REGRESSION ANALYSIS OF UAV COLLECTED COTTON CROP DATA FOR YIELD PREDICTION

A Thesis

by

BIANCA BRIANNE LOPEZ

BS, Texas A &M University-Corpus Christi, 2019

Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in

MATHEMATICS

Texas A&M University-Corpus Christi
Corpus Christi, Texas

May 2022

May 2022

REGRESSION ANALYSIS OF UAV COLLECTED COTTON CROP DATA FOR YIELD PREDICTION

A Thesis

by

BIANCA BRIANNE LOPEZ

This thesis meets the standards for scope and quality of
Texas A&M University-Corpus Christi and is hereby approved.

Alexey Sadovski, PhD
Chair

Beate Zimmer, PhD
Committee Member

Nothabo Dube Davis, PhD
Committee Member

May 2022

ABSTRACT

Prediction of cotton yield can enable farmers to make more beneficial planning, budgeting, and intervention decisions. The objective of this thesis was to assess the performance of principal component regression (PCR), partial least squares regression (PLSR), Ridge regression, and least absolute shrinkage and selection operator (LASSO) regression for predicting cotton yield. During the 2016 growing season, excess greenness index (ExG), normalized difference vegetation index (NDVI), canopy height (CH), and canopy volume (CV) were calculated weekly from UAS (unmanned aerial systems) collected RGB (red, green, blue) and multispectral images of an experimental cotton field located at the Texas A&M AgriLife Research Center in Corpus Christi, Texas, USA (27° 46' 57.08" N, 97° 33' 40.94" W). Irrigation was taken as a categorical variable, with the field split into two approximately equal sections of dry and irrigated plots. Data were split into 80 percent training data and 20 percent testing data for all models and a 10-fold cross validation was performed to find the optimal number of principal components for the PCR, latent variables for PLSR, and the hyperparameters of the LASSO and Ridge regressions. All models were trained with the weekly time series variables ExG, NDVI, CH, and CV and the categorical variable irrigation. Each model was also trained with the same time series variables up to 67 days after planting and irrigation. The set of models trained on the entire season resulted in the following test set mean squared error values and R-squared scores, respectively; PCR with $\sim$2.83 and $\sim$0.48, PLSR with $\sim$1.00 and $\sim$0.80, LASSO regression with $\sim$0.94 and $\sim$0.81, and Ridge regression with $\sim$1.32 and $\sim$0.73. The models trained on the first 67 days subset obtained the following mean squared error values and R-squared scores, respectively; PCR with $\sim$2.88 and $\sim$0.47, PLSR with $\sim$1.54 and $\sim$0.70, LASSO regression with $\sim$1.60 and $\sim$0.67, and Ridge regression with $\sim$1.61 and $\sim$0.67. LASSO regression fit best out of the four regressions used to model the entire season's data with the highest R-squared value and lowest MSE score. This model could be useful for decision-making in preparation for future growing seasons. The PLSR model trained on the first 67 days after planting subset resulted in

the lowest MSE and the highest R-squared of $\sim$0.70. Decisions for intervention could be made with reasonable accuracy at 67 days after planting based on the PLSR model.

DEDICATION

This thesis is dedicated to the loving memory of my abuela, Natalia Garcia Acevedo.

ACKNOWLEDGEMENTS

I would like to sincerely thank my committee chair, Dr. Alexey Sadovski, for recognizing my strengths, providing me with many valuable opportunities, and helping me to achieve my academic goals.

I am also thankful to Dr. Beate Zimmer for her guidance and help during my writing and editing processes and for being an excellent professor throughout my time in the graduate program. I greatly appreciate all the knowledge I gained from her rigorous and interesting courses.

I would also like to thank Dr. Nothabo Dube Davis for allowing me to be involved in this research, for her encouragement, and all that I learned from her during my time as a research assistant.

I am thankful to my mom, dad, family, and friends who supported me while I completed my degree. I also acknowledge the excellent mathematics and computer science professors I learned from throughout my academic career. Lastly, I'd like to thank Alyssa Good for providing me with the tools and support I needed to complete my degree and confidently transition into the next chapters of my career.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

Page

LIST OF TABLES

Page

CHAPTER I: INTRODUCTION

Cotton (*Gossypium hirsutum L.*) is one of the most grown fiber crops in the world [32]. In the United States, Texas grows approximately 40% of the country's cotton, harvesting approximately 6 million bales of cotton in the year 2020 [22]. Predicting cotton yield can allow farmers to make more informed decisions when planning for future seasons or choosing to intervene during the growing season. Collecting crop data can be difficult due to the long collection period for each growing season, field areas being too large to sample accurately by hand, and obscured satellite images leading to missing or inaccurate data. The availability and affordability of Unmanned Aerial Systems (UAS) have provided an alternative to data collection from satellite imagery [4]. This thesis uses data computed from red, green, and blue (RGB) and multispectral sensor-equipped UAS imagery to build and compare four cotton yield prediction models that solve the problem of multicollinearity among data variables. The regression models assessed were LASSO, Ridge, principal component, and partial least squares regression. In the second chapter of this thesis, we review the biological process of cotton growth and the mathematical concepts used to build our suggested models. These concepts include a review of principal component analysis and the mathematical background of the four regression models assessed. In Chapter III, we describe the location and climate of the experimental cotton field and the process of computing the predictor variables from UAS images. Chapter IV explores the data set, describes the model assessment process and selects the best model for our data. Finally, a discussion of this thesis and possible future research is given in Chapter V.

CHAPTER II: REVIEW OF THE LITERATURE

2.1 Introduction

In this chapter, we lay the groundwork for our problem statement by providing pertinent information about the areas of study relevant to the creation of our model. We include information on related work, cotton cultivation, principal component analysis, principal component regression, multiple linear regression, Ridge regression, LASSO regression, and partial least squares regression.

2.2 Related Work

Farmers have always sought to increase crop yield. Analysis of crop growth behavior and prediction of yield can equip farmers with valuable information. Recent advancements in remote sensing and machine learning have enabled researchers to make more accurate predictions of crop yield [24], [15]. The development of models based on UAV collected data has become increasingly popular as drones have become more accessible and affordable [16], [19]. In 2019, multi-temporal UAS data were used to compare the effects of tillage and no-tillage management practices on cotton growth and development [4]. In 2020, UAS imagery was used to develop a deep learning-based object detection framework to assess plant population [25]. Multiple studies have used neural networks to make predictions on cotton yield [36], [3], [9]. Neural networks are more efficient with larger data sets which are not always available. This thesis continues research with data calculated from UAV collected imagery used in [3] where researchers developed a machine learning framework for crop yield estimation. The intention of this thesis is to develop a more interpretable model with comparable prediction accuracy on the relatively small data set.

2.3 The Cotton Crop

Cotton (*Gossypium hirsutum L.*) is a perennial plant domesticated to grow as an annual crop; it prefers warmer climates and produces cotton lint and seeds in its mature fruit. Cotton is one of the most popularly grown fiber crops in the world and has been grown for harvest since 6000

B.C. [32]. Cotton has the unique quality of producing both fiber and food as a crop; farmers use cottonseed as livestock feed, and cottonseed oil is present in various foods for human consumption [18]. In the United States, cotton is grown almost entirely in seventeen states that create the "Cotton Belt" this "belt" spans the south of the United States from Virginia to California [22]. From August 2019 to July 2020, this area produced nearly 20 million bales of cotton, valued at about 7 billion dollars in cotton fiber and cottonseed [22]. Texas has been the leading state for cotton production in the United States since the 1880s, currently producing approximately 40 percent of the country's cotton [22][31]. In 2020, Texas harvested approximately 6 million bales of cotton or 2.9 billion pounds of cleaned cotton lint [22]. For perspective, the cotton lint from one bale of cotton has the potential to produce more than 200 pairs of jeans or 1,200 t-shirts [22]. The data for this study were collected in the Coastal Bend region of Texas and, more specifically, Nueces county. Nueces county consists of flat coastal grasslands with an area of 847 square miles and an elevation range from 0 to 180 feet above sea level [17]. This area experiences high humidity with a subtropical climate. When cultivating cotton, an understanding of the anatomy and growth stages of the plant is essential for effective decision-making. In the following sections, we describe the basic anatomy of a cottonseed and the growing stages of the plant from germination to boll maturation.

2.4 Cotton Seed Anatomy and Germination

Each cotton seed has an outer shell called a seed coat; this seedcoat has a pointed, narrow end, called the micropyle, and a flatter wide end, the chalaza cap. The chalaza is the ovule of the seed and will absorb water and oxygen during the germination process. Inside the seed coat is the embryo, which contains the radicle, the hypocotyl, the epicotyl, and the two cotyledons, all shown in Figure 2.1 taken from [18]. We can see the makings of the cotton seedling even before germination, with each part of the embryo becoming a vital part of the seedling emergence process. After planting, water and oxygen enter the chalaza triggering a chain of physical and chemical reactions. First, soil moisture travels through the chalaza into the seed towards the

radicle; the seed swells as water permeates and softens the seed coat and embryo. The seed eventually ruptures, and the radicle pushes through the micropyle and downward into the soil, becoming the root system's first shoot. In the "crook stage", the hypocotyl emerges from the soil first after elongating from the radicle and forming a hook-like structure that pushes out of the soil, becoming the initial stem. Once the seedling is above the surface, the cotyledons unfurl into two initial leaves, turn green, and synthesize chlorophyll for the developing root system. Next, the bud located above the cotyledons, called the apical meristem, emerges and will bear all future vegetation and reproductive growths.

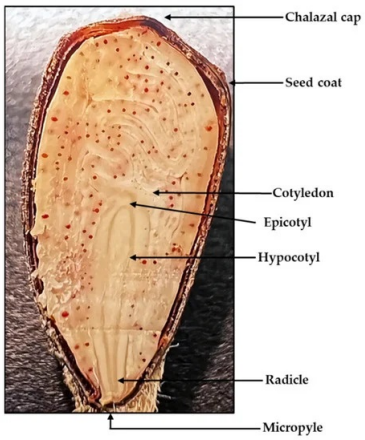


Figure 2.1
Anatomy of a Cotton Seed (Maeda 2021)

Depending on conditions, the seedling emergence process can last anywhere from 4 to 14 days after planting; as emergence time increases, so does the risk of death and lower yield. Cotton seeds germinate well in soil with high oxygen concentration, temperatures at or above 60°F, and adequate moisture [7]. The planting and early development stages are critical for healthy, productive cotton plants. The first 30 to 40 days after planting indicates the future productivity of the crop [6]. If conditions are suitable for germination, the hypocotyl will be visible four to five days after planting. Many factors can inhibit the successful emergence of seedlings and, in turn, diminish the total yield of the crop. Some possible dangers to the germination process include too much or too little moisture present in the soil, soil temperatures below 60°F, or low

oxygen concentration; these risks make the timing of planting very important for early season plant health [6]. Deterling advises that there should be a minimum average soil temperature of 60° at eight inches deep for at least ten days, with temperature measurements being taken daily at 8 in the morning [6]. It is also essential that the seeds be planted at 1 to 2.5 inches deep, depending on soil type and moisture. The crop's success will depend on obtaining a "good stand"; where healthy and evenly distributed seedlings indicate a "good stand". After establishing a stand, there is no way to increase yield, only maintain or decrease it [6]. A poor stand leaves the seedlings vulnerable to disease, stress from the elements, and pests.

2.5 The Organs of the Cotton Plant

After the initial planting and sprouting stages, the cotton plants form four organs that bring the plant to maturity. The four organs of the cotton plant are the roots, stem, leaves, and fruits.

The Root System

The cotton plant's root system consists of the primary taproot, the lateral roots extending from the taproot, and root hairs along with the lateral roots. The root hairs are responsible for collecting most of the plant's soil-based nutrients and hydration. The initial taproot established during germination can grow downward into the soil for several days and reach up to 9 inches of depth before the cotyledons begin to emerge above ground [6]. As the cotyledons emerge and unfurl, the taproot develops lateral roots and eventually root hairs. Most roots will be confined to the first 1 to 3 feet of soil but may penetrate further than this. The primary root system will be established around 8 to 10 weeks after planting [6]. The size, depth, and shape of the root system depend on the compaction of soil, soil moisture, and weather conditions. The roots will continue to grow through the flowering stage of the plant, then begin to decline as carbohydrates are redirected to the maturing bolls.

Development and Utility of the Main Stem

The stem of the cotton plant consists of nodes and internodes and supports the vegetation and fruit. The first true leaf develops above the cotyledons approximately seven days after seedling

establishment [30]. The cotyledons are the bottom-most leaves and are the only two leaves that develop across from each other on the main stem. The apical meristem will determine the pattern of development for all vegetative growth. If this terminal bud is damaged, the branch below it will take over as the main stem but may suffer abnormalities, and a decrease in productivity [6].

Types of Cotton Leaves and their Growth Patterns

Two types of leaf vegetation grow from the main stem of the cotton plant, namely main stem leaves and subtending leaves. The main stem leaves, attached by a stem-like structure called a petiole, are directly connected to the main stem. Subtending leaves grow from the fruiting branch above a main stem leaf. The cotton plant leaves appear on the main stem in a spiral pattern, with each main stem leaf and branch growing on the nodes of the main stem, with a new node growing every three days on average in the early period of the growing season [30]. After approximately 20 days of growth, the cotton leaves reach their photosynthetic capacity and then begin to decline [30].

Square Formation and Flowering

Roughly 35 days after planting, the cotton plant begins the square bud formation process, called "squaring" [30]. Square buds are fruiting buds and are the first signs of reproductive growth. The first fruiting branch will usually develop on the fifth to the seventh node of the stem, and this branch will develop the first square. Each fruiting branch will average 1 to 4 fruiting positions. The first visual indication of square development is the appearance of the epicalyx; these are leaf-like bracts that will eventually surround the plant's flowers.

The anthesis or flower bloom, shown in Figure 2.2 appears around three weeks after the first square develops. Each cotton plant has female and male organs with which the flowers are produced and fertilized. Pollination of the flowers occurs during the first few hours after blooming, and a pollen tube germinates after the flower's stigma is pollinated. After pollination, the pollen tube extends through the style and micropyle and penetrates the ovule chamber leading to fertilization. The cotton plant's flowers are white immediately after blooming, then

Figure 2.2

Flower Blooms (Author's photo)

Figure 2.3

Open Bolls (Author's photo)

progress to pink and dark pink in five to seven days. After darkening, the blooms will wilt and usually detach from the plant. The fallen flowers reveal the developing cotton bolls that will hold the plant's fruit.

2.6 Cotton Boll Development Phases

The cotton bolls will open under favorable conditions roughly 106 days after planting [30]. The stages of boll development occur in this period, namely the enlargement phase, the filling phase, and the boll maturation phase.

The Enlargement Phase

During the approximately three-week enlargement phase, the plant produces and elongates fibers inside the cotton boll. The fibers grow to maximum length from epidermal cells on the seed coat,

seeds, and bolls reach maximum volume at this stage. Adverse conditions can heavily influence maximum fiber length in the enlargement phase. Inadequate water, unfavorable weather conditions, and lack of nutrients can cause harm to the developing fibers.

The Filling Phase

The next stage, the filling phase, begins during the fourth week after flowering and lasts until the sixth week after pollination. During the filling stage, we see the growth of the cotton fibers stop and the beginning of the formation of the secondary fiber walls. The hollow part of the straw-like fibers receives a deposit of cellulose every day in this stage, and eventually, the space fills. The productivity at this stage is also sensitive to the adverse conditions as in the enlargement phase [6].

The Boll Maturation Phase

In the boll maturation phase, boll size and weight reach a maximum, and bolls split open as the fruit matures fully. The boll capsules dry and then eventually crack, allowing the boll to open. Defoliant use encourages the cotton bolls to open and control harvest time and regrowth. Farmers apply defoliants roughly two weeks before harvest or when approximately 60% of cotton bolls open. The cotton will be fully mature and harvested around 160 days after planting [30]. In the Coastal Bend, most farmers harvest cotton with cotton strippers machines and then process the cotton for production. In the next section, we give an overview of multiple linear regression.

2.7 Mathematical Background

In this section we describe the mathematical background of the models developed in this thesis. An understanding of linear algebra and statistics is assumed.

Multiple Linear Regression

Multiple linear regression (MLR) models the relationship between a single response variable and two or more predictor variables. These predictor variables can be numeric, nominal, or ordinal. MLR analysis assumes that there is a linear relationship between the response and predictor

variables, residuals are normally distributed, and the predictor variables are not highly correlated. The general parametric equation of the response variable when using MLR is given below.

$$y = f(\mathbf{x}) + \varepsilon$$

Where y is the response variable, $f(\mathbf{x})$ is an unknown function, and ε is the error term. Here $f(\mathbf{x})$ is given by,

$$f(\mathbf{x}) = \xi_0 + \xi_1 x_1 + \xi_2 x_2 + \cdots + \xi_k x_k$$

and the estimated value of the response variable is given by,

$$\hat{y} = \hat{\xi}_0 + \hat{\xi}_1 x_1 + \hat{\xi}_2 x_2 + \ldots \hat{\xi}_k x_k$$

Where $\hat{y}$ is the estimation of the response variable, $x_1, x_2, \ldots, x_k$ are the predictor variables (also called features or columns) of the data set, and $\xi_1, \xi_2, \ldots, \xi_k$ are the estimated coefficients of the predictor variables and ξ_0 is the intersection. The intersection is the mean for the response when all predictor variables are set to zero. We define the following matrices to represent this multiple regression in matrix form.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ \vdots & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \qquad \boldsymbol{\xi} = \begin{bmatrix} \xi_0 \\ \xi_1 \\ \vdots \\ \xi_k \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \xi_k \end{bmatrix} \tag{2.1}$$

Where $\mathbf{y}$ is an n dimensional column vector containing the response variables, $\mathbf{X}$ is an $n \times (k+1)$ dimensional matrix with ones in the first column and the values of each predictor

variable in the remaining columns. $\boldsymbol{\xi}$ is a $(k+1)$ dimensional column vector containing the intercept and the coefficients of the predictor variables. Finally, $\boldsymbol{\varepsilon}$ is an n dimensional column vector containing the error terms. The first column of $\mathbf{X}$ is filled with ones and the matrix has $k+1$ columns to account for the intercept value ξ_0. We can now write the equation $\mathbf{y} = \boldsymbol{\xi}\mathbf{X} + \boldsymbol{\varepsilon}$ in matrix form and after performing the matrix multiplication and addition we obtain,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \xi_0 + \xi_1 x_{11} + \xi_2 x_{12} + \cdots + \xi_k x_{1k} + \varepsilon_1 \\ \xi_0 + \xi_1 x_{21} + \xi_2 x_{22} + \cdots + \xi_k x_{2k} + \varepsilon_2 \\ \xi_0 + \xi_1 x_{31} + \xi_2 x_{32} + \cdots + \xi_k x_{3k} + \varepsilon_3 \\ \vdots \\ \xi_0 + \xi_1 x_{n1} + \xi_2 x_{n2} + \cdots + \xi_k x_{nk} + \varepsilon_n \end{bmatrix} \tag{2.2}$$

It can be seen that the entries of equation 2.2 have the form of the multiple regression equation. Now we have the task of estimating the coefficients; this is done by minimizing the least squares criterion, defined below. Consider an n dimensional column vector $\mathbf{e}$ such that,

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 - \hat{y_1} \\ y_2 - \hat{y_2} \\ y_3 - \hat{y_3} \\ \vdots \\ y_n - \hat{y_n} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \hat{y_1} \\ \hat{y_2} \\ \hat{y_3} \\ \vdots \\ \hat{y_n} \end{bmatrix} = \mathbf{y} - \hat{\mathbf{y}}$$

Then the residual sum of squares (RSS) equation is defined as,

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 = \sum_{i=1}^{n} e_i^2 \tag{2.3}$$

then from 2.3 it follows that,

$$RSS = \mathbf{e^T e}$$

$$RSS = (\mathbf{y} - \hat{\mathbf{y}})^T (\mathbf{y} - \hat{\mathbf{y}})$$

$$RSS = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\xi}})^T (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\xi}})$$

$$RSS = (\mathbf{y^T} - \hat{\boldsymbol{\xi}}^{\mathbf{T}}\mathbf{X^T})(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\xi}})$$

After expanding we obtain,

$$RSS = \mathbf{y^T y} - \mathbf{y^T X}\hat{\boldsymbol{\xi}} - \hat{\boldsymbol{\xi}}^{\mathbf{T}}\mathbf{X^T y} + \hat{\boldsymbol{\xi}}^{\mathbf{T}}\mathbf{X^T X}\hat{\boldsymbol{\xi}} \tag{2.4}$$

To estimate the MLR equation's coefficients we must find the vector $\hat{\boldsymbol{\xi}}$ that minimizes equation 2.4. To find the minimum of an equation we wish to find $\hat{\boldsymbol{\xi}}$ such that the derivative (or slope) of the function is zero. So, we take the partial derivative of RSS with respect to $\hat{\boldsymbol{\xi}}$ set equal to zero and obtain the following,

$$\frac{\delta(RSS)}{\delta\hat{\boldsymbol{\xi}}} = 0 \qquad \text{(Taking the partial derivative w.r.t } \hat{\boldsymbol{\xi}}\text{)}$$

$$\frac{\delta}{\delta\hat{\boldsymbol{\xi}}}(\mathbf{y^T y} - \mathbf{y^T X}\hat{\boldsymbol{\xi}} - \hat{\boldsymbol{\xi}}^{\mathbf{T}}\mathbf{X^T y} + \hat{\boldsymbol{\xi}}^{\mathbf{T}}\mathbf{X^T X}\hat{\boldsymbol{\xi}}) = 0$$

$$0 - \mathbf{X^T y} - \mathbf{X^T y} + \mathbf{2X^T X}\hat{\boldsymbol{\xi}} = 0$$

$$2X^T X\hat{\boldsymbol{\xi}} = 2\mathbf{X^T y} \qquad \text{(Simplifying and isolating the } \hat{\boldsymbol{\xi}} \text{ term)}$$

$$\mathbf{X^T X}\hat{\boldsymbol{\xi}} = \mathbf{X^T y}$$

$$(\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T X}\hat{\boldsymbol{\xi}} = (\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T y} \qquad \text{(Isolating } \hat{\boldsymbol{\xi}} \text{ vector)}$$

Finally, we obtain the following estimate of the $\hat{\boldsymbol{\xi}}$ column vector,

$$\hat{\boldsymbol{\xi}} = (\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T y} \tag{2.5}$$

Now, given $\mathbf{X}$ and $\mathbf{y}$, we can obtain estimates of the coefficients to calculate the predicted response, $\hat{\mathbf{y}}$, of $\mathbf{y}$. However, if multicollinearity is present in the data set, the prediction model may over-fit, making our MLR model inefficient at predicting with data it has not seen before. Multicollinearity is defined as the presence of multiple highly correlated predictor variables in a data set. In the following four sections, we describe several alternative methods with advantages over MLR.

Ridge Regression

Ridge regression or L^2 norm regularization is a shrinkage method that constrains or "shrinks" all coefficient estimates towards zero, but never exactly zero. Ridge is similar to MLR except that it uses a penalty term when estimating the coefficients to avoid over-fitting. The penalty term penalizes the sum of squares of the coefficients. Finding the coefficients with this method is equivalent to the method of Lagrange multipliers which is used to find a maximum or minimum function. We want to find $\hat{\boldsymbol{\xi}}$ that minimizes the following equation (we define all vectors and matrices as in 2.1).

$$\mathscr{L} = RSS + \lambda \hat{\boldsymbol{\xi}}^T \hat{\boldsymbol{\xi}} \tag{2.6}$$

Where λ is a nonnegative tuning parameter or the Lagrangian multiplier, $\hat{\boldsymbol{\xi}}$ is the vector of coefficient estimates, and RSS is the residual sum of squares. We can see from the equation above that if $\lambda = 0$, we minimize the RSS, and the Ridge regression be equivalent to the MLR. Furthermore, the estimated coefficients will approach zero if $\lambda \to \infty$. To find the estimates of $\hat{\boldsymbol{\xi}}$

we take the derivative of $\mathscr{L}$, set the derivative equal to zero and solve for $\hat{\boldsymbol{\xi}}$.

$$\frac{\delta(\mathscr{L})}{\delta\hat{\boldsymbol{\xi}}} = 0 \qquad \Longrightarrow$$

$$\frac{\delta}{\delta\hat{\boldsymbol{\xi}}}(RSS + \lambda\hat{\boldsymbol{\xi}}^T\hat{\boldsymbol{\xi}}) = 0 \qquad \Longrightarrow$$

$$\frac{\delta}{\delta\hat{\boldsymbol{\xi}}}(\mathbf{y^T y} - \mathbf{y^T X}\hat{\boldsymbol{\xi}} - \hat{\boldsymbol{\xi}}^{\mathbf{T}}\mathbf{X^T y} + \hat{\boldsymbol{\xi}}^{\mathbf{T}}\mathbf{X^T X}\hat{\boldsymbol{\xi}} + \lambda\hat{\boldsymbol{\xi}}^T\hat{\boldsymbol{\xi}}) = 0 \qquad \text{(By 2.4)}$$

$$0 - \mathbf{X^T y} - \mathbf{X^T y} + \mathbf{2X^T X}\hat{\boldsymbol{\xi}} + 2\lambda\hat{\boldsymbol{\xi}} = 0 \qquad \Longrightarrow$$

$$2\mathbf{X^T X}\hat{\boldsymbol{\xi}} + 2\lambda\hat{\boldsymbol{\xi}} = 2\mathbf{X^T y} \qquad \Longrightarrow$$

$$\mathbf{X^T X}\hat{\boldsymbol{\xi}} + \lambda\hat{\boldsymbol{\xi}} = \mathbf{X^T y} \qquad \Longrightarrow$$

$$(\mathbf{X^T X} + \lambda\mathbf{I})\hat{\boldsymbol{\xi}} = \mathbf{X^T y} \qquad \text{(Isolating } \hat{\boldsymbol{\xi}}\text{)}$$

$$(\mathbf{X^T X} + \lambda\mathbf{I})^{-1}(\mathbf{X^T X} + \lambda\mathbf{I})\hat{\boldsymbol{\xi}} = (\mathbf{X^T X} + \lambda\mathbf{I})^{-1}\mathbf{X^T y}$$

After applying the inverse of $(\mathbf{X^T X} + \lambda\mathbf{I})$ to both sides we find the estimates of the Ridge regression coefficients to be,

$$\hat{\boldsymbol{\xi}}^{Ridge} = (\mathbf{X^T X} + \lambda\mathbf{I})^{-1}\mathbf{X^T y} \tag{2.7}$$

Least Absolute Shrinkage and Selection Operator

The least absolute shrinkage and selection operator (LASSO) regression can select a subset of a given data matrix by finding highly correlated variables and shrinking the coefficient of all but one to zero. LASSO solves nearly the same optimization problem as Ridge regression, except the penalty term is the ℓ^1 norm instead of the L^2 norm. So we want to find $\hat{\boldsymbol{\xi}}$ that minimizes the following equation.

$$\mathscr{L} = RSS + \lambda\|\xi\|_1 \tag{2.8}$$

Then we wish to solve the following to obtain the lasso coefficients,

$$\hat{\boldsymbol{\xi}}^{lasso} = \min_{\boldsymbol{\xi}}\{\frac{1}{2}\sum_{i=1}^{n}(y_i - \xi_0 - \sum_{j=1}^{m} x_{ij}\xi_j)^2 + \lambda \sum_{j=1}^{m} |\xi_j|\} \tag{2.9}$$

Where y_i are the elements of the response vector, ξ_0 is the intercept of the regression equation, ξ_j are the elements of the coefficient vector, and λ is a nonnegative tuning parameter by the sum of absolute values of the coefficients (the ℓ^1 norm). LASSO regression has the ability to form a subset of the predictor variables while maintaining their original state. LASSO regression is a useful alternative to MLR and can solve the problem of multicollinearity.

Principal Component Analysis

Principal component analysis (PCA) is a widely used multivariate statistical procedure invented in 1901 by Karl Pearson and subsequently named and developed by Harold Hotelling in the 1930s [28][11]. PCA is used in agriculture [33], data denoising [8], neuroscience [27], and several other critical fields. PCA is attractive for its simplicity; the non-parametric method extracts underlying patterns and relationships from complicated data sets [34]. PCA can effectively reduce the dimension of a data set where multicollinearity exists, which is attractive for regression analysis. In regression modeling, the statistical significance of the variables diminishes when multicollinearity is present. PCA reduces m number of correlated predictor variables into p number of uncorrelated, or nearly uncorrelated, linear combinations of the original variables, where $n, p \in \mathbb{Z}^+$. These linear combinations form a set of orthogonal variables called principal components. The first principal component direction, consisting of projections of the data points onto a unit vector, "explains" the most variance [12]. Geometrically, the projections onto the principal component directions explain more variance the more the projections onto the vector are spread out. If projections of the data points are close together, they vary less and will therefore capture less information about the data set. The components are sorted into descending order of explained variance, starting with the first component. These components form an orthogonal basis; in other words, the components are

uncorrelated, which can solve the problem of high multicollinearity in a data set [34].
Let us take our $n \times m$ set of features as the matrix $\mathbf{X}$. Computationally PCA is performed by standardizing $\mathbf{X}$, computing the covariance matrix then computing the eigendecomposition of the standardized matrix, sorting the resulting eigenvalues in decreasing order, and projecting the data points onto the new subspace. When PCA is performed with technology it is easy to find results and analyze them but what enables us to perform this analysis? The methods find their roots in linear algebra and calculus [21]. Consider a unit vector $\vec{v} \in \mathbb{R}^{n\times 1}$ in our n-dimensional vector space and let the standardized version of $\mathbf{X}$ be the $n \times m$ matrix $\mathbf{A}$ with the form,

$$A = \begin{bmatrix} \frac{x_{11}-\mu_1}{S_1} & \cdots & \frac{x_{1m}-\mu_m}{S_m} \\ \vdots & \ddots & \vdots \\ \frac{x_{n1}-\mu_1}{S_1} & \cdots & \frac{a_{nm}-\mu_m}{S_m} \end{bmatrix} \tag{2.10}$$

where $\mu_1, \ldots, \mu_m$ and $S_1, \ldots, S_m$ are the means and the standard deviations of each column of $\mathbf{A}$, respectively. Now, each sample point's strength in the direction of v can be expressed as the projection of the point onto v. The projections onto the unit vector are calculated with the following formula,

$$Proj_{\vec{v}} a^i = (a^i \vec{v})\vec{v} \tag{2.11}$$

where a^i is the i^{th} column vector of the scaled matrix $\mathbf{A}$ and $(a^i\vec{v})$ is the coordinate of a^i in the direction of $\vec{v}$. Considering all m samples we can express their squared strength of variance in the direction of v as the following equation.

$$\kappa = \sum_{i=}^{m} |\vec{v}^T a^i|^2 = \sum_{i=1}^{m} \vec{v}^T a^i (a^i)^T \vec{v} = \vec{v}^T (\sum_{i=1}^{m} a^i (a^i)^T)\vec{v} = \vec{v}^T A A^T \vec{v} \tag{2.12}$$

Then to find the maximum variance, we must solve the following optimization problem.

$$\max_{\vec{v}} \kappa(\vec{v}) \quad \text{such that,} \quad |\vec{v}| = 1 \tag{2.13}$$

To find $\vec{v}$ that maximizes κ in 2.13 we can use the method of Lagrange multipliers. The Lagrangian equation of 2.13 is,

$$\mathscr{L}(\vec{v}) = \vec{v}^T \mathbf{A}\mathbf{A}^T \vec{v} - \lambda(\vec{v}^T \vec{v} - 1) \tag{2.14}$$

Then to maximize 2.14, we find the stationary points of $\mathscr{L}(\vec{v})$ which gives the conditions below.

$$\frac{\partial \mathscr{L}}{\partial \vec{v}} = 0, \quad 2\mathbf{A}\mathbf{A}^T \vec{v} - 2\lambda \vec{v} = 0 \tag{2.15}$$

$$\frac{\partial \mathscr{L}}{\partial \lambda} = 0, \quad \vec{v}^T \vec{v} - 1 = 0 \tag{2.16}$$

From the second equation of 2.15 we obtain,

$$\mathbf{A}\mathbf{A}^T \vec{v} = \lambda \vec{v} \tag{2.17}$$

The form of equation 2.17 implies that the vector $\vec{v}$ is an eigenvector of $\mathbf{A}\mathbf{A}^T$. Since the product of a matrix and its transpose is a real symmetric and positive semi-definite matrix $\mathbf{A}\mathbf{A}^T$ has only real non-negative eigenvalues and real eigenvectors. Now, without loss of generality we are able to arrange the eigenvalues of $\mathbf{A}\mathbf{A}^T$ in descending order such that,

$$\lambda_1 \geq \lambda_2 \geq \ldots \lambda_k \geq \cdots \geq \lambda_n \tag{2.18}$$

$$\tag{2.19}$$

with the corresponding orthogonal eigenvectors

$$\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_k, \ldots, \vec{v}_n \tag{2.20}$$

Now, the equation 2.12 is maximized in the direction of $\vec{v}_1$ associated with the eigenvalue λ_1. Then equation 2.17 can be rewritten as,

$$\mathbf{A}\mathbf{A}^T = \vec{v}\lambda\vec{v}^T \implies \tag{2.21}$$

$$\tag{2.22}$$

$$\mathbf{A}\mathbf{A}^T = \begin{pmatrix} \vec{v}_1 & \vec{v}_2 & \vec{v}_3 \ldots & v_k \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ 0 & 0 & \lambda_3 & \ldots & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \lambda_k \end{pmatrix} \begin{bmatrix} \vec{v}_1 \\ \vec{v}_2 \\ \vec{v}_3 \\ \vdots \\ v_k \end{bmatrix} \implies \tag{2.23}$$

$$\tag{2.24}$$

$$\mathbf{A}\mathbf{A}^T = \lambda_1\vec{v}_1\vec{v}_1{}^T + \lambda_2\vec{v}_2\vec{v}_2{}^T + \lambda_3\vec{v}_3\vec{v}_3{}^T + \cdots + \lambda_k\vec{v}_k\vec{v}_k{}^T + \cdots + \lambda_m\vec{v}_m\vec{v}_m{}^T \tag{2.25}$$

After diagonalizing $\mathbf{A}\mathbf{A}^T$ we can choose $k < m$ terms of 2.25 then project the data points of $\mathbf{A}$ onto the principal components to achieve dimension reduction. Here, the eigenvectors

$$\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_k, \ldots, \vec{v}_k \tag{2.26}$$

describe the direction of the principal components with respect to the original data features. The elements of each eigenvector are called loadings and will become the coefficients of the original data features. These loadings tell us the amount of correlation each variable has to the eigenvector. Let the matrix $\mathbf{V}$ contain the chosen k eigenvectors used to lower the dimension of

A such that,

$$\mathbf{V} = \begin{bmatrix} v_{11} & \dots & v_{1k} \\ \vdots & \ddots & \vdots \\ v_{m1} & \dots & v_{mk} \end{bmatrix} \tag{2.27}$$

Then we can project the data points of the scaled matrix **A** by performing the following matrix multiplication.

$$\mathbf{A}' = \mathbf{AV} \quad \Longrightarrow \tag{2.28}$$

$$\mathbf{A}' = \begin{bmatrix} a_{11} & \dots & a_{m1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} v_{11} & \dots & v_{1k} \\ \vdots & \ddots & \vdots \\ v_{m1} & \dots & v_{mk} \end{bmatrix} \tag{2.29}$$

Then $\mathbf{A}'$ is the $n \times k$ projected data matrix whose elements represent the length of each original data point onto the origin. We have now lowered the dimensions of our data set by replacing the m features with the chosen number of principal components. The principal components have the form,

$$\vec{p}_z = \mathbf{A}\vec{v}_z \tag{2.30}$$

where $\vec{p}_z$ is the z^{th} principal component for $\vec{z} \in 1, 2, ...k$ and $\vec{v}_z$ is the $z^t h$ column vector in $\vec{v}$. Note that because the principal components are orthogonal to each other, they are uncorrelated. This lack of correlation solves the problem of multicollinearity in the data set while retaining the information of the original features in the principal components.

Principal Component Regression

Principal component regression (PCR) regresses a response variable with a chosen number of principal components and predictor variables. This method can be used as an alternative to

multiple linear regression. In PCR the data matrix has been formed by a selected k principal components. The transformed data matrix has the form,

$$\mathbf{A}' = \begin{bmatrix} \vec{p}_1 & \vec{p}_2 & \dots & \vec{p}_k \end{bmatrix} \tag{2.31}$$

where $\vec{p}_1, \vec{p}_2, \dots, \vec{p}_k$ are the chosen first k components. Recall from the previous section that the principal components are uncorrelated so there is no problem of multicollinearity. Then the response variable y is estimated as,

$$\hat{\mathbf{y}} = \boldsymbol{\xi}\mathbf{A}' \tag{2.32}$$

where $\boldsymbol{\xi}$ is a vector containing the intercept and the estimated coefficients that are estimated by,

$$\boldsymbol{\xi} = ((\mathbf{A}')^{\mathbf{T}}\mathbf{A})^{-\mathbf{1}}(\mathbf{A}')^{\mathbf{T}}\mathbf{y} \qquad \text{(By 2.5)}$$

So, we estimate the response variable with a lower number of predictor variables and little to no correlation between the predictor variables. A downside to PCR is that it is unsupervised, meaning the response variable is not considered when extracting the principal components. In the next section, we discuss a more robust model that is similar to principal component analysis called partial least squares regression (PLSR).

Partial Least Squares Regression

Partial least squares regression, also called projection to latent structures (PLS), is a supervised method that can be used to predict one or multiple response variables. The method was created by Herman O.A. Wold in the 1960s and has some advantages over PCR [29]. Like PCR, PLSR involves forming linear combinations of the original data column vectors called latent variables or components. However, unlike PCR, we find components that best explain the predictor and response variables. Let the $n \times m$ matrix, $\mathbf{X}$, be the set of predictor variables, $\mathbf{y}$ be the response

variable, and **A** be the standardized matrix of **X**. In PLSR, scores are extracted from both **A** and **y** with the objective that the scores have maximal covariance. The process of developing a PLSR model involves regressing every column of **A** onto the score vector $\vec{u}$ of **y**, filling a vector $\vec{w}$ with the resulting regression coefficients, normalizing $\vec{w}$, regressing each row of **A** onto $\vec{w}$, storing the regression coefficients in the score vector $\vec{t}$ of **A**, regressing **y** onto $\vec{t}$, storing the coefficients in $\vec{c}$, and regressing each row of **y** onto $\vec{c}$. To start, we set $\vec{u}$ to the first column of **y**, which in our case is the only column of **y**. Now we compute the first set of regressions and store them in $\vec{w}$ we calculate the vector,

$$\vec{w} = \frac{1}{\vec{u}^T\vec{u}} \cdot \mathbf{A}^T\vec{u} \tag{2.33}$$

This results in a vector containing the regression coefficients of $\mathbf{A}^T\vec{u}$. The weights of $\vec{w}$ are large if the columns of **A** and the vector $\vec{u}$ are strongly correlated. Now, to regress **A** onto $\vec{w}$ and store the coefficients in $\vec{t_a}$ we calculate,

$$\vec{t_a} = \frac{1}{\vec{w}^T\vec{w}} \cdot \mathbf{A}^T\vec{w} \tag{2.34}$$

The values of $\vec{t}$ are large when the rows of **A** are similar to the $\vec{w}$ and go to zero it the rows are very different. Next we regress **y** onto the vector $\vec{t}$ and store the resulting regression coefficients in $\vec{c}$ by calculating,

$$\vec{c} = \frac{1}{\vec{t}^T\vec{t}} \cdot \mathbf{y}^T\vec{t} \tag{2.35}$$

In the last step before deflating process we regress the rows of **y** onto

$$\vec{u} = \frac{1}{\vec{c}^T\vec{c}} \cdot \mathbf{y}\vec{c} \tag{2.36}$$

We iterate through these steps until $\vec{u}$ converges. When $\mathbf{y}$ is a 1-dimensional column vector, as is the case in this thesis, convergence is achieved in one iteration. After convergence the $\mathbf{A}$ and $\mathbf{y}$ loadings are as follows,

$$\vec{p} = \frac{1}{\vec{t}^T\vec{t}} \cdot \mathbf{A}^T\vec{t} \quad \text{for } \mathbf{X} \text{ and} \tag{2.37}$$

$$\vec{q} = \frac{1}{\vec{u}^T\vec{u}} \cdot \mathbf{y}^T\vec{u} \quad \text{for } \mathbf{y} \tag{2.38}$$

Then our goal is to maximize the covariance between $\vec{t}$ and $\vec{u}$,

$$Cov(\vec{t},\vec{u}) = Correlation(\vec{t},\vec{u}) \times \sqrt{\vec{t}^T\vec{t}} \times \sqrt{\vec{u}^T\vec{u}} \tag{2.39}$$

Maximizing Equation 2.39 gives the best explanation of the predictor variables by maximizing the variance of $\vec{t}$, the best explanation of the response by maximizing the variance of $\vec{u}$, and the greatest relationship between the predictor space and response variable my maximizing the correlation between $\vec{t}$ and $\vec{u}$.

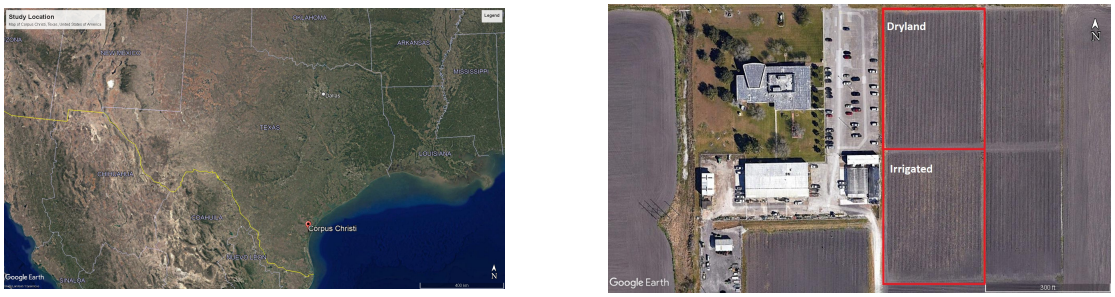
CHAPTER III: METHODOLOGY

3.1 Introduction

In this chapter, we describe the geography and climate of the study area, the specifications of tools used to process the UAV collected imagery, and detail the computation of our predictor variables.

3.2 Study Area Location

The data set utilized in this thesis was collected from an experimental field in Corpus Christi, Nueces County, Texas, USA, during the 2016 growing season. The experimental field, shown in Figure 3.4, is located at the Texas A&M AgriLife Research Center in Corpus Christi, Texas, USA (27° 46' 57.08" N, 97° 33' 40.94" W). Corpus Christi is in a humid subtropical region with flat coastal grasslands located in the Coastal Bend of southern Texas, bordering the Gulf of Mexico, as can be seen in Figure 3.4.



(a) Map of Corpus Christi, Texas, USA

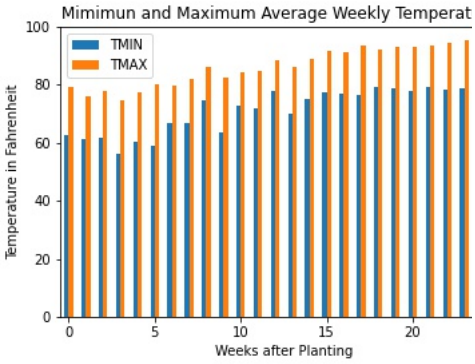
(b) Aerial View of Study Area

Figure 3.4

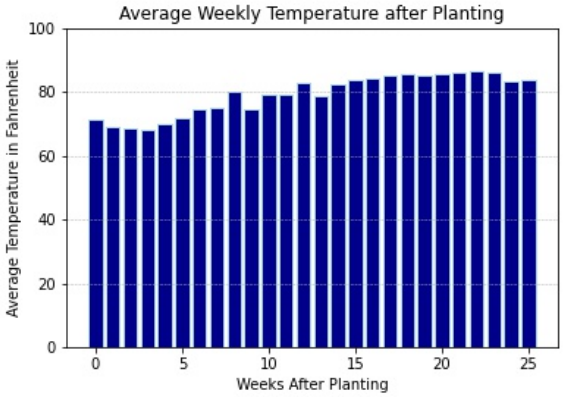
(a) Google Earth aerial image of study location. (b) Google Earth image of experimental field location.

3.3 Study Area Climate

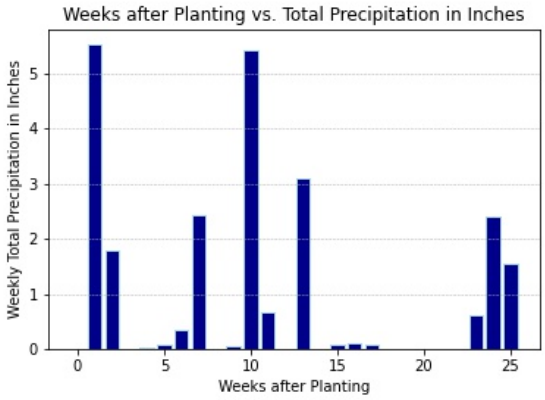
The climate data used here was provided by the National Oceanic and Atmospheric Administration (NOAA) and collected by the Corpus Christi National Weather Station and Corpus Christi international airport station; the stations are 5.6 miles and 5.9 miles from the experimental field, respectively. Temperatures in the area are favorable for cottonseed



(a) Minimum and Maximum Temperature



(b) Average Weekly Temperature



(c) Total Weekly Precipitation in Inches

Figure 3.5

(a) A bar plot displaying the mean minimum and maximum weekly temperatures in °F. Data provided by NOAA and collected from the National Weather Service in Corpus Christi. (b) A scatter plot showing the average weekly temperature in °F. Data provided by NOAA and collected from the Corpus Christi International Airport (c) A bar plot displaying the total inches of rainfall by weeks after planting. Data provided by NOAA and collected from the National Weather Service in Corpus Christi.

germination and sprout emergence, averaging above 60°F consistently in the early weeks of the season. The area experiences high temperatures in the summer; temperatures reach 90°F or higher regularly during the late growing season. During the early growing season, temperatures are moderate. Temperatures in the area reach well into the 90s during the middle to the late

growing season, as is illustrated in Figure 3.5b and Figure 3.5a. During the day, the cotton plant can open pores on its leaves called stomates, allowing it to cool itself [26]. However, the cotton plant closes its pores during the night and can no longer cool its bolls. The high temperatures in the study area may inhibit cotton quality and lint yield if temperatures during the night exceed 80 °F. The area also experiences high humidity during the boll maturation phases, making it difficult for the plants to regulate their temperature. When these adverse conditions occur, the plants rely on the evaporation of soil moisture to cool. The plant's ability to cool is very important during the boll development phases. If the plant cannot cool itself, it may respond by producing smaller bolls or dropping bolls. Irrigation can aid in soil moisture, but for dry plots, precipitation is necessary, or farmers must intervene. Precipitation in this area is sporadic, as shown in Figure 3.5c. Consequently, farmers may have to supplement the precipitation with sprinklers or other methods.

3.4 Collection of Data

Traditionally, researchers collect agricultural data from hand measurements. Hand measurements in large fields can be inefficient and tedious. More recently, satellite imagery has been used to solve this problem, but visual obstructions may be a problem. Clouds can cause the Earth's surface to appear warped by their shadows or outright block a satellite's view of the Earth. Drones offer a solution to this problem and have recently become relatively inexpensive and more accessible. Camera-equipped drones can collect imagery efficiently, avoid most visual obstructions, and provide multispectral images of fields throughout the growing seasons. These images can then be processed and used to calculate physical plant attributes and vegetation indices.

Drone, Senors, and Platform Specifications

Drones have become attractive for precision farming due to their availability and relatively low cost. This thesis uses data computed from sensor-equipped unmanned aerial vehicle (UAV) collected imagery. UAV, commonly known as drones, are defined as any aircraft that does not

require a pilot on board to fly. Two sensor-equipped drones collected RGB and multispectral imagery of the experimental field with flights beginning on April 7, 2016, and ending on August 18, 2016, with a total of 30 flights [3]. A DJI Phantom 2 drone equipped with a 3-axis gimbal-stabilized RGB sensor camera with an image size of 4384 x 3288 pixels collected RGB imagery [3]. A 3-axis gimbal allows the camera on the drone to rotate on three axes while keeping object-orientation steady during rotation. A multispectral sensor camera (Tetracam ADC snap sensor) mounted to a 3DR IRIS drone captured red, green, and near-infrared spectral bands with an image size of 1280 x 1024 pixels [3]. Based on weather conditions and availability of equipment, the collection of multispectral data occurred every 7 to 10 days, while the RGB data was collected every 4 to 7 days [3].

Structure from Motion Photogrammetric Processing

Photogrammetry is the science of processing photos to acquire a 3-dimensional model of an object of interest, drawing, map, or measurement. Researchers use structure from motion photogrammetry to extract 3D object information from images that overlap by more than 80 percent without the need for precise camera calibration. The process of extracting information from the raw UAV collected images is described by [3] as follows. Proceeding each drone flight, the SfM-based commercial software Agisoft Photoscan Pro 1.3.0 processed the raw RBG and multispectral images [3]. First, a feature matching algorithm called scale-invariant feature transform (SIFT) identified corresponding key features in the set of overlapping raw images [3]. A bundle block adjustment process reconstructed the exact camera position and orientation. The bundle block adjustment receives key features and ground control point coordinates to calculate the 3D coordinates of the matching points. The matching points' 3D coordinates and color information are stored as a set of 3D points called the densified point cloud. Next, a triangulated irregular network (TIN) generates a digital surface model (DSM). A DSM represents the Earth's surface and the objects on it. A triangulated irregular network is a tiling of non-overlapping triangles representing surface morphology. Lastly, an orthorectified image mosaic is created

using the DSM to project every image pixel. Orthorectification is the process of creating a 3D model that can measure distance accurately; it does this by removing distortions such as terrain from raw imagery. In the next section, the computation of the predictor variables from the DSM is detailed.

Computing Predictor Variables

The temporal canopy attributes considered in this study included canopy height and canopy volume and the temporal vegetative indices (VIs), excess greenness index, and normalized difference vegetation index. We also consider the qualitative parameter irrigation, classifying plots as irritated or dry.

Prior to planting, a digital terrain model (DTM) of the study area was generated as a base for canopy height computation. This base DTM was then subtracted from each DSM generated from images collected to compute canopy height throughout the season; the average canopy height of each plot was then computed from the canopy height model maps [3].

The following equation was used to compute canopy volume per plot.

$$CV = \sum_{i} (H_i \times GSD^2) \tag{3.40}$$

Here, H_i is the height of i^th pixel of each plot, and GSD is the ground sample distance within each plot [3]. GSD is defined as the distance between two pixels in an image.

NDVI images were created using equation 3.41 from Rouse et al. (1974). The average NDVI per plot was computed by taking the average of all pixel values within each plot [3].

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)} \tag{3.41}$$

The variables on the right side of equation 3.41 are spectral bands of the electromagnetic spectrum, namely the red and near-infrared (NIR) bands. The red band falls into the visible light spectrum while the NIR is just beyond the visible light spectrum [2]. When a plant has a high

level of chlorophyll, it will reflect more NIR energy, so a healthier plant will have a higher NIR [2].

To calculate excess greenness index images, the following equation from [37] was used then the average ExG of the pixels of each plot was taken.

$$ExG = 2G - R - B \tag{3.42}$$

$$G = \frac{green}{max(green)}; B = \frac{blue}{max(blue)}; R = \frac{red}{max(red)} \tag{3.43}$$

Where green, blue, and red are the corresponding spectral bands in the visible light spectrum.

CHAPTER IV: FINDINGS/RESULTS

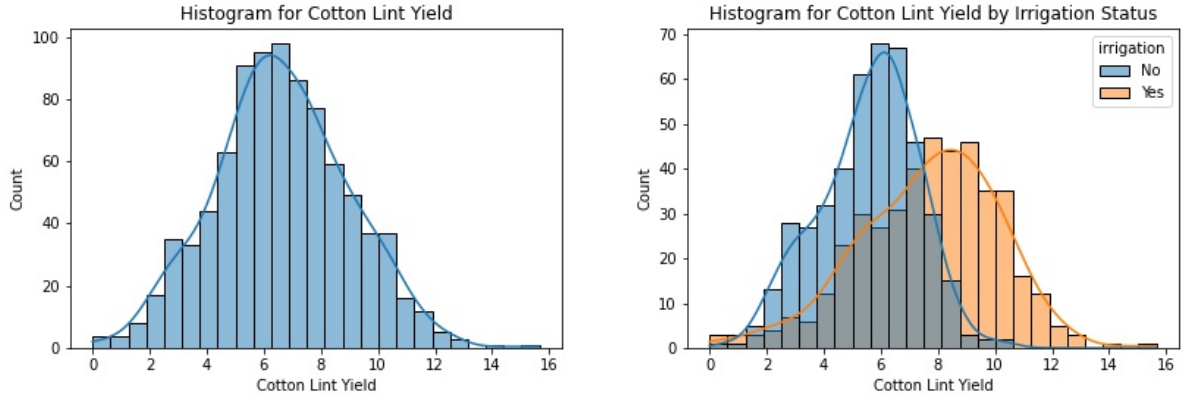
4.1 Exploration of Data Set

In this section, we describe key features of the data set, visualize relationships between variables, and provide an analysis of our results. The preliminary analysis of the data set will allow us to uncover any underlying patterns and help select an appropriate model. We executed the analysis of the data set using the programming language Python version 3.10.0. We begin our analysis by defining the variables of our data set below.

Variable Name	Symbol	Variable Definition
Order	order	Number assigned to each point by row and column values starting at the northwest corner of the field.
Row	row	Row number from east to west.
Column	col	Column number from north to south.
Plot Number	rep	Plot number.
Seed Type	Seed_Type	Seed genotype name or code
Irrigation	irrigation	irrigation status, either yes or no.
Actual Yield per row	Yield	Cotton lint yield in pounds.
Excess greenness Index	ExG(mmdd)	Excess greenness index, RBG sensor-based index. Each of these variable names is followed by a date in the format (mmdd).
Normalized Difference Vegetation Index	NDVI(mmdd).	Normalized Difference Vegetation Index, multispectral sensor-based index. Each of these variable names is followed by a date in the format (mmdd).
Canopy Height	CH(mmdd)	Height of cotton plant in meters, each of these variable names is followed by a date in the format (mmdd).
Canopy Volume	CV(mmdd)	Canopy volume in m^3, each of these variable names is followed by a date in the format (mmdd).

Table 4.1
Description of Data Set Variables

We will take the variable Yield, seen in Table 4.1, as our response variable for our regression models. Here, we measure Yield in pounds per row in every plot. Below, we provide two histograms that display the distribution of the response variable.



(a) Cotton Lint Yield Histogram (b) Cotton Lint Yield Histogram by Irrigation

Figure 4.6

(a) A histogram of cotton lint Yield per row (b) A histogram of cotton lint Yield per row by irrigation status.

From Figure 4.6 we see that Yield has a mean near 6 pounds, the Yield values fall between approximately 0 and 16 pounds, and the distribution of Yield appears to be relatively symmetric. We separate Yield by irrigation in Figure 4.6, there are some clear differences in distribution. The Yield of irrigated rows ranges from roughly 0 to 15 pounds, while the dry rows' yields range from roughly 0 to 11 pounds. In addition, the mean of the irrigated group is about 8 pounds, and the center of the dry group is about 6 pounds.

In the box and whisker plot shown in Figure 4.7, there is an indication of some possible outliers for both dry and irrigated Yield values. The spread of the irrigated group is larger than the dry group; approximately 25 percent of the irrigated Yield values are larger than the dry Yield values.

In Table 4.2 we see that there are 434 irrigated rows and 441 dry rows. The irrigated plots have a mean of 7.68 pounds with a standard deviation of 2.46, and the dry plots have a mean of 5.57

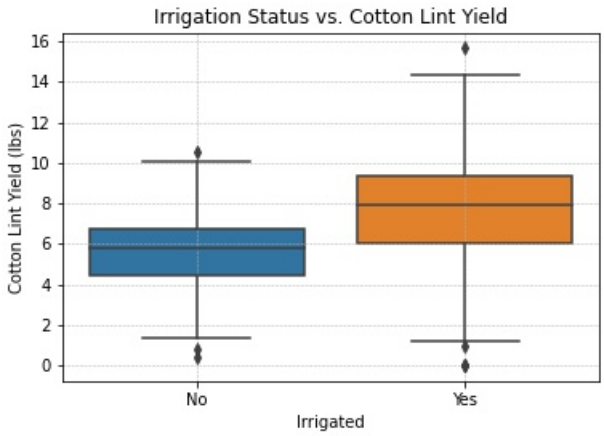


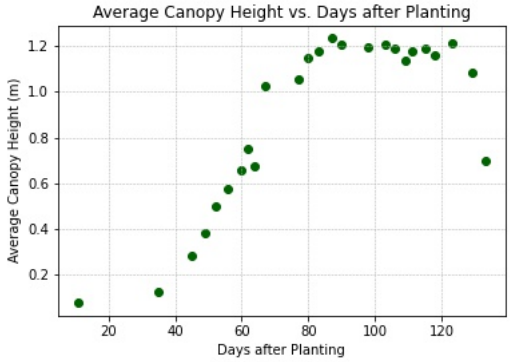
Figure 4.7
Yield grouped by Irrigation Box Plot

	Yield (lbs) Irrigated	Yield (lbs) Dry
N	434	441
Mean	7.68	5.57
S.D.	2.46	1.73
Min	0	0.40
Median	7.93	5.76
Maximum	15.65	10.54

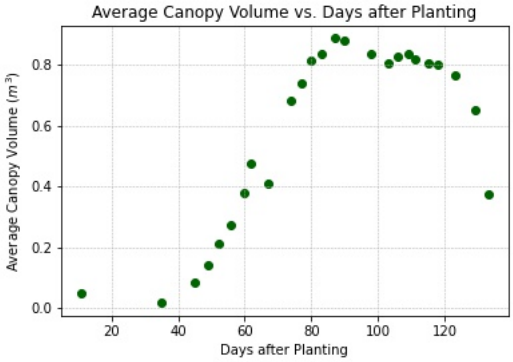
Table 4.2
Yield by Irrigation Summary Statistics

pounds with a standard deviation of 1.73. The summary statistics given in Table 4.2 indicate the irrigated plots had more successful Yields than the dry plots. High temperatures and humidity may explain the difference in Yield in the area during boll development. Irrigated plots may have had more access to more soil moisture which could have aided in maintaining a favorable boll temperature.

In Figure 4.8 two scatter plots illustrate the average canopy height in meters and average canopy volume in cubic meters of the entire field against days after planting. There is a clear pattern for both height and volume; both plots increase as the days increase, reaching a maximum before they decline. The two plots have very similar shapes. The similarities between the height and volume are not surprising as the calculation of canopy volume involves canopy height. The eventual decline may coincide with the application of defoliants which force the plants to shed



(a) Average Canopy Height vs. Time



(b) Average Canopy Volume vs. Time

Figure 4.8

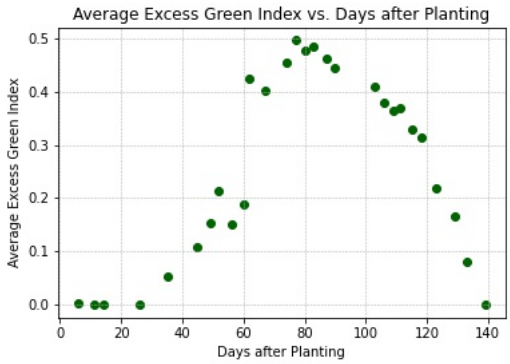
(a) A scatter plot representing change in average canopy height of the field over the growing season. Canopy height is given meters and time expressed as days after planting. (b) A scatter plot representing change in average canopy volume of the field over time. Canopy volume is given cubic meters and time expressed as days after planting.

their leaves and increase boll production before harvesting.

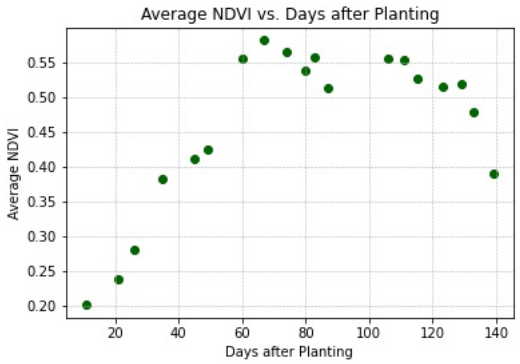
Figure 4.9 we see a similar pattern to the canopy height and volume. Multi-spectral sensor-equipped drones flew less, resulting in fewer NDVI points nonetheless, an increase occurs, reaching a maximum of roughly 0.58, then the points begin to decline with time. The average excess greenness index points are similar to NDVI, peak at around 0.5, then decline.

The scatter plots in Figure 4.10 display the relationship between the early, middle, and late-season canopy height, in meters, and Yield. We averaged the first nine weeks of the canopy height variables as Early Canopy Height, the next nine weeks as Mid Canopy Height, and the remaining nine weeks as Late Canopy Height. All three graphs appear to have a positive correlation, but the early and mid-season plots have a slightly stronger correlation of 0.61 and 0.59, respectively than the late-season canopy height correlation of 0.48. Early detection of low Yield can be very useful for farmers, allowing them to treat specific crop areas.

The three scatter plots presented in Figure 4.11 illustrate the early, middle, and late-season average canopy volume against Yield. We averaged the first nine weeks of the canopy volume



(a) Average Excess Greenness Index vs. Time

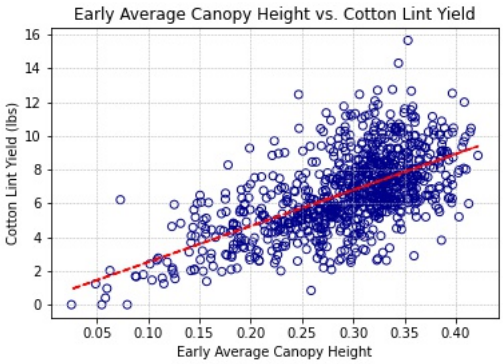


(b) Average NDVI vs. Time

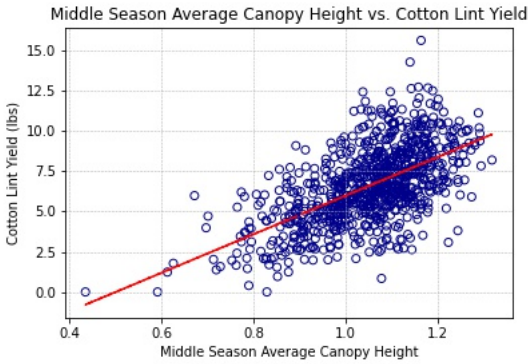
Figure 4.9

(a) A scatter plot representing change in average excess greenness index of the field over time. Time is expressed as days after planting. (b) A scatter plot representing change in average normalized difference vegetation index (NDVI) of the field over time. Time is expressed as days after planting.

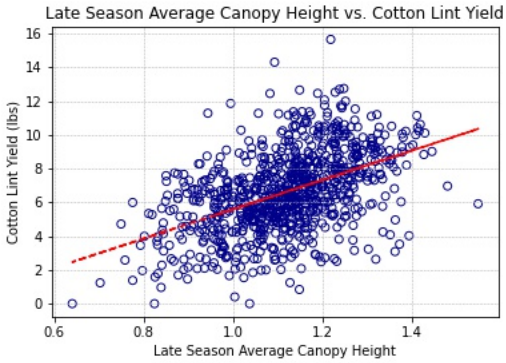
variables as Early Canopy Volume, the next nine weeks as Mid Canopy Volume, and the remaining nine weeks as Late Canopy Volume. While a positive trend is visible in all three plots, the Mid Canopy Volume plot in Figure 4.11b has a slightly higher correlation value of 0.54, while the Early Canopy Volume plot and Late Canopy Volume plot have correlation values of 0.53 and 0.52 respectively. Similar to the average canopy height, greater canopy volume indicates healthy plants and future Yield. More leaf area will increase the plant's productivity in the early and middle growing season by providing the plant with more food.



(a) Early Canopy Height vs Yield



(b) Mid Canopy Height vs. Yield

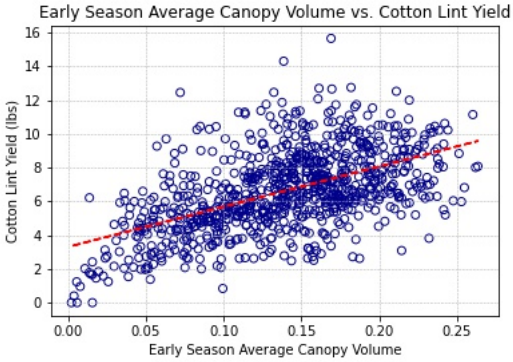


(c) Late Canopy Height vs. Yield

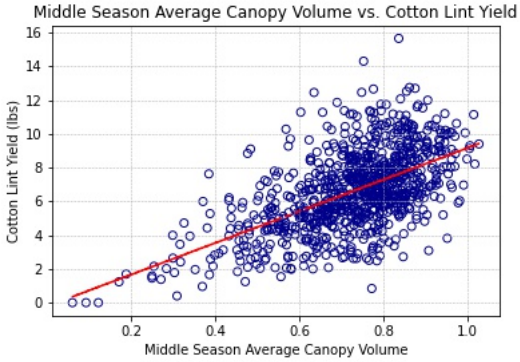
Figure 4.10

Scatter plots displaying the average (a) early season, (b) mid=-season, and (c) late season canopy height in meters versus Yield in pounds.

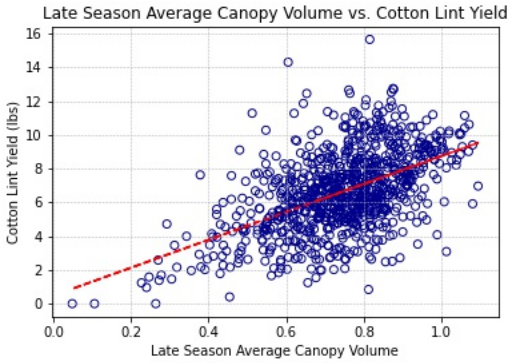
Next, in Figure 4.12, we see the average early, middle, and late-season excess greenness index against Yield. There were fewer weeks of ExG collected, so the separation into average early, middle, and late-season were slightly different from CH and CV. The first seven weeks were averaged for Average Early ExG ending on the same date as Early CH and CV. The next nine weeks were averaged for Average Mid ExG and the last nine weeks for Average Late ExG. We see a positive slope between the average excess greenness index and Yield in the early and mid-season. The Average Early ExG has a correlation of 0.49, Average Mid ExG has a correlation of 0.57, and Average Late ExG has a correlation of -0.02. During the season's last weeks, the plants turn brown after losing their leaves to increase boll production, which may



(a) Average Early Canopy Volume vs. Yield



(b) Average Mid Canopy Volume vs. Yield



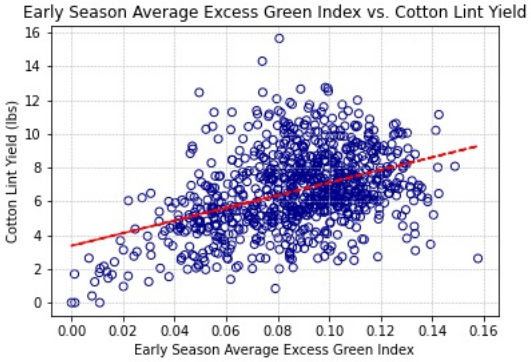
(c) Average Late Canopy Volume vs. Yield

Figure 4.11

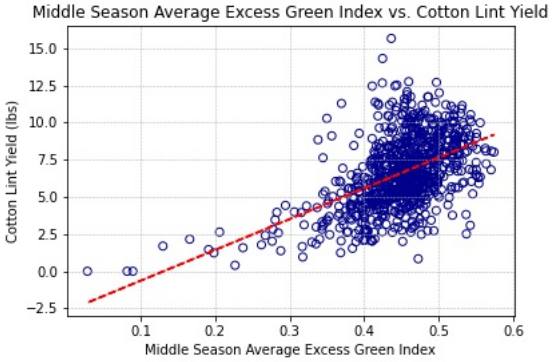
Scatter plots displaying the average (a) early season, (b) mid-season, and (c) late season canopy volume measured in m^3 versus Yield in pounds.

explain the lack of correlation to yield in the late part of the season.

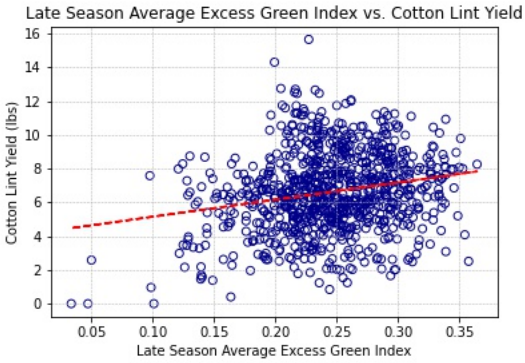
Figure 4.13 contains scatter plots of the early, late, and mid-season average NDVI against Yield. Similar to ExG, fewer flights were made throughout the season. The first seven weeks were averaged for Average Early NDVI ending on the same date as Early CH and CV. The next nine weeks were averaged for Average Mid NDVI and the last nine weeks for Average Late NDVI. There is a clear positive correlation of 0.44 between the variables with NDVI values in the early season plot. Average Mid NDVI has a correlation value of -0.55, and Average Late NDVI has a correlation value of -0.51. Similar to Late ExG the late season plants turn brown which may explain the negative correlation between Late NDVI and Yield.



(a) Average Early ExG vs. Yield



(b) Average Mid ExG vs. Yield

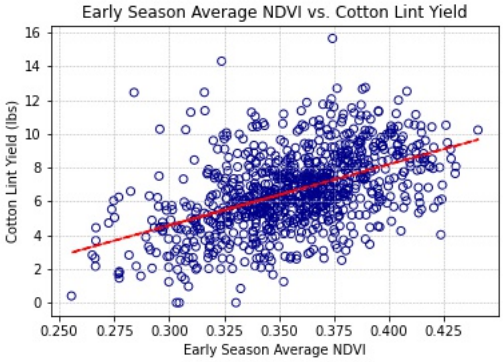


(c) Average Late ExG vs. Yield

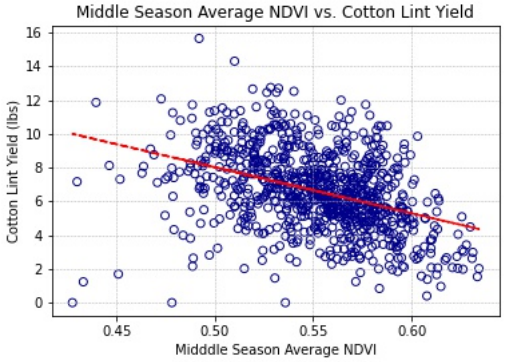
Figure 4.12

Scatter plots displaying the average (a) early season, (b) mid-season, and (c) late season excess greenness index versus Yield in pounds.

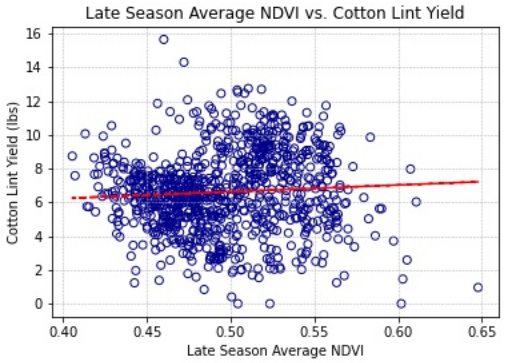
Figure 4.14 shows the early, mid, and late-season predictor variables against the Yield, irrigated plot Yield, and dry plot Yield. Several highly correlated variables are visible; we expect this because the predictor variables are indicators of plant health. We can see from the heat map that canopy height and canopy volume have moderate to high correlations. To investigate these relationships further, we provide the correlation matrices corresponding to the heat maps.
In Figure 4.15 we see the correlation values higher than 0.5 highlighted in green. We see the average Mid ExG of the entire field and Yield correlation coefficient of $\sim$0.51. Early CH and Mid CH have correlation coefficients with Yield of $\sim$0.61 and $\sim$0.59, respectively. Early CV, Mid CV, and Late CV have correlation scores to yield of $\sim$0.54, $\sim$0.57, and $\sim$ 0.52, respectively.



(a) Average Early NDVI vs. Yield



(b) Average Mid NDVI vs. Yield



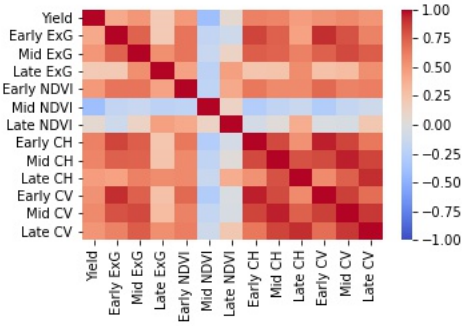
(c) Average Late NDVI vs. Yield

Figure 4.13

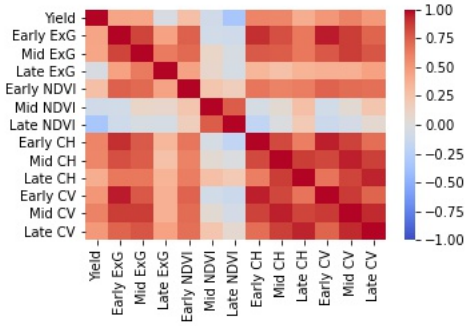
Scatter plots displaying the average (a) early season, (b) mid-season, and (c) late season normalized difference vegetation index versus Yield in pounds.

We also see a very high correlation of ∼0.95 between Early CV and Early CH, ∼0.95 between Mid CV and Mid CH, and ∼0.91 between Late CV and Late CH. These high correlations between plant height and volume are expected because healthy plants will grow taller and have more leaf area. Notice that Early CV and Early ExG also have a very high correlation of ∼0.91. Excess greenness index is an indicator of good plant health, so it follows that it would be highly correlated with plant volume.

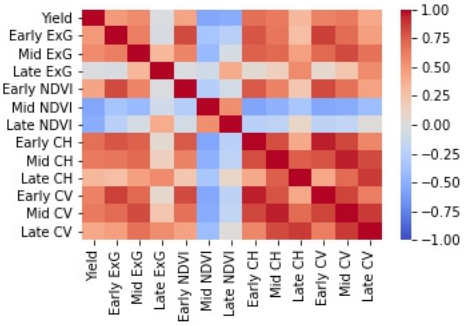
In 4.16 we see Early CH, Mid CH, Early CV, and Mid CV have high correlations to average irrigated plot Yield values of ∼ 0.60, ∼ 0.59, ∼0.51, and ∼0.60 respectively. There is a similar pattern to Figure 4.15 where there are very high correlations between Early CH and Early CV,



(a) Early, Mid, Late Season Predictor Variables vs. Yield



(b) Early, Mid, Late Season Predictor Variables vs. Irrigated Plots Yield



(c) Early, Mid, Late Season Predictor Variables vs. Dry Plots Yield

Figure 4.14

Heat maps displaying the relationships of the early, middle, and late season predictor variables and yield.

Mid CH and Mid CV, and Late CH and Late CV of ∼0.95, ∼0.95, and ∼0.94, respectively. We also see a very high correlation of ∼0.92 between Mid CV and late CV, ∼0.91 between Early ExG and Early CH, and ∼0.96 between Early CV and Early ExG.

In Figure 4.17 we see high correlations between average dry plot Yield values and Mid ExG, Mid NDVI, Late NDVI, Early CH, Mid CH, Early CV, and Mid CV of ∼0.57, ∼-0.55, ∼-0.51, ∼0.68, ∼0.63, ∼0.61, and ∼0.64, respectively. There are very high correlations of ∼0.95 between Early CH and Early CV and ∼0.95 between Mid CH and Mid CV.

	Yield	Early ExG	Mid ExG	Late ExG	Early NDVI	Mid NDVI	Late NDVI	Early CH	Mid CH	Late CH	Early CV	Mid CV	Late CV
Yield	1	0.417948	0.513038	0.208364	0.493779	-0.393329	0.064905	0.609809	0.594086	0.481231	0.537024	0.57404	0.523
Early ExG	0.417948	1	0.74086	0.204028	0.664706	-0.17312	-0.103094	0.851048	0.743592	0.457884	0.91432	0.799876	0.605
Mid ExG	0.513038	0.74086	1	0.548724	0.665633	-0.145473	0.132032	0.754124	0.740884	0.615372	0.748183	0.835098	0.762
Late ExG	0.208364	0.204028	0.548724	1	0.450026	-0.212542	0.46455	0.239295	0.254842	0.558307	0.268159	0.298033	0.547
Early NDVI	0.493779	0.664706	0.665633	0.450026	1	-0.195339	0.41521	0.654637	0.574018	0.574545	0.714046	0.607398	0.621
Mid NDVI	-0.393329	-0.17312	-0.145473	-0.212542	-0.195339	1	0.126655	-0.266254	-0.189493	-0.134444	-0.280366	-0.165111	-0.1
Late NDVI	0.064905	-0.103094	0.132032	0.46455	0.41521	0.126655	1	-0.064389	0.023722	0.392772	-0.026208	-0.043673	0.22
Early CH	0.609809	0.851048	0.754124	0.239295	0.654637	-0.266254	-0.064389	1	0.830818	0.536684	0.949893	0.8482	0.651
Mid CH	0.594086	0.743592	0.740884	0.254842	0.574018	-0.189493	0.023722	0.830818	1	0.798361	0.824888	0.946857	0.846
Late CH	0.481231	0.457884	0.615372	0.558307	0.574545	-0.134444	0.392772	0.536684	0.798361	1	0.566551	0.744554	0.914
Early CV	0.537024	0.91432	0.748183	0.268159	0.714046	-0.280366	-0.026208	0.949893	0.824888	0.566551	1	0.866409	0.691
Mid CV	0.57404	0.799876	0.835098	0.298033	0.607398	-0.165111	-0.043673	0.8482	0.946857	0.744554	0.866409	1	0.888
Late CV	0.522886	0.604963	0.762069	0.547284	0.6205	-0.103575	0.220413	0.651164	0.845929	0.914115	0.691343	0.887876	1

Figure 4.15

Average of early, mid, and late season predictor variables and yield for entire field with correlations above 0.5 highlighted

	Yield	Early ExG	Mid ExG	Late ExG	Early NDVI	Mid NDVI	Late NDVI	Early CH	Mid CH	Late CH	Early CV	Mid CV	Late CV
Yield	1	0.444402	0.433002	-0.038145	0.279535	-0.09345	-0.329546	0.598134	0.589112	0.392585	0.508658	0.604469	0.495372
Early ExG	0.444402	1	0.848534	0.453268	0.747657	-0.083066	-0.096459	0.907501	0.812052	0.64875	0.959077	0.870108	0.72669
Mid ExG	0.433002	0.848534	1	0.638526	0.710997	0.107158	-0.040563	0.770705	0.761998	0.655495	0.775111	0.867829	0.788844
Late ExG	-0.038145	0.453268	0.638526	1	0.445603	0.086178	-0.048476	0.339665	0.271217	0.354116	0.375887	0.377631	0.4596
Early NDVI	0.279535	0.747657	0.710997	0.445603	1	0.240538	0.164517	0.660371	0.601782	0.629105	0.736398	0.696935	0.683708
Mid NDVI	-0.09345	-0.083066	0.107158	0.086178	0.240538	1	0.759528	-0.053778	0.052134	0.27806	-0.079772	0.05018	0.24107
Late NDVI	-0.329546	-0.096459	-0.040563	-0.048476	0.164517	0.759528	1	-0.183403	-0.018176	0.203821	-0.121045	-0.070178	0.069478
Early CH	0.598134	0.907501	0.770705	0.339665	0.660371	-0.053778	-0.183403	1	0.834582	0.620249	0.949266	0.855586	0.689694
Mid CH	0.589112	0.812052	0.761998	0.271217	0.601782	0.052134	-0.018176	0.834582	1	0.869681	0.841901	0.950622	0.864809
Late CH	0.392585	0.64875	0.655495	0.354116	0.629105	0.27806	0.203821	0.620249	0.869681	1	0.671207	0.849588	0.938254
Early CV	0.508658	0.959077	0.775111	0.375887	0.736398	-0.079772	-0.121045	0.949266	0.841901	0.671207	1	0.881291	0.732868
Mid CV	0.604469	0.870108	0.867829	0.377631	0.696935	0.05018	-0.070178	0.855586	0.950622	0.849588	0.881291	1	0.924223
Late CV	0.495372	0.72669	0.788844	0.4596	0.683708	0.24107	0.069478	0.689694	0.864809	0.938254	0.732868	0.924223	1

Figure 4.16

Average of early, mid, and late season predictor variables and Yield for the irrigated plots with correlations above 0.5 highlighted

4.2 Model Assessment and Selection

Data Cleaning and Specifications

This study is interested in predicting Yield with predictor variables collected throughout the growing season and with predictor variables collected over the first 67 days. The full season data matrix included 875 rows and 103 columns, including all ExG, NDVI, CV, and CH variables, Yield, and irrigation. The data matrix formed with data from the first 67 days included 875 rows and 44 columns included all ExG, NDVI, CV, and CH variables up to June 7^{th}, irrigation, and Yield. In both data subsets, each row in the data matrix represented a single row in each plot, with each plot having two rows. All models had their predictor variables standardized, the

	Yield	Early ExG	Mid ExG	Late ExG	Early NDVI	Mid NDVI	Late NDVI	Early CH	Mid CH	Late CH	Early CV	Mid CV	Late CV
Yield	1	0.49425	0.570111	-0.022649	0.4381	-0.554493	-0.509362	0.684811	0.634742	0.327931	0.60819	0.636534	0.423497
Early ExG	0.49425	1	0.622775	-0.035749	0.821484	-0.348622	-0.238012	0.784169	0.661025	0.281843	0.866528	0.701751	0.474877
Mid ExG	0.570111	0.622775	1	0.335677	0.599293	-0.42516	-0.04931	0.742714	0.711295	0.466753	0.715268	0.820688	0.673584
Late ExG	-0.022649	-0.035749	0.335677	1	-0.018057	-0.086914	0.405989	0.048265	0.158742	0.566341	0.0736	0.238361	0.557581
Early NDVI	0.4381	0.821484	0.599293	-0.018057	1	-0.257991	-0.077415	0.773354	0.605435	0.23328	0.827703	0.673979	0.451208
Mid NDVI	-0.554493	-0.348622	-0.42516	-0.086914	-0.257991	1	0.551247	-0.555413	-0.485687	-0.339991	-0.542727	-0.522459	-0.39844
Late NDVI	-0.509362	-0.238012	-0.04931	0.405989	-0.077415	0.551247	1	-0.240642	-0.20707	0.108808	-0.211713	-0.183483	0.020133
Early CH	0.684811	0.784169	0.742714	0.048265	0.773354	-0.555413	-0.240642	1	0.819258	0.420157	0.949212	0.839192	0.583362
Mid CH	0.634742	0.661025	0.711295	0.158742	0.605435	-0.485687	-0.20707	0.819258	1	0.752687	0.793515	0.947966	0.824548
Late CH	0.327931	0.281843	0.466753	0.566341	0.23328	-0.339991	0.108808	0.420157	0.752687	1	0.42247	0.700545	0.879692
Early CV	0.60819	0.866528	0.715268	0.0736	0.827703	-0.542727	-0.211713	0.949212	0.793515	0.42247	1	0.847696	0.621754
Mid CV	0.636534	0.701751	0.820688	0.238361	0.673979	-0.522459	-0.183483	0.839192	0.947966	0.700545	0.847696	1	0.878438
Late CV	0.423497	0.474877	0.673584	0.557581	0.451208	-0.39844	0.020133	0.583362	0.824548	0.879692	0.621754	0.878438	1

Figure 4.17

Average of early, mid, and late season predictor variables and yield for the dry plots with correlations above 0.5 highlighted in green and below -0.5 highlighted in red

response variable was left in its original form. Any small negative values in the data matrices were set to zero. We chose to zero these values because CV and CH cannot be negative, and a negative ExG or NDVI value indicates no plant vegetation. Negative values only occurred in the first few weeks for the variables ExG and NDVI. Dummy variables were created for the irrigation variable, with 0 representing a dry plot and 1 representing an irrigated plot. Both subsets of the original data matrix were used to build four regression models, including Ridge regression, LASSO regression, PCR, and PLSR. A 10-fold cross-validation was performed to select the number of components used in PCR and PLSR and the LASSO and Ridge regression hyperparameters. All models were trained with 80% of the data and tested with 20% of the data. We compare each model's mean squared error and R-squared score for selection. We also use the MATLAB application Regression Learner to train various machine learning models and simple statistical models then compare them to our models.

Ridge Regression Performance

Ridge regression requires us to find an appropriate λ to shrink the coefficients the predictor variables that contribute little to the response variable. To find a λ that shrinks the coefficients efficiently a 10-fold cross validation was used. In Figure 4.18 we see that when lambda increases, the coefficients go towards zero. The coefficients will never reach exactly zero in Ridge regression but can get very small. The cross-validation resulted in $\lambda \approx 0.5749$. The

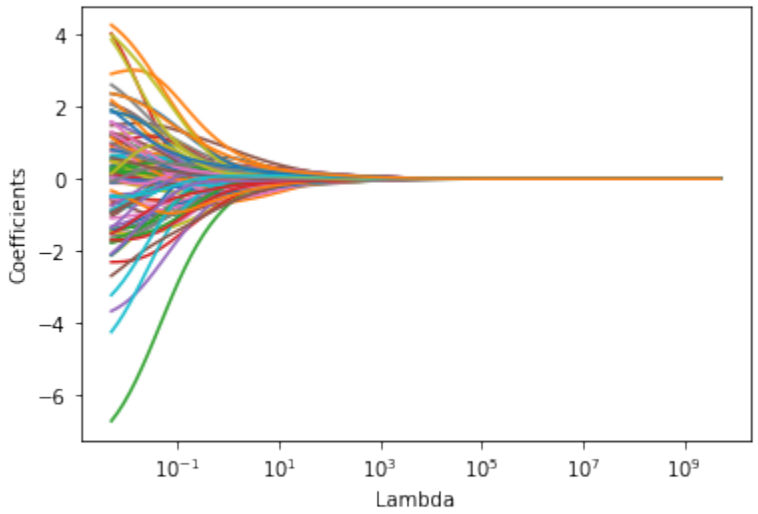


Figure 4.18

Visualization of the behavior of the coefficients in Ridge regression as λ gets very large.

model was trained with the given λ and the training data. The model was then tested with the testing set resulting in an MSE of $\sim$1.32 and an R-squared value of $\sim$0.73.

A second Ridge regression model was trained on the data matrix containing the time series variables up to June 7^{th}, irrigation, and Yield. Data were split into 80% training, 20% testing sets, and a 10-fold cross-validation was performed to find an appropriate λ. A plot of the behavior of the coefficients as λ gets very large is shown in Figure 4.19.

The cross-validation resulted in $\lambda \approx 0.0153$. The model was trained with the given λ and the training data set. The model was then tested with the testing set resulting in an MSE of $\sim$1.61 and an R-squared value of $\sim$0.67.

LASSO Regression Performance

LASSO regression is similar to Ridge regression, except that the shrinkage term λ can shrink coefficients to zero, effectively creating a subset of the variables in the data set. A 10-fold cross-validation was performed to find an appropriate shrinkage term; the cross-validation resulted in $\lambda \approx 0.00195$. A visualization of the behavior of the coefficients as λ gets very large is shown in Figure 4.20. In comparison to Ridge regression, the coefficients in LASSO shrink

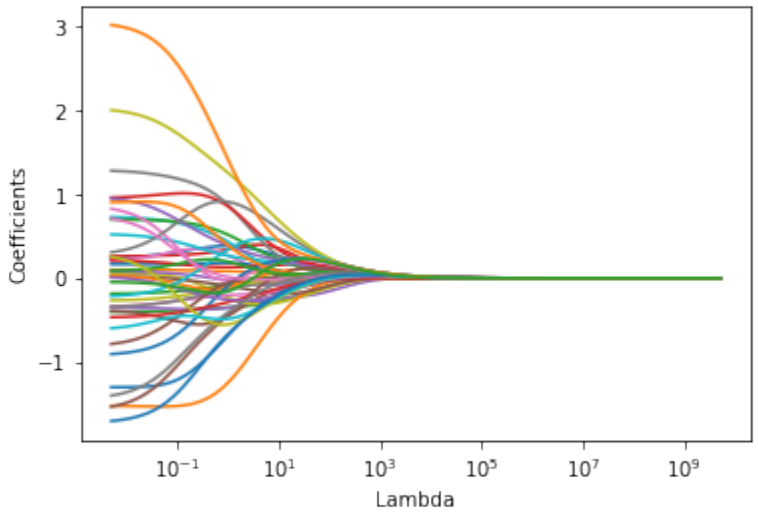


Figure 4.19

Visualization of the behavior of coefficients in Ridge regression (67 days after planting subset) as λ becomes very large.

much faster, as we can see from Figure 4.20. We trained the model with the λ found in the

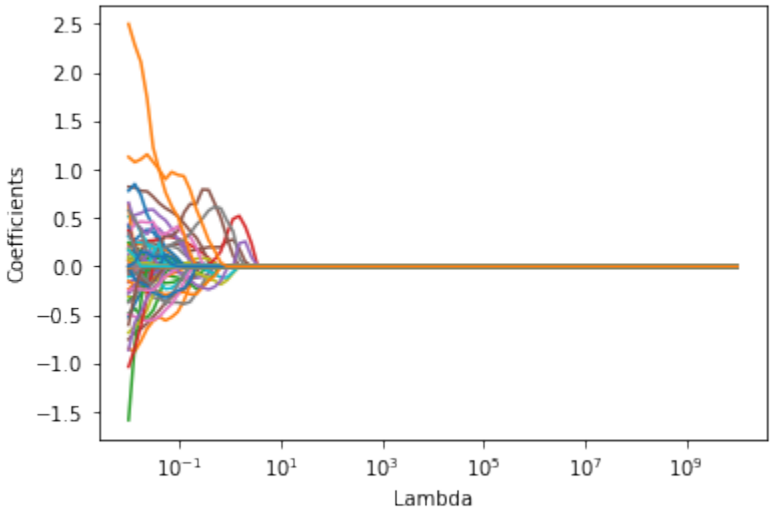


Figure 4.20

Visualization of the behavior of coefficients in LASSO regression as λ becomes very large.

cross-validation and the training set. Then we validated the model with the testing set resulting in an MSE of $\sim$0.94 and an R-squared score of $\sim$0.81.

A second LASSO regression model was trained on the 67 days after planting subset with the same methods. The resulting λ from the cross-validation was $\sim$0.00158. A visualization of the behavior of the coefficients as λ gets very large is shown in Figure 4.21 After training the model with the training set, we validated the model with the testing set. Validation on the testing set resulted in an MSE of $\sim$1.60 and an R-squared value of $\sim$ 0.67.

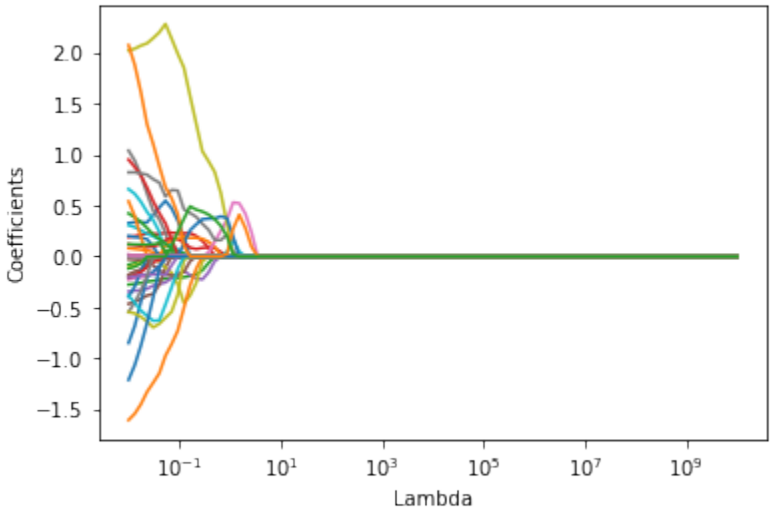


Figure 4.21

Visualization of the behavior of coefficients in LASSO regression as λ becomes very large (67 days after planting subset).

Principal Component Analysis

In this section, we perform a principal component analysis on our data set. The variables selected from the original data set include all NDVI, ExG, CH, and CV variables and the categorical variable irrigation. The irrigation predictor variable was transformed using dummy variables where a 0 indicated a dry plot and a 1 indicated an irrigated plot. After standardizing our data set and isolating the predictor variables, we performed PCA as detailed in Chapter II. In the scree plot displayed in Figure 4.22, the percentage of explained variance is visualized for the first 13 components. After PC12, every PC explains less than 1% of variability.

In the scree plot displayed in Figure 4.22, the percentage of explained variance is visualized for

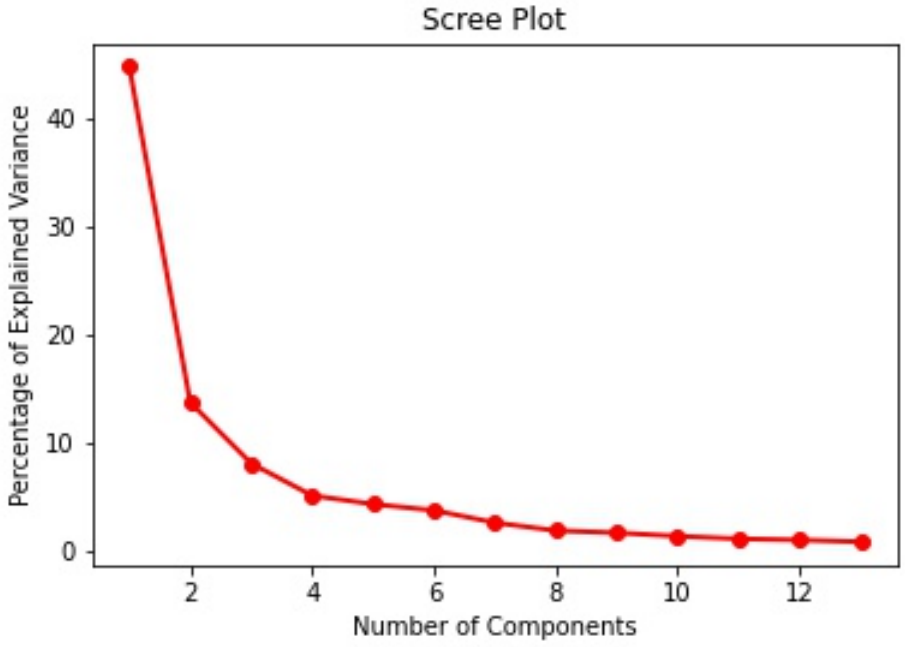


Figure 4.22
Scree Plot for the first 10 principal components.

the first 13 components. After PC12, every PC explains less than 1% of variability.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
%Explained Variance	44.837	13.793	8.150	5.174	4.382	3.802	2.662	1.929	1.728	1.414	1.174	1.070	0.902

Table 4.3
Explained variance of first 13 principal components.

The percent of explained variance of the first 13 principal components are given in Table 4.3. The first principal component explains approximately 44.837% of variability in the data set and the first six PC explain approximately 80.138% of the cumulative variability in the data.
In Table 4.4 we see that many of the loadings are low, indicating that most of the predictor variables only make a small contribution to the principal component. In the fifth PC, we see that NDVI0412 has a loading of $\sim$0.3125, NDVI0607 has a loading of $\sim$ 0.2998, NDVI0614 has a loading of $\sim$0.2243, and NDVI0620 has a loading of $\sim$0.3042. All other loadings in all six principal components are very low.
From Table 4.5 we can see that PC6 had a loading of $\sim$0.3231 for ch0412 and a loading of

Variable	PC1	PC2	PC3	PC4	PC5	PC6
NDVI0412	-0.013661	0.009241	0.017175	-0.062199	0.312456	0.100506
NDVI0422	-0.028244	0.170269	0.005931	0.067177	0.097929	0.101165
NDVI0427	-0.025228	0.167851	-0.017412	0.042291	0.107284	0.139544
NDVI0506	-0.085936	0.103961	0.066458	-0.036195	0.022517	-0.075958
NDVI0516	-0.113276	0.025053	0.066675	0.138802	0.064939	-0.103511
NDVI0520	-0.112338	0.033124	0.074504	0.094009	0.046193	-0.140027
NDVI0531	-0.044403	0.116311	-0.080385	0.024889	0.086362	-0.189237
NDVI0607	-0.014672	-0.126396	-0.096354	-0.018678	0.299776	-0.059469
NDVI0614	0.003426	-0.184386	-0.11401	-0.002429	0.224279	-0.084237
NDVI0620	0.042279	-0.005259	-0.054561	-0.014841	0.304231	-0.168804
NDVI0623	0.019206	-0.153984	-0.181609	0.06875	0.186986	0.006321
NDVI0627	0.04347	-0.084703	-0.151602	0.002679	0.189349	-0.056617
NDVI0716	0.049657	-0.001248	-0.103891	0.164119	0.274201	0.073676
NDVI0721	0.003852	0.024294	-0.155376	0.124479	0.255033	-0.038605
NDVI0725	-0.007856	0.108373	-0.134791	0.18775	0.27333	0.026766
NDVI0802	0.033153	0.120694	-0.152041	0.048277	0.08854	0.056037
NDVI0808	0.02269	0.024273	-0.160572	0.05487	0.015122	-0.056759
NDVI0812	-0.016117	0.231117	-0.094821	0.12628	0.06773	0.003723
NDVI0818	-0.016812	0.166336	-0.135101	0.116233	0.132156	-0.01169

Table 4.4
Loadings of NDVI with values higher than 0.2 and lower than -0.2 highlighted

$\sim$0.3763 for ch0415. All other loadings were below 0.3 and higher than -0.3.

In Table 4.6, we see all loadings in PC1, PC5, and PC6 are lower than 0.2 and higher than $\sim$-0.2 PC2 has moderate loadings of $\sim$0.2300, $\sim$0.1966, and $\sim$0.2491 with ExG0407, ExG0808, and ExG0812. PC3 has moderate loadings of $\sim$-0.2077, $\sim$-.2038, and $\sim$-.2022 with variables ExG0719, ExG0721, and ExG0725. PC4 has loadings of $\sim$-0.2026, $\sim$-0.2374, and $\sim$-0.2694 for variables ExG0719, ExG0721, and ExG0728.

Table 4.7 PC2 has a loading of $\sim$0.2250 for cv0812 and $\sim$0.2491 for irrigation. PC4 has a loading of $\sim$-0.2211 for cv0412. PC6 has a loading of $\sim$0.3250 and $\sim$0.3460 with cv0412 and cv0415, respectively. In the next section, we perform a principal component regression where these principal components will be used as our predictor variables.

Principal Component Regression Performance

Before fitting our data to the PCR, we separated the data into 80% training and 20% testing sets. A 10-fold cross-validation was performed with the training data to select an appropriate number of principal components based on MSE. Figure 4.23 shows the mean squared error score using the cross-validation for the first components starting with the intercept (no PCs) and then adding

Variable	PC1	PC2	PC3	PC4	PC5	PC6
ch0412	0.002222	-0.079278	-0.072875	-0.213767	0.018964	0.323138
ch0415	0.003453	0.044936	-0.063173	-0.102186	0.13503	0.376318
ch0427	-0.012818	0.124367	0.09567	-0.14365	-0.011145	0.105541
ch0506	-0.114903	0.074381	0.132032	-0.07481	0.071389	0.068531
ch0516	-0.128906	0.023865	0.128409	-0.040193	0.054537	0.062155
ch0520	-0.13026	0.031336	0.122466	-0.024114	0.055902	0.078387
ch0523	-0.130211	0.032746	0.116878	0.002985	0.046421	0.079377
ch0527	-0.134831	0.013574	0.101846	-0.003123	0.034573	0.042925
ch0531	-0.136242	-0.004449	0.092667	-0.005936	0.038898	0.032116
ch0602	-0.137336	0.020488	0.055555	-0.022848	0.051041	0.051663
ch0607	-0.097151	0.121588	-0.055266	0.114279	-0.107181	0.160183
ch0614	-0.120993	0.09033	0.039102	0.054382	-0.004451	-0.057127
ch0617	-0.137777	-0.034124	0.051915	0.040236	-0.006902	0.023592
ch0620	-0.134525	-0.022941	0.050104	0.060898	-0.045929	-0.011672
ch0623	-0.123878	-0.103159	-0.015544	0.103553	-0.023069	0.111077
ch0627	-0.118843	-0.106145	-0.04049	0.123968	-0.040141	0.099783
ch0630	-0.114142	-0.094056	-0.037451	0.141113	-0.045487	0.127297
ch0708	-0.129037	-0.072992	-0.043472	0.075463	-0.050381	0.005901
ch0713	-0.116318	-0.132945	-0.078658	0.057416	-0.022565	0.020705
ch0716	-0.115983	-0.129448	-0.088564	0.066119	-0.03123	0.007509
ch0719	-0.125069	-0.075605	-0.100901	0.087014	-0.054427	0.033902
ch0721	-0.119767	-0.107516	-0.090656	0.081116	-0.039775	0.005643
ch0725	-0.119857	-0.079348	-0.086968	0.102413	-0.055702	0.004343
ch0728	-0.120368	-0.04373	-0.138685	0.086595	-0.079327	0.030936
ch0802	-0.118124	-0.064767	-0.122432	0.119824	-0.053839	0.020055
ch0808	-0.061108	0.103621	-0.194603	0.065785	-0.154	0.085603
ch0812	-0.027192	0.233876	-0.072491	0.108994	-0.009469	0.056186

Table 4.5
Loadings of CH with values higher than 0.2 and lower than -0.2 highlighted

one component after each iteration. The first 6 PCs were selected based on the MSE values becoming nearly constant after six components and the percent of variance they explain. These PCs were then used as the predictor variables in a multiple linear regression. The model had an MSE of $\sim$2.83 and an R-squared score of $\sim$ 0.48 on the testing data.

The same process was used to build a model based on the subset of predictor variables up to 67 days after planting. Figure 4.24 shows the results of 10-fold the cross validation used to estimate an appropriate number of principal components. Based on the cross validation we take 8 PCs are our predictor variables. The linear regression was fit with the first 8 PCs of the training data then the first 8 PCs of the test data were used to validate the model. Like the first PCR, this model performed poorly on the testing data with an MSE of $\sim$2.88 and an R-squared value of $\sim$0.47. Although PCR can lower the dimension of the data, there may be a few reasons why the model predicts poorly with the testing data. PCR does not consider the response variable when calculating PCs, and the selected PCs may not be the most useful. There could be different

Variable	PC1	PC2	PC3	PC4	PC5	PC6
ExG0407	-0.019654	0.229963	-0.001777	0.081601	-0.065479	-0.127854
ExG0412	-0.036249	0.124023	0.022723	0.018776	0.002932	-0.128769
ExG0415	-0.0288	0.032408	0.010388	0.106603	-0.081359	-0.022134
ExG0427	-0.039512	0.064044	0.078586	0.057402	0.03663	0.023555
ExG0506	-0.088391	-0.025274	0.043975	-0.046638	-0.019109	-0.106172
ExG0516	-0.119332	0.014696	0.158544	-0.03845	0.013525	-0.060844
ExG0520	-0.128267	0.026601	0.115507	0.011325	0.05369	-0.037114
ExG0523	-0.119735	0.011826	0.127468	-0.050547	0.071544	-0.030509
ExG0527	-0.119023	-0.057051	0.116073	-0.098892	0.068481	-0.066911
ExG0531	-0.107953	-0.048085	0.095806	-0.002842	0.193029	-0.112521
ExG0602	-0.112525	0.010816	0.057262	-0.014667	0.109173	-0.132963
ExG0607	-0.122515	-0.02214	0.035544	-0.051547	0.084452	-0.128367
ExG0614	-0.128882	0.011873	-0.027528	-0.069789	0.032607	-0.097507
ExG0617	-0.120359	0.056699	-0.054666	-0.041758	0.082849	0.019935
ExG0620	-0.108057	0.071322	-0.04174	-0.120111	0.071577	-0.00584
ExG0623	-0.111219	0.080159	-0.069603	-0.102994	0.037706	0.021328
ExG0627	-0.092613	0.114773	-0.067719	-0.16453	0.022703	-0.019159
ExG0630	-0.090826	0.107529	-0.037922	-0.124151	-0.016492	-0.027009
ExG0713	-0.086934	0.057805	-0.09954	-0.245462	0.063655	-0.106598
ExG0716	-0.07933	-0.049657	-0.128026	-0.245645	0.084678	-0.072486
ExG0719	-0.06792	-0.000785	-0.207665	-0.202566	-0.064246	-0.039105
ExG0721	-0.066482	0.016113	-0.203832	-0.237367	-0.054454	-0.064865
ExG0725	-0.058483	0.036478	-0.202221	-0.194949	-0.055978	-0.116153
ExG0728	-0.051354	0.074164	-0.165059	-0.269439	-0.042596	-0.111426
ExG0802	-0.032262	0.165766	-0.151056	-0.132411	-0.083538	-0.046714
ExG0808	-0.038697	0.19661	-0.123757	-0.082696	-0.100839	-0.098047
ExG0812	-0.023419	0.249145	-0.043536	0.034689	-0.010391	-0.015156
ExG0818	0.012417	0.111976	-0.021717	0.074905	-0.050266	0.066584

Table 4.6
Loadings of ExG with values higher than 0.2 and lower than -0.2 highlighted

variables that may better fit our data.

Partial Least Squares Regression Performance

After standardizing the data, we separated the predictor and response variables and performed a 10-fold cross-validation to search for an appropriate lambda. After each iteration, a component was added and given an MSE score. Figure 4.25 displays the number of components, or latent variables, versus the MSE of the model. We chose 13 components and then trained the model on the training data, resulting in an MSE of $\sim$1.00 and an R-squared value of $\sim$0.80.

The same process was used to build a model for the data set containing the first 67 days of the times series variables, irrigation, and Yield. The results of the 10-fold cross-validation are shown in 4.26. We selected 21 components based on the MSE slightly increasing after this point. The model was trained with the 21 components and the training data. We validated the model with the test data and attained an MSE of $\sim$1.54 and an R-squared value of $\sim$0.70.

Variable	PC1	PC2	PC3	PC4	PC5	PC6
cv0412	0.003806	-0.069625	-0.0773	-0.221102	0.011842	0.325009
cv0415	0.004149	0.028873	-0.076758	-0.07939	0.145928	0.346047
cv0427	-0.016998	0.089835	0.090884	-0.148444	-0.013584	0.084628
cv0506	-0.099646	0.092156	0.119539	-0.116951	0.05924	0.05228
cv0516	-0.12837	0.04431	0.109914	-0.021852	0.054389	0.076561
cv0520	-0.1293	0.020261	0.126996	-0.03181	0.05126	0.070836
cv0523	-0.128102	0.043308	0.124451	-0.017105	0.034591	0.04772
cv0527	-0.135678	0.028126	0.096364	-0.00936	0.024264	0.001213
cv0531	-0.138479	-0.006362	0.089547	0.000507	0.039562	0.003101
cv0602	-0.141174	0.011552	0.062367	-0.022017	0.054172	0.016338
cv0607	-0.117274	0.100181	-0.034342	0.087171	-0.081642	0.116293
cv0614	-0.134708	0.071122	0.027895	0.017074	-0.013072	-0.069394
cv0617	-0.143468	-0.036119	0.028597	-0.0033	-0.010422	-0.000636
cv0620	-0.14279	-0.033723	0.0302	-0.000378	-0.02672	-0.032392
cv0623	-0.135603	-0.085909	-0.022056	0.034153	-0.010101	0.050096
cv0627	-0.133803	-0.090361	-0.034306	0.043385	-0.021201	0.051379
cv0630	-0.132657	-0.080415	-0.019256	0.058198	-0.023665	0.0722
cv0708	-0.139029	-0.060276	-0.046024	0.000118	-0.02925	-0.050025
cv0713	-0.132671	-0.089874	-0.076901	0.005087	-0.013007	-0.024239
cv0716	-0.128847	-0.111519	-0.068082	-0.002273	-0.012823	-0.03147
cv0719	-0.134197	-0.073084	-0.093066	0.020232	-0.043879	0.001236
cv0721	-0.13405	-0.079884	-0.078571	0.020374	-0.033072	-0.023847
cv0725	-0.134712	-0.038522	-0.101502	0.030499	-0.059304	-0.017776
cv0728	-0.129431	-0.06678	-0.108099	0.018089	-0.055094	-0.002656
cv0802	-0.129041	-0.031378	-0.121261	0.039506	-0.059876	-0.004034
cv0808	-0.091639	0.085867	-0.180251	0.052203	-0.132973	0.049197
cv0812	-0.04613	0.225042	-0.057174	0.117358	0.022469	0.021087
irrigation_Yes	-0.014897	0.249095	-0.032081	0.105499	0.004155	0.035976

Table 4.7

Loadings of CV and irrigation with values higher than 0.2 and lower than -0.2 highlighted

Comparing the Models

In Table 4.8 we see the mean squared errors and R-squared values of Ridge, LASSO, PCR, PLSR, and several machine learning models where both the response and predictor variables were standardized. Ridge regression performed moderately well with an R-squared of $\sim$0.73, and PCR performed poorly with an R-squared of $\sim$0.48. We elaborate on why the PCR model may have performed poorly in the Chapter V. We see that LASSO and PLSR had similar R-squared values of $\sim$0.81 and $\sim$0.80, respectively. The two models also had similar MSE of $\sim$0.94 for LASSO and $\sim$1.00 for PLSR. Both LASSO and PLSR outperformed the machine learning models ran in MATLAB shown In Table 4.8 we see that LASSO and PLSR outperformed a MLR, support vector regression, and several neural networks. The LASSO regression was favored over the PLSR because of the interpretability of the LASSO regression.

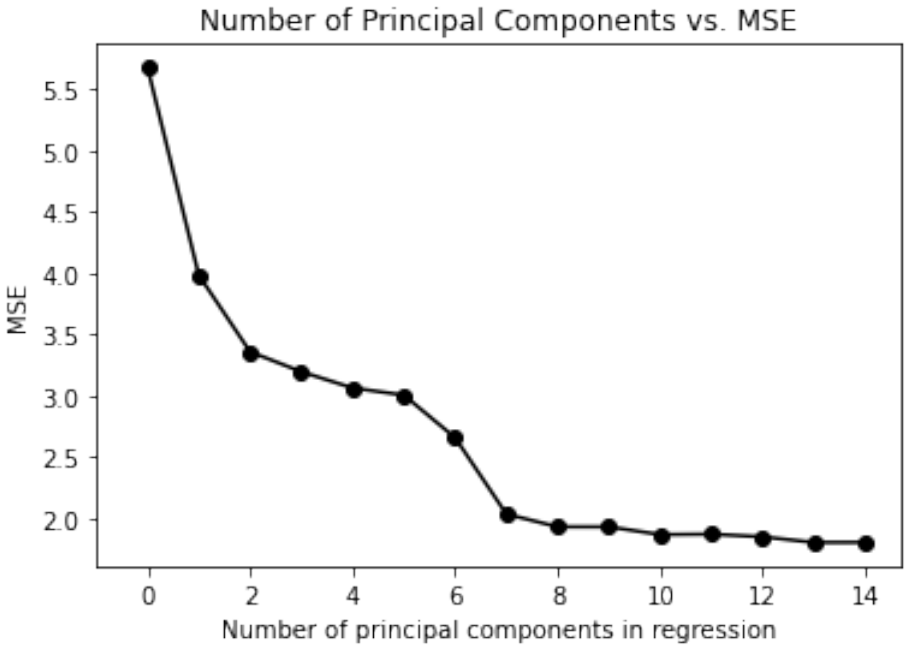


Figure 4.23
Number of Principal Components vs. MSE

Variables are still in their original form in LASSO regression, but some have been removed by receiving a coefficient of zero. The remaining variables are easy to interpret when used to predict with new data, so the LASSO regression model was selected as the best model.

	MSE(Test)	R-Squared(Test)	MSE(Train)	R-Squared(Train)
LASSO Regression	0.94	0.81	1.10	0.81
PLSR	1.00	0.80	1.14	0.80
Ridge Regression	1.32	0.73	1.56	0.73
PCR	2.83	0.48	2.57	0.54
MLR	1.41	0.75	1.44	0.74
Regression Tree	2.89	0.50	2.78	0.50
SVM	1.27	0.78	1.36	0.76
Wide Neural Network	1.29	0.77	1.70	0.69
Medium Neural Network	1.73	0.70	2.43	0.56
Narrow Neural Network	2.69	0.53	3.37	0.40
Bilayered Neural Network	3.03	0.47	2.97	0.47
Trilayered Neural Network	2.32	0.60	2.64	0.53

Table 4.8

R-squared and MSE of the four models trained on full season data compared to several regression models

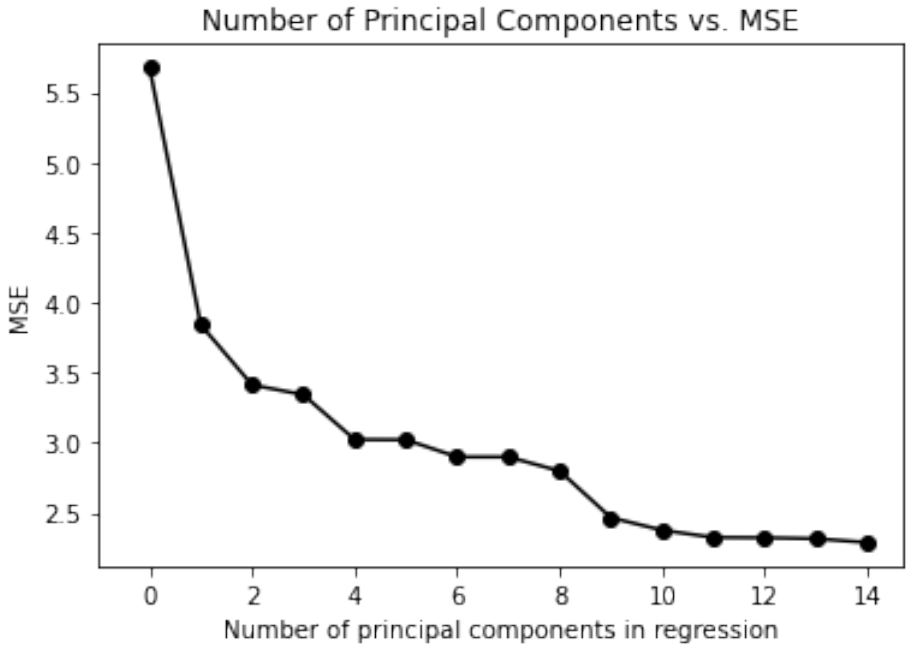


Figure 4.24
Number of principal components vs. MSE for PCR

The MSE and R-squared scores of the second group of models trained and tested on the data from the first 67 days after planting are listed in Table 4.9. The Ridge and LASSO regression performed nearly the same, with MSE values of $\sim$1.61 and $\sim$1.60 and R-squared values of $\sim$0.67 and $\sim$0.67, respectively. The PCR performed poorly on this data set MSE value of $\sim$2.88 and an R-squared value of $\sim$0.47. PLSR performed the best on this data subset with the lowest MSE score of $\sim$1.54 and highest R-squared score of $\sim$0.70. In Table 4.9 we see the PLSR outperformed a MLR, support vector regression, and several neural networks.

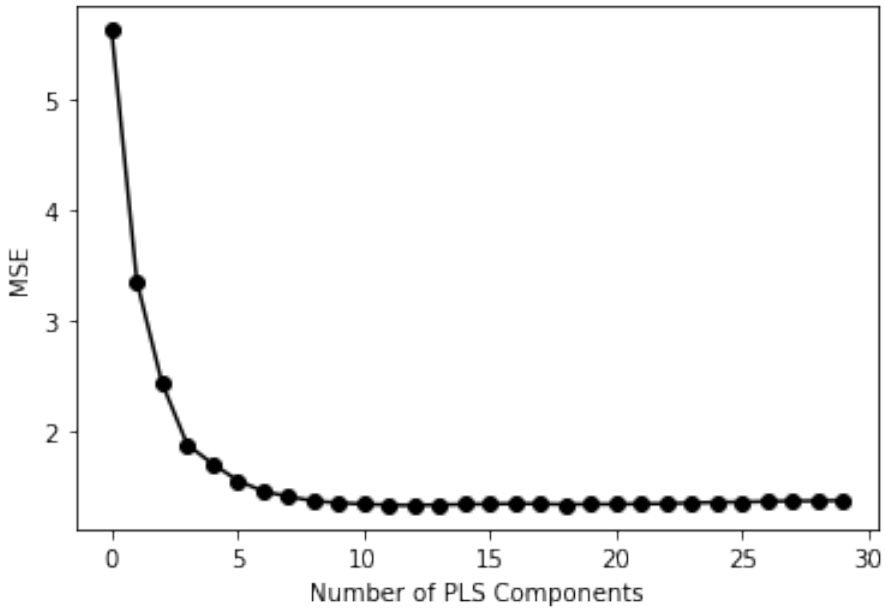


Figure 4.25
Plot of MSE calculated during cross validation and number of PLS components

	MSE(Test)	R-Squared(Test)	MSE(Train)	R-Squared(Train)
LASSO Regression	1.60	0.67	1.80	0.67
PLSR	1.54	0.70	1.79	0.69
Ridge Regression	1.61	0.67	1.84	0.68
PCR	2.88	0.47	2.69	0.52
MLR	2.48	0.51	1.81	0.69
Regression Tree	3.77	0.25	3.03	0.48
SVM	2.37	0.53	1.83	0.68
Wide Neural Network	3.42	0.32	3.01	0.48
Medium Neural Network	5.74	-0.14	6.23	-0.08
Narrow Neural Network	3.35	0.14	4.05	0.30
Bilayered Neural Network	5.73	-0.14	4.15	0.28
Trilayered Neural Network	4.53	0.10	6.01	-0.04

Table 4.9

R-squared and MSE of the four models trained on the 67 days after planting data subset compared to several regression models

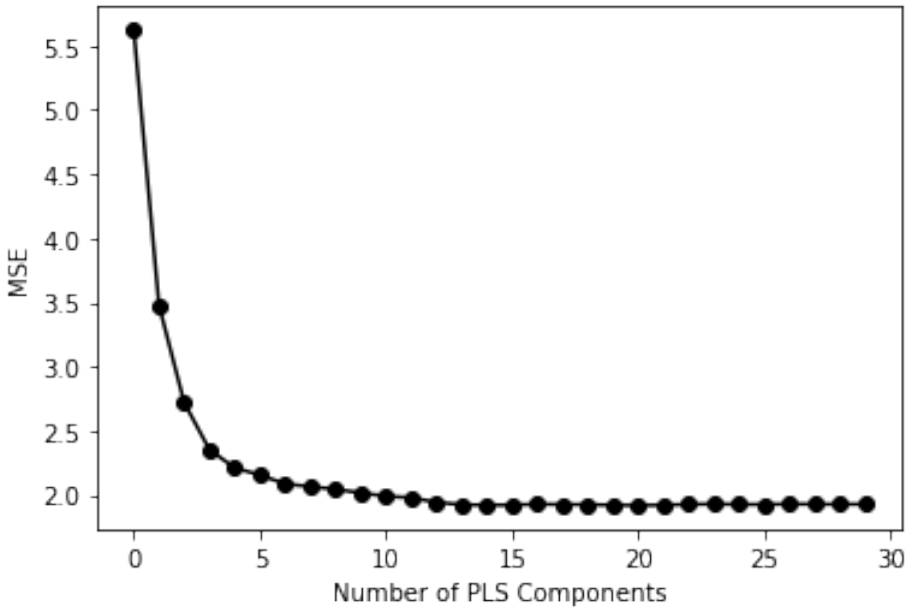


Figure 4.26

Plot of MSE calculated during cross validation and number of PLS components (67 days after planting subset)

CHAPTER V: DISCUSSION AND FUTURE RESEARCH

This thesis had the objective of predicting cotton yield based on data calculated from UAV collected imagery of an experimental cotton field at the Texas A&M AgriLife Research Center in Corpus Christi, Texas, USA (27° 46' 57.08" N, 97° 33' 40.94" W). We explored models appropriate for relatively small data sets that could make predictions comparable to those of neural networks seen in current literature. The performance of four models was compared, namely PCR, PLSR, Ridge regression, and LASSO regression. The LASSO model selected for the full season of time series variables (ExG, NDVI, CH, and CV) and irrigation. The LASSO regression predicted cotton yield per plot with an R-squared value of $\sim$0.81 and an MSE of $\sim$0.94. The LASSO was selected over the PLSR (MSE $\approx$ 1.00, R-squared $\approx$ 0.80) because of its simplicity and comparable performance. The PLSR model uses latent variables to make predictions while the LASSO creates a subset of the original data set by shrinking some of the coefficients to zero based on its penalty term, the ℓ^1 norm, making LASSO easier to interpret when predicting with future data.

The PLSR model was selected to predict cotton yield using time series variables (ExG, NDVI, CH, and CV) for the first 67 days after planting and irrigation. The PLSR model performed best out of the four models, with an R-squared score of $\sim$0.70 and an MSE value of $\sim$1.54. PLSR is similar to PCR; both are calculated using latent variables, but, unlike PCR, PLSR considers the correlations between the set of predictor variables and the response variable. The inclusion of the response when training provides the model with more information about the system, making it a more robust model. We found that the PCR model performed poorly for both sets of data; we believe that performance would improve if the principal components should be rotated to obtain more accurate explanations of variance. PCR performance may also be improved by excluding variables such as canopy height while including highly correlated variables like canopy volume. Outliers may have also affected the accuracy of the PCR model. A larger data set would enable us to use neural networks that were not appropriate for the size of our current data set. More

effective hyperparameter tuning methods, such as quadratic optimization may improve the accuracy of our models. Reorganization of the early, middle, and late season subsets to coincide with the growth periods of cotton may improve results. Predictor variables such as average percentage of soil nutrients and soil moisture per row may improve the outcome of future models as these factors directly affect the growth and health of plants.

REFERENCES

[1] Herve Abdi. Partial least squares regression and projection on latent structure regression ´ (pls regression). *Wiley interdisciplinary reviews: computational statistics*, 2(1):97–106, 2010.

[2] National Aeronautics and Science Mission Directorate Space Administration. Reflected near-infrared waves, 2010. Retrieved: 2022-04-4, from NASA science website: http://science.nasa.gov/ems/08 nearinfraredwaves.

[3] Akash Ashapure, Jinha Jung, Anjin Chang, Sungchan Oh, Junho Yeom, Murilo Maeda, Andrea Maeda, Nothabo Dube, Juan Landivar, Steve Hague, et al. Developing a machine learning based cotton yield estimation framework using multi-temporal UAS data. ISPRS *Journal of Photogrammetry and Remote Sensing*, 169:180–194, 2020.

[4] Akash Ashapure, Jinha Jung, Junho Yeom, Anjin Chang, Murilo Maeda, Andrea Maeda, and Juan Landivar. A novel framework to detect conventional tillage and no-tillage cropping system effect on cotton growth and development using multi-temporal UAS data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152:49–64, 2019.

[5] Akash Ashapure, Sungchan Oh, Thiago G Marconi, Anjin Chang, Jinha Jung, Juan Landivar, and Juan Enciso. Unmanned aerial system based tomato yield estimation using machine learning. In *Autonomous Air and Ground Sensing Systems for Agricultural Optimization and Phenotyping IV*, volume 11008, page 110080O. International Society for Optics and Photonics, 2019.

[6] Del Deterling. How a cotton plant grows, Nov 2011. Retrieved: 2022-03-4, from: https://cottonbugs.tamu.edu/development-and-growth-monitoring-of-the-cotton-plant/.

[7] Jennifer Dorsett. Cotton production figures for 2019 released, May 2020. https://texasfarmbureau.org/cotton-production-figures-for-2019-released/.

[8] Ercan Gokgoz and Abdulhamit Subasi. Effect of multiscale pca de-noising on emg signal classification for diagnosis of neuromuscular disorders. *Journal of medical systems*, 38(4):1–10, 2014.

[9] Amir Haghverdi, Robert A Washington-Allen, and Brian G Leib. Prediction of cotton lint yield from phenology of crop indices using artificial neural networks. *Computers and Electronics in Agriculture*, 152:186–197, 2018.

[10] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[11] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[12] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[13] Jeff Jauregui. Principal component analysis with linear algebra. *Philadelphia: Penn Arts & Sciences*, 2012.

[14] Jinha Jung, Murilo Maeda, Anjin Chang, Mahendra Bhandari, Akash Ashapure, and Juan Landivar-Bowles. The potential of remote sensing and artificial intelligence as tools to improve the resilience of agriculture production systems. *Current Opinion in Biotechnology*, 70:15–22, 2021.

[15] Lakshmi Sirisha Kommareddi and Luo Sha. Machine learning using satellite remote sensing to predict agricultural yield of cash crops in usa. In *IRC-SET 2020*, pages 585–598. Springer, 2021.

[16] Bo Li, Xiangming Xu, Li Zhang, Jiwan Han, Chunsong Bian, Guangcun Li, Jiangang Liu,

and Liping Jin. Above-ground biomass estimation and yield prediction in potato by using uav-based rgb and hyperspectral imaging. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:161–172, 2020.

[17] Christopher Long. Nueces County, 2020. Updated: June 9, 2020, Retrieved from: https://www.tshaonline.org/handbook/entries/nueces-county.

[18] Andrea B Maeda, Leslie W Wells, Monica A Sheehan, and Jane K Dever. Stories from the greenhouse—a brief on cotton seed germination. *Plants*, 10(12):2807, 2021.

[19] Ali Masjedi, Jieqiong Zhao, Addie M Thompson, Kai-Wei Yang, John E Flatt, Melba M Crawford, David S Ebert, Mitchell R Tuinstra, Graeme Hammer, and Scott Chapman. Sorghum biomass prediction using UAV-based remote sensing data and crop model simulation. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 7719–7722. IEEE, 2018.

[20] Tahir Mehmood and Bilal Ahmed. The diversity in the applications of partial least squares: an overview. *Journal of Chemometrics*, 30(1):4–17, 2016.

[21] Mike Mei. Principal component analysis. *The University of Chicago: Chicago, IL, USA,* 2009.

[22] Leslie Meyer. Cotton sector at a glance, 2020. https://www.ers.usda.gov/topics/crops/cotton-wool/cotton-sector-at-a-glance/.

[23] Petteri Nevavuori, Nathaniel Narra, Petri Linna, and Tarmo Lipping. Crop yield prediction using multitemporal UAV data and spatio-temporal deep learning models. *Remote Sensing*, 12(23):4000, 2020.

[24] Petteri Nevavuori, Nathaniel Narra, and Tarmo Lipping. Crop yield prediction with deep convolutional neural networks. *Computers and electronics in agriculture*, 163:104859,

2019.

[25] Sungchan Oh, Anjin Chang, Akash Ashapure, Jinha Jung, Nothabo Dube, Murilo Maeda, Daniel Gonzalez, and Juan Landivar. Plant counting of cotton from UAS imagery using deep learning-based object detection framework. *Remote Sensing*, 12(18):2981, 2020.

[26] Derrick Oosterhuis, Tom Kerby, and Kater Hake. Leaf physiology and management. *Physiology Today*, May 1990.

[27] Rich Pang, Benjamin J Lansdell, and Adrienne L Fairhall. Dimensionality reduction in neuroscience. *Current Biology*, 26(14):R656–R660, 2016.

[28] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science,* 2(11):559–572, 1901.

[29] Dante M Pirouz. An overview of partial least squares, October 2006. Available at SSRN: https://ssrn.com/abstract=1631359.

[30] Glen Lorin Ritchie, Craig W Bednarz, Philip H Jost, and Steve M Brown. Cotton growth and development. Technical report, University of Georgia, 2007.

[31] John RC Robinson and Dean A McCorkle. Trends and prospects for Texas cotton. *Texas; Connecting the Old West to the New East,* 2006.

[32] Tayyaba Shaheen, Nabila Tabbasam, Muhammad Atif Iqbal, Muhammad Ashraf, Yusuf Zafar, Andrew H Paterson, et al. Cotton genetic resources. a review. *Agronomy for sustainable development*, 32(2):419–432, 2012.

[33] Amir Shakeel, Irfan Talib, Muhammad Rashid, Asif Saeed, Khurram Ziaf, and M Farrukh Saleem. Genetic diversity among upland cotton genotypes for quality and yield related traits. *Pak. J. Agric. Sci*, 52(1):73–77, 2015.

[34] Jonathon Shlens. A tutorial on principal component analysis. *CoRR*, abs/1404.1100, 2014. http://arxiv.org/abs/1404.1100.

[35] Ewout W Steyerberg. Overfitting and optimism in prediction models. In *Clinical Prediction Models*, pages 95–112. Springer, 2019.

[36] Danilo Tedesco-Oliveira, Rouverson Pereira da Silva, Walter Maldonado Jr, and Cristiano Zerbato. Convolutional neural networks in predicting cotton yield from images of commercial fields. *Computers and Electronics in Agriculture*, 171:105307, 2020.

[37] David M Woebbecke, George E Meyer, Kenneth Von Bargen, and David A Mortensen. Color indices for weed identification under various soil, residue, and lighting conditions. *Transactions of the ASAE*, 38(1):259–269, 1995.

[38] Junho Yeom, Jinha Jung, Anjin Chang, Akash Ashapure, Murilo Maeda, Andrea Maeda, and Juan Landivar. Comparison of vegetation indices derived from UAV data for differentiation of tillage effects in agriculture. *Remote Sensing*, 11(13):1548, 2019.

APPENDIX A

PYTHON CODE

GitHub repository link: https://github.com/BiancaBrianne/Cotton-Yield-Estimation.git