks`.

Following, the example of the generated output for oat dataset is shown:

```
glimpse(rank_analysis_info[[7]])
```

```
## Rows: 10
## Columns: 3
## $ GENOTYPE          <chr> "G1", "G10", "G2", "G3", "G4", "G5", "G6", "G7", "G8~
## $ Observed_Rank     <int> 6, 10, 3, 2, 5, 7, 8, 4, 1, 9
## $ ProbabilisticRank <int> 6, 10, 3, 2, 5, 8, 7, 4, 1, 9
```

4.4.1.2 Visualization Function and its Outputs

Figure 4.3 is showing the CIs of mean yield ranks (input argument `method` is defined as `mean_phenotype`) for 80% confidence level (`p` = 0.8). It is the output of `rank_visualization` when the output object of `rank_analyzer` is fed to the function as `rank.analyzer.output`.

```
Rank_CI = rank_visualization(input_name = rank.analyzer.output,
                             p = 0.8,
                             n_top = 10,
                             method = 'mean_phenotype')
```

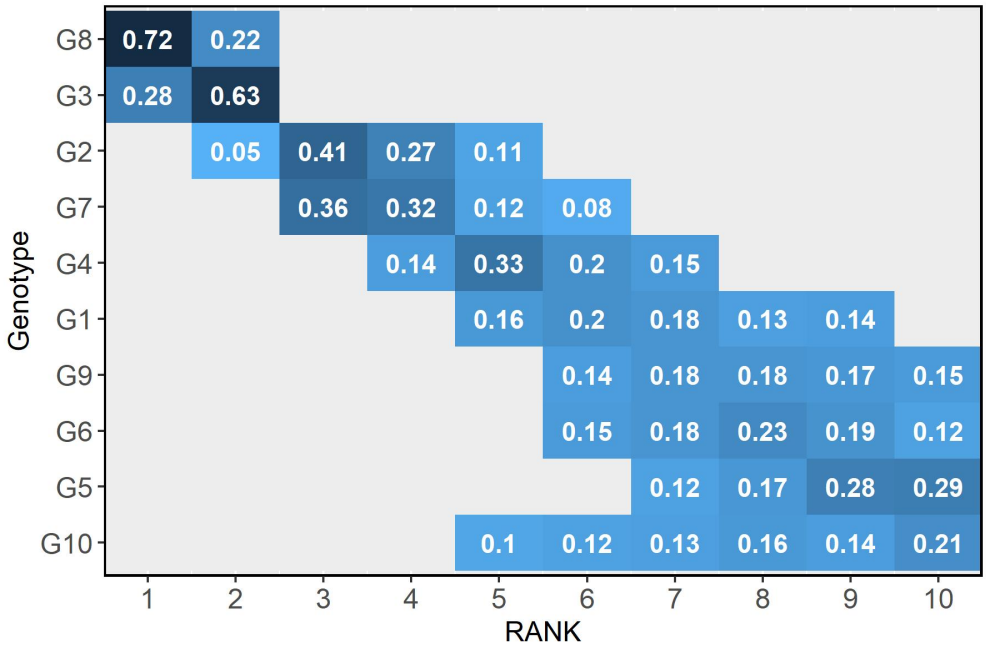

Figure 4.3: Confidence intervals of rank when ranking oat cultivars according to mean yield across the target environments.

To see how reliable the new tool is compared to ranks based on observed yield, the rank confidence interval of probabilistic ranks is also provided as an output for this function in Figure 4.4. It should be noted that to create this plot, the function needs to run `rank_analyzer` repeatedly which is embedded inside the function.

As can be seen, the CIs are tighter for genotypes' probabilistic ranks, and it is suggested that one can be more certain when reporting the probabilistic ranks compared to observed yield ranks. This affirms the gap in the literature and the importance of the new approach we have proposed in our endeavor to fill in the gap in plant breeding MET analysis to address the uncertainties rooted in G×E interactions of genotypes when comparing them.

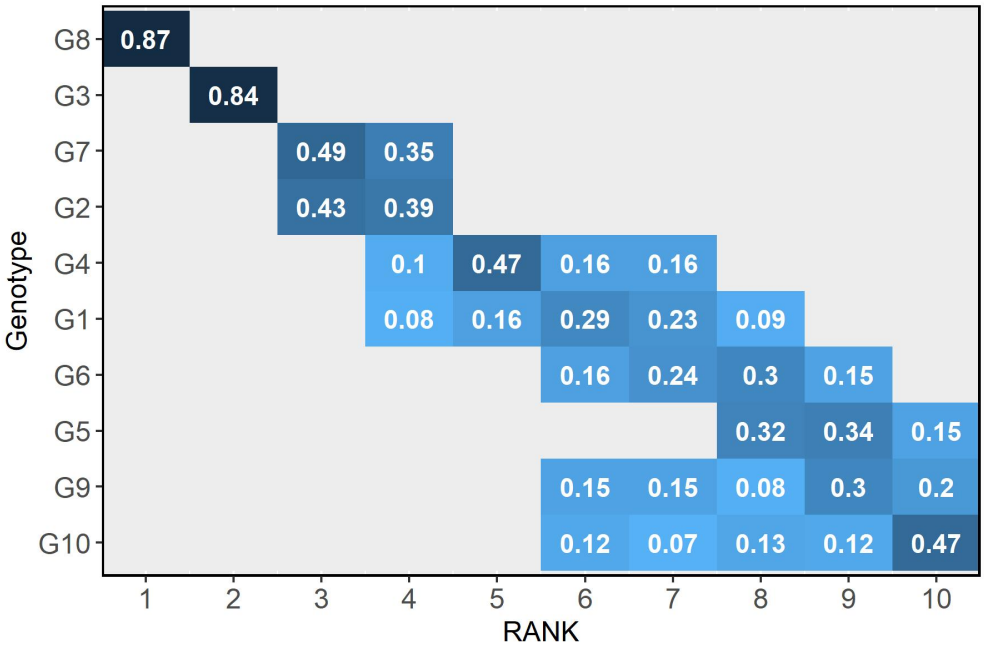

Figure 4.4: Confidence intervals of rank when ranking oat cultivars according to which is the most likely to have higher yield across the target environments.

4.4.1.3 Comparison of probabilistic ranking and other selection methodologies

As mentioned before, at its core, this complex selection process comes down to making pair-wise comparisons between two genotypes. A pair-wise comparison may be converted into a ranked list. Table 4.1 demonstrates the results of ranked list of genotypes according to probabilistic ranking along with ranked lists with mean phenotype ranking and three classic approaches, i.e., Kang's rank-sum method (Kang, 1988), stability variance (Shukla, 1972), and superiority measure (LIN and BINNS, 1988) for comprehensive exploration. The very first method measure calculates the stability variation as an unbiased estimate of G×E. The second measure adds the ranks of mean yield with stability measure of variance defined by Shukla so as to achieve relatively stable genotype with higher yield by way of lower rank-sum. The latest one is a stability measure that finds the average mean square differences of genotypes' response with the best response over the locations. Superiority indices are defined as Pi_a, Pi_f, Pi_u in the table for all, favorable and unfavorable environments respectively.

Table 4.1: Oat Genotypes' ranks based on probabilistic and some traditional ranking methods.

Genotype	Ranking Method						
	Probabilistic Rank	Mean Yield	Shukla's Variance	Kang's Rank-Sum	Superiority Measure		
					All Env.	Favorable	Unfavorable
G8	1	1	6	2(7)	2	2	1
G3	2	2	1	1(3)	1	1	2
G7	3	4	8	6(12)	4	6	3
G2	4	3	7	4(10)	3	3	5
G4	5	5	5	4(10)	6	5	4
G1	6	6	2	3(8)	5	4	6
G6	7	8	3	5(11)	7	7	8
G5	8	7	4	5(11)	8	8	7
G9	9	9	9	7(18)	9	9	9
G10	10	10	10	8(20)	10	10	10

4.4.1.4 Convergence of Bootstrap Sampler

In this section, we evaluate the convergence for the bootstrap estimates of the desired probabilities with respect to the bootstrapping sample size. Specifically, we try to show that the probability estimates converge as the number of samples increases. The convergence plot shown in Figures 4.5 and 4.6 suggest a practical recommendation on how long the user need to run the model. As shown in Figure 4.5, the number of iterations needed is a function of the mean difference between genotypes and in order to compare genotypes, we may need increased precision depend on the similarity of genotypes. It means that for the similar genotypes, we may need a large number of of bootstrap samples while if we have a substantial gap between genotypes, less precision is needed. As can be detected in Figure 4.5, to have a *reliable* estimation on how G8 outperform G3 in oat data set (e.g., $\hat{P}(G8 > G3) > \frac{1}{2}$), i.e., have its win probability converged, we might need to have more than 8000 bootstrap samples.

On the other hand, when comparing top genotype, i.e., G8, to another genotype, e.g., G5, it can be detected that the convergence starts in much lesser number of bootstrap samples. It is shown in Figure 4.6 that the trend gets almost stable after around 2000 iterations. This procedure has been reproduced three times for validation as shown in Figures 4.5 and 4.6. This suggests that bootstrapping as having good asymptotic properties can account for a proper comparison when one consider to draw enough samples to have accurate probabilities.

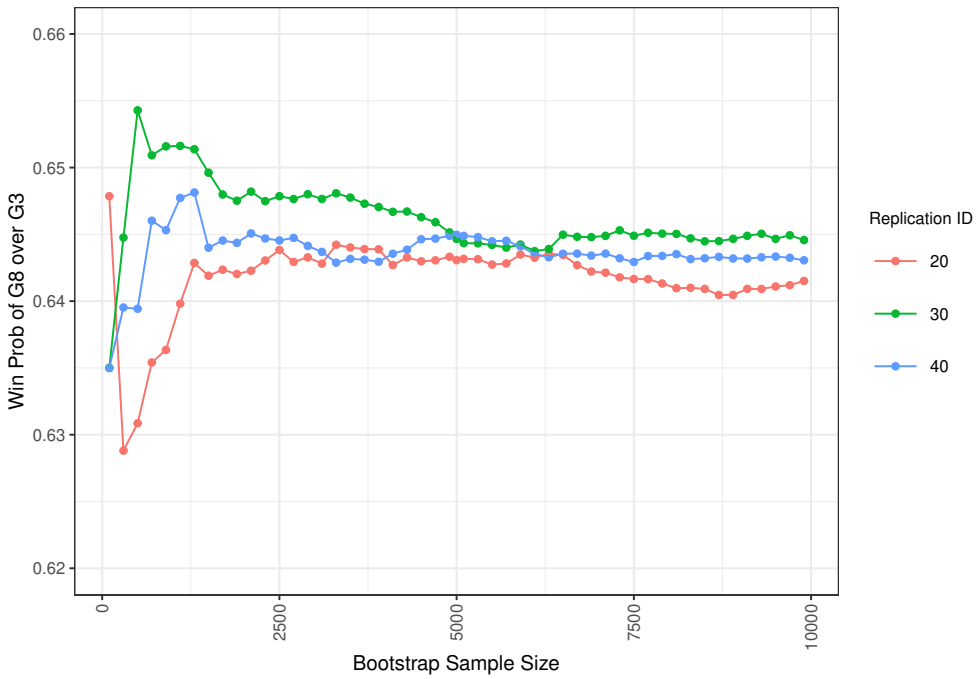

Figure 4.5: Convergence of the estimated win probability based on pairwise comparison for genotypes G8 and G3.

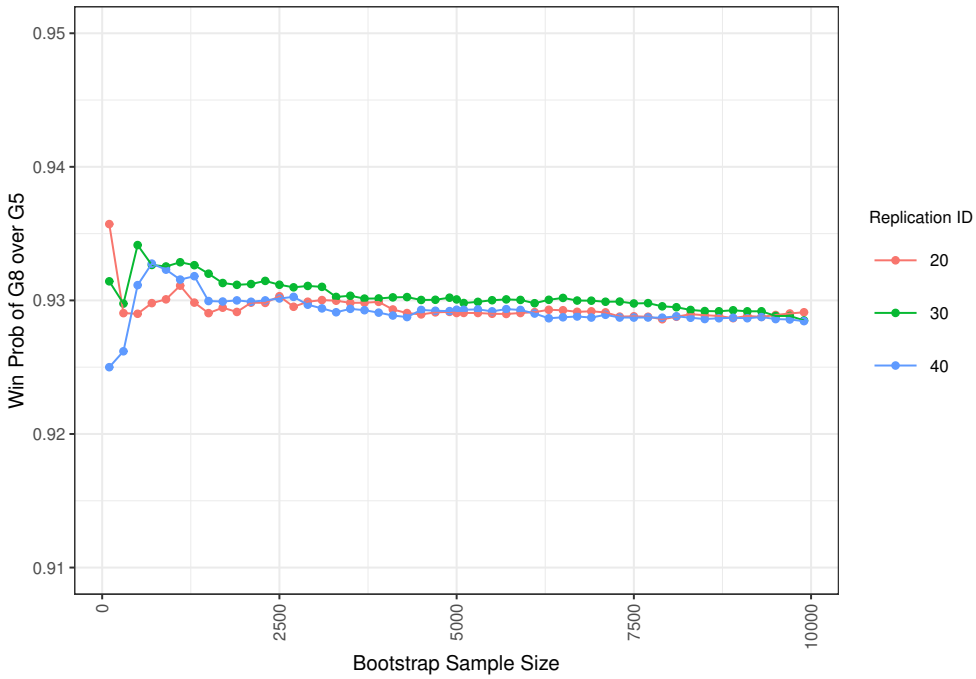

Figure 4.6: Convergence of the estimated win probability based on pairwise comparison for genotypes G8 and G5.

4.4.2 Rapeseed Dataset

In this part, we use rapeseed yield multi-environment trial that has been described by (Shafii et al., 1992) and is available through `agridat` package (Wright, 2021). It contains 6 cultivars (genotypes) grown in 14 locations across 3 years of 1987, 1988, and 1989.

4.4.2.1 Main Functions and Their Outputs

Here, a same procedure done for oat dataset in previous part, is implemented for rapeseed dataset. Below is a summary of `data_preparation` output dataframe with a brief information of each column, their data and structures.

```
rapeseed_data = data_preparation(data)
glimpse(rapeseed_data)
```

```
## Rows: 648
```

```
## Columns: 9
## Groups: GENOTYPE, ENVIRONMENT [162]
## $ ENVIRONMENT    <chr> "GGA 87", "GGA 87", "GGA 87", "GGA 87", "GGA 87", "GGA ~
## $ YEAR           <int> 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87,~
## $ LOCATION       <fct> GGA, GGA, GGA, GGA, GGA, GGA, GGA, GGA, GGA, GGA, GGA, ~
## $ REPNO          <fct> R1, R2, R3, R4, R1, R2, R3, R4, R1, R2, R3, R4, R1, R2,~
## $ GENOTYPE       <fct> Bienvenu, Bienvenu, Bienvenu, Bienvenu, Bridger, Bridge~
## $ PT             <dbl> 960.61, 1329.39, 1781.11, 1698.16, 1605.13, 1211.69, 13~
## $ P_Env          <dbl> 1396.081, 1396.081, 1396.081, 1396.081, 1396.081, 1396.~
## $ Corrected_PT   <dbl> -435.47125, -66.69125, 385.02875, 302.07875, 209.04875,~
## $ M_Corrected_PT <dbl> 46.23625, 46.23625, 46.23625, 46.23625, -32.78625, -32.~
```

As mentioned in the previous section, the `experiment_data` is returned as the first output of `rank_analyzer` package concerning traceability. Therefore, we bring the review for the rest of the outputs.

```
glimpse(rank_analysis_info[[2]])
```

```
## Rows: 6
## Columns: 9
## $ GENOTYPE       <fct> Glacier, Bienvenu, Bridger, Jet, Cascade, Dwarf
## $ `1`            <dbl> 0.71, 0.10, 0.19, 0.04, 0.02, 0.01
## $ `2`            <dbl> 0.18, 0.28, 0.35, 0.14, 0.04, 0.03
## $ `3`            <dbl> 0.07, 0.26, 0.26, 0.22, 0.11, 0.11
## $ `4`            <dbl> 0.02, 0.22, 0.09, 0.29, 0.19, 0.17
## $ `5`            <dbl> 0.01, 0.10, 0.07, 0.24, 0.24, 0.31
## $ `6`            <dbl> 0.00, 0.04, 0.04, 0.07, 0.39, 0.37
## $ M_Corrected_PT <dbl> 113.51, 119.68, 1.34, -89.51, -45.74, -99.28
## $ Rank           <dbl> 2, 1, 3, 5, 4, 6
```

The above table shows the probability of each cultivar being ranked from 1 to 6 along with their mean yield and rank. The third object returned from `rank_analyzer` is a list off genotypes with their most probable rank derived from the bootstrap resampling procedure.

```
glimpse(rank_analysis_info[[3]])
Rows: 6
Columns: 2
$ GENOTYPE          <fct> Glacier, Bienvenu, Bridger, Jet, Cascade, Dwarf
$ MostProbable_rank <dbl> 1, 2, 2, 4, 6, 6
```

The forth object is a table of rank frequency of genotypes in all locations throughout all bootstrap resamples. This output can be potentially useful when analyzing a subset of environments, if needed.

```
glimpse(rank_analysis_info[[4]])
```

```
## Rows: 300
## Columns: 7
## $ GENOTYPE <fct> Bienvenu, Glacier, Bridger, Cascade, Jet, Dwarf, Bienvenu, Gl~
## $ `1`      <int> 2, 22, 0, 0, 4, 0, 2, 15, 10, 1, 0, 1, 27, 1, 0, 0, 0, 0, ~
## $ `2`      <int> 1, 2, 16, 0, 8, 2, 1, 12, 13, 1, 1, 2, 0, 21, 1, 0, 4, 3, 11,~
## $ `3`      <int> 10, 2, 2, 8, 4, 4, 10, 0, 2, 2, 9, 9, 0, 1, 8, 0, 10, 9, 10, ~
## $ `4`      <int> 7, 0, 3, 11, 2, 4, 4, 0, 0, 3, 5, 12, 0, 2, 7, 0, 13, 5, 4, 0~
## $ `5`      <int> 6, 1, 1, 6, 9, 2, 9, 0, 0, 2, 9, 3, 0, 2, 11, 0, 0, 10, 1, 0,~
## $ `6`      <int> 1, 0, 5, 2, 0, 15, 1, 0, 2, 18, 3, 0, 0, 0, 0, 27, 0, 0, 1, 0~
```

The fifth output derived from `rank_analyzer` is a dataframe of genotypes' mean phenotype in all bootstrap resamples.

```
glimpse(rank_analysis_info[[5]])
```

```
## Rows: 300
## Columns: 2
## $ GENOTYPE  <chr> "Bienvenu", "Glacier", "Bridger", "Cascade", "Jet", "Dwarf",~
## $ MeanYield <dbl> -11.387, -35.096, 59.356, -93.081, 97.228, -17.020, -21.710,~
```

The pairwise comparison probabilities of rapeseed genotypes is the sixth output returned by `rank_analyzer`.

```
glimpse(rank_analysis_info[[6]])
```

```
## Rows: 30
## Columns: 3
## Groups: GENOTYPE [6]
## $ GENOTYPE    <chr> "Bienvenu", "Bienvenu", "Bienvenu", "Bienvenu", "Bienvenu"~
## $ Compared_to <chr> "Bridger", "Cascade", "Dwarf", "Glacier", "Jet", "Bienvenu~
## $ Probability <dbl> 0.440, 0.613, 0.693, 0.273, 0.660, 0.560, 0.667, 0.634, 0.~
```

The last object returned by this function is the table of probabilistic ranks for rapeseed data.

```
glimpse(rank_analysis_info[[7]])
Rows: 6
Columns: 3
$ GENOTYPE          <chr> "Bienvenu", "Bridger", "Cascade", "Dwarf", "Glacier", "Jet"
$ observed_Rank     <int> 1, 3, 4, 6, 2, 5
$ ProbabilisticRank <int> 3, 2, 4, 6, 1, 5
```

The outputs of `rank_analyzer` are then feed into the visualization function that incorporate valuable insights to users, especially breeders, from the proposed bootstrapping approach.

4.4.2.2 Visualization Function and its Outputs

Figure 4.7 is showing the rank CIs of rapeseed from `rank_visualization` for the mean yield ranks with 80% confidence level ($\mathtt{p} = 0.8$).

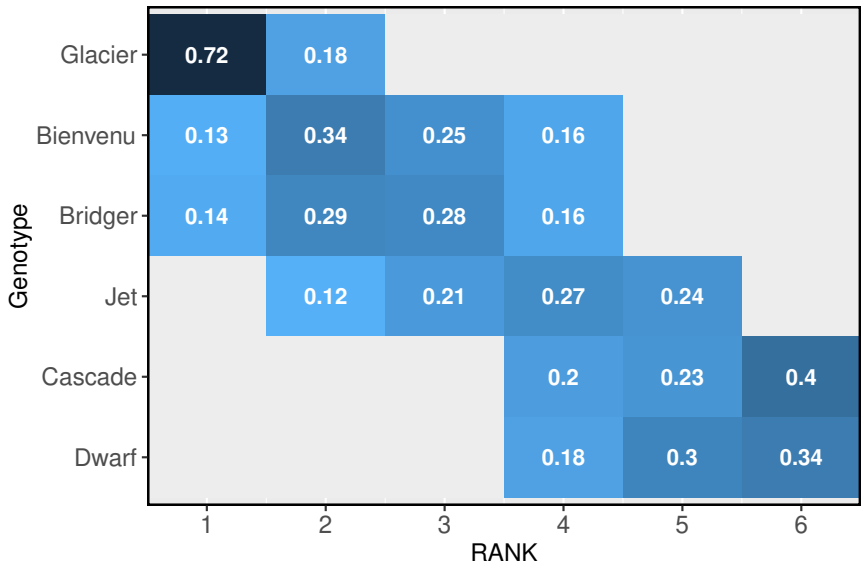

Figure 4.7: Confidence intervals of rank when ranking rapeseed cultivars according to mean yield across the target environments.

Regarding the performance comparison of the genotypes in rapeseed data, Shafii et al. (1992) have shown that the Bridger and Bienvenu cultivars have strong interaction effects with the environment whereas the Glacier cultivar had the least interaction effects. They have also shown that Glacier is the most stable cultivar, whereas Bridger and Bienvenu are the least stable (Shafii and Price, 1998).

We can see that confidence intervals of rank can detect their findings in Figure 4.8. It also shows how Bienvenu and Bridger are adaptive and have a chance to perform very well in their desirable environments, which might become misleading when compared with a good-performing stable cultivar if the observed environment distribution gives the preference toward them. This brings up concerns when decision-making. The literature has also tried to explain it (Shafii et al., 1992; Shafii and Price, 1998; Tai, 1971).

One other output we find an insightful visualization is the CIs we get from probabilistic ranks. When using probabilistic rank, as it captures both genetic effect superiority and stability simultaneously, one can consider these decision criteria being reflected in probabilistic rank and be more confident about their decision because probabilistic ranking tries to capture the

uncertainty stems in limited environments observed. This is the case for the rapeseed dataset when we visualize the probabilistic rank CIs, shown in Figure 4.8, where we get tighter CIs for the top cultivars that implies a greater degree of precision for selecting Glacier.

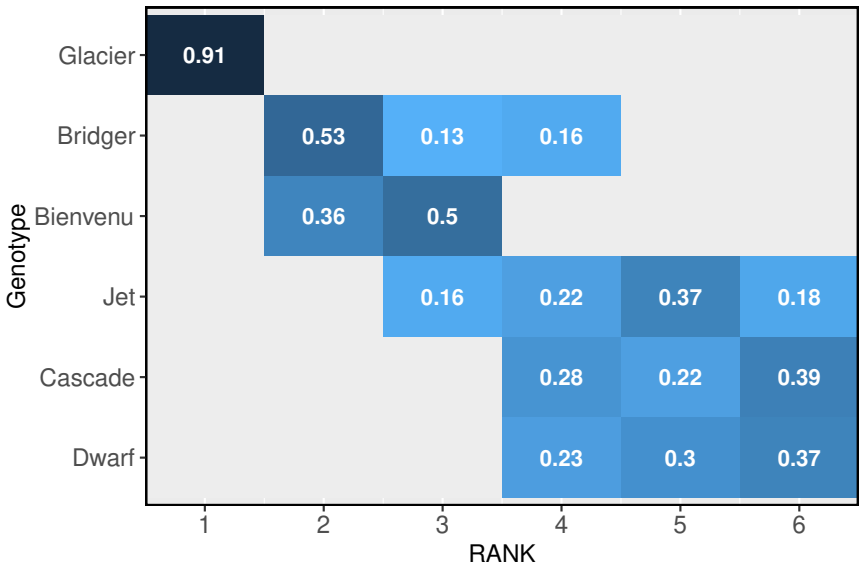

Figure 4.8: Confidence intervals of rank when ranking rapeseed cultivars according to which is the most likely to have higher yield across the target environments.

Looking at Table 4.2, the Bienvenu cultivar ranks the highest based on mean yield, followed closely by the Glacier cultivar (2487.95 and 2481.78 kg/ha). It is therefore predictable that the probabilistic analysis ranks Glacier ahead of Bienvenu due to its superior stability. An interesting observation is that the cultivar that is third according to mean yield rank, namely Bridger, also ranks above Bienvenu according to the probabilistic analysis. The mean yield of Bridger is almost 118 kilograms smaller than Bienvenue and only 101 kilograms more than the Dwarf cultivar that has the smallest yield. Furthermore, the analysis reported by (Shafii and Price, 1998) does not show differences in stability between the two cultivars. Nonetheless, the probabilistic comparison can show that Bridger is more likely to perform better than Bienvenue across these environments.

Table 4.2: Rapeseeds' ranks based on probabilistic and some traditional ranking methods.

Genotype	Ranking Method						
	Probabilistic Rank	Mean Yield	Shukla's Variance	Kang's Rank-Sum	Superiority Measure		
					All Env.	Favorable	Unfavorable
Glacier	1	2	1	1(3)	2	2	3
Bridger	2	3	5	3(8)	5	6	1
Bienvenu	3	1	6	2(7)	1	1	4
Cascade	4	4	4	3(8)	3	4	2
Jet	5	5	3	3(8)	6	5	5
Dwarf	6	6	2	3(8)	4	3	6

4.4.2.3 Convergence of Bootstrap Sampler

Here, the convergence of pairwise comparisons with respect to the bootstrap sample sizes is explored for rapeseed data. Figure 4.9 is showing how method converges fast when the cultivars (genotypes) are far from each other (with respect to their genetic effects) and the model does not need large bootstrapping. On the contrary, Figure 4.10 shows that more than 2000 bootstraps are required for a convergence of Bridger win probability over Bienvenu. This concept needs to be considered by the user when inserting the bootstrap samples matrix of `boots_matrix` to `rank_analyzer` function.

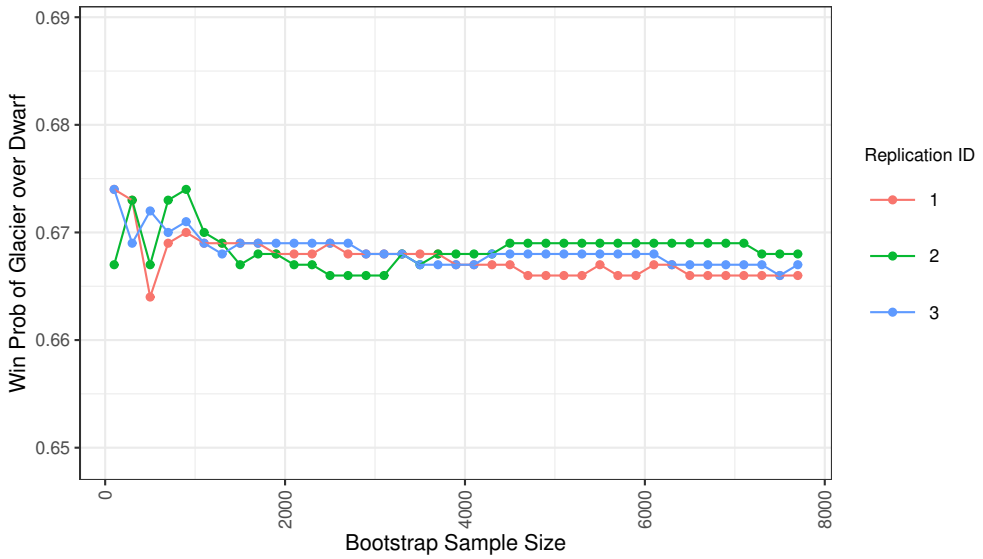

Figure 4.9: Convergence of the estimated win probability based on pairwise comparison for Glacier and Dwarf.

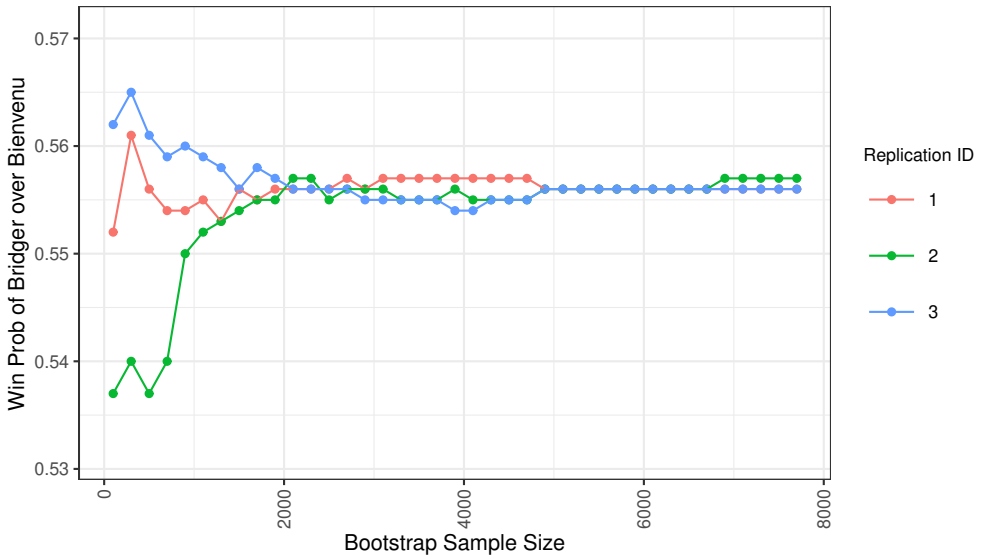

Figure 4.10: Convergence of the estimated win probability based on pairwise comparison for Bridger and Bienvenu.

To sum up the understanding from analyzing the rapeseed dataset, one can notice that along with the literature that has done a detailed analysis of this data using the AMMI model, biplots based on the principal components obtained from that model, and thorough investigation of the environment-by-environment performance of cultivars, the user can detect the same conclusion from Figure 4.8 when using the new proposed method. To the best of our knowledge, no other method has pulled together all the information and shown this conclusion in a summary plot.

4.5 Conclusion

There exists a lot of tools that are being used for analysis of multi-environment trial data. In this chapter, we try to add a new one for analyzers to investigate METs from a new perspective. It would be particularly interesting for breeders to make sure they have considered both the traditional and newly proposed approaches when decision-making. METanalyzeR package can be used on real world large datasets as well as synthetic datasets for specific research explorations. The development version of this package is on Github for usage at (Bijari). It can also be installed using `devtools` as below:

```
# install.packages("devtools") run this line to install devtools
devtools::install_github("ReyhanehBijari/METanalyzeR")
library(METanalyzeR)
```

As the next step, we will upload an stable version of METanalyzeR package on CRAN.

4.6 References

Abberton, M. (2012). Molecular plant breeding. by y. xu. wallingford, uk: Cabi (2012), pp. 752, £59.95. isbn 9781845939823. *Experimental Agriculture*, 48(3):461–461.

Bijari, R. MS Windows NT kernel description.

Cooper, M. and DeLacy, I. H. (1994). Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theoretical and Applied Genetics*, 88:561–572.

DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3).

Dudewicz, E. J. (1980). Ranking (ordering) and selection: An overview of how to select the best. *Technometrics*, 22:113.

Eberhart, S. A. and Russell, W. A. (1966). Stability parameters for comparing varieties. *Crop Science*, 6:36–40.

Efron, B. (1987). Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, 82(397):171–185.

Gibbons, J. D., Olkin, I., and Sobel, M. (1979). An Introduction to Ranking and Selection. *The American Statistician*, 33(4):185–195.

Govindarajulu, Z. and Harvey, C. (1974). Bayesian procedures for ranking and selection problems. *Annals of the Institute of Statistical Mathematics*, 26(1):35–53.

Kang, M. S. (1988). A rank-sum method for selecting high-yielding, stable corn genotypes. *Cereal Research Communications*, 16(1/2):113–115.

Laird, N. M. and Louis, T. A. (1989). Empirical Bayes Ranking Methods. *Journal of Educational Statistics*, 14(1):29–46.

LIN, C. S. and BINNS, M. R. (1988). A superiority measure of cultivar performance for cultivar × location data. *Canadian Journal of Plant Science*, 68(1):193–198.

Malosetti, M., Ribaut, J. M., and 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 MAR.

Olivoto, T. and L'ucio, A. D. (2020). metan: an r package for multi-environment trial analysis. *Methods in Ecology and Evolution*, 11(6):783–789.

Pour-Aboughadareh, A., Khalili, M., Poczai, P., and Olivoto, T. (2022). Stability indices to deciphering the genotype-by-environment interaction (gei) effect: An applicable review for use in plant breeding programs. *Plants*, 11(3).

Sa'diyah, H. and Hadi, A. F. (2016). Ammi model for yield estimation in multi-environment trials: A comparison to blup. *Agriculture and Agricultural Science Procedia*, 9:163–169. International Conference on Food, Agriculture and Natural Resources, IC-FANRes 2015.

Seong-Hee Kim and Nelson, B. L. (2007). Recent advances in ranking and selection. In *2007 Winter Simulation Conference*. IEEE.

Shafii, B., Mahler, K. A., Price, W. J., and Auld, D. L. (1992). *Crop Science*, 32.

Shafii, B. and Price, W. J. (1998). Analysis of genotype-by-environment interaction using the additive main effects and multiplicative interaction model and stability estimates. *Journal of Agricultural, Biological, and Environmental Statistics*, 3.

Shukla, G. K. (1972). Some statistical aspects of partitioning genotype-environmental components of variability. *Heredity*, 29:237–245.

Tai, G. C. C. (1971). Genotypic stability analysis and its application to potato regional trials. *Crop Science*, 11.

Vargas, M., Combs, E., Alvarado, G., Atlin, G., Mathews, K., and Crossa, J. (2013). Meta: A suite of sas programs to analyze multienvironment breeding trials. *Agronomy Journal*, 105:11–19.

Wright, K. (2021). *agridat: Agricultural Datasets*. R package version 1.19.

CHAPTER 5. GENERAL CONCLUSION

5.1 Summary

This dissertation is devoted to help solving a real world problem innovative using data science. There have been lots of efforts in the area of plant breeding to improve the quality of decisions made in such programs. While the use of new techniques has increased in this area, there exist lots of limitations in these programs that tie to unavoidable uncertainties that need to be taken into account for proper analysis. We believe plant breeding would benefit from the body of this work as it tries to fill the gap in the analysis of METs data.

The novel methods presented in this dissertation have two themes in common. First, we aim to account for the uncertainty due to the limited number of environments being observed. This is particularly important in plant breeding because the interaction effects between genotype main effect and the environment are typically large, and it is impossible to ever observe all environments of interest. The year effect is the larger part of the environmental interactions and it is simply impossible to experiment with tens of years. And even within a year, experimenting with a large number of locations is expensive and thus rarely feasible. The second common theme is how the uncertainty is estimated. We propose the use of resampling with replacement of environments, that is, bootstrap resampling, to estimate the probabilities of interest. To the best of our knowledge, this has not been investigated before in the plant breeding context, but our results indicate that it works well and could thus perhaps be utilized further for such data.

More specifically we investigate two new methods. In Chapter 2 we have proposed a novel method for constructing approximate rank confidence intervals when ranking experimental genotypes. The relevant probabilities are estimated using the above-mentioned bootstrap approach and the confidence intervals thus capture the uncertainty due to the selection of environments observed. We show that the empirical coverage of the confidence intervals is good,

that is, they work well in practice, and compare the use of rank confidence intervals to a standard approach.

In Chapter 3 we propose an entirely new method for making pair-wise comparisons between genotypes, namely to prefer the genotype that is more likely to be better across a sample of environments rather than the genotype with the better mean. Again, the probabilities are estimated using bootstrapping. We show that in many cases decisions made using such probabilistic comparison differ from those made using means-based comparison; and this is especially true of cases that are of the most practical interest, namely when the main effects are close and the two genotypes interact with the environment in different manners.

Finally, Chapter 4 pulls together the material from Chapter 2 and Chapter 3. We describe an R package developed so that users can apply these two new methods to support advancement decisions. We further convert the probabilistic comparison into probabilistic rank and compare this new ranking method to standard methods.

5.2 Future Work

As a new perspective to solve the research problem we defined, there is a lot into this research for future work. As probabilistic comparison is not directly estimating any component of the phenotypic response, it combines their effects within itself. The next step can be to investigate whether and how the probabilistic selection includes more than the existing models (whole distribution of phenotype), i.e., it includes both elements of mean prediction (BLUP, AMMI) as traditional selection procedures and stability measures. An insightful future work can extend the applicability of probabilistic comparisons by comparing them with traditional stability measures. Another research can be applying predictive models for explaining when probabilistic selection is different than mean selection.