A MACHINE LEARNING REGRESSION MODEL FOR PREDICTING THE OCCURRENCE OF AFLATOXIN IN PEANUT FIELDS USING ENVIRONMENTAL AND FIELD-DEPENDENT VARIABLES

by

SUNAAB KUKAL

(Under the Direction of George Vellidis and Thirimachos Bourlai)

ABSTRACT

Machine learning techniques were used to develop a mathematical model that predicts the occurrence of aflatoxin in peanut grown in rainfed fields. Data for training and testing the model were collected from three farmers' fields in southern Georgia, USA. The study found that soil temperature, soil texture and meteorological parameters like solar radiation, air temperature, and vapor pressure deficit (VPD) can be used to explain the variation in aflatoxin production using a random forest regression model paired with feature engineering methods (recursive feature elimination) and cross-validation techniques, with a coefficient of determination of 27% and an RMSE of 0.65.

INDEX WORDS: Machine Learning, Random Forest, Aflatoxin, Peanut, Decision Support Tools

A MACHINE LEARNING REGRESSION MODEL FOR PREDICTING THE OCCURRENCE OF AFLATOXIN IN PEANUT FIELDS USING ENVIRONMENTAL AND FIELD-DEPENDENT VARIABLES

By

SUNAAB KUKAL

B.Tech., Punjab Agricultural University, India, 2021

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2024

A MACHINE LEARNING REGRESSION MODEL FOR PREDICTING THE OCCURRENCE OF AFLATOXIN IN PEANUT FIELDS USING ENVIRONMENTAL AND FIELD-DEPENDENT VARIABLES

by

SUNAAB KUKAL

Major Professors:	George Vellidis
	Thirimachos Bourlai
Committee:	Robert Kemerait
	Alicia Peduzzi
	Cristiane Pilon

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2024

ACKNOWLEDGEMENTS

I am grateful to my major advisor, Dr. George Vellidis, for the time, effort, and wisdom he has invested in me and this thesis. His professionalism and prioritization of my work have made a significant impact on my academic and personal growth. For all the meetings, discussions, and critiques that have led to this thesis's completion, I owe a profound debt of gratitude to Dr. Vellidis. He has been patient with me despite my shortcomings at times. I would like to appreciate the role of my committee (Cristiane Pilon, Thirimachos Bourlai, Alicia Peduzzi, Robert Kemerait) in creating an interdisciplinary environment of learning through my M.S. and lending their respective expertise towards my research. I would like to thank my parents, Mrs. Charanjeet Kaur and Dr. Surinder Singh Kukal, who believed in me and always cared for me, irrelevant to the distance between us.

This study was made possible with funding from Georgia Peanut Commission. I would like to thank Matthew Gruver, Morgan Sysskind and Sarah Maktabi for their unconditional help and support through countless days of field work. I would also like to mention my friends Orestis Giannopoulos, Giannis Gallios, Mohammad Usman Khalid, Ved Prakash and Kamal Dhillon who flooded me with their love and endless support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

Mycotoxins are toxic secondary metabolites produced by fungi that are harmful to human and animal health. These compounds may contaminate foods and feeds, either in the field or postharvest, creating a major food safety risk. Because of their ability to cause illness or fatalities at very low concentrations of $\mu g\ kg^{-1}$ (commonly referred to as parts per billion or ppb), mycotoxin levels on foods and feeds are strictly regulated by food safety agencies around the world. The Food and Agriculture Organization of the United Nations has estimated that 25% of the world's crops are affected by mycotoxins each year, with losses of ≈1 billion tons of food products annually (Schmale & Munkvold, 2009). In the US and Canada, direct loss estimates to mycotoxins (without considering human health impacts) vary between $0.5 and $5 billion per year (Schmale & Munkvold, 2009). Global and regional increases in temperature and the increased focus on consumer health and low mycotoxin tolerances will exacerbate the mycotoxin problem in the future.

1.2 Aflatoxin in peanut and its impacts

Peanut (*Arachis hypogea* L.), widely recognized as one of the most extensively cultivated legumes on a global scale, occupies a pivotal position in agriculture due to its widespread production, consumption, and international trade. However, despite its

economic significance and nutritional value, peanut face a persistent challenge linked to fungal contamination that has far-reaching implications for both human health and market access. This challenge arises from *Aspergillus flavus* and *Aspergillus parasiticus* (hereafter referred to collectively as *A. flavus*) which are saprophytic soil fungi that infect and contaminate preharvest and postharvest seed crops with the carcinogenic secondary metabolite aflatoxin (Klich, 2007; Amaike and Keller, 2011).

Aflatoxins, characterized by their potent carcinogenic properties, pose serious health risks to humans and animals alike. These compounds have been associated with an array of adverse health effects, including liver damage, stunted growth, and weakened immune systems (Kumar et al., 2017) and are classified as Group 1 carcinogens by the International Agency for Research on Cancer (IARC) (IARC, 2002). The resulting ailment, aflatoxicosis, presents a significant public health concern in regions where peanuts constitute a staple food. What magnifies this issue is the resilience of aflatoxins to conventional processing methods and even high temperatures. This persistence underscores the challenge of effectively eliminating aflatoxins from contaminated crops. As a result, the development and implementation of rigorous strategies for proper harvesting, storage, drying, and processing are pivotal in reducing aflatoxin levels in peanut, thereby mitigating the associated health risks. An additional layer of complexity is introduced by the fact that aflatoxin contamination is not limited to peanut alone. Aflatoxins can also be found in other major crops, including almonds, maize, wheat, and cottonseed, further amplifying the concern for food safety and security. As a result, the need to effectively manage aflatoxin contamination has implications beyond the peanut industry and extends to the broader context of global food supply chains.

Aflatoxin-contaminated peanuts pose a significant food safety risk. The presence of aflatoxins in peanuts can lead to the contamination of peanut-derived products such as peanut butter, candy bars, and other products, amplifying the health hazards (Bennett & Klich, 2003). Peanuts with aflatoxin contamination face severe economic losses in the agricultural and food processing sectors. Contaminated crops are often rejected or downgraded, resulting in financial setbacks for farmers (Dorner, 2008). Losses due to aflatoxin contamination have led to an economic loss of somewhere between $500 million and $1.5 billion every year. These losses consist of reduced prices of low-grade produce, pre-harvest (in-field) and post- harvest (storage) aflatoxin management, and research costs to control aflatoxin production (Robens and Cardwell 2003). Lamb and Sternitzke (2001) reported that in the southeastern United States (U.S.), aflatoxin in peanuts cost farmers and shellers on average, from 1993 to 1996, $70/ha. Aflatoxins can accelerate spoilage and reduce the shelf-life of peanuts. They contribute to off-flavors, odors, and visible mold growth, making the peanuts unappealing and unsafe for consumption (Dorner, 2009).

One of the major impacts of quality degradation due to aflatoxin production in peanut is the loss of market for peanut exports to the European Union (E.U.). E.U. enforces a stringent low-tolerance policy for aflatoxin contamination in imported peanut leading to significant economic repercussions for peanut-exporting countries like the U.S., which stands fourth globally in peanut production. This scenario underscores the urgency of addressing aflatoxin contamination to safeguard international trade relationships, economic stability, and public health. In response to the multifaceted challenge of aflatoxin contamination, researchers have embarked on comprehensive studies to unravel the complexities of aflatoxin presence throughout the peanut production cycle. These studies

encompass various stages of production, spanning from the characterization of crop traits before planting to in-field monitoring of critical parameters like soil type and water stress, pre-harvest risk analysis, and post-harvest storage management strategies (Torres et al., 2014).

Aflatoxin is a mycotoxin of particular interest in Georgia because peanut is one of that state's most important crops. In 2022, Georgia peanut farmers harvested 275,000 ha producing 725,000 tonnes of peanuts which was approximately 52% of the entire U.S. production (Georgia Peanut Commission, 2023). In some years, losses to aflatoxin are significant. For example, the peanut industry-wide loss to aflatoxin in Georgia in 2019 was estimated at 24% of the crop. One shelling company, Premium Peanut LLC, reported their losses to aflatoxin at $150 million that year.

To mitigate this issue, aflatoxin must be studied qualitatively and quantitatively. This study is focused on predicting aflatoxin production in peanut pods in Georgia fields by correlating easily sensed field parameters to aflatoxin concentrations. Using statistical and machine learning regression methods, this study evaluates the complex relationships between field and crop attributes such soil, water, canopy light reflection, and weather and aflatoxin production in peanuts.

1.3 Literature Review

1.3.1 Pathways for aflatoxin contamination from *A. flavus*

A. flavus are ubiquitous soil fungi responsible for aflatoxin contamination in various crops, posing significant risks to food safety and human health. Recent research studies have delved into understanding the biology, ecology, and control strategies for *A. flavus* and aflatoxin contamination. A study by Cary et al. (2000) focused on the genetic

regulation of aflatoxin biosynthesis in *A. flavus*. This study identified key genes involved in the production of aflatoxins, unveiling potential targets for controlling aflatoxin contamination. Another study by Cotty and Jaime-Garcia (2007) examined the diversity of *A. flavus* strains and their population dynamics in agricultural fields. This research focused on the importance of understanding the genetic diversity of *A. flavus* for developing effective management strategies. In the context of controlling aflatoxin contamination, a comprehensive review by Abbas et al. (2017) highlighted various biological, chemical, and physical methods used to mitigate aflatoxin contamination in crops. This review provided insights into the practical approaches for reducing aflatoxin contamination in crops.

The colonization of peanut kernels by of *A. flavus* has been a concern for farmers, buyers, and the industry for a very long time, due to the potential contamination of peanut kernels with aflatoxins, highly toxic and carcinogenic secondary metabolites produced by this fungus. Researchers have conducted numerous studies to understand the dynamics and factors influencing *A. flavus* infestation in peanut crops. Despite the ubiquitous presence of these fungi in soil, extensive invasion by *A. flavus* and contamination of the peanut crop with aflatoxin occurs primarily when the plant is subjected to drought stress and high soil temperatures (Hill et al. 1983; Sanders et al. 1985; Clevenger et al. 2016). Peanuts grown under drought stress may also be predisposed to subsequent aflatoxin contamination during harvest, handling, and storage.

A study by Fountain et al. (2015) investigated the impact of irrigation practices on *A. flavus* contamination in peanuts. The research reported that irrigation management plays a crucial role in reducing *A. flavus* infestation and subsequent aflatoxin contamination, emphasizing the importance of proper water management practices in peanut cultivation.

Additionally, a study by Probst et al. (2014) studied the influence of environmental factors such as temperature and moisture on aflatoxin production by *A. flavus*. This study reported the significance of studying regional variations in environmental factors to control or predict in-field aflatoxin production Another paper by Chang et al. (2017) explored the genetic diversity of *A. flavus* populations in peanuts across different regions. This study's findings suggested that understanding the genetic variation of *A. flavus* strains is essential for developing region-specific strategies to combat infestation and aflatoxin contamination in peanut crops.

The work of Isleib et al. (2013) focused on breeding efforts to develop peanut varieties with enhanced resistance to aflatoxin contamination from *A. flavus*. Breeding programs targeting genetic resistance have the potential to provide long-term solutions to mitigate infestation and ensure peanut crop safety. In addition to genetic approaches, a study by Shrestha et al. (2017) delved into the biocontrol of *A. flavus* in peanuts using non-toxigenic strains of the fungus. This research demonstrated the promise of biological control strategies in reducing *A. flavus* infestation and aflatoxin contamination in peanut crops. In conclusion, research on *A. flavus* and aflatoxin contamination is being carried out through various studies in different disciplines, including genetic regulation, population dynamics, control methods, and environmental influences. These studies collectively contribute to our understanding of the fungi and provide a foundation for developing effective strategies to mitigate aflatoxin contamination in crops.

Previous work also showed that soil type and crop rotation affect colonization of peanut kernels by *A. flavus* (Griffin and Garren 1974; Abbas et al. 2004) and that populations of the fungus in soil exhibit a moderate degree of spatial structure (Abbas et

al. 2004). Vellidis et al. (2006) found that aflatoxin contamination was spatially aggregated within a rainfed peanut field. In that study, aflatoxin levels measured on peanut kernels sampled systematically throughout the field were used to create an interpolated map of the aflatoxin distribution in the field, which showed several areas with high concentrations or "hotspots" (Figure 1.1a). Maps of normalized difference vegetation index (NDVI) (Figure 1.1b) and ECa (Figure 1.1c) of the field were also developed. NDVI is an indicator of plant biomass and plant vigor and was assessed with a tractor-mounted multispectral sensor. ECa is a surrogate for soil texture with low ECa values indicating sandy soils and high ECa values indicating soils with higher clay content. Geostatistical analysis indicated spatial correlation among aflatoxin concentration, NDVI, and ECa. Generally, higher levels of aflatoxin were observed in areas with lower ECa and lower NDVI. This indicates that aflatoxin is more prevalent in areas that may have experienced physiological water stress due to the lower water-holding capacity and the drought-prone characteristic of sandier soils. Thus, the study by Vellidis et al. (2006) documented that aflatoxin is spatially clustered within peanut fields and may be correlated with easily measurable field parameters. While insightful, the study's limited scope prompted the need for further research to validate these findings across different contexts. Subsequent studies expanded on these findings. Li et al. (2018) conducted a cross-regional study involving various fields and growth cycles, confirming the spatial distribution trend observed by Vellidis et al. (Li et al., 2018). This research also underscored the influence of climate variables on aflatoxin concentrations. Clevenger et al. (2020) used hyperspectral imaging to differentiate aflatoxin levels in individual peanut seeds (Clevenger et al., 2020). In the field of technology-driven research, Leite et al. (2021) demonstrated the potential of machine-

learning techniques for aflatoxin detection in peanuts using near-infrared spectroscopy. Their study showcased the accuracy and efficiency of machine-learning algorithms in trying to explain the variation in aflatoxin production based on remotely and proximally sensed field parameters (Leite et al., 2021).

1.3.2 Microclimatic conditions

As reported by Probst et al. (2014), microclimatic conditions in peanut fields play a crucial role in influencing aflatoxin production in peanuts. Understanding the impact of microclimate on aflatoxin production is essential for developing effective strategies to mitigate contamination. Several key factors within the microclimate contribute to aflatoxin production are (i) Air and soil temperature: Elevated temperatures promote the growth of *A. flavus* and enhance aflatoxin production. Hot and dry conditions can create a favorable environment for the development of *A. flavus* and toxin synthesis. Studies such as Dorner (2009) have shown a positive correlation between high air and soil temperatures and increased aflatoxin contamination.; (ii) Moisture: Adequate soil moisture during the growing season is essential for peanut development but must be managed carefully. Excessive moisture can lead to fungal invasion, especially during pod maturation. Probst and Callicott (2012) emphasize the importance of monitoring moisture levels to prevent aflatoxin contamination.; (iii) Relative Humidity: High relative humidity during pod maturation can encourage *A. flavus* infection and aflatoxin product ion. Proper aeration and drying of peanuts are critical to reducing humidity-related risks (Eshelli et al., 2017).; (iv) Drought Stress: Prolonged drought stress can weaken peanut plants, making them more susceptible to A. flavus invasion. Strategies to mitigate drought stress, as discussed by Wang et al. (2016), can indirectly reduce aflatoxin contamination. Proper temperature,

moisture, and humidity management, along with strategies to combat drought stress, are essential for minimizing the risk of aflatoxin contamination in peanuts.

1.3.3 Physical damage

Insects, nematodes, and other soil biota may damage the peanut pod shell creating a pathway for the fungus to reach the kernels (Diener, 1987). The lesser cornstalk borer is an example of an insect that penetrates peanut pod shells and exposes peanut kernels, making them drier and thus more favorable for the fungus to parasitize. Hill et al. (1983) reported that sound mature kernels (SMKs) did not show any signs of contamination and peanuts that were damaged due to insect infestation had higher concentrations of aflatoxin. A study by Tillman et al. (2016) investigated the feeding habits of *Blissus insularis*, a common burrower bug species in peanuts and found that these insects preferentially feed on immature peanut pods. While this feeding behavior may cause physical damage to the pods, it does not directly implicate them in aflatoxin contamination. While there is some evidence of physical damage caused by burrower bugs to peanut pods, more research is needed to establish a direct link between burrower bug infestations and increased aflatoxin contamination in peanuts. The impact may vary depending on factors like bug population density, environmental conditions, and the presence of other factors promoting fungal growth.

1.3.4 Post-contamination management

Aflatoxin management in peanut after contamination in the field is a critical step to ensure food safety and minimize economic losses. Contaminated peanuts can be sorted and graded using specialized equipment to remove visibly moldy or damaged kernels. This reduces the overall aflatoxin content (Dorner, 2008). Accurate testing methods, such as

high-performance liquid chromatography (HPLC) or enzyme-linked immunosorbent assays (ELISA), are employed to quantify aflatoxin levels. This helps determine the extent of contamination and whether the peanuts meet regulatory limits (Dorner, 2008). Certain mechanical and chemical methods can be applied to reduce aflatoxin levels in peanuts. These include blanching, roasting, and surface treatments that can reduce aflatoxin concentrations (Klich, 2007). Proper storage practices are crucial to prevent further aflatoxin production and contamination. Maintaining low moisture levels and storing peanuts in well-ventilated conditions can help minimize fungal growth (Sobolev & Dorner, 2009). Research explores the use of aflatoxin-reducing agents, such as specific microorganisms or additives, to detoxify aflatoxins in peanuts (Rasooli et al., 2010). These approaches show promise in reducing aflatoxin levels. Beneficial microorganisms, such as atoxigenic strains of *A. flavus*, can be applied to peanut fields to competitively exclude aflatoxin-producing strains (Atehnkeng et al., 2008). Managing aflatoxin contamination in peanuts involves a combination of physical, chemical, and biological methods to reduce aflatoxin levels and ensure that contaminated peanuts meet safety standards. These approaches aim to salvage the economic value of contaminated peanuts while safeguarding human and animal health.

Collectively, these studies described so far signify the progression of research to estimate aflatoxin production in peanuts. They provide insights into the spatial distribution of aflatoxin concentrations, validate the correlation with various parameters, and introduce innovative techniques for accurate estimation. As the field continues to evolve, incorporating interdisciplinary approaches and cutting-edge technologies such as machine

learning models, our ability to effectively estimate and mitigate aflatoxin production in peanuts is advancing, ultimately safeguarding global food safety.

1.3.5 Aflatoxin measurement techniques:

The Enzyme-Linked Immunosorbent Assay (ELISA) and VICAM (Waters-VICAM, Massachusetts) methods are commonly employed to detect aflatoxins in peanuts, crucial due to the carcinogenic nature of aflatoxins. The ELISA method utilizes antibodies specific to aflatoxins to quantify their presence in a sample. It involves the binding of aflatoxins to these antibodies, which are then detected using an enzyme-conjugated secondary antibody. The resulting color change, upon substrate addition, quantitatively indicates the aflatoxin level. This method is favored for its rapidness, sensitivity, and specificity, making it suitable for screening large numbers of samples efficiently (Hafez, 2021).

The VICAM method involves using immunoaffinity columns to clean up and concentrate aflatoxins from peanut extracts. This method is highly selective and sensitive, allowing for the precise measurement of aflatoxins. After passing through these columns, aflatoxins are eluted and quantified using high-performance liquid chromatography (HPLC), providing accurate results. The VICAM method is particularly noted for its precision and ability to handle complex food matrices like peanuts. Both methods are integral in maintaining food safety, with ELISA providing a fast, initial screening option, and VICAM offering high precision in detailed analysis. Each method has its advantages and is chosen based on the specific requirements of the testing scenario, such as the number of samples, required sensitivity, and available resources.

1.3.6 Predicting aflatoxin contamination in the field with mathematical models

As described earlier, studies show that aflatoxin concentrations are dependent on a variety of environmental factors. Current knowledge indicates that the most important factors are soil moisture and soil temperature. Craufurd et al. (2006), Diener et al. (1987), and Hill et al. (1983) all reported that *A. flavus* produced increased aflatoxin concentrations in conditions with drier soil moisture and elevated soil temperatures. These studies concluded that these correlations can be used to develop decision-support tools (DSTs) for in-field aflatoxin risk prediction. In a study by Chauhan et al. (2010), a basic model was developed that was dependent on soil temperature in the form of an aflatoxin risk index (ARI) that showed a direct correlation with observed aflatoxin concentration measurements. Figure 1.2 shows a curvilinear relationship between observed values of aflatoxin concentrations and the estimated ARI from this study. (Diener et al. 1987) also found that elevated temperatures and dry soil moisture conditions contributed to higher aflatoxin production in peanut. These studies indicated that the spatial distribution of aflatoxin may be explained using the spatial and temporal variation of field parameters such as soil characteristics, weather variables, plant growth and pathological parameters.

Mathematical simulation tools which are sometimes referred to as DSTs are increasingly used in agriculture to make data-driven decisions. DSTs range from simple to complex. A complex DST is the Decision Support System for Agrotechnology Transfer (DSSAT) which includes dynamic crop growth simulation models for over 42 crops including peanuts (Boote 2019). DSSAT provides a framework to conduct research for understanding the effect of various management practices and changes in environmental conditions on the growth and yield of crops by evaluating the relative response of different scenarios (Hoogenboom et al. 2004).

1.3.7 Machine learning models

Emerging DSTs in agriculture are machine learning regression models (MLRs) that quantify the relationship between a dependent variable and a set of independent variables by producing a mathematical equation that best represents the relationship that can be used for the prediction of the dependent variable (Kahane 2008). There are multiple types of MLRs such as linear regression models, multivariate regression models, logistic regression models, etc. One of these types of MLRs is decision trees. Decision trees are predictors that use branching based on the analogy of that of a tree, with a random classifier (a function of either single input variable-univariate splitting or multiple input variables-multivariate splitting) at each node leading to a decision (output variable) at the end of a series of random classifiers (Rokach and Maimon, 2005). A decision tree's attributes are its number of nodes and depth which contribute to the extent of generalization error of the decision tree. When multiple decision trees are developed at the same time for regression analysis of a dataset and the results are combined to obtain one regression model, also called an ensemble model – a model that combines results from various methods to obtain an optimal regression model, it is called a random forest regression model. Breiman (2001) reported that combined results from decision trees performed competitively when compared to other ensemble methods and performed better under some circumstances. The combining of the decision trees is done by averaging or taking a weighted average depending on the methodology and datasets used (Biau and Scornet, 2016). Breiman (2001) discusses that random forests can be used for regression by replacing class labels at each node with numerical values resulting in numerical output values. The accuracy requirement of random forests is to have a low correlation between low-error trees and

residuals. This can be done by increasing the number of features which leads to a slow increase in correlation and hence, low generalization error.

Random forest regression models have been used in various studies focusing on the development of regression models to quantify the dependency of dependent variables on independent variables (Hoffman, Kemanian, and Forest, 2018); (Burton and Kemanian, 2022). To adapt the model to the dataset to obtain the least generalization error, attributes of the model (namely, number of trees- ntree, number of variables sampled at each split - mtry, and number of predictors) were altered. Hoffman (2018) found the optimal number of trees by increasing the number of trees until the error was observed to be not decreasing. mtry was set by hit and trial until the best results were found. Finally, the number of predictors was found by analyzing the variable importance and partial dependency plots and then eliminating the predictors that showed no contribution to the model or provided useful information about the prediction. Variable importance is a measure of the impact of the predictor variables on the estimation of the response (Strobl et al. 2008). Hoffman (2018) estimated the variable importance of each predictor which helped to visualize the effect of each yield predictor used in the model and reported the partial dependency plots that depicted the effect of various predictors on the response. These plots were then used to prioritize the parameters that showed heavy effect on the prediction of the response and to eliminate or reduce the effect of the parameters that showed little or no contribution to the prediction of the response from the test dataset.

1.3.8 Machine learning regression models applied to agriculture

MLR models have found numerous applications in agriculture. These models leverage large datasets and advanced algorithms to make predictions, optimize resource

allocation, and improve yield. In recent years, the application of MLR models in agriculture has gained significant attention. Farmers and agricultural scientists are turning to these data-oriented technologies to enhance their decision-making processes and achieve more sustainable and efficient crop production.

MLR models, such as Linear Regression, Decision Trees, and Random Forests, have the potential to provide valuable insights and predictions based on historical data and real-time information. One of the key applications of machine learning in agriculture is precision agriculture. This approach involves the use of various sensors and data collection methods to monitor soil conditions, weather patterns, and crop health. These data are then integrated into MLR models, which can predict crop yields and recommend precise resource allocation. For example, a study by Klompenburg et al. (2020) demonstrated the use of MLR models to predict crop yields based on soil moisture, temperature, and historical yield data. MLR models are also useful in optimizing resource management. Through the analysis of historical data and environmental factors, these models can recommend the optimal use of water, fertilizers, and pesticides, reducing waste and environmental impact. A study by Abioye et al. (2020) explored the use of MLR models to predict optimal irrigation schedules, resulting in significant water savings and improved crop quality.

Detecting and managing pests and diseases is a crucial aspect of agriculture. MLR models can aid in the early detection of these issues by analyzing sensor data and historical patterns. In their study, Han et al. (2020) used regression models to predict the spread of a fungal disease in wheat crops. The early warnings provided by the model allowed farmers to take timely preventive measures. While MLR models have shown great promise in

agriculture, there are challenges such as the need for high-quality data, model interpretability, and scalability. Researchers continue to address these issues to make these models more accessible to farmers and agricultural professionals. However, the integration of machine learning regression models into agriculture has opened new possibilities for precision farming, resource management, and pest control.

1.4 Project goals

Studies done by Hill (1983), Vellidis (2006), and Chauhan (2010) are a gateway to more extensive research on the quantification of relationships between aflatoxin and environmental parameters that can function as indicators of aflatoxin production in peanut. However, the persistent challenge of aflatoxin contamination, driven by the presence of *A. flavus*, demands a multifaceted approach to address the problem. The overall goal of the proposed study was to address research gaps in the understanding of the environmental and field-dependent factors influencing the contamination of peanut kernels by *A. flavus* and to use this knowledge to develop a predictive model capable of identifying potential aflatoxin hotspots within peanut fields. A substantial collection of field parameters such as soil variables, plant physiological and pathological parameters, and weather variables are needed to train complex mathematical models.

An aflatoxin predictive model could revolutionize peanut cultivation by equipping growers with the tools necessary to proactively manage aflatoxin contamination. By enabling growers to know that conditions are conducive to aflatoxin contamination of their peanut crop prior to the harvest season, they could take targeted actions, such as segregating peanuts from identified high-risk areas, thereby preserving both the quality of

the harvest and the integrity of the food supply chain. Such a development could lead to safer food products, enhanced market access, and improved global food security.

1.5 Hypothesis and objectives

This study hypothesized that aflatoxin contamination in growing peanuts can be predicted by using an ensemble of easily measurable field and environmental parameters. The specific objectives used to test the hypothesis and achieve the project's goal were to:90

- Conduct on-farm field studies to collect data sets that can be used to train a predictive model.
- Develop and evaluate a machine learning model that accurately predicts the occurrence of aflatoxin in the field using environmental and field-dependent variables.

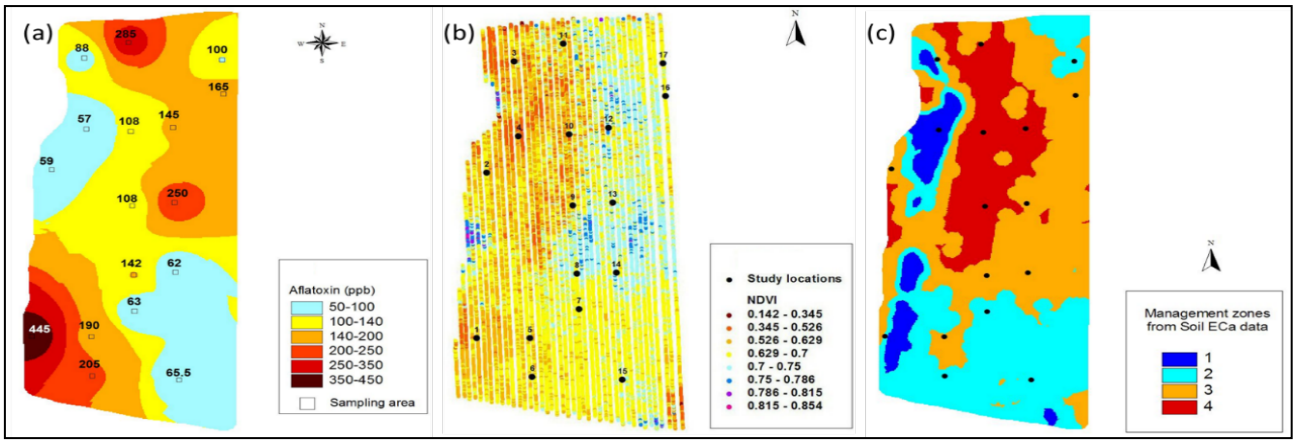

Figure 1. 1 Aflatoxin concentration map from data collected in a 14-ha peanut field in Tift Co., GA in 2006 (a); NDVI map from a Crop Circle reflectance sensor (b); and soil ECa map created with a Veris 3100 EC mapper (c). Blues and orange/red indicate higher clay and sand content, respectively.

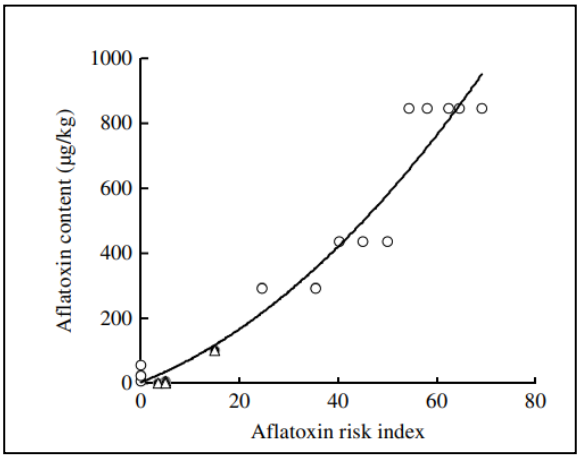

Figure 1. 2 Curvilinear relationship between aflatoxin concentrations and Aflatoxin Risk Index (ARI) (Diener et al. 1987)

CHAPTER 2

MATERIALS AND METHODS

The work reported here is a component of a larger ongoing project focusing on understanding and predicting the occurrence of aflatoxin contamination in rainfed peanut fields in southern Georgia, USA. The larger project began in 2022 but the work reported here focuses only on data collected during the 2023 growing season. However, for completeness and for cataloguing data for future use, procedures and some data are reported for both growing seasons.

2.1 Field data collection

For each year (2022 and 2023), three rainfed grower-managed fields were selected in the state of Georgia. In both years, two fields were in southcentral Georgia (SC) and one in southwestern Georgia (SW). Figure 2.1 shows the general location of the farmer fields in both years. The field names, locations, and size are summarized in Table 2.1. The 2023 fields are shown in Figures 2.2 – 2.4. Figure 2.5 shows a timeline of the field components of the study.

2.1.1 Soil ECa

Soil apparent electrical conductivity (ECa) is a measure of the soil's ability to conduct electricity and is affected by soil texture, soil moisture, soil salinity, and other factors. Soil ECa is often used as a surrogate for soil texture. Soil ECa was the first

measurement taken in all fields using a Veris 3100 (Veris Technologies - Salina, KS) instrument with RTK guidance. Data were collected continuously in 9 m parallel swaths. In the configuration used, the Veris 3100 instrument collected integrated values of soil ECa for 0-0.3m and 0-0.9m. Soil ECa and elevation data were then used in the Management Zone Analyst (MZA) software (Fridgen et al. 2004) to create soil ECa and elevation-based management zones (MZs) in the field. MZA uses fuzzy means clustering to group like values. Soil ECa MZ maps of the fields are shown in Figure 2.6-2.8. The MZs were used to assess the size of the sampling grid that was overlain on each field. Based on the observed spatial variability, a 0.5 ha (1.2 ac) sampling grid was selected for all three fields in 2022 (Table 2.1, Figures 2.9 – 2.11). while a 0.4 ha (1 ac) sampling grid was selected for all three fields in 2023 (Table 2.1, Figures 2.5 – 2.6). The size of the cells was optimized to be small enough to capture the spatial variability of the fields but also to account for the labor needed to collect data. Specifically, 2022A contained 11 grid cells, 2022B contained 31 cells and 2022C contained 21 cells, 2023A contained 30 cells (Figure 2.2), 2023B contained 17 cells (Figure 2.3) and 2023C field contained 26 cells (Figure 2.4). A sampling area of 6 m of radius around the center point of each grid cell was then delineated for non-destructive and destructive plant measurements.

2.1.2 Soil texture

Soil parameters play an important role in how the crops, soil moisture, soil temperatures and micro-organisms behave in the profile. Variation in the texture of the soil can lead to various changes in soil health and crop growth. For example, water holding capacity, infiltration rates, and soil texture can accelerate aflatoxin production (Dorner, 2009). Sandy soil textures and inadequate water availability can lead to water stress making

peanut plants susceptible to *A. flavus* infestation. To quantify soil texture in the research fields, a 90 cm intact soil core was collected from the center point of each grid cell (Figure 2.12). The soil cores were segmented into 15 cm sections and analyzed for soil texture by Waters Agricultural Laboratories (Camilla, Georgia). Soil texture was obtained in percentages of sand, silt, and clay.

2.1.3 Soil water tension and soil temperature measurement

Soil matric potential is essentially a measure of the energy required by plants to extract soil water from the soil matrix and is measured in units of centibars (cb) or kiloPascals (kPa). Soil matric potential values are negative. For simplicity, the absolute value of matric potential is often used and is referred to as soil water tension (SWT). The UGA Smart Sensor Array system (UGA SSA) was used to continuously measure SWT in approximately half of the grid cells in each field during the 2022 and 2023 growing seasons.

The UGA SSA is a wireless soil moisture sensing system that allows for a high density of sensor nodes and was developed by the UGA Precision Ag Team. It is described in detail by Vellidis et al. (2008; 2013) and has been used to monitor soil moisture and schedule irrigation in corn, cotton, peanuts, soybean, vegetables (eggplant, pepper, tomato, watermelon), and blueberry. The UGA SSA consists of smart sensor nodes and a base station. The term sensor node refers to the combination of electronics and sensor probes installed within a field at one location. The electronics include a circuit board for data acquisition and processing and a radio frequency (RF) transmitter. Each soil moisture probe integrates up to three Watermark® soil moisture sensors as shown in Figure 2.13. In addition, each node supports two thermocouples for measuring soil and/or canopy temperature. For field crops like cotton or corn, the sensors on the probe are arranged so

that when installed they are 20, 40, and 60 cm (8, 16, 24 in) below the soil surface although any combination of depths is possible. For peanuts, the sensors are installed at 10, 20, and 40 cm (4, 8, 16 in). The UGA SSA reports soil moisture in terms of SWT in units of kPa. The RF transmitter is responsible for transmitting sensor data. To overcome the attenuating effect of the plant canopy, the RF transmitter antenna is mounted on a spring-loaded telescoping fiberglass rod (Figure 2.13). Variable antenna heights are used to ensure that the antenna is always above the crop canopy. For example, a height of 2.5 m is adequate for low-growing crops like cotton, soybeans, and peanuts while a height of 4.5 m is used for tall crops like corn. This design allows field equipment such as sprayers and tractors to pass directly over the sensors without damaging them. The UGA SSA nodes are powered by two 1.5 V alkaline batteries which have a life of more than 120 days. This typically spans an entire growing season.

At a convenient location at the edge of each field, a base station receives the data from all nodes at hourly intervals. The base station stores the data on a solar-powered netbook computer and then transmits the data via a cellular modem to a cloud server where they are stored, managed, manipulated, and visualized. The data can also be downloaded in spreadsheet format from the server. Figures 2.14 and 2.15 show soil temperature and SWT graphs, respectively, from one of the sensor nodes in the fields. To consolidate hourly SWT and soil temperature measurements for 3 different depths into one measurement that represents SWT and soil temperature in the profile, we used weighted sum of SWT and soil temperature at 4, 8 and 16 inches of depths using 0.5, 0.3 and 0.2 weights respectively. We used the mean of values for 14 days prior to aflatoxin sampling date to consolidate data into one measurement that represents the variation in the period.

2.1.4 Physiological measurements

As part of the overall project, physiological measurements were made in all the fields. Although the physiological data were not used for the model developed in this thesis, a brief description of the measurements is provided for completeness. Beginning with 60 days after planting (DAP), plant physiological data were collected biweekly using a LICOR-600 Porometer/Fluorometer (LI-COR – Lincoln, NE) (Figure 2.16) for physiological parameters and a METER LP-80 Ceptometer (METER – Pullman, WA) (Figure 2.17) for leaf area index (LAI). Physiological parameters were classified into porometer readings (stomatal conductance, transpiration, leaf vapor pressure) and fluorometer readings (minimum and maximum fluorescence in light and dark, quantum efficiency in dark and light, and leaf light absorptance). Physiological parameters were measured at three locations within the sampling area around each cell's center point (Figure 2.18). For physiological parameters, a leaf was clamped onto the LI-COR-600's infrared leaf sensor which measured leaf temperature, and the fluorescence detector measured fluorescence signals from two fluorometer LEDs reflected from the leaf through the 0.75 cm diameter aperture. LAI was measured by laying the LP-80's probe across the row and measuring LAI which was formulated by the instrument from the below-canopy photosynthetically active radiation (PAR) sensed from the probe and the light intensity measured by an external sensor placed above-canopy at the reading moment and location.

To account for proper representation of the sensed area, multiple measurements were taken in the sampling area, specifically, three for physiological parameters and four for LAI. These multiple observations were averaged to obtain one data point for each grid cell in each of the fields. The same locations were used at each sampling event. This dataset

was not used for aflatoxin hotspot estimation as physiological parameters have a complex correlation with each other and must be studied separately to determine the relationship with A. flavus contamination.

2.1.5 Aflatoxin measurements

At the same time as the physiological measurements, ten whole plant samples including roots and pods were collected for subsequent aflatoxin analyses. From the designated sampling area in each plot, three clusters of three to four plants were obtained at random. After the plants were collected, they were returned to the laboratory where the peanut pods were removed from the vines, shelled, and analyzed for aflatoxin by Waters Agricultural Laboratories (Camilla, Georgia) using the ELISA kit method for aflatoxin (Kolosova et al., 2009). This testing kit is manufactured by Neogen-Veratox® (Lansing, Michigan) and has a detection limit between 1.4-50 ppb. The measured aflatoxin concentrations were reported in units of parts per billion (ppb). From 90 DAP to harvesting, a total of four observations were obtained for plant physiological parameters and peanut aflatoxin concentrations at each sampling point in all fields. Samples were also collected from each location at harvest. The sampling dates for each field are reported in Table 2.2. The last tissue sampling for aflatoxin measurement was done on the harvesting of each field as shown in Table 2.2.

2.1.6 Yield Data

Harvest at each field was done on the dates shown in Table 2.2. Yield data were collected by harvesting 100 ft. of one peanut bed (2 rows), 50 ft. on each side of the sampling point, with a 2-row bagging combine. Two to three bags were collected at each sampling point. The bags were weighed in the field and their mass recorded. Samples of

peanut pods were extracted from one bag for aflatoxin analysis, burrower bug damage assessment, moisture content, and foreign material content. Figure 2.19 shows a visualization of the harvesting method.

2.1.7 Remotely Sensed Data

Planet Explorer, which is a platform developed by Planet Labs (San Francisco, CA) makes available multispectral imagery of the earth's surface at a spatial resolution of 3 m and temporal resolution of one day. This platform was used to download multispectral images of the study fields on or around the field sampling days. When cloud-free images were not available for the sampling day, images from ±2 days from the sampling date were used. These multispectral images reported reflectance data in four wavebands: Blue (455-515 nm), Green (500-590 nm), Red (590-670 nm) and Near Infrared (780-860 nm). The reflectance rasters were used to develop the Normalized Difference Vegetation Index (NDVI) for each field for each sampling date.

NDVI (Equation 1) corresponds to vegetation greenness, density and plant vigor and biomass with the index value ranging from -1 to 1, in which negative values correspond to water and clouds, values near zero correspond to bare soil, low positive values correspond to low biomass and/or poor vigor while and higher positive values indicate higher biomass and good vigor.

$$NDVI = [NIR - RED] / [NIR + RED] \quad (1)$$

The NDVI values were in a 3 m raster that corresponded to the spatial resolution of the images. The rasters were used to extract NDVI values for each sampling area by

creating a 6 m buffer at the center point of the sampling area and taking the arithmetic mean of all the rasters within the vicinity.

2.1.8 Meteorological Parameters

Precipitation (mm), air temperature (K), solar radiation (W/m2) and vapor pressure deficit (kPa) were meteorological parameters used in the models. For 2022 growing season, these data were extracted from an API that offers daily high-spatial resolution (1/24th degree ~ 4 km) surface meteorological dataset ranging from 1980-2022, called gridMET. This API uses climatically aided interpolation on gridded climate data from PRISM to produce high temporal and spatial resolution datasets. To obtain raw tabular data from gridMET, various functions in R, developed and compiled by 'mikejohnson51' (GitHub) in packages "AOI" and "climateR", were used to download required parameters for the fields. To represent the meteorological parameters on the day physiological parameters and tissue samples were taken, arithmetic mean of fourteen days was taken for all variables except precipitation, for which total amount of precipitation in the corresponding interval was taken. The final dataset of meteorological parameters consists of four observations of four meteorological variables for each field. For the 2023 growing season, the meteorological parameters were obtained from an in-field ATMOS 41 weather station (Figure 2.20) manufactured by METER (Pullman, WA), at each field. These parameters were recorded and uploaded using telemetric methods and were extracted at the end of the crop season.

2.2 Data Analysis

2.2.1 Field Data

Of all the data sets collected for the overall project, five were used for this work – soil texture, meteorological parameters, soil water tension, NDVI, and aflatoxin concentrations. These five datasets were appended, cleaned, and sorted to prepare them for use in the modeling component of the study. A brief description of each data set follows.

1. Soil Texture: The soil texture dataset consisted of percentages of sand, silt, and clay. The amount of clay present in the soil was taken as the representative of soil texture. Clay percentages at 5 different depths were compiled for each plot at a temporal resolution of the full crop season.
2. Meteorological Parameters: The four variables that represented the contribution of weather to the estimation of aflatoxin production were solar radiation (W/m2), precipitation (mm), ambient temperature (K) and vapor pressure deficit (VPD). These were compiled at a temporal resolution of 15 days before each aflatoxin sampling date. 15 days of data for solar radiation, ambient temperature and VPD were averaged, and the sum of precipitation was taken to represent the effect of weather on aflatoxin production at each field location.
3. Soil Water Tension and Soil Temperature: SWT and soil temperature was recorded at an hourly rate by the UGA SSA but were consolidated by taking an average of 7 days before each aflatoxin sampling. One value of soil water tension was taken to represent soil moisture conditions at 3 depths (10, 20, and 40 cm) for 7 days and soil temperature was recorded and appended in the final dataset at 10 cm depth.

4. Normalized Difference Vegetation Index: NDVI was taken as one value at each sampling point for each sampling date by taking the average value of all pixels lying in the vicinity of 6m from the sampling point.
5. Aflatoxin concentration: Aflatoxin values for each plot for each field at each sampling date were appended to the final dataset. These numeric aflatoxin levels were transformed into categorical values by considering values greater than 1 ppb as 1 and values less than 1 ppb as 0 to develop classification models.

2.2.2 Predictive Models

Three modeling approaches were used to understand the relationships between the physical and environmental data collected during the project and the presence of aflatoxin in the peanut pods at the end of the growing season. The methods employed were evaluated based on R^2 and RMSE obtained from the predicted and observed values from the models developed. R^2 (coefficient of determination) is a measure of the goodness of fit of the model. It is the proportion of variance in the dependent variable that can be explained by the predictor set in a regression model. On the other hand, RMSE is a measure of the average difference between the values predicted by the model and the observed values. The three approaches are described below. The corresponding codes for models in Section 2.2.2 are presented in full in the Appendix.

2.2.2.1 Linear Regression

A linear regression model was developed using R coding through RStudio (Boston, Massachusetts). Initial steps involved creating a heatmap using the "ggplot2" package to understand the correlation between all the indicators to be input in a basic linear model. The indicators used were those described in the section 2.1.

To visualize the correlation between certain indicators, basic linear regression was done using the "lm" function from the "stats" package. The pairs of indicators with correlation coefficients greater than 0.5 were subset and were input in the linear model as interactions between the indicators. To train and test the linear model, a random split of 75% and 25% of the dataset was done. The obtained model was evaluated by using mean square error (MSE) and r-squared (r^2) estimated by regressing observed and estimated aflatoxin values.

2.2.2.2 Random Forest Regression Model

A random forest regression model using the "randomForest" package in RStudio was developed by again splitting the data into training and testing set and creating a loop to iterate over 25250 combinations of values for the number of trees ("ntree"), number of variables split ("mtry") and maximum number of nodes in a tree ("maxnodes"). "ntree" ranged from one to 501 with an interval of five, "mtry" ranged from one to five with an interval of one and "maxnodes" ranged from one to fifty with an interval of one. Post-development of all these models, evaluation was done using two parameters again, MSE and r^2. The model with the maximum sum of two parameters was chosen as the best-tuned random forest model. The chosen random forest model was iterated a hundred times to ensemble the different results from this model being run multiple times with a certain set seed. This gave us a further tuned random forest model with more stability in performance and estimation. The importance parameter (IncNodePurity) was estimated for each variable along with the partial dependency plots.

2.2.2.3 Recursive Feature Elimination combined with Random Forest Regression

The random forest regression model (Section 2.2.2.2) alone was not capable of subsetting important variables from the original predictor set that contribute to the skill of the model. Although the random forest modeling has the capability to recognize and delineate inter-variable collinearity by selecting a subset of the variables at each node-split in each tree, it was not able to not only identify the predictors that were actually adding to the variance explained in the aflatoxin concentration, but also eliminate predictors that were not adding any unique information to the model or, in cases, were a source of noise in the dataset. To do so, the random forest model was combined with a more-informed feature engineering approach called Recursive Feature Elimination (RFE).

This approach was implemented using "RecursiveFeatureElimination" function from a Python library called "feature_engine". An 80-20 ratio was used to randomly subset the dataset into training and testing datasets. The models developed were evaluated based on the R^2 value obtained in each iteration and a three-fold cross validation was used to increase the robustness of the model. Using the mean R^2 value from the three-fold cross validation model as a baseline, certain variables were retained or dropped from the predictor set based on their contribution to the model R^2 values relative to the baseline. Prior to RFE, all the predictors were scalarized, i.e., removing the mean and scaling to unit variance. This was accomplished to reduce computational resources used in the modeling process. The RFE procedure is described in detail as follows:

1. Use three-fold cross validation to train the model and set the mean cross-validation performance as the baseline (R^2_{baseline}).

2. Re-train the model without each variable to evaluate the contribution of the corresponding variable by observing the change as $R^2_{baseline} - R^2_{new}$. If the change is positive, it means the variable was important for the model to explain the variance and if the change is negative, it means that removing the variable resulted in a better mean performance of the model.

3. Eliminating the variables with a negative change gives a new a set of predictors that now have their own interactions. This process is repeated until the smallest change in $R^2 < 0.001$.
4. The final set of variables is a collection of all the variables that have a positive contribution to the model's performance and the importance of each variable is evaluated based on the extent of positive change in R^2.
5. The $R^2_{new} - R^2_{baseline}$ quantity was interpreted as the feature importance criterion for this method as opposed to the IncNodePurity criterion used for the Random Forest Regression Model in section 2.2.2.2.
6. To evaluate if changing the predictor set affects the predictor interaction and model performance, steps 2-5 were repeated to see if one or more predictors are eliminated. If not, the obtained set of predictors include all those that contribute some unique information to the model.

Once the reduced predictor set was identified, these predictors were used in a random forest model to predict aflatoxin concentration using procedures like section 2.2.2.2. The predictions using the engineered set of predictors were compared against independent testing data to document model performance (R^2, MSE, and RMSE).

Additionally, the physical relationships between aflatoxin concentration and the RFE-retained predictors were visualized using partial dependence plots.

2.2.2.4 Testing of Regression and Classification Random Forest model

To further test the performance of the Random Forest regression model with RFE, it was applied to the 2022 data set. Aflatoxin concentrations from the 2022 fields had very high values. Because the samples were stored in poor conditions for an extended period before being analyzed for aflatoxin, we have low confidence in the measured concentrations. Consequently, the model was not tested against measured concentrations. Rather, it was tested for its ability to indicate the presence of aflatoxin at the same sampling locations for which we had positive samples in 2022.

The developed model was employed on the 2022 testing dataset that comprised of the same 8 predictors used in 2023 as subset by RFE technique. The output from the regression model tested on 2022 dataset was used to develop interpolated maps for predicted values of aflatoxin concentration. In addition to this, a classification was developed as well, using the same attributes and predictor set to delineate the efficacy of the model to predict the presence of aflatoxin contamination at the sampling points in 2022 trial fields. This was done by transforming the zero values of aflatoxin concentrations to the absence of aflatoxin contamination and non-zero values to the presence of aflatoxin contamination. This helped us understand the ability of the model to detect a series of patterns that can indicate to the presence of aflatoxin in the fields. To evaluate this ability of the model, we plotted a confusion matrix, that helped us visualize and understand its performance.

Tables:

Table 2.1. Table summarizing the design of field trials implemented for studies in 2022-2023.

Field Name	Year	Location	Size (ha)	Grid Cell Size (ha)	Number of Sampling Points
2022A	2022	South-Central Georgia	6.2	0.5	11
2022B	2022	South-Central Georgia	16.9	0.5	31
2022C	2022	South-Western Georgia	10.5	0.5	21
2023A	2023	South-Central Georgia	12.1	0.4	30
2023B	2023	South-Central Georgia	7.3	0.4	17
2023C	2023	South-Western Georgia	10.5	0.4	26

Table 2.2. Aflatoxin whole-plant sampling dates for field trials in 2022-2023.

Field Name	**Sampling Dates**	**Harvest Date**
2023A	July 5, 2023	September 20, 2023
	July 26, 2023	
	August 8, 2023	
	August 22, 2023	
	September 20, 2023	
2023B	July 7, 2023	October 10, 2023
	July 25, 2023	
	August 9, 2023	
	August 22, 2023	
	October 10, 2023	
2023C	August 1, 2023	September 25, 2023
	August 15, 2023	
	September 9, 2023	
	September 25, 2023	

Figures:

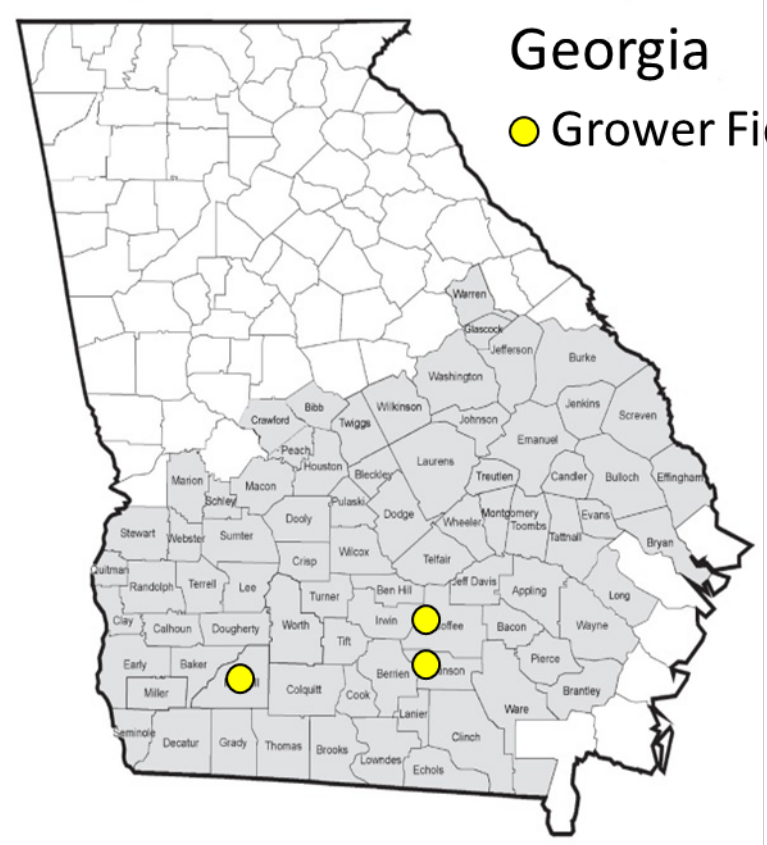

Figure 2. 1 Map of Georgia showing the general location of the three grower fields used in 2022 and 2023 (yellow circles) within the Georgia peanut growing area (gray counties).

Figure 2.2. Map of the 2023A field in south-central Georgia showing the 0.4ha grid and numbered sampling locations. The area shown in forest within the field boundary was cleared by the landowner prior to the 2022 growing season.

Figure 2.3. Map of the 2023B field in south-central Georgia showing the 0.4ha grid and numbered sampling locations.

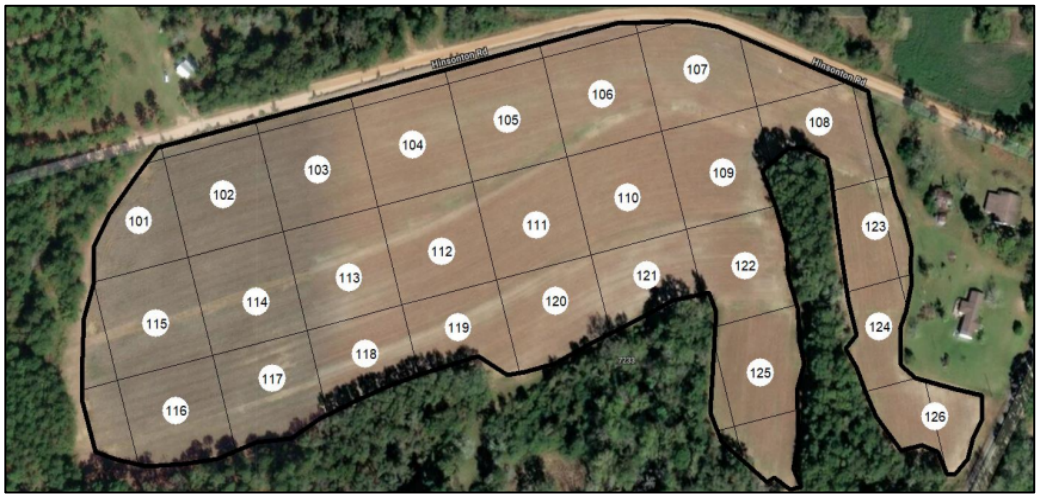

Figure 2.4. Map of the 2023C field in southwestern Georgia showing the 0.4ha grid and numbered sampling locations.

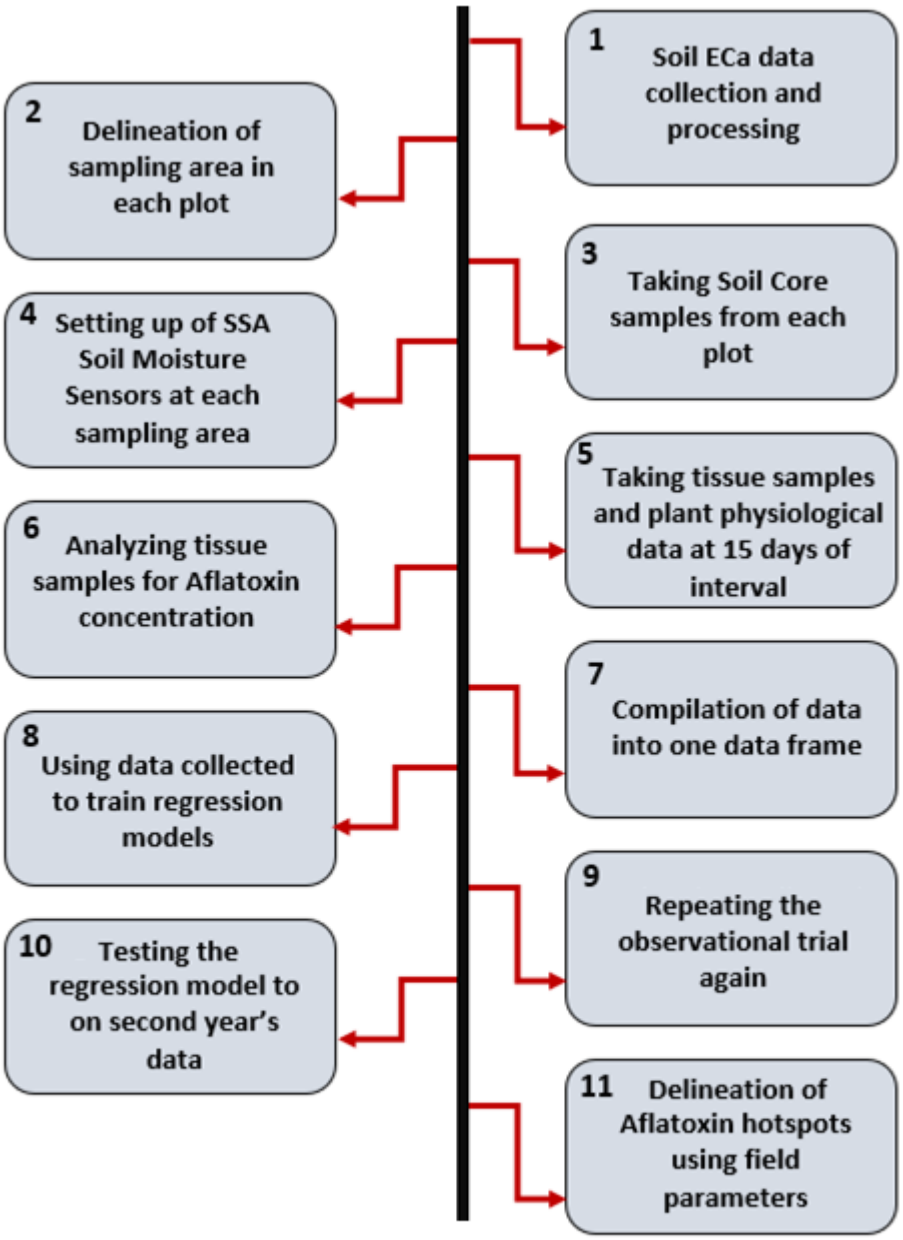

Figure 2.5. Workflow timeline of the study.

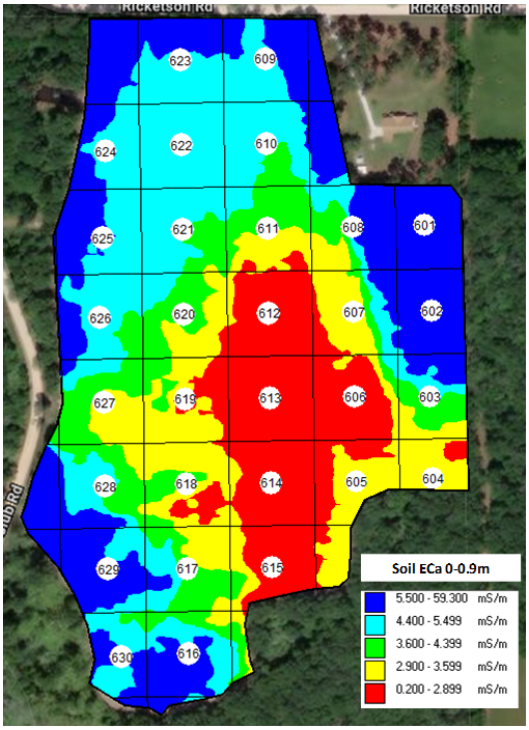

Figure 2.6. Soil ECa map of 2023A field collected with a Veris 3100. Data shown are for an integrated depth of 0-0.9m.

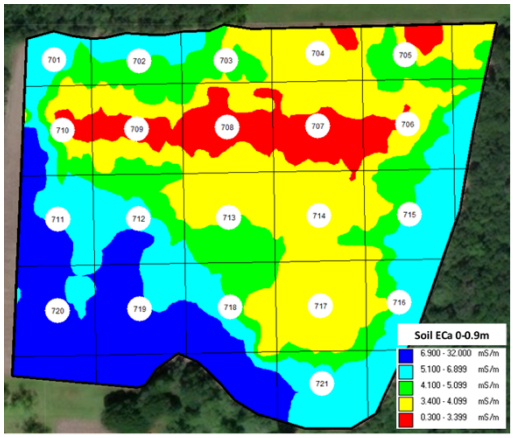

Figure 2.7. Soil ECa map of 2023B field collected with a Veris 3100. Data shown are for an integrated depth of 0-0.9m.

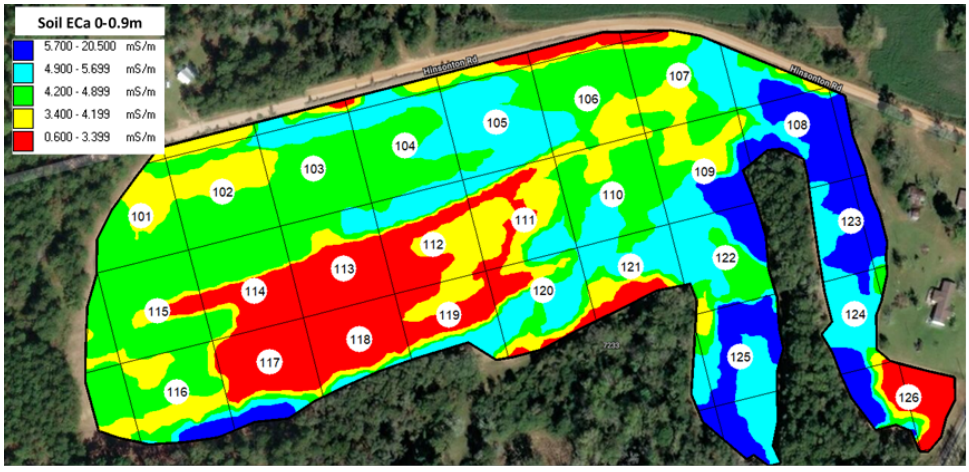

Figure 2.8. Soil ECa map of 2023C field collected with a Veris 3100. Data shown are for an integrated depth of 0-0.9m.

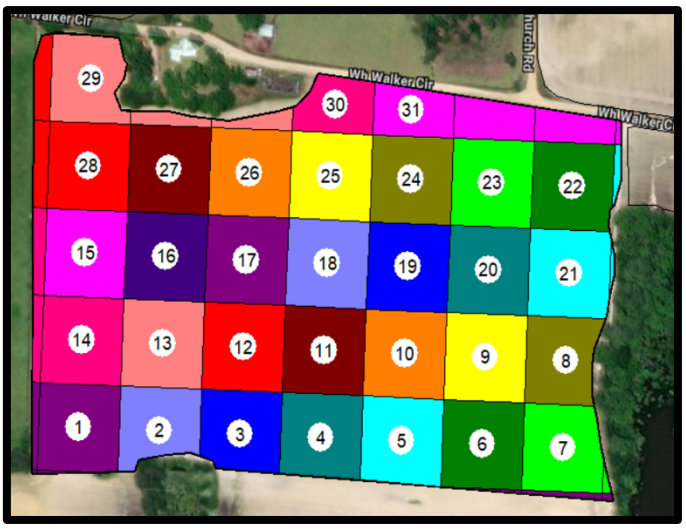

Figure 2.9. Map of the 2022A field in south-central Georgia showing the 0.4ha grid and numbered sampling locations.

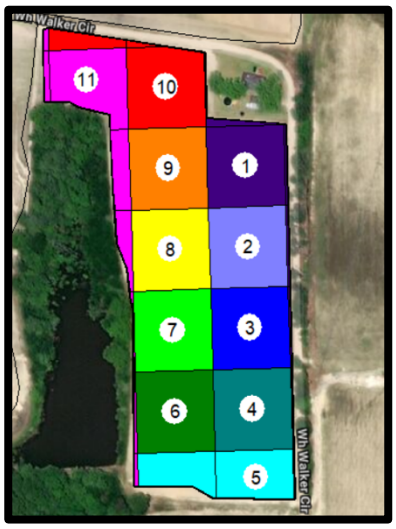

Figure 2.10. Map of the 2022B field in south-central Georgia showing the 0.4ha grid and numbered sampling locations.

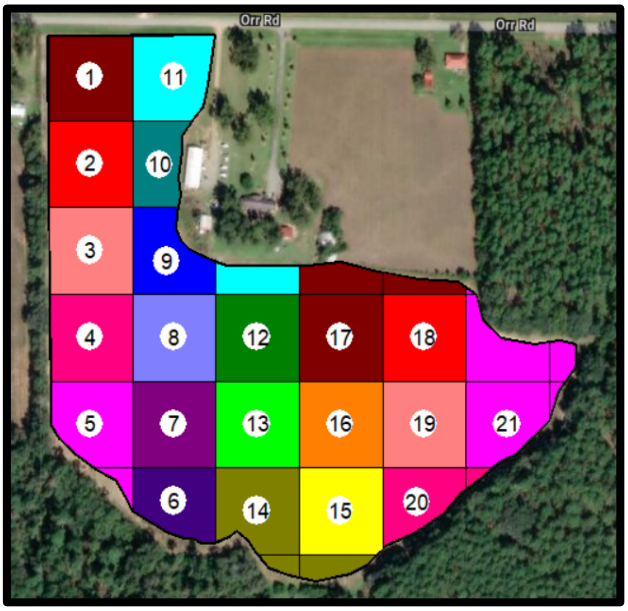

Figure 2.11. Map of the 2022C field in south-central Georgia showing the 0.4ha grid and numbered sampling locations.

Figure 2.12. This is an intact extracted soil core sample that was segmented and later separated for soil texture analysis.

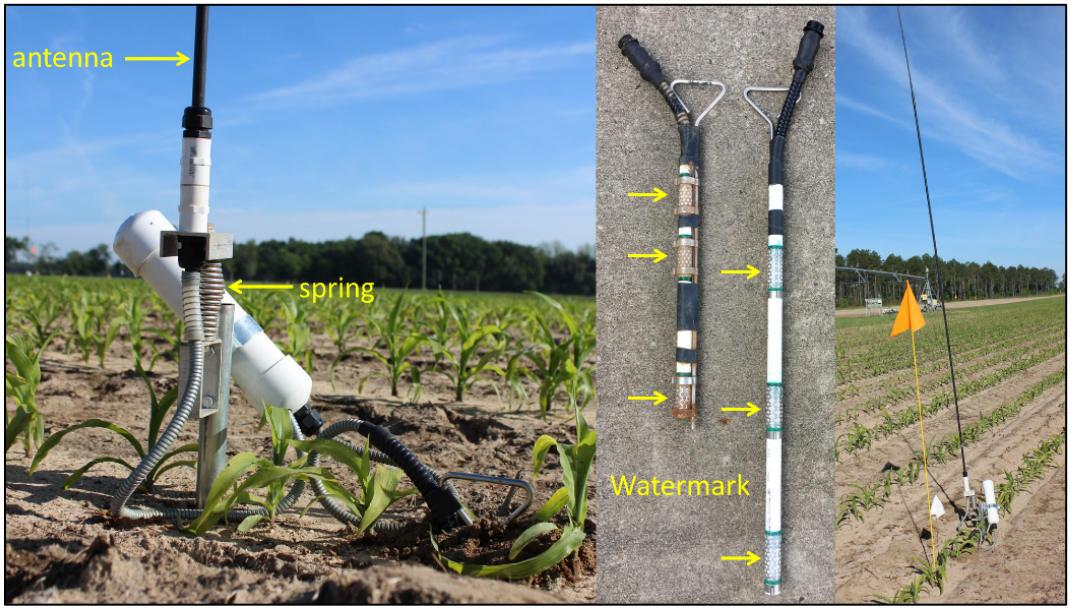

Figure 2.13. A UGA SSA sensor node installed in the field and two versions of the probe – one for shallow-rooted crops and one for deep-rooted crops.

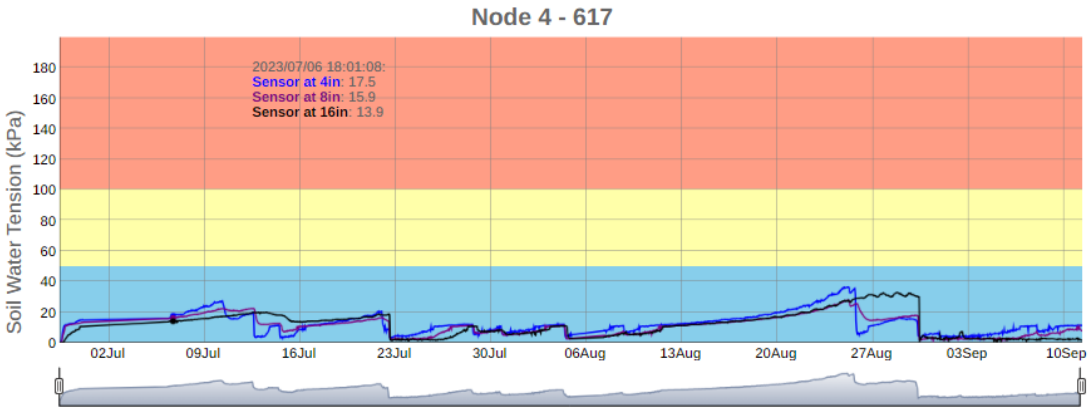

Figure 2.14. Soil Water Tension graph that was developed for one node in field 2023A.

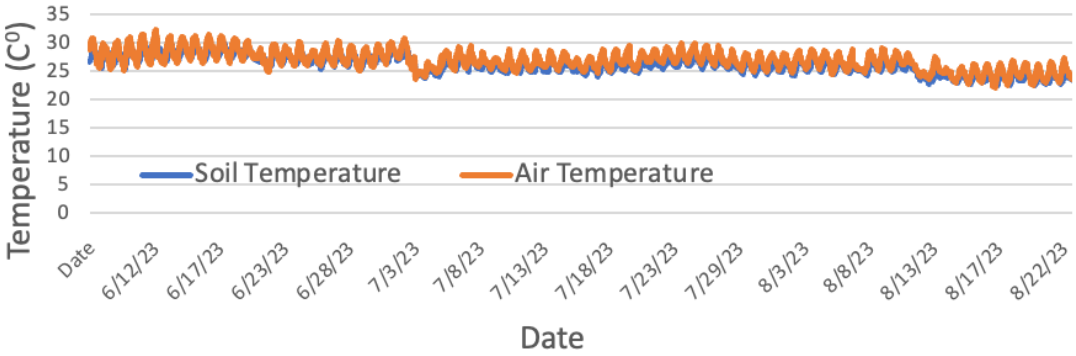

Figure 2.15. Soil Temperature graph that was developed for one node in the one of the fields.

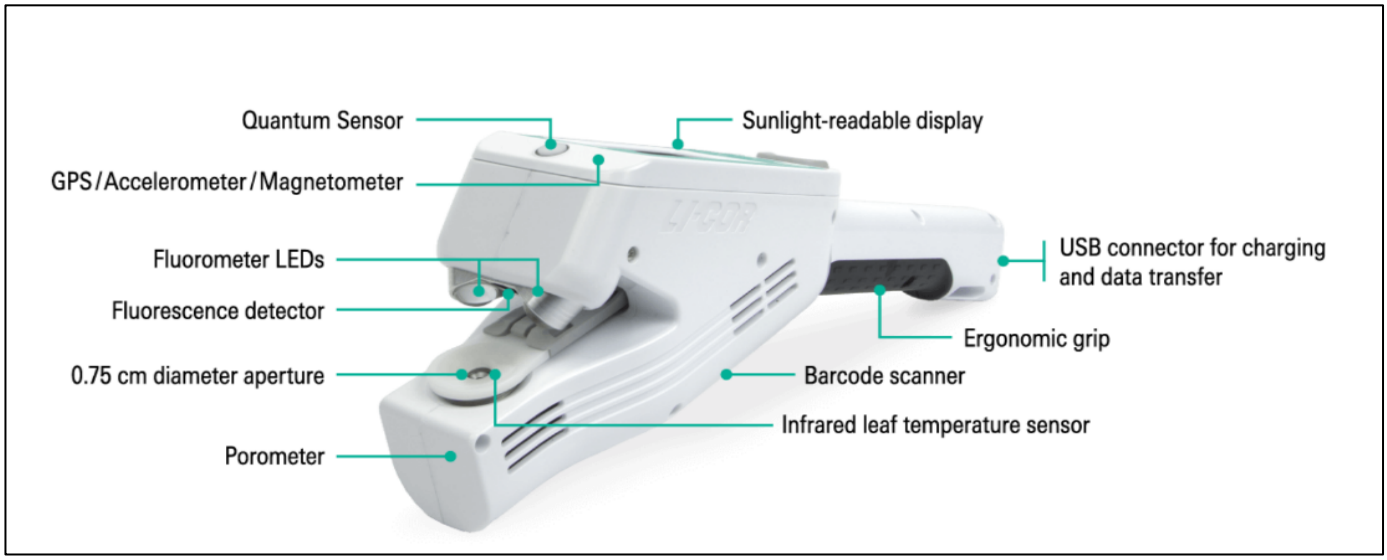

Figure 2.16. Detailed image of a LI-COR 600 and its components (https://www.licor.com/env/products/LI-600/)

Figure 2.17. Image of a METER LP-80(https://www.metergroup.com/en/meter-environment/products/accupar-lp-80-canopy-interception-par-leaf-area-index)

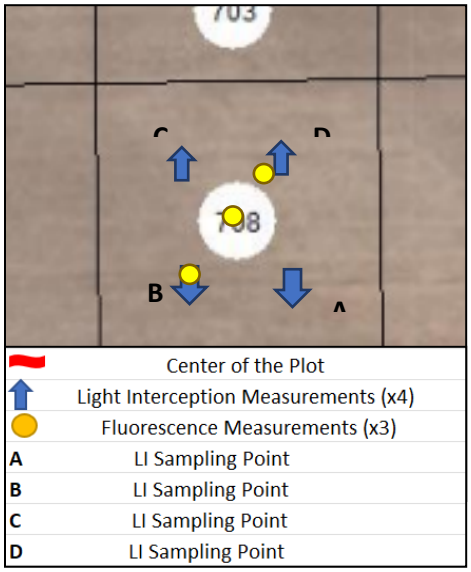

Figure 2.18. Physiological parameter sampling method.

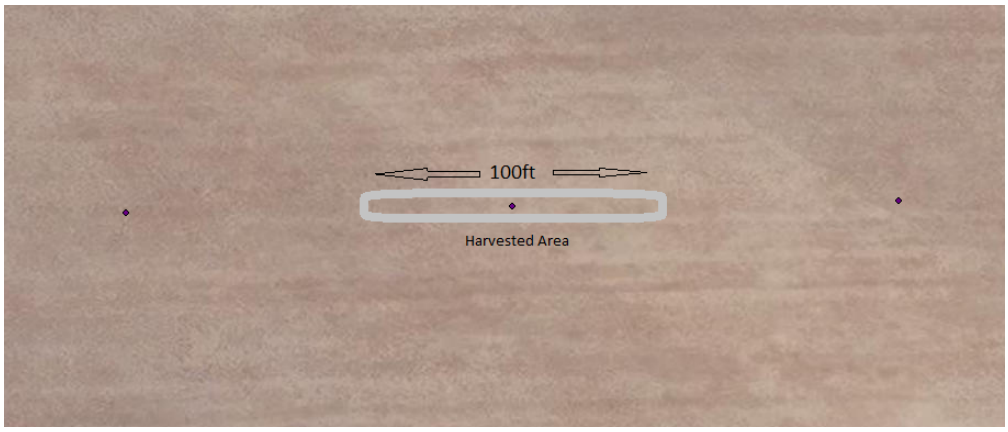

Figure 2.19. Visualized image of field harvesting at sampling points. 50ft of harvested peanuts were collected on each side of the sampling point in the same twin row.

Figure 2.20. Installed ATMOS 41 weather station used to sense meteorological parameters.

CHAPTER 3

RESULTS AND DISCUSSION

3.1 Results

3.1.1 Field Data

The paragraphs below provide a summary of the physical and environmental data that were collected during the growing season and used in the two models.

3.1.1.1 Soil Texture

Soil Texture for all the fields was analyzed by creating interpolated maps (Figure 3.1-3.6) using Inverse Distance Weighted methods to sense the spatial distribution of clay percentage of soils at different depths (0-15cm and 75-90cm). Since, the target is to quantify soil texture to be used an identifier, clay percentage was used as an indicator of soil texture as clay particles have a significant impact on the textural properties of the soil such as water retention, cation exchange capacity, plasticity and others. Figures 3.7-3.12 show histograms of the % clay distribution in all the 0-15cm and 75-90cm soil core samples collected in each field. These histograms show the density of pixels for a certain raster value. This helped us understand how much clay was present in the field at different depths. For example, we observed that soil clay percentage in field 2023A was found more at northern end of the field at 0-15cm of depth (Figure 3.1) whereas high clay content was only found at south-west end of the field at 75-90cm of depth (Figure 3.2). Figure 3.7

shows that most of the shallow profile of 2023A had clay content of ~ 4% and at deeper profiles, majority of the clay content ranged from 5-16%. This helped us understand the spatial distribution of the fields' soil texture and quantify the variability of soil texture that made us choose the fields in the first place before the trial started.

3.1.1.2 Weather Variables

Weather variables were observed from in-field weather sensors daily. Periodic in-field observations were taken to assure the quality of the weather data. VPD and air temperature were used as the indicators of the water demand of the atmosphere to categorize different weather conditions.

Field meteorological parameters were visualized by plotting linear graphs to see the temporal trends in the various parameters. Figure 3.13 shows precipitation for each field through the 2023 growing season. The temporal period used was between 2023/06 and 2023/10, as this was the period where the aflatoxin sampling was done. To visualize the temporal variation of solar radiation, air temperature and VPD throughout the season for all three fields, Figures 3.14, 3.15 and 3.16 respectively were developed. Through the season, sensor output of the weather station in field 2023A for actual vapor pressure (e_a) was 0 starting early July up to the end of the season. This caused the VPD (e_s - e_a) to only output e_s for that period. This can be visualized as the sudden peak in the trend for VPD in field 2023A (Figure 3.16). To worked around this loss of data, the VPD values for field 2023A were replaced by the values obtained from the weather station at 2023B as they were at a distance of 860m which is insignificant in context of spatial variability in VPD.

3.1.1.3 Soil Water Tension and Soil Temperature

Soil water tension values were used as the absolute indicator of soil moisture and was used in synergy with other parameters to quantify the soil profile conditions. The

weekly average of the soil water tension was used as the indicator of the moisture conditions in the soil profile corresponding to each aflatoxin sampling date and spatial point. SWT changed with changing weather parameters, mostly precipitation. In an event of precipitation, the field terrain governed the SWT in the profile for the days until next precipitation event. Parts of the field with higher clay percentages retained moisture for longer leading to low SWT. On the other hand, the parts of the field with lower clay percentages showed higher water loss rate and dried up sooner causing comparatively higher SWT. Figures 3.17-3.22 show the contrast between the SWT trends in two different spots in each field through the season. The highlighted points in Figures 3.1-3.6 and Figures 3.23-3.25 show the locations of the SWT sensors that were used to observe the spatial variation in SWT trends through an overlapping visualization with clay % and NDVI. The figures show how different soil texture, terrain and other factors affected SWT trends and how they change conditions for aflatoxin production. Figure 3.26 shows a scatterplot between SWT and measured aflatoxin concentrations in peanut pods throughout the growing season. The estimates trendline has a positive slope and a correlation coefficient of 0.09. This indicates that there is an insignificant positive relationship between increasing SWT (drier soil) and increasing aflatoxin concentrations.

3.1.1.4 NDVI

NDVI maps were developed to understand plant biomass distribution and changes in all fields. Figures 3.24-3.26 show examples of NDVI maps developed and Figures 3.27-3.29 show the histograms created for these NDVI rasters to understand the distribution of the index values in these fields. These spatial data were compiled in a tabular form (Table 3.1-3.3) which consisted of values of the index for each grid cell in all three fields for the days physiological measurements were taken. From these tables, it was observed that the

values for NDVI through all three fields ranged between 0.7 and 1. This implies that NDVI values came to a saturation limit and were not able to scale plant health and canopy density at this point. Still, to use NDVI rasters to observe any reflectance signals to correlate with aflatoxin, we implemented these datasets to train our models. Figure 3.24-3.26 represents the NDVI for each field at a spatial resolution of 3m in mid-season. Figure 3.24 is the map for NDVI values for the field 2023B developed using a set of satellite images taken on July 25, 2023. NDVI for certain parts of the field is lower than the rest. This depicts the absence of dense biomass which could be due to inadequate soil moisture, varying soil profile and terrain or it could be just bare soil as the image was taken before significant overlapping was observed.

3.1.1.5 Aflatoxin Concentrations in Peanut Pods

A. flavus contaminated peanuts can have aflatoxin concentrations ranging from below detection limits to hundreds of ppb. Beginning in early August and ending at harvest, a total of 318 samples were collected and analyzed for aflatoxin concentrations from the three 2023 fields. 256 (80.50%) of the samples were within the Veratox® ELISA method's quantitation range while 62 (19.5%) were below the method's detection limits (1.4 ppb). The mean concentration was 0.84 ppb, standard deviation was 0.75 ppb and the maximum measured concentration was 3.5 ppb. Table 3.4 summarizes the overall results as well as the results for individual fields.

To evaluate the range of concentrations measured in the study, a density plot of 152 observed aflatoxin values from the 2023 fields was developed (Figure 3.30). Only aflatoxin values for which the other variables used in the regression and random-forest model were available were included. The mean aflatoxin concentration for this subset of samples was 0.98 ppb while the median concentration was 0.9 ppb and the standard deviation was 0.815

ppb. The highest measured concentration was 2.3 ppb while 17% of the values were below the detection limit. 53% of the values were between the detection limit and the mean of the dataset. The density plot was then used to establish a "cutoff" value for the modeling. The cutoff value was established on the basis of the standard acceptable aflatoxin concentrations in international markets. Since, our data's range was not as wide as the real market, and we decided to keep the cutoff as 1 ppb instead of 4 ppb as suggested by USDA. The idea of having a cutoff was to split the data into safe and unsafe categories. Concentrations below the cutoff were considered "safe" or non-indicative of potential aflatoxin contamination problems once the peanuts were harvested. Concentrations above the cutoff were considered "unsafe" and indicative of potential contamination problems.

Figures 3.31, 3.32 and 3.33 show the interpolated maps for the aflatoxin concentrations at harvest in each of the fields. When comparing to maps of the other measured variables, it was observed that parts of the field where high clay content were found, higher values of aflatoxin were observed at harvest. Figure 3.34 shows a scatterplot between clay percentage at 75-90cm of depth and aflatoxin indicating the positive relationship between the two variables. One explanation that can justify this implication is that higher clay content leads to poor drainage conditions at deeper depths of the profile and adversely affects peanut pod growth and development which may leave the pods susceptible to *A. flavus* colonization and subsequent aflatoxin development. Figures 3.35-3.37 show maps of % clay, aflatoxin concentrations at harvest, and mid-season NDVI of each field for easier comparison of spatial trends. It can be observed from

3.1.2 Linear Model

A linear model is used to gain a basic quantified regressed relationship between a certain number of indicators and a target variable. For our datasets, ten interactions within the indicators were found significant (correlation coefficient > 0.5) using a heatmap (Figure 3.38). Significant correlations were found between clay % (depth 2), clay % (depth 3), and clay % (depth 4) with each other, NDVI with clay % (depth 4 and 5) and temperature with both solar radiation and soil water tension. A linear model was developed by using all variables and significant interactions as input. Apart from this, the heatmap also included aflatoxin as the predicted variable to look for any direct correlations with the predictos. Even though weak, aflatoxin showed correlations with SWT, NDVI and clay percentages at different depths. The linear model resulted in an r^2 value of 0.33, with an MSE of 0.57 and a confidence interval of 99%. The regression between the observed aflatoxin and estimated aflatoxin resulted in an equation with intercept of 0.65 and slope of 0.32 with an r^2 of 0.03. Figure 3.39 shows the visualization of the performance of the linear model using a scatter plot between the observed and estimated values of aflatoxin. These results imply that the linear regression method was not able to find a significant or determinant relationship between aflatoxin and the input variables.

3.1.3 Random Forest Model

A Random Forest model is primarily defined by the number and type of trees it consists of. A tree in a random forest regression model can have various attributes that can be appropriate towards contrasting applications. A tree's performance is defined by the number of splits, also called the number of nodes, and the number variables split at each node. At each node, at least one variable is used by the model to split the dataset into two

regions which gives the least entropy, in other words, the most information gained. For our dataset, a random forest regression model was applied to investigate if the former can present improved performance when compared to the linear model described in Section 3.1.2. The random forest regression model was able to explain the variation and the relationship between aflatoxin and other parameters better relative to the linear model (r^2 = 0.03). Upon evaluating model parameters amongst 25250 model runs, it was found that the best-performing random forest model had 6 trees with 20 nodes in each tree and 1 variable split at each node. This means that in this model, 6 trees were developed, in which each of them had 20 nodes and at each node, only one variable was split.

To visually assess the accuracy and performance of the model, the observed and predicted values were regressed against each other (Figure 3.40). The final random forest model was evaluated, with an RMSE of 0.74 and an r^2 value of 0.16. This implies that together, all the 22 predictors that were included in the model were able to explain 16% of the variance observed in aflatoxin concentration. The variable importance parameter, IncNodePurity, was the highest for clay percentage at depth 5 (90 cm), followed by NDVI, soil water tension, VPD and other variables (Figure 3.41).

3.1.4 Recursive Feature Elimination

As intended, recursive feature elimination (RFE) applied to the Random Forest model was able to limit the original predictor (variable) set to fewer and relatively more meaningful predictors from a statistical standpoint. On the first RFE iteration the model identified five predictors (silt % at 15-30 cm depth, clay % at 30-45 cm of depth, sand % at 30-45 cm of depth, sand % at 45-60 cm of depth, NDVI) as not sufficiently contributing to overall model performance based on a threshold of 0.001 change in R^2. Because the

interactions between predictors change when some predictors are removed from the model, additional iterations were performed until all remaining predictors were retained in the model. The second iteration resulted in an additional five predictors being eliminated (sand % at 0-15 cm of depth, silt % at 0-15 cm of depth, clay % at 45-60 cm of depth, silt % at 45-60 cm of depth, Soil Water Tension), retaining 12 predictors. During the third iteration, four additional predictors were eliminated (clay % at 0-15 cm of depth, sand % at 15-30 cm of depth, clay at 15-30 cm of depth, precipitation), retaining eight predictors.

Further RFE iterations did not eliminate any more predictors, meaning that the remaining eight predictors all contributed unique information and were necessary to achieve optimal performance. The final set of predictors in the order of their importance based on the change in R^2 was silt % at 60-75 cm of depth, vapor pressure deficit, clay % at 60-75 cm of depth, sand % at 60-75 cm of depth, silt % at 30-45 cm of depth, soil temperature, solar radiation, and air temperature. Using these eight predictors, resulted in improved model performance ($R^2 = 0.27$ and RMSE = 0.65) compared to using a Random Forest model without RFE (Figure 3.42). As evident, the RFE model enabled improved prediction of aflatoxin concentration despite only relying on a third of the original number of predictors collected.

Because the model relies on only 8 predictor variables, we visualized the functional relationships between aflatoxin risks and these predictors (Figures 3.43-3.50). The y-axes on these plots represents the partial dependency of aflatoxin on each of the predictors and the x-axes represents, on each plot, the range of 5^{th} to 9^{t5h} percentile of predictor dataset to minimize any misinterpretations caused by the outliers. The ticks on x-axis of each plot represents the deciles from each predictor's data distribution. Silt % at 75-90 cm of depth

(Figure 3.50), silt % at 30-45 cm of depth (Figure 3.46), solar radiation (Figure 3.44), air temperature (Figure 3.43), and clay % at 75-90 cm of depth (Figure 3.48) showed a positive partial dependency in the model whereas, sand % at 75-90 cm of depth (Figure 3.47), soil temperature (Figure 3.45), and vapor pressure deficit (Figure 3.49) showed a negative trend for the same.

It was observed that after eliminating predictors using RFE approach, the final set of predictors consisted of just weather variables, soil texture parameter and soil temperature. This implies that the model can be run without any in-field observations if soil temperature was removed as a predictor from the final set. To evaluate the performance of the model without soil temperature as a predictor, another model was developed. This model had a R^2 value of 0.13 and RMSE of 0.84. Without soil temperature, the model does not perform nearly as well. This indicates that as expected, soil temperature is a critical variable in predicting aflatoxin presence.

3.1.5 Testing of Regression and Classification Random Forest model

The model output a set of predicted aflatoxin values using the 2022 predictor dataset (Table 3.5). As the model was trained on values from 2023, it was unable to predict aflatoxin values greater than values in 2023. So, to compare the observed and predicted values, we developed interpolated maps for the two. Upon visualization of the interpolated maps (Figure 3.51 – 3.56) we observed that the predicted maps showed a weak similarity to the observed aflatoxin maps. One sound reason could be that model was trained on a limited range of aflatoxin concentrations, so it predicts within that range only. The model does not recognize any unique information from the datasets that can lead the predictions to be as high as they are. To understand the model's processing and ability to delineate

relationships between the predictor variables, we developed a classification model to predict the presence of aflatoxin contamination in 2022 trial fields. This enabled us to quantify the model's performance in recognizing relationships and information patterns that indicate aflatoxin's contamination. Upon observing the model's performance through the confusion matrix developed (Figure 3.57), we found that model was 96% accurate in predicting aflatoxin contamination. This implied that even though the model was unable to accurately estimate aflatoxin value due to the limited range of training dataset for target variable, the model was able to delineate the relationships which indicate aflatoxin production using these 8 variables.

3.2 Discussion and Conclusions

The use of common performance metrics (R^2, RMSE) enables one-to-one comparison among how the effectively the three models affect aflatoxin concentration. It was established that the models employing random forest regression were superior to multiple linear regression model. Further, amongst the two random forest techniques, the one employing RFE performed better (R^2 = 0.27 and RMSE = 0.65) compared to the random forest model without RFE (R^2 = 0.16 and RMSE = 0.74). The random forest model with RFE explained 27% of aflatoxin concentration variance using 8 variables some of which can be measured once for each field (soil texture) or easily measured during the growing season (meteorological data and soil temperature). This provides some promise that with further refinement with data from additional fields, this type of model may be able to predict aflatoxin occurrence at the field level with a reasonable degree of confidence.

The primary reason the random forest models superseded the linear model is that the linear model overlooks the possibility of any non-linear relationships that may be present between the predictors and predicted variables. The partial dependence plots (Figures 3.43-3.50) serve as evidence of presence of non-linearity for at least some of the predictors. A more obvious reason for this performance gap is the inherent difference between the models: Tree-based learning (Random Forest models) versus monotonic slopes (linear model). Because a decent proportion of the predictor variables were moderately to strongly correlated (Figure 3.38) to each other, it is highly likely the linear model suffers from multi-collinearity. On the other hand, random forest models or machine learning models, for that matter, are particularly suited for application where multi-collinearity is present or likely. A good example of possibly highly correlated variables are depth specific soil properties.

Further, the performance gap of the random forest model versus random forest combined with RFE and its likely causes deserve emphasis. Besides just performance, a distinct difference between the two approaches is model parsimony. A parsimonious model is one that achieves the desired level of prediction skill with as few predictors as possible. There is substantial literature that highlights the importance of parsimony even when a slight loss of performance is incurred. A parsimonious model is often attractive because it presents the user with the luxury of measuring only a few parameters of interest, relieving pressure on time and resources. In our case, model parsimony did not just provide with that luxury but also substantially improved the performance (R^2 increased from 0.16 to 0.27). This observation insinuates that inclusion of 19 out of 27 predictors was negatively impacting model skill. Furthermore, RFE concluded that the model does not view SWT as

a relevant predictor. This is significant, given the fact, that it is one of costliest and most laborious variables to measure in the field (using in-situ sensors) and drought is consistently identified as a causal agent of aflatoxin contamination. Models that use a smaller predictor set are more likely to be adopted over complex models by agricultural managers.

When change in performance is used as a criterion for feature importance, air temperature emerged as the most important variable followed by solar radiation and soil temperature. Soil temperature cycles follow air temperature (diurnal to longer periods) and their relationship is a function of soil water state and flux. Because solar radiation (incoming shortwave) is a major driver of net radiation, available at a field, it influences elements of the energy budget used to heat air (sensible heat) and the soil (ground heat flux). Thus, not only the drivers themselves but the interactions between them as explained above are useful for the model to extract valuable information. The high importance of these predictors implies that soil temperature acts as a major driver of aflatoxin production in peanut fields as suggested by Dorner (2009). Despite low importance, comparative to predictors mentioned above, clay, silt and sand percentage at 75-90 cm of depth, and silt percentage at 30-45 cm of depth play a crucial role in the model's understanding.

A possible reason that soil texture plays an integral role in the aflatoxin production is that soil texture contributes crucially to a soil profile's water retention ability. Dorner (2009) also discusses the presence of such a soil texture at deeper depths that allow optimum water drainage as a crucial occurrence. This provides peanut plants with an optimal soil profile for healthy growth when drought conditions appear towards the end of the season which can be a cause for accelerated aflatoxin production (Wang et al., 2016).

At shallower depths, the presence of optimum silt % is crucial as this is where peanut pods reside. Probst and Callicott (2012) suggest practices that allow optimal soil moisture conditions in rooting depths to prevent conditions favorable to aflatoxin production. All above mentioned variables, directly or indirectly, correspond to just either humidity or temperature whereas, VPD acts as variable that incorporates humidity and temperature into one parameter that can be used in model's understanding of aflatoxin production. After this discussion, an argument might arise about the relative importance of moisture and temperature in aflatoxin production. The presence of solar radiation and vapor pressure deficit in the final set of predictors implies that the interaction between moisture and temperature plays a crucial role in aflatoxin production as well.

It is equally useful to assess which predictors were not regarded as important by the model. One of the major parameters that was eliminated by RFE was the soil texture in depths that were not as important for aflatoxin production in peanut pods. Due to the collinearity in the properties of soil at different depth, the model included only the texture properties from the depth that mattered the most. Another variable eliminated by the RFE approach was NDVI, for which a plausible explanation could be that although, NDVI is a measure of greenness (hence, plant vigor), there is no literature showing the correlation between contamination in peanut pods below the surface and surface reflection from the canopy. So, NDVI as a predictor did not present any unique input to the model's performance. However, it is also possible that NDVI was a weak predictor because the NDVI values used in the model were all collected at least 90 DAP to coincide with the collection of physiological parameters and whole plant samples for aflatoxin testing. By then, NDVI values were mostly above 0.75 indicating a closed canopy and within the range

of NDVI saturation. A closed canopy and NDVI saturation are likely not a good assessment of plant biomass present in the field. NDVI maps from earlier in the growing season before canopy closure should be investigated as potentially better predictors.

A thought-provoking elimination was that of precipitation since, past studies and farmers' experiences have suggested that precipitation and aflatoxin production are inversely proportional. This may have been caused by the fact that only one year of data were used in the model and so between-year comparisons with different precipitation patterns were not possible. Another possible explanation is that precipitation at a field scale might help in reducing aflatoxin contamination, a lot of factors play part in doing so. Spatially varying field parameters like field terrain and soil infiltration rates can lead to spatially varied impact from precipitation events, even when assuming precipitation did not vary in the field. This implies that direct correlation between aflatoxin and precipitation might be weak and can be explained with other dependent variables like VPD and soil temperature in the model.

A probable reason why SWT was not included by RFE as a predictor in the model is that the soil water may not have been depleted beyond a physiological threshold. Peanut plants tend to reduce their transpirative rates in the season when drought conditions show up, making peanuts insensitive to drier conditions at some level. Even though there was substantial temporal variation in SWT, it was not extreme enough to cause physiological stress in peanuts. For example, Figure 2.11 shows that SWT was below 50 kPa throughout the growing season at this node in Field 2023A. Therefore, it can be concluded that water availability during the study period was not limiting making SWT irrelevant, although, it may have played a more important role in water stressed conditions.

The developed PDPs (Figure 3.43-3.50) were used to individually visualize each predictor's relationship with estimated aflatoxin concentrations. The model learned the relationship between aflatoxin concentration and the 8 variables in a 9-dimension response space (8 predictors and aflatoxin concentration). Since, a 9-dimensional response space cannot be easily plotted and interpreted (at least with the use of tools available to us), PDPs helped us plot the relationship into 8, 2-dimensional, graphs. Starting from the two most important features (based on change in R^2), the PDP for air temperature (Figure 3.43) showed a linear increasing trend and solar radiation (Figure 3.44) did not show an effect on aflatoxin but had a directly proportional trend for higher values, implying that with increasing temperature and ambient heat, aflatoxin concentrations increased as well. This justifies the fact that elevated temperatures are favorable conditions for *A. flavus* to produce aflatoxin. Although increased ambient temperatures cause soil temperature to increase as well, peanut plants show an overlapping canopy phenomenon towards the peak of peanut plant growth that hinders the direct contact of soil surface with solar radiation and shows a small dip in the average soil temperature when visualized across time (Figure 3.58). Nevertheless, this does not lower aflatoxin concentrations as the peanut pods are already developed and are susceptible to contamination. This phenomenon makes the model understand that lowered soil temperature, on a seasonal scale, increases aflatoxin concentrations. It can be observed clearly in Figure 3.45 that partial dependence of aflatoxin concentration on soil temperature decreases in a certain range of increasing soil temperature but after all increases at very high values. Discussing soil texture properties, it can be interpreted from the PDP for silt % at 30-45 cm depth (Figure 3.46) that it hindered aflatoxin contamination at optimum ranges only. This, as mentioned earlier, is a result of

the fact that peanut pods are developed in this range of the profile, so, to maintain optimum moisture conditions throughout the season, certain soil texture properties are required. On the other hand, at a depth of 75-90 cm, a soil texture with lower clay (Figure 3.48), higher sand particles (Figure 3.47.) and lower silt particles (Figure 3.50) can lead to lowered levels of aflatoxin concentrations. This is because at deeper depths, proper drainage of infiltrated water is necessary for plant roots to breathe and develop peanut pods that are less prone to aflatoxin contamination. PDP developed for VPD (Figure 3.49) depicts a positive relationship, implying that as the ambient conditions grow drier and hotter, or in other words, when evapo-transpirative demand increases in the field, aflatoxin production increases. This concludes all the above mentioned PDPs for weather-dependent variables that as the ambient temperature grows and moisture in the air decreases, aflatoxin contamination rises as well.

Model performance might have been impeded by the presence of spatial variability in predictor variables and aflatoxin across the field. Capturing spatial variation in aflatoxin appropriately requires high-granularity soil sampling and analysis, due to its dependency on certain conditions favorable to the fungi and being highly sensitive to even small changes in its influencers. Although the model uses a wide range of space-varying predictors to explain variation in aflatoxin concentration in peanut kernels, it does not account for spatial trends.

In the model framework, locations and time of sampling were pooled together into one training dataset, making space and time equivalent for training. We did not attempt to tease apart spatial and temporal part of variation in the model framework, because considering space time equivalency allowed us to increase the data available to train on.

Combining spatial and temporally collected datapoints will increase the volume of training dataset as opposed to a scenario where each is considered individually. Explicit inclusion of location latitude and longitude would have provided opportunity to quantitatively address spatial variability using continuous variables. However, if spatial variability was to be accounted for in the model, it would have been more reasonable to have a greater spatial resolution of sampling locations to accurately capture finer level spatial variability that is realistic for a commercial production field. The reason we chose not to explicitly account for spatial attributes is to maintain universality for model application. For example, if field-specific attributes in fact were included as predictors, the model's applicability would have been restricted to the experimental field. Whereas the intention was to develop a model that could be used irrespective of the target location.

Tables:

Table 3.1. Extracted NDVI values for sampling points in field 2023A by sampling date.

Field 2023A Plot Number	NDVI by Sampling Date			
	7/5/2023	7/26/2023	8/8/2023	8/22/2023
1	0.645	0.824	0.831	0.823
2	0.689	0.818	0.848	0.820
3	0.781	0.859	0.855	0.854
4	0.815	0.863	0.852	0.854
5	0.808	0.865	0.863	0.860
6	0.790	0.858	0.863	0.855
7	0.762	0.849	0.855	0.849
8	0.710	0.850	0.845	0.848
9	0.730	0.828	0.834	0.818
10	0.690	0.816	0.835	0.807
11	0.712	0.836	0.841	0.828
12	0.790	0.873	0.856	0.848
13	0.794	0.874	0.862	0.865
14	0.813	0.875	0.864	0.861
15	0.806	0.880	0.875	0.870
16	0.852	0.880	0.874	0.874
17	0.849	0.881	0.875	0.874
18	0.820	0.874	0.864	0.871
19	0.796	0.877	0.875	0.869
20	0.800	0.874	0.869	0.865
21	0.822	0.871	0.863	0.868
22	0.787	0.860	0.860	0.854
23	0.740	0.833	0.849	0.832
24	0.774	0.854	0.849	0.840
25	0.798	0.841	0.856	0.842
26	0.799	0.858	0.862	0.852
27	0.842	0.871	0.864	0.860
28	0.857	0.880	0.867	0.865
29	0.849	0.868	0.857	0.863
30	0.845	0.878	0.865	0.863

Table 3.2. Extracted NDVI values for sampling points in field 2023B by sampling date.

Field 2023B Plot Number	NDVI by Sampling Date			
	7/7/2023	7/25/2023	8/9/2023	8/22/2023
1	0.734	0.726	0.890	0.810
2	0.728	0.744	0.868	0.823
3	0.741	0.808	0.871	0.814
4	0.752	0.784	0.891	0.815
5	0.736	0.745	0.893	0.782
6	0.757	0.783	0.900	0.784
7	0.785	0.823	0.911	0.829
8	0.769	0.822	0.912	0.824
9	0.765	0.780	0.915	0.826
10	0.755	0.726	0.892	0.819
11	0.761	0.821	0.915	0.829
12	0.767	0.844	0.915	0.810
13	0.760	0.795	0.897	0.805
14	0.771	0.815	0.900	0.804
15	0.817	0.833	0.917	0.837
16	0.799	0.821	0.913	0.828
17	0.782	0.840	0.914	0.818
18	0.734	0.726	0.890	0.810
19	0.728	0.744	0.868	0.823
20	0.741	0.808	0.871	0.814
21	0.752	0.784	0.891	0.815

Table 3.3. Extracted NDVI values for sampling points in field 2023C by sampling date.

Field 2023C Plot Number	NDVI by Sampling Date		
	8/1/2023	8/15/2023	9/7/2023
1	0.944	0.862	0.895
2	0.950	0.872	0.901
3	0.939	0.875	0.896
4	0.945	0.867	0.896
5	0.941	0.866	0.892
6	0.943	0.866	0.894
7	0.944	0.862	0.884
8	0.932	0.860	0.898
9	0.930	0.858	0.907
10	0.942	0.874	0.905
11	0.947	0.873	0.894
12	0.950	0.886	0.894
13	0.946	0.868	0.887
14	0.946	0.864	0.895
15	0.936	0.855	0.896
16	0.935	0.870	0.902
17	0.934	0.854	0.892
18	0.930	0.869	0.893
19	0.943	0.866	0.898
20	0.937	0.869	0.905
21	0.880	0.841	0.890
22	0.939	0.868	0.897
23	0.939	0.858	0.899
24	0.931	0.849	0.898
25	0.944	0.862	0.898
26	0.936	0.858	0.891

Table 3.4. Table showing summary of aflatoxin concentrations observed in each field.

Field	Aflatoxin Concentration (ppb)		
	Mean	Standard Deviation	Maximum
2023A	0.73	0.63	2.30
2023B	1.03	0.80	3.10
2023C	0.83	0.81	3.50
All Fields	0.86	0.75	3.50

Table 3.5. Table showing Observed and Predicted Aflatoxin values

Observed Aflatoxin (ppb)	Predicted Aflatoxin (ppb)
0	1.018
60	1.043
390	0.97
3.3	1.016
68	0.853
30	1.034
130	0.992
42	0.915
270	1.174
1	0.925
4.8	0.813
120	1.06
50	1.179

Observed Aflatoxin (ppb)	Predicted Aflatoxin (ppb)
75	1.45
100	0.972
190	1.609
2	0.966
430	1.495
88	1.537
48	1.012
57	1.067
55	1.116
58	1.598
28	1.22
220	1.104
75	1.282
0	1.52
0.041	0.918
92	1.562
62	1.716
0	1.067

Figures:

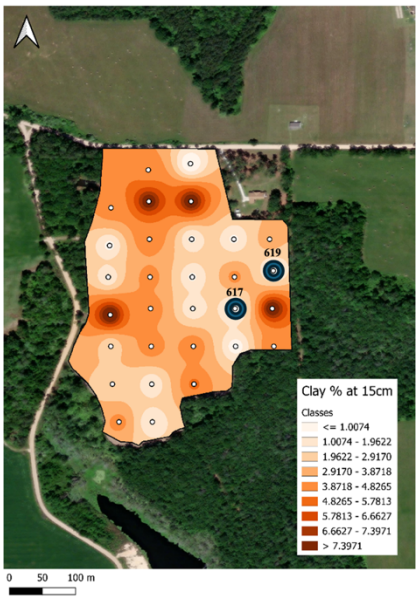

Figure 3.1. Interpolated map developed for clay percentage at 0-15cm depth in field 2023A. The depicted points are the locations of the SWT sensors from Figure 3.19 and 3.20.

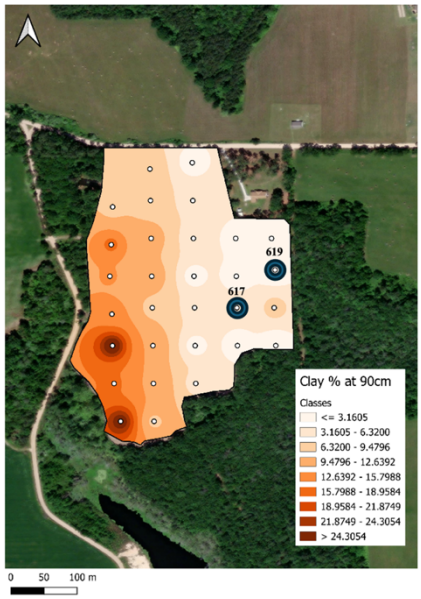

Figure 3.2. Interpolated map developed for clay percentage at 75-90cm depth in field 2023A. The depicted points are the locations of the SWT sensors from Figure 3.19 and 3.20.

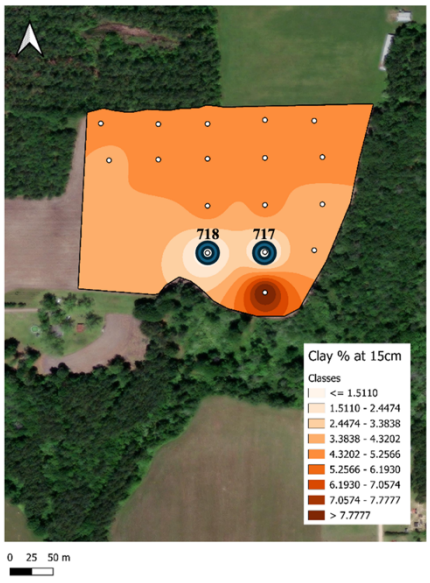

Figure 3.3. Interpolated map developed for clay percentage at 0-15cm depth in field 2023B. The depicted points are the locations of the SWT sensors from Figure 3.17 and 3.18

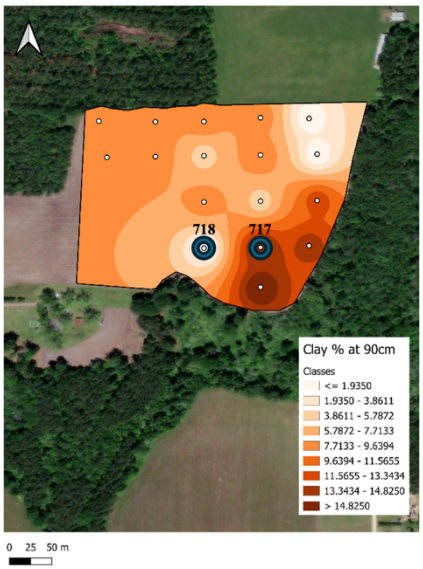

Figure 3.4. Interpolated map developed for clay percentage at 75-90cm depth in field 2023B. The depicted points are the locations of the SWT sensors from Figure 3.17 and 3.18

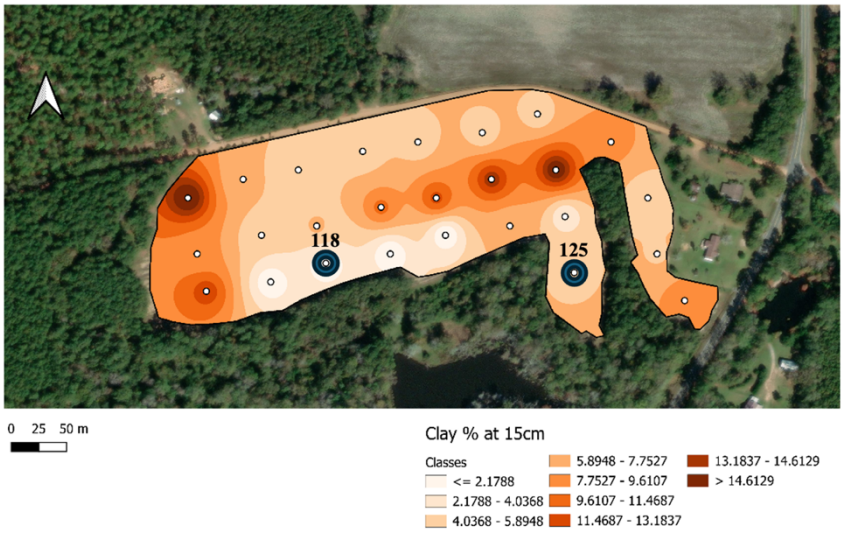

Figure 3.5. Interpolated map developed for clay percentage at 0-15cm depth in field 2023C. The depicted points are the locations of the SWT sensors from Figure 3.21 and 3.22

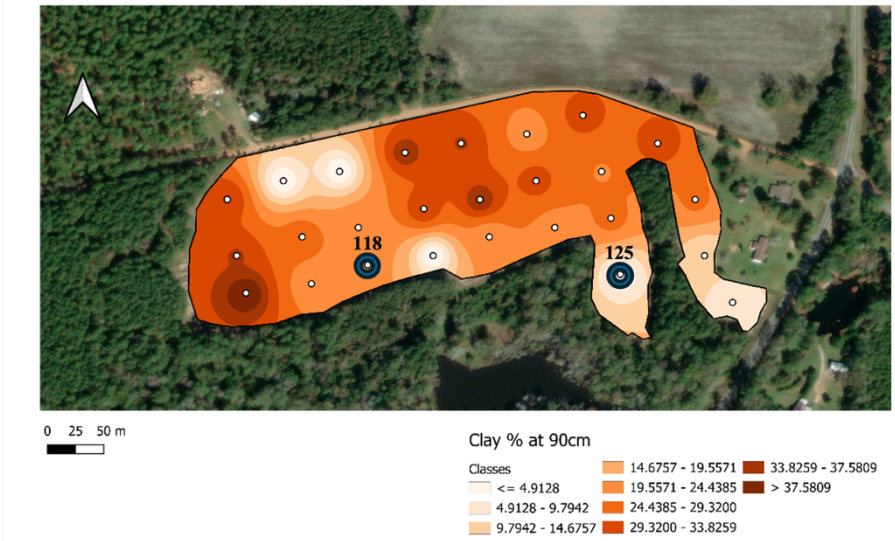

Figure 3.6. Interpolated map developed for clay percentage at 75-90cm depth in field 2023C. The depicted points are the locations of the SWT sensors from Figure 3.21 and 3.22

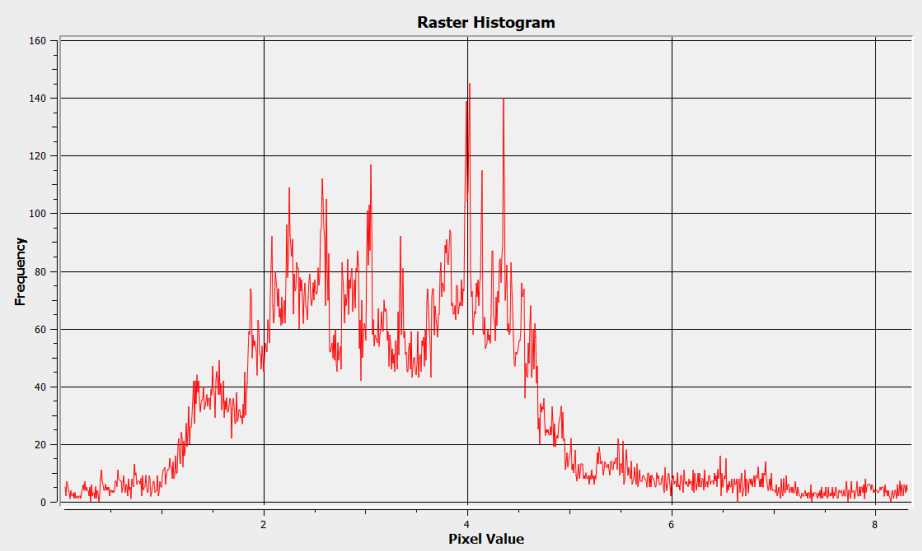

Figure 3.7. Raster Histogram developed for clay percentage at 0-15cm depth in field 2023A.

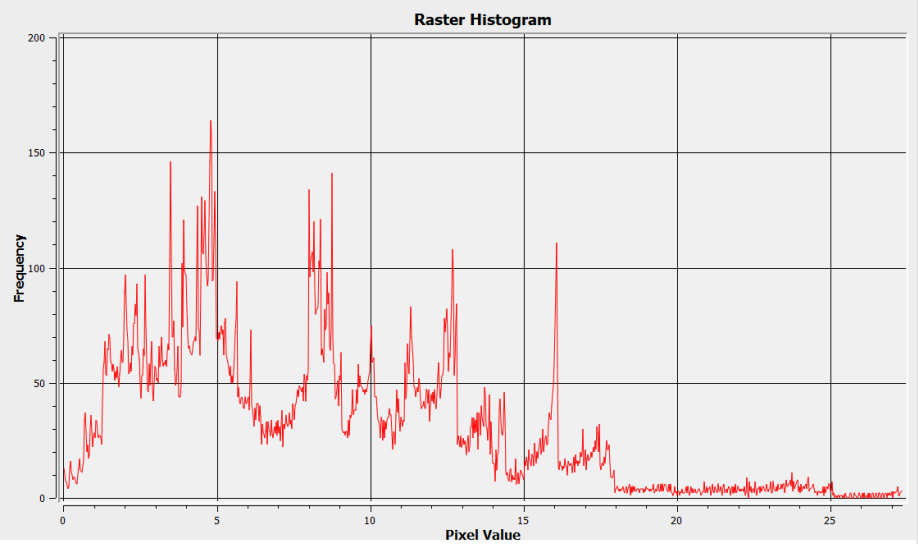

Figure 3.8. Raster Histogram developed for clay percentage at 75-90cm depth in field 2023A.

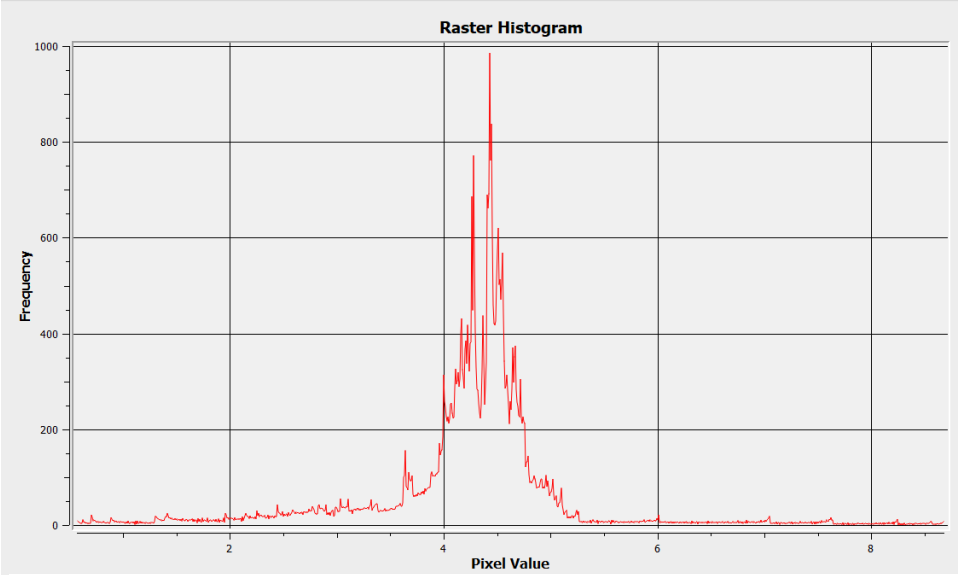

Figure 3.9. Raster Histogram developed for clay percentage at 0-15cm depth in field 2023B.

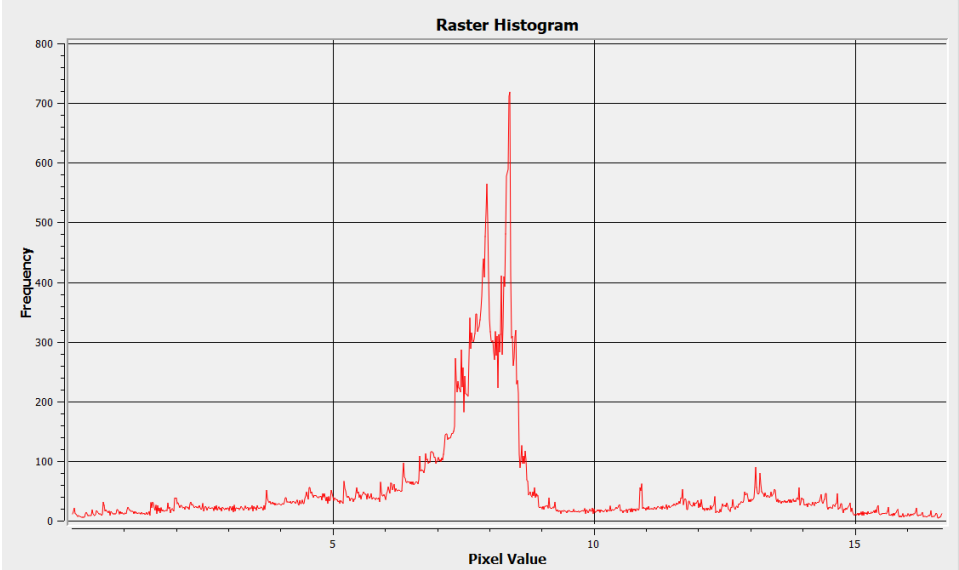

Figure 3.10. Raster Histogram developed for clay percentage at 75-90cm depth in field 2023B.

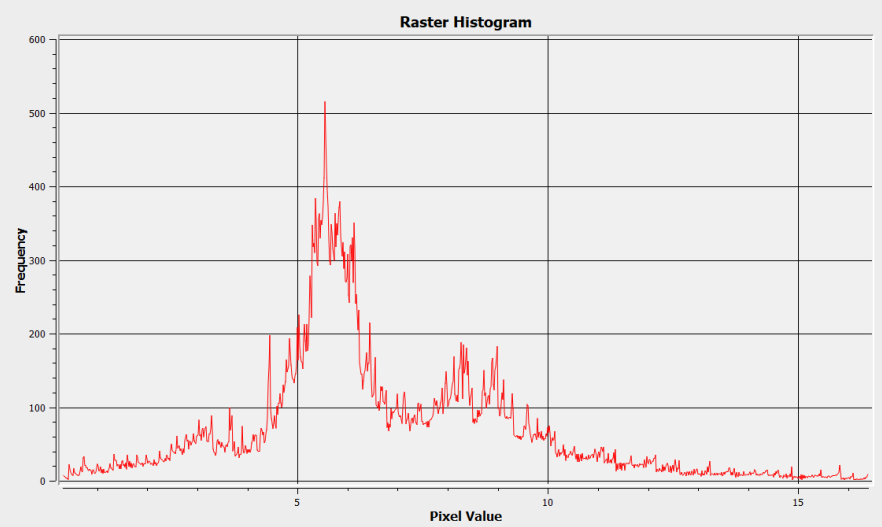

Figure 3.11. Raster Histogram developed for clay percentage at 0-15cm depth in field 2023C.

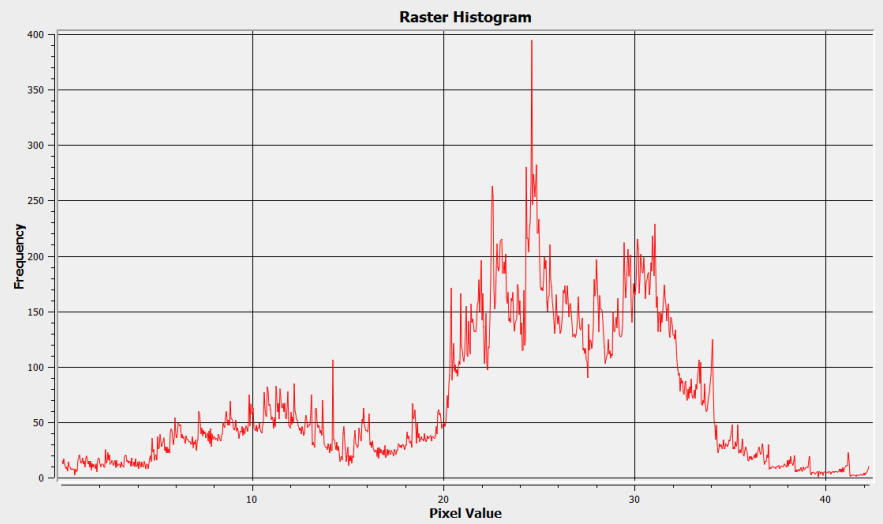

Figure 3.12. Raster Histogram developed for clay percentage at 75-90cm depth in field 2023C.

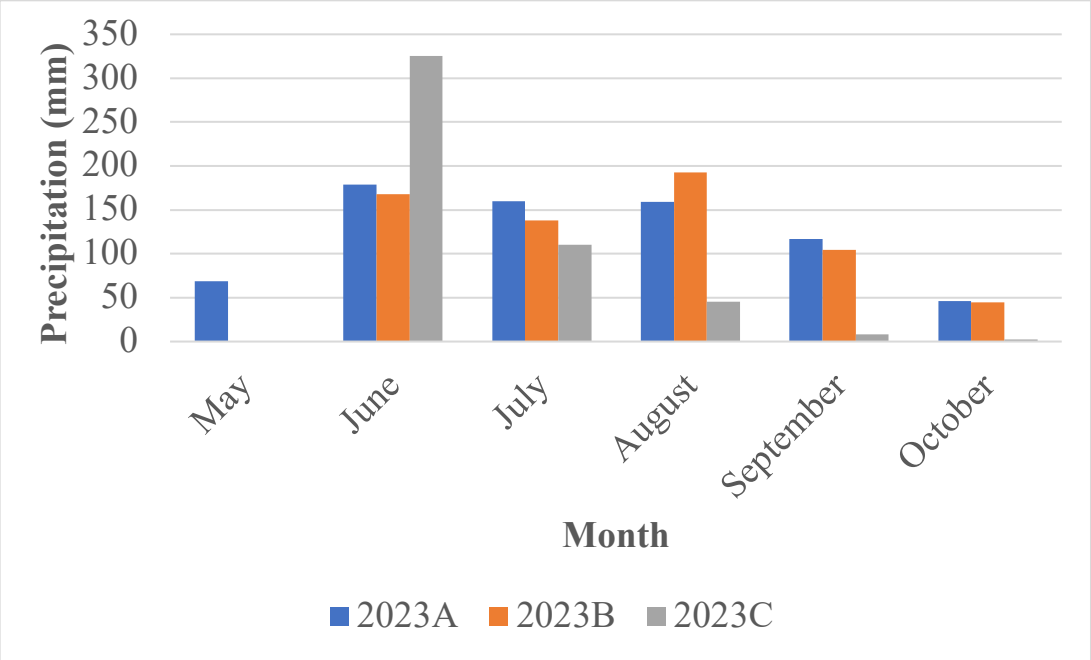

Figure 3.13. Bar graph showing seasonal trends in each field for precipitation.

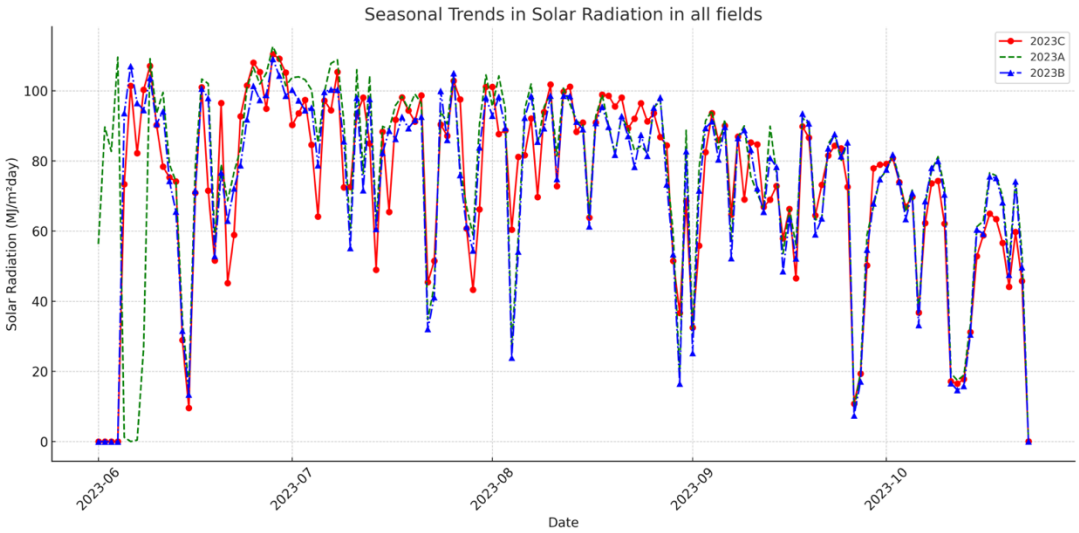

Figure 3.14. Trendline plot for Daily total solar radiation in each of the fields.

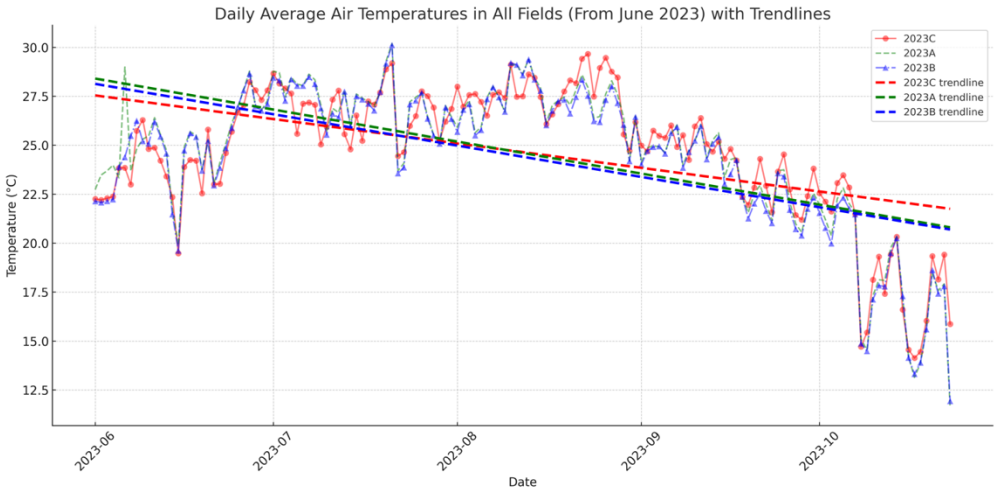

Figure 3.15. Trendline plot for Daily average air temperature in each of the fields along with regression lines.

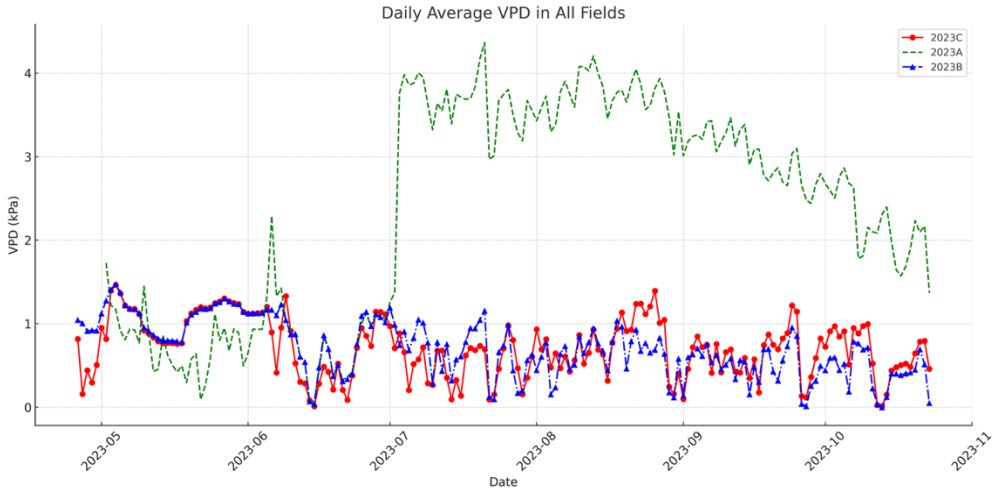

Figure 3.16. Trendline plot for Daily average vapor pressure deficit in each of the fields

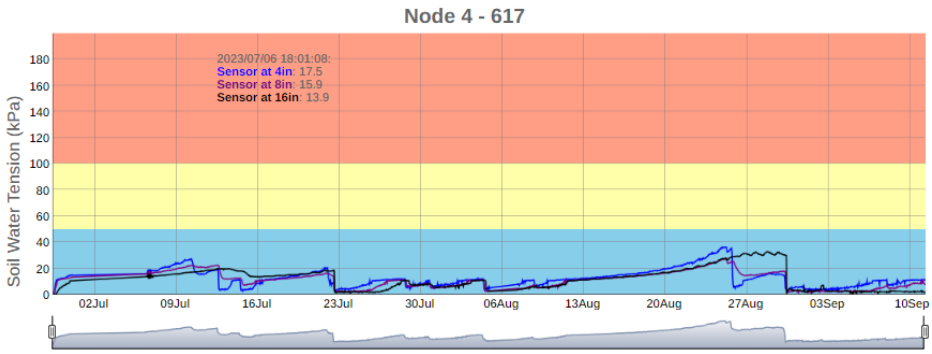

Figure 3.17. Temporal trend plot with low water tension conditions developed for SWT in Field 2023A.

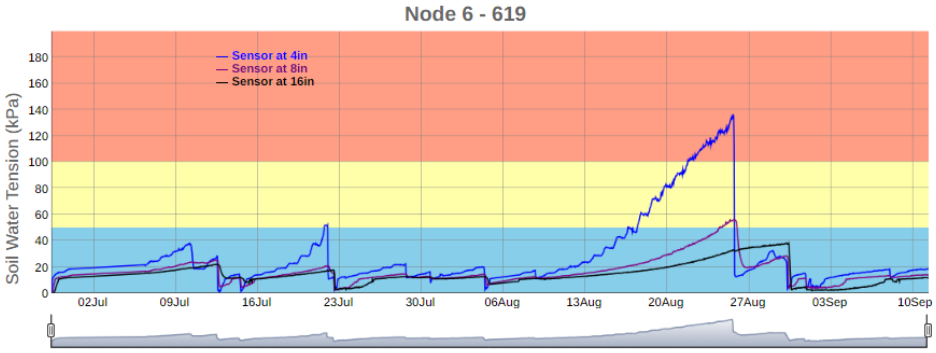

Figure 3.18. Temporal trend plot with high water tension conditions developed for SWT in Field 2023A.

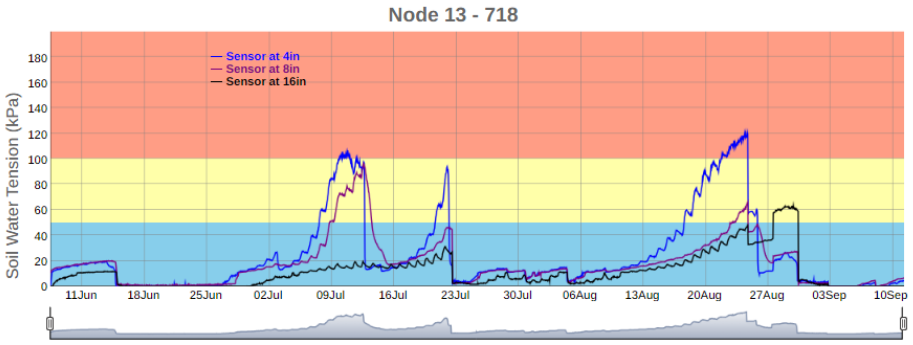

Figure 3.19. Temporal trend plot with high water tension conditions developed for SWT in Field 2023B.

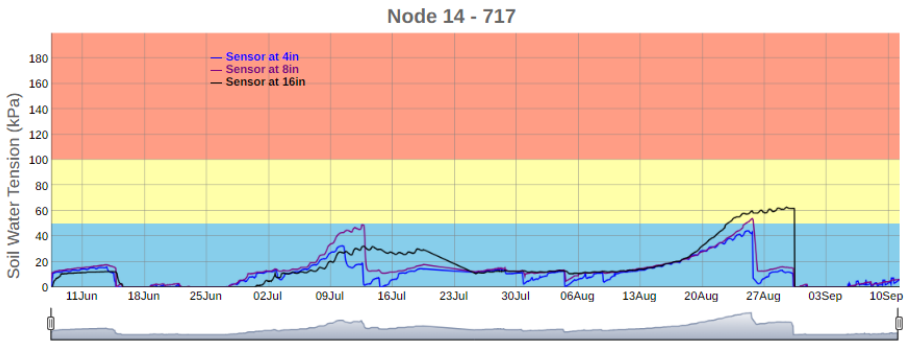

Figure 3.20. Temporal trend plot with low water tension conditions developed for SWT in Field 2023B.

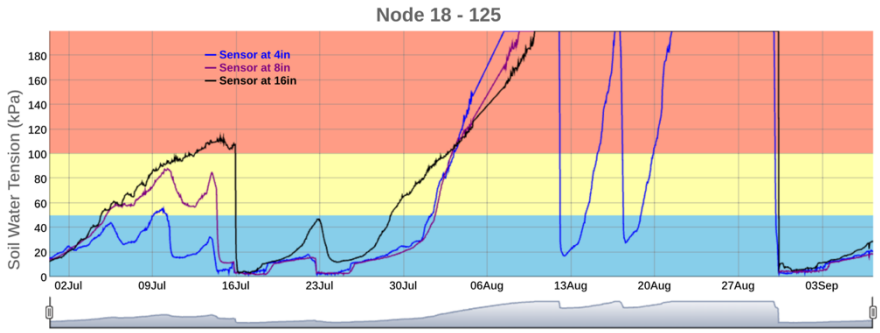

Figure 3.21. Temporal trend plot with high water tension conditions developed for SWT in Field 2023C.

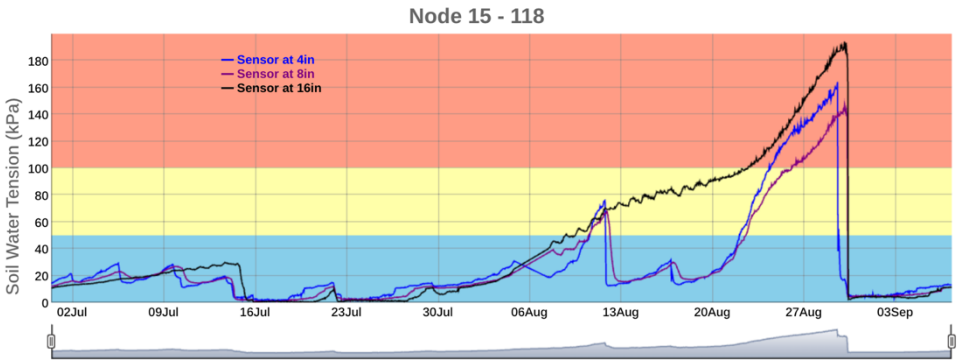

Figure 3.22. Temporal trend plot with low water tension conditions developed for SWT in Field 2023C.

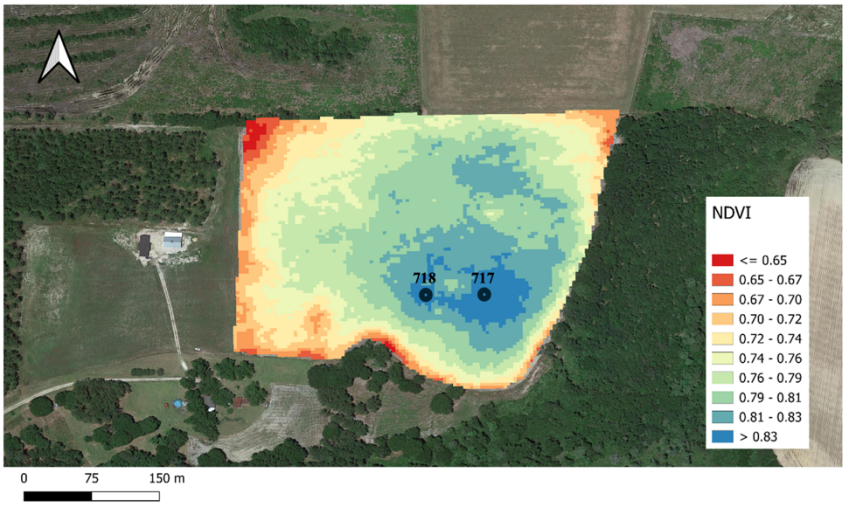

Figure 3.23. NDVI map developed for field 2023B from satellite imagery taken on August 9, 2023. The depicted points are the locations of the SWT sensors from Figure 3.19 and 3.20

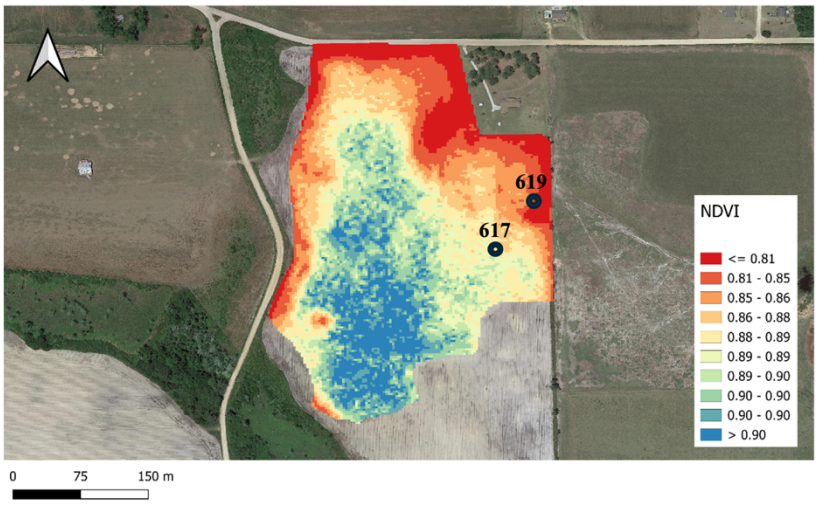

Figure 3.24. NDVI map developed for field 2023C from satellite imagery taken on August 15, 2023. The depicted points are the locations of the SWT sensors from Figure 3.17 and 3.18

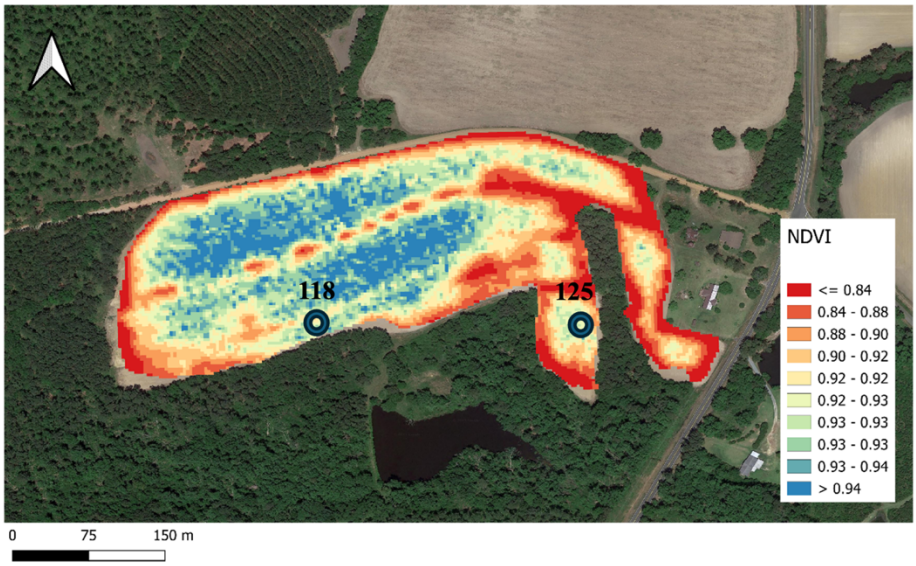

Figure 3.25. NDVI map developed for field 2023A from satellite imagery taken on August 8, 2023. The depicted points are the locations of the SWT sensors from Figure 3.21 and 3.22

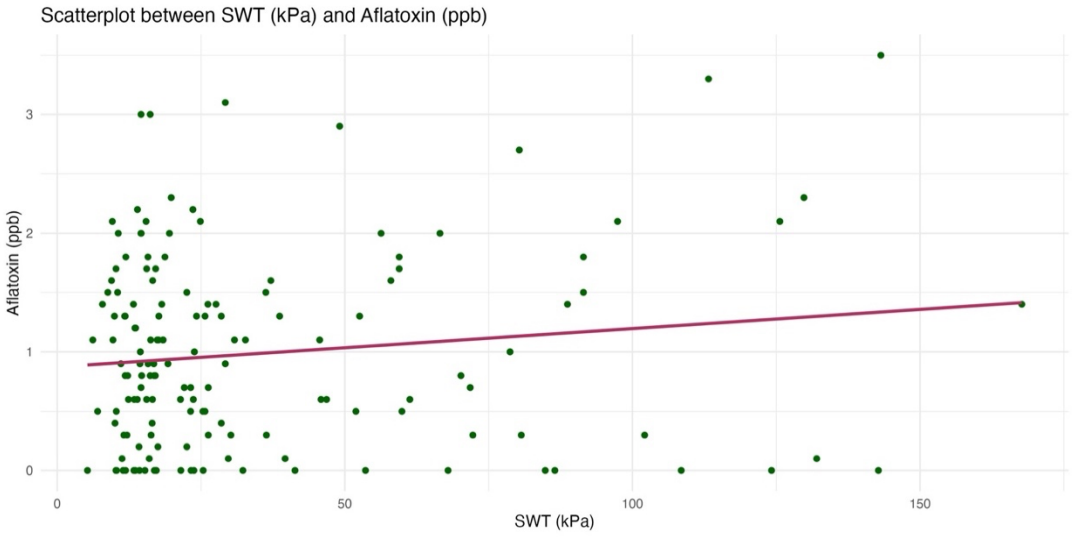

Figure 3.26. Scatterplot with a linear regression line between SWT (kPa) and Aflatoxin (ppb).

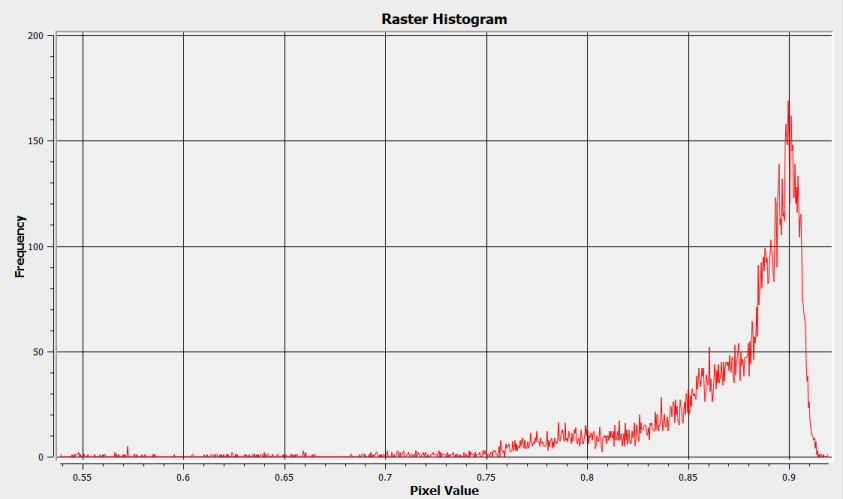

Figure 3.27. Raster Histogram developed for NDVI from image taken on July 25, 2023 in field 2023B.

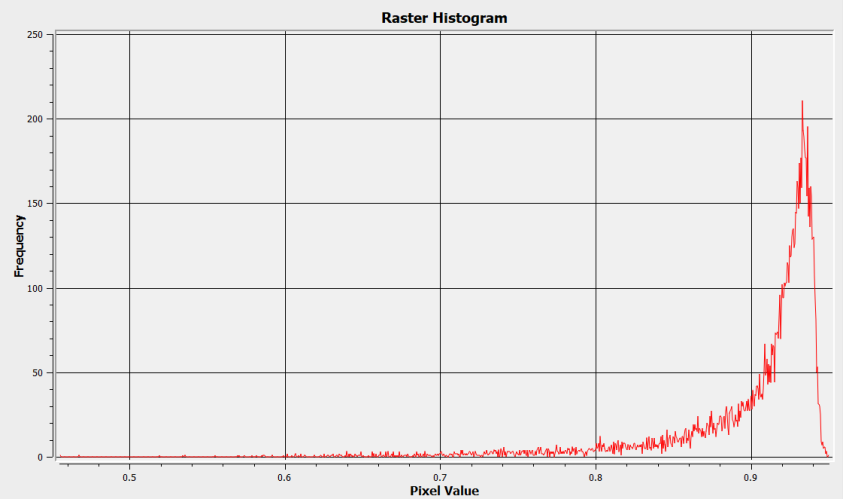

Figure 3.28. Raster Histogram developed for NDVI from image taken on August 1, 2023 in field 2023C.

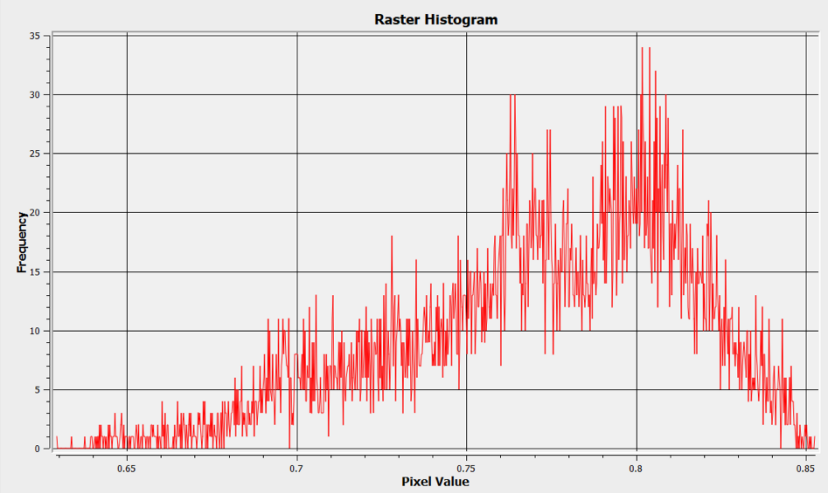

Figure 3.29. Raster Histogram developed for NDVI from image taken on July 26, 2023 in field 2023A.

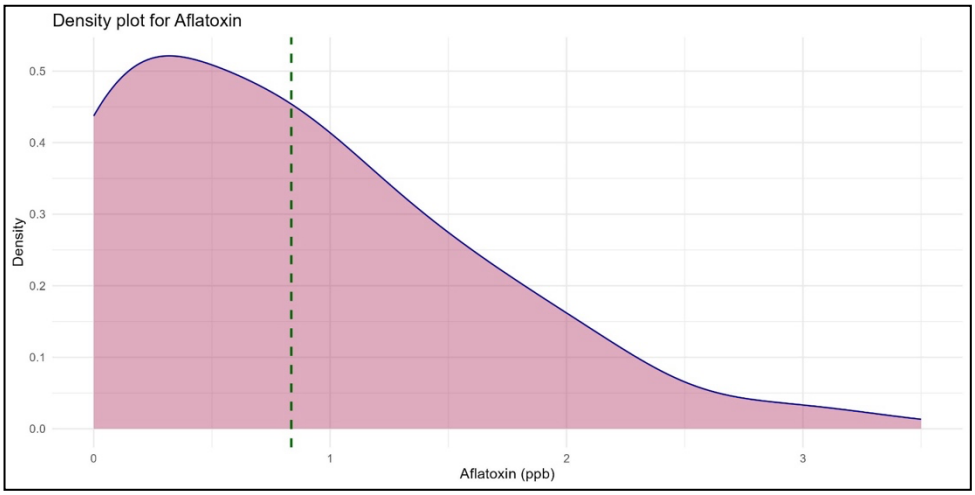

Figure 3.30. Density plot for Aflatoxin distribution in database with a vertical mean line

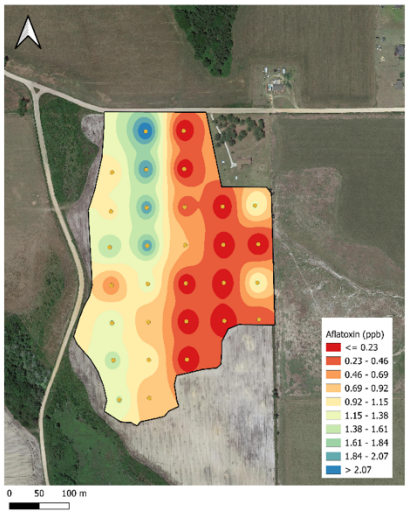

Figure 3.31. Interpolated map of aflatoxin concentrations at the harvest for field 2023A

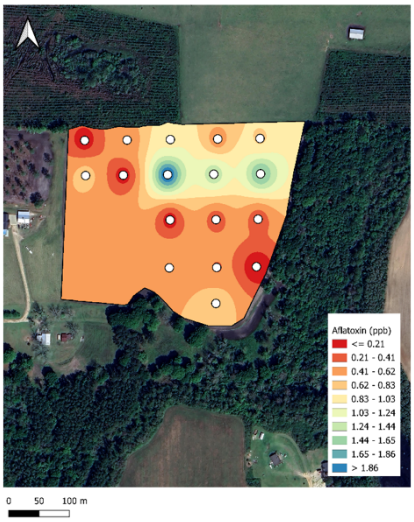

Figure 3.32. Interpolated map of aflatoxin concentrations at the harvest for field 2023B

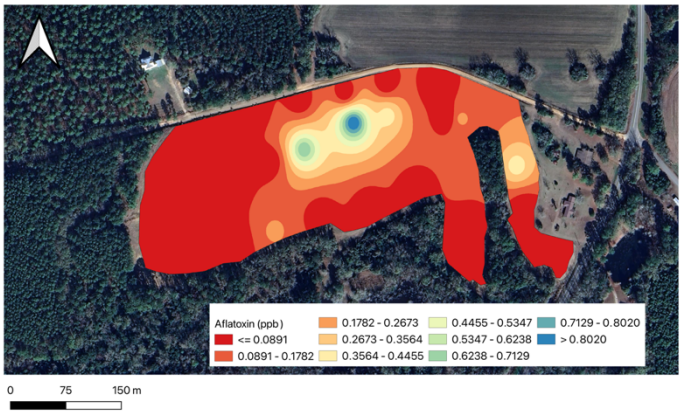

Figure 3.33. Interpolated map of aflatoxin concentrations at the harvest for field 2023C

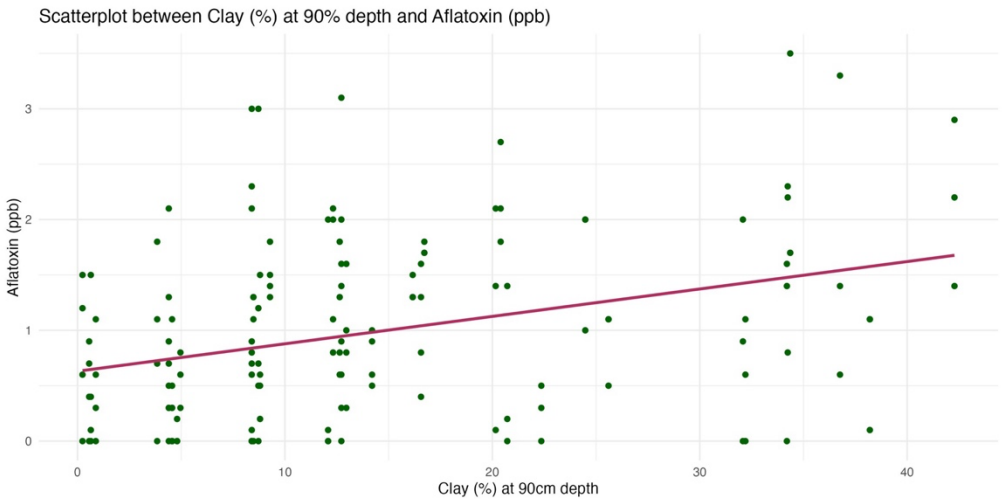

Figure 3.34. Scatter plot between clay percentage at 75-90cm of depth and aflatoxin showing positive trend.

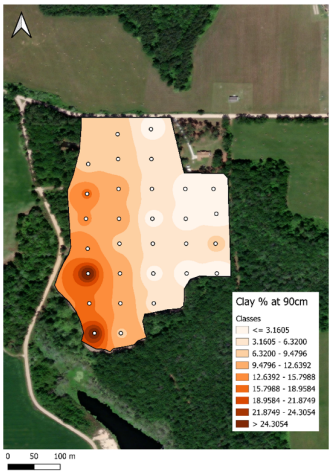

Figure 3.35 (a). Interpolated map developed for clay percentage at 15cm depth in field 2023A.

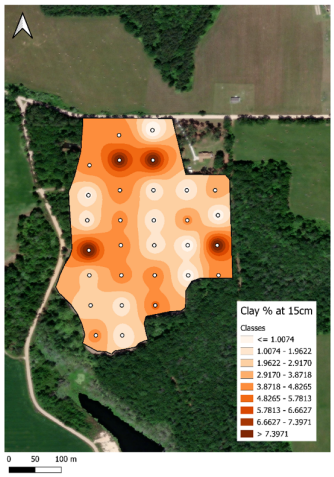

Figure 3.35 (b). Interpolated map developed for clay percentage at 90cm depth in field 2023A.

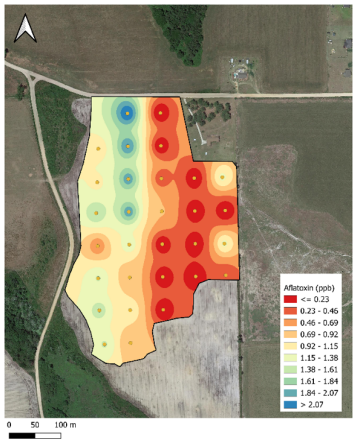

Figure 3.35 (c). Interpolated map of aflatoxin concentrations at the harvest for field 2023A

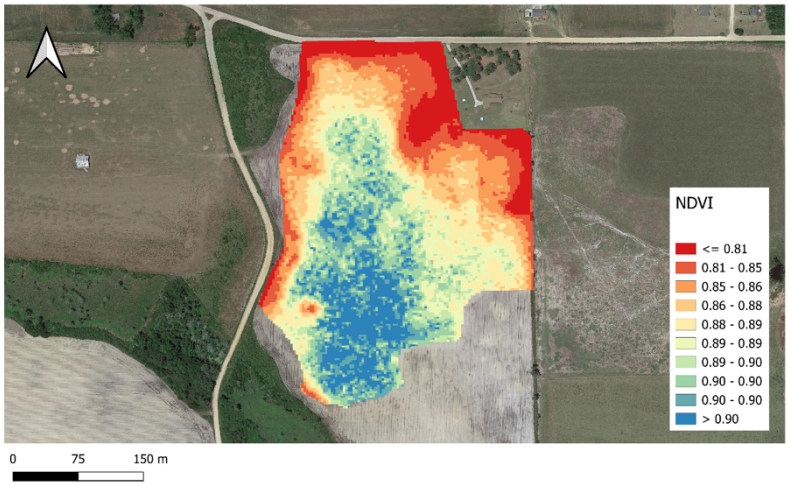

Figure 3.35 (d). NDVI map developed for field 2023C from satellite imagery taken on August 15, 2023

Figure 3.35 Complied maps for (a) Clay% - 90cm ; (b) Clay% - 15cm; (c) Aflatoxin concentrations; (d) NDVI for the 2023A field

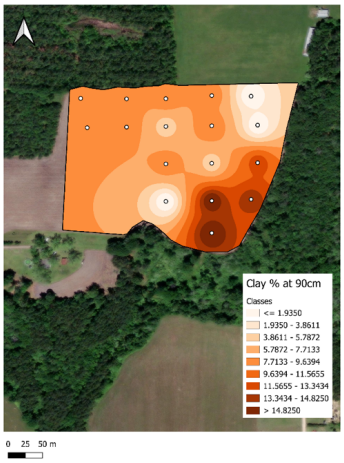

Figure 3.36 (a). Interpolated map developed for clay percentage at 90cm depth in field 2023B.

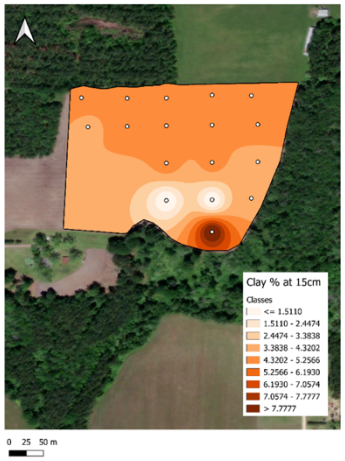

Figure 3.36 (b). Interpolated map developed for clay percentage at 15cm depth in field 2023B.

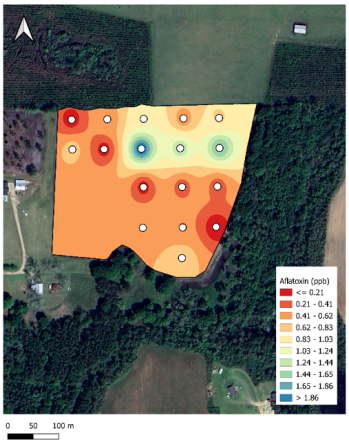

Figure 3.36 (c). Interpolated map of aflatoxin concentrations at the harvest for field 2023B

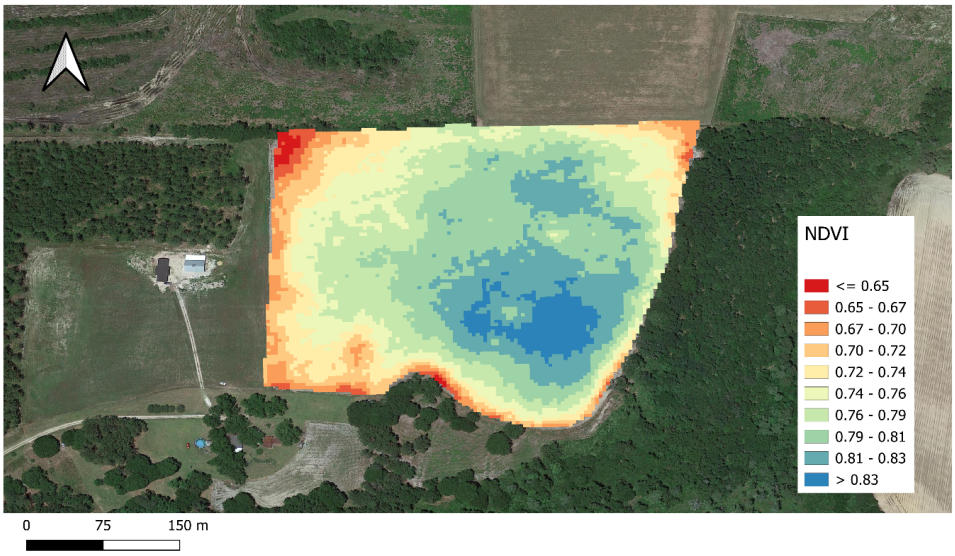

Figure 3.36 (d). NDVI map developed for field 2023B from satellite imagery taken on August 9, 2023

Figure 3.36. Complied maps for (a) Clay% - 90cm ; (b) Clay% - 15cm; (c) Aflatoxin concentrations; (d) NDVI for the 2023B field

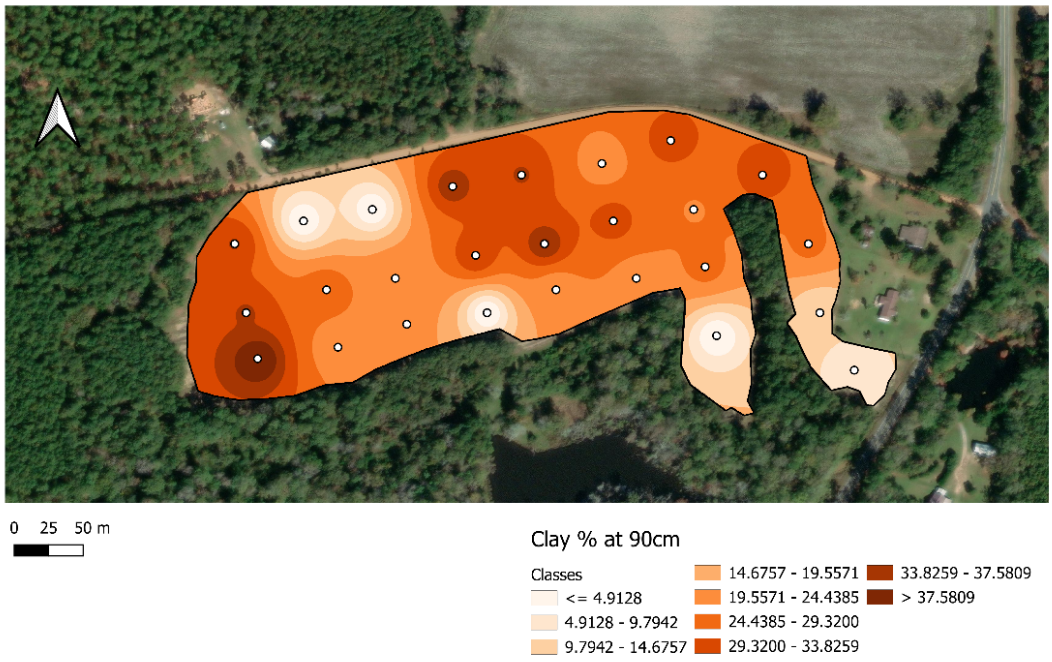

Figure 3.37 (a). Interpolated map developed for clay percentage at 90cm depth in field 2023C.

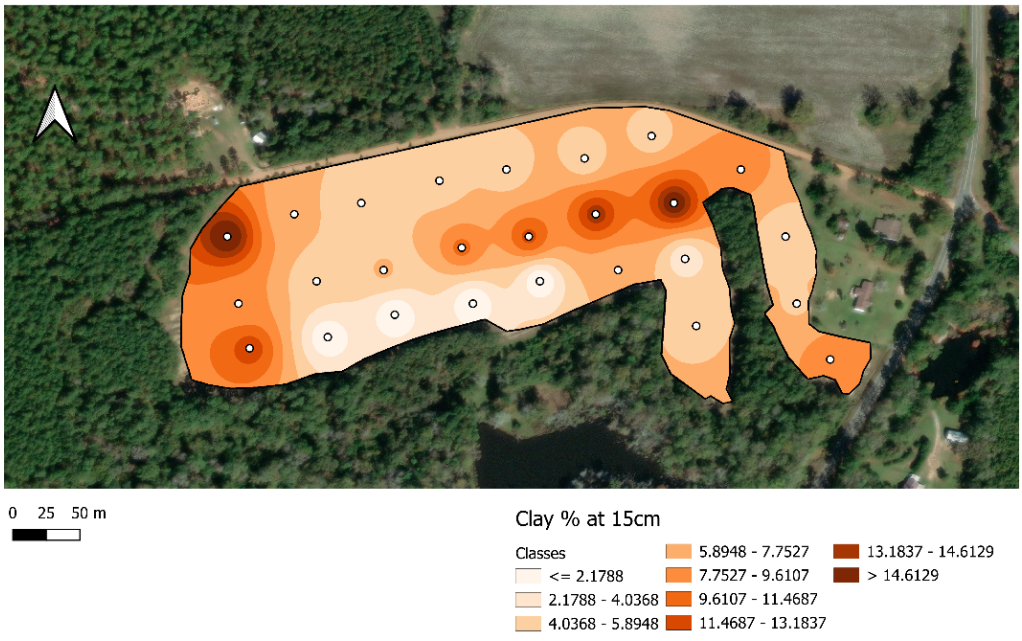

Figure 3.37 (b). Interpolated map developed for clay percentage at 15cm depth in field 2023C.

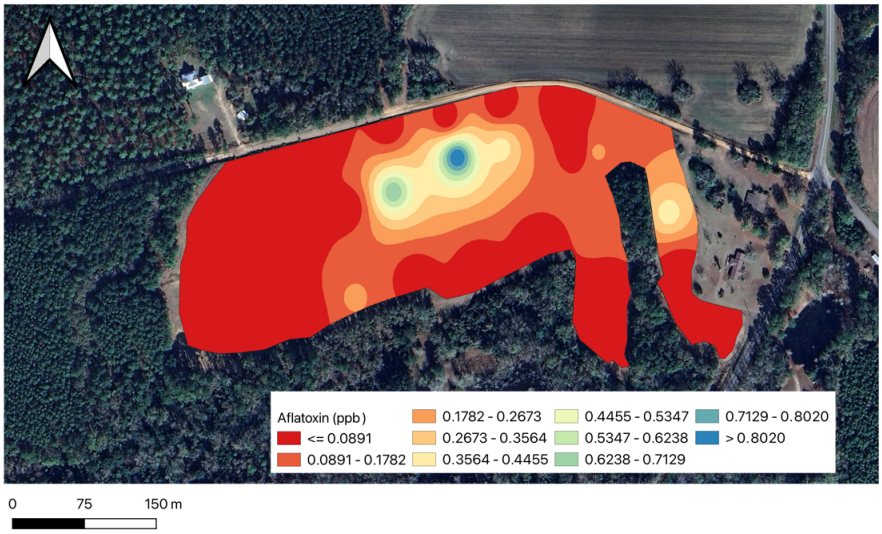

Figure 3.37 (c). Interpolated map of aflatoxin concentrations at the harvest for field 2023C

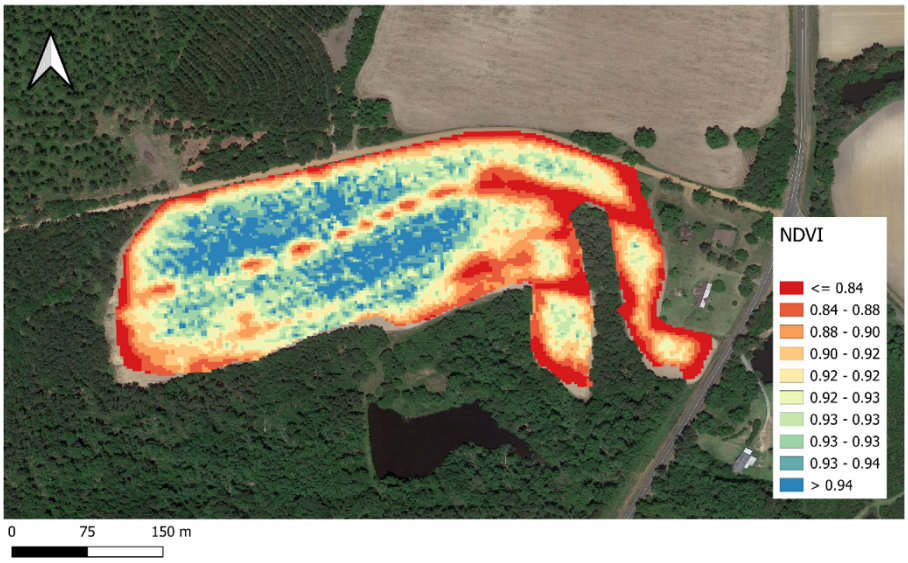

Figure 3.37 (d). NDVI map developed for field 2023A from satellite imagery taken on August 8, 2023

Figure 3.37 Complied maps for (a) Clay% - 90cm ; (b) Clay% - 15cm; (c) Aflatoxin concentrations; (d) NDVI for the 2023C field

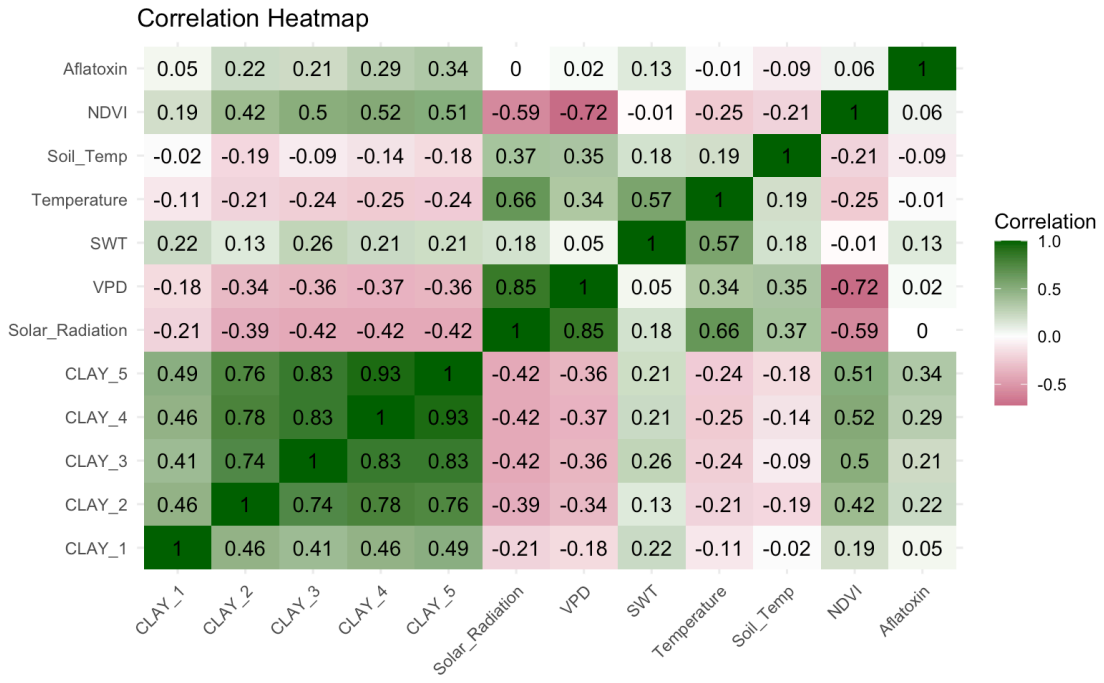

Figure 3.38. Correlation heatmap was developed to quantify interactions between independent variables.

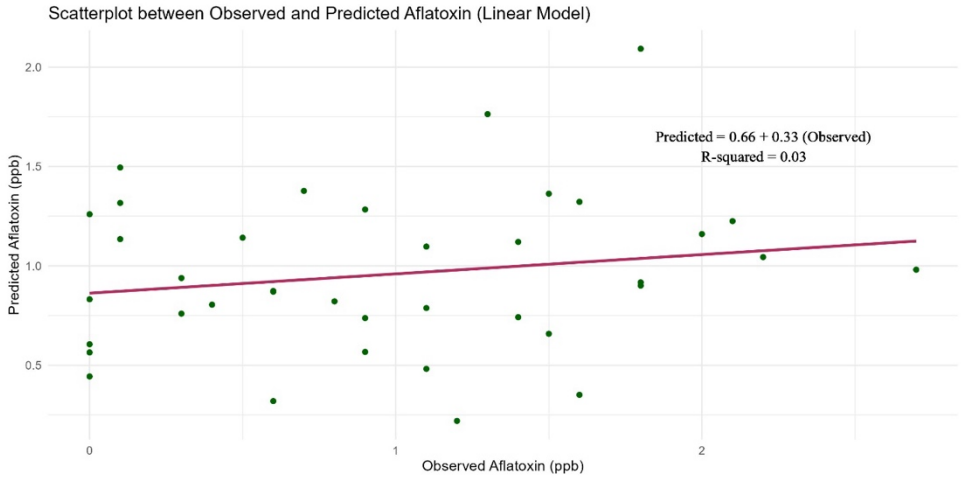

Figure 3.39. Scatter plot between observed and estimated values of aflatoxin from the developed linear model.

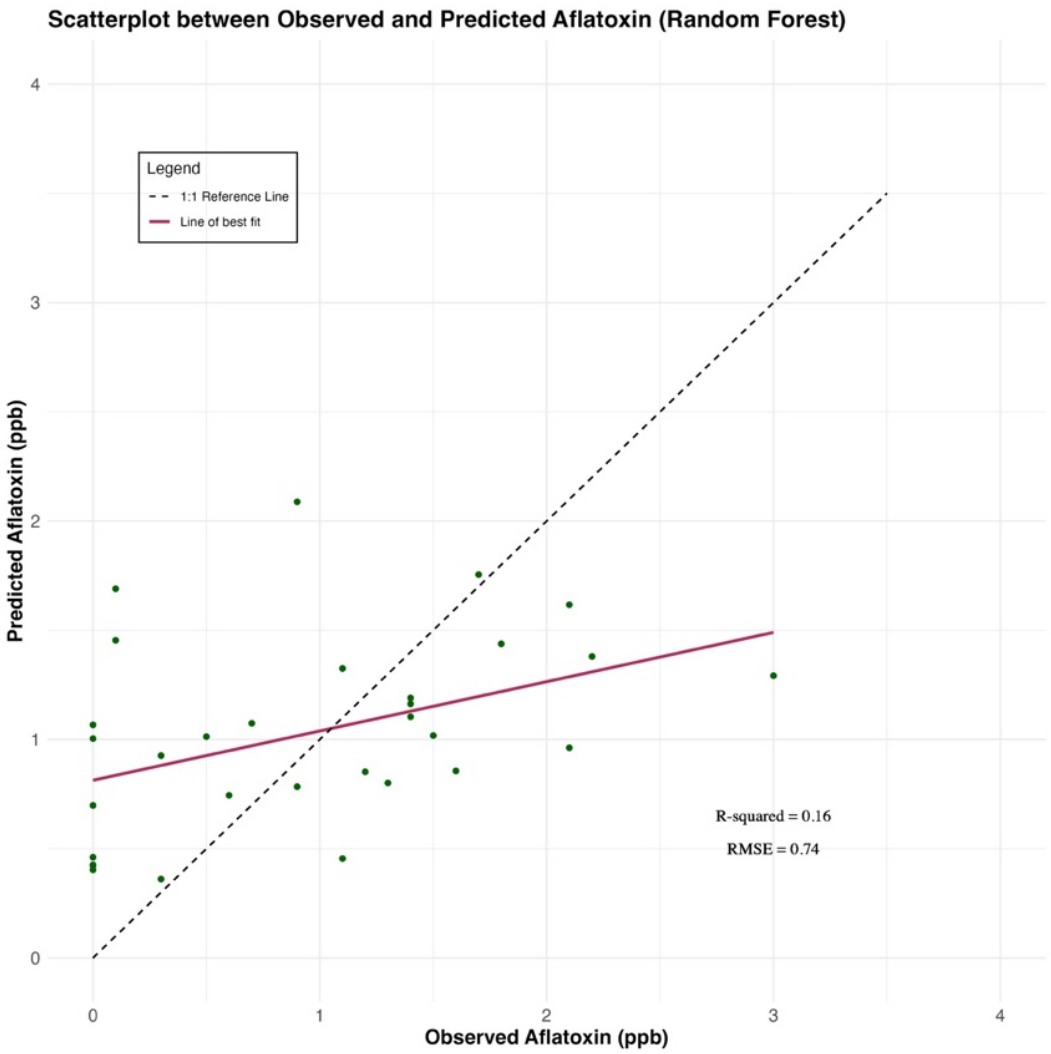

Figure 3.40. Scatter plot between observed and estimated values of aflatoxin from the developed random forest model.

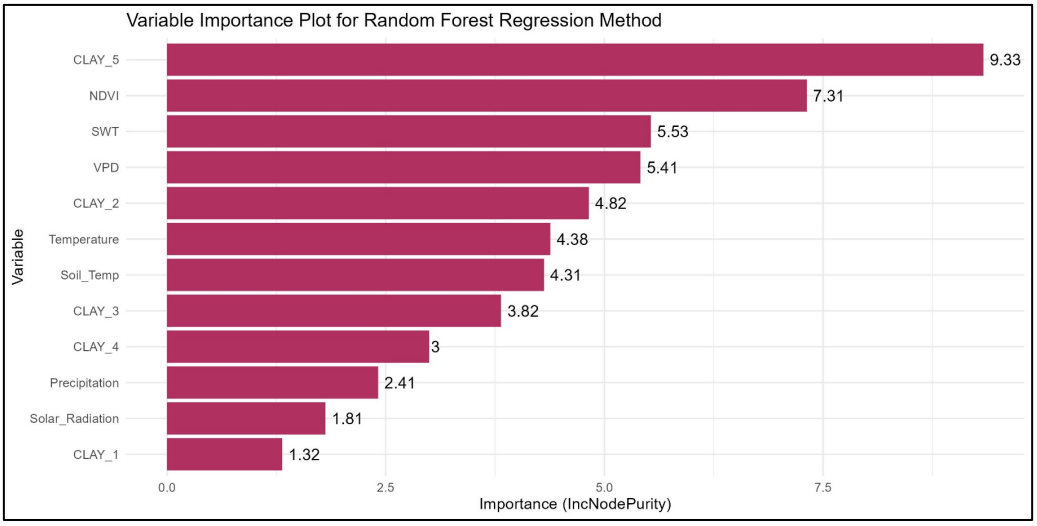

Figure 3.41. Variable Importance Plot for Random Forest Regression Method

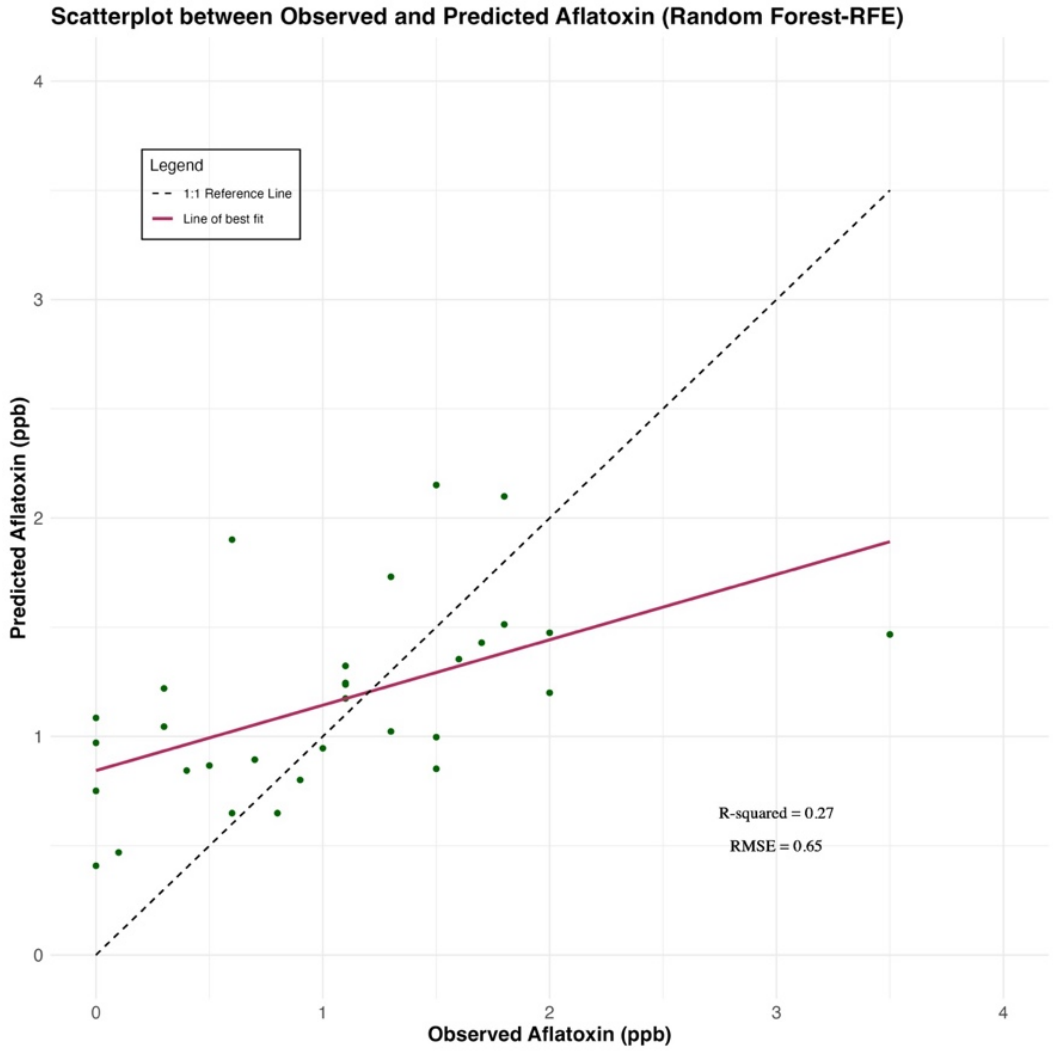

Figure 3.42. Scatter plot between observed and estimated values of aflatoxin from the developed random forest model after feature elimination method

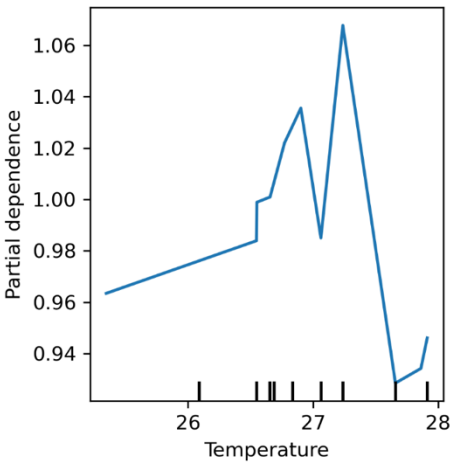

Figure 3.43. Partial Dependence of Random Forest Model on air temperature (^{0}C)

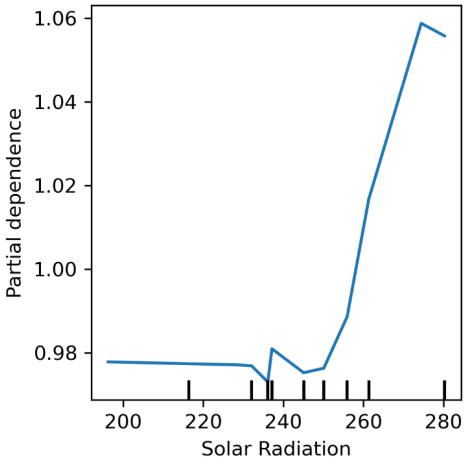

Figure 3.44. Partial Dependence of Random Forest Model on Solar Radiation

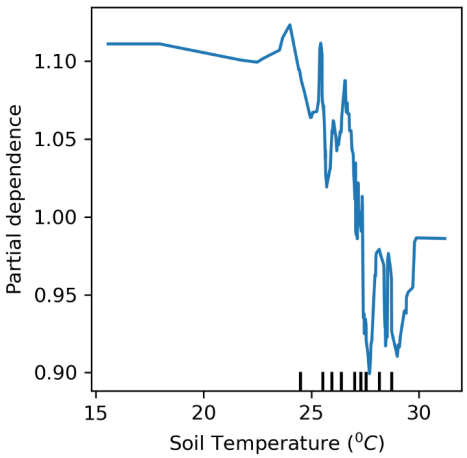

Figure 3.45. Partial Dependence of Random Forest Model on Soil Temperature

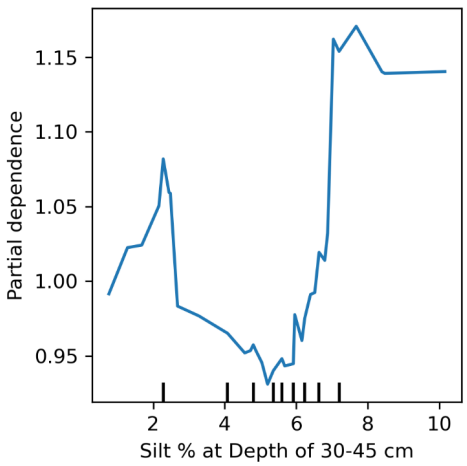

Figure 3.46. Partial Dependence of Random Forest Model on Silt % at depth of 30-45 cm

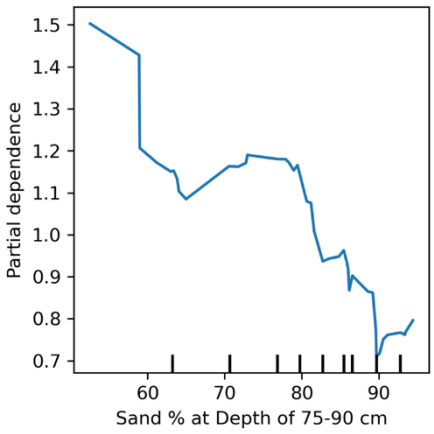

Figure 3.47 Partial Dependence of Random Forest Model on Sand % at depth of 75-90 cm

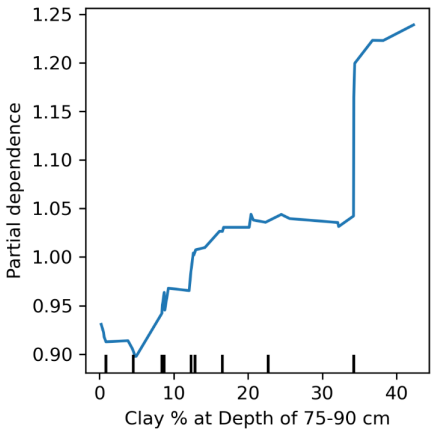

Figure 3.48 Partial Dependence of Random Forest Model on Clay % at depth of 75-90 cm

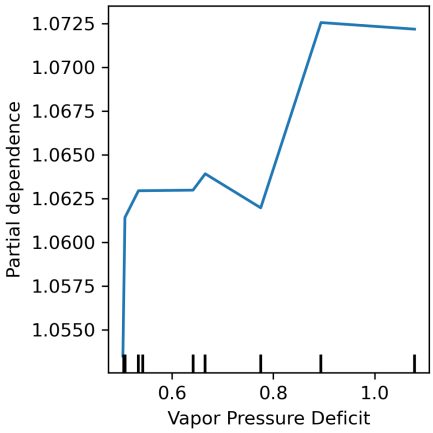

Figure 3.49. Partial Dependence of Random Forest Model on vapor pressure deficit

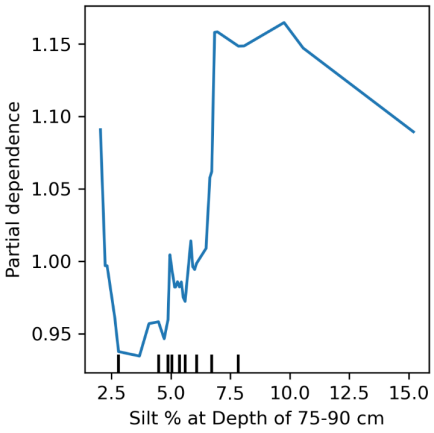

Figure 3.50. Partial Dependence of Random Forest Model on Silt % at depth of 75-90 cm

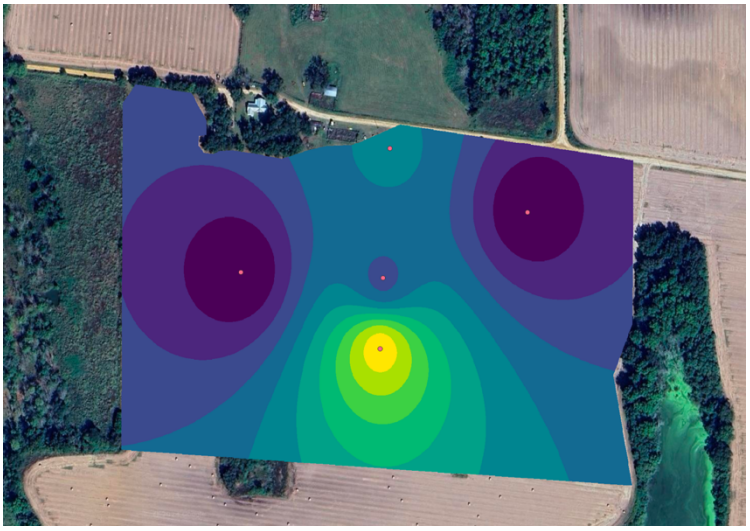

Figure 3.51. Interpolated map for observed aflatoxin values for field 2022A

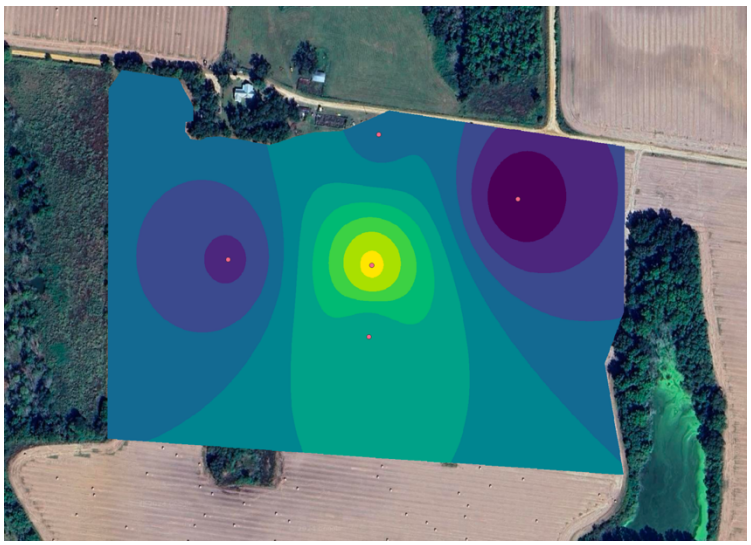

Figure 3.52. Interpolated map for predicted aflatoxin values for field 2022A

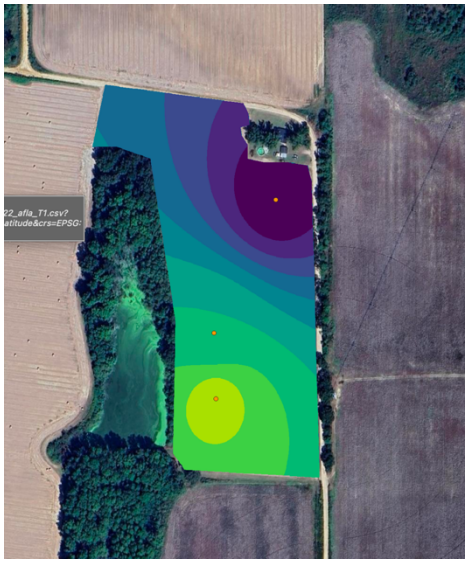

Figure 3.53. Interpolated map for observed aflatoxin values for field 2022B

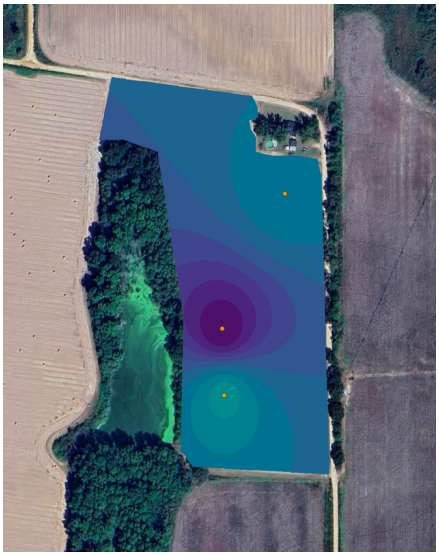

Figure 3.54. Interpolated map for predicted aflatoxin values for field 2022B

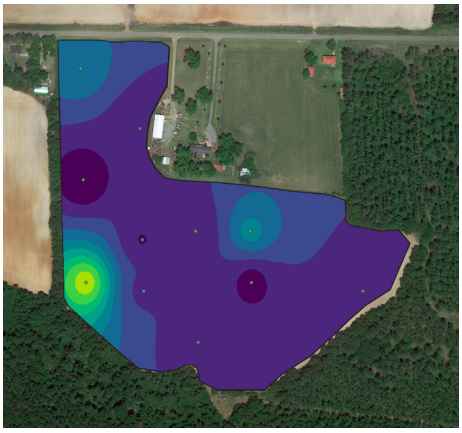

Figure 3.55. Interpolated map for observed aflatoxin values for field 2022C

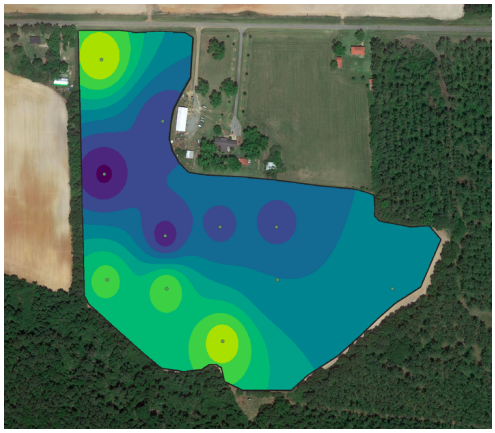

Figure 3.56. Interpolated map for predicted aflatoxin values for field 2022C

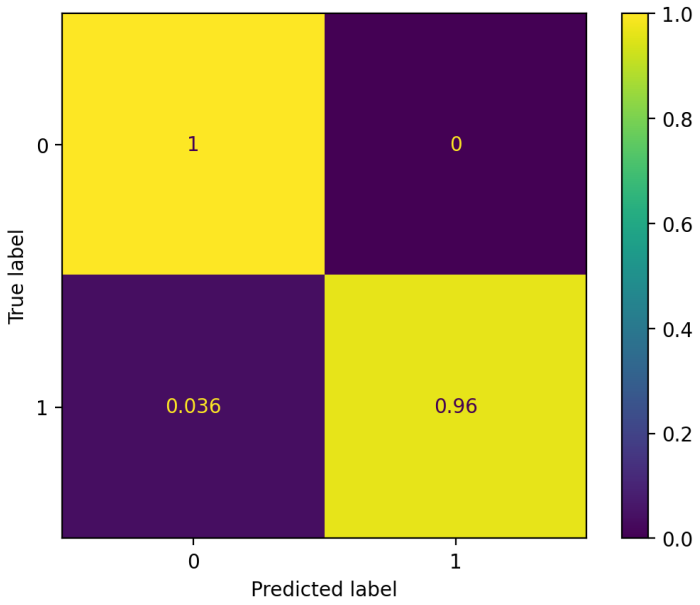

Figure 3.57. Confusion Matrix developed to visualize the performance of the classification model trained on 2023 data and tested on 2022 data.

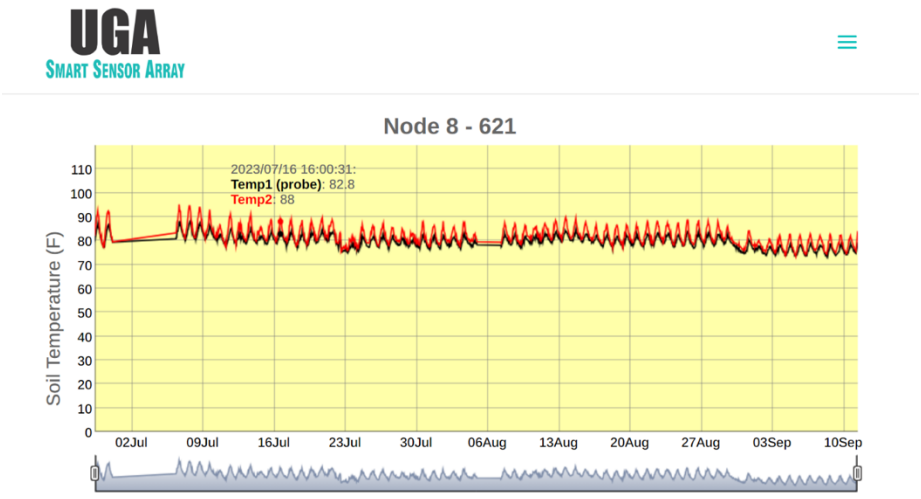

Figure 3.58 Trend graph showing gradual decrease in soil and air temperature in Field 2023A.

CHAPTER 4

FINAL CONCLUSIONS AND FUTURE ASPECTS

This study's goal was to develop a mathematical model that can be used to estimate forecast the presence of aflatoxin in rainfed peanut fields using in-field parameters. Our model used 8 detected predictors and was able to explain 27% variation. Starting with 22 variables, the feature engineering procedures revealed that only 8 predictors were sufficiently important to achieve the explained variance through the data collected. Although, the performance of the model was moderate at best, it is important to realize the possibly sizeable uncertainties associated with the target variable, collection and synthesis of predictor variables, and logistical realities. First, the growing seasons that were monitored for this study were not conducive for severe aflatoxin risk. This is evident from the fact that aflatoxin concentrations measured in this study ranged from 0 to 3.5 ppb. This limitation arose because this study was done on farmers' fields without any inoculation in the topsoil. In the light of this fact, it must be understood that the predictor variables were responsible for only explaining variance between this narrow range. It is not unreasonable to expect that in conditions where aflatoxin risk is high, the performance maybe better as there would be large gradients of both predictors and predicted variables for the model to train on. This limitation can be hurdled over by creating more study trials and inoculating the trial fields with *Aspergillus flavus* and observing/validating the relationships from earlier studies. This study must be carried forward for longer to observe the variation in

aflatoxin production throughout the seasons and have a larger dataset. Obtaining a larger dataset throughout the years not only helps improve the performance of the random forest model developed, but it also allows other, more complex machine learning and deep learning models to be implemented that require a bigger dataset for significant results. With these future studies, it would be worthwhile to consider increasing the spatial and temporal resolution of data (predicted and predictor variables) collection.

The results from this study help us pave a pathway for a better understanding of the contamination behavior of aflatoxin. In this study, we developed a Random Forest regression model to elucidate the complex relationships between aflatoxin concentration and a suite of field parameters, including soil-related factors, meteorological conditions, and spectral reflectance indices. By leveraging this robust machine learning method, we quantified the contributory degree of each parameter, offering a subtle understanding of the conditions that predispose peanut crops to aflatoxin contamination, thereby compromising the integrity of the harvest. It incorporates readily observable field parameters to forecast aflatoxin concentration levels with a demonstrated high degree of accuracy, as indicated by validation metrics such as R-squared and root mean square error values. For agronomists and farmers, the practical deployment of this model could translate to more strategic cultivation practices, specifically tailored to mitigate the risk of aflatoxin contamination. Furthermore, the model's utility extends to research contexts, wherein scientists may employ its predictive capabilities to refine experimental designs. By focusing on variables with significant impact on aflatoxin levels, researchers can optimize data collection strategies, thus enhancing the efficiency of field studies.

Our use of feature engineering procedures allowed for drastically reducing the number of predictors to be measured to predict aflatoxin risk by both researchers and farmers, which is a significant contribution of this study. By eliminating the need for measuring redundant parameters that add none-to-little information to the model, limited resources can be reallocated to improve statistical power (sampled fields, resolution aspects) Consequently, this can result in the acquisition of large-scale, high-fidelity datasets at a reduced financial and labor investment, contributing to the advancement of precision agriculture and plant pathology disciplines. This investigation advances a foundational model elucidating spatial and temporal variability in aflatoxin synthesis. Despite certain constraints inherent to modeling complex biological systems, this framework significantly enhances our comprehension of the intricate dynamics between aflatoxin production and agronomic factors. The model's design prioritizes the acquisition of data with superior spatial and temporal granularity, particularly for those variables demonstrating a heightened influence on predictive accuracy within aflatoxin estimation paradigms.

References

Abbas, H. K., et al. (2017). Aflatoxin contamination in sorghum: Factors and management strategies. Toxins, 9(3), 805.

Abbas, H.K., Zablotowicz, R.M., and Locke, M.A. 2004. Spatial variability of Aspergillus flavus soil populations under different crops and corn grain colonization and aflatoxins. Can. J. Bot. 82:1768-1775.

Abioye EA, Hensel O, Esau TJ, Elijah O, Abidin MSZ, Ayobami AS, Yerima O, Nasirahmadi A. Precision Irrigation Management Using Machine Learning and Digital Farming Solutions. AgriEngineering. 2022; 4(1):70-103. https://doi.org/10.3390/agriengineering4010006

Amaike, S., N.P. Keller. 2011. Aspergillus flavus. Annual Review of Phytopathology. 49:107-133.

Anfossi, L., Baggiani, C., Giovannoli, C., & Giraudi, G. (2011). Occurrence of aflatoxin M1 in dairy products. Aflatoxins-detection, measurement and control, 1-20.

Atehnkeng, J., et al. (2008). Biological control of aflatoxins in Africa: Current status and potential challenges in the face of climate change. World Mycotoxin Journal, 1(3), 317-327.

Bennett, J. W., & Klich, M. (2003). Mycotoxins. Clinical Microbiology Reviews, 16(3), 497-516.

Biau, Gérard, and Erwan Scornet. 2016. "A Random Forest Guided Tour." TEST 25 (2): 197–227. https://doi.org/10.1007/s11749-016-0481-7.

Boote, Kenneth, ed. 2019. Advances in Crop Modelling for a Sustainable Agriculture. 0 ed. Burleigh Dodds Science Publishing. https://doi.org/10.1201/9780429266591.

Breiman, Leo. 2001. "[No Title Found]." Machine Learning, Random Forests, 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Burton, Amanda B., and Armen R. Kemanian. 2022. "Assessing a Century of Maize and Soybean Polyculture for Silage Production." Agronomy Journal 114 (3): 1615–26. https://doi.org/10.1002/agj2.21006.

Cary, J. W., et al. (2000). Characterization of aflJ, a gene required for conversion of pathway intermediates to aflatoxin. Applied and Environmental Microbiology, 66(1), 120-126.

Chang, P. K., et al. (2017). Understanding aflatoxin contamination and reducing aflatoxin content in peanuts through metabolomics-based studies. Toxins, 9(5), 141.

Chauhan, Y. S., G. C. Wright, R. C. N. Rachaputi, D. Holzworth, A. Broome, S. Krosch, and M. J. Robertson. 2010. "Application of a Model to Assess Aflatoxin Risk in

Peanuts." The Journal of Agricultural Science 148 (3): 341–51. https://doi.org/10.1017/S002185961000002X.

Clevenger, J., Marasigan, K., Liakos, V., Sobolev, V., Vellidis, G., Holbrook, C., and Ozias-Akins, P. 2016. RNA sequencing of contaminated seeds reveals the state of the seed permissive for pre-harvest aflatoxin contamination and points to a potential susceptibility factor. Toxins 8(11):317.

Cotty, P. J., & Jaime-Garcia, R. (2007). Influences of climate on aflatoxin-producing fungi and aflatoxin contamination. International Journal of Food Microbiology, 119(1-2), 109-115.

Diener, U L, R J Cole, T H Sanders, G A Payne, L S Lee, and M A Klich. 1987. "Epidemiology of Aflatoxin Formation by Aspergillus Flavus*." Annual Review of Phytopathology 25 (1): 249–70. https://doi.org/10.1146/annurev.py.25.090187.001341.

Dorner, J. W. (2008). Management and prevention of mycotoxins in peanuts. Food Additives and Contaminants, 25(2), 203-208.

Dorner, J. W. (2008). Management and prevention of mycotoxins in peanuts. Food Additives & Contaminants: Part A, 25(2), 203-208.

Dorner, J. W. (2008). Management and prevention of mycotoxins in peanuts. Food Additives & Contaminants: Part A, 25(2), 203-208.

Dorner, J. W. (2009). Biological control of aflatoxin contamination in corn using a nontoxigenic strain of Aspergillus flavus. Journal of Food Protection, 72(4), 801-804.

Dorner, J. W. (2009). Biological control of aflatoxin contamination in corn using a nontoxigenic strain of Aspergillus flavus. Journal of Food Protection, 72(4), 801-804.

Fountain, J. C., et al. (2015). Peanut (Arachis hypogaea L.) response to water deficit stress: Proteomic analysis of the peanut leaf proteome. Scientific Reports, 5, 11036.

Griffin, G.J., and Garren, G.H. 1974. Population levels of Aspergillus flavus and the A. niger group in Virginia peanut field soils. Phytopathology 64:322-325.

Hafez E, Abd El-Aziz NM, Darwish AMG, Shehata MG, Ibrahim AA, Elframawy AM, Badr AN. Validation of New ELISA Technique for Detection of Aflatoxin B1 Contamination in Food Products versus HPLC and VICAM. Toxins. 2021; 13(11):747. https://doi.org/10.3390/toxins13110747

Han J, Zhang Z, Cao J, Luo Y, Zhang L, Li Z, Zhang J. Prediction of Winter Wheat Yield Based on Multi-Source Data and Machine Learning in China. Remote Sensing. 2020; 12(2):236. https://doi.org/10.3390/rs12020236

Hill, R A, P D Blankenship, R J Cole, and T H Sanders. 1983. "Effects of Soil Moisture and Temperature on Preharvest Invasion of Peanuts by the Aspergillus Flavus Group and Subsequent Aflatoxin Development." Applied and Environmental Microbiology 45 (2): 628–33. https://doi.org/10.1128/aem.45.2.628-633.1983.

Hill, R. A., Blankenship, P. D., Cole, R. J., & Sanders, T. H. (1983). Effects of soil moisture and temperature on preharvest invasion of peanuts by the Aspergillus flavus group and subsequent aflatoxin development. Applied and environmental microbiology, 45(2), 628-633.

Hoffman, Alexis L., Armen R. Kemanian, and Chris E. Forest. 2018. "Analysis of Climate Signals in the Crop Yield Record of Sub-Saharan Africa." Global Change Biology 24 (1): 143–57. https://doi.org/10.1111/gcb.13901.

Hoogenboom, G., J. W. Jones, P. W. Wilkens, C. H. Porter, W. D. Batchelor, L. A. Hunt, K. J. Boote, U. Singh, O. Uryasev, and W. T. Bowen. 2004. "Decision Support System for Agrotechnology Transfer Version 4.0." University of Hawaii, Honolulu, HI (CD-ROM).

International Agency for Research on Cancer (IARC). (2002). Some traditional herbal medicines, some mycotoxins, naphthalene and styrene. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, 82, 301-366.

Isleib, T. G., et al. (2013). Registration of 'NemaTAM' peanut. Journal of Plant Registrations, 7(2), 176-179.

Kahane, Leo. 2008. Regression Basics. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc. https://doi.org/10.4135/9781483385662.

Klich, M. A. (2007). Aspergillus flavus: the major producer of aflatoxin. Molecular Plant Pathology, 8(6), 713-722.

Klompenburg T, Kassahun A, Catal A, Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, Volume 177, 2020, 105709, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2020.105709.

Kolosova, A.Y., Shim, WB., Yang, ZY. et al. Direct competitive ELISA based on a monoclonal antibody for detection of aflatoxin B1. Stabilization of ELISA kit components and application to grain samples. Anal Bioanal Chem 384, 286–294 (2006). https://doi.org/10.1007/s00216-005-0103-9

Kumar, P., Mahato, D. K., Kamle, M., Mohanta, T. K., & Kang, S. G. (2017). Aflatoxins: A global concern for food safety, human health and their management. Frontiers in microbiology, 7, 2170.

Lamb, M. C., & Sternitzke, D. A. (2001). Cost of aflatoxin to the farmer, buying point, and sheller segments of the southeast United States peanut industry. Peanut Science, 28(2), 59-63.

Li, X. et al. (2020). "Optimal irrigation scheduling using machine learning regression models." Agricultural Water Management, 238, 106210.

Mehl, H. L., Jaime, R., Callicott, K. A., Probst, C., Garber, N. P., Ortega-Beltran, A., ... & Cotty, P. J. (2012). Aspergillus flavus diversity on crops and in the environment can

be exploited to reduce aflatoxin exposure and improve health. Annals of the New York Academy of Sciences, 1273(1), 7-17.

Probst, C., et al. (2014). Environmental factors contribute to the risk of Aflatoxin B1 contamination in Chilean maize fields. Food Additives & Contaminants: Part A, 31(10), 1660-1670.

Rasooli, I., et al. (2010). Control of Aspergillus flavus growth and aflatoxin production in maize grains using active packaging with volatile thyme oil. Journal of Food Safety, 30(3), 503-515.

Robens, J., & Cardwell, K. (2003). The costs of mycotoxin management to the USA: management of aflatoxins in the United States. Journal of Toxicology: Toxin Reviews, 22(2-3), 139-152.

Robens, Jane, and Kitty Cardwell. 2003. "The Costs of Mycotoxin Management to the USA: Management of Aflatoxins in the United States." Journal of Toxicology: Toxin Reviews 22 (2–3): 139–52. https://doi.org/10.1081/TXR-120024089.

Rokach, Lior, and Oded Maimon. 2005. "Decision Trees." In Data Mining and Knowledge Discovery Handbook, edited by Oded Maimon and Lior Rokach, 165–92. Boston, MA: Springer US. https://doi.org/10.1007/0-387-25465-X_9.

Sanders, T.H., Cole, R.J., Blankenship, P.D., and Hill, R.A. 1985. Relation of environmental stress duration to Aspergillus flavus invasion and aflatoxin production in preharvest peanuts. Peanut Sci. 12:90-93.

Schmale, David G., and Gary P. Munkvold. "Mycotoxins in crops: A threat to human and domestic animal health." The plant health instructor 3.3 (2009): 340-353.

Shrestha, S., et al. (2017). Pre-harvest aflatoxin contamination in drought tolerant and drought susceptible maize hybrids. Toxins, 9(7), 239.

Sobolev, V. S., & Dorner, J. W. (2009). Cleanup procedure for determination of aflatoxins in major agricultural commodities by liquid chromatography with fluorescence detection. Journal of AOAC International, 92(3), 659-666.

Sobolev, V. S., & Dorner, J. W. (2009). Cleanup procedure for determination of aflatoxins in major agricultural commodities by liquid chromatography with fluorescence detection. Journal of AOAC International, 92(3), 659-666.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. "Conditional Variable Importance for Random Forests." BMC Bioinformatics 9 (1): 307. https://doi.org/10.1186/1471-2105-9-307.

Tillman, P. G., et al. (2016). Groundnut seed and pod damage by Blissus insularis (Hemiptera: Blissidae) and associated kernel infection by Aspergillus parasiticus and Aflatoxin Accumulation. Journal of Economic Entomology, 109(2), 793-800.

Torres, A. M., Barros, G. G., Palacios, S. A., Chulze, S. N., & Battilani, P. (2014). Review on pre-and post-harvest management of peanuts to minimize aflatoxin contamination. Food Research International, 62, 11-19.

Vellidis, G., Tucker, M., Perry, C., Reckford, D, Butts, C., Henry, H., et al. 2013. A soil moisture sensor-based variable rate irrigation scheduling system. In: J.V. Stafford (Ed.), Precision Agriculture 2013. Wageningen Academic Publishers, Wageningen.

Vellidis, George, M. Tucker, C. Perry, C. Kevin, and C. Bednarz. 2008. "A Real-Time Wireless Smart Sensor Array for Scheduling Irrigation. Computers and Electronics in Agriculture 61 (1): 44–50. https://doi.org/10.1016/j.compag.2007.05.009.

Vellidis, G., B. Ortiz, C. Perry. 2006. Unpublished data.

APPENDIX

Code:

- **R:**
    - **This code refers to the libraries used for all functions used in the analysis code.**

```
#reprtree
options(repos='http://cran.rstudio.org')
have.packages <- installed.packages()
cran.packages <- c('devtools','plotrix','randomForest','tree')
to.install <- setdiff(cran.packages, have.packages[,1])
if(length(to.install)>0) install.packages(to.install)
library(devtools)
if(!('reprtree' %in% installed.packages())){
  install_github('munoztd0/reprtree')
}
for(p in c(cran.packages, 'reprtree')) eval(substitute(library(pkg), list(pkg=p)))
#reading libraries
library(readxl)
library(ggplot2)
library(gplots)
library(pheatmap)
library(tidyverse)
library(tidyr)
library(dplyr)
library(Metrics)
library(glmnet)
```

```
library(randomForest)
library(pdp)
library(ggthemes)
library(reshape2)
library(party)
library(reprtree)
```

- **This is the code to import and clean the dataset and does initial analysis and visualization of the dataset.**

```
#reading and processing raw data
data <- read.csv("Aflatoxin_23.csv")
data <- data[,-c(21,22,23)]
data <- subset(data, complete.cases(data))
max(data$Aflatoxin)
min(data$Aflatoxin)
write.csv(data, "Aflatoxin_final.csv")
data <- data %>%
  mutate(afla_cut = ifelse(Aflatoxin >= 1, 1, 0))
data <- na.omit(data)
mean(data$Aflatoxin)
median(data$Aflatoxin)
sd(data$Aflatoxin)

#density plot
p <- ggplot(data, aes(x = Aflatoxin))+
  geom_density(color="darkblue", fill = "maroon", adjust = 1.5, alpha=0.4)+
  geom_vline(aes(xintercept=mean(Aflatoxin)),
             colour = "darkgreen", linetype="dashed", size=0.8)+
```

```
  theme_minimal()+
  ggtitle("Density plot for Aflatoxin")+
  ylab("Density")+
  xlab("Aflatoxin (ppb)")
p

ggsave("Result_figures/Density.jpeg", width = 10, height = 5)



#heatmap matrix
data_heatmap <- as.matrix(data[,c("CLAY_1",
                  "CLAY_2",
                  "CLAY_3",
                  "CLAY_4",
                  "CLAY_5",
                  "Solar_Radiation",
                  "VPD",
                  "SWT",
                  "Temperature",
                  "Soil_Temp",
                  "NDVI",
                  "Aflatoxin")])
data_heatmap <- na.omit(data_heatmap)
correlation_matrix <- cor(data_heatmap, method = "spearman")
melted_correlation <- melt(correlation_matrix)
custom_colors <- c("maroon", "white", "darkgreen")
```

```
#Plotting heatmap
q <- ggplot(melted_correlation, aes(Var1, Var2, fill = value, label = round(value, 2))) +
  geom_tile() +
  scale_fill_gradient2(low = custom_colors[1], mid = custom_colors[2], high = custom_colors[3]) +
  geom_text(color = "black") +
  labs(title = "Correlation Heatmap", x = "", y = "", fill = "Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
q
ggsave("Result_figures/heatmap.jpeg", width = 10, height = 5)


#example correlation between clay_2 and NDVI
lm_model <- lm(SWT~ Aflatoxin, data = data)
intercept <- coef(lm_model)[1]
slope <- coef(lm_model)[2]
r_squared <- summary(lm_model)$r.squared
r_squared
slope
intercept
#plotting the regression between clay_2 and NDVI
r <- ggplot(data, aes(SWT, Aflatoxin)) +
  geom_point(color = "darkgreen") +  # Add the scatterplot points
  geom_smooth(method = "lm", se = FALSE, color = "maroon") +
  #geom_text(family = "serif", size = 6, aes(x = 50, y = 3.7, label = paste("Aflatoxin (ppb) =", round(intercept, 2), "+", round(slope, 2), "(SWT(kPa))"))) +
  #geom_text(family = "serif", size = 6, aes(x = 22.4, y = 3.5, label = paste("R-squared =", round(r_squared, 2))))+
```

```
  labs(title = "Scatterplot between Clay (%) at 90% depth and Aflatoxin (ppb)",
    x = "Clay (%) at 90cm depth",
    y = "Aflatoxin (ppb)") +
  theme_minimal()
r
ggsave("Result_figures/scatter_clay_afla.jpeg", width = 10, height = 5)


#filtering correlations
melted_correlation_1 <- filter(melted_correlation, Var1 != Var2 & value > 0.5)


#datasplit for modelling
set.seed(123)
training_fraction <- 0.75
training_size <- floor(nrow(data) * training_fraction)
training_indices <- sample(1:nrow(data), size = training_size)
training_data <- data[training_indices, ]
testing_data <- data[-training_indices, ]
```

- **This part of the code develops the linear model explained in Section 2.2.2.1**

```
#linear model
lm_model1 <- lm(Aflatoxin ~ CLAY_1 +
                  CLAY_2 +
                  CLAY_3 +
                  CLAY_4 +
                  CLAY_5 +
                  Solar_Radiation +
                  Precipitation+
```

```
            Temperature +
            VPD +
            SWT +
            Soil_Temp +
            NDVI+
            CLAY_2:CLAY_3+
            CLAY_2:CLAY_4+
            CLAY_2:CLAY_5+
            CLAY_3:CLAY_4+
            CLAY_3:CLAY_5+
            CLAY_4:CLAY_5+
            NDVI:CLAY_4+
            NDVI:CLAY_5+
            Temperature:Solar_Radiation+
            Temperature:SWT
            ,training_data)
summary(lm_model1)

testing_data$predictions <- predict(lm_model1, testing_data)
gof(testing_data$predictions, testing_data$Aflatoxin)

mse <- mse(testing_data$predictions, testing_data$Aflatoxin)
print(mse)
plot(testing_data$Aflatoxin, testing_data$predictions)
fit = lm(Aflatoxin~predictions, data = testing_data) #Create the linear regression
for validation
summary(fit) #Review the results
abline(fit)
```

```
intercept_1 <- coef(fit)[1]

slope_1 <- coef(fit)[2]

r_squared_1 <- summary(fit)$r.squared

intercept_1

slope_1

r_squared_1



#plotting the observed with estimated aflatoxin values

s <- ggplot(testing_data, aes(x = Aflatoxin, y = predictions))+

  geom_point(color = "darkgreen")+

  geom_smooth(method = "lm", se = FALSE, color = "maroon")+

  geom_text(family = "serif", size = 4, aes(x = 2.2, y = 1.65, label =
paste("Predicted =", round(intercept_1, 2), "+", round(slope_1, 2),
"(Observed)"))) +

  geom_text(family = "serif", size = 4, aes(x = 2.17, y = 1.55, label = paste("R-
squared =", round(r_squared_1, 2))))+

  labs(title = "Scatterplot between Observed and Predicted Aflatoxin (Linear
Model)",

       x = "Observed Aflatoxin (ppb)",

       y = "Predicted Aflatoxin (ppb)") +

  theme_minimal()

s

ggsave("Result_figures/scatter_lm_model.jpeg", width = 10, height = 5)
```

- **This part of the code was written to develop the 25250 random forest model iterations as mentioned in Section 2.2.2.2**

```
#random forest

set.seed(123)

training_fraction <- 0.75
```

```
training_size <- floor(nrow(data) * training_fraction)
training_indices <- sample(1:nrow(data), size = training_size)
training_data <- data[training_indices, ]
testing_data <- data[-training_indices, ]

ntree_values <- seq(from = 1, to= 501, by =5)
mtry_values <- seq(from = 1, to = 5, by = 1)
maxnodes_values <- seq(from = 1, by = 1, to = 50)

#looping for best parameters
set.seed(123)
results_df <- data.frame(ntree = integer(0),
                         mtry = integer(0),
                         maxnodes = integer(0),
                         mse = numeric(0),
                         r_squared = numeric(0))



for(ntree in ntree_values){
  for(mtry in mtry_values){
    for(maxnodes in maxnodes_values){
      rf_model <- randomForest(Aflatoxin ~ CLAY_1 +
                               CLAY_2 +
                               CLAY_3 +
                               CLAY_4 +
                               CLAY_5 +
                               Solar_Radiation +
```

```
                    Precipitation+
                    Temperature +
                    VPD +
                    SWT +
                    Soil_Temp +
                    NDVI,
                    training_data,
                    ntree = ntree,
                    mtry = mtry,
                    maxnodes = maxnodes)
      predictions <- predict(rf_model, newdata = testing_data)
      ml = lm(predictions~testing_data$Aflatoxin)
      mse <- mean((predictions - testing_data$Aflatoxin)^2)
      r_squared <- summary(ml)$r.squared

      results_df <- rbind(results_df,
                  data.frame(ntree = ntree,
                        mtry = mtry,
                        maxnodes = maxnodes,
                        mse = mse,
                        r_squared = r_squared))
    }
  }
}

best_model <- results_df[which.max(results_df$mse + results_df$r_squared), ]
best_model #parameters with the best results
plot(results_df$maxnodes, results_df$r_squared)
```

```
#best model running one time
set.seed(123)
best_rf <- randomForest(Aflatoxin ~ CLAY_1 +
                        CLAY_2 +
                        CLAY_3 +
                        CLAY_4 +
                        CLAY_5 +
                        Solar_Radiation +
                        Precipitation+
                        Temperature +
                        VPD +
                        SWT +
                        Soil_Temp +
                        NDVI,
                       training_data,
                       ntree = 266,
                       mtry = 5,
                       maxnodes = 1)
predictions_b <- predict(best_rf, newdata = testing_data)
gof(predictions_b, testing_data$Aflatoxin)


#emsembled results
num_models <- 100
seed_values <- 1:num_models
predictions <- matrix(0, nrow = nrow(testing_data), ncol = num_models)
```

```
for (i in 1:num_models) {
  set.seed(seed_values[i])
  test_rf_model <- randomForest(Aflatoxin ~ CLAY_1 +
                                CLAY_2 +
                                CLAY_3 +
                                CLAY_4 +
                                CLAY_5 +
                                Solar_Radiation +
                                Precipitation+
                                Temperature +
                                VPD +
                                SWT +
                                Soil_Temp +
                                NDVI,
                               training_data,
                               ntree = 266,
                               mtry = 5,
                               maxnodes = 1)
  predictions[, i] <- predict(test_rf_model, newdata = testing_data)
}

averaged_predictions <- rowMeans(predictions)
gof(averaged_predictions, testing_data$Aflatoxin)
plot(averaged_predictions, testing_data$Aflatoxin)
test <- data.frame(pred = averaged_predictions, afla = testing_data$Aflatoxin)
```

```
df <- data.frame(pred = averaged_predictions, obs =testing_data$Aflatoxin)
write.csv(df, "test.csv")

ggplot(data = test, aes(x =afla, y =pred))+
  geom_point()+
  geom_smooth(method = lm)+
  xlim(0,3)+
  ylim(0,3)+
  abline()

model_t <- lm(afla~pred, data = test)
plot(test$afla, test$pred)
abline(model_t)



evaluate <- lm(pred ~ afla, data = test)
summary(evaluate)
plot(evaluate)
gof1 <- gof(averaged_predictions, testing_data$Aflatoxin, norm = "maxmin")
gof1

mse_e <- mean((averaged_predictions - testing_data$Aflatoxin)^2)
r_squared_e <- 1 - (var(averaged_predictions) / var(testing_data$Aflatoxin))
mse_e
r_squared_e
```

```
plot(test_rf_model)
reprtree:::plot.getTree(test_rf_model)
```

- **This code was written to visualize PDPs mentioned in Section 2.2.2.2**

```
#Importance plots
importances <- importance(test_rf_model)
sorted_importances <- sort(importances, decreasing=TRUE)
print(sorted_importances)
varImpPlot(test_rf_model)



importance_df <- data.frame(
  Variable = rownames(importances),
  Importance = importances[,"IncNodePurity"]  # Adjust this based on your needs
)


t <- ggplot(importance_df, aes(x = reorder(Variable, Importance), y =
Importance)) +
  geom_bar(stat = "identity", fill = "maroon")+
  labs(x = "Variable", y = "Importance") +
  coord_flip()+
  labs(title = "Variable Importance Plot for Random Forest Regression Method",
       y = "Importance (IncNodePurity)",
       x = "Variable")+
  geom_text(aes(label = round(Importance, 2)), hjust = -0.2, size = 4)+
  theme_minimal()
t
ggsave("Result_figures/varimp.jpeg", width = 10, height = 5)
```

```
#partial dependence plots for each variable

partial_clay5<- as.data.frame(partialPlot(test_rf_model, x.var = CLAY_5,
pred.data = testing_data, plot = F))

u_clay5 <- ggplot(partial_clay5, aes(x = x, y = y)) +

  geom_line(size = 2, color = "maroon") +

  labs(x = "Clay (Depth 5)", y = "Partial Dependence")+

  labs(title = "Partial Dependence of Random Forest Model on Clay (Depth 5)")+

  theme_minimal()

u_clay5

ggsave("Result_figures/partial_clay5.jpeg", width = 10, height = 5)


partial_NDVI<- as.data.frame(partialPlot(test_rf_model, x.var = NDVI, pred.data
= testing_data, plot = F))

u_ndvi <- ggplot(partial_NDVI, aes(x = x, y = y)) +

  geom_line(size = 2, color = "maroon") +

  labs(x = "NDVI", y = "Partial Dependence")+

  labs(title = "Partial Dependence of Random Forest Model on NDVI")+

  theme_minimal()

u_ndvi

ggsave("Result_figures/partial_ndvi.jpeg", width = 10, height = 5)


partial_SWT<- as.data.frame(partialPlot(test_rf_model, x.var = SWT, pred.data =
testing_data, plot = F))

u_SWT <- ggplot(partial_SWT, aes(x = x, y = y)) +

  geom_line(size = 2, color = "maroon") +

  labs(x = "SWT", y = "Partial Dependence")+

  labs(title = "Partial Dependence of Random Forest Model on SWT")+

  theme_minimal()
```

```
u_SWT
ggsave("Result_figures/partial_SWT.jpeg", width = 10, height = 5)


partial_VPD<- as.data.frame(partialPlot(test_rf_model, x.var = VPD, pred.data =
testing_data, plot = F))
u_VPD <- ggplot(partial_VPD, aes(x = x, y = y)) +
  geom_line(size = 2, color = "maroon") +
  labs(x = "VPD", y = "Partial Dependence")+
  labs(title = "Partial Dependence of Random Forest Model on VPD")+
  theme_minimal()
u_VPD
ggsave("Result_figures/partial_VPD.jpeg", width = 10, height = 5)


partial_clay2<- as.data.frame(partialPlot(test_rf_model, x.var = CLAY_2,
pred.data = testing_data, plot = F))
u_clay2 <- ggplot(partial_clay2, aes(x = x, y = y)) +
  geom_line(size = 2, color = "maroon") +
  labs(x = "Clay (Depth 2)", y = "Partial Dependence")+
  labs(title = "Partial Dependence of Random Forest Model on clay2")+
  theme_minimal()
u_clay2
ggsave("Result_figures/partial_clay2.jpeg", width = 10, height = 5)


partial_temp<- as.data.frame(partialPlot(test_rf_model, x.var = Temperature,
pred.data = testing_data, plot = F))
u_temp <- ggplot(partial_temp, aes(x = x, y = y)) +
  geom_line(size = 2, color = "maroon") +
  labs(x = "Air Temperature", y = "Partial Dependence")+
```

```
  labs(title = "Partial Dependence of Random Forest Model on Air
Temperature")+

  theme_minimal()

u_temp

ggsave("Result_figures/partial_temp.jpeg", width = 10, height = 5)


partial_stemp<- as.data.frame(partialPlot(test_rf_model, x.var = Soil_Temp,
pred.data = testing_data, plot = F))

u_stemp <- ggplot(partial_stemp, aes(x = x, y = y)) +

  geom_line(size = 2, color = "maroon") +

  labs(x = "Soil Temperature", y = "Partial Dependence")+

  labs(title = "Partial Dependence of Random Forest Model on Soil
Temperature")+

  theme_minimal()

u_stemp

ggsave("Result_figures/partial_stemp.jpeg", width = 10, height = 5)


partial_ppt<- as.data.frame(partialPlot(test_rf_model, x.var = Precipitation,
pred.data = testing_data, plot = F))

u_ppt <- ggplot(partial_ppt, aes(x = x, y = y)) +

  geom_line(size = 2, color = "maroon") +

  labs(x = "Precipitation", y = "Partial Dependence")+

  labs(title = "Partial Dependence of Random Forest Model on Precipiation")+

  theme_minimal()

u_ppt

ggsave("Result_figures/partial_ppt.jpeg", width = 10, height = 5)


partial_solar<- as.data.frame(partialPlot(test_rf_model, x.var = Solar_Radiation,
pred.data = testing_data, plot = F))

u_solar <- ggplot(partial_solar, aes(x = x, y = y)) +
```

```
  geom_line(size = 2, color = "maroon") +
  labs(x = "Solar Radiation", y = "Partial Dependence")+
  labs(title = "Partial Dependence of Random Forest Model on Solar Radiation")+
  theme_minimal()
u_solar
ggsave("Result_figures/partial_solar.jpeg", width = 10, height = 5)
```

- **Python:**
    - **Recursive Feature Elimination :**

```
import pandas as pd
import matplotlib.pyplot as plt

from sklearn.datasets import fetch_california_housing
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from feature_engine.selection import RecursiveFeatureElimination
data = pd.read_csv('../Afla_final.csv')
data = data.loc[:, ~data.columns.str.contains('^Unnamed')]
X = data.drop('Aflatoxin', axis=1)
y = data['Aflatoxin']

# Train/test set generation
X_train, X_test, y_train, y_test = train_test_split(
   X, y, test_size=0.2, random_state=1121218
)

scaler = StandardScaler()
scaler.fit(X_train)

X_train_std = pd.DataFrame(scaler.transform(X_train), columns =
X_train.columns)
X_test_std = pd.DataFrame(scaler.transform(X_test), columns = X_test.columns)
```

```
# Scale train and test sets with StandardScaler
# X_train_std = StandardScaler().fit_transform(X_train)
# X_test_std = StandardScaler().fit_transform(X_test)

# Fix the dimensions of the target array
y_train = y_train.values.reshape(-1, 1)
y_test = y_test.values.reshape(-1, 1)
model = RandomForestRegressor(
    random_state=10,
)
RFE_model = RecursiveFeatureElimination(
    estimator = model, # the ML model
    scoring = 'r2',
    threshold = 0.001,
    cv=3,
)
RFE_model.fit(X_train_std, y_train)
RFE_model.initial_model_performance_
RFE_model.feature_importances_.plot.bar(figsize=(10, 5))
plt.ylabel('Feature Importance')
plt.title('Feature Importance (IncNodePurity)')
plt.show()
pd.Series(RFE_model.performance_drifts_).plot.bar(figsize=(10, 5))
plt.title('Performance change after removing features recursively')
plt.ylabel('$R^2$ change when feature was removed')
plt.show()
pd.Series(RFE_model.performance_drifts_).sort_values().plot.bar(figsize=(20,5))
plt.ylabel('Change in performance ($R^2$) when removing feature')
```

```
plt.show()
RFE_model.features_to_drop_
X_1 = RFE_model.transform(X_train_std)
RFE_model.performance_drifts_
X_1.head()
# drop variables
X_train1 = RFE_model.transform(X_train_std)
X_test1= RFE_model.transform(X_test_std)
RFE_model1 = RecursiveFeatureElimination(
   estimator = model, # the ML model
   scoring = 'r2',
   threshold = 0.001,
   cv=3,
)
RFE_model1.fit(X_train1, y_train)
RFE_model1.initial_model_performance_
pd.Series(RFE_model1.performance_drifts_).plot.bar(figsize=(10, 5))
plt.title('Performance change after removing features recursively')
plt.ylabel('$R^2$ change when feature was removed')
plt.show()
X_train2 = RFE_model1.transform(X_train1)
X_test2= RFE_model1.transform(X_test1)
RFE_model2 = RecursiveFeatureElimination(
   estimator = model, # the ML model
   scoring = 'r2',
   threshold = 0.001,
   cv=3,
)
```

```
RFE_model2.fit(X_train2, y_train)
RFE_model2.initial_model_performance_
pd.Series(RFE_model2.performance_drifts_).sort_values().plot.bar(figsize=(10,5)
)
plt.title('Performance change after removing features recursively')
plt.ylabel('Change in performance ($R^2$) when removing feature')
plt.show()
RFE_model2.performance_drifts_
RFE_model2.features_to_drop_
X_train3 = RFE_model2.transform(X_train2)
X_test3= RFE_model2.transform(X_test2)
RFE_model3 = RecursiveFeatureElimination(
   estimator = model, # the ML model
   scoring = 'r2',
   threshold = 0.001,
   cv=3,
)
RFE_model3.fit(X_train3, y_train)
RFE_model3.initial_model_performance_
RFE_model3.features_to_drop_
X_train4 = RFE_model3.transform(X_train3)
X_test4= RFE_model3.transform(X_test3)
RFE_model4 = RecursiveFeatureElimination(
   estimator = model, # the ML model
   scoring = 'r2',
   threshold = 0.001,
   cv=3,
)
```

```
RFE_model4.fit(X_train4, y_train)

RFE_model4.initial_model_performance_

RFE_model4.features_to_drop_

pd.Series(RFE_model4.performance_drifts_).sort_values().plot.bar(figsize=(10,5)
)

plt.title('Performance change after removing features recursively')

plt.ylabel('Change in performance ($R^2$) when removing feature')

plt.savefig('Plt_imp.tif',dpi = 300,bbox_inches='tight')

plt.show()

RFE_model4.performance_drifts_
```

- **Scatter Plot :**

```
import numpy as np

import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import accuracy_score

from sklearn.metrics import ConfusionMatrixDisplay

from sklearn.tree import DecisionTreeClassifier
```

```
def presence(Aflatoxin):
    if Aflatoxin > 0:
        return 1
    else:
        return 0



data = pd.read_csv('Afla_final_cv3.csv')
data = data.loc[:, ~data.columns.str.contains('^Unnamed')]
X = data.drop('Aflatoxin', axis=1)
y = data['Aflatoxin']

# Train/test set generation
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=1121238
)
data_2022 = pd.read_csv("2022_final.csv")
y_test = data_2022["Aflatoxin"]
X_test = data_2022.drop("Aflatoxin", axis = 1)

model = RandomForestRegressor(random_state=10).fit(X_train, y_train)

predictions = model.predict(X_test)
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import gaussian_kde
from sklearn.metrics import r2_score, mean_squared_error
```

```
xy = np.vstack([y_test, predictions])
z = gaussian_kde(xy)(xy)

idx = z.argsort()

fig,ax = plt.subplots()
ax.scatter(y_test, predictions,c=z,s=10,edgecolor='none')



coefficients = np.polyfit(y_test, predictions, 1)
m, b = coefficients

# Add a line of best fit to the plot
x_range = np.linspace(min(y_test), max(y_test), 100)
y_fit = m * x_range + b
line = ax.plot(x_range, y_fit, color='red', label='Line of Best Fit')

# Add a 1:1 line to the plot
min_val = min(min(y_test), min(predictions))
max_val = max(max(y_test), max(predictions))
line_1to1 = ax.plot([min_val, max_val], [min_val, max_val], color='blue',
linestyle='dashed', label='1:1 Line')

# Calculate R-squared
r_squared = r2_score(y_test, predictions)

# Calculate Mean Squared Error (MSE)
```

```
mse = mean_squared_error(y_test, predictions)

# Calculate R Mean Squared Error (RMSE)
rmse = mean_squared_error(y_test, predictions, squared=False)



# Add R-squared value and MSE as text on the plot with bounding box
r_squared_text = f'R-squared: {r_squared:.2f}'
mse_text = f'MSE: {mse:.4f}'
rmse_text = f'RMSE: {rmse:.4f}'
ax.text(0.80, 0.05, r_squared_text, transform=ax.transAxes, fontsize=8, va='top',
bbox=dict(facecolor='white', edgecolor='black', boxstyle='round,pad=0.5'))
ax.text(0.835, 0.12, mse_text, transform=ax.transAxes, fontsize=8, va='top',
bbox=dict(facecolor='white', edgecolor='black', boxstyle='round,pad=0.5'))
ax.text(0.82, 0.19, rmse_text, transform=ax.transAxes, fontsize=8, va='top',
bbox=dict(facecolor='white', edgecolor='black', boxstyle='round,pad=0.5'))

# Create the equation string
equation = f'Line of Best Fit: y = {m:.2f}x + {b:.2f}'

# Add equation as text on the plot with bounding box
ax.text(0.555, 0.27, equation, transform=ax.transAxes, fontsize=9, va='top',
bbox=dict(facecolor='white', edgecolor='black', boxstyle='round,pad=0.5'))

# Adjust the position of the legend
ax.legend(loc='upper left', fontsize=9)
```

```
# Add labels and title to the plot
ax.set_xlabel('Observed Aflatoxin concentration (ppb)')
ax.set_ylabel('Predicted Aflatoxin concentration (ppb)')
ax.set_title('Observed vs Predicted Aflatoxin concentration - 2022')
plt.savefig("2022_results.tif", dpi = 200, bbox_inches = 'tight')
# Show the plot
plt.show()
```

- **Partial Dependency Plots :**

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.tree import DecisionTreeClassifier


data = pd.read_csv('../Afla_final.csv')
data = data.loc[:, ~data.columns.str.contains('^Unnamed')]
X = data.drop('Aflatoxin', axis=1)
y = data['Aflatoxin']

# Train/test set generation
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=1121218
)
```

```
model = RandomForestRegressor(random_state=10).fit(X_train, y_train)

from matplotlib import pyplot as plt

from sklearn.inspection import PartialDependenceDisplay

plt.rcParams["figure.figsize"] = (3.5,3.5)

feat_name = 'Temperature'

display=PartialDependenceDisplay.from_estimator(model, X, [feat_name],
grid_resolution=200, percentiles=(0.05, 0.95))

plt.tight_layout()

plt.savefig("pdp_Temp.png", dpi=300)

plt.show()

#plt.style.use('default')

from matplotlib import pyplot as plt

from sklearn.inspection import PartialDependenceDisplay

plt.rcParams["figure.figsize"] = (3.5,3.5)

feat_name = 'Solar_Radiation'

display=PartialDependenceDisplay.from_estimator(model, X, [feat_name],
grid_resolution=200, percentiles=(0.05, 0.95))

plt.xlabel("Solar Radiation")

plt.tight_layout()

plt.savefig("pdp_SolRad.png", dpi=300)

plt.show()

#plt.style.use('default')

from matplotlib import pyplot as plt

from sklearn.inspection import PartialDependenceDisplay

plt.rcParams["figure.figsize"] = (3.5,3.5)

feat_name = 'Soil_Temp'
```

```
display=PartialDependenceDisplay.from_estimator(model, X, [feat_name],
grid_resolution=200, percentiles=(0.05, 0.95))

plt.xlabel("Soil Temperature ($^0C$)")

plt.tight_layout()

plt.savefig("pdp_SoilTemp.png", dpi=300)

plt.show()

#plt.style.use('default')

from matplotlib import pyplot as plt

from sklearn.inspection import PartialDependenceDisplay

plt.rcParams["figure.figsize"] = (3.5,3.5)

feat_name = 'CLAY_5'

display=PartialDependenceDisplay.from_estimator(model, X, [feat_name],
grid_resolution=200, percentiles=(0.05, 0.95))

plt.xlabel("Clay % at Depth of 75-90 cm")

plt.tight_layout()

plt.savefig("pdp_Clay5.png", dpi=300)

plt.show()

#plt.style.use('default')

from matplotlib import pyplot as plt

from sklearn.inspection import PartialDependenceDisplay

plt.rcParams["figure.figsize"] = (3.5,3.5)

feat_name = 'VPD'

display=PartialDependenceDisplay.from_estimator(model, X, [feat_name],
grid_resolution=200, percentiles=(0.05, 0.95))

plt.xlabel("Vapor Pressure Deficit")

plt.tight_layout()

plt.savefig("pdp_VPD.png", dpi=300)

plt.show()

#plt.style.use('default')
```

```
from matplotlib import pyplot as plt

from sklearn.inspection import PartialDependenceDisplay

plt.rcParams["figure.figsize"] = (3.5,3.5)

feat_name = 'NDVI'

display=PartialDependenceDisplay.from_estimator(model, X, [feat_name],
grid_resolution=200, percentiles=(0.05, 0.95))

plt.xlabel("NDVI")

plt.tight_layout()

plt.savefig("pdp_NDVI.png", dpi=300)

plt.show()

#plt.style.use('default')

from matplotlib import pyplot as plt

from sklearn.inspection import PartialDependenceDisplay

plt.rcParams["figure.figsize"] = (3.5,3.5)

feat_name = 'SWT'

display=PartialDependenceDisplay.from_estimator(model, X, [feat_name],
grid_resolution=200, percentiles=(0.05, 0.95))

plt.xlabel("Soil Water Tension")

plt.tight_layout()

plt.savefig("pdp_SWT.png", dpi=300)

plt.show()

#plt.style.use('default')

from matplotlib import pyplot as plt

from sklearn.inspection import PartialDependenceDisplay

plt.rcParams["figure.figsize"] = (3.5,3.5)

feat_name = 'Precipitation'

display=PartialDependenceDisplay.from_estimator(model, X, [feat_name],
grid_resolution=200, percentiles=(0.05, 0.95))

plt.xlabel("Precipiation")
```

```
plt.tight_layout()
plt.savefig("pdp_ppt.png", dpi=300)
plt.show()
#plt.style.use('default')
```