Chance-constrained Optimization Models for Agricultural Seed Development and Selection

by

Ozkan Meric Ozcan

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved June 2019 by the
Graduate Supervisory Committee:

Dieter Armbruster, Co-Chair
Esma Gel, Co-Chair
Jorge Sefair

ARIZONA STATE UNIVERSITY

August 2019

ABSTRACT

Breeding seeds to include desirable traits (increased yield, drought/temperature resistance, etc.) is a growing and important method of establishing food security. However, besides breeder intuition, few decision-making tools exist that can provide the breeders with credible evidence to make decisions on which seeds to progress to further stages of development. This thesis attempts to create a chance-constrained knapsack optimization model, which the breeder can use to make better decisions about seed progression and help reduce the levels of risk in their selections. The model's objective is to select seed varieties out of a larger pool of varieties and maximize the average yield of the "knapsack" based on meeting some risk criteria. Two models are created for different cases. First is the risk reduction model which seeks to reduce the risk of getting a bad yield but still maximize the total yield. The second model considers the possibility of adverse environmental effects and seeks to mitigate the negative effects it could have on the total yield. In practice, breeders can use these models to better quantify uncertainty in selecting seed varieties.

TABLE OF CONTENTS

Page

LIST OF TABLES iii

LIST OF FIGURES iv

CHAPTER

1.0 INTRODUCTION 1

1.1 Overview 1

1.2 Background 2

1.3 Goals and Limitations 3

1.4 Knapsack Problem 3

1.5 Findings 5

1.7 Organization 5

2.0 LITERATURE REVIEW 6

2.1 Agricultural Seed Development 6

2.2 Robust and Chance-Constrained Optimization 8

3.0 METHODOLOGY 15

3.1 Baseline Scenarios 16

3.1.1 Naïve Knapsack 16

3.1.2 Risk Reduction Knapsack 17

3.2 Adverse Environmental Cases 19

CHAPTER Page

4.0 RESULTS 21

4.1 Set-Up..... 21

4.2 Results of the Baseline Case 22

4.4 Adverse Environment Cases 25

4.4.1 Results of Drought Scenario..... 25

4.4.2 Results of Rainy Scenario..... 32

4.4.3 Results of Extreme Temperature Scenario 37

4.5 SAA Method 42

4.6 Conclusions..... 43

REFERENCES 45

APPENDIX

A BASELINE DATA SET……..…….48

B ADVERSE ENVIROMENT BASELINE DATA SET ……50

C ADVERSE ENVIROMENT DROUGHT DATA SET……..…52

D ADVERSE ENVIROMENT RAIN DATA SET …….54

E ADVERSE ENVIROMENT EXTREME TEMPERATURE DATA SET……56

List of Tables

Table Page

1. Potential Responses to Adverse Environments 19

2. Summary of Results for Baseline Case 22

3. Summary of Results for Results for Optimal Knapsack without Drought 26

4. Summary of Results for Optimal Knapsack with Drought 28

5. Changes in the Naïve and Risk Knapsack with Drought 30

6. Summary of Results for Optimal Knapsack without Rain 33

7. Summary of Results for Optimal Knapsack with Rain 34

8. Changes in the Naïve and Risk Knapsack with Rain 37

9. Summary of Results for Optimal Knapsack without Extreme Temperature 38

10. Summary of Results for Optimal Knapsack with Extreme Temperature 38

11. Changes in the Naïve and Risk Knapsack with Extreme Temperature 42

12. SAA Method 42

List of Figures

Figure Page

1. Normal Curve for Naive Knapsack and (691;20%) Risk Knapsack..... 24

2. Normal Curve for Naïve Knapsack and all Risk Scenarios..... 25

3. Normal Curve for Naïve Knapsack and Risk Scenarios without Drought 27

4. Normal Curve for Naïve Knapsack and Risk Scenarios with Drought 29

5. Normal Curves for Changes in the Drought/No Drought Naïve and (522; 20%) Risk Constraint Knapsack..... 31

6. Normal Curve for Naïve Knapsack and Risk Scenarios without Rain 33

7. Normal Curve for Naïve Knapsack and Risk Scenarios with Rain 35

8. Normal Curves for Rain/No Rain Naïve and 1399,20% Risk Constraint Knapsack.... 36

9. Normal Curve for Naïve Knapsack and Risk Scenarios without Extreme Temperature 39

10. Normal Curve for Naïve Knapsack and Risk Scenarios with Extreme Temperature 40

11. Normal Curve for Naïve and 1340,20% Risk Constraint Knapsack 41

Chapter 1:

Introduction

1.1 Overview

The current world population is 7.6 billion, and it is expected to grow above 9.8 billion by 2050 (UN). According to the population reference bureau, the current rate of population increase is 1.2%. This means that the world's population of 7 billion will double to 14 billion within the next 60 years. This upward growth is expected to continue and place an even greater burden on the earth's ability to provide enough food for its inhabitants. The Food and Agriculture Organization of the UN states that more than 66% of the world's population is malnourished. This is a 300% increase in the number of people who are malnourished when compared to 1950.

Crop farming is a reliable way of growing food, however simply growing more crops is not a practical solution due to the extra strain this will place on the earths supply of arable land. Also, climate change has been shown to reduce wheat yields by 6% for every 1-degree Celsius increase in global temperatures. Rice yields are also reduced as nighttime temperatures increase (Environmental Health News). Additionally, most farming is dependent on irrigation water which is gained from snowmelt, usually stored in mountain snow packs. Climate change has been shown to reduce snow packs and therefore lessen the availability of irrigation water.

In order to overcome these issues, countries and companies around the world have increasingly focused on developing crops with higher bushel yields (US NEWS). For example, some Asian countries have been able to keep up food production in part due to developing new breeds of Asian rice. (FOA)

1.2 Background

When it comes to creating new seeds, farming and agricultural groups set the expectations. These groups advise farmers on what to look for in varieties and how to make good choices when selecting them. For example, Iowa State University's Department of Agronomy is one of these groups. They advise farmers to look for five things when selecting soybean varieties. They are, in order of importance, yield, disease resistance, maturity group, grain composition, hedging and lodging.

Another example is the Rice Knowledge Bank, which as its name suggests advises farmers on what to look for when selecting rice seeds. The Rice Knowledge Bank recommends that farmers find seeds which meet market expectations like color, shape and cooking characteristics, provide adequate yield, have disease resistance, have adequate tillering capacity and resistance to lodging. This advice causes there to be a market demand for certain traits, and it is the breeder's job to come up with new varieties that can satisfies these demands.

To develop new varieties of seeds, breeders use stage-gating. Breeders will start with a large pool of seeds and each have some desirable trait that the breeders have developed in the seeds. During a stage, breeders will plant each seed in a different location around the United States for one growing season. At the end of the growing season, a varieties performance will be judged based on the varieties total yield. The seeds with the best performance are promoted to the next stage and this cycle continues until three stages are completed. The seeds remaining at the end of the third stage are released to the market. This process takes several years to complete from start to finish, so the development of a poor performing seed is very costly. Progression decisions are largely based on the

expertise of the breeders and the results of the three seasons of planting. This is a risky method of progression because there is no way to confirm or disprove the assumptions made by the breeders and the three seasons of planting may misrepresent the real performance of the seeds. As a result, mathematical models are developed to help breeders in making more objective decisions and reduce the risk of selecting poor performing seeds.

1.3 Goals and Limitations

The goal of this research is to assess the potential for using robust optimization methodologies in helping seed breeders make better seed advancement decisions. Since breeders evaluate seed performance based on yield, this thesis is focused on creating an optimization model which helps select seed varieties which maximize the average yield of the selected group. The study also considers potential variance in the yield caused by any number of factors. This way, the solution to the model is a knapsack which both maximizes the yield but also acts to hedge against uncertainty. This study is limited by the small amount of data available on seed variety performance. Another limitation is the lack of computing power when trying to solve more complex optimization problems.

1.4 Knapsack Problem

The breeders' job is to select a subset of seed varieties that they think will succeed in the market. Usually, there will be some constraints on which seeds they can select and how many, usually it is a group of seeds which are progressed together. A group of seeds is progressed instead of a single variety because seeds are progressed through stages, and each stage eliminates more seeds until the most elite seeds are selected. This structure defines the knapsack problem, which is an optimization model which selects a subset of

items out of a larger set based on some objective function and constrained, usually, by a size limit on the subset. The most basic deterministic knapsack set up is as follows,

$$\max \sum_{i \in I} a_i x_i$$

$$\text{s.t.} \sum_{i \in I} x_i \leq C$$

$$x_i \in \{0,1\}$$

In this case, x is a binary decision variable on whether to include item i in a set I of alternative items in the knapsack. The objective of this knapsack model is to maximize the total knapsack. The equation $\sum_{i \in I} a_i x_i$ represents the total value of the knapsack, here a_i is the value of item i. C is the maximum size of the knapsack and the constraint $\sum_{i \in I} x_i \leq C$ prevents the model from selecting more items then the total size of the knapsack. Additional constraints can be added to this model make it more specialized.

The chance-constrained knapsack is simply the knapsack model above with one or more chance constraints. In a basic form it looks like this

$$\max \sum_{i \in I} a_i x_i$$

$$\text{s.t.} \sum_{i \in I} x_i \leq C$$

$$Pr(W \geq R) \geq \alpha$$

$$x_i \in \{0,1\}, \quad \forall i \in I$$

Here Pr() represents the probability of an event occurring, and W and R are some values of interest and α is the probability of success.

1.5 Findings

The novelty of this work is in showing the importance of considering uncertainty in seed selection and how using a chance-constrained knapsack optimization model can help breeders make better selection decisions. We show that using a naïve (selection based only on the mean) approach to the seed selection problem can result in very low means when uncertain events occur. Therefore, we use optimization to find a compromise between finding varieties which maximize the knapsack value and hedging against uncertainty. This research shows that knapsacks selected by this optimization has a smaller total value, but also increases the chances of performing better when uncertain events occur. Using this model, breeders can select their own requirements such as minimum yield requirements and identify which seeds maximize the knapsack's yield.

1.7 Organization

The first chapter of this thesis gives an introduction of the thesis. It explains the motivation of the work and basic topic. The second chapter is the literature review and it explains the current work being done in this field. It also details the work which has gone into the fields of robust optimization and chance constraints. The third chapter talks about the methodology of the thesis. It goes into detail about which models are used and how, plus what assumptions are being made. The fourth chapter shows the findings of the thesis and discusses the implications of the findings. The fifth chapter wraps up the thesis and reiterates the importance and results of the work.

Chapter 2:

Literature Review

2.1 Agricultural Seed Development

Historically, most research involved in the crop yield improvement was revolved around improving the environment that the crop was planted in. However, more recently, agricultural research has started to include more genetics. A widely used tool for agricultural research is optimization. In the case of genetic research this includes optimizing which genes are selected and how many new genes are utilized in a single seed. This technique coupled with useful metrics can help seed breeders breed seeds with interesting and valuable properties.

Before researchers considered genetics, most research was done on the environment. Optimization was used to properly use farm resources, like water, sun light and nutrients in the soil. For example, Goulding's (2008) paper on resource management focuses on nitrogen use in the soil. Nitrogen is a critical nutrient for a variety of plants and needs to be replaced in the soil each year. However, an overabundance of nitrogen in the soil can also be detrimental to crop yields. Goulding (2008) describes the twin problems of countries how have too much nitrogen in their soil and countries without enough. He also mentions the importance of proper nitrogen balance in protecting the environment. Goulding (2008) suggests several best practices in proper nitrogen balance. These include regular data collection, the use of lime, proper fertilizer use calculation and careful irrigation. Another example is the use of optimization in selecting which type of crops to plant and how much land to allocate to each crop type. Boles (1955) sets up this problem

as an optimization problem, where the objective is to select the combination of crops and land allocation which maximizes the yearly profit of the farm. Each crop and allocation plan have a cost associated with it. He sets up a linear program which when solved provides the best of combination of these two variables.

More recently, researchers have also been including genetics into their research. Generally, the goals of such research are to discover new ways of adding new genes the gene pool of an already existing seed. Gene stacking or gene pyramiding is a popular problem currently being researched; it is defined as trying to introduce several desirable genes from many parents into a single genotype for a specific trait. Beukelaer (2015) explains how marker-assisted gene pyramiding problems are usually solved using integer programming. Beukelaer (2015) then goes on to say that the heuristic method which he helped developed can work in smaller cases. In larger cases this method can provide good approximations. Other researchers like Canzar's (2011) use mathematical optimization to solve the gene pyramiding problem. Canzar (2011) attempts to minimize the number of generations and population size required to have a seed exhibit the properties of a certain gene. He shows that the general problem is NP-hard but that the problem difficulty can be reduced by taking advantage of the combinatorial structure of the problem.

Alongside methods on how to introduce genes to seeds, researchers have also been working on how to decide which genes to include. Traditionally, gene selection is arbitrary and based mostly on the breeder's intuition or what the breeder believes to be the market's demand. Researchers like Byrum et. al (2017), explain how they developed a metric which can help measure genetic gain. Genetic gain is the amount of increase in performance, in this case yield performance, that is achieved through the addition of genes. This metric can

help identify which genes assist in yield performance the most and breeders can use it as a way of more rigorously selecting which genes are important to pursue.

Byrum et. al (2016) also discusses several operations research tools that are used in seed development decision making. The trait introgression tool evaluates the time, cost and probability of success during the variety design phase. This tool uses discrete-event simulation to asses these outcomes. This is an important step in improving the design phase of seed development. Using this tool, breeders can plan future steps in the design phase and measure the consequences of those decisions before implementing them. Another tool used during the variety design phase is the breeding project lead tool, this tool helps the project lead determine the best use of resources like fields, facilities and plant materials. It also uses discrete-event simulation. For field trails, a tool called the yield trail design optimizer is used to identify the optimal number of varieties, locations and replications to test. This tool helps the breeders quickly decide which varieties meet commercial expectations. This tool requires three probability distributions in order to function correctly. In requires the variability of intrinsic yield, variability due to location and variability across replications. These distributions used to simulate the yield of varieties at each location and each replication. By randomly generating values from those distributions, the model will calculate the mean of the variety and top performing varieties are selected for trail testing.

2.2: Robust and Chance-Constrained Optimization

Using deterministic optimization, which assumes that the parameters of the problem are known with certainty can result in poor or unrealistic results. Robust

optimization is a method of quantifying uncertainty into an optimization model (Pinar 2005). It is a wort-case oriented methodology and does not require prior knowledge of a distribution and is useful when parameters are uncertain, and when the objective function value is highly sensitive to parameter changes (Ben-Tal 2009). In the case of this research, it is assumed that the underlying distribution is known, so chance-constrained optimization is used.

Chance-constrained optimization is a major approach to solving optimization problems that deal with uncertainty and is used in many industries, including finance, water management and renewable energy. Chance-constrained optimization works by replacing the uncertain parameters with random variables. This alteration makes the problem more accurate by better modeling the problem but changes the structure of the model which causes certain problems in coming to a solution (Gelute 2012). For one, chance-constrained problems are both non-convex and non-linear (Henrion & Strugarek 2008). Second, calculating correct probability distributions can be very difficult and slight inaccuracies in the probabilities can greatly alter the solution of the model (Uryasev 1995). Certain solution methodologies have been proposed to overcome these problems, including approximating distributions and different ways of formulation (Ahmed & Shapiro 2008).

Stochastic or chance-constrained optimization was first described by Charnes and Cooper (1959). They described the concept as a problem where certain random numbers are selected as a function of a random variable with a known distribution in such a way that it maximizes some objective function subject to constraints that the probability be maintained above or at some value.

The generic way to express the stochastic constraint is

$$\min_{x \in X} f(x) \quad \text{subject to } Pr\{G(x, \omega) \leq 0\} \geq 1 - \alpha,$$

where ω is the random variable and $Pr\{G(x, \omega) \leq 0\}$ is the probability that a certain constraint will be greater than or equal to 0. The program seeks to identify a decision vector x, which minimizes the objective function and satisfies the constraint with a probability of $1 - \alpha$, where $0 \leq \alpha \leq 1$.

Gelute (2012) describes in further detail the various applications, properties and numerical issues that come with chance-constrained optimization problems. Gelute (2012) mentions that the classical applications for chance-constrained optimization includes water reservoir management, optimal power flow and reliability engineering and that modern applications include unmanned drone navigation and reliable wind and power generation. He also mentions the difficulty of calculating the probability distribution for ω. However, he does mention a special case:

If $G(x, \omega) = a^T x + b - \omega, \omega \in \mathbb{R}$ and $\omega \sim N(\mu, \sigma_2)$, then

$$Pr\{G(x, \omega) \leq 0\} \geq \alpha \leftrightarrow \varphi^{-1}(1 - \alpha) - (a^T x + b) \geq 0.$$

Gelute (2012) explains two approximation strategies, all three strategies attempt to circumvent the computational problems (non-convexity, non-linearity, calculating the probability distribution) with chance-constrained problems. The first method is called back-mapping. The idea is to find a monotonic relationship between the function for the random variable and the true distribution of the random variable, which can help create direct representation of the chance constraints. The disadvantage is that such a relationship may not exist for every problem.

The second method is called sample average approximation (SAA), which was created by Ahmed and Shapiro (2008). SAA removes two of the main difficulties in solving chance-constrained optimization problems, which is the non-convex nature of chance-constrained optimization problems and difficulty checking for feasibility. These problems are avoided by replacing the true distribution of the random variables with an empirical distribution created through a Monte Carlo simulation. SAA is a useful approximation and can create good approximate solutions to the exact case. However, even though SAA removes some of the problems with chance constrained optimization, it is still a NP-hard problem. To resolve this issue, Ahmed and Shapiro use a mixed-integer program to solve the SAA formulation. The formulation is

$$\min f(x)$$

$$\text{subject to } G\left(x, \omega^j\right) \leq M_j z_j$$

$$\sum_{j=1}^{N} z_j \leq \gamma N$$

$$z_j \in \{0,1\}$$

$$x \in X$$

where $G\left(x, \omega^j\right)$ is a function of the decision variables and random distribution, γ is the allowable risk factor, N is the total number of samples and z is the count of failures, j is the set of samples, X is the set of decision variables, and M is any very large number. This formulation "counts" the number of failures within the number of samples and keeps this number below a certain risk factor decided beforehand. This allows us to use chance-constrained constraints without having to calculate the probability distribution.

Henrion (2004) further explains the benefits of chance-constrained optimization by explaining how deterministic solutions to some problems are unstable. In his cash matching problem example, he shows how small changes to the initial payments can make a big difference in the optimal value and decision variables. The cash matching problem is the problem of trying to pay pension costs by financing these costs through the purchase of three types of bonds. The goal is to maximize the amount of money remaining at the end of each year.

Henrion (2004) also makes the point that because of the potential for real-world situations, depending on exact cash payments is risky. Rather, he proposes a more robust solution by treating the initial cash payments as functions of a random variable and creating constraints which enforce the rule that the probability of having a positive cash flow exceeds 95%. This problem, which has random parameters on the right-side of the inequality, can be simplified by taking advantage of the normal distribution. By using the 95-quantile and multiplying it by the expected value of the cash payment, it can effectively enforce the 95% constraint. This makes the stochastic constraint into a simple linear programming one. The constraint transforms in this fashion

$$Pr(\sum_{i=1}^{n} a_{ij}x_i \geq \omega_j) \geq p \Leftrightarrow \sum_{i=1}^{n} a_{ij}x_i \geq b_j + \grave{\sigma}_j q_p$$

where q_p is the percentile used from the normal distribution.

There is also an important distinction to be made between the two styles of writing stochastic constraints. It can be rewritten so that all constraints in the problem must pass with a probability of α or that every constraint $i \in I$ where I is the set of constraints must

pass with a probability of α_i. They are known as joint and single chance constraint respectively and can be written as

Single Chance Constraint

$$\text{subject to } Pr\{G_i(x,\omega) \leq 0\} \geq \alpha_i, i = 1, \ldots\ldots, m,$$

Joint Chance Constraint

$$\text{subject to } Pr\{G_i(x,\omega) \leq 0, i = 1, \ldots\ldots, m\} \geq \alpha,$$

In example of the cash matching problem, individual chance constraints are used to guarantee that the probability of having a positive cash flow for each individual year is over 95%. However, by formulating the constraints in this fashion, the overall (across all years) probability of a positive cash flow is lower than 95%. For example, for five years, each year is expected to have positive cash flow 95% of the time, then the probability that all five years is over 95% is 77%, which less than the 95% that was wanted. In the case where the overall probability must be over 95%, joint probability constraints should be used by replacing the many individual chance constraints with the single joint constraint. While this set up has fewer constraints, it is more difficult to solve than the many individual constraints.

Chance-constrained optimization problems can sometimes have problems with convexity (Henrion 2004). Convexity is an important property in optimization because it is a key factor in whether the problem will converge to a single solution. A chance-constrained optimization problem is convex if F_ω, the distribution of the random vector ω, is a quasiconcave function within the constraint

$$\{x|Pr(\omega < x) \geq p\} = \{x|F_\omega(x) \geq p\}.$$

However, it turns out that if the function has the property of being log-concave (which also means that it is quasiconcave), this is enough to say that the problem is convex. Fortunately, many useful and commonly used distributions have the property of being log-concave (multivariate normal, Pareto). (Henrion 2004)

A second problem has to do with the stability of the optimization problem. In theory, chance-constrained problems are constructed using a known probability distribution (Charnes & Cooper 1959). However, finding the correct probability distribution is often difficult or impractical, hence empirical or approximated distributions are often used. The main concern with using an empirical distribution is that by using an approximated distribution, the problem solution does not give the "true" answer. Ultimately, since better approximations will give more accurate results, the final solution to the chance-constrained optimization problem is highly dependent on the accuracy of the probability distribution.

Chapter 3:

Methodology

The problem that many breeders face is trying to decide which seeds will perform well in real world conditions. Currently, breeders mainly look at the average bushel yield of the seed to evaluate its performance. However, looking at mean yields is a risky way of conducting seed selection. There are many factors which can influence the yields of seeds, and this method puts the breeder at risk of these factors lowering the yield. A better way of seed selection is by also considering the potential variance in the seeds yield. This way breeders can make more relevant decisions on seed performance and reduce the risk of getting lower than expected yields.

This thesis uses a chance-constrained knapsack optimization model to optimize the total yield of the selected knapsack, but also hedge against the uncertainty in the seed's yield. This model is relevant for this problem because breeders select only a few seeds to move on to the next stage and want this group of seeds to be the highest performing seeds possible. Therefore, the objective function of the model is to maximize the total yield of the selected knapsack. The constraints include restricting the size of the knapsack to be selected and certain chance constraints which force the model to meet minimum requirements. The knapsack capacity restriction models the fact that breeders only choose a few seeds out of the pool of seeds, and the chance-constraint is what hedges against uncertainty. We are assuming that the distributions of each varieties yields are available.

3.1: Baseline Scenarios

We want to create a knapsack model which maximizes the yield for N different varieties. Here we consider two different cases, one where the knapsack selection is only based on the mean, and another where the knapsack selection considers the uncertainty of the yield. The second knapsack models the impact of variation due to genetic effects, different agricultural practices, local soil or field conditions and local weather phenomena. Variety yields are assumed to be independent and normally distributed.

3.1.1: Naïve Knapsack

This knapsack model only considers the mean of the variety's yield. In most cases, this version of the knapsack problem can be reduced to the rank ordering problem, where every variety is order based on their mean yields the top n number of varieties are selected. As a result, this problem is trivial to solve. The formulation for the naïve knapsack is

$$\max \sum_{i \in I} \mu_i x_i$$

$$\text{s.t.} \sum_{i \in I} x_i \leq C$$

$$x_i \in \{0,1\}, \qquad \forall i \in I$$

Here we are tying to maximize the total knapsack value. μ_i is the mean of the variety i where i is a variety in the set of alternative seed varieties I. The variable x_i is a binary variable and is equal to 1 if variety i is selected. C is the knapsack capacity, representing the maximum number of varieties to be selected.

3.1.2: Risk Reduction Knapsack

In the risk reduction knapsack, we consider the potential yield risk that could be caused by uncertainty and variation. We would again like to select C varieties from a given set, I of alternative seeds, each with an estimated yield randomly distributed (according to a known distribution, which in this case is assumed to be normal) with means μ_i and variances σ_i^2 . We would like to maximize the total average yield that the selected varieties offer (i.e., total average knapsack yield), subject to a risk constraint. We set a minimum weight level, called "required minimum weight" and denoted by R. We would like to ensure that the probability that the total knapsack yield is greater than or equal to R with a probability of at least α.

The model can be stated as

$$\max \sum_{i \in I} \mu_i x_i$$

$$\text{s.t.} \sum_{i \in I} x_i \leq C$$

$$P(W \leq R) \leq 1 - \alpha$$

$$x_i \in \{0,1\}, \qquad \forall i \in I$$

where W is the total knapsack yield which is defined as the sum of the yields of all selected varieties and x_i indicates if a variety is selected for the knapsack. The variable x_i is 1 if the variety is selected, 0 otherwise. W is a random variable given by the sum of the yields of the varieties in the chosen knapsack. When the variety yields are independent and identically distributed, the distribution of the total knapsack yield (which is a random

variable that is a sum of the random variety yields of the varieties in the knapsack) is also normally distributed with mean μ and variance σ^2 that is equal to

$$\mu = \sum_{i \in I} \mu_i x_i$$

$$\sigma^2 = \sum_{i \in I} \sigma_i^2 x_i$$

Hence, when the variety yields are normally distributed and independent, it is possible to rewrite the above model as

$$\max \sum_{i \in I} \mu_i x_i$$

$$\text{s.t.} \sum_{i \in I} x_i \leq C$$

$$\sum_{i \in I} \mu_i x_i + \varphi^{-1}(1-\alpha) \sqrt{\sum_{i \in I} \sigma_i^2 x_i^2} \geq R$$

$$x_i \in \{0,1\}, \quad \forall i \in I$$

where φ denotes the standard normal CDF, and the constraint is due to the fact that

$$P(W \leq R) = P(Z \leq \frac{R-\mu}{\sqrt{\sigma^2}}) = \varphi(\frac{R-\mu}{\sqrt{\sigma^2}}) \leq 1-\alpha \Leftrightarrow \frac{R-\mu}{\sqrt{\sigma^2}} \leq \varphi^{-1}(1-\alpha)$$

Substituting the value of μ and σ^2 gives the stated constraint.

This knapsack should give a different result when compared to the naïve knapsack, since the addition of a constraint reduces the feasible region of this problem, the solution to this problem should be less than that of the naïve knapsack. This also makes intuitive sense, because varieties with high variances but otherwise high means are penalized because of their larger variances and therefore not selected. For this research, we are

assuming that our data is normal, however this will not always be the case. In the results section we will discuss other methods that could be used to achieve similar results to the exact formulation used above using sampling.

3.2: Adverse Environmental Cases

This model considers the potential influences of extreme weather conditions. They are systematic and correlated and are modeled by a random but correlated shift of the probability distribution for the yields of each variety. We consider three cases reflecting the probability of a drought year, the probability of a rainy year (e.g., el Nino year) and the probability of a year with extreme temperatures. For each of these cases we will we still sample in an iid fashion from the shifted probability distributions since the local and uncorrelated random effects mentioned in the baseline scenario are still present. This model has a different baseline than the previous models since a different dataset was used. Each environmental case is compared to the baseline case.

Three cases are established, they are drought, rain, and extreme temperature. There are four possible responses to the scenarios. They are explained below in Table 3.1:

Table 3.0.1: Potential Responses to Adverse Environments

Code	Meaning
(+,+)	Mean Increases, Variance Increases
(+,-)	Mean Increases, Variance Decreases
(-,+)	Mean Decreases, Variance Increases
(-,-)	Mean Decreases, Variance Decreases

Each variety has a 25% chance of having any of the four responses to the adverse environment scenario. The exception to this is the drought scenario, which causes all varieties to experience the (-, +) response. Also, the model used in this design had to be adjusted to

$$\max \sum_{i \in I} \mu_{1i} x_i$$

$$\text{s.t.} \sum_{i \in I} x_i \leq C$$

$$\sum_{i \in I} \mu_{2i} x_i + \varphi^{-1}(1-\alpha) \sqrt{\sum_{i \in I} \sigma_{2i}^2 x_i^2} \geq R$$

$$x_i \in \{0,1\}, \qquad \forall i \in I$$

where μ_{1i} is the baseline mean of variety $i \in I$ and μ_{2i} and σ_{2i}^2 are the response mean and variance of variety $i \in I$ of alternative seeds. x_i is the decision variable and is equal to 1 if variety i is selected, R is the required yield and φ denotes the standard normal CDF.

Chapter 4:

Results

4.1 Set-Up

Each of the 50 varieties was assigned a randomly generated value for the mean and the variance. The means were generated using a uniform distribution between 40 and 80. All means were generated to three decimal places; this was done to avoid multiple alternative optimal solutions due to too many varieties having the same mean yield. The variances were generated using a uniform distribution between 25 and 900. These were also generated to three decimal places.

For the first case, two models, the naïve and the risk reduction knapsack were compared. The naïve knapsack did not use any risk mitigation constraints, but rather found the highest possible knapsack value for a knapsack capacity of C. This set the baseline that the risk reduction knapsacks can be compared to. The risk reduction knapsack found the optimal knapsack value given a risk constraint. For example, the model could be run to find the optimal knapsack yield given a knapsack capacity of 5, with a minimum yield of 500 which can't be violated more than 10% of the time.

In the risk reduction knapsack there were three adjustable parameters for the model. These were the knapsack capacity, the required yield and the allowable percentage of failure. For this study, the knapsack capacity was set to 10, and the allowable percentages of failure were 20%, 10%, 5%, and 1%. The required yield used for each model was empirically found for each percentage. It was found by starting from a required yield of 0 and increasing it by 1, until a required yield which caused the problem to become infeasible was found. The required yield immediately before infeasibility was used for the problem

parameter. This required yield is the highest possible minimum yield for the specific knapsack, and so it is the most desirable minimum yield.

4.2 Results of the Baseline Case

The first entry in the table is the results of the "no risk" or naïve knapsack calculation. It is calculated by removing the chance-constraints in the optimization model. This model will select the C seed varieties with the highest yields, resulting in the knapsack with the highest possible total knapsack value. In this case, $C = 10$ and the naïve knapsack calculation resulted in a knapsack with a total yield of 734.84 and a standard deviation of 53.23. Knapsack yield and standard deviation was calculated using the following formulas.

$$Knapsack\ Yield\ = \sum \mu_s \quad \text{where S is in the set of select varieties.}$$

$$Knapsack\ Standard\ Deviation$$

$$= \sum \sqrt{\sigma_s^2} \quad \text{where S is in the set of select varieties.}$$

Table 4.1: Summary of Results for Baseline Case

Required Yield	Risk	Varieties in the Selected Knapsack	Average Knapsack Yield	Knapsack Standard Deviation
-	-	{X7, X10, X12, X15, X20, X24, X28, X29, X31, X39}	734.84	53.23
691	20	{X6, X7, X10, X15, X20, X24, X28, X29, X31, X39}	732.47	47.91
672	10	{X6, X7, X10, X15, X20, X24, X29, X31, X39, X48}	729.34	43.28
662	5	{X6, X7, X10, X15, X20, X25, X29, X31, X39, X48}	718.43	34.21

The second entry is the results of adding the constraint that the knapsack yield had to be greater than or equal to a value of 691 at least 80% of the time. This resulted in a

knapsack yield of 732.47 and a standard deviation of 47.9. Compared to the naïve knapsack, the (691; 20%) knapsack exchanges the X12 variety for the X6 variety. Here (691; 20%) indicates that this knapsack's yield does not fall below 691 bushels more than 20% of the time. This results in a knapsack yield loss of 2.37 bushels; however, it improves the standard deviation by 5.32. The difference in the selected set of varieties can be attributed to the addition of the risk constraint. This risk constraint forces the model to select less risky varieties and as a result, it attempts to lower the standard deviation of the knapsack. Since the naïve knapsack gives the highest possible knapsack value, it is reasonable to expect to lose some knapsack value in exchange for choosing a less risky knapsack. Figure 4.1 shows the normal curve of the naïve knapsack in blue, and the curve for the (691;20%) knapsack is in orange. We can see that compared to the naïve curve, the 691,20% curve is further to the left but has a narrower curve.

The 3rd and 4th entries in the table are the (672;10%) and (662;5%) risk constraints, respectively. Notice that as the risk percentage decreases, the required yield, knapsack yield and standard deviation decrease as well. The strictest risk constraint is the (662;5%) constraint, which means that the model is finding a knapsack which has a total yield of 662 or higher more than 95% of the time. Compared to the naïve knapsack, this knapsack exchanges varieties X12, X24, X28 with varieties X6, X25, X48. The exchange results in a knapsack yield of 718.43 and a standard deviation of 34.21 bushels, which means a difference in knapsack yield of 16.41 bushels and a difference in standard deviation of 19.02. This represents about a 2% loss in yield when compared to the naïve knapsack, but about a 35% reduction in the standard deviation. This is an important result because it

shows even extremely risk adverse breeders can select safe knapsacks without sacrificing too much of the knapsack yield.

Figure 4.1: Normal Curve for Naive Knapsack and (691;20%) Risk Knapsack

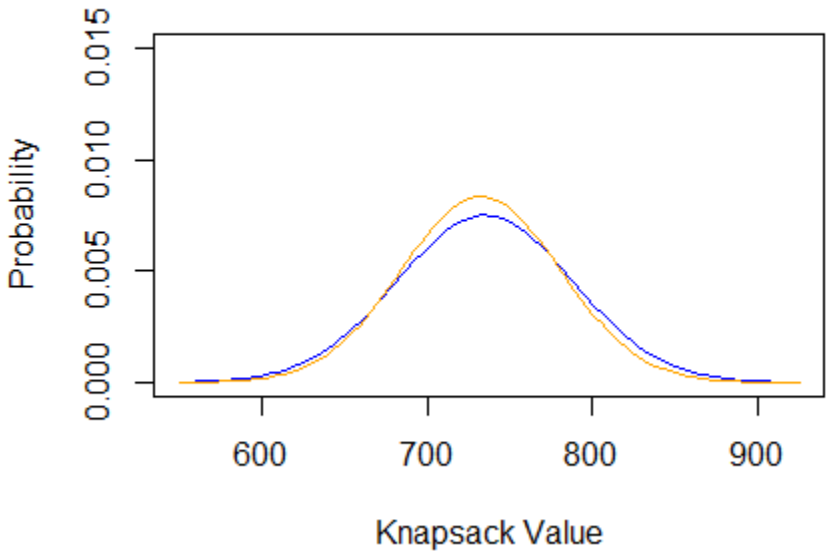


Figure 4.1 is the normal curve for naïve knapsack and the knapsack constrained by the (691;20%) constraint. We can see that by adding a constraint to reduce the risk of getting low yields has an effect on the curve of the knapsack. Here, the mean is decreased, however so is the variance. This observation shows the relationship between reducing the risk of getting low yields and the knapsack value.

Figure 4.2: Normal Curve for Naïve Knapsack and all Risk Scenarios

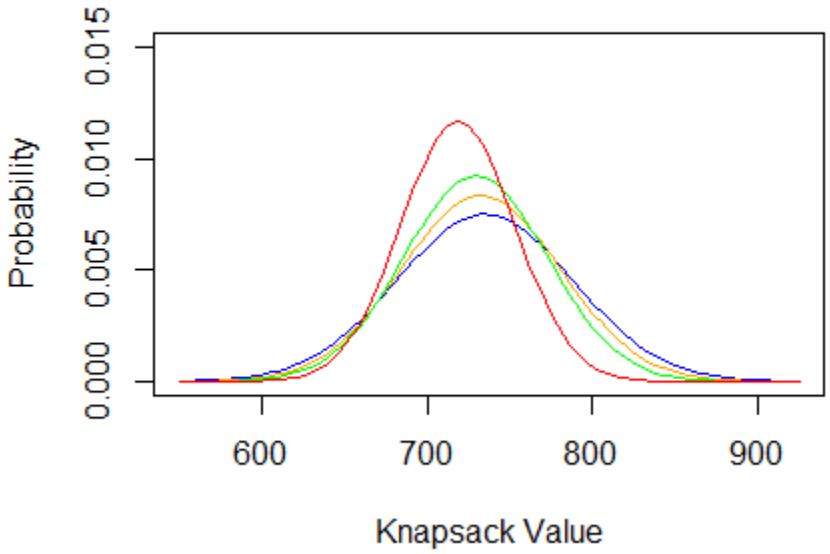


Figure 4.2 shows the curves of all knapsacks in Table 4.1. Here the pattern between reducing the risk and the affects it has on the curve is better shown. Clearly, in exchange for reducing the risk of getting a low yield, total knapsack value must be sacrificed.

4.4 Adverse Environment Cases

4.4.1 Results of Drought Scenario

In the drought scenario, we reduce the means and increase the variances of every variety. Two tables catalog the results of the drought scenario. Table 4.2 summarizes the results of the drought scenario, given there was not a drought. Table 4.3 summarizes the results of the drought scenario, given there was a drought.

Table 4.2: Summary of Results for Results for Optimal Knapsack without Drought

Required Yield	Risk	Varieties in the Selected Knapsack	Average Knapsack Yield	Knapsack Standard Deviation
-	-	{X19, X34, X45, X49, X50, X10, X40, X18, X21, X41}	970	63.03
522	20	{X17, X20, X23, X28, X30, X31, X32, X34, X48, X49}	760	61.87
490	10	{X17, X20, X23, X28, X30, X31, X32, X34, X48, X49}	760	61.87
463	5	{X17, X20, X23, X28, X30, X31, X32, X34, X48, X49}	760	61.87
415	1	{X17, X23, X28, X3, X30, X31, X32, X34, X48, X49}	753	59.19

The model for the drought scenario maximizes the total knapsack yield but tries to reduce the risk of selecting a bad knapsack given that there was a drought. Therefore, the knapsack yields in Table 4.2 are much higher than the required yield. The required yield values are chosen based on the scenario that a drought does happen. For example, the second entry in Table 4.2 has a knapsack value of 760 and a standard deviation of 61.87. The constraint that finds this is the (522;20%) constraint, meaning in the event of a drought, the selected varieties in the knapsack will have a knapsack value of 522 or more, more than 80% of the time.

Figure 4.3: Normal Curve for Naïve Knapsack and Risk Scenarios without Drought

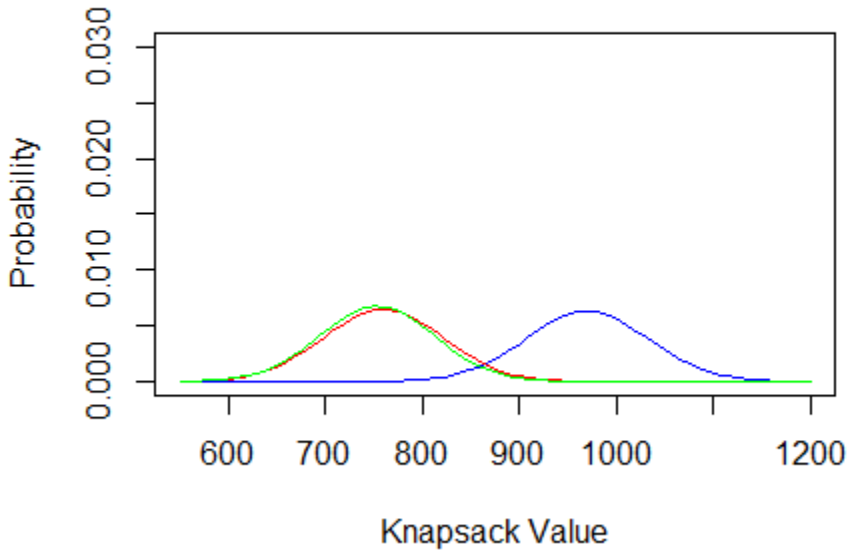


Figure 4.3 shows the normal curves of the entries in Table 4.2. The blue curve is the naïve knapsack, the red is both the (522;20%) and (490;10%) constraints, and the green is the (463;5%) constraint. There is a large discrepancy between the three curves. This is likely due to the constraint using the drought distribution instead of the no drought distribution.

Table 4.3: Summary of Results for Optimal Knapsack with Drought

Required Yield	Risk	Varieties in the Selected Knapsack	Average Knapsack Yield	Knapsack Standard Deviation
-	-	{X17, X23, X30, X34, X28, X32, X48, X20, X49, X31}	585	73.89
522	20	{X17, X20, X23, X28, X30, X31, X32, X34, X48, X49}	585	73.89
490	10	{X17, X20, X23, X28, X30, X31, X32, X34, X48, X49}	585	73.89
463	5	{X17, X20, X23, X28, X30, X31, X32, X34, X48, X49}	585	73.89
415	1	{X17, X23, X28, X3, X30, X31, X32, X34, X48, X49}	578	69.54

However, if there is not a drought, it will have a knapsack value of 760 and a standard deviation of 61.87. To see the knapsack yield of this set of varieties if there is a drought, we can look at Table 4.3. Here, there is still the same constraint (522;20%), but now the knapsack value is 585 and the standard deviation is 73.89. This means that if a drought occurs, the value of the knapsack drops by 175 bushels but is still higher than the required yield.

Figure 4.4: Normal Curve for Naïve Knapsack and Risk Scenarios with Drought

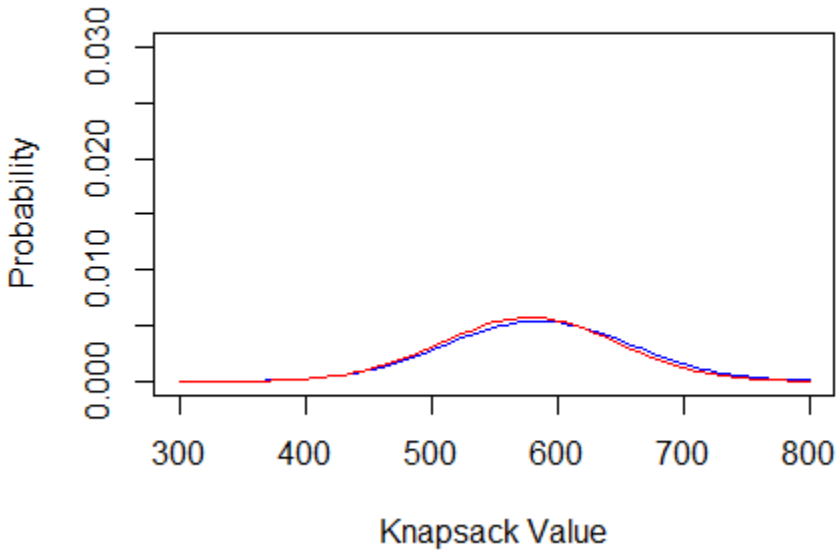


Figure 4.4 shows the normal curves for the entries in Table 4.3. The blue curve represents both the naïve knapsack and the (522;20%), (490;10%), (463;5%) constraints. The red curve is the (415;1%) constraint. These two curves are closer together in comparison to Figure 4.3. This is likely because the constraints use the drought distribution and, in this case, so does the objective function.

Table 4.4: Changes in the Naïve and Risk Knapsack with Drought

	Varieties	No Drought Yield	Drought Yield
1	{X19, X34, X45, X49, X50, X10, X40, X18, X21, X41}	970	465
2	{X17, X20, X23, X28, X30, X31, X32, X34, X48, X49}	760	585
3	{X17, X23, X30, X34, X28, X32, X48, X20, X49, X31}	760	585

Table 4.4 shows how the naïve and risk knapsack change when there is and is not a drought and it's a good example of how neglecting to account for uncertainty will result in risky knapsacks. Here we can see that the varieties that come up with the naïve knapsack when optimizing for when there is not a drought, labeled here as (1), suffers when there is a drought. This knapsack loses almost half of its yield, and in comparison, the knapsack (2) which hedges against drought is larger by 120 bushels. Interestingly, the hedging knapsack is the optimal knapsack when trying to maximize the yield in the event of a drought (3). This makes sense, because the model will choose to select the varieties that perform best during a drought.

Figure 4.5: Normal Curves for Changes in the Drought/No Drought Naïve and (522; 20%) Risk Constraint Knapsack

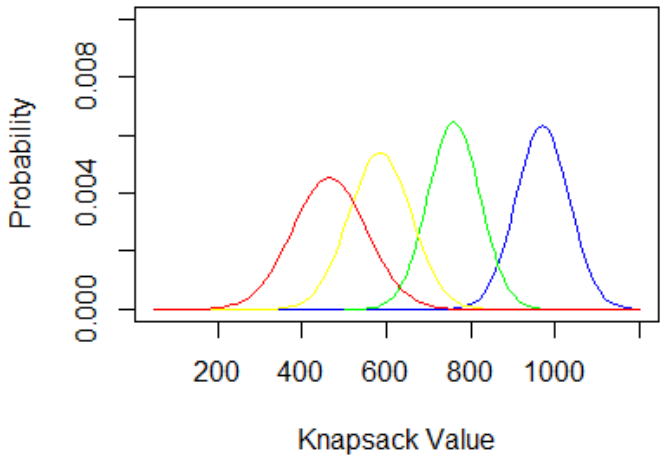


Figure 4.5 further shows the relationship described in Table 4.4. Here the blue curve is the naïve knapsack created using the values from the baseline scenario. The baseline scenario assumes that there will be no drought and so it does not hedge against drought. The red curve shows the performance of the varieties selected from the baseline if a drought does occur. The blue and red curves together show how a naïve knapsack can be very risky. The breeder who selects this knapsack make themselves very vulnerable to the adverse effects of drought, since in the event of a drought the breeder will lose a third of their total yield. The green curve shows the results of adding the (522;20%) risk constraint to the model using the baseline means. The yellow curve shows what happens to this knapsack if a drought occurs. Here we can see that the green curve has a smaller mean than the blue curve, meaning that because we hedged against the drought, we will lose some yield in the event a drought does not occur. However, if a drought does occur, this knapsack has a yield of 585 bushels, which is higher than the red curves yield of 465 bushels. So, if there is a

drought does occur, the breeder who selected the hedging knapsack will have 120 more bushels. On the other hand, if a drought does not occur than the person who selected the hedging knapsack will lose 210 bushels in yield.

The results of the other entries in Table 4.3 are what we expect to see. As the percent of success increases, the required yield, knapsack value and standard deviation are all decreased. Again, showing that the model will give up some knapsack value to reduce the total standard deviation. In the simplest case the deciding factor for selecting the risk reduction knapsack is the probability of drought and how the breeder can mitigate the potential loss due to drought versus the opportunity loss. In a more complex case, this creates an interesting economic opportunity. If the drought reduces the yields of soybeans across many farms, the reduction in supply could increase the selling price. Therefore, there could be additional benefits to hedging against the drought in such a market situation.

4.4.2 Results of Rainy Scenario

In the rainy scenario, each variety can respond in four different ways. Each has an equal chance of increasing or decreasing both the variety's mean and variance. Table 4.5 and 4.6 summarizes the results of the rain scenario. In this scenario, the presence of extra rain has the chance to increases the mean yields of some varieties and as a result, the knapsack values where rain does occur is much larger, see Figure 4.6. The results are calculated in a similar way to the drought scenario. Here the model tries to maximize the knapsack value when there is not any rain but meet a required yield when there is rain.

Table 4.5: Summary of Results for Optimal Knapsack without Rain

Required Yield	Risk	Varieties in the Selected Knapsack	Average Knapsack Yield	Knapsack Standard Deviation
-	-	{X19, X34, X45, X49, X50, X10, X40, X18, X21, X41}	970	63.03
1399	20	{X3, X9, X14, X15, X19, X30, X31, X36, X38, X40}	750	51.79
1377	10	{X3, X9, X14, X15, X19, X30, X31, X36, X38, X40}	750	51.79
1360	5	{X3, X9, X14, X15, X19, X30, X31, X36, X38, X40}	750	51.79
1328	1	{X3, X9, X15, X19, X30, X31, X36, X38, X40, X44}	749	55.98

Figure 4.6: Normal Curve for Naïve Knapsack and Risk Scenarios without Rain

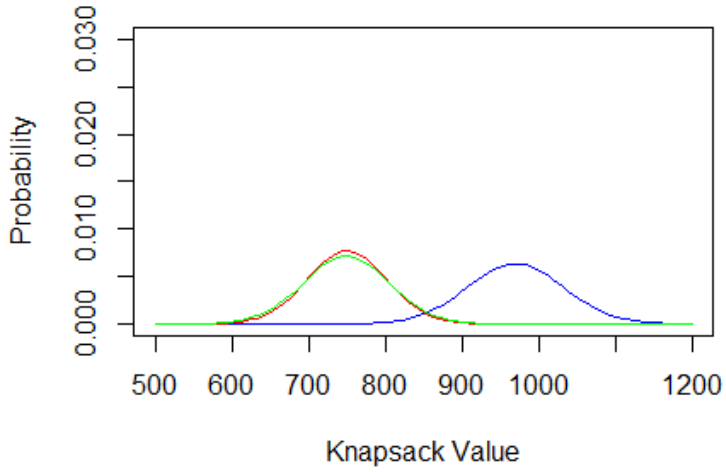


Figure 4.6 shows the normal curves of the entries in Table 4.5. The blue curve is the naïve knapsack, the red is the (1399;20%), (1377;10%), (1360;10%) constraints, and the green is the (1328;1%) constraint.

Table 4.6: Summary of Results for Optimal Knapsack with Rain

Required Yield	Risk	Varieties in the Selected Knapsack	Average Knapsack Yield	Knapsack Standard Deviation
-	-	{X40, X38, X9, X15, X36, X3, X31, X26, X14, X19}	1445	57.68
1399	20	{X3, X9, X14, X15, X19, X30, X31, X36, X38, X40}	1441	49.29
1377	10	{X3, X9, X14, X15, X19, X30, X31, X36, X38, X40}	1441	49.29
1360	5	{X3, X9, X14, X15, X19, X30, X31, X36, X38, X40}	1441	49.29
1328	1	{X3, X9, X15, X19, X30, X31, X36, X38, X40, X44}	1434	45.24

The required yields for the rain scenario are much larger when compared to the required yields of the drought scenario. This is because, as mentioned before, some varieties will have their yields increased when there is an increase in rain. For example, in entry one, the (1399;20%) constraint results in a knapsack with a knapsack value of 750 and a standard deviation of 51.79, when there is no extra rain, but a knapsack value of 1441 when there is extra rain. The relationship between the rain and no rain knapsacks is different from that of the drought no drought knapsack. Because of increase in mean yields, rain scenario can be thought of like the required opportunity knapsack. Therefore, the constraint can be thought of as securing the opportunity of getting more than 1399 knapsack value, at least 80% of the time.

Figure 4.7: Normal Curve for Naïve Knapsack and Risk Scenarios with Rain

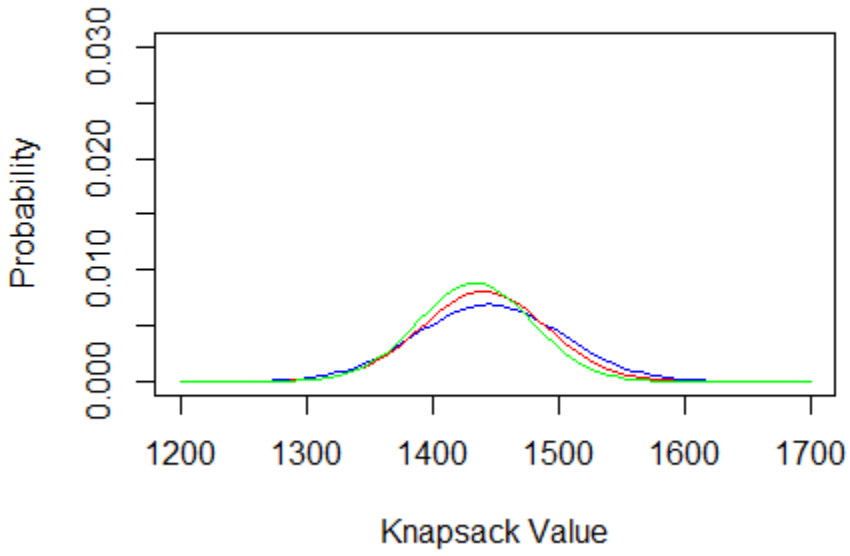
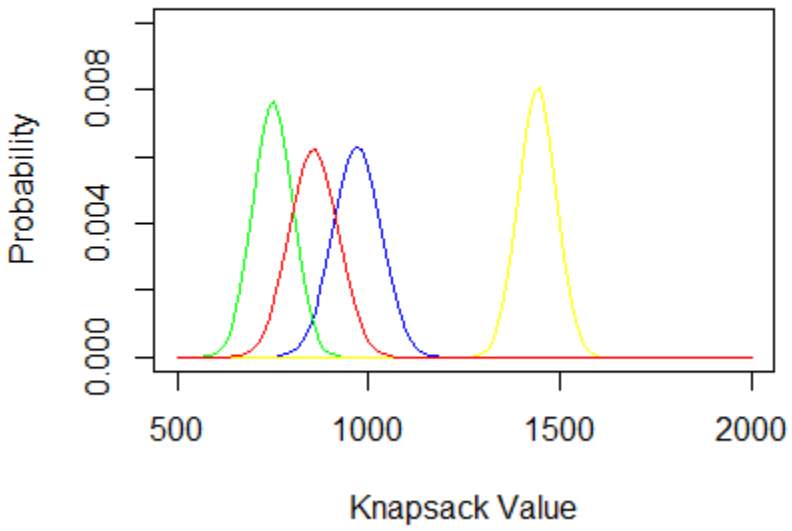


Figure 4.7 shows the normal curves for the entries in Table 4.6. The blue curve represents both the naïve knapsack, the red is the (1399;20%), (1377;10%), (1360;10%) constraints, and the green is the (1328;1%) constraint.

Figure 4.8: Normal Curves for Rain/No Rain Naïve and 1399,20% Risk Constraint Knapsack



Also, because of the different relationship, the graph for the naive knapsacks and the risk constraint looks different. In Figure 4.8, the blue curve is the naïve knapsack using the baseline (assuming no drought) means. The green curve is how this knapsack reacts to the rain. Here we can see that the mean of the knapsack goes down by 220 bushels. The red curve is the hedged knapsack using the baseline means and the yellow curve is how this knapsack reacts when there is rain. If there is not rain then the hedged knapsack does not perform as well as the naïve knapsack, however when there is rain, it greatly outperforms the blue curve. This is because the hedged knapsack can take advantage of the varieties which have an increase in yield due to the rain.

Table 4.7: Changes in the Naïve and Risk Knapsack with Rain

	Varieties	No Rain Yield	Rain Yield
1	{X19, X34, X45, X49, X50, X10, X40, X18, X21, X41}	970	857
2	{X3, X9, X14, X15, X19, X30, X31, X36, X38, X40}	750	1441
3	{X40, X38, X9, X15, X36, X3, X31, X26, X14, X19}	743	1445

Table 4.7 shows how the naïve and risk knapsacks change based on the environment. Here again we see that the knapsack optimized for no rain suffers when there is rain. In this case, the knapsack is not able to take advantage of the extra growth available due to the rain. In this case the difference is even more than the in the drought case. Here if a person chooses the hedged knapsack (2) and it does not rain, they will lose 220 bushels of yield. However, if they select the hedged knapsack and it does rain, they will gain 584 bushels of yield. The benefit of this hedged knapsack is that it is able to take advantage of the varieties which may have had an increased yield due to the rain.

4.4.3 Results of Extreme Temperature Scenario

The extreme temperature scenario is like rainy scenario in that each variety's mean and variance has an equal chance of increasing and/or decreasing.

Table 4.8: Summary of Results for Optimal Knapsack without Extreme Temperature

Required Yield	Risk	Varieties in the Selected Knapsack	Average Knapsack Yield	Knapsack Standard Deviation
-	-	{X19, X34, X45, X49, X50, X10, X40, X18, X21, X41}	970	63.03
1340	20	{X1, X3, X7, X11, X15, X17, X33, X34, X37, X38}	689	57.10
1316	10	{X1, X3, X7, X11, X15, X17, X33, X34, X37, X38}	689	57.10
1297	5	{X1, X3, X7, X11, X14, X15, X33, X34, X37, X38}	668	53.05
1266	1	{X1, X3, X7, X11, X14, X15, X33, X34, X37, X38}	668	53.05

Table 4.9: Summary of Results for Optimal Knapsack with Extreme Temperature

Required Yield	Risk	Varieties in the Selected Knapsack	Average Knapsack Yield	Knapsack Standard Deviation
-	-	{X38, X3, X33, X15, X1, X7, X37, X17, X11, X34}	1386	54.55
1340	20	{X38, X3, X33, X15, X1, X7, X37, X17, X11, X34}	1386	54.55
1316	10	{X38, X3, X33, X15, X1, X7, X37, X17, X11, X34}	1386	54.55
1297	5	{X1, X3, X7, X11, X14, X15, X33, X34, X37, X38}	1372	45.21
1266	1	{X1, X3, X7, X11, X14, X15, X33, X34, X37, X38}	1372	45.21

Tables 4.8 and 4.9 summarize the results of the extreme temperature scenario. Like the rain scenario, this scenario also has the possibility to increase the mean yield of a variety. This is also why the required yield values are so much higher than the drought scenario. The results are also like that of the rain scenario. They are not identical because each variety has a different probability of reacting a certain way to each adverse environmental scenario.

Figure 4.9: Normal Curve for Naïve Knapsack and Risk Scenarios without Extreme Temperature

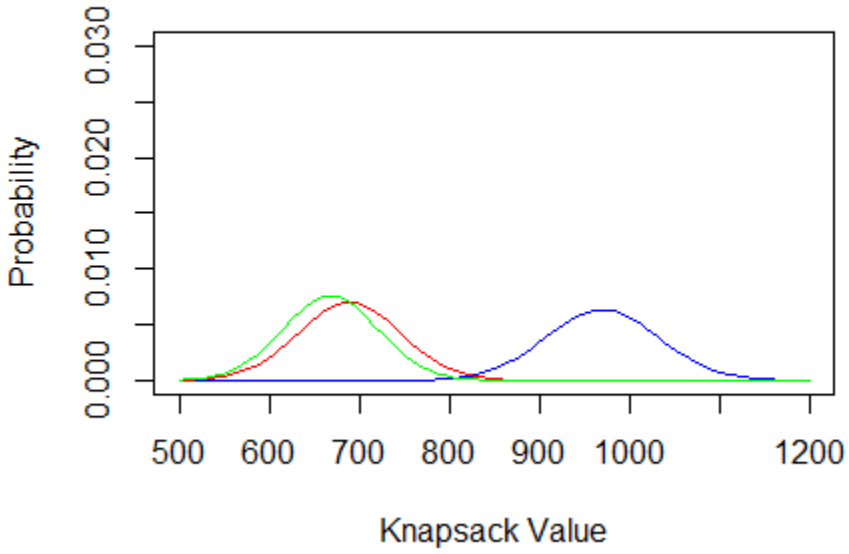


Figure 4.9 shows the normal curves of the entries in Table 4.8. The blue curve is the naïve knapsack, the red is the (1340;20%), (1316;10%) constraints, and the green is the (1297;5%), (1266;1%) constraints.

Figure 4.10: Normal Curve for Naïve Knapsack and Risk Scenarios with Extreme Temperature

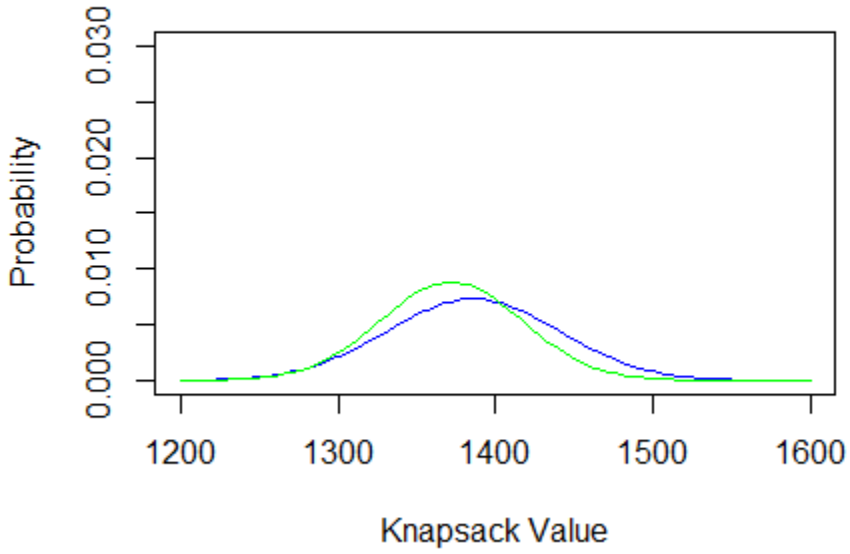
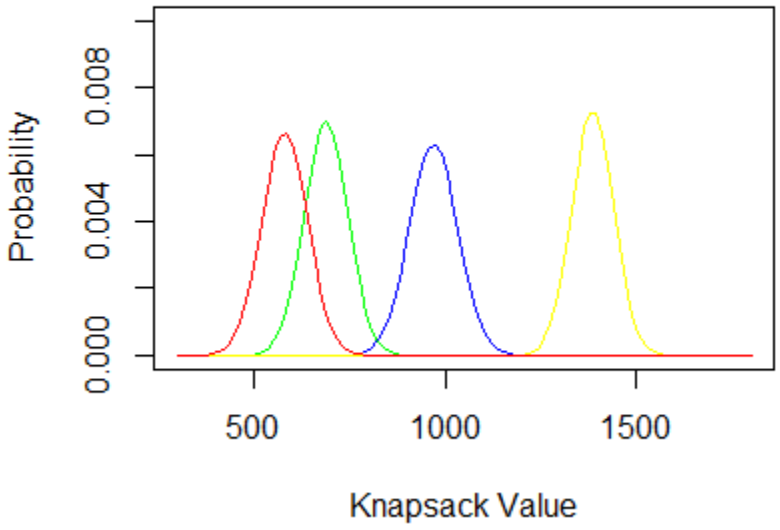


Figure 4.10 shows the normal curves of the entries in Table 4.9. The blue curve is the naïve knapsack, and the (1340;20%), (1316;10%) constraints, the green is the (1297;5%), (1266;1%) constraints.

Figure 4.11: Normal Curve for Naïve and 1340,20% Risk Constraint Knapsack



In Table 4.8, we can see that both the required yield and the knapsack values are lower than those in the rain scenario, again this likely because of the different response probabilities. Here the blue curve is the naïve knapsack using the baseline means, and the red curves shows how the knapsack reacts to the extreme temperature. The green curve is the hedged knapsack using the baseline means and the yellow curve shows that variety responds to the extreme temperature.

Table 4.10: Changes in the Naïve and Risk Knapsack with Extreme Temperature

	Varieties	No Extreme Temperature	Extreme Temperature
1	{X19, X34, X45, X49, X50, X10, X40, X18, X21, X41}	970	581
2	{X1, X3, X7, X11, X15, X17, X33, X34, X37, X38}	689	1386
3	{X1, X3, X7, X11, X15, X17, X33, X34, X37, X38}	689	1386

Table 4.10 shows the changes in the naïve and risk knapsacks for extreme temperature. This is similar to the rain scenario, where the hedged knapsack takes advantage of the fact that some varieties improve due to the weather condition. The implications are like the rain scenario.

4.5 SAA Method

In cases where the true distribution is unknown, or too complex to calculate correctly, sampling can be substituted for the distribution. In the SAA method, the accuracy of the answer depends on the number of samples available, with a larger number of samples providing a more accurate solution. This is because a larger sample size will more closely represent the underlying distribution. Table 4.11 shows the difference in values for solutions found with a range of sample sizes. Here the risk constraint is the (691;20%) constraint and we are using the same data used in the baseline case.

Table 4.11: SAA Method

Sample Size	**Knapsack Value**
Exact Formulation	732.47
1000	729.34
2000	729.34
3000	732.47

Here we can see that as the number of samples increases, the solution to the model gets closer to the exact formulation. In general, it is best to have as many samples as possible, but we have found that for this research about 3000 samples is a good approximation for the true distribution.

4.6 Conclusions

In summary the purpose of this work is to establish a tool that breeders can use to make better decisions when it comes to seed progression. This tool is a chance-constrained knapsack optimization model, which selects the N varieties which maximize the average yield of the knapsack. This model is constrained by yield requirements and a minimum chance of success. The mean and the variance of the yields of the seed varieties are assumed to be normal and are used to create the chance-constraint. However, this method assumes that the breeder has very good knowledge of the variety's distribution, mean and variance. Since, this isn't always the case, the breeders can use sampling to replace the need of calculating a probability distribution.

This thesis shows the risks of selecting knapsacks based only on the mean yields of the varieties. This method can result in reduced yields due to factors such as weather conditions, gene variability and soil type. By creating a risk reduction knapsack, we can select varieties that ensure a higher mean yield when uncertain events do occur. This research shows that by considering uncertainty, we select varieties which provide a smaller total yield, but have a smaller total variance. This means that even when the varieties perform poorly, they will still meet certain minimum requirements.

Additionally, we see that when trying to hedge against poor weather conditions, we select varieties which perform well both when there is or is not a drought. The total yield for this knapsack is smaller than that of the naïve knapsack, but if there is a drought, then this knapsack performs much better than the naive one. Similarly, the knapsacks for the rain and extreme temperature case result in a combination of varieties which perform well whether the poor weather condition occurs.

This is a novel contribution which can be used as a tool for breeders to make better decisions during seed progression. This will save time and money in seed trials and allow the breeders to provide higher yielding seeds to the market. Currently, breeders risk progressing dud seeds far into the 3-year span of testing seeds. Using this tool, breeders can lower the chance of progressing bad seeds.

In this work we used data pulled from a normal distribution and showed that given normally distributed data it is possible to create a chance-constrained optimization model. In reality, it is not always possible to have normally distributed data and sampling should be used to approximate the distribution of the data. Yield data could be affected by different factors like weather, soil, past yields, genetic traits. To create a more accurate data distributions data from these different places could be pooled together and assimilated into a single distribution.

References

1. "8 Ways to Fix the Global Food Crisis." *U.S. News & World Report*, U.S. News & World Report, www.usnews.com/news/articles/2008/05/09/8-ways-to-fix-the-global-food-crisis.

2. Ahmed, Shabbir, and Alexander Shapiro. "Solving Chance-Constrained Stochastic Programs via Sampling and Integer Programming." *State-of-the-Art Decision-Making Tools in the Information-Intensive Age*, 2008, pp. 261–269., doi:10.1287/educ.1080.0048.

3. Ben-Tal, Aharon, et al. *Robust Optimization*. Princeton University Press, 2009.

4. Beukelaer, Herman De, et al. "Heuristic Exploitation of Genetic Structure in Marker-Assisted Gene Pyramiding Problems." *BMC Genetics*, vol. 16, no. 1, 2015, p. 2., doi:10.1186/s12863-014-0154-z.

5. Boles, James N. "Linear Programming and Farm Management Analysis." *Journal of Farm Economics*, vol. 37, no. 1, 1955, p. 1., doi:10.2307/1234071.

6. Byrum, Joseph, et al. "Advanced Analytics for Agricultural Product Development." *Interfaces*, vol. 46, 2016, pp. 5–17.

7. Byrum, Joseph, et al. "Genetic Gain Performance Metric Accelerates Agricultural Productivity." *Interfaces*, vol. 47, no. 5, 2017, pp. 442–453., doi:10.1287/inte.2017.0909.

8. Canzar, Stefan, and Mohammed El-Kebir. "A Mathematical Programming Approach to Marker-Assisted Gene Pyramiding." *Lecture Notes in Computer Science Algorithms in Bioinformatics*, 2011, pp. 26–38., doi:10.1007/978-3-642-23038-7_3.

9. "Chance-Constraint Method." *Optimization*, optimization.mccormick.northwestern.edu/index.php/Chance-constraint_method.

10. Charnes, A., and W. W. Cooper. "Chance-Constrained Programming." *Management Science*, vol. 6, no. 1, 1959, pp. 73–79., doi:10.1287/mnsc.6.1.73.

11. Ehrlich, Paul R. "Analysis: Pessimism on the Food Front." *EHN*, EHN, 26 Apr. 2018, www.ehn.org/what-are-the-threats-to-future-food-security-2562981347.html.

12. Gelute, Abebe. "Chance constrained optimization - applications, properties and numerical issues" LIlmenau University of Technology, May 31, 2012

13. Goulding, Keith, et al. “Optimizing Nutrient Management for Farm Systems.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, The Royal Society, 12 Feb. 2008, www.ncbi.nlm.nih.gov/pmc/articles/PMC2610177/.

14. Henrion, R.; Strugarek, C. Convexity of chance constraints with independent random variables. Comput. Optim. Applic. 41, 262-276, 2008.

15. Henrion, René. *Introduction to Chance Constrained Optimization*. 2004

16. Hincks, Joseph. “We Are Headed for a World Food Crisis. Here's How to Stop It.” *Time*, Time, 28 Mar. 2018, time.com/5216532/global-food-security-richard-deverell/.

17. Kopcso, David, et al. “Artificial Neural Networks as Alternatives to Statistical Quality Control Charts in Manufacturing Processes.” *Computational Statistics*, 1992, pp. 309–315., doi:10.1007/978-3-642-48678-4_39.

18. Mccorkle, Chester O. “Linear Programming as a Tool in Farm Management Analysis.” *Journal of Farm Economics*, vol. 37, no. 5, 1955, p. 1222., doi:10.2307/1234021.

19. Pedersen, Palle. *Variety Selection*. crops.extension.iastate.edu/files/article/VarietySelection_000.pdf.

20. Pimentel, David. “World Overpopulation.” *SpringerLink*, Springer Netherlands, 3 Jan. 2012, link.springer.com/article/10.1007/s10668-011-9336-2.

21. Pınar, Mustafa. *Robust Optimization*. Sept. 2005, www.ie.bilkent.edu.tr/~mustafap/courses/rt4.pdf.

22. Plaxico, James S. “[Simplified Presentation and Logical Aspects of Linear Programming Technique]: Discussion.” *Journal of Farm Economics*, vol. 36, no. 5, 1954, p. 1048., doi:10.2307/1234314.

23. Sadik, Nafis. “Population Growth and the Food Crisis.” *Population Growth and the Food Crisis*, www.fao.org/3/U3550t/u3550t02.htm.

24. “Seed-Selection Resources to Help Farmers Maximize Yield Potential, Profits.” *United Soybean Board*, unitedsoybean.org/article/seed-selection-resources-to-help-farmers-maximize-yield-potential-profits.

25. “Selecting Seed Varieties.” *The Miracle Bean*, www.themiraclebean.com/selecting-seed-varieties/.

26. Uryasev, S. Derivatives of probability functions and some applications. Ann Oper Res, 56, 287{311, 1995}

27. “Variety Selection.” *Variety Selection - IRRI Rice Knowledge Bank*, www.knowledgebank.irri.org/training/fact-sheets/crop-establishment/item/variety-selection-fact-sheet.

28. “Variety Selection.” *Washington Crop*, washingtoncrop.com/seed/variety-selection.

29. “World Population Projected to Reach 9.8 Billion in 2050, and 11.2 Billion in 2100 | UN DESA Department of Economic and Social Affairs.” *United Nations*, United Nations, www.un.org/development/desa/en/news/population/world-population-prospects-2017.html.

APPENDIX A

BASELINE DATA SET

Variety	Mean	Variance
X1	59.95	120.52
X2	41.81	158.45
X3	55.10	73.44
X4	57.36	275.64
X5	51.03	245.45
X6	68.84	30.79
X7	71.39	68.64
X8	60.06	37.17
X9	48.12	729.44
X10	68.93	115.99
X11	40.48	745.22
X12	71.21	568.47
X13	64.61	526.54
X14	62.76	110.07
X15	71.76	25.68
X16	47.72	214.78
X17	62.59	130.38
X18	53.00	63.56
X19	65.75	275.79
X20	73.18	31.27
X21	55.49	657.44
X22	61.61	613.14
X23	66.38	514.91
X24	77.50	786.65
X25	66.59	83.68
X26	57.34	28.79
X27	58.50	670.47
X28	71.42	461.13
X29	79.87	146.08
X30	43.17	480.85
X31	78.00	321.96
X32	54.49	42.52
X33	41.16	656.93
X34	42.74	847.96
X35	59.07	540.15

Variety	Mean	Variance
X36	64.37	312.94
X37	43.42	96.60
X38	56.95	743.69
X39	71.58	307.59
X40	62.97	549.18
X41	63.32	59.68
X42	49.97	37.08
X43	62.10	26.82
X44	40.33	127.47
X45	54.24	187.15
X46	60.27	816.93
X47	48.44	282.85
X48	68.29	38.62
X49	40.82	670.32
X50	41.17	41.41
X50	41.17	41.41

APPENDEX B

ADVERSE ENVIROMENT BASELINE DATA SET

Variety	AverageYield1	Variance1
X1	42	326
X2	65	542
X3	57	266
X4	70	122
X5	84	434
X6	76	333
X7	83	141
X8	65	363
X9	91	199
X10	97	491
X11	61	98
X12	49	119
X13	51	364
X14	54	100
X15	72	242
X16	63	227
X17	75	546
X18	93	585
X19	100	579
X20	64	590
X21	93	118
X22	43	371
X23	59	261
X24	79	140
X25	82	624
X26	63	344
X27	53	450
X28	60	566
X29	49	92
X30	70	232
X31	73	165
X32	84	434
X33	64	524
X34	100	487
X35	74	422
X36	77	130
X37	75	408
X38	60	223

Variety	Average Yield1	Variance
X39	87	389
X40	X41	92
X41	X42	56
X42	X43	64
X43	X44	53
X44	X45	100
X45	X46	52
X46	X47	64
X47	X48	75
X48	X49	100
X49	X50	99
X50	99	161

APPENDEX C

ADVERSE ENVIROMENT DROUGHT DATA SET

Variety	AverageYield2	Variance2
X1	38	627
X2	32	321
X3	50	299
X4	46	410
X5	30	641
X6	23	579
X7	53	932
X8	34	951
X9	43	604
X10	34	572
X11	26	535
X12	51	668
X13	20	696
X14	40	955
X15	25	53
X16	24	610
X17	60	145
X18	33	810
X19	45	959
X20	57	922
X21	52	625
X22	43	89
X23	60	641
X24	26	610
X25	21	514
X26	53	979
X27	34	762
X28	59	896
X29	21	816
X30	60	366
X31	54	280
X32	59	349
X33	44	519
X34	60	951
X35	34	668
X36	20	691
X37	46	608
X38	38	861

Variety	Average Yield 2	Variance 2
X39	21	346
X40	52	748
X41	37	475
X42	28	446
X43	20	664
X44	24	862
X45	54	916
X46	27	451
X47	24	123
X48	59	151
X49	57	759
X50	41	930

APPENDEX D

ADVERSE ENVIROMENT RAIN DATA SET

Variety	AverageYield2	Variance2
X1	38	627
X2	122	150
X3	143	340
X4	26	124
X5	30	641
X6	23	579
X7	53	932
X8	92	516
X9	150	221
X10	114	203
X11	121	647
X12	90	774
X13	86	75
X14	133	510
X15	149	81
X16	24	610
X17	43	44
X18	52	117
X19	132	145
X20	41	180
X21	52	625
X22	32	85
X23	53	148
X24	83	37
X25	37	150
X26	135	910
X27	44	215
X28	118	65
X29	94	135
X30	131	12
X31	143	90
X32	97	604
X33	27	169
X34	107	864
X35	109	298
X36	149	730
X37	46	608

Variety	AverageYield2	Variance2
X38	154	138
X39	86	142
X40	157	163
X41	37	475
X42	28	446
X43	81	111
X44	126	127
X45	84	217
X46	27	451
X47	99	851
X48	88	616
X49	81	331
X50	41	930

APPENDEX E

ADVERSE ENVIROMENT EXTREME TEMPERATURE DATA SET

Variety	AverageYield2	Variance2
X1	145	527
X2	32	321
X3	153	147
X4	46	410
X5	30	641
X6	46	124
X7	141	89
X8	34	951
X9	49	138
X10	37	182
X11	121	647
X12	90	774
X13	22	153
X14	110	19
X15	149	81
X16	38	126
X17	124	951
X18	33	810
X19	38	118
X20	57	922
X21	52	625
X22	43	89
X23	60	641
X24	26	610
X25	37	150
X26	38	179
X27	34	762
X28	59	896
X29	88	779
X30	104	470
X31	54	280
X32	22	174
X33	153	122
X34	116	83
X35	34	668
X36	20	691
X37	130	191

Variety	AverageYield2	Variance2
X38	154	138
X39	97	741
X40	52	748
X41	37	475
X42	34	59
X43	20	664
X44	24	862
X45	84	217
X46	26	55
X47	95	60
X48	21	213
X49	81	331
X50	51	125