Precision nitrogen management: Site-specific management grids, soil nitrogen estimation, and optimum sampling

by

Wubengeda Admasu Yilma

B.S., Haramaya University, 2009
M.S., IHE Delft Institute for Water Education, 2017

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Agronomy
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Abstract

Integrating macro-scale variability in soil and micro-variability in crops and translating it into a practical management unit is a complex task. Past studies on precision nitrogen (N) management have focused primarily on employing large-scale variability in soil. More recently, sensing N in crop canopies to characterize micro-scale variability has been used for customized fertilizer applications. However, macro- and micro-scale variability have not been effectively coupled and adapted to match current fertilizer application technologies (i.e., individual nozzle level control and application of fertilizer). Likewise, temporal and micro-spatial uncertainty of data for crop variables, influenced by environmental and crop management practices, have not been adequately addressed.

A pre-requisite to most N prescription maps is quantification of soil NO_3-N spatial distribution in a field, which mandates intensive soil sampling and analysis. The complex mobility of nitrogen in soil, coupled with time, labor, and costs of analysis, underscores the necessity for an alternative method. Application of non-imaging hyperspectral sensing, characterized by high spectral resolution (1-3 nm bandwidth) of a spectroradiometer, facilitates a more comprehensive spectral analysis. This approach would have the potential to estimate surface soil NO_3-N rapidly and inexpensively without compromising accuracy in soil characterization.

Most often, commercial soil sampling designs uses a simplified random and/or stratified design, which entails collection of soil samples without accounting for spatial dependency. In instances where ancillary data such as proximal soil and grain yield are available, leveraging a sampling technique that accounts for geographic and feature space optimization of sampling configuration would be essential. In this context, spatial representations emerge as a critical facet

of site-specific management protocols, strategically designed to account for inherent complexities of within-field heterogeneity and spatial autocorrelation.

The objectives of this study were to: (i) delineate site-specific management grids that account for macro-scale variability in soil and micro-scale variability in crop; (ii) estimate surface soil NO_3-N using non-imaging hyperspectral sensing; and (iii) optimize soil sampling configuration with conditioned Latin Hypercube Sampling (cLHS) technique and hybrid cLHS-Sparse Sensor Placement Optimization for Reconstruction (SSPOR) techniques. Studies were conducted for each objective across two to four locations in Kansas and Colorado spread over four to six site years. For objective 1, a hybrid fuzzy inference system (H-FIS) was developed to delineate SSMGs that account for macro-scale variability in soil and micro-scale variability in crops. A proximal soil sensor, Veris-MSP3 was used to acquire soil variables. The Unmanned Aerial Vehicle (UAV) equipped with on-board MicaSense Rededge-3 multispectral sensor was used to collect multi-spectral images during the crop growing season. Adaptive Neuro-Fuzzy Inference System (ANFIS) was used to generate a tuned membership function (MFs) within the H-FIS framework. Crafted fuzzy rules and expert knowledge combined fuzzy rules were defined with H-FIS model output to generate SSMGs. A combination of selected soil and crop variables and crop grain yield was used to tune MFs with $R^2 > 0.7$. The H-FIS model was used to couple micro-scale variability in soil and macro-scale variability in crop with the selected soil and crop variables (Elevation, Slope, ECa, OM, and Normalized Difference Red Edge (NDRE)). The H-FIS output data were used to map the SSMGs. A classification with ten classes of grid-based management units was successfully delineated for all site years. A similar spatial trend was observed between SSMGs delineated based on coupled macro-scale variability in soil and micro-scale variability in crops and grain yield (corn and soybean) for all site years. Overall, the complex coupling of soil and crop variables accounting for both macro- and micro-scale variability was achieved. Each management grid represented a distinct area with specific productivity potential characterized into multiple classes. These classes are expected to be used as a 'weighted-coefficient' based on SSMGs to adjust N prescription maps and optimize applications of the N rate.

For objective 2, soil samples were analyzed for surface soil NO_3-N and spectral measurements were acquired with ASD FieldSpec3 spectroradiometer. Spectral preprocessing techniques were applied to refine the spectral data and remove noise. Effective spectral windows were determined such that spectral demonstrated sensitivity to changes in soil NO_3-N. The selected spectral windows: W1 (1460-1690 nm), W2 (1730-1780 nm), and W3 (1940-2150 nm) were used to estimate soil NO_3-N. Partial Least Squares Regression (PLSR) and Random Forest Regression (RFR) were used to estimate soil NO_3-N using selected spectral features and measured soil NO_3-N. Estimated surface soil NO_3-N maps were developed. Map of measured soil NO_3-N showed a similar spatial trend with estimated soil NO_3-N maps. The findings of potential hyperspectral data-based surface soil NO_3-N estimation models and maps are expected to provide a cost-effective and practical technique to improve soil fertility management practices.

For objective 3, an optimized soil sampling design was developed by integrating multiple years of grain yield data with proximal soil data (with Veris-MSP3 soil sensor). The hybrid cLHS-SSPOR and cLHS techniques with proximal soil+yield data sources were used to optimize soil sampling sites. Results with different optimal soil sample sizes for site year-1 and site year-2 indicated the need for site-specific sampling strategies hybrid cLHS-SSPOR and cLHS techniques. In this study, Kullback-Leibler divergence of hybrid cLHS-SSPOR-based technique indicated its capability to characterize local spatial heterogeneity compared to the cLHS. These findings emphasize the importance of tailored sampling techniques for different study areas to accurately generate a soil sampling design that can characterize soil spatial variability.

Overall, the results of this study introduced a data fusion method based on evidential reasoning that integrates macro-scale variability in soil and micro-scale variability in crops to delineate SSMGs. Non-imaging hyperspectral that uses spectroradiometer proves the potential of developing an accurate surface soil NO_3-N estimation model using selected spectral features. The importance of tailored sampling techniques was emphasized for different study locations to accurately generate a soil sampling design.

Precision nitrogen management: Site-specific management grids, soil nitrogen estimation, and optimum sampling

by

Wubengeda Admasu Yilma

B.S., Haramaya University, 2009
M.S., IHE Delft Institute for Water Education, 2017

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Agronomy
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2024

Approved by:

Major Professor
Dr. Raj Khosla

Copyright

Abstract

Integrating macro-scale variability in soil and micro-variability in crops and translating it into a practical management unit is a complex task. Past studies on precision nitrogen (N) management have focused primarily on employing large-scale variability in soil. More recently, sensing N in crop canopies to characterize micro-scale variability has been used for customized fertilizer applications. However, macro- and micro-scale variability have not been effectively coupled and adapted to match current fertilizer application technologies (i.e., individual nozzle level control and application of fertilizer). Likewise, temporal and micro-spatial uncertainty of data for crop variables, influenced by environmental and crop management practices, have not been adequately addressed.

A pre-requisite to most N prescription maps is quantification of soil NO_3-N spatial distribution in a field, which mandates intensive soil sampling and analysis. The complex mobility of nitrogen in soil, coupled with time, labor, and costs of analysis, underscores the necessity for an alternative method. Application of non-imaging hyperspectral sensing, characterized by high spectral resolution (1-3 nm bandwidth) of a spectroradiometer, facilitates a more comprehensive spectral analysis. This approach would have the potential to estimate surface soil NO_3-N rapidly and inexpensively without compromising accuracy in soil characterization.

Most often, commercial soil sampling designs uses a simplified random and/or stratified design, which entails collection of soil samples without accounting for spatial dependency. In instances where ancillary data such as proximal soil and grain yield are available, leveraging a sampling technique that accounts for geographic and feature space optimization of sampling configuration would be essential. In this context, spatial representations emerge as a critical facet

of site-specific management protocols, strategically designed to account for inherent complexities of within-field heterogeneity and spatial autocorrelation.

The objectives of this study were to: (i) delineate site-specific management grids that account for macro-scale variability in soil and micro-scale variability in crop; (ii) estimate surface soil NO_3-N using non-imaging hyperspectral sensing; and (iii) optimize soil sampling configuration with conditioned Latin Hypercube Sampling (cLHS) technique and hybrid cLHS-Sparse Sensor Placement Optimization for Reconstruction (SSPOR) techniques. Studies were conducted for each objective across two to four locations in Kansas and Colorado spread over four to six site years. For objective 1, a hybrid fuzzy inference system (H-FIS) was developed to delineate SSMGs that account for macro-scale variability in soil and micro-scale variability in crops. A proximal soil sensor, Veris-MSP3 was used to acquire soil variables. The Unmanned Aerial Vehicle (UAV) equipped with on-board MicaSense Rededge-3 multispectral sensor was used to collect multi-spectral images during the crop growing season. Adaptive Neuro-Fuzzy Inference System (ANFIS) was used to generate a tuned membership function (MFs) within the H-FIS framework. Crafted fuzzy rules and expert knowledge combined fuzzy rules were defined with H-FIS model output to generate SSMGs. A combination of selected soil and crop variables and crop grain yield was used to tune MFs with $R^2 > 0.7$. The H-FIS model was used to couple micro-scale variability in soil and macro-scale variability in crop with the selected soil and crop variables (Elevation, Slope, ECa, OM, and Normalized Difference Red Edge (NDRE)). The H-FIS output data were used to map the SSMGs. A classification with ten classes of grid-based management units was successfully delineated for all site years. A similar spatial trend was observed between SSMGs delineated based on coupled macro-scale variability in soil and micro-scale variability in crops and grain yield (corn and soybean) for all site years. Overall, the complex coupling of soil and crop variables accounting for both macro- and micro-scale variability was achieved. Each management grid represented a distinct area with specific productivity potential characterized into multiple classes. These classes are expected to be used as a 'weighted-coefficient' based on SSMGs to adjust N prescription maps and optimize applications of the N rate.

For objective 2, soil samples were analyzed for surface soil NO_3-N and spectral measurements were acquired with ASD FieldSpec3 spectroradiometer. Spectral preprocessing techniques were applied to refine the spectral data and remove noise. Effective spectral windows were determined such that spectral demonstrated sensitivity to changes in soil NO_3-N. The selected spectral windows: W1 (1460-1690 nm), W2 (1730-1780 nm), and W3 (1940-2150 nm) were used to estimate soil NO_3-N. Partial Least Squares Regression (PLSR) and Random Forest Regression (RFR) were used to estimate soil NO_3-N using selected spectral features and measured soil NO_3-N. Estimated surface soil NO_3-N maps were developed. Map of measured soil NO_3-N showed a similar spatial trend with estimated soil NO_3-N maps. The findings of potential hyperspectral data-based surface soil NO_3-N estimation models and maps are expected to provide a cost-effective and practical technique to improve soil fertility management practices.

For objective 3, an optimized soil sampling design was developed by integrating multiple years of grain yield data with proximal soil data (with Veris-MSP3 soil sensor). The hybrid cLHS-SSPOR and cLHS techniques with proximal soil+yield data sources were used to optimize soil sampling sites. Results with different optimal soil sample sizes for site year-1 and site year-2 indicated the need for site-specific sampling strategies hybrid cLHS-SSPOR and cLHS techniques. In this study, Kullback-Leibler divergence of hybrid cLHS-SSPOR-based technique indicated its capability to characterize local spatial heterogeneity compared to the cLHS. These findings emphasize the importance of tailored sampling techniques for different study areas to accurately generate a soil sampling design that can characterize soil spatial variability.

Overall, the results of this study introduced a data fusion method based on evidential reasoning that integrates macro-scale variability in soil and micro-scale variability in crops to delineate SSMGs. Non-imaging hyperspectral that uses spectroradiometer proves the potential of developing an accurate surface soil NO_3-N estimation model using selected spectral features. The importance of tailored sampling techniques was emphasized for different study locations to accurately generate a soil sampling design.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my advisor, Dr. Raj Khosla, to whom I am very grateful for providing me this opportunity. His consistent advice, encouragement and guidance during my graduate study have been instrumental in making this research possible. He has taught me a unique way of thinking in solving challenges which also helps me in my everyday life. I would like to thank my committee members, Dr. Nathan Nelson, Dr. Gerard Kluitenberg, and Dr. Ajay Sharda, for their invaluable support throughout my study. I also wish to express my sincere appreciation to Dr. Jim Ippolito, who provided essential assistance at the beginning of my PhD studies.

Acknowledgment is made to the BARD US-Israel Agricultural Research and Development fund for generously sponsoring my research. This support has been a key enabler in the pursuit of my academic goals and the advancement of my research.

I am deeply thankful to the Dr. Raj Khosla Precision Ag Lab for providing a working environment, vital resources, and continuous support. The lab's good team spirit has significantly contributed to my personal and professional development. I extend my gratitude to the members of the lab; I would also like to thank Dr. Dipankar Mandal for his unwavering support and guidance. I am grateful Dr. Mandal for listening my ideas and encouraging me to learn new skills. I acknowledge Dr. Jeffery Siegfried, for teaching me UAV operations and image processing. My sincere thanks also go to Ross Unruh for his help in collecting research data and managing laboratory operations.

Lastly, my profound gratitude goes to my family and friends. Their patience, support, guidance, and encouragement have been the backbone of my journey.

Dedication

I would like to dedicate this work to all my family members that have encouraged me to dream for big. I would like to dedicate it to my grandmother Abaynesh Nigatu and late mother Kelemua Argaw.

Chapter 1 - Coupling Macro and Micro-scale Variability for Delineation of Site-Specific Management Grid

1.1 Abstract

Managing spatial and temporal variability in soil properties and crop characteristics have been studied widely for precision nutrient management. Quantifying variability in soils often provides a macro-scale characterization of soil properties. Conversely, high-resolution reflectance-based characterization of crop canopy has enabled quantification of crop characteristics at a micro-scale. This study aims to delineate a site-specific management unit that accounts for both macro-scale-scale spatial variability in soil and micro-scale-scale spatial variability in crops. Translating coupled multi-scale variables to a management grid is complex due to lack of a clear threshold to determine the scale of variability. Furthermore, the dynamic influence of the environment and crop management practices on crop Nitrogen (N) demand brings uncertainty to data. Therefore, this study employed a hybrid fuzzy inference system (H-FIS) to couple soil and crop variables. The specific objectives of this study were: (1) to assess the influence of spatial patterns of micro-scale and macro-scale variables on crop grain yield; (2) to generate H-FIS based Site-Specific Management Grids that accounts for macro and micro variability in soil and crop properties; and (3) To validate H-FIS based SSMGs developed in objective (2). The study was conducted over four site years at Kansas River Valley (KRV) experimentation fields in Topeka and Rossville, Kansas. A proximal soil sensor, Veris-MSP3 was used to acquire soil variables. The Unmanned Aerial Vehicle (UAV) on-board MicaSense Rededge-3 multispectral sensor was used to collect multi-spectral images. These imageries were subsequently processed to calculate vegetation indices including NDVI (Normalized Difference Vegetation Index), SAVI (Soil Adjusted Vegetation Index), NDRE (Normalized Difference Red Edge) as proxy to crop growth. Soil and crop variables selection was conducted to choose the most influential variables to develop H-FIS model. Adaptive Neuro-Fuzzy Inference System (ANFIS) was used to generate a tuned membership function (MFs) within the H-FIS framework. Crafted fuzzy rules and expert knowledge combined fuzzy rules were delineated with H-FIS model output to generate SSMGs. A combination of selected soil and crop variables and yield was used to tune MFs with $R^2 > 0.7$. H-FIS model used to couple micro-scale variability in soil and macro-scale variability in crop with the selected soil and crop variables (Elevation, Slope, ECa Shallow, ECa Deep, soil organic matter, and NDRE). The H-FIS output data were used to map the SSMGs. A classification with ten classes of grid-based management units was successfully delineated for all site years. A similar spatial trend was observed between coupled variable SSMGs and grain yield (corn and soybean) for all site years. The degree of agreement between coupled soil-crop variables based SSMGs and solely macro-scale variability based SSMGs indicated a dissimilarity with kappa coefficient (κ) ranging between -0.013 and 0.16. Coupled variables based SSMGs resulted an expected weaker spatial autocorrelation indicating a varied spatial pattern (not clustered as zones) than solely macro-scale variability based SSMGs. SSMGs were validated with soil ground truth data, as a result OM and CEC were significantly different between a potential low and high SSMGs. H-FIS produced a grid-based management units with and without constrained classes, suitable for sprayers. Overall, the complex coupling of soil and crop variables accounting for both macro- and micro-scale variability was achieved.

1.2 Introduction

Natural variation in soil and crop management practices are major causes of spatial and temporal variability in an agricultural field (Mzuku et al., 2005; Cordero et al., 2019). Understanding the reason and scale of spatial variability in a field constitutes a first step in designing an effective precision nutrient management strategy (Khosla et al., 2002). Optimum or efficient nutrient management, especially nitrogen (N) management is crucial for sustaining crop yields under the changing climate. The global nitrogen use efficiency (NUE) is less than 50% (Omara et al., 2019; McKay Fletcher et al., 2022). Most of the surplus N, which refers to the amount of N applied to soils not taken up by plants, is released into the environment and undergoes transformations, impacting different ecosystem compartments through the N cascade process. In the United States, the US Midwest (also known as the Corn Belt) which produces about 33% of the world's corn for food, feed, and biofuel (USDA and FAS, 2023), low productivity potential area contributed 44% of N loss to the environment (Basso et al., 2019). To mitigate losses, several precision N management practices have been proposed in literature. It suggests matching crop requirements with N applications. In the last two decades, several in-season N management strategies have been developed and tested with precision agriculture practices (Khosla et al., 2002; Cao et al., 2017; Cordero et al., 2019). A typical method to manage nutrient variability in soil is to classify fields into management units and apply fertilizer according to productivity potential of each management unit. In literature, these management units have been referred to as Site-specific management zones (SSMZ), which are areas in a field with similar soil properties that can be managed uniformly for growing crops (Mulla and Khosla, 2015). Accordingly, defining and optimizing homogeneous management units is vital to achieving appropriate use of soil and agricultural resources. Soil-based SSMZ-guided variable rate fertilizer applications were found to be efficient for areas with large-scale spatial variability of soil properties (Cordero et al., 2019).

In the literature, several methods have been proposed to delineate management zones based on soil properties (Derby et al., 2007; Peralta et al., 2015; Haghverdi et al., 2015; Tripathi et al., 2015; Gili et al., 2017; Rossi et al., 2018), topographic attributes (Fraisse et al., 2001), and remote sensing data over bare soil (Mulla, 2013; Georgi et al., 2018), these methods accounted primarily macro-scale variability in soils. Longchamps and Khosla (2017) reported that out of the 162 delineation techniques reviewed, 37% of them used soil physical and chemical properties, followed by data collected with proximal and remote sensing methods. Numerous studies showed that soil-based SSMZs characterized macro-scale spatial variability of soil (Koch et al., 2004; Schepers et al., 2004; Corwin and Lesch, 2005).

With the advancement of precision agricultural machineries (e.g., GPS-equipped grain yield monitors) and rapid evolution of remote sensing technologies to acquire a high spatial, spectral, and temporal-resolution imagery, management units have also been delineated using crop yield maps (Diker et al., 2004; Derby et al., 2007; Damian et al., 2017), and remote sensing data (Jin et al., 2017; Breunig et al., 2020; Rasmussen et al., 2021). Both, yield maps and high-resolution imagery, enabled characterization of micro-scale variability of crops. Likewise, sprayer technology for fertilizer application has undergone significant development in recent years. From a single rate uniform application across the entire sprayer boom to variable rate application, to section control systems, to more recently advanced nozzle control systems, which allow separately controlling each nozzle, and adjusting the rate in real-time (Kitchen et al., 2010; Guerrero and Mouazen, 2021). Methods for varying N within a field are justified by the spatially variable nature of N mineralization and potential N losses over nonuniform agricultural landscapes. Uniform N applications across field ignore that N supply from soil, crop N uptake, and response to N are not spatially uniform (Inman et al., 2005).

Accounting only macro-scale variability in soil for management units is not sufficient because it fails to account for in-season crop nitrogen (N) demand influenced by weather and crop management practices (Shanahan et al., 2008). Even though quantifying micro-scale variability has been studied with crop N sensors, there are very few studies that tried to integrate macro-scale and micro-scale variability for delineation of a site-specific management unit. Assimilating soil with crop information was assessed by combination of datasets that represents macro-scale variability in soil and micro-scale variability in crop (Derby et al., 2007; Breunig et al., 2020). In principle, the best predictors of management units should be spatially correlated with crop yield (Gavioli et al., 2016). Coupling macro-variability with micro-variability can create an opportunity to assess and manage both, macro-scale and micro-scale variability, of soil and crop parameters, respectively (Cordero et al., 2019). Integrating crop data with soil information can help identifying soil-based constraints and tailor appropriate management strategies. In a multi-scale data fusion perspective, soil and crop parameters can be coupled together by assigning spatial weighting factors for each parameter (Gavioli et al., 2016). An optimal nutrient management strategy stands to gain advantage by integrating both macro-scale variability in soil and micro-scale variability observed in crop canopy. Soil-plant relationship often exhibits non-linearity exhibiting variation in crop growth in response to soil variability (Shatar and Mcbratney, 1999). This variation is attributed to the influence of other factors such as plant diseases, management practices, weed competition, and pests, which may be more relevant in specific areas.

Analyzing non-linearity in both macro-scale and micro-scale variables is contingent on the context and lacks a clear threshold to determine scale of variability. The spatial crop N demand

and uptake are highly dependent on the weather and their respective interaction effects. However, weather conditions are complex with a high level of uncertainty (Hasan et al., 2008) and are influential to micro-scale variability of crops. Embracing ambiguity and uncertainty in soil and crop data may help in building a holistic and adaptive approach to managing agricultural fields. Assimilation of these datasets can be accomplished in a fuzzy inference system (Guillaume, 2001), which is recognized for its effectiveness in handling complicated interaction systems particularly under ambiguity and uncertainty conditions. Few studies had incorporated fuzzy inference system for agronomic decisions related works to account probabilistic predictions (Urretavizcaya et al., 2014; Vallentin et al., 2020; Heiß et al., 2023; Bökle et al., 2023). Consequently, the delineation of management units can consider a spectrum of possibilities and their respective likelihoods, moving away from single deterministic prediction. Such an approach provides more precise management units, tailored based on grid size applicable for farm implement width and also accounts for macro and micro-variability in the system. Considering the size of these smaller management units to match individual nozzle level fertilizer delivery system, these management units are more appropriately referred to as 'Site-Specific Management Grid (SSMG)'.

1.2.1 Hypothesis

The macro-scale spatial variability in soil and micro-scale spatial variability in crop canopies can be accounted for by delineating site-specific management grids.

1.2.2 Objectives

The overall objective of this study was to delineate site-specific management grids that accounts for macro-scale variability in soil and micro-scale variability in crop. The specific objectives were to:

i. Assess the influence of micro-scale and macro-scale variables on spatial patterns of grain yield.

ii. Generate fuzzy inference system based site-specific management grids that accounts for macro and micro variability in soil and crop properties, and

iii. Validate fuzzy inference system based site-specific management grid developed in objective (ii).

1.3 Methodology

1.3.1 Study Sites

This study was conducted over a period of two years (2021 and 2022) at two locations in Kansas, over a total of four site-years. The sites are distributed across Kansas River Valley (KRV) experiment fields at Topeka (39° 4’ 34.04"N, 95° 46’ 11.98"W) and Rossville (39° 7’ 6.36"N, 95° 55’ 37.35"W) stations (Figure 1.1). Studies conducted at KRV Topeka experiment field were designated as ‘site year-1’ and ‘site year-2’ for 2021 and 2022, respectively, and studies conducted at KRV Rossville experiment field were labelled as ‘site year-3’ and ‘site year-4’ for 2021 and 2022, respectively. Site year-1 and 2 covers an area of 3.36 ha, while site year-3 and site year-4 acreage is 7.76 ha. Both study sites are equipped with lateral move irrigation systems.

(a) (b) (c)

Figure 1.1. Study site locations over Kansas. (a) Kansas Research Valley Rossville and Topeka stations highlighted as yellow stars over Landsat composite image (Courtesy: Google Earth); (b) Kansas Research Valley Rossville experiment field (Site year-3 and Site year-4), and (c) Kansas Research Valley Topeka experiment field (Site year-1 and Site year-2) highlighted with red boundary.

Weather data was acquired from the closest Kansas Mesonet station. (Appendix B - Weather information for all four site years.). For site year-1, daily average temperature varied between a minimum of -20.9°C to a maximum of 31.5°C. Similar trend was observed for site year-2, with a minimum of -15.95°C and maximum of 31.3°C. A precipitation of 816 mm in 2021 and 756 mm in 2022 was recorded at for site year-1 and site year-2, respectively. Precipitation recorded for site-year-1 and site year-2 were comparable to the 10-year average precipitation. Daily average temperature for site year-3 ranged between a minimum of -21.1°C to a maximum of 31.6°C. A similar trend was reported for site year-4, with a minimum of -15.95°C and maximum of 31.4°C. A precipitation of 814 mm and 783 mm was recorded for site year-3 and site year-4, respectively.

Study sites soils were mapped as Eudora-Bismarckgrove silt loam formed from alluvial and silty alluvium parent materials (Soil Survey Staff, 2024) Even though flooding occurs occasionally, the KRV Topeka field exhibits good drainage with a varying slope from 0 to 1%. Soil salinity levels range from non-saline to slightly saline, with measurements between 0 to 2 mmhos/cm. The soil for site year-3 and site year-4 consists of four distinct soils, each exhibiting unique characteristics. The Northwest part was mapped as Kimo silty clay loam and Bismarckgrove-Kimo. It was characterized as poorly to moderately drained nature and a gentle slope of 0 to 1%. The central part of study area exhibits a diagonal ridge of Belvue silt loam with well-drainage properties and slope of 0 to 1%. The southern section of the field lies on Eudora silt loam, characterized as a class with well-drainage with a gentle slope of 0 to 1%. The soil salinity fore site year-3 and site year-4 consists non-saline to very slightly saline classification, with measurements between 0 to 2 mmhos/cm.

1.3.2 Crop Cultivation and Management Practices

Best crop management practices were employed across all site years. Efforts were made to keep yield-limiting factor as minimum as possible, uniform infield management practices were applied. Crop rotation of corn (*Zea mays* L.) and soybean (Glycine max) was practiced on both study sites. Corn fields received fertilizer on two occasions: dry fertilizer (11-104-104-26 (Su)-1.74 (Zn)) in the fall season (in month of October) and 68 kg N ha^{-1} as anhydrous ammonia (NH_3) in the spring season (in month of May). (See Appendix Table C.1. Crop cultivation and management practices).

1.3.3 Data Collection

1.3.3.1 Soil Data Collection

Trimble *Real-time kinematic* positioning (RTK) system equipped, John Deere 5055E tractor was coupled with Veris-MSP3 (Veris Technologies, Inc., Salina, KS, USA) system to acquire apparent soil Electrical conductivity at depth of 0-30 cm (ECa shallow), apparent Electrical conductivity at depth of 0-90 cm (ECa deep) and organic matter (OM) (Figure 1.2 a). The Veris-MSP3 data were collected for site year-1 and 2 and site year-3 and 4 on 29 March and 31 March in 2021, respectively.

The Veris-MSP3 was operated at a speed of 4 to 5 km/h with transect intervals of 10 to 12 meters (Figure 1.2 b). In a transect, data recording was set in an interval of 1 to 1.5 m in the direction of plant rows. Optical sensor mounted on VERIS-MSP3 unit was to detect soil reflectance in both red and near-infrared bands (Gundy and Dille, 2022). ECa at shallow and deep depth, OM, CEC, altitude, and slope were processed in FieldFusion (Veris Technologies Inc.) web platform. ECa and OM were calibrated with ground truth soil data (OM, pH, and soil texture).

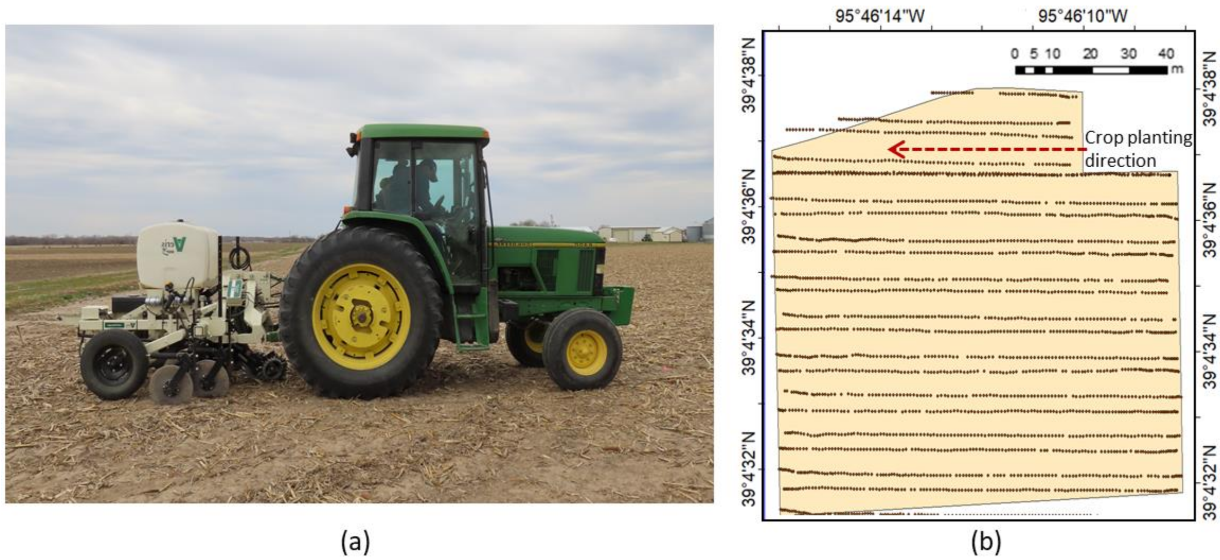


Figure 1.2. Veris-MSP3 mapping unit (a) Veris-MSP3 scanning soil to collect ECa at shallow and deep (with disk coulter), and OM (with optical sensor) at Kansas Research Valley Topeka experiment field (Site year-1 and Site year-2); (b) field data collection scheme overlaid on field boundary map.

1.3.3.2 Multi-spectral Images Collection

For both the study sites, aerial imagery was collected with a MicaSense Rededge-3 multispectral sensor (MicaSense, Seattle, WA, USA) mounted on a DJI Matrice 100 quadcopter (DJI, Nanshan, Shenzhen, China) unmanned aerial vehicle (UAV) as presented in Figure 1.3 a-b. The multi-spectral sensor has 5 wavebands: Blue (475±20 nm), Green (560±20 nm), Red (668±10 nm), Red Edge (717±10 nm), and Near-infrared (840±40 nm). Photogrammetry missions assisted with UgCS v.4.18 for DJI were flown in a grid pattern with 75% front and side overlap at 50 m above ground level (See Appendix D - UAV flight plan preparation).

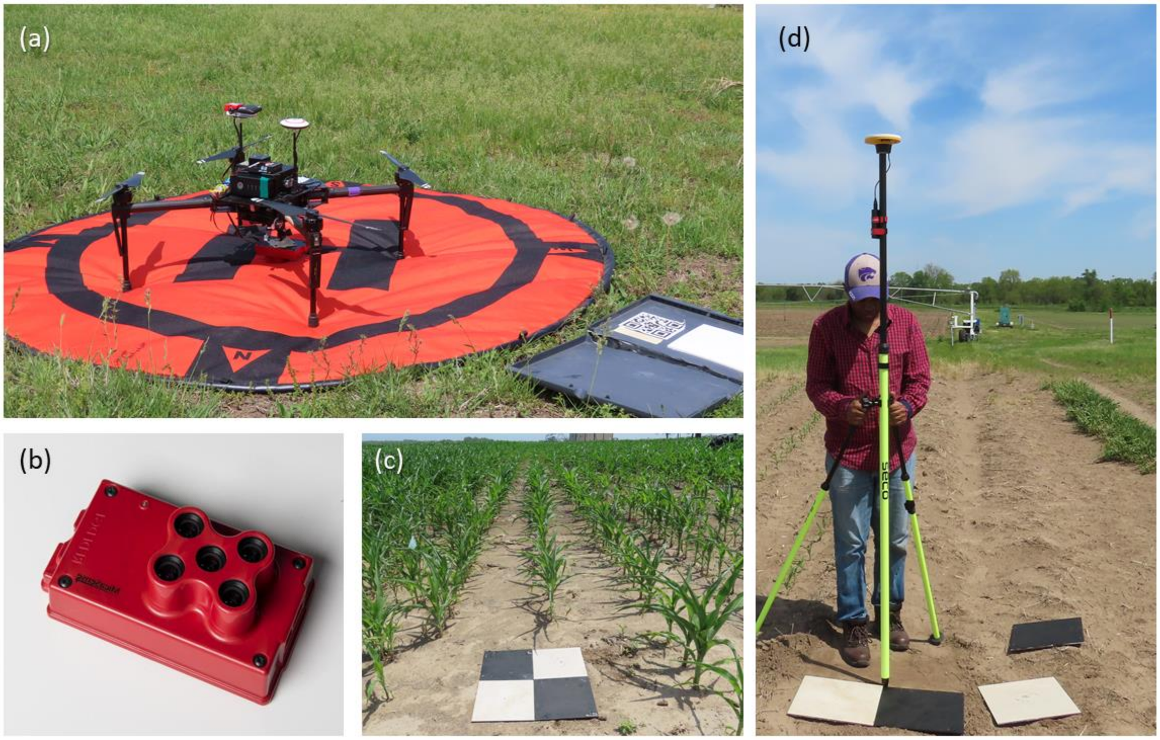


Figure 1.3. Multispectral image acquisition (a) DJI Matrice 100 quadcopter UAV on launching pad and calibration reflection panel; (b) MicaSense Rededge-3 multispectral sensor; (c) Ground control points (GCP) placed in crop rows; (d) Colleting GCP coordinates with Trimble DA2 GNSS receiver.

The calibration reflectance panel images were acquired before takeoff to be used for calibrating multi-spectral images acquired from the crop field (Figure 1.3 a). In addition, ground control points (GCP) were used for the aerial triangulation process for a series stack of UAV multi-spectral images. The Trimble DA2 GNSS receiver with Real-Time Kinematic (RTK) corrections (±1.0 cm accuracy) was employed to acquire GCP coordinates for geometric calibration (Figure 1.3 c and d). To enhance visibility, GCPs were prepared with four black and white tiles to create a checkerboard pattern (Figure 1.3 c). The GCPs were strategically placed at four corners and in the center of the field. The coordinates were collected from the center of GCP tiles. UAV data acquisitions dates for all four site years are presented in Appendix E - UAV imagery collection.

1.3.4 Multi-spectral Image Processing

Imageries collected on each flight date were processed with Pix4D mapper Pro software (Pix4D, Lausanne, Switzerland) to create georeferenced and orthorectified mosaic images. The image processing was executed in a three-step workflow: initial processing, generation of dense point cloud and mesh, and creation of Digital Surface Model (DSM), orthomosaic and index generation. As a preprocessing stage, data preparation was performed by removing images taken outside target area, blurred images, and duplicated calibration panel images. During the initial processing step, key-point were automatically identified from each image to establish a relative position between images. Additionally, geometric, and radiometric correction processes were performed to rectify multi-spectral images for spatial distortion and ensure UAV images reflect true intensity, respectively. The geometric correction process starts with importing GCP coordinates (latitude, longitude, and elevation), followed by manual marking of GCP centers to link GCPs with known locations. Radiometric correction was performed on Pix4D mapper with camera and sun irradiance correction approach. The calibration panel images were imported to Pix4D mapper, and then marked an area on center of panel image. The known reflectance value of each band was applied to each marked area to calculate a conversion factor to transform Digital Number (DN) values into reflectance values. In the second processing phase, a point cloud was created by matching key-point. The point cloud was used to create a 3D textured mesh to generate DSM, orthomosaic and indices at third and final stage phase. Vegetation indices were calculated with Pix4D mapper index calculator tool using the respective band reflectance.

Multi-spectral images collected for the study sites were processed to generate NDVI (Normalized Difference Vegetation Index), SAVI (Soil Adjusted Vegetation Index), NDRE

(Normalized Difference Red Edge), and DSM maps. The formulae used to calculate NDVI, SAVI, and NDRE indices (Sishodia et al., 2020; Niu et al., 2023) are presented in Table 1.1.

Table 1.1. Vegetation indices generated from multi-spectral imagery.

Vegetation indices	Formula
Normalized difference vegetation index (NDVI)	$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{Red}}{\mathrm{NIR} + \mathrm{Red}}$
Soil Adjusted vegetation index (SAVI)	$\mathrm{SAVI} = \frac{\mathrm{NIR} - \mathrm{Red}}{\mathrm{NIR} + \mathrm{Red} + \mathrm{L}} \times (1 + \mathrm{L})$
Normalized Difference Red Edge (NDRE)	$\mathrm{NDRE} = \frac{\mathrm{NIR} - \mathrm{RedEdge}}{\mathrm{NIR} + \mathrm{RedEdge}}$

Note: In the SAVI formula, L is an adjustment factor that minimizes soil brightness influences and varies based on vegetation density. L is set to 0.5 for high vegetation cover and 0 for bare soil conditions. NIR stands for Near-Infrared Red.

1.3.5 Soil Sample Collection and Analysis

Soil samples were collected from all four site years. Soil sampling locations were identified with a stratified random soil sampling design at 0.202 ha grid size. One composite soil sample, consisting of 8 to 10 cores, were collected from 0 to 30 cm from each grid cell and recorded with a differential global positioning system (DGPS). The collected soil samples were air-dried and passed through a 2-mm sieve and were analyzed at a commercial lab. (Complete list of soil parameters and measurement techniques are presented in Appendix F - Soil parameters estimated from samples collected.

1.3.6 Crop Yield Data Collection

Corn and soybean grain yield maps were collected with a CASE IH Axial-Flow 2344 (Case IH, Racine, WI) combine harvester equipped with a DGPS receiver and yield monitor. Table 1.3

show the harvesting date, average moisture, and estimated dry grain yield. The SMS AgLeader and ArcGIS Pro 3.1.1 software were used to process harvested raw data according to the guidelines provided by Sudduth and Drummond, 2007. The grain yield was adjusted to 15.5 % and 11.5% commercially acceptable moisture content of corn and soybean, respectively. (Summarized grain yield monitor data for all sites are presented in Appendix Table H.1).

1.3.7 Data Pre-processing for Soil and Crop Variables

Three major categories of data were used for this study, which include (a) Topography data: DSM and slope (%), (b) soil data: Soil ECa (mS/m) at both deep and shallow depth, OM (%), and CEC, and (c) multi-spectral images: NDVI, SAVI, and NDRE to delineate SSMZs/SSMGs. The geolocations of Veris measurement points at every 10 to 12 m spacing (Figure 1.2 b) was laid out as a reference to extract pixel value from DSM, slope (%), NDVI, SAVI, and NDRE maps. The pixel values were extracted for all site years with ArcGIS Pro 3.1.1 software. Due to the sparse canopy presence at early stage of the crop, iso-clustering and masking techniques were employed on RGB images to segment the crop from ground. A total of 12 to 18 soil and crop variables vector point data were used for further analysis. The spatial variability of soil variables was quantified and visualized by plotting semi-variance against the lag-distance for all observations. A variogram model was fitted to compute semi-variogram using “gstat” and “sp” packages in R. This analysis determined the spatial range beyond which observations showed spatial independence. The estimated spatial range was used to confirm the expected macro-scale spatial variability of soil variables.

1.3.8 Model Development and Statistical Analysis

The schematic workflow presented in Figure 1.4 show the statistical analysis and model development for delineation of site-specific management grids by coupling macro-scale variability of soil and micro-scale spatial variability crop. The methodological details are illustrated in following subsections.

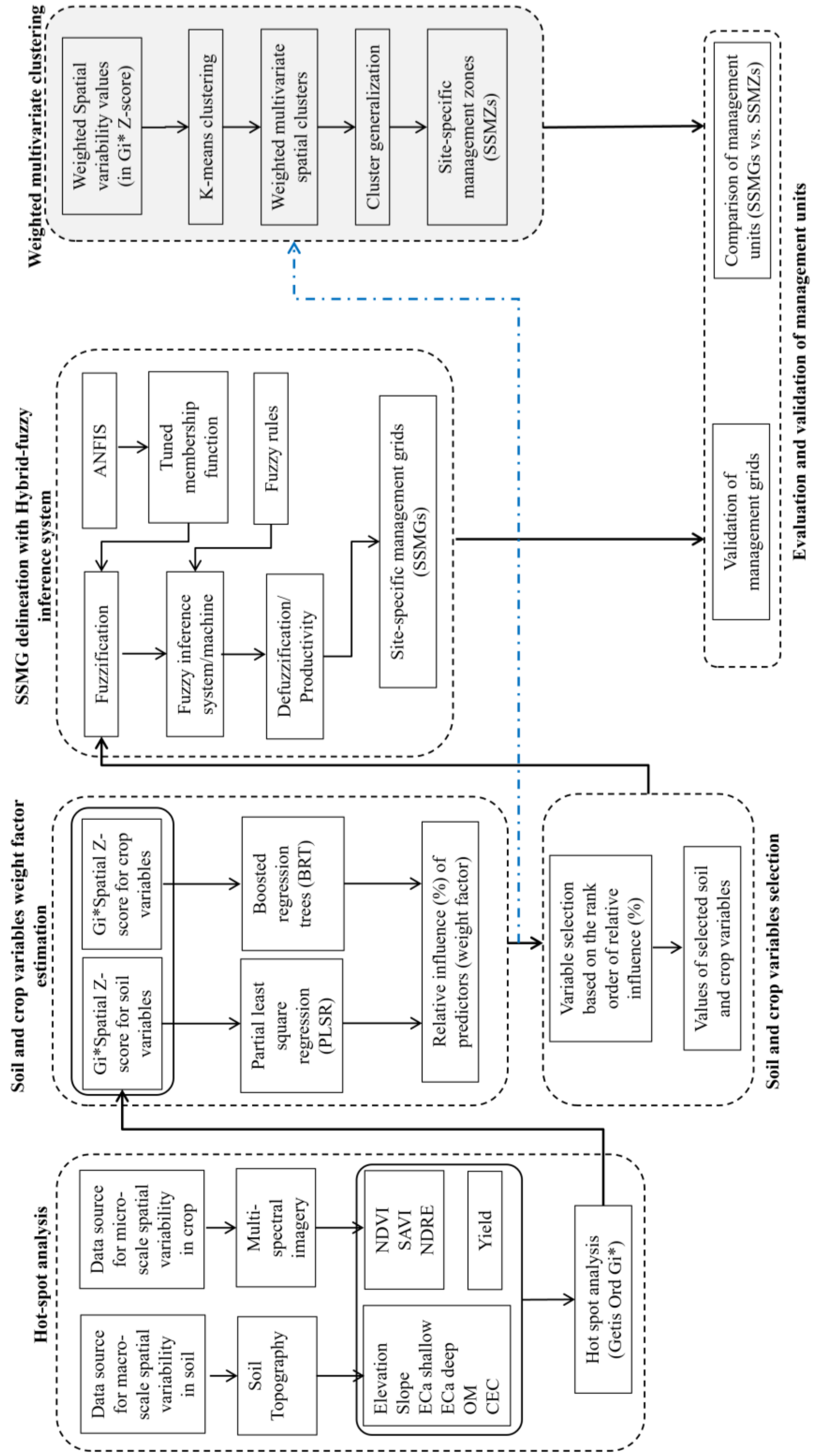


Figure 1.4. Schematic workflow to delineate site specific management grid based on macro-scale variability in soil and micro-scale variability in crop canopy.

1.3.8.1 Hot-spot Analysis

The degree of auto-correlation for soil, crop, and grain yield variables were estimated with Getis-Ord Gi* hot-spot analysis. Getis-Ord Gi* hot-spot analysis measured degree of auto-correlation between spatial points, thereby generating a z-score and p-value for each point to reveal the strength and direction of association with neighboring points (Getis and Ord, 1992). Hot spot analysis was conducted using Getis-Ord Gi* hot-spot statistical method using python programming language (Python, version 3.7.1, Python Software Foundation) and ArcGIS pro 3.1.1 arcpy module. This analysis resulted in a spatially characterized soil, crop and yield variables expressed in standardized Getis-Ord Gi* z score. The Getis-Ord Gi* z-score values were mapped with ArcGIS pro 3.1.1 to visualize statistically significant clustering of high and low values across study sites.

1.3.8.2 Soil and Crop Variables Selection and Weighting Factor Estimation

Soil and crop variables selection were made through a ranking variables based on their relative influence on spatial variability of grain yield. The standardized Getis-Ord Gi* z score values of crop variables were used as an input to estimate their respective impact on standardized Getis-Ord Gi* z score of yield. A study by Gebbers et al. (2011) reported that soil variables collected with Veris MSP3 shows a high multicollinearity and recommended to use Partial Least Squares Regression (PLSR) to examine relative importance of soil variables on grain yield. Conversely, relative influence of predictor variables on the spatial variability of grain yield was estimated with Gradient Boosted Regression Tree (GBRT) (Ohana-Levi et al., 2019). GBRT proved to be an appropriate multivariate model for a dataset with less frequency of multicollinearity (Elith et al., 2008). Therefore, PLSR and GBRT models were used to examine the relative importance of soil and crop predictor variables, respectively. PLSR and GBRT were computed with R programming language with "gbm", and "plsr" function from "dplyr" package.

Initially, caret package was employed to split the datasets to 60% and 40% random training and testing sets, respectively (Kuhn, 2008). The data split was established with the purpose to create robustness in the model and avoid overfitting. The estimated relative influence of predictor variables was plotted with "ggplot2" in R. In addition, crop variables with a Pearson correlation coefficient above 0.85 were removed to avoid multicollinearity (Farrar and Glauber, 1967). R programming language with "cor" function was used to estimate the Pearson correlation coefficient. The estimated relative influence was ranked to determine the most influential three to four crop variables that influenced the spatial variability of grain yield and were selected to delineate SSMGs using weighted multivariate clustering and fuzzy logic system. Weighting factors for soil and crop variables were set to 0.5. The estimated relative influence of each variable was designated as weight factor to estimate weighted Getis-Ord Gi* z-score values, which was used subsequently for weighted multivariate clustering. It must be noted that the weighted Getis-Ord Gi* z-score values were used only for weighted multivariate clustering, not for fuzzy logic system.

1.3.8.3 SSMG Delineation using Hybrid-Fuzzy Inference System (H-FIS)

A total of seven to eight soil and crop variables that influenced spatial variability of yield were selected per site-year to develop fuzzy logic system. For each site-year, soil and topography variables including, DSM (m), slope (%), ECa (mS/m) at shallow, ECa (mS/m) at deep depth, OM (%), and CEC in addition to one or two vegetation indices (NDVI, SAVI, and NDRE) used as input variables and their respective yield/productivity as an output variable. The delineation of H-FIS based SSMZs coupled variables that dignify macro-scale variability in soil and micro-scale variability in crop. H-FIS uniquely integrates membership functions generated via an ANFIS with expert-defined fuzzy rules and output functions. The system was designed and executed with Matlab scripted code within Fuzzy Logic Designer environment. The Fuzzy logic system was performed with three major steps of data processing: (a) fuzzification, (b) rule base and inference system, and (c) defuzzification.

Fuzzification

Fuzzification was executed to transform the crisp, exact values of soil and crop variables into fuzzy values using membership functions. Fuzzy values use a linguistic category that could provide a more human-reasoning that embraces uncertainty of soil and crop data due to natural systems. For example, instead of strictly categorizing a single ECa shallow value as "low", "medium" or "high", a membership function was used to assign a 70% association (0.7 degree of membership) with the "high" category and a 50% association (0.50 degree of membership) with the "medium" category and 35% association (0.35 degree of membership) with the "low" category. These fuzzy values are derived from predefined fuzzy sets (linguistic category), such as "low", "medium", and "high", allowing for a nuanced representation that captures the inherent variability and uncertainty present in soil parameter.

Membership functions (MFs)

A MF is a curve that maps each value of soil and crop variables to a degree of membership between 0 and 1. A Gaussian bell-shaped MF was used to define fuzzy sets into High, Medium, and Low for soil and crop input variables. The fuzzy sets were pre-defined based on an expert opinion that High, Medium, and Low represent a distinction in fuzzy logic system. The Gaussian bell-shaped MF is represented by Eq. (1) and the parameters x, σ and c characterize the membership function. On the other hand, the productivity MFs were set with zero-order Sugeno-type system, which characterized a constant output MFs. As a result, the output was constant value for each rule in the fuzzy inference system.

$$\boldsymbol{f}\ (\boldsymbol{x}; \boldsymbol{\sigma}, \boldsymbol{c}) = \boldsymbol{e}^{-\frac{(\boldsymbol{x}-\boldsymbol{c})^2}{2\boldsymbol{\sigma}^2}} \qquad (\mathbf{1})$$

Where:

c is the value where the membership function reaches its maximum (or the center of the curve),

σ standard deviation, that controls the width of the bell curve.

Setting up parameters for the membership functions is non-trivial. Tuning is essential to better characterize the MFs.

Tuning membership functions

The parameters of MFs were tuned using ANFIS. DSM, slope, ECa at shallow depth, ECa at deep depth, OM, CEC, and NDVI, SAVI, and NDRE used as input variables while their respective yield as an output variable. ANFIS model, a hybrid learning algorithm of neural-network and fuzzy inference system structure employed five layers (Figure 1.5). The first layer holds the membership layer which categorizes fuzzy sets into High, Medium, and Low. The second layer evaluates the firing strength with "AND" operation, which results in weight of each rule. Under layer 3, the firing strengths (or rule weights) were normalized. The fourth layer calculates system output for each rule using defuzzification. The last layer aggregates the outputs to produce a single crisp yield output. The defuzzification layer consists of multiple polynomials (e.g., f1, f2...as indicated in Figure 1.5), which are mapping functions of input variables. Coefficients of these polynomials, i.e., pi and qi were adjusted during ANFIS training phase. The adjusted coefficients tuned down membership function while minimizing error between predicted and actual yield.

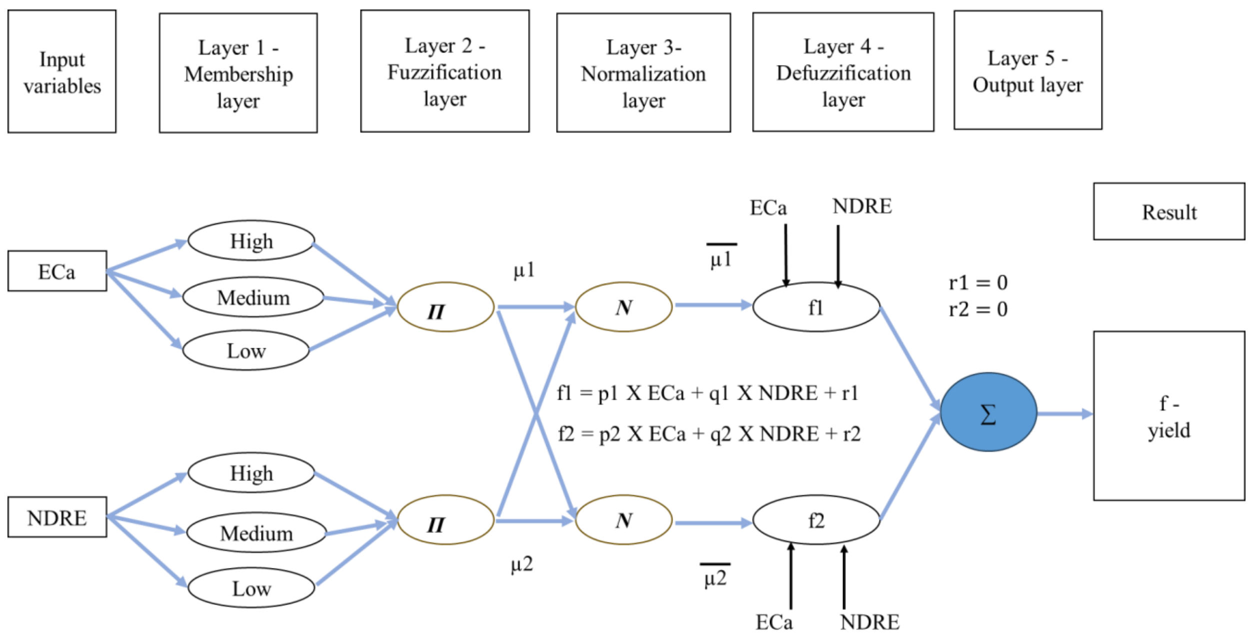


Figure 1.5. The architecture of Adaptive Neuro-Fuzzy Inference System (ANFIS) for tuning membership functions.

ANFIS adjusted the parameters (c and standard deviation to define the center and width of Gaussian curve) of membership functions in antecedent (input) and consequent (output). A combination of the least squares method and backpropagation gradient descent methods were used for parameter optimization to tune membership function (Naderloo et al., 2012; Taghavifar and Mardani, 2014). The training process tuned membership functions to result in a fuzzy system's output that was closely approximated to the output. The predicted yield from ANFIS was compared with actual yield to evaluate model performance. The trained ANFIS model was evaluated with error metrics. The tuned membership function which resulted in a good yield prediction, was chosen as acceptable tuned membership function. The prediction accuracy of yield by ANFIS model was indicated by calculated RMSE and r^2 value. The ANFIS tuned membership function was further used to create a Takagi-Sugeno based fuzzy inference model to generate the potential productivity as an output.

Fuzzy rule base and inference system

Fuzzy rule could be generated with a machine learning based on the relationship between input and output variables (Vallentin et al., 2020). In this study, ANFIS based fuzzy rules were prepared with [IF … THEN] format. The fuzzy rule's antecedent (input) had multiple conditions combined using "AND" operators and conditions were characterized with linguistic variables. The firing strength was evaluated with the use of minimum function (i.e., AND operation Eq. (2)). The minimum of the coupled membership degree was selected, embodying the result of AND operation, as shown in Figure 1.6. The AND operation for an example firing strength is expressed as minimum function:

$$Firing\ Strength\text{: }\min\{\mu_{High}(EC_a), \mu_{Medium}(NDRE), \dots, \mu_{An}(x_n)\} \quad (2)$$

Where:

μHigh (ECa): membership value of ECa under low fuzzy set,

μMedium (NDRE): membership value of NDRE under high fuzzy set,

μAn (xn): membership value of the input xn in the fuzzy set An.

Normalization of firing strength ($\bar{\mu}_n$) were performed for each fuzzy rule, as illustrated in Figure 1.6. On the other hand, the consequent (output) functions were generated with the adjusted coefficients (p_i, q_i and r_i as discussed in Sec 1.3.8.3). For instance, $f_1 = p_1 \times EC_a + q_1 \times NDRE + r_1$ is consequent (output) function for rule 1. This type of representing input-consequent inference is known as Takagi-Sugeno fuzzy inference system.

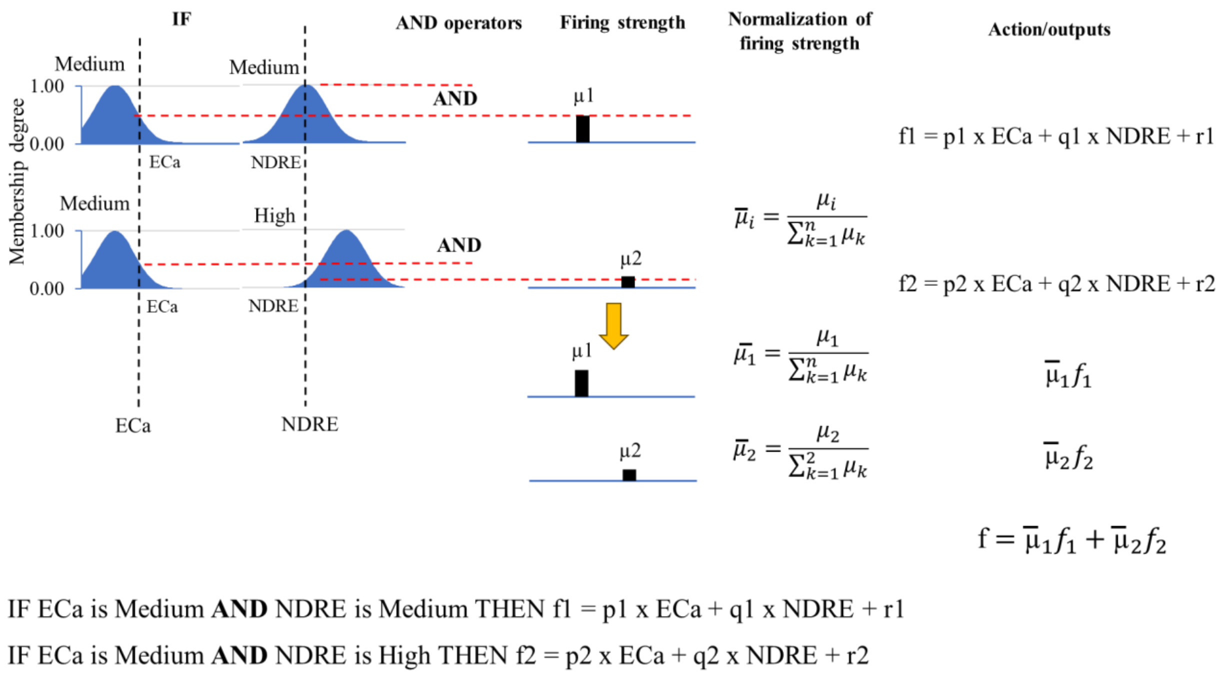


Figure 1.6. A schematic example of Takagi-Sugeno fuzzy inference system. Only two rules and two input variables (ECa and NDRE) are presented for representation.

Defuzzification

Post the fuzzy-rule application, WTAVR (Weighted Technique for Attaining Values Represented) method was used to convert the fuzzy values into a crisp value. The final crisp value was calculated as the sum of the normalized firing strength and the consequent (output) function shown in Eq. (3).

$$f = \bar{\mu}_1 f_1 + \bar{\mu}_2 f_2 + \cdots + \bar{\mu}_n f_n \qquad (3)$$

1.3.9 Developing Hybrid-Fuzzy Inference System (H-FIS)

ANFIS tuned MFs were used as an input to develop H-FIS model. Tuned MFs (gaussian bell-shaped membership function) from ANFIS converted values of soil and crop variables into fuzzy values. Unlike ANFIS fuzzy rules (data-driven fuzzy rules), H-FIS model required manually crafted fuzzy rules. Therefore, data and expert knowledge combination based fuzzy rules were prepared for each study site-year. Fuzzy rules were expected to capture complex relationships between input variables and grain yield. The input variables (topographic, soil, and crop variables) were inferred and color coded to low, medium, and high category to reflect the corresponding sorted normalized yield values as an output. A study by Bökle et al. (2023) grouped multiple input variables into yield, soil, and remote sensing to reduce the number of fuzzy rules and weight of each contribution while using corresponding variables. Rule generations were performed by listing all possible combinations of input and outputs with [IF … THEN] format. These combinations were inspected and refined with expert knowledge. A visual inspection and comparison between input variables and yield spatial maps were used as a methodological tool to improve the fuzzy rules when rules contradict the known logic. The H-FIS model executed with a scripted Matlab code using "evalfis" function and compute fuzzification output as a result. The Takagi-Sugeno fuzzy inference system was employed to compute the outputs. This stage was critical for understanding the model's ability beyond the training dataset, avoiding overfitting, and thereby ensuring the applicability of the FIS model.

1.3.10 Delineation of SSMG using H-FIS

The fuzzy outputs were point data spaced one meter, which later transformed to a grid of one meter. The range of the fuzzy output lies between 0 and 1, denoting 0 to be potentially a low productivity area and 1 to be a high productivity. The SSMG were delineated based on the need to have a unit level decision capability that could be practical and relatable with the available field machine for crop management. Agricultural machineries take grid based classified information than a continuous map, as a result it is crucial to develop a classification map (Vallentin et al., 2020). On the other hand, zone-based approach to delineate management units aggregates the contribution of micro-scale variability with a potential loss of information. Therefore, the fuzzy outputs were classified into 10 quantile-based interval classes of grid-based site-specific management maps using ArcGIS Pro 3.1.1 software. A reclassification of the 10 class was made indicating 10 (dark green) as a potentially high management grid and 1 (dark red) as a potentially low management grid.

1.3.11 Weighted Multivariate Clustering (WMC)

In addition to the management unit delineation using H-FIS, another conventional approach, Weighted Multivariate Clustering (WMC), was employed, and a subsequent comparative evaluation between the two methods was conducted. The multiplication of Getis-Ord Gi* z-score values and weight factor determined from GBRT relative influence analysis resulted weighted Getis-Ord Gi* z-score values. The weighted values used as input to generate cluster zones for a potential delineation of SSMZ using k-means clustering. K-means clustering method is a centroid-based clustering algorithm that calculates distances among input data points and iteratively identifies their k centroids. This method was used to partition dataset into k different clusters. Determining optimal number of clustering classes for k-means was performed with Calinski-Harabasz (C-H) index in Matlab environment. The C-H index provides a measure of the cluster's compactness and separation, where higher values generally indicate better-defined clusters (Peeters et al., 2015). This index was used to determine an optimal number of clusters (k) in an iterative approach by varying k values between 3 to 6. For each potential k value, dataset was clustered, and then cluster validity was evaluated using C-H index as,

$$\mathbf{CH}\ (\mathbf{k}) = (\mathbf{B}(\mathbf{k})/(\mathbf{k}-\mathbf{1}))/(\mathbf{W}(\mathbf{k})/(\mathbf{N}-\mathbf{k})) \qquad (\mathbf{4})$$

Where:

B(k) is the inter-cluster sum of squares,

W(k) is the within-cluster sum of squares, N is the number of data points.

The k values against corresponding C-H index values were plotted with Matlab, and the k values corresponding to highest CH value was selected as optimal number of clusters. At final stage, the clustered zones were characterized with two years of yield map to assign high, medium, and low management zones.

1.3.12 Evaluation and Validation of Management Units

In the process of management unit delineation, it is crucial to ensure the consistency of clusters and evaluate inter-region variability. This study assessed the spatial distribution of site-specific management units (SSMUs) developed with WMC, and H-FIS methods. The efficacy of the SSMUs methods were evaluated using the Moran's I index and the Moran's Z-score (Ohana-Levi et al., 2019). Moran's, I index, and Z-score are widely used statistical methods to assess the spatial auto-correlation. R programming language with libraries, "clusterSim", "spdep", and "sp' were used to perform evaluation analysis.

Statistical analysis to validate SSMUs was conducted with R programming language. ANOVA and a Tukey's Honest Significance Difference (HSD) test were conducted for a combination of soil properties (Organic Matter (%), pH, and CEC) from lab analysis) and SSMUs to examine the statistical difference among units.

1.4 Results and Discussion

1.4.1 Influence of micro-scale and macro-scale variables on Spatial Pattern of Grain Yield

1.4.1.1 Hot-spot Analysis

The hot-spot analysis of each variable at each measured location was performed to access the strength and direction of association with neighboring points. The hot- and cold-spot maps presented in Figure 1.7 and Figure 1.8 were generated with Getis-Ord Gi* methodology to estimate the degree of auto-correlation for topography, soil, crop and yield variables. Based on the Getis-Ord Gi* z-score values, each measurement point with red-blue shades corresponds with specific confidence levels of 90%, 95%, and 99%. Areas with deep blue dot shades show a statistically significant (Cold spot with 99% confidence) spatial clusters of low values. Similarly, deep red dot shades indicate a statistically significant (Hot spot with 99% confidence) spatial cluster of values, while the white point shades present non-significant spatial clustering.

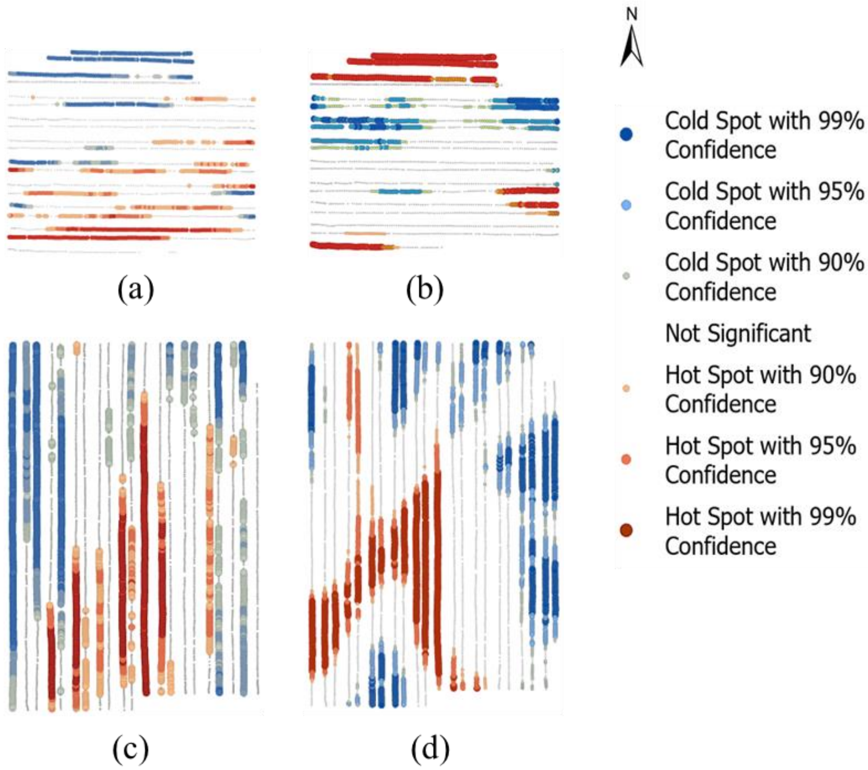


Figure 1.7. Getis-Ord Gi* hot-spot map for topography variables at two sites, (a) elevation at) Kansas Research Valley Topeka experiment field (Site year-1 and Site year-2); (b) slope at) Kansas Research Valley Topeka experiment field (Site year-1 and Site year-2) ; (c) elevation at Kansas Research Valley Rossville experiment field (Site year-3 and Site year-4); and (d) slope at Kansas Research Valley Rossville experiment field (Site year-3 and Site year-4) .

The hot- and cold-spot analysis of altitude and slope shows the presence of a distinct spatial pattern in field features. The North and central part of site year-1, site year-2, site year-3 and site year-4 exhibit a spatial clustering of high slope values which indicates steep inclines and declines (Figure 1.7 b and d). The hot spot with 99% of confidence of elevation could be interpreted as the presence of a raised part of the field, which is the south part of all study site years (Figure 1.7 a and c).

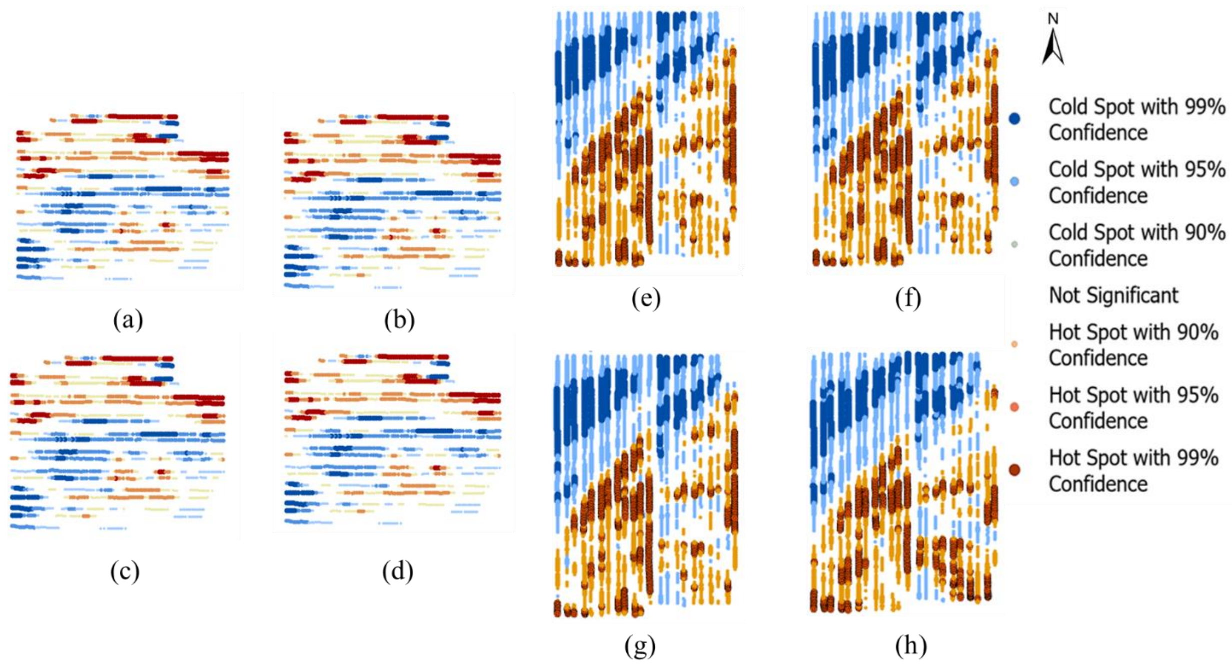


Figure 1.8. Getis-Ord Gi* hot-spot map of other soil properties, (a) Organic matter (OM), (b) Cation Exchange Capacity (CEC), (c) Apparent Electrical Conductivity at shallow (ECa shallow), (d) Apparent Electrical Conductivity at deep (ECa deep) for Kansas Research Valley Topeka experiment field (Site year-1 and Site year-2). (e) Organic Matter (OM), (f) Cation Exchange Capacity (CEC), (g) Apparent Electrical Conductivity at shallow (ECa shallow), and (h) Apparent Electrical Conductivity at deep (ECa deep) for Kansas Research Valley Rossville experiment field (Site year-3 and Site year-4).

The Getis-Ord Gi* hot-spot analysis across various parameters, including OM, CEC, and ECa at both shallow and deep depths, reveals distinct spatial patterns across all study site years. The Getis-Ord Gi* hot-spot analysis (Figure 1.8 a and e) indicates that the central and southern part of the study sites, respectively, exhibits a clustering of low organic matter content values, shown by the deep blue dot shades, with a 99% confidence. For CEC, there was a notable similarity in the dispersion of clustering of low CEC and clustering of low organic matter content values (Figure 1.8 b and f). In site year-3 and site year-4 (Figure 1.8 f), higher CEC values concentrate on the northwest part. The OM, CEC, and ECa Getis-Ord Gi* hot-spot analysis at shallow and deep depth for site year-3 and site year-4 (Figure 1.8 e, f, g, and h) demonstrates a change from a significant clustering of high values to low values in a diagonal gradient from northwest to southeast. The diagonal gradient (from high-low value) from northwest to southeast for site year-3 and site year-4 align with the USDA NRCS soil classification (see Sec. 1.3.1) across the study site.

The Getis-Ord Gi* hot-spot of topography and soil proved macro-scale variability in soil properties across the fields. The spatial patterns indicate potential of soil properties in influencing delineation of SSMGs. Similarly, Getis-Ord Gi* hot-spot values of crop variables (vegetation indices - NDVI, SAVI, and NDRE) and grain yield were generated. The values were extracted with similar point vector locations as soil and crop variables (Veris MSP3 – ECa, CEC, OM, Elevation and Slope). The Getis-Ord Gi* z-score values were used in determining their relative influence on the spatial variability of yield. However, high resolution crop variables were not used to estimate Getis-Ord Gi* z-score values. The main reason was the potential mis-representation of hot and cold spots due to high sensitivity of scale of analysis and no need in practicality of identifying hot and cold spots from raster images.

1.4.1.2 Soil and Crop Variables Selection and Weight Factor Estimation

Pre-processing of soil and crop variables for multicollinearity were conducted. The Pearson correlation coefficient reveled the respective values of soil and crop variables as reported in Figure 1.9 a and b, respectively. The Pearson correlation coefficient of soil variables shows a multicollinearity for all soil properties. Therefore, the PLSR model was used to assess the relative influence of soil variables on spatial variability of yield. Soil properties collected with Veris MSP3 showed a high multicollinearity, as a result PLSR was recommended in predicting yield (Gebbers et al., 2011). Conversely, partial multicollinearity of crop variables reported as shown in Figure 1.9 b. Therefore, crop variables with Pearson correlation coefficient above 0.85 were removed to avoid multicollinearity, as suggested by Farrar and Glauber (1967). GBRT based relative weight of multivariable on yield resulted in a weighted multivariate spatial clustering with high autocorrelation (Ohana-Levi et al., 2019). Therefore, relative influence of crop variables on spatial variability of yield were estimated with GBRT model.

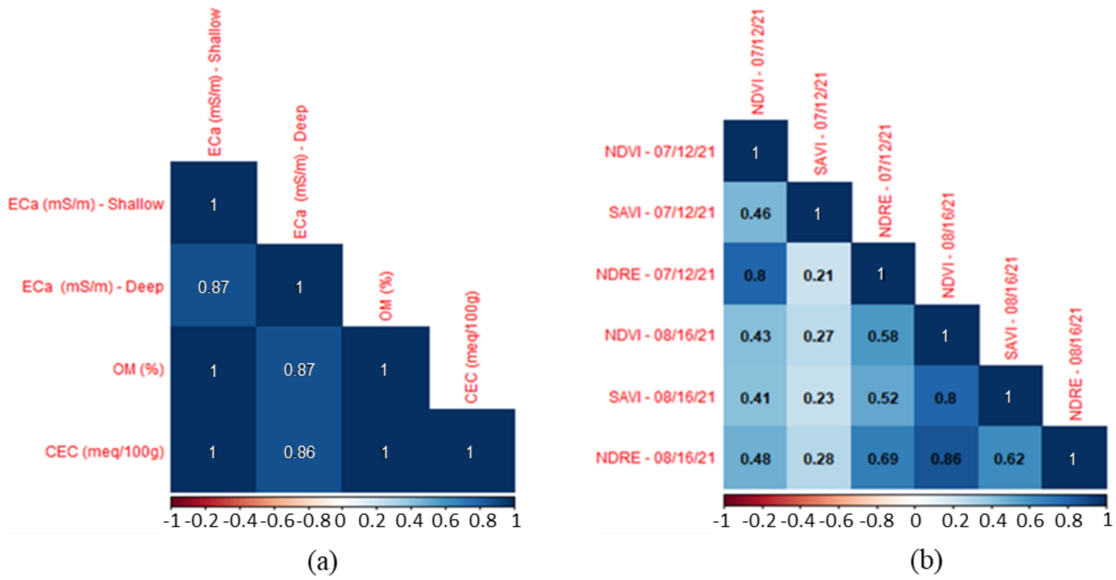


Figure 1.9. Correlation matrix of (a) soil, and (b) crop variables for site year-1.

The relative importance of soil variables on spatial variability of grain yield was investigated using PLSR model over four site-years. For site year-1 and 2, OM consistently showed a higher influence regardless of the crop type cultivated with value of 32.54% and 33.20%, for site year-1 and 2, respectively (Figure 1.10 a). At site year-3 and 4, ECa Shallow emerged as the most influential soil property in affecting spatial variability of corn and soybean yield. While corn cultivation in site year-1 and 4 showed OM and ECa Shallow influence on spatial variability yield, it was not the same for soybean in site year-2 and 3.

The relative influence of crop variables on spatial variability of gran yield was determined by GBRT. In 3 out of 4 years, NDRE variable stood out as the primary influencer with a relative influence ranging from 33.86% to 76.05%. For site year-3, "NDVI - 08/16/2021" led with a 48.58% influence, closely followed by "NDRE - 08/16/2021" at 33.86%. In site year-4, "NDRE -

08/10/2022" at 35.51% and "NDRE - 07/19/2022" with 29.06% were the top two variables with highest influence on spatial variability of grain yield. The top ranked out variables, NDRE and NDVI were selected to be used as an input variable to delineate SSMGs with weighted multivariate clustering and fuzzy logic system. The relative influences of each variable were used as weight factor to generate weighted Getis-Ord Gi* z-score values.

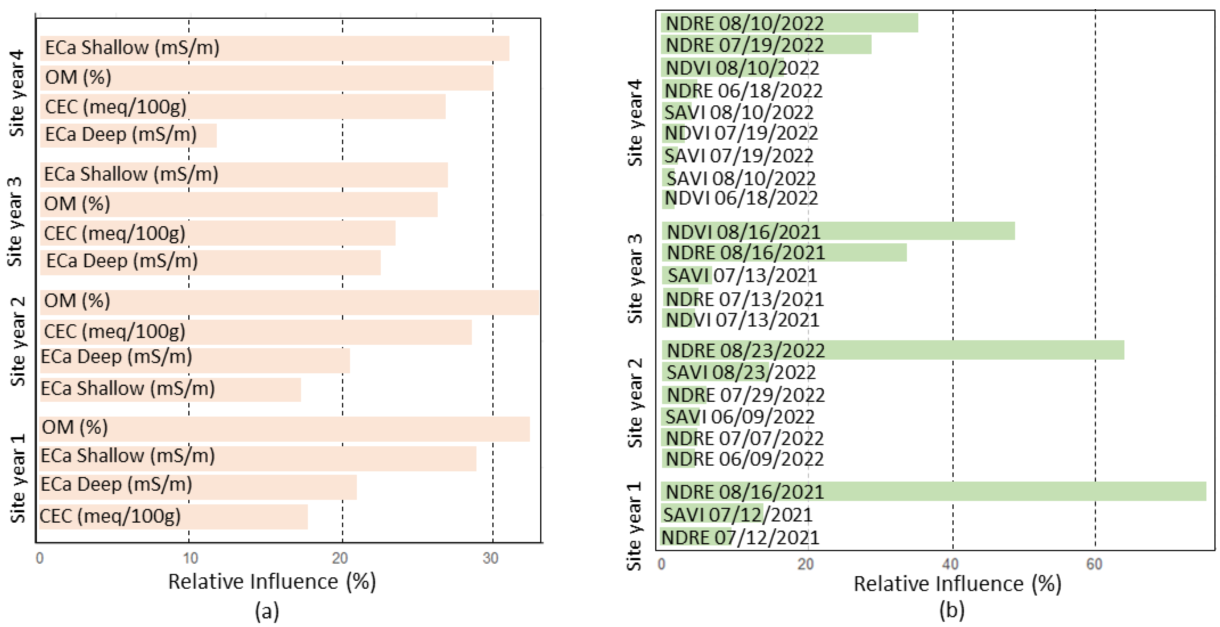


Figure 1.10. Relative influences of different variables on spatial variability of yield for four site years, (a) soil variables based on partial least square regression (PLSR) model; and (b) crop variables based on gradient boosted regression trees (GBRT) model.

The GBRT models consistently identified NDRE and NDVI as the top variables. Despite signal saturation concerns with increasing crop leaf area index (Gitelson et al., 2003), NDVI has frequently been utilized to correlate with grain yield (Johnson, 2014; Martin et al., 2007). However, Paiao et al., (2021) reported a lower predictive power of NDVI for grain yield estimation at various crop development stages compared to NDRE. Similarly, Torino et al., (2014) observed that NDVI measurements were less accurate in predicting grain yield than NDRE, especially in the later stages of vegetative development. These findings align with previous studies that analyzed

VI to yield correlation as the season progresses. Oglesby et al., (2022) comparing data acquisition across different growth stages of corn with a MicaSense sensor, reported that relative NDRE is the most suitable stage for grain yield prediction within vegetative stages between VT-R1 stage, with an R^2 of 0.78 and RMSE of 1.07 t/ha.

1.4.2 Fuzzification and Accuracy of ANFIS Estimates

Fuzzification of input variables was performed with tuned membership functions. Tuned membership function was achieved with ANFIS model training, and the respective high yield prediction performance (high R^2) was used to ascertain tuned membership function. The ANFIS model was trained for four site years. Tuned membership function result based on the selected soil and crop variable (input variables) and corn yield as an output for site year-1 is presented in Figure 1.11.

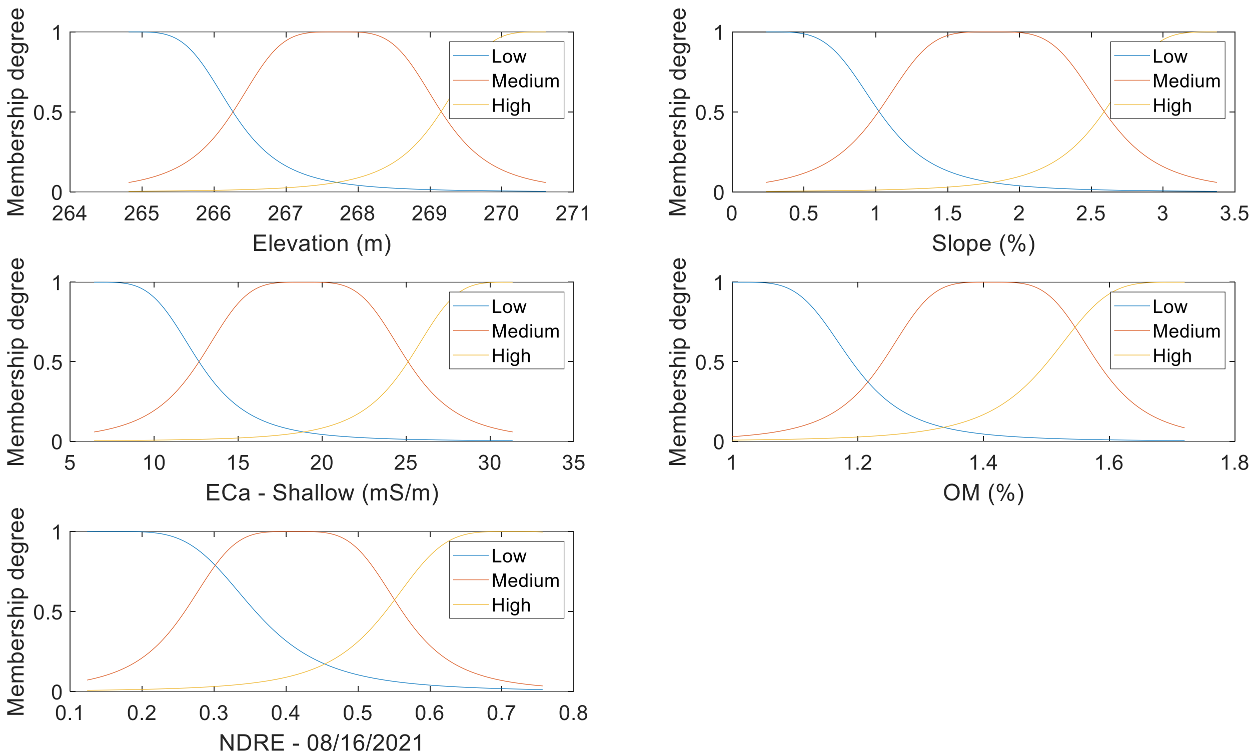


Figure 1.11. Membership function of soil and crop variables for site year-1. Gaussian bell-shaped MF's parameters were tuned using Adaptive Neuro-Fuzzy Inference System (ANFIS).

Input variables, elevation, slope, ECa at shallow, and OM parameters represent macro-scale variability in soil for site year-1. Micro-scale variability in crops were quantified with selected variable of NDRE from 08/16/2021. Tuned membership functions were achieved on 70th epochs where the model reached minimum training RMSE of 1.65 t/ha and a validation RMSE of 1.20 t/ha. Error metrics results suggested that the model was a good fit with minimal deviations. Figure 1.12 shows the performance of the ANFIS model for site year-1. The model's performance was quantified by the coefficient of determination, R^2, which stands at 0.80. Tuning membership function with yield prediction performance using ANFIS resulted in 82% accuracy as reported by (Qaddoum et al., 2011). Similar results with ANFIS were also obtained for other site years. A detailed results and analysis for other site years are presented in Appendix I - Membership functions and ANFIS output.

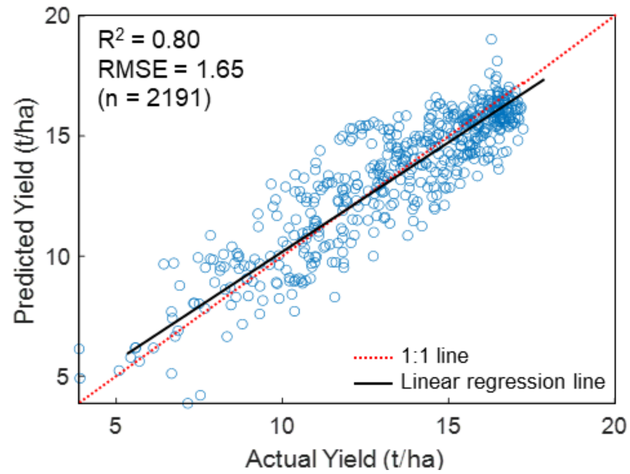


Figure 1.12. Scatter plot of actual vs. Adaptive Neuro-Fuzzy Inference System (ANFIS) predicted corn yield for site year-1.

Table 1.2. Adaptive Neuro-Fuzzy Inference System (ANFIS) training accuracy for four site years.

Year	**Study site**	**Crop**	**Selected input Variables**	**Training Accuracy**	
				RMSE (t/ha)	**R^2**
2021	Site year-1	Corn	Elevation, slope, ECa shallow, OM, NDRE 08/16/2021	1.65	0.80
	Site year-2	Soybeans	Elevation, slope, ECa shallow, OM, NDVI 08/16/2021 NDRE 08/16/2021	0.60	0.74
2022	Site year-3	Soybeans	Elevation, slope, ECa shallow, OM, CEC, NDRE 08/23/2022	0.29	0.77
	Site year-4	Corn	Elevation, slope, ECa shallow, OM, NDRE 07/19/2022 NDRE 08/10/2022	1.41	0.77

1.4.3 Delineation of SSMGs using H-FIS

H-FIS based SSMGs using coupled soil and crop input variables are shown in Figure 1.13 a and b, for site year-1 and 2. SSMG classes show grid productivity scales, denoting 10 as high productive grid and 1 as low productive grid. For site year-1, a noticeable low productivity grids occupied the centre of field encapsulated by a medium productivity grids and high productivity grids bounding the medium productivity grids (Figure 1.13 a). On the contrary, for soybean cultivated site year-2, a high-productivity grids seem more dominant compared to site year-1 SSMG (Figure 1.13 b) and the low-productivity grids appears to have shrunk in number of grids and divided into smaller regions.

Macro-scale variability in soil was spatially quantified with only OM for site year-1, whereas OM and CEC were used for site year-2. However, the difference in addition of CEC variable did not bring any changes to spatial pattern of management grid (Figure 1.13 c and Figure 1.13 d). Such an observation suggests that differences in inclusion of NDRE of corn and soybean played an influential role in SSMGs delineation. Therefore, NDRE of corn and soybean were a potential variable for the difference in spatial pattern of SSMGs for site year-1 and site year-2, as shown in Figure 1.13 a and b. Visually, coupled macro-scale and micro-scale variability based SSMGs presented both macro-scale and micro-scale variability in the delineated SSMGs (Figure 1.13 a and b), whereas soil based SSMGs only captured macro-scale variability (Figure 1.13 c and d).

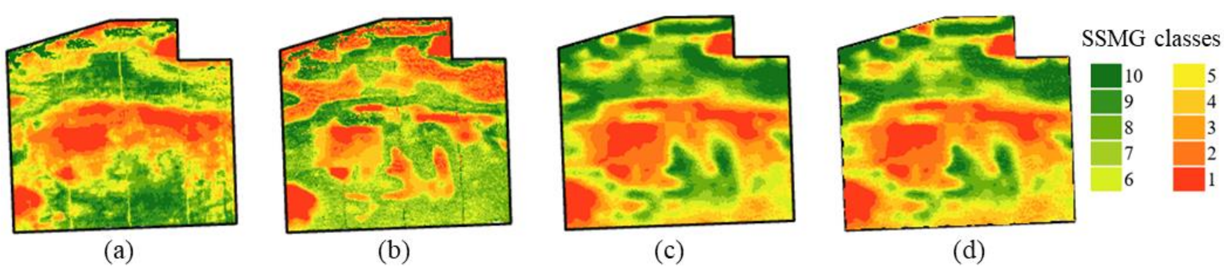


Figure 1.13. Hybrid-Fuzzy Inference System (H-FIS) based site-specific management grids with coupling of macro-scale spatial variability in soil and micro-scale variability in crop for (a) site year-1, (b) site year-2; and site-specific management grids with soil alone for (c) site year-1, (d) site year-2.

Grain yield for corn and soybean were presented in Figure 1.14 a and b, respectively. As visually observed, there was a clear pattern in area association of crop grain yield and SSMGs maps. Figure 1.14 a show that corn yield for site year-1 dominated with high grain yield, spread with smaller patches of medium and low corn yield at the center. In contrast, the yield for soybeans in site year 2 displays a more pronounced variability of high, medium, and low soybean grain yield.

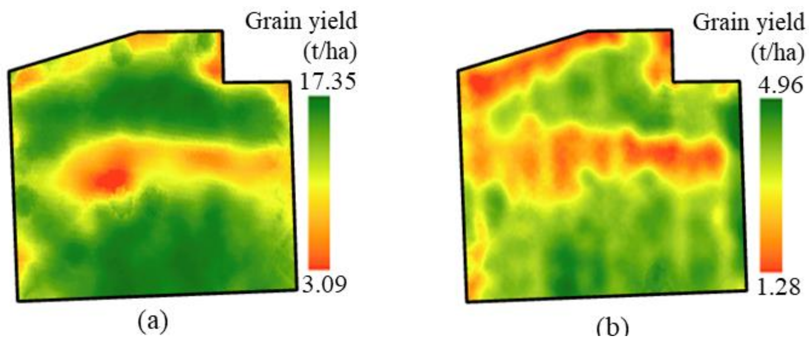


Figure 1.14. Crop grain yield data (a) corn for 2021, (b) soybean for year 2022 at Kansas Research Valley Topeka experiment field (Site year-1 and Site year-2).

Unlike the soil fertility data-based management unit delineation (Ortega and Santibáñez, 2007; Shashikumar et al., 2023), utilizing in-season crop growth variables (vegetation indices) acquired remotely produced enhanced accuracy and reliability (Gallardo-Romero et al., 2023). However, the utilization of a singular data source carries inherent risks, particularly in scenarios where the information density required for precise interpretation is low (Wu et al., 2002). Consequently, the integration of data fusion methods stands as a valuable augmentation to the array of techniques for delineating management units. The effectiveness of H-FIS in accurately mapping human thought processes and argumentations significantly advances the process of micro- and macro-scale variability data fusion (multi-scale data fusion). The rule-based inference engine of H-FIS allowed different types of data to be related to each other and their information content is enhanced.

SSMGs maps for site year-3 and site year-4 illustrated distinct high and medium productivity potential grids at northwest part of the field while low productivity grids dominated east and southwest part of field (Figure 1.15 a-b). The noticeable medium-productivity grid at central west part of field seems to have contracted considerably in site year-4 (Figure 1.15 b). Under macro-scale variability in soil based SSMGs (Figure 1.15 c and d), the vast majority of low productivity grids dominated west and south boundary of the field. On the other hand, coupled macro-scale and micro-scale variability based SSMGs added a mix of high and medium productivity grids to the same part of the field. The shift illustrates that micro-scale variability in crops influences spatial pattern and subsequent demarcation of productivity grids. Therefore, it is possible to suggest that areas considering low productivity potential delineated with solely macro-scale variability in soil could result in a different field nutrient management strategy. The significant finding of this study is that an integration of macro- and micro-scale variability for delineation of SSMGs, increased number of high and medium productivity grids for all site years.

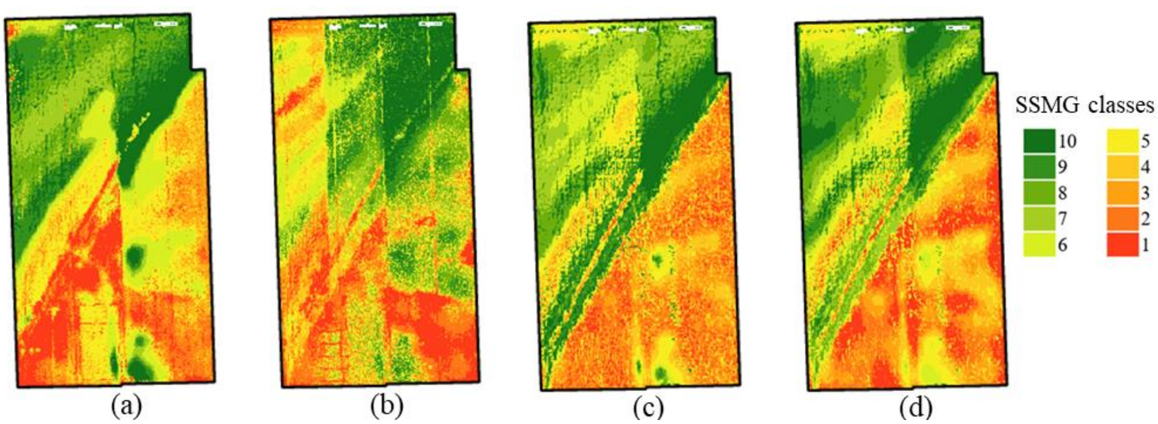


Figure 1.15. Hybrid-Fuzzy Inference System (H-FIS) based site-specific management grid with coupling of macro-scale spatial variability in soil and micro-scale variability in crop for (a) site year-3, (b) site year-4; and site-specific management grid with soil alone for (c) site year-3, (d) site year-4.

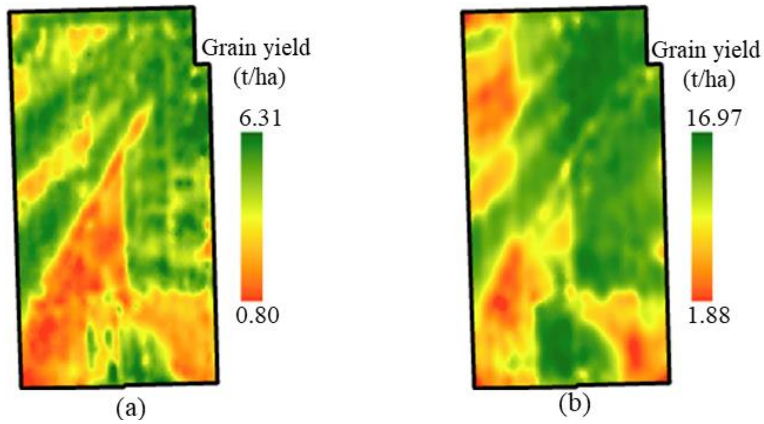


Figure 1.16. Crop grain yield data (a) soybean, (b) corn at Kansas Research Valley Rossville experiment field (Site year-3 and Site year-4).

The visual observation of H-FIS based SSMGs were supported by quantitative assessment with kappa coefficient (κ) based on the agreement between the SSMGs. Quantitative comparison between coupled variables based SSMGs and macro-scale variability based SSMGs, for both site year-1 and site year-2 reported a kappa coefficient of 0.16 and -0.013, demonstrating the dissimilarities between SSMGs. Similarly, site year-3 and site year-4 SSMGs comparison resulted a kappa coefficient value of 0.09 and 0.08, respectively. The kappa coefficient explicitly shows that there was no agreement between SSMGs maps, indicating a dissimilarity created based on the data sources used for delineating the management grids. A study by Bökle et al. (2023) reported dissimilarities among management units between those who included and omitted remote sensing information for class based prescription maps using FIS system. This study suggested that management grids dissimilarities occurred mainly due to the difference in data sources used to delineate the SSMGs. Therefore, integrating micro-scale variability on the macro-scale variability to delineate a management grid resulted a complete different spatial pattern. Consequently, the expected nitrogen management strategy would change due to changes of grids demarcation pattern and increase in number of grids with high and medium potential productivity. Therefore, delineating a SSMGs by integrating micro-scale variability in crop with macro-scale variability in soil would have an impact on potential change in quantity of nitrogen application. Further studies regarding the impact of SSMGs on the increase of nitrogen use due to the introduction of new medium and high management grids would be necessary.

1.4.4 Weighted Multivariate Clustering (WMC)

SSMZ maps were generated with WMC technique, as illustrated in Figure 1.17 and Figure 1.18. Evaluation of WMC zones solely based macro-scale variability in soil (Figure 1.17 and Figure 1.18) shows that there was a resemblance in their zoning pattern. For site year-3 and 4, there was a resemblance between WMCs delineated with coupled macro-scale and micro-scale variability (Figure 1.18 a and b). Coupled macro-scale and micro-scale variability based WMC zones of site year-1 (Figure 1.17 a) were dominated by zone 1 and 3, in contrast to solely macro-scale variability based WMC (Figure 1.17 c), where it showed a more heterogenous spread of all three zones.

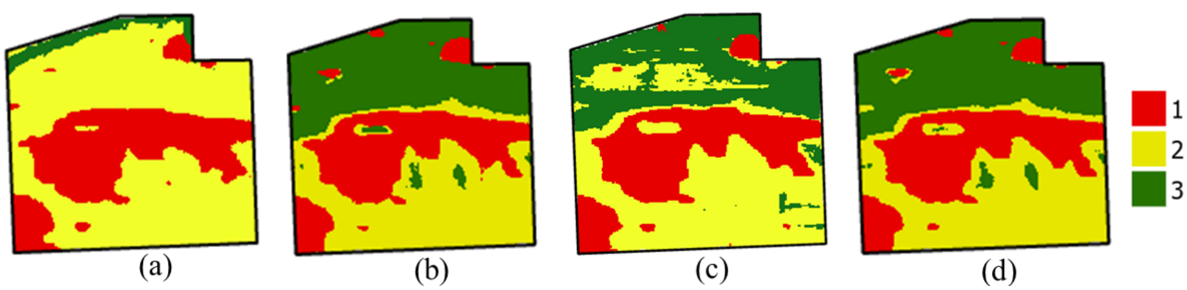


Figure 1.17. Weighted Multivariate Clustering (WMC) based site-specific management zones with coupling of macro-scale spatial variability in soil and micro-scale variability in crop for (a) site year-1, (b) site year-2; and site-specific management zones with macro-scale variability in soil alone for (c) site year-1, (d) site year-2.

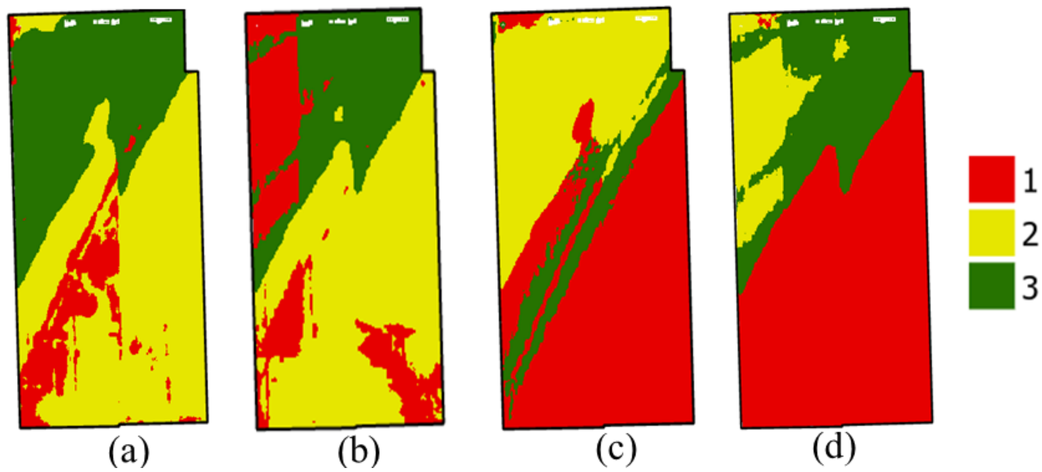


Figure 1.18. Weighted Multivariate Clustering (WMC) based site-specific management zones with coupling of macro-scale spatial variability in soil and micro-scale variability in crop for (a) site year-3, (b) site year-4; and site-specific management zones with macro-scale variability in soil alone for (c) site year-3, (d) site year-4.

1.4.5 Evaluation, Comparison, and Validation of SSMUs

1.4.5.1 Evaluation of SSMGs Delineated with H-FIS

SSMGs of ten classes denoting 0 as low productivity grids and 10 as high productivity grids, as shown in Figure 1.13 a and b. SSMGs were evaluated for spatial distribution and compactness of management grids produced.

Table 1.3. Evaluation of Hybrid-Fuzzy Inference System (H-FIS) based site specific management grids delineated with coupling of macro-scale spatial variability in soil and micro-scale variability in crop and macro-scale spatial variability in soil.

SSMGs	Site year	Moran's I index	Moran's I Z – Score
SSMGs using soil + crop	1	0.843	465.145
	2	0.716	395.858
	3	0.877	791.215
	4	0.767	692.025
SSMGs using soil	1	0.909	501.744
	2	0.904	499.626
	3	0.863	778.58
	4	0.862	777.466

As shown in Table 1.3, Moran's I index values of SSMGs shows a strong positive spatial correlation, across all site years. However, the Moran's I index of SSMGs with coupled macro and micro-scale variability showed a relatively lower Moran's I index as compared to solely macro scale variability based SSMGs. Relatively smaller Moran's I index shows a weaker spatial pattern due to introduction of variables to capture micro-scale variability in crops. H-FIS showed that the micro-scale variability in crop reflected a relatively lower Moran's I index. Lower Moran's I index implies macro-scale and micro-scale variability were accounted for with H-FIS delineated management grids.

Under H-FIS SSMG delineation, soil and crop variables were processed under a point data level. The point data were translated into grid unit, an output to map the SSMG. On the other hand, WMC clustering partitions the spatial data into distinct clusters based on the mean pixel values that lead to oversimplification. Therefore, it failed to capture micro-scale variability in crops. Moran's, I index value close to 1 indicates a strong positive spatial autocorrelation while index value close to 0 shows a random spatial pattern. The lower Moran's I index value of H-FIS based SSMG using soil plus crop corresponds to a more random spatial pattern as shown in Figure 1.19 a and b. H-FIS based SSMG using soil plus crop shows a micro-scale spatial pattern (Figure 1.19 a and b) as compared to a more distinct larger zones (Figure 1.19 c and d). Therefore, it is possible to conclude that H-FIS using soil and crop-based management grid delineation approach characterized micro-scale variability of the field while WMC units failed to capture.

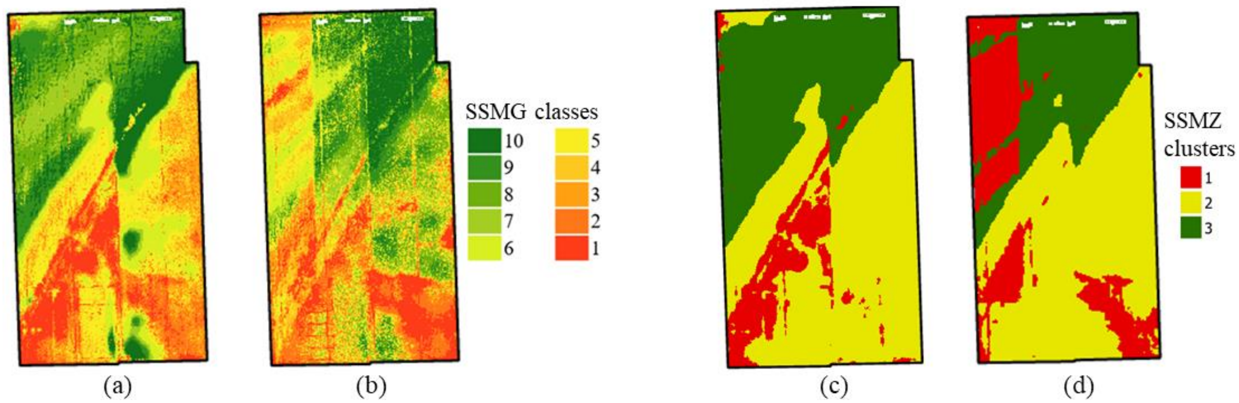


Figure 1.19. Coupled macro-scale variability in soil and micro-scale variability in crop based Weighted Multivariate Clustering (WMC) using Hybrid-Fuzzy Inference System (H-FIS) method (H-FIS based SSMGs using soil + crop) for (a) site year-3, (b) site year-4; macro-scale variability in soil based WMC using k-means method (k-means based WMC using soil + crop) for (c) site year-3, (d) site year-4.

1.4.5.2 Evaluation of SSMUs Delineated with WMC

Evaluation of SSMUs delineated with WMC method was illustrated in Table 1.4 . SSMUs with coupled soil and crop variables and solely soil variables were compared with quantified degree of autocorrelation. Moran's I index showed a highly significant spatial auto-correlation for both SSMUs across four study site years (Table 1.4).

Table 1.4. Evaluation of Weighted Multivariate Clustering (WMC) based SSMUs delineated with coupling of macro-scale spatial variability in soil and micro-scale variability in crop (Site-Specific Management Units (SSMUs) using soil + crop) and macro-scale spatial variability in soil (Site-Specific Management Units (SSMUs) using soil).

WMC SSMZs	Site year	Moran's I index	Moran's I Z-Score
WMC SSMUs using soil + crop	1	0.942	301.465
	2	0.926	296.544
	3	0.889	530.723
	4	0.936	558.866
WMC SSMUs using soil	1	0.948	303.117
	2	0.963	308.486
	3	0.972	580.620
	4	0.936	558.589

The integration of macro-scale variability in soil and micro-scale variability in crop (soil and crop variables) to delineate SSMUs were expected to result relatively lower Moran's I index due to the nature of a more heterogenous spatial pattern because of the micro-scale variability in crop. However, as shown in Table 1.4, SSMUs exhibited a comparable average Moran's I index value of 0.92 and 0.95, respectively. This result indicates that WMC method failed to account for micro-scale variability in crops.

1.4.5.3 Validation of H-FIS based SSMGs

For the purpose of validation with measured soil data, the H-FIS based SSMG were aggregated into three classes as shown in Figure 1.20. It is known that slope, elevation, ECa, OM, CEC, NDVI, and NDRE are well established data sources for delineation of management units (Moral et al., 2023). For Site year-1 and site year-3, H-FIS used coupled soil and crop variables (elevation, slope, ECa shallow, OM, NDVI 08/16/2021and NDRE 08/16/2021) to delineate SSMG (Figure 1.20 a-d). Management grids were validated with measured soil properties, OM, pH, and CEC.

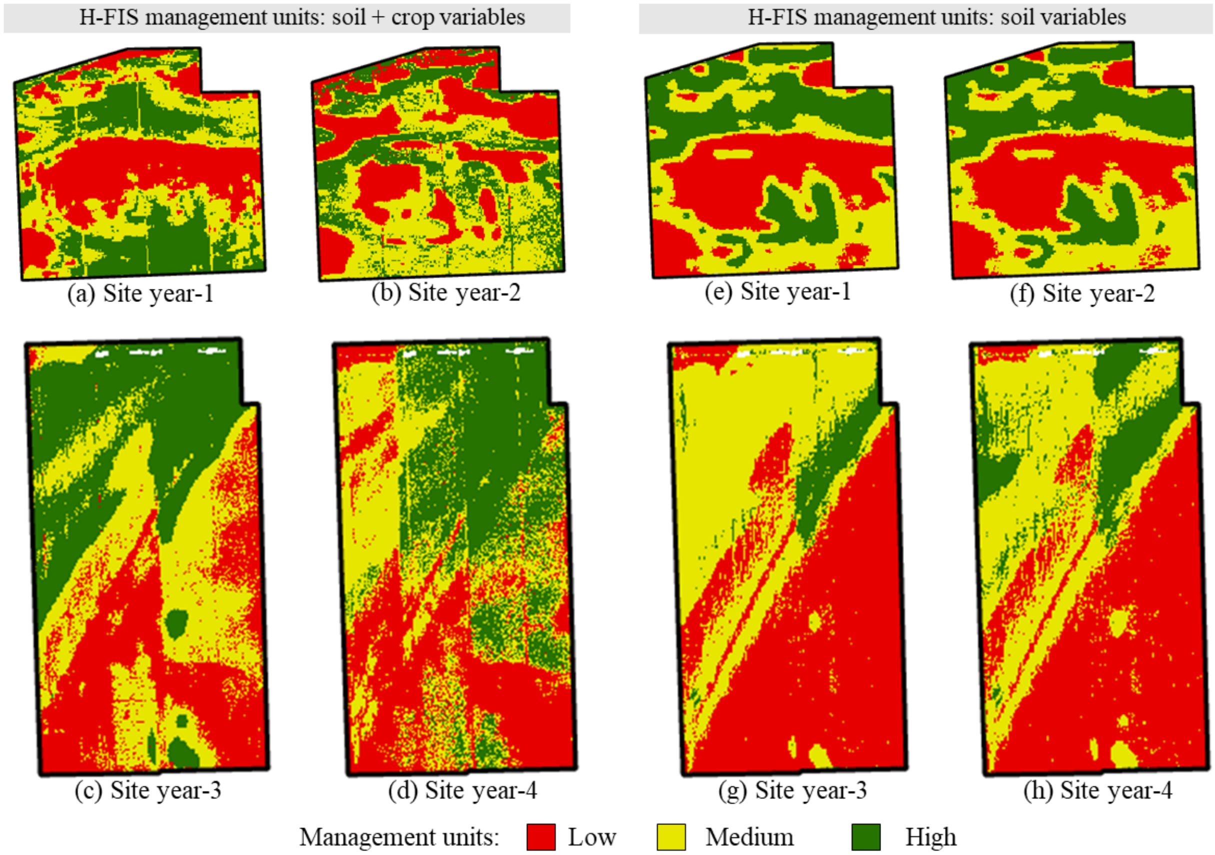


Figure 1.20.Aggregated management units derived from Hybrid-Fuzzy Inference System (H-FIS) Site-Specific Management Grids (SSMGs) for four site years based on (a-d) soil + crop variables, and (e-h) soil variables.

Table 1.5. Differences in mean values of soil Organic matter (OM%), pH and Cation Exchange Capacity (meq $100g^{-1}$) (soil samples collected) across different productivity zones to validate H-FIS) based SSMGs using soil + crop and soil. Statistical significance is presented with Tukey's HSD post hoc test (T-HSD) group symbology.

Site years	Management Unit	Soil + crop-based management units			Soil based management units		
		OM (%)	CEC (meq $100g^{-1}$)	pH	OM (%)	CEC (meq $100g^{-1}$)	pH
Site year-1	Low	1.07±0.21 a	7.75±1.27 a	6.65±0.29 a	1.25±0.31 a	9.36±2.75 a	6.68±0.28 a
	Medium	1.56±0.11 b	12.08±1.82 b	6.44±0.35 a	1.44±0.32 a	12.84±1.43 b	6.88±0.70 a
	High	1.44±0.36 ab	12.55±2.22 b	6.94±0.72 a	1.53±0.32 a	11.06±2.11 b	6.36±0.15 a
Site year-2	Low	1.22±0.17 a	9.38±2.15 a	6.80±0.59 a	1.15±0.10 a	8.42±1.13 a	6.53±0.12 a
	Medium	1.40±0.22 a	10.45±1.65 a	6.68±0.51 a	1.36±0.22 ab	9.44±1.31 a	6.56±0.19 ab
	High	1.27±0.17 a	8.83±1.18 a	6.56±0.15 a	1.38±0.19 b	11.18±1.74 b	6.68±0.80 b
Site year-3	Low	1.08±0.21 a	7.36±1.52 a	6.72±0.39 a	1.12±0.20 a	7.44±1.64 a	6.72±0.43 a
	Medium	1.24±0.38 a	8.45±2.97 a	6.83±0.35 a	1.50±0.39 b	10.04±2.73 b	6.87±0.41 a
	High	1.53±0.43 b	10.0±2.83 b	6.84±0.32 a	1.1±0.28 a	8.25±0.49 a	6.95±0.07 a
Site year-4	Low	1.23±0.36 a	9.12±2.65 a	6.69±0.46 a	1.17±0.20 a	8.12±1.34 a	6.35±0.35 a
	Medium	1.33±0.28 a	9.89±3.05 a	6.47±0.56 a	1.43±0.33 b	10.99±2.94 b	6.72±0.59 a
	High	1.31±0.16 a	9.28±1.87 a	6.32±0.35 a	1.6±0.32 ab	13.0±1.56 ab	6.3±0.51 a

For quantitative analysis, mean values of soil OM, pH and CEC (soil samples collected) across different aggregated management units were compared for statistical significance (see Table 1.5). Mean OM and CEC are found to be significantly different between low and high SSMGs for site year-1 and site year-3 ($p < 0.05$). High, medium, and low values of OM and CEC correspond to high, medium, and low productivity grids. This aligns with the hypothesis that macro macro-scale-scale spatial variability in soil and micro-scale-scale spatial variability in crop can be accounted for by delineating site-specific management unit. Soil plus crop variables-based H-FIS SSMGs accounted for multiple scales of variability while having OM and CEC significantly different for two site years (site year-1 and site year-3). Such a finding is supported by previous studies where higher values of OM and CEC have been reported in higher productivity zones (Miao et al., 2018; Moral et al., 2023).

A study by Denora et al., (2022) reported that OM was significantly different between defined management units with > 32% changes in value between units. At the same time, the significance difference of CEC between medium and high and low and medium were insignificant. Similar findings reported by (Inman et al., 2005), and suggested that medium productivity units may have small inclusions of both high and low productivity units with a potential of fuzzing the distinctions between the units. On the contrary, OM and CEC were not significantly different between management grid ($p < 0.05$). Except OM for site year-1, OM and CEC showed a significant difference between low and medium productivity units for all site years ($p < 0.05$). Soil-based H-FIS SSMGs (Figure 1.20 e-h) used variables that capture macro-scale variables in soil (ECa, OM, CEC). This result was expected because soil based SSMGs are delineated to characterize macro-scale variability in soil, which is proved in this validation. It would be fair to validate soil plus crop variables based SSMGs with a combined soil and crop variable ground truth data to acquire a direct validation.

1.5 Conclusions and Perspectives

This study introduces a data fusion method based on evidential reasoning that integration of micro and macro-scale variability in crop and soil would result in more accurate delineation of management units. The management units underwent spatial clustering, focusing on the scale of data characterized by data resolution. Using the newly proposed H-FIS, crop growth information (derived from vegetation indices using remote sensing imagery) and soil proximal data can be fused independently of their units and spatial resolution, leading to the generation of productivity potential maps. These productivity potential maps, representing continuous data, were subsequently utilized to create a management grid (classified map) for precision farming applications. This classification is important as it accounts micro-scale variability in the field, allowing for the implementation of precision farming measures. The use of a classified map is not only more interpretable than continuous data but also aligns well with the operational principles of the latest agricultural machines with variable rate applications, such as individual nozzle level control.

Tuned membership functions were ascertained with high performance of predicted yield for all site years. The H-FIS delineated site-specific management grid for all study site years. Soil and crop variables based H-FIS management grid characterized macro-scale variability in soil and micro-scale variability in crop. On the other hand, solely soil based H-FIS management grid failed to capture the micro-scale variability in crops. Scattering of management grids are highly dependent on vegetation indices, as a result there is a clear indication of changes of management grid as changes of crops cultivated. Additionally, H-FIS based site-specific management grid delineated productivity zones. On the contrary, weighted multivariate clustering based management zone did not characterize micro-scale variability and cluster nomenclature for productivity grid. It is apparent from H-FIS based management grid evaluation that, lower Moran's I index values showed an increases spatial variability within a field with more varied pattern compared to soil variable based H-FIS grids. A significant difference between low and high productivity zones of OM and CEC for soil and crop was observed for H-FIS management grids.

Expert knowledge based fuzzy rules employed to develop H-FIS model. This study suggested future related works should generate fuzzy rules through grouping of all variables into soil and crop categories to reduce the complexity and number of fuzzy rules to be crafted. At the same time, this approach could minimize the steps of developing FIS model by removing the PLSR and GBRT based soil and crop input variable selection stage.

The integration of micro and macro-scale data fusion using H-FIS involves meticulous tasks related to data formatting and expert knowledge integration. Despite its labor-intensive nature, this method is distinguished by its transparency and understandable fusion logic, in contrast to algorithms operating as black boxes. Notably, H-FIS possesses the ability to autonomously develop a standard ruleset based on soil or crop variability gleaned from yield data maps and the corresponding evidence sources. This adaptability renders it suitable for diverse fields. This method can be customized to suit individual agricultural fields and their yield-relevant characteristics. In conclusion, SSMGs successfully translated coupling of macro-scale variability in soil and micro-scale variability in crops to clustered zones that could potentially require similar management practices. Delineation of SSMGs is the initial stage in optimizing fertilizer usage and successful SSMGs guided fertilizer applications.

Chapter 2 - Hyperspectral Sensing to Estimate Soil Nitrogen and Reduce Soil Sampling Analysis

2.1 Abstract

Optimal crop management relies on accurate measurements of soil nitrogen and a comprehensive understanding of its spatial variability. However, spatially characterized information on soil nitrate-nitrogen (NO_3-N) is not always readily available during fertilizer application practices. The complex mobility of soil NO_3-N, coupled with the labor requirements and high costs linked to soil chemical analysis, underscores the necessity for an alternative method. An improved approach is warranted to achieve precise soil NO_3-N estimation, ensuring it remains cost-effective and straightforward. In this study, non-imaging hyperspectral sensor with *visible to short*-wave infrared spectrum were used to estimate soil NO_3-N content. The objectives of this study were: (1) to estimate surface soil NO_3-N using non-imaging hyperspectral sensing; and (2) to observe spatial pattern of reduced soil sampling data set in reproducing a comparable spatial variability of soil surface soil NO_3-N. Soil samples were analyzed for soil NO_3-N and spectral measurements were acquired with ASD FieldSpec3 spectroradiometer. Spectral preprocessing techniques were applied to refine the spectral data and remove noise. Effective spectral windows were determined such that spectral demonstrated sensitivity to changes in soil NO_3-N. The selected spectral windows: W1 (1460-1690 nm), W2 (1730-1780 nm), and W3 (1940-2150 nm) were used to estimate soil NO_3-N. Partial Least Squares Regression (PLSR) and Random Forest Regression (RFR) were used to estimate soil NO_3-N using selected spectral features and measured soil NO_3-N. For full and reduced soil sampling data set, RFR model resulted in soil NO_3-N estimation with smallest R^2 0.84 for site year-6, and highest R^2 values of 0.99 for site year-1 and site year-3. Estimation of NO_3-N with RFR model using reduced soil sampling data set resulted low RMSE

and MAE error compared with PLSR model. Estimated soil NO_3-N map were developed. Map of measured soil NO_3-N showed a similar spatial trend estimated with estimated soil NO_3-N maps. The findings of potential of hyperspectral data-based soil NO_3-N estimation models provide a cost-effective and practical technique to improve soil fertility management practices.

2.2 Introduction

Soil is a fundamental component in farming. It is a foundational growing medium for plants and a source of essential nutrients to crops (Barrett et al., 2016). To meet crops nutritional demand, fertilizer applications are made at planting and during the growing season to maintain optimum supply for nutrients to crops. To achieve a higher level of precision in effective crop management, growers require information about soil physical and chemical characteristics at every location of their fields. Compared to other macro and micronutrients, nitrogen (N) due to its complex behavior in soil, development of a reliable soil test for plant N availability has been difficult (Masclaux-Daubresse et al., 2010). In practice, soil nitrate-nitrogen (NO_3-N) is generally estimated and used along with crop grain yield goal and other nitrogen credits such as soil organic matter-OM%, previous leguminous crop, or manure application, for N fertilizer recommendations (Leikam et al., 2003; Magdoff, 1991; Morris et al., 2018). Moreover, soil samples acquired at a sampling density of one sample per ha (i.e., common commercial practice for field mapping) often fail to characterize the spatial variability of soil properties that occur at a much smaller scale (Lin et al., 2005; Mzuku et al., 2005; Longchamps et al., 2015).

The conventional soil sampling and laboratory-based routine chemical analysis to determine soil nitrogen are labor-intensive, expensive, and time-consuming (Janik et al., 1998). Furthermore, they typically necessitate the utilization of hazardous chemicals, such as concentrated acids and alkalis, which introduce potential safety hazards. Therefore, assessing soil nitrogen status needs an indirect technique that is capable of reducing soil sampling data set to make it less labor intensive, as well as cheaper, quicker, non-destructive, and chemical free. Many researchers have used multiple tools including electrochemical sensors and optical techniques (spectroscopy) to rapidly, inexpensively, and indirectly estimate soil nitrogen (Adamchuk et al., 2004). Electrochemical sensors, such as ion-selective field effect transistors and ion-selective electrodes have been used to estimate soil NO_3-N. On-the-go (mobile) soil nutrient sensors, such as nitrate ion-selective membrane approaches were effective in detecting soil nitrate and potassium nitrate within 10 to 30 mg NO_3-N kg^{-1} soil at a dilution ratio of 10:1, showing good selectivity over other ions (Kim et al., 2006). Laboratory experiments demonstrated the feasibility of using NO_3^- and K^+ ion-selective electrodes (ISEs) for measuring NO_3-N and K^+ in moist soil samples (Adamchuk et al., 2002). The nitrate electrodes showed stable readings quickly but faced challenges in electrode-soil contact quality and calibration drift.

On the other hand, based on electromagnetic wave interaction with soil physical and chemical properties, many researchers estimated soil NO_3-N with a spectrophotometer, spectroradiometer, X-ray fluorescence (XRF), and fluorescence. A study by Shibusawa, (2003) demonstrated that soil spectrophotometer accurately predicted soil NO_3-N content, achieving a high R^2 value of 0.94, but reported inconsistency in maintaining spatial correlation for NO_3-N across the field. A study by Reeves et al. (2010) documented an estimation of soil C and N with R^2 of 0.92 using diffuse reflectance infrared Fourier transform spectroscopy (DRIFTS) tool. Mid-infrared (MIR) spectroscopy method’s accuracy varies with soil composition and moisture, indicating potential inconsistency across different soil types. With MIR spectroscopy, nitrate concentration in soil can be accurately measured with minimal errors (Linker, 2004). Fourier transform infrared Attenuated Total Reflectance (ATR) Spectroscopy effectively estimates soil water content and nitrate concentration, but it had a paste standardization problem that resulted in a nitrate determination error ranging from 6 to 14 mg N kg^{-1} of dry soil depending on soil type (Borenstein et al., 2006). In contrast, a study by Jahn et al. (2006), reported that MIR can effectively detect nitrate in various soil types, documenting high accuracy in laboratory and field experiments. Additionally, as with every other method, it requires soil-specific calibration due to varying carbonate contents in soils, which can interfere with nitrate peak detection and necessitate different calibration equations for accurate predictions. Also, MIR spectrometers are too expensive while VIS and SWIR spectrometers are relatively cheap.

In recent times, reflectance spectroscopy is being used to quantify soil nitrogen (Demattê et al., 2004; Insausti et al., 2012; Jia et al., 2017; Kawamura et al., 2017; J. P. Schmidt et al., 2017; Vibhute et al., 2020a; Vohland et al., 2014; Xiao et al., 2018). These techniques offer numerous advantages over traditional methods, including non-destructive sample analysis, cost-efficiency, user-friendliness, and rapid processing of large sample volumes with minimal sample preparation (Demattê et al., 2004). Spectral reflectance curves have been utilized to analyze various soil properties, such as organic matter, total iron, silt, sand, and mineralogy, which significantly influence spectral data set and features (Demattê et al., 2004). Nevertheless, the accuracy of nitrogen detection using reflectance spectroscopy can be impacted by factors like soil pretreatment, water content, and particle size (Nie et al., 2017). Elevated soil water content and extreme particle sizes (either too large or too small) can diminish the accuracy of NIR predictions (Jia et al., 2016; Vohland et al., 2014). Moreover, selecting effective spectral windows for nitrogen detection can pose challenges due to uncertainties in spectral absorption peaks or reflectance changes (Wang et al., 2021). To address these challenges, various strategies have been laid out to encompass soil pretreatment, spectral reflectance data processing, and feature band selection.

A pivotal challenge is discerning the precise spectral bands or windows that exhibit sensitivity to variations in soil NO_3-N. The spectral features of the Visible-Near Infra-Red (VNIR) regions on the electromagnetic spectrum are associated with vibration modes of molecular functional groups (Ben-Dor and Banin, 1995). Reflectance received from each single functional group will contain several wavelengths per energetic delta of the various excited states. However, reflectance received from various molecules with multiple functional groups results in a complex reflectance profile that needs to be treated accordingly using the techniques outlined by (Adamchuk et al., 2004; Wold et al., 1987). Spectrum decomposition is one of the techniques that can differentiate between various chemical components of the sample (Jolliffe, 2011). Spectral bands under Short-Wave Infra-red region (SWIR) correspond to the energy differences between excited and ground levels of functional groups (Viscarra Rossel et al., 2006). Additionally, it is now accepted that the visual spectrum obtains minuscule perturbations belonging to macronutrients and interference of water absorption in this region (Chang et al., 2001; Lunt et al., 2005; Summers et al., 2011). The application of non-imaging hyperspectral sensing, characterized by high spectral resolution (1-3 nm bandwidth), facilitates a more comprehensive spectral analysis. This approach emerges as an efficient tool for indirect estimation of soil nitrogen, demonstrating effectiveness even in scenarios necessitating high sampling densities.

Enhancing crop productivity through optimized nutrient input aligned with soil spatial variability has been demonstrated (Schmidt et al., 2017). However, traditional soil sampling methods pose challenges in cost, expansion, and in the implementation of precision agriculture. The application of spectroradiometer-based soil scanning offers a promising solution for monitoring soil macro-nutrient spatial variability (Clark and Roush, 1984; Chang et al., 2001; Vohland et al., 2014). This approach has potential to reduce the required number of soil samples and associated costs without compromising accuracy in soil characterization. Soil sampling, particularly at high spatial resolutions necessary for precision agriculture, is a time-consuming, laborious, and an expensive process. By minimizing soil sampling data set, labor and financial costs can be significantly reduced (Kawamura et al., 2017; Linker et al., 2005). This reduction is crucial for making precision agriculture more accessible and affordable for growers. The study aims to evaluate precision of non-imaging hyperspectral data in predicting soil NO_3-N concentrations through meticulous spectral window selection. This study undertakes the challenge of estimating soil nitrogen while reducing soil sampling data set, leveraging spectral similarities through application of the hyperspectral sensing tool.

2.2.1 Hypotheses

This study presents two hypotheses: (1) non-imaging hyperspectral sensing can accurately estimate surface soil NO_3-N; and (2) Spectral matching technique can effectively reduce soil sampling data set.

2.2.2 Objectives

The objectives of this study were to:

i. Estimate surface soil NO_3-N using non-imaging hyperspectral sensing.
ii. Observe spatial pattern of reduced soil sampling data set in reproducing a comparable spatial variability of soil NO_3-N

2.3 Materials and Methods

2.3.1 Study Sites

This study was conducted over two years (2021 and 2022) across three study locations in the state of Kansas, and one location in Fort Collins, Colorado (Figure 2.1). The study conducted in 2021 at KRV Topeka field was designated as ‘Site Year-1’. Similarly, studies conducted at KRV Rossville field in 2021 and 2022 were designated as ‘Site Year-2’ and ‘Site Year-3’. The Lonsinger location was designated as ‘Site Year-4’ for 2021. The Fort Collins location was designated as ‘Site Year-5’ and ‘Site Year-6’ for 2021 and 2022, respectively (see Appendix A - Study Area).

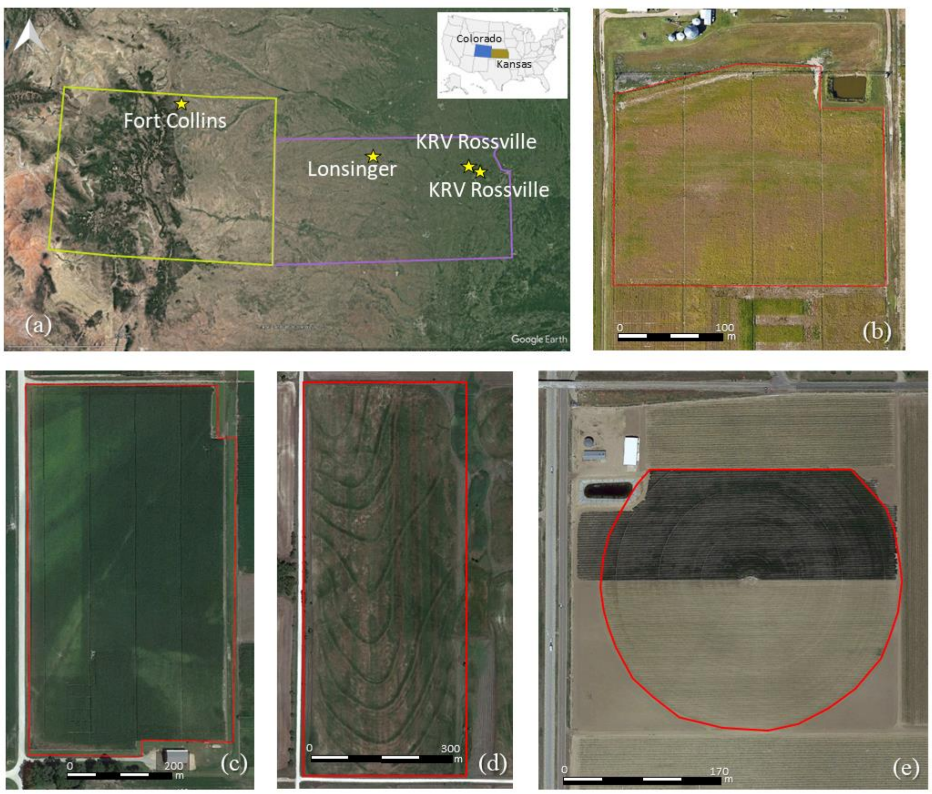


Figure 2.1. Study locations across Kansas and Colorado. (a) Map of four field locations is highlighted with yellow stars; (b) Kansas Research Valley Topeka experiment field (site year-1); (c) Kansas Research Valley Rossville experiment field (site year-2) and (site year-3); (d) Lonsinger research site (site year-4); and (e) Fort Collins site is marked with red boundary (site year-5) and (site year-6).

2.3.2 Data Collection

2.3.2.1 Reference Soil Sample Collection and Analysis

Soil samples were collected from all six site years. Within field soil sampling locations were identified with a stratified random soil sampling design. A composite soil sample consisting of 8 to 10 soil cores was collected per 0.202 ha grid cell size for each study area. Soil sampling locations were recorded with a differential global positioning system (DGPS). Composite sample consisting of multiple soil cores were collected from 0 to 30 cm. The collected surface soil samples were air-dried and passed through a 2-mm sieve. (see Appendix Table F.1) at a commercial soil testing laboratory (AgSource Laboratories in Lincoln, Nebraska, USA).

2.3.2.2 Soil Spectra Collection and Analysis

Spectral measurements were acquired for the soil samples (as illustrated in Sec. 1.2.1) at the Precision Ag Lab, Kansas State University (Figure 2.2), using a FieldSpec3 spectroradiometer (Analytical Spectral Devices, Boulder, CO, USA). The ASD FieldSpec3 has a spectral range from 350 nm to 2500 nm. The spectral resolution (bandwidth) varies across lower to higher wavelengths. In the spectral range of 700-1400 nm, the spectral resolution is 3 nm, while the spectral resolution is 10 nm for 1400 to 2100 nm wavelength range. The FieldSpec3 spectroradiometer was calibrated using spectral reading taken for a spectralon white reference panel. The instrument was recalibrated after every 10th sample. This step was crucial for obtaining reliable reflectance measurements and mitigating potential instrument-related biases. To further improve the accuracy and repeatability of spectral readings, measurements were conducted in a completely dark room with a Tungsten Quartz Halogen b/1500 source of illumination (ASD muglight), which minimized stray light and specular reflection errors. The muglight was heated for 3 hours before taking the measurement. The spectral reading was measured for soil samples following the protocol suggested by the Soil Spectroscopy Group and the Development of a Global Soil Spectral Library (Viscarra Rossel et al., 2006).

For spectral readings, 45 gm soil samples were placed on borosilicate petri dishes and positioned on top of a Muglight for illumination during scanning. The spectrometer features a 25° field of view and a 12 mm spot diameter, facilitating comprehensive and consistent data acquisition. Four replications of the readings were taken for each soil sample, rotating the petri dish by 90 degrees between each replication. The average of these four replications was calculated for further stesps. This procedure accounts for possible variations in reflectance due to sample heterogeneity or instrument positioning, eventually yielding a robust and reliable dataset for further analysis.

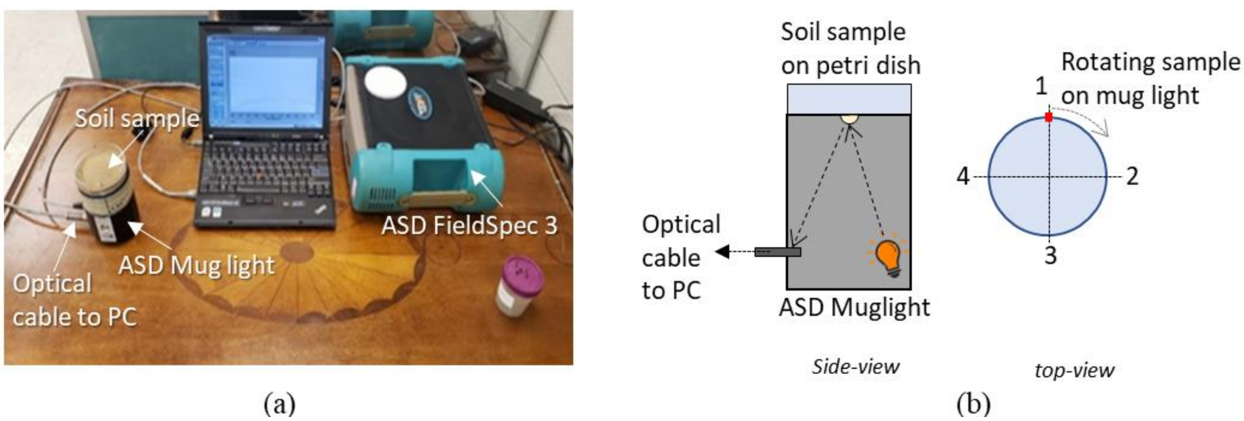


Figure 2.2. Instrument setup for soil spectral data acquisition. (a) FieldSpec 3 spectrometer with mug light, soil sample on borosilicate petri dish, and spectralon (white reference); (b) Four replications of readings for each soil samples with rotating the petri dish by 90 degrees on mug light.

2.3.3 Model Development and Statistical Analysis

The schematic workflow in Figure 2.3 shows the statistical analysis and model development for developing Nitrate-nitrogen maps. The methodological details are illustrated in following subsections.

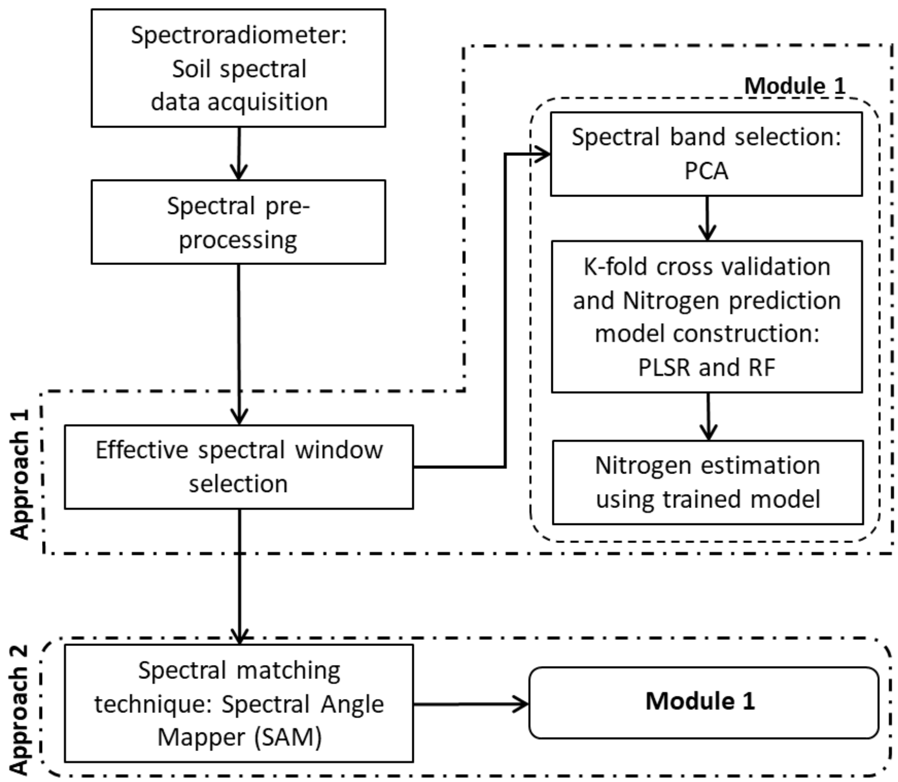


Figure 2.3. Schematic workflow for statistical analysis and model development to estimate soil NO_3-N. Approach 1 and 2 highlight methodological flow with initial high sampling data set and reduced sampling data set soil data, respectively.

2.3.3.1 Spectral Pre-processing

The spectral pre-processing techniques applied were objective oriented and case sensitive. Splice correction, removal of reflectance below 40 nm, Savitzky–Golay smoothing, Multiplicative scatter correction (MSC) and continuum removal were applied to minimize the influence of noise, remove spectral artifacts, and enhance the signals present in the dataset.

Splice correction was used to remove the inconsistencies through correcting the spectral discontinuity between three detectors during data collection. Splice corrections were applied at 1000 nm and 1830 nm to guarantee that the data remains consistent across the entire wavelength range. R studio used to perform the spectral pre-processing, starting with a "spliceCorrection" function from the "ithir" package in R (Colin et al., 2017) to smoothen the transitions between the detectors. Absorbance values are more linearly related to the concentration of the ingredients in the soil sample. Linearity makes absorbance data more suitable for quantitative analysis. Therefore, the spectral data of reflectance to absorbance were converted with the following equation:

$$\text{Absorbance} = \log_{10} \frac{1}{\text{Reflectance}} \qquad (2.1)$$

A two-step pre-processing approach was applied to refine the spectral data and eliminate unwanted noise, which involved Savitzky-Golay (SG) smoothing followed by Multiplicative Scatter Correction (MSC). SG smoothing was applied to reduce the noise in the spectra while preserving the shape of the absorption features. The smoothing was achieved by fitting a polynomial of a specified degree to a moving window of data points (Savitzky & Golay, 1964). In this study, a window size of 15 and a polynomial of degree 0 were used. MSC was then applied to correct for the multiplicative effects of scattering. This step contributed by minimizing the influence of the soil physical properties of the sample, such as particle size (Rasskazov et al., 2019).

The continuum removal (CR) technique was applied to highlight the absorption features of the elements in the soil spectra. The continuum is an envelope fitted to the spectra, and the removal process involves dividing the original spectra by the continuum. This step normalizes the spectra and enhances the absorption features. The continuum removal was applied on the pre-processed multiplicative scatter correction. R code with “continuumRemoval” function was used.

2.3.3.2 Effective Spectral Window Selection Experiment

Two separate experiments were conducted in this study to select effective spectral windows that are sensitive to Nitrate-nitrogen. The ASD FieldSpec3 (spectroradiometer with Muglight) was used to measure the soil spectra of each soil sample prior to mixing it with liquid and after it was dried. Soils samples were collected and processed as described in Section 2.3.2.1. The experiment was conducted with two replications of five soil samples for site year-1 and site year-5. From each sample, 40 g air-dry soil was taken on a borosilicate glass petri dish, and ASD spectral readings were collected in the dark room setup. Samples from the same soil were taken on another 4 petri dishes. One soil was mixed with deionized water (40 ml volume), while three soil samples were treated with sodium nitrate ($NaNO_3$) solution at three different concentration levels of 6.2 ppm, 62 ppm, and 620 ppm of NO_3^{-}, respectively. Spectra readings were measured from the wet soil every six to twelve hours until it was completely dry. The spectral window of treated soil samples sensitive to different levels of sodium nitrate was selected as a potential effective spectral window of NO_3^{-}. The effective spectral window selection was performed with visual observations and careful comparison of sensitive spectra regions to NO_3^{-}, and sensitive spectra regions to moisture.

2.3.3.3 Spectral Matching Techniques

The Spectral Angle Mapper (SAM) is an advanced analytical technique employed in this research to assess the similarity between spectral signatures derived from soil samples. This method interprets each spectrum as a vector in a high-dimensional space, each dimension representing a spectral band. SAM technique uses the calculation of the angle between two spectral vectors, a measure that is robust to variations in illumination and sensor response. The function calculates the angle as,

$$\Theta = \cos^{-1}\left(\frac{\sum_{k=1}^{m} S_{i.k}S_{j,k}}{\sqrt{\sum_{k=1}^{m} S_{i,k}^2}\sqrt{\sum_{k=1}^{m} S_{j,k}^2}}\right) \tag{2.2}$$

Where:

θ is the spectral angle in radians,

S_i and S_j are the ith and jth spectral vector, respectively,

$S_{i,\,k}$ and $S_{j,\,k}$ are the spectral reflectance values at the kth band for spectra i an j,

m is the total number of bands.

A systematic pairwise evaluation was performed on the entire dataset. The SAM values for each unique spectral pair were stored in an n × n matrix, with n being the total number of spectra. Due to the symmetrical property of the angle between two vectors (i.e., the angle from S_i to S_j is the same as from S_j to S_i), the computation was efficiently executed only for $i<j$, thereby reducing the computational burden by nearly half. Following the pairwise comparison, a critical threshold of radians was applied to the matrix of SAM values to identify spectral redundancies. Spectra pairs with SAM values below specific threshold were deemed similar enough to be considered redundant, leading to potential biases in subsequent analyses. Therefore, these spectra were set aside for removal. A threshold that resulted in relatively highest R^2 in estimation of soil NO_3-N was selected, suggesting it as an appropriate cutoff for distinguishing between different soil types based on their spectral signatures.

2.3.3.4 Spectral Band Selection with PCA

Principal Component Analysis (PCA) is a technique for reducing data that is frequently employed for evaluating remotely sensed data. PCA were used for spectral analysis to select the most appropriate wavelength(s). Narrowing the spectrum down to a smaller range of wavelengths simplifies the model while yielding results that are either comparable to or superior to those obtained using the whole spectrum (Damiano et al., 2019). PCA-loading was carried out to choose meaningful wavelengths to determine which are most relevant for prediction of soil Nitrogen from the spectral reflectance. According to a study by Xie et al. (2018), these selected wavelengths to determine the quantity of nitrogen contained in maize may result in good performance.

2.3.3.5 K-fold Cross Validation and Nitrogen Estimation Model Construction

In this study, 10-fold cross validation was adopted for model construction, meaning that 10 data partition schemes were used to model the same set of input-target data (Kucheryavskiy et al., 2020; Ludwig et al., 2018). The soil spectral data and associated soil nitrogen (chemical analysis) information were used as input and target variable. The partial least squares (PLS) regression, and random forest regression (RF) algorithms were adopted to develop the model. The prediction models were evaluated with root mean absolute error (*MAE*), root mean square error of prediction (RMSE), modeling coefficient of determination (R^2) as expressed in Eq. (2.3) - (2.5). Based on the evaluation of the prediction model (taking non-imaging hyperspectral data as predictors and soil ground truth data collected and processed as target variables), the best performing model with high values of R^2 was chosen to estimate N rate.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \qquad (2.3)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (2.4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (2.5)$$

Where:

n: number of data points,

y_i: actual soil nitrogen content,

$\hat{y}_i$: predicted soil nitrogen content,

$\bar{y}$: the mean actual soil nitrogen content.

2.3.4 Prediction Methods for Estimated Nitrogen

Geostatistical analyses were conducted to evaluate spatial variability and pattern of measured and estimated surface soil NO_3-N. Semivariogram Modeling: Spherical, exponential, and Gaussian semivariogram models under simple and ordinary kriging methods were compared to identify best fit for modeling spatial distribution of soil NO_3-N. The best fitted model was selected as the most appropriate based on comparable Root Mean Square Error (RMSE) and Average Standard Error (ASE) that results a lower standardized error. The study investigated spatial dependence of soil NO_3-N using the nugget to sill ratio derived from the semivariogram models. This analysis used to classify the spatial dependence into strong, moderate, or weak, according to established thresholds. The classification was further correlated with the type of crop previously cultivated. This was used to highlight the impact of crop rotation and management practices on soil NO_3-N values. This geostatistical method allowed for the interpolation of soil NO_3-N values across the study sites, facilitating the visualization of spatial variability. All soil

NO_3-N maps were visualized using a green-to-red color scheme to represent the concentration levels, with green indicating higher concentrations and red indicating lower concentrations.

2.4 Result and Discussion

2.4.1 Nitrogen Estimation with High Soil Sampling Data set

Spectral data were passed through spectral pre-processing, selection of effective spectral window and PCA before estimating surface soil NO_3-N. Spectral pre-processing reduced the influence of noise and enhance signals present in dataset. Effective spectral windows were selected to identify the sensitive window to Nitrate-nitrogen. PCA was used further to select spectral bands that contributed most significantly to the variability of soil NO_3-N.

2.4.1.1 Spectral Pre-processing

The spectral preprocessing techniques corrected inconsistent and noisy spectra data. It transformed the reflectance to absorbance and reduced the noise by removing the short wavelength below 400 nm. The Savitzky-Golay (SG) smoothing technique effectively reduced the random noises in the spectral data by applying a moving average filter method. A similar smoothing method was effectively used by (Vohland et al., 2014). The MSC reduced the variations in baseline of the spectra. In a separate study, (Kooistra et al., 2003) also recommended spectral processing with MSC, where corrected spectra produced a best prediction of organic matter and clay content when compared with spectra without MSC preprocessing. Therefore, the spectral features related to soil composition became more prominent than the physical properties of soil. Spectral data at different correction steps are presented in Figure 2.4 , which includes 42 soil samples collected from KRV Topeka site.

The continuum-removed spectra result is presented in Figure 2.4, which displays more distinct and well-defined absorption features compared to the spectra after Savitzky-Golay smoothing and MSC processing. A continuum removal step enhances the contrast between the absorption features, which were critical for understanding the composition and properties of the soil samples (Clark and Roush, 1984). The application of the preprocessing resulted in a significant improvement in the spectral data. This enhanced spectral data to facilitate a reliable analysis of the soil nitrogen estimation. Spectra data was processed with the first derivative to enhance the resolution of spectral resolution. The first derivative of absorption spectrum makes it easier, with visibility to detect the peaks and troughs of the spectra. The sensitivity of spectra due to high concentration of NO_3^- were visually observed from the first derivative output (see Sec. 2.4.1.2).

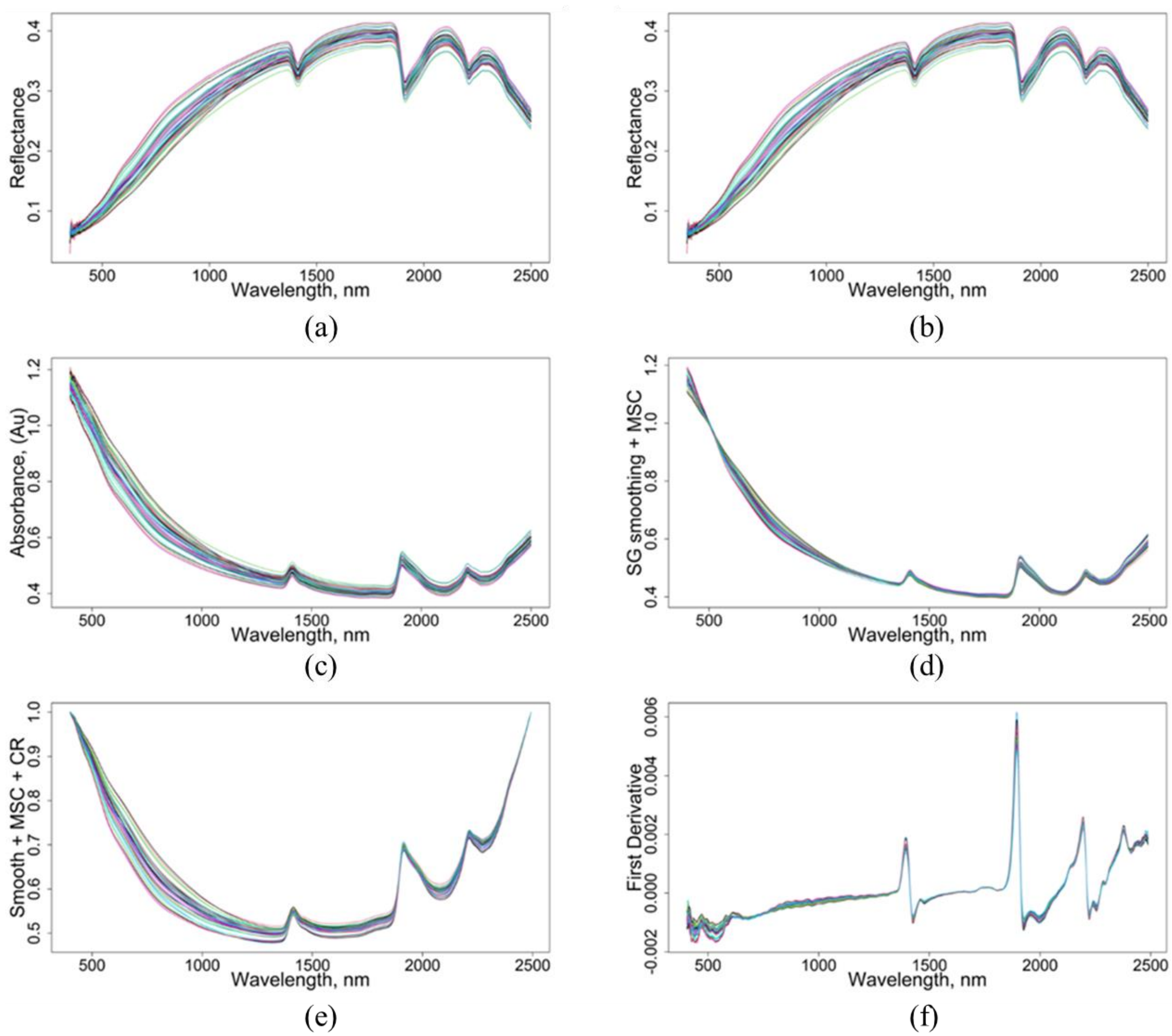


Figure 2.4. Spectral analysis of site year-1. (a) reflectance vs. wavelength; (b) splice corrected reflectance vs. wavelength; (c) absorbance (below 400 nm removed) vs. wavelength; (d) Savitzky-Golay smoothing + Multiplicative scatter correction applied absorbance vs. wavelength; (e) Savitzky-Golay smoothing + Multiplicative scatter correction + continuum-removal applied absorbance vs. wavelength; (f) first derivative vs. wavelength.

2.4.1.2 Effective Spectral Windows

The effective spectral windows are highlighted in Figure 2.6 for both the site years 1 and 2. Soil samples pretreated with high and medium concentrations of $NaNO_3$ presented peaks and troughs. Figure 2.6 illustrates the first derivative of the Visible and Near-Infrared (VNIR) spectral region corresponding to four different treatments. Among these treatments, three were subjected to pre-treatment with different concentrations of sodium nitrate, categorized as low, medium, and high. Analysis over two site years indicates that soil treated with a low concentration of sodium nitrate (6.2 ppm) exhibited a first derivative spectral response more akin to that of dry soil. Conversely, the spectral response of soil for medium (62 ppm) and high (620 ppm) concentrations of sodium nitrate exhibits a similar trend. The similar spectral feature pattern suggests a detection of sensitivity of soil samples spectral reading when treated with $NaNO_3$. A study by Ehsani et al. (1999) confirmed that the absorption peak at 1942 nm in the Near-Infrared (NIR) spectrum, seen at low nitrate concentrations, is specific to nitrate ions and not due to ammonium or calcium contaminants. Detection of low concentration of soil NO_3^- (samples with 62 ppm of $NaNO_3$) allows to scale up these effective spectral windows for soils that are not pretreated (in practical consideration) and very low concentration of NO_3-N presence. Differential spectral features were apparent for soil mixed with high and medium $NaNO_3$ levels as shown in Figure 2.6 . The sensitivity of NO_3^- ions were observed at three spectral windows of 1460 nm to 1690 nm, 1730 nm to 1780 nm, and 1940 nm to 2150 nm. The first and third spectral windows were relatively broad (Figure 2.6). Whereas the effective second spectral window selected was shorter (Figure 2.6). Similar spectral ranges were observed for the site year 2. Broader spectral windows were visible for the first and third effective spectral windows as compared to the second. The range of the effective spectral window selected includes wavelengths: W1 (1460-1690 nm), W2 (1730-1780 nm), and W3 (1940-2150 nm).

The fundamental modes of vibration for the NO_3- molecule predominantly occur in the mid-infrared wavelength regime. The NO_3- molecule is planar and has three equivalent N−O bonds. The fundamental peaks are associated with NO_3 molecular bond stretching and bending (Figure 2.5). Additional peaks arising from overtones and combinations, present in the near-infrared (NIR) region, are generally weaker than the fundamental peaks in the mid-IR region (Jahn et al., 2005). These NIR peaks often get overshadowed by absorbances attributed to carbonate and water molecules. In a soil nitrate sensitivity study in NIR region, (Ehsani et al., 1999) found 1942 nm wavelength is sensitive to nitrate, when $NaNO_3$ was mixed with dimethyl sulfoxide solvent. On the contrary, the present study indicated a broad range of sensitivity at W3 window while water was used as solvent.

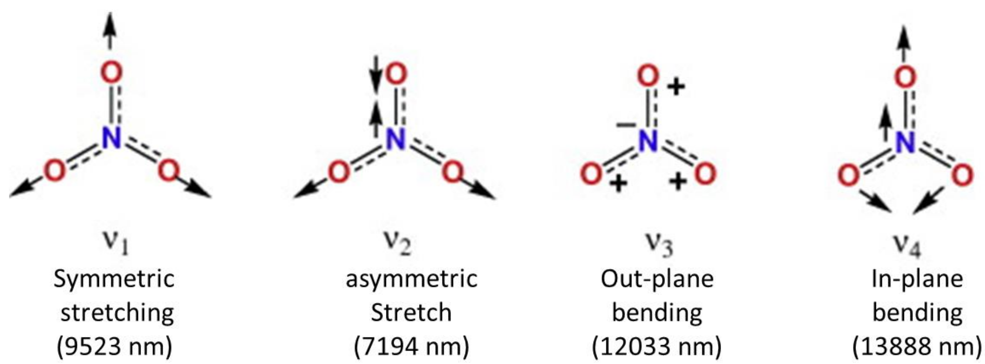
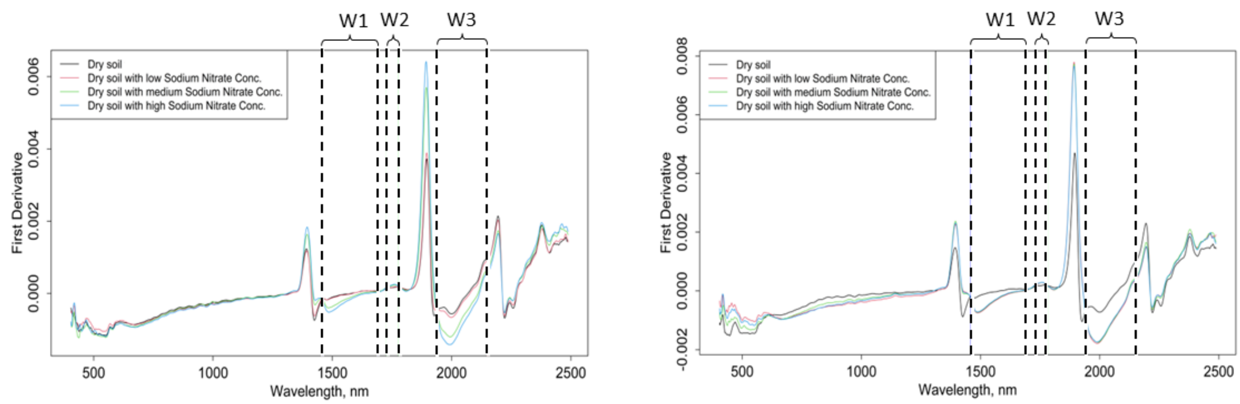


Figure 2.5. Fundamental modes of vibration in a NO3- molecule (adapted (Jahn et al., 2005)**).**

Specific wavelengths including 1472 nm, 1795 nm, 2210 nm, and 2296 nm were also reported to have a higher correlation (R^2 of 0.95) while estimating soil NO_3-N (Vibhute et al., 2020b). A study by (Anom et al., 2000) also reported that 1681 nm reflectance feature was significantly correlated with soil NO_3-N, which corresponds with the findings of this study.



(a) (b)

Figure 2.6. First derivative spectra of soils in controlled experiment with and without pretreatment of sodium nitrate. (a) site year-1, and (b) site year-2. Effective spectral windows are highlighted with black vertical dotted lines and marked as W1 (1460-1690 nm), W2 (1730-1780 nm), and W3 (1940-2150 nm).

2.4.1.3 Principal Component Analysis (PCA) of Spectral Data

PCA was used in this study to reduce the dimensionality of the spectral data. In PCA analysis, the spectral bands that contributed most significantly to the variability of the soil $NO_3^- - N$ were used (See Sec. 2.4.2). PCA revealed that the majority of variance was accounted for by the initial principal components. For site year-1, 2 and 3, more than 99.8 % of variance was explained by the first three dimensions (Table 2.1). Similar patterns were observed for site years 4 to site 6, where the first three dimensions capture above 99.6% to 100% of variance (Table 2.1). The dominance of PC1 for all site years, in explaining the variance suggests a strong common factor influencing the soil parameters across the dataset. The presence of a soil $NO_3^- - N$ in the selected effective spectral windows strongly influences soil spectral characteristics. The second principal component (PC2), while accounting for a considerably smaller portion of the variance compared to PC1, still represents a significant part. This implies the existence of another, less dominant, set of factors or conditions influencing soil properties. The very low levels of variance accounted for by subsequent dimensions (Dim3 and onwards) suggest that these factors contribute minimally to the variability in the soil data. It implies that most of the soil properties' variability is captured by PC1 and PC2.

Table 2.1. Principal component analysis results for spectral data collected for all six site years.

Sites		Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Dim8	Dim9	Dim10
Site year 1	Variance	324.9	46.4	30.0	0.5	0.1	0.1	0.0	0.0	0.0	0.0
	% of var.	80.8	11.5	7.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0
	Cumulative %of var.	80.8	92.4	99.8	99.9	100.0	100.0	100.0	100.0	100.0	100.0
Site year 2	Variance	368.7	26.4	6.7	0.1	0.1	0.0	0.0	0.0	0.0	0.0
	% of var.	91.7	6.6	1.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Cumulative % of var.	91.7	98.3	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Site year 3	Variance	374.5	24.5	2.7	0.1	0.1	0.0	0.0	0.0	0.0	0.0
	% of var.	93.2	6.1	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Cumulative % of var.	93.2	99.3	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Site year 4	Variance	369.0	31.4	1.4	0.1	0.1	0.0	0.0	0.0	0.0	0.0
	% of var.	91.8	7.8	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Cumulative % of var.	91.8	99.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Site year 5	Variance	152.6	38.0	0.6	0.4	0.2	0.1	0.0	0.0	0.0	0.0
	% of var.	79.5	19.8	0.3	0.2	0.1	0.0	0.0	0.0	0.0	0.0
	Cumulative % of var.	79.5	99.3	99.6	99.8	99.9	100.0	100.0	100.0	100.0	100.0
Site year 6	Variance	165.8	24.9	0.7	0.4	0.2	0.0	0.0	0.0	0.0	0.0
	% of var.	86.3	13.0	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0
	Cumulative % of var.	86.3	99.3	99.7	99.9	100.0	100.0	100.0	100.0	100.0	100.0

2.4.1.4 Accuracy of Nitrogen Estimation

Estimated NO_3-N values using two machine learning models (PLSR and RFR) are presented on scatter plots for each site year in Figure 2.7 and Figure 2.8. Estimated NO_3-N values were generated based on the soil samples collected with high data set (high sampling density).

Three error metrics of RMSE, MAE and R^2 were used to measure the model fitness while comparing the estimated with observed NO_3-N values.

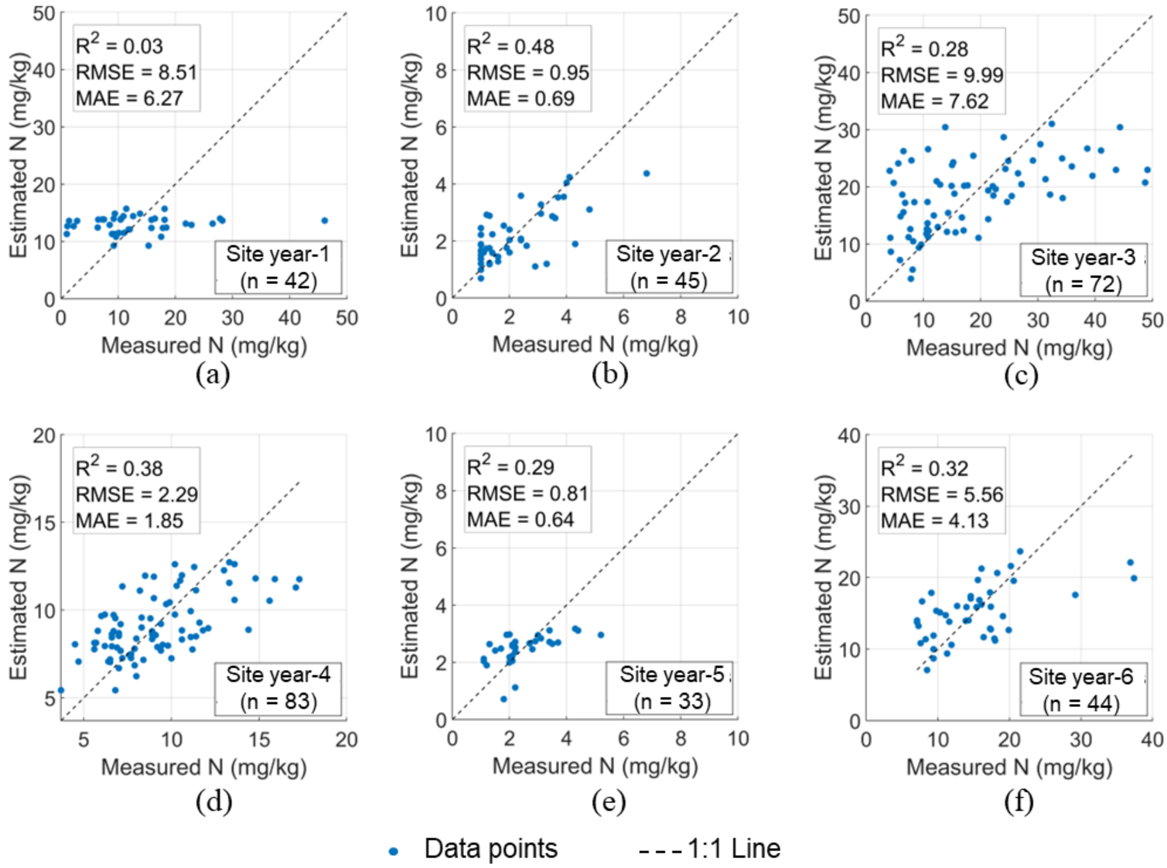


Figure 2.7. Accuracy of partial least square regression model in soil NO_3-N estimation for different site year data collected with high sampling data set. (a) Site year-1 (Topeka 2021); (b) Site year-2 (Rossville 2021); (c) Site year-3 (Rossville 2022); (d) Site year-4 (Lonsinger 2021); (e) Site year-5 (Fort Collins 2021); (f) Site year-6 (Fort Collins 2022). Total number of samples for each site year is represented with the variable n.

The accuracy assessment in Figure 2.7 a-f showed that the PLSR model with high soil sampling data set performed relatively well for site year-2, with R^2 values of 0.48 compared to the rest of site years (Figure 2.7 b). However, the results obtained for site year-1 to site year-6 were in a similar range, indicating that PLSR model for estimation of NO_3-N is not satisfactory. Site year-1, site year-3 and site year-6 indicated high error rates with RMSE of 8.51 mg/kg, 9.99 mg/kg, and 5.16 mg/kg, respectively for PLSR model. The high RMSE values relate with the presence of high soil NO_3-N as compared to site year-2, site year-4, and site year-6, where the soil NO_3-N range was smaller. This difference in RMSE was due to the difference in crop cultivated in the season before soil sampling. For example, soybean was cultivated prior to soil sampling for site year-1, site year-3 and site year-6. Soybean being a leguminous crop contributed to fixation of atmospheric N (N_2), as result it increased the concentration of soil NO_3-N (Córdova et al., 2019).

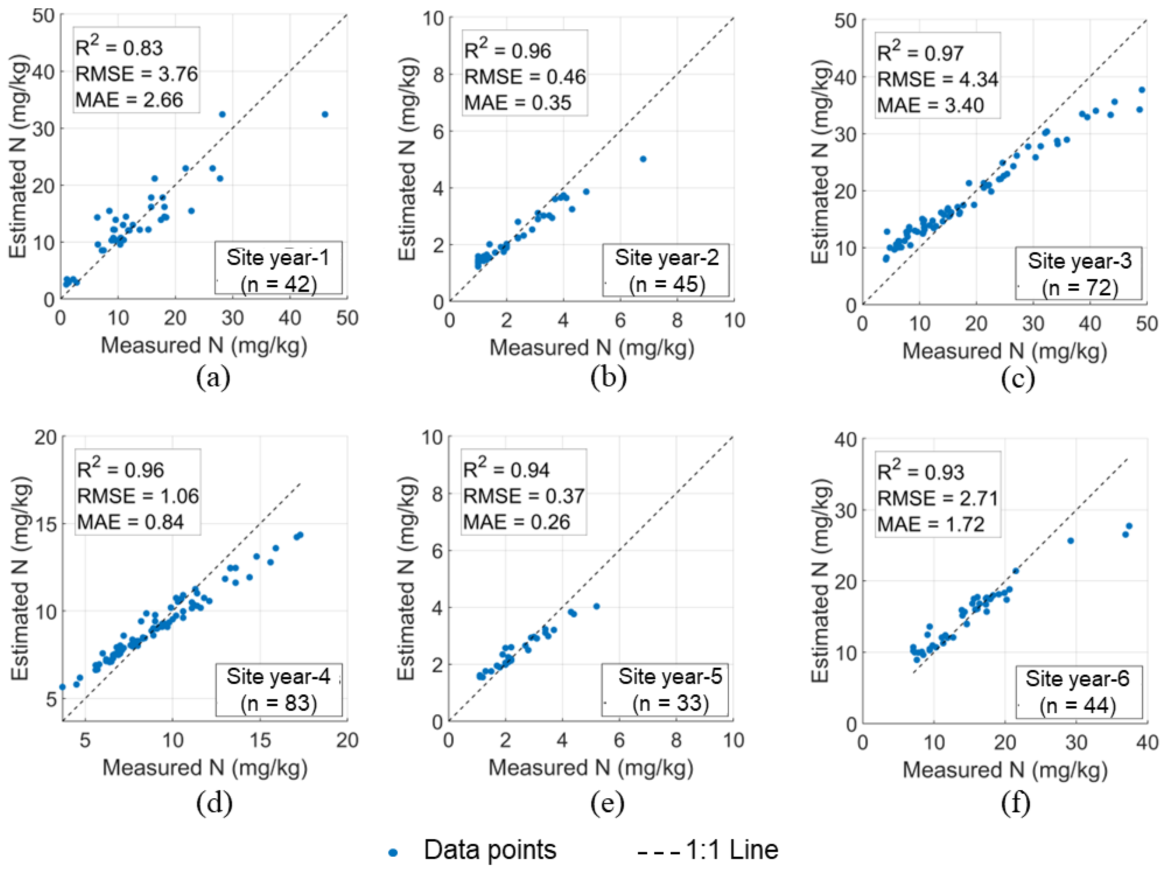


Figure 2.8. Accuracy of random forest regression model in soil N estimation for different site year data collected with high sampling data set. (a) Site year-1 (Topeka 2021); (b) Site year-2 (Rossville 2021); (c) Site year-3 (Rossville 2022); (d) Site year-4 (Lonsinger 2021); (e) Site year-5 (Fort Collins 2021); (f) Site year-6 (Fort Collins 2022). Total number of samples for each site year is represented with the variable n.

As shown in Figure 2.8 a-f, RFR model performed best for site year-3 with R^2 values of 0.97, RMSE (0.434 mg/ka), and MAE (3.4 mg/ka) followed by site year-4 and site year-2 with R^2 values of 0.96. The RMSE and MAE values of 1.06 mg/ka and 0.84 mg/ka for site year-4 were obtained, followed by 0.46 mg/ka and 0.35 mg/ka for site year-2 (Figure 2.8). The lowest model performance in the estimation of NO_3-N using RFR was for site year-1 with R^2 (0.83), RMSE (2.73mg/kg), and MAE of 1.72 mg/kg (Figure 2.8 a). However, RFR model performs better without being affected by amount of NO_3-N present in the soil. In general, the comparison between, PLSR and RFR model in the estimation of NO_3-N revealed that, RFR model performed best compared to the PLSR model. Using site year-3, RFR model resulted in better NO_3-N estimation with R^2 values of 0.97, compared to R^2 values of 0.28 for PLSR. Similarly, RFR based NO_3-N estimation reduced the RMSE and MAE by 56.5 % and 55.3 % from PLSR model. RFR model demonstrated a robust and reliable performance in estimating soil NO_3-N across all site years.

Xu et al. (2020) also reported the superiority of RFR with an R^2 value of 0.65. The overall accuracy of RFR was found to surpass that of PLSR across all site years, a result attributed to the enhanced generalization ability and noise resistance of RF models under limited data conditions. Therefore, RFR emerges as a reliable modeling method for constructing an estimation model of soil NO_3-N using hyperspectral data. Variations in soil nitrogen (N) estimation accuracies across different site years can be attributed to their soil type and crop-rotation practices. Song et al. (2019) highlighted that the disparity in soil spectral reflectance between two crop rotation systems (corn-wheat) may be partly induced by the incorporation of straw back into the field. Continuous decomposition of straw is anticipated to release water-soluble substances and crude protein, thereby increasing carboxyl content. Notably, corn straw often retains stable structures of tannin and lignin, making it resistant to decomposition. In contrast, within the corn-soybean-corn system, soybean straw decomposition rates are generally high, influencing the physical and chemical properties of the soil. Abundant lignin and cellulose in crop stubble have been identified as causing absorption characteristics in the 1710 and 2100–2300 nm wavelengths (Li et al., 2015; Serrano et al., 2015). These observations contribute to the understanding of the intricate relationship between crop rotation systems, straw decomposition dynamics, and their impact on soil properties, ultimately influencing the accuracy of soil nitrogen estimation.

Soil nitrogen estimation at high accuracies utilizing machine learning algorithms has been reported in several other studies, employing a spectral range of 300-2500 nm (Madari et al., 2006; Morellos et al., 2016; Santos et al., 2020; Vohland et al., 2014). However, these studies typically calculated total soil nitrogen, as opposed to the focus on soil NO_3-N in the present study. The superior performance in nitrogen estimation can be attributed to the effective utilization of spectral windows that exhibit minimal overlap with other spectral features, particularly those related to NO_3-N. Notably, PLSR exhibited an overestimation in soil NO_3-N values compared to RFR. This overestimation by the PLSR model has also been observed by others (Morellos et al., 2016) when estimating soil total nitrogen.

The comparison of different models reveals that RF machine learning techniques yield highly satisfactory results for predicting soil NO_3-N, but limitations exist regarding the generalization of findings of this study. This limitation arises because, for machine learning techniques, model generalization is highly dependent on the variability present in the calibration and testing samples. Consequently, there is a need to incorporate samples from a broader diversity of soil types and different fields to enhance the models' generalizability. Transferring a model derived from one site year to another for nitrogen estimation presents a considerable challenge. Ongoing efforts to develop soil property models using extensive soil databases are underway, based on the assumption that each chromophore affects all soil types equally (Rossel et al., 2016). However, building a generic model solely based on soil type is subjective. Recent studies indicate that a preliminary division of the dataset into subsets or clusters may enhance the prediction performance of soil properties (Araújo et al., 2014). On the contrary to soil NO_3-N estimation with high sampling data set, NO_3-N estimation models were further developed with lesser sampling data set data produced by spectral angle mapper (SAM). These reduced sampling data set were then used for PLSR and RFR model development. Soil samples those have the same characteristics produce a similar soil spectral reflectance. Therefore, pairwise soil spectral data reading is performed to quantify soil samples characteristics. SAM was used to quantify spectral similarities with radian measurements. Pair of soil samples that produces smaller radians showed a similar characteristic, as a result spectral angle (in radian) threshold were produced to remove soil spectra with similar characteristics. As a result, high soil sampling intensities were reduced with radian threshold values of between 0.001 and 0.003, for all site years. A heat map of pairwise comparison of spectral angle (in radians) and selected threshold of 0.002 for site year-2 is shown in Figure 2.9. The spectral angle threshold was selected based on results which produced the highest R^2 during model training for RFR. The selected threshold reduced the high soil sampling data set from 45 soil samples to 23 soil samples for site year-2.

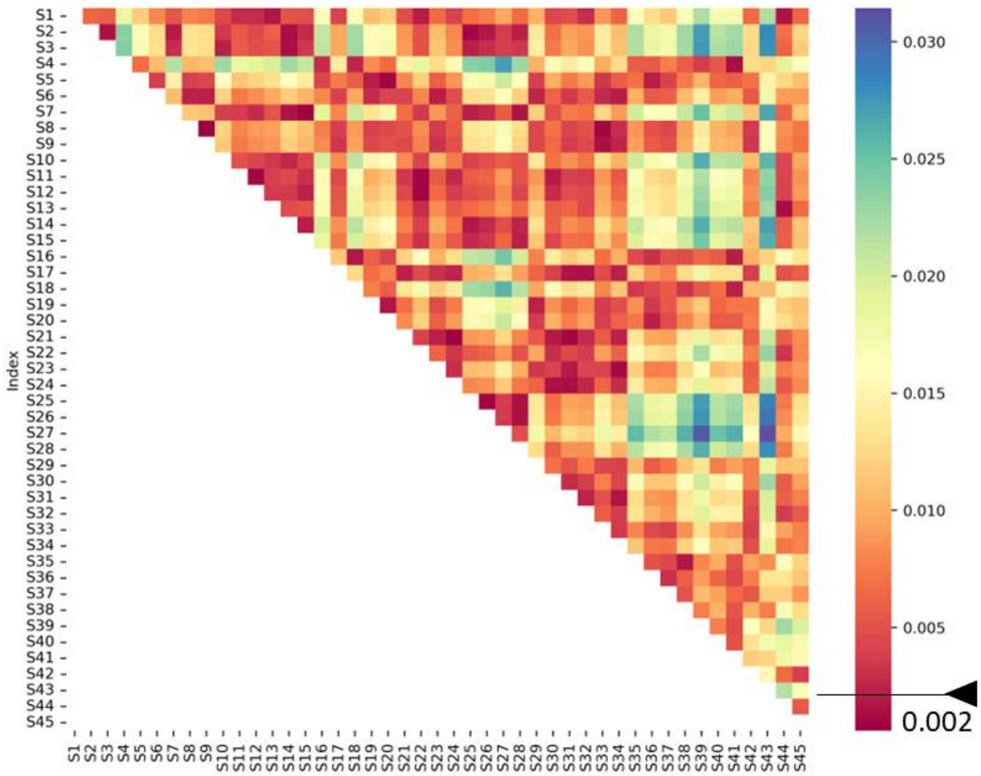


Figure 2.9. Pairwise evaluation of spectral angles (in radian) among the soil samples for Site year-2 (n = number of soil samples = 45). Threshold value is highlighted with black arrow head.

Spectral matching commonly involves the derivation of spectral angles among various samples, followed by clustering (Ren et al., 2020). Similarly, a threshold-based approach on spectral angles is employed in this study for the selection of data samples. Traditionally, the determination of threshold angles has relied on trial-and-error, a conventional method. However, a computational method for semi-automatic determination, with less biased threshold angles for the SAM technique, can be further explored (Shahriari et al., 2014). Nevertheless, noise emerges as a critical factor that could impact the accuracy of spectral similarity estimates. The denoising of soil spectral data by smoothing the curves while preserving the original spectral features is a central challenge in numerous studies (Zeng et al., 2023).

SAM reduced soil sampling data set for site year-4, site year-5 and site year-6 from 83, 33, and 44 to 26, 17 and 12 soil samples, respectively. For soil NO_3-N estimation with PLSR model, site year-2 resulted a relative higher R^2 of 0.46, the second lowest RMSE and MAE with value of 1.10 mg/kg and 0.83 mg/kg (Figure 2.10). In similar manner to soil NO_3-N estimation with high soil sampling data set, the presence of high soil NO_3-N in the soil reported relatively the highest RMSE and MAE for site year-1, site year-3 and site year-6.

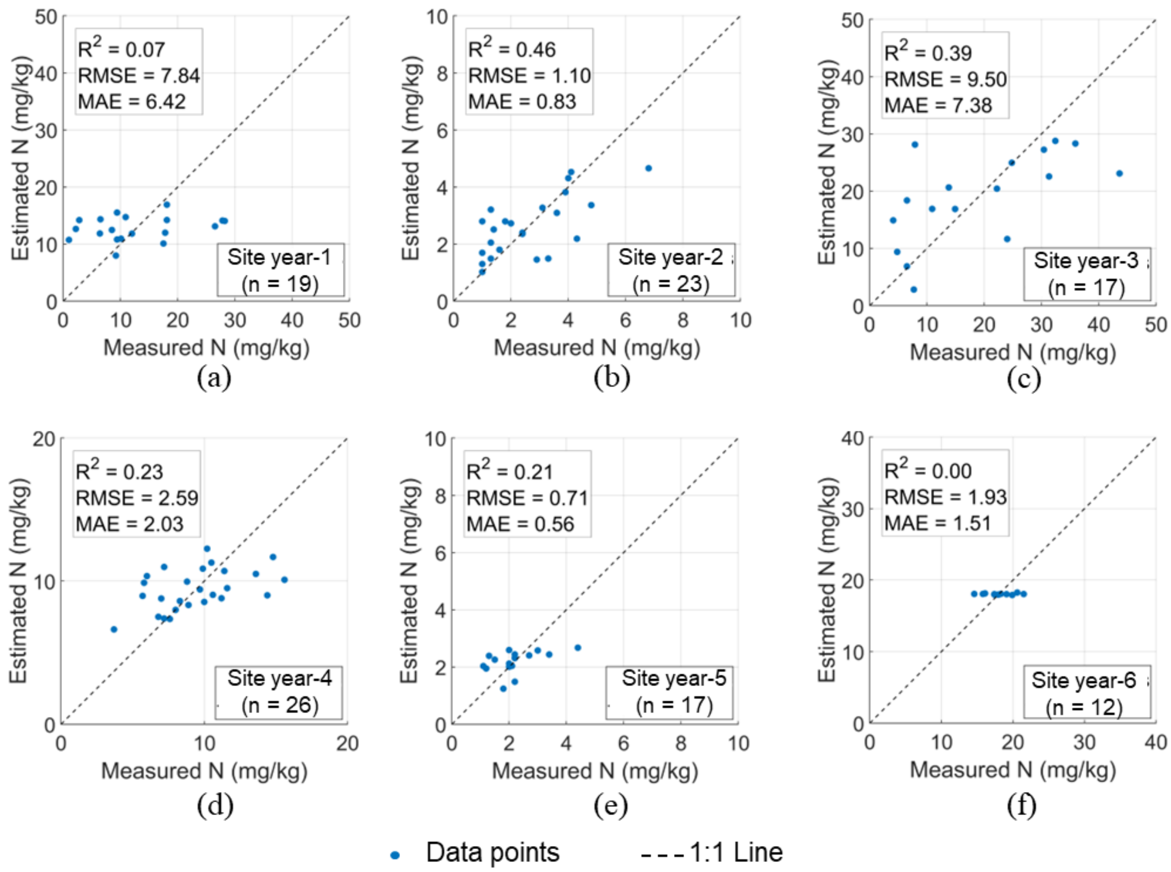


Figure 2.10. Accuracy of partial least square regression model in soil N estimation for different site year data with reduced (spectral angle mapper based) sampling data set. (a) Site year-1 (Topeka 2021); (b) Site year-2 (Rossville 2021); (c) Site year-3 (Rossville 2022); (d) Site year-4 (Lonsinger 2021); (e) Site year-5 (Fort Collins 2021); (f) Site year-6 (Fort Collins 2022). Total number of samples for each site year is represented with the variable n.

On the other hand, the RFR model resulted in soil NO_3-N estimation with smallest R^2 0.84 for site year-6, and highest R^2 values of 0.99 for site year-1 and site year-3 (Figure 2.11). Estimation of NO_3-N with RFR model using reduced soil sampling data set reduced the RMSE and MAE error metrics by 52.2 %, 49.18 % when compared with PLSR model, for site year-1 and site year-3, respectively.

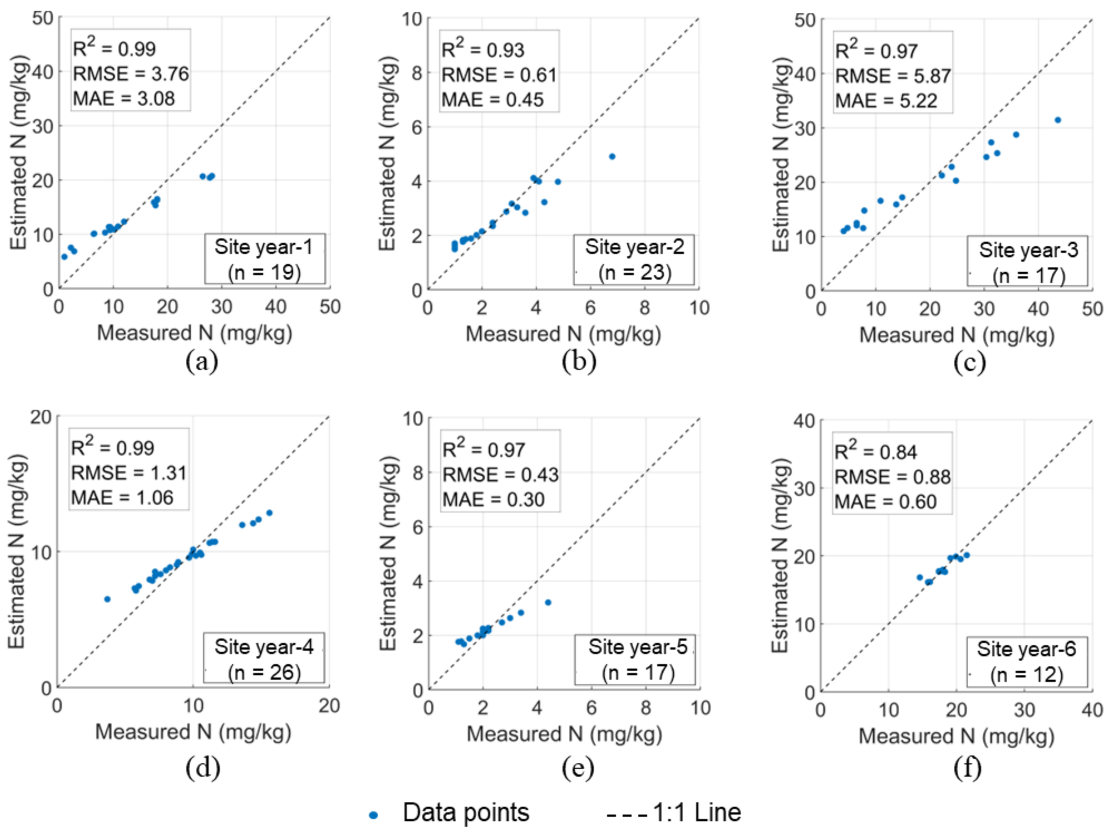


Figure 2.11. Accuracy of random forest regression model in soil N estimation for different site year data with reduced (spectral angle mapper based) sampling data set. (a) Site year-1 (Topeka 2021); (b) Site year-2 (Rossville 2021); (c) Site year-3 (Rossville 2022); (d) Site year-4 (Lonsinger 2021); (e) Site year-5 (Fort Collins 2021); (f) Site year-6 (Fort Collins 2022). Total number of samples for each site year is represented with the variable n.

2.4.2 Spatial Pattern of Soil Nitrate Estimates

Spatial pattern analysis investigated spatial variability of measured soil NO_3-N and estimated soil NO_3-N with a reduced soil sampling data set (see Appendix Table J.1 and Appendix

Table J.2). In the analysis of spatial pattern among samples, spherical model was selected as the best-fitting semivariogram model as compared to exponential and gaussian models. The best fitting semivariogram model was subsequently used to map both measured and estimated soil NO_3-N (Figure 2.12). Spatial maps of measured soil NO_3-N and RF based estimated soil NO_3-N, with reduced sampling data set produced with simple kriging method. All soil NO_3-N maps were classified with green to red color scheme, where green represents higher values of soil NO_3-N, and red side represents lower soil NO_3-N values (mg/kg). The total number of site year for this study were six. However, spatial pattern analysis of soil NO_3-N was performed only for the first four site years. The reason to discard site year-5 and site year-6 (Fort Collins study sites) data was attributed to a poor prediction occurred from unrealistic spatial discontinuity attributed to imposed fertilizer experiments in previous years.

Spatial structure of soil NO_3-N were investigated with nugget to sill ratio. Nugget to sill ratio less than 25 % reflects a strong spatial dependence. Soil NO_3-N variable considered to have a moderate spatial dependence when nugget to sill ratio is between 25 % and 75 % (Yanl et al., 2007). A weaker spatial dependence of soil NO_3-N occurs if nugget to sill ratio is more than 75 %. This spatial dependence classification approach has been used by various studies (Yanl et al., 2007). This study finds that semivariogram indicated a spatial dependence of high to moderate class for measured and RF estimated soil NO_3-N. Nugget to sill ratio of 11% were characterized for both site year-2 and site year-4. A moderate spatial dependence of 30% and 50% exhibited for site year-1 and site year-3, respectively. RF based estimated soil NO_3-N with reduced sampling data set has a nugget to sill ratio of less than 25%, which reflects strong spatial dependence for all four site years. The presence of strong spatial dependence during specific site years where corn was cultivated as previous year crop. Moderate spatial dependence was correlated to site years where soybean was cultivated as previous crop. Spatial dependences were correlated with previous crop rotation and management practices. Soil NO_3-N concentration increases after soybean harvest and relatively lower residual soil NO_3-N after corn harvest. Soybean as a leguminous crop, it fixes atmospheric nitrogen and elevate soil nitrogen concentration. Additionally, soybean has lower nitrogen removal rates compared to corn, as a result, soybean cultivation increases soil residual soil NO_3-N (Hirsh & Weil, 2019; Villarini et al., 2016). A study by Smith et al., (2022) finds a strong spatial dependence of soil N with a low soil N concentration. Shorter spatial autocorrelation range of 92 m and longer spatial range autocorrelation of 277 m were reported for site year-1 and site year-2. In comparison, spatial range were shorter for RF based estimated soil NO_3-N with reduced sampling data set, spatial range autocorrelation was 85 m for site year-1, and 261 m site year-2. Spatial structure analysis proves that there was strong spatial dependence for RF based estimated soil NO_3-N regardless of reduced soil data set. This indicates a degree of correlation of RF based estimated soil NO_3-N concentrations over spatial scale.

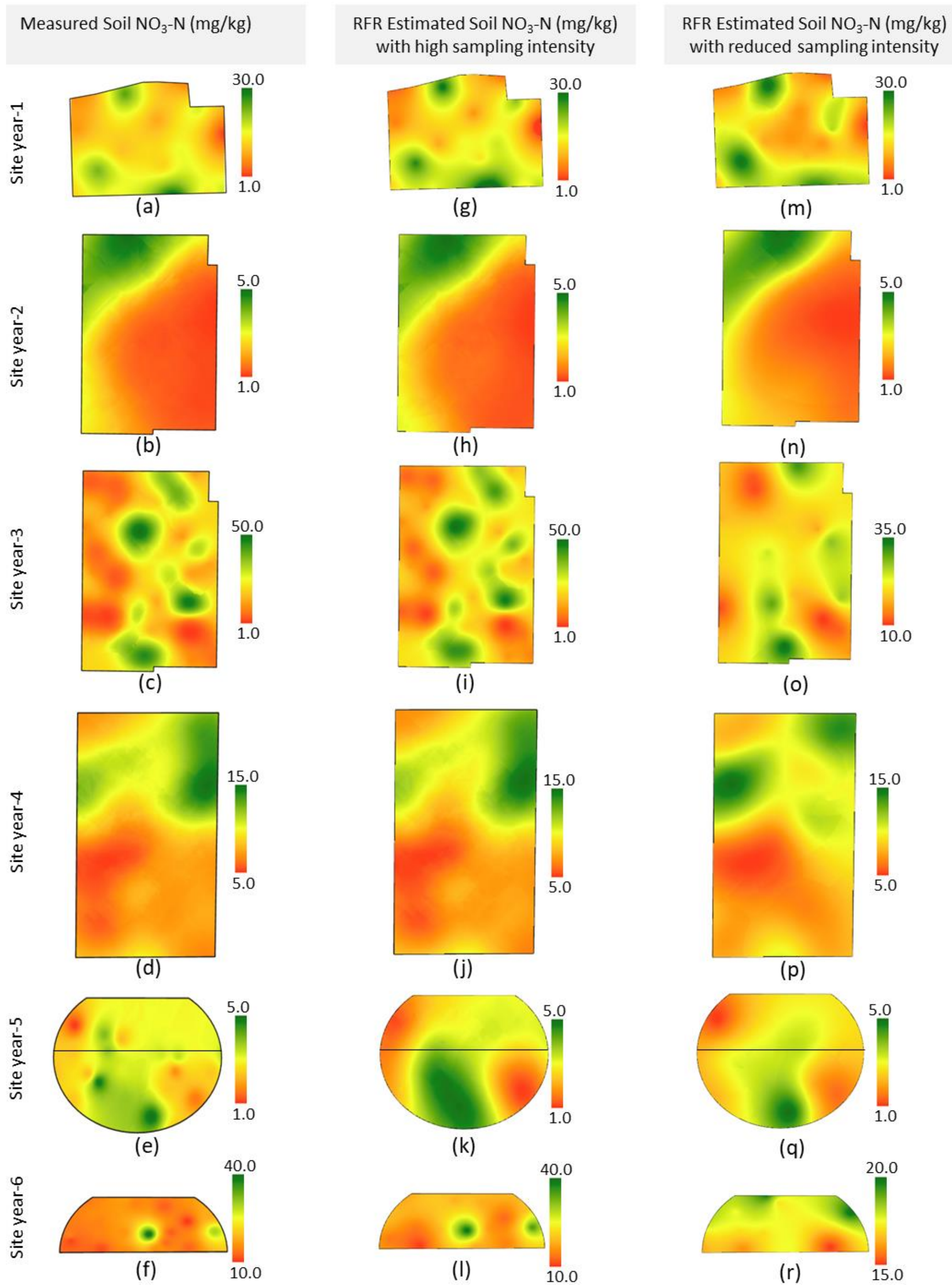


Figure 2.12. Geostatistical modeling results for six site years (a-f) measured soil NO_3-N; (g-l) random forest regression estimated soil NO_3-N with high sampling data set; and (m-r) random forest regression estimated soil NO_3-N maps with reduced sampling data set data.

The soil NO_3-N spatial variability with high (Figure 2.12 g) and reduced data set (Figure 2.12 m) data were apparently showing a similar spatial trend with measured soil NO_3-N maps (Figure 2.12 a) for Site year-1. For site year-1, the regions along the north and southwest (two green spots) boundary of the field with high measured soil NO_3-N for site year-1 have a similar pattern for soil NO_3-N estimated under high and reduced soil sampling data set. For site year-2, high estimated soil NO_3-N were observed in the northwest part of the field, while the lowest soil NO_3-N were on the east and southeast boundary (Figure 2.12 b, h, and n). Under site year-3, estimated soil NO_3-N shows a shrinking of high and low soil NO_3-N regions with reduced soil sampling data set, while the high soil data set map had a similar spatial pattern with the measured NO_3-N. The reduced soil data set-based estimation reduced the soil NO_3-N value range across with high and low values, as a result the range was reduced to 35 mg kg^{-1} to 10 mg kg^{-1} (Figure 2.12 c and o). Similar results were also observed for site year-6. High and reduced sampling data set based estimated soil NO_3-N spatial distribution patterns were consistent when compared to measured soil NO3-N maps.

2.5 Conclusion and Perspectives

This study demonstrated the potential of estimating surface soil NO_3-N with non-imaging hyperspectral data and reducing sampling data set at the same time without compromising the accuracy of soil NO_3-N estimation. To find effective spectral wavelengths particularly sensitive to change in soil NO_3-N, this study first demonstrated an experimental method by adding varying concentrations of $NaNO_3$ in range of 62 and 620 ppm. These concentrations are in range of NO_3 commonly available in soil and which makes it practical to detect NO_3-N sensitive wavelength windows. Non-imaging hyperspectral data with ASD FieldSpec3 has a spectral range from 350 nm to 2500 nm and manages to estimate soil NO_3-N under high and reduced soil sampling data set. Considering the regression model, RFR resulted to the best model in estimating soil NO_3-N under high and reduced soil sampling data set compared to PLSR. Spectral matching techniques, spectral angle mapper (SAM) was adopted to reduce the soil sampling data set of all study site years in a considerable manner.

The reduction in soil sampling data set was solely determined through the comparison of spectral data between samples collected from different locations. The proposed method did not account for spatial context, which resulted in an unconstrained selection of reduced sampling locations and lag distances. Introducing a spatially constrained spectral angle threshold scheme could prove beneficial in limiting the reduction of sampling locations. Furthermore, the reduced samples indicated a narrow range of soil NO_3-N values in a few cases. Hence, a thorough investigation into a robust spectral matching approach is warranted. Nonetheless, employing hyperspectral sensors for surface soil NO_3-N estimation generation offers a swifter and more convenient approach to discerning variable nitrogen application needs for growers. In contrast to the conventional method, where growers physically send soil samples to a lab for analysis, and then create a prescription, reflectance spectroscopy emerges as a promising non-destructive and expeditious alternative for producing prescription maps.

Chapter 3 - Optimal Soil Sampling Design using Hybrid cLHS-SSPOR Technique

3.1 Abstract

Soil sampling design is the first step to make an informed decision on soil and crop management. Most often commercial soil sampling design uses a simplified random and/or stratified design without consideration of geographic and feature space optimization of ancillary data. With the conditioned Latin Hypercube Sampling (cLHS) technique, ancillary data (e.g., proximal soil data) have been utilized to design soil sampling strategy by optimizing solely feature space. However, determining the optimal number of samples and their geo-locations, i.e., the sampling site configuration, remains a central challenge in digital soil mapping. The objectives of this study were: (1) to determine optimal soil sampling size with proximal soil data, and crop grain yield information; and (2) to configure soil sampling sites, i.e., spatial arrangement of sampling sites; and (3) to validate spatial accountability of the optimized soil sampling designs. The study was conducted on data collected over two years (2021-2022) at Kansas River Valley (KRV) Topeka and Rossville. A Veris-MSP3 soil scanner, and a combine harvester equipped with GPS were used for data collection. Kullback-Leibler Divergence (KL divergence) was used as a statistical quantifier metric to determine the optimal sample size based on the similarities between population and sample size. This study used cLHS and hybrid cLHS-SSPOR to configure soil sampling sites. Ripley's K function was used to discern patterns of clustering or dispersion of soil sampling locations identified using proximal soil and proximal soil + yield covariates. Soil sampling design generated with cLHS and hybrid cLHS-SSPOR techniques were validated with quantitate comparison of sampled soil data points and all possible sampling data points (population). Soil sampling size determined with proximal soil covariates were 15 and 20,

respectively for site-year-1 and site year-2. Inclusion of crop yield covariates optimized soil sample size to 11 and 19 sites for site year-1 and site year-2, respectively. The addition of yield data as covariate provided a better perspective in capturing the feature space variability, which resulted in reduction of soil sample size. The observed K values for site year-2 shows that there was no clustering of soil sampling configuration using cLHS method using proximal soil covariates. Clustering of configured sites occurred at less than 10 meters with inclusion of yield covariates. For hybrid cLHS-SSPOR based point pattern analysis of sampling sites, the observed K values for proximal soil and proximal soil + yield resulted in a strong clustering effect at short distance, < 10 meters. The cLHS and hybrid cLHS-SSPOR site configuration methods were improved due to addition of crop yield covariates. Soil sampling design with proximal soil + yield covariates produce sampling sites clustered at a distance that guarantees spatial dependence of soil properties. Soil sampling design validation reported low KL divergence values of soil parameter space for coupled proximal soil and yield data source regardless of soil sampling design technique. The inclusion of yield covariates in sampling design improved representative sample sites configuration. The findings of this study emphasize the importance of tailored soil sampling design in relation to unique soil spatial properties.

3.2 Introduction

Soil sampling design is the first step to make an informed decision on soil and crop management with site-specific management units (SSMUs). Soil sampling is labor-intensive, time consuming and cost prohibitive (Koch et al., 2004). Devising an optimal soil sampling plan is a daunting problem if the objective is to account for minimal number of soil samples while providing maximal information on the spatial variability. Optimized soil sampling is crucial to create reliable digital soil maps (Zhang et al., 2022). Soil sampling design strategies become more significant for SSMUs, where the focus is to develop multiple zones with similar soil and crop management needs. Properly developed soil sampling has the potential to cover variability of soil properties (Shaner et al., 2008). Within the field spatial variability of soil properties needs to be assessed, as it is the foundation for tailored soil and crop management decisions making.

In the past, traditional soil sampling designs have been adopted to collect soil samples, treating fields as having homogenous areas with similar soil fertility. Random sampling, grid sampling, transect sampling, nested soil sampling, and random stratified sampling were tested to assess spatial variability of soil properties (Shaner et al., 2008; Khosla et al., 2010). SSMUs and soil maps developed based on traditional soil sampling could result inaccurate application of N fertilizers if it is limited to characterize soil spatial variability. Traditional soil sampling methods consider cost of soil sampling collection and processing. However, an effective Soil sampling design provides a soil sampling size and collection site that can accurately reflect the nutrient spatial variability. Another challenge to the traditional soil sampling could be high soil variability within the field. This affects the effectiveness of soil sampling strategies and limits it to an assumption that soil analysis assessment hinges to the extent of within field spatial variability of soil properties.

An optimized soil sampling design is crucial for a high spatial variability of soil properties. The essence of soil sampling statistics is based on the understanding of how to select the number of samples that accurately represent the whole population. High density soil sampling is recommended to capture the spatial variability of soil properties. Cost of soil sampling and processing needs to be considered as main factor to develop soil sampling design. However, low-density soil sampling could reduce the cost of soil sampling while compromising the accuracy of characterizing soil spatial properties distribution.

Soil sampling design is expected to incorporate geographic and/or feature space optimization. Feature space refers to range of soil variables of measured or observed well-represented in the soil samples collected. The representation effectiveness is expressed if an accurate estimation of regression parameters occurred. Geographical distribution of soil sampling points across the interest area expressed spatial coverage. Geographic space coverage ensures the spread-out sampling sites across the target area, which could also result to accurate interpolated soil maps.

In a recent review of soil sampling strategies, it was noted that 40% of extension sources recommended the adoption of grid sampling, for site-specific agricultural management (Lawrence et al., 2020). Grid soil sampling often employs a design-based methodology for its simplicity, cost-effectiveness, and objectivity. One such design-based approach involves utilizing classical statistical concepts like randomization and stratification (de Gruijter et al., 2006). In this method, inference is primarily reliant on the sampling design and the independence of sampling locations introduced by the random selection of sampling units. It does not consider the spatial independence of-underlying phenomenon. In instances where multiple data layers are available for determining the best grid soil sampling design, leveraging Latin Hypercube Sampling (Minasny and McBratney, 2006) can offer advantages for stratifying the feature space. The Conditioned Latin Hypercube Sampling (cLHS) algorithm is a stratified random procedure that provides sampling design based on multiple variables and their distribution. The technique optimizes the n sites to be included in the sample, such that the multivariate distribution of the ancillary data is maximally stratified. For grid soil sampling, the variables of interest may include soil data, remote sensing imagery, yield monitor measurements, or other ancillary data to efficiently stratify using traditional methods.

The cLHS technique aims to maximize the stratification of the feature space in the sample, and yet it may produce an uneven coverage of the geographical space, overall resulting in a low estimation quality (Israeli et al., 2019). The method is proficient at describing overall statistical features and determining the optimal quantity of soil samples to collect. However, it may not effectively characterize changes in object attributes concerning spatial location, essential for deciding where to conduct soil sampling. Spatially arranged sampling locations can be expressed as site configuration. Nonetheless, prediction accuracy may be improved by covering both the feature and the geographical spaces in a sample (Hengl, 2007). A similar issue has been addressed while determining sparse sensor locations concerning environmental covariates (Kelp et al., 2022; Saito et al., 2021). The primary goal was to deploy sensors across extensive areas of the field to characterize as many properties of interest as feasible. Therefore, to determine the most suitable sensor positions across a field involved leveraging the Sensor Configuration Optimization (SCO) method. Manohar et al. (2018) introduced a greedy optimization algorithm that utilizes QR pivoting decomposition of the reduced-order basis (e.g., proper orthogonal decomposition - POD modes) of input variables. QR technique decomposes a given matrix into two Q and R components. This QR pivoting approach acts as a “greedy” selection algorithm, demonstrating computational efficiency in identifying sensor locations that are near-optimal. Greedy approaches are often favored over other optimization techniques, given that achieving the true optimal solution frequently involves combinatorial intractable optimization. Considering the high correlation between crop yield and physical properties and chemical nutrients in soil, integrating crop yield data into soil sampling plans can provide valuable insights for optimizing crop nutrient management (Lark, 2016). Despite these advancements, the role of crop yield data in soil sampling strategies remains relatively unexplored. The study presents an opportunity to assess the effectiveness of yield data layers from multiple years, in conjunction with proximal data layers (such as apparent electrical conductivity (ECa), organic matter (OM%), and Cation exchange capacity (CEC)) in optimizing soil sampling design.

3.2.1 Hypothesis

The study hypothesized that coupling multiple years of grain yield data with proximal soil data could optimize soil sampling design.

3.2.2 Objectives

i. To determine optimal soil sampling size with proximal soil data, and crop grain yield information.
ii. To configure soil sampling sites, i.e., spatial arrangement of sampling sites, and
iii. To validate spatial accountability of the optimized soil sampling designs.

3.3 Materials and Methods

3.3.1 Study area

This study was conducted over two years (2021 and 2022) across two study areas in Kansas. The sites are located across Kansas River Valley (KRV) experiment fields at Topeka (39° 4’ 34.04"N, 95° 46’ 11.98"W) and Rossville (39° 7’ 6.36"N, 95° 55’ 37.35"W) (Figure 3.1). Studies conducted at KRV Topeka experiment field were designated as ‘Site Year-1’. Similarly, ‘Site Year-2’ was labelled for studies conducted at KRV Rossville experiment field. Site year-1 covers an area of 3.36 ha, while site year-2 acreage is 7.76 ha. However, SSPOR preferred a regular-shaped study area to avoid the complex mathematical model and algorithms that account for irregularities in shape. Therefore, study areas were clipped to a regular rectangle shape for both study sites, region of 2.01 ha and 5.35 ha were clipped from site year-1 and site year-2 site, respectively.

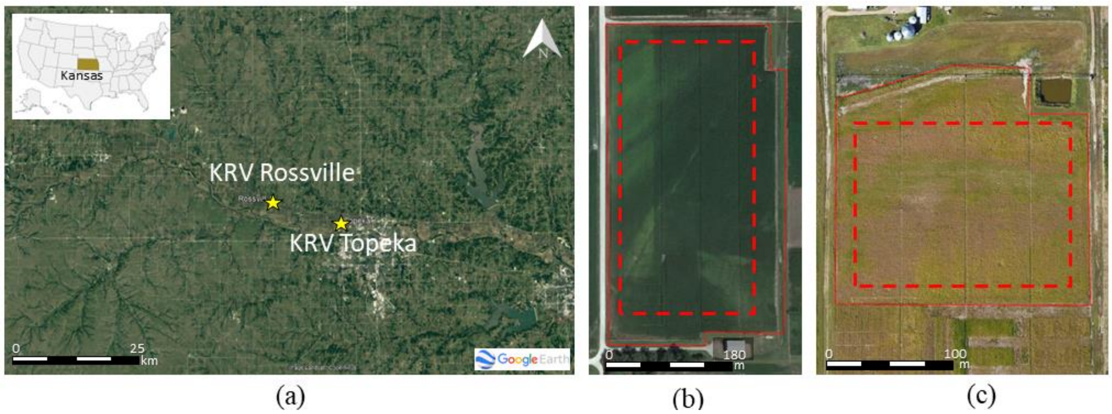
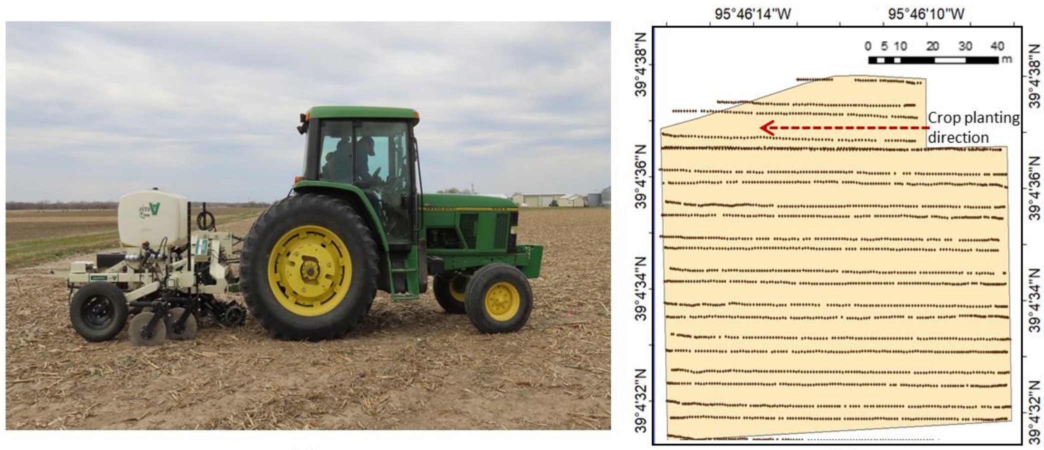


Figure 3.1. Study site locations in Kansas. (a) Kansas River Valley Topeka and Rossville stations highlighted as yellow stars over Landsat composite image (Courtesy: Google Earth); (b) Site year-2 (Kansas River Valley Rossville experiment field), and (c) Site year-1(Kansas River Valley Topeka) experiment field highlighted with red boundary. Dotted red boundaries indicate clipped study regions.

3.3.2 Proximal Soil Data Collection

A proximal soil sensor, Veris-MSP3 (Veris Technologies, Inc., Salina, KS, USA) was used to map geo-referenced soil properties. The Trimble *Real-time kinematic* positioning (RTK) system equipped, John Deere 5055E tractor connected with Veris-MSP3 system to acquire apparent Electrical conductivity at depth of 0-30 cm (ECa shallow), apparent Electrical conductivity at depth of 0-90 cm (ECa deep) and organic matter (OM) (Figure 3.2 a). The Veris-MSP3 data were collected at site year-1 and site year-2 sites on 29 March and 31 March in 2021, respectively.



(a) (b)

Figure 3.2. Veris-MSP3 mapping unit (a) Veris-MSP3 scanning soil to collect ECa shallow and deep (with disk coulter), and OM (with optical sensor) at site year-1; (b) field data collection scheme overlaid on field boundary map.

3.3.3 Soil Sample Collection and Analysis

Soil samples were collected in 2023 from both study sites for validation of the optimized soil sampling designs. Soil sampling locations in the field were identified with a stratified random sampling design method. Soil sampling locations were recorded with a differential global positioning system (DGPS). Composite soil samples were collected from each location per 0.101 ha to 0.202 ha grid cell size for Topeka and Rossville. Soil samples were collected from 0 to 30 cm depth. The collected soil samples were air-dried, grounded, and passed through a 2-mm sieve. The samples were sent to a commercial lab, AgSource Laboratories in Lincoln, Nebraska, USA for analysis. (Complete list of soil parameters analyzed and measurement techniques used are presented in Appendix Table F.1).

3.3.4 Crop Yield Data Collection

Corn (*Zea mays* L) and soybean, (*Glycine max* L) grain yield data were collected with a CASE IH Axial-Flow 2344 (Case IH, Racine, WI) combine harvester equipped with a DGPS receiver and yield monitor. Table 3.1 shows the harvesting date, average moisture, and estimated dry yield. The SMS AgLeader and ArcGIS Pro 3.1.1 software were used to process the raw harvest data and corrected in accordance with guidelines presented Sudduth and Drummond, 2007.

Table 3.1. Summarized yield monitor data for all site years.

Year	Study site	Crop	Area (ha)	Harvesting date	Estimated average dry yield (t/ha)
2021	Site year-1	Corn	3.36	15-Sep	14.91
	Site year-2	Soybeans	7.73	18-Oct	3.78
2022	Site year-1	Soybeans	3.36	10-Oct	3.62
	Site year-2	Corn	7.73	27-Sep	14.64

3.3.5 Data Processing and Statistical Analysis

The schematic workflow in Figure 3.3 shows the data preparation and statistical analysis for developing soil sampling design.

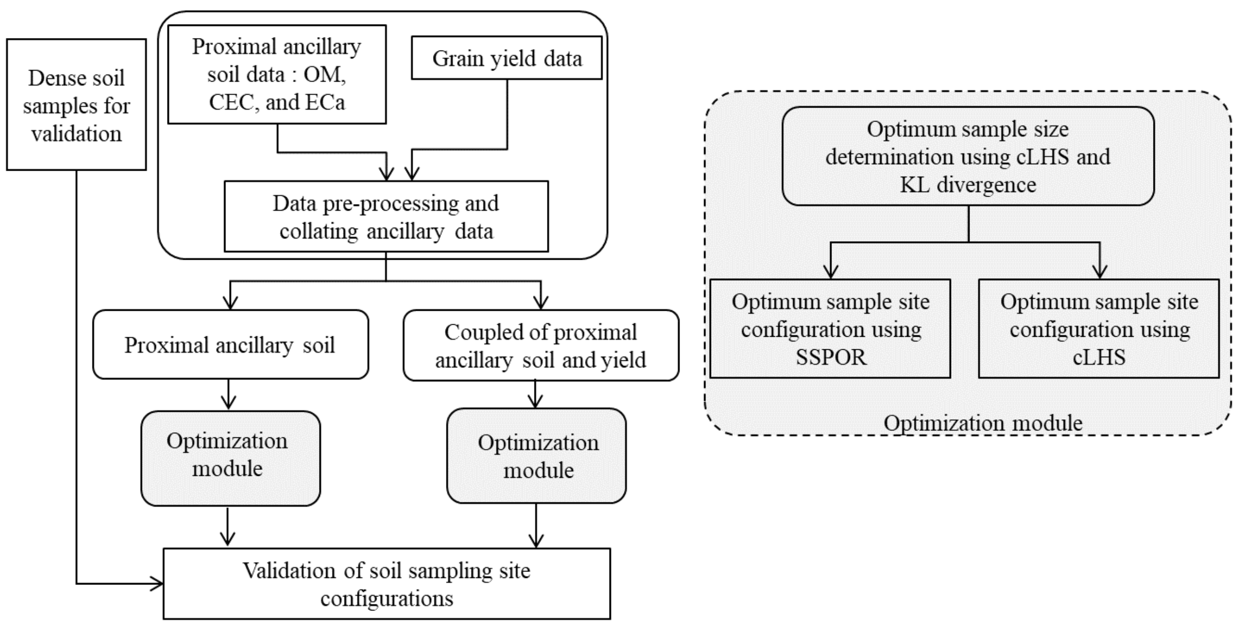


Figure 3.3. Schematic workflow for data preparation and statistical analysis to optimize design of soil sampling strategy using multiple ancillary data sources. The optimization module is highlighted in grey shades.

Sparse Sensor Placement Optimization for Reconstruction (SSPOR) and conditioned Latin hypercube sampling (cLHS) are two approaches used to identify an optimal distribution of soil sampling sites. Proximal soil data, ECa shallow, ECa deep, OM and CEC (in point vector files) were interpolated with kriging technique to estimate soil information at unsampled locations. The interpolation was performed with ArcGIS pro 3.1.1 and RStudio. ArcGIS software conducted a cross-validation process based on the assumption of a missing value, interpolation was performed using each sampling location and compared with measured values. The interpolated proximal soil data and crop grain yield raster maps were resampled to a 3 m pixel size. The resampling pixel size was selected to inference composite soil sampling region.

For the cLHS, data preparation started with constructing a quantile matrix for covariate data. The feature space of the covariate data was divided into 25 bins to represent the entire range of all soil and yield values (Malone et al., 2019). A covariate data hypercube was created through iterating all data points to determine the quantile bins they belong for each covariate. The covariate data hypercube consisted of a multidimensional grid for each cell with different combinations of quantiles for all variables. It resulted in structured stratified spread data with a unique combination of quantile bins for all variables to ensure cLHS has a spread-out sampling site. This study used R package ‘cseq’ to define the sequence of sample sizes. Sample sizes were set up in a sequence of numbers from 5 to 100 in steps of 5 sample sizes, in increments. Each sample size was iterated 10 times to ensure sample representativeness in capturing full dataset. The ‘clhs’ R function was used to perform cLHS using a multivariate dataset (both proximal soil and yield) to generate the soil sample size.

3.3.6 Soil Sample Size Determination

The cLHS technique (see Sec. 3.3.7). was adopted to determine soil sample size from empirical distribution function. A statistical quantifier metric is essential to determine similarities between population and sample size. This study used Kullback-Leibler Divergence (KL divergence) metrics (a detailed formulation is presented in Sec. 3.3.8) as a function of soil sample size to determine the optimal sample size (Malone et al., 2019). In an iterative approach, soil sample sizes (sub-sample) were reduced from original samples (population), and corresponding KL divergence values were obtained. The optimal sample size was determined when plotted KL divergence metrics stabilized and showed a minimal change in value (See Sec. 3.4.1). A Python script was used to identify the "elbow point" from the second derivatives of KL divergence metric values. Elbow point is where the trend changes significantly which indicates the point of optimal soil sampling size (Saurette et al., 2022a; Sood et al., 2022; Yan et al., 2023). The sample size is identified where the second derivatives change sign from positive to negative or reach zero.

3.3.7 Conditioned Latin Hypercube Sampling (cLHS)

The cLHS aims to locate samples geographically in a manner that replicates the empirical distribution functions of ancillary information pertaining to the samples (Clifford et al., 2014). Unlike Latin hypercube sampling, the cLHS considers a joint distribution of variables (k number of variables) in addition to the normal distribution of variables (Minasny & McBratney, 2006b). The cLHS optimizes x, which is a sub-sample of size n, given N ($n<N$) sites with ancillary variables (X) in order that x forms a Latin hypercube, or the multivariate distribution of X is maximally stratified. The cLHS methodology as introduced by Minasny and McBratney, 2006 is presented in the following steps:

Step 1- Stratification and Sampling: The ancillary data X (with size N×k) were divided into n equally probable strata for each variable. It can be noted that n also indicates the optimum number of sampling sites (generated using KL divergence comparison as illustrated in Sec. 3.3.8). This stratification was based on the quantile distribution of each variable in X, with the quantiles calculated as $q_{i,\,j},\ldots,\ q_{n+1,j}$, for j=1,…, k, and i = 1,…, n+1. A sub-sample x (size n×k) was then selected from X. A matrix Z was defined which counts the number of x that fall into each of the defined stratum:

$$Z = \begin{bmatrix} Z_{11} & \cdots & Z_{1k} \\ \vdots & \ddots & \vdots \\ Z_{n1} & \dots & Z_{nk} \end{bmatrix}_{n\times k} \qquad (3.1)$$

A true LHS would have values of 1 for all cells of matrix Z. The challenge of cLHS lies with finding x that has the value of Z close or equal to 1. Additionally, the correlation matrix C was also calculated for X.

Step 2- Random sampling of n: Out of N population, n random samples were generated. The x (i,…, n) indicated sampled sited, subsequently r (i,…, N-n) became unsampled sites. Additionally, the correlation matrix T was also calculated for x.

Step 3- Objective Functions: As the samples' attributes within each variable (k) were randomly selected and combined to generate new feature space (x), samples satisfying the feature vector might not exist in the population. (Minasny & McBratney, 2006a) proposed the cLHS based on the LHS by redesigning the objective function and tailored optimization algorithm. The algorithm introduced several objective functions to ensure comprehensive sampling. The global objective function O was adapted as presented in Eq. (3.2).

$$O = w_1 O_1 + w_2 O_2 \tag{3.2}$$

O1 and O2 represent the sub-objective function for continuous variables, and the correlation matrix difference between samples and the population. The w1 and w2 are the corresponding weights of each sub-function, which are generally set to 1. The sub-objectives were:

$$O1 = \sum_{i=1}^{n} \sum_{j=1}^{k} \left| Z\left(q_{i,j} \leq x_j \leq q_{i+1,j}\right) - 1 \right| , \tag{3.3}$$

$$O2 = \sum_{i=1}^{k} \sum_{j=1}^{k} |c_{ij} - t_{ij}|, \tag{3.4}$$

Where, Z counts the number of x_j that fall in between quantiles, $q_{i,j}$ and $q_{i+1,j}$. The terms c_{ij} and t_{ij} are elements of the correlation matrices C and T, respectively.

Step 4- Optimization and sampling adjustments: The global objective function was optimized using an annealing schedule based on the Metropolis algorithm. In an iterative process, this algorithm picks a sample randomly from *x* and swap it with a random site from unsampled site *r*. Based on criteria set, it also removes sample(s) from x which has the largest $Z\left(q_{i,j} \leq x_j \leq q_{i+1,j}\right)$, and replace it with random site(s) from r. All steps from 2 were iteratively repeated until the objective function value meet a predetermined stopping criterion or a defined number of iterations. The number of iterations were selected as 10000 (Malone et al., 2019).

3.3.8 Kullback-Leibler Divergence

The divergence of one probability distribution from another is measured in terms of Kullback-Leibler divergence, which originated from the concept of information theory (Kullback and Leibler, 1951). A dissimilarity measurement was made between population and sample distributions (Wan et al., 2021). The optimal sample sizes were determined with the KL divergence that indicates the smallest deviation from the overall data distribution (Eq. (3.5)). To achieve this, the probability of sampling points was calculated as shown below. The divergence was computed with 'KL.empirical' function in R-package.

$$KL(P||\,Q) = \sum_{i=1}^{n} P(x_i) \log \frac{P(x_i)}{Q(x_i)} \tag{3.5}$$

Where:

$P(x_i)$ is original probability distribution of the ancillary data for the entire field,

$Q(x_i)$ is sample probability distribution of the ancillary data at locations given by a candidate sample,

KL (P || Q) is the Kullback-Leibler divergence between the two probability distributions.

3.3.9 Optimum Sampling Site Configuration

In this study, optimized sampling locations were identified using the Sparse Sensor Placement Optimization for Reconstruction (SSPOR) approach proposed by (Manohar et al., 2018). It utilizes signal reconstruction based on a tailored library of features extracted from training data (Figure 3.4). Optimized sparse sampling locations were identified using the greedy optimizations based on QR factorization and Proper Orthogonal Decomposition (POD).

For the given data composite, it can be represented by 2-dimensional system with m measurements (e.g., $Yield_{year1}$, $Yield_{year2}$), i.e., X is an observation data matrix. The observation matrix was transformed to POD subspace as linear combinations of orthonormal eigenmodes ψ_k (POD modes) that define a low-dimensional embedding space, yielding a reduced representation. A low-dimensional representation of X in terms of POD coefficients a_k can be transformed back to the full state with a linear combination of POD modes as:

$$X = \sum_{k=1}^{m} a_k \psi_k \qquad (3.6)$$

The eigenmodes ψ_k and POD coefficients a_k were obtained from the singular value decomposition (SVD) and data training. The SVD is the optimal least-squares approximation to the data for a given rank m, as demonstrated by the Eckart–Young theorem (Manohar et al., 2018; Shahriari et al., 2014).

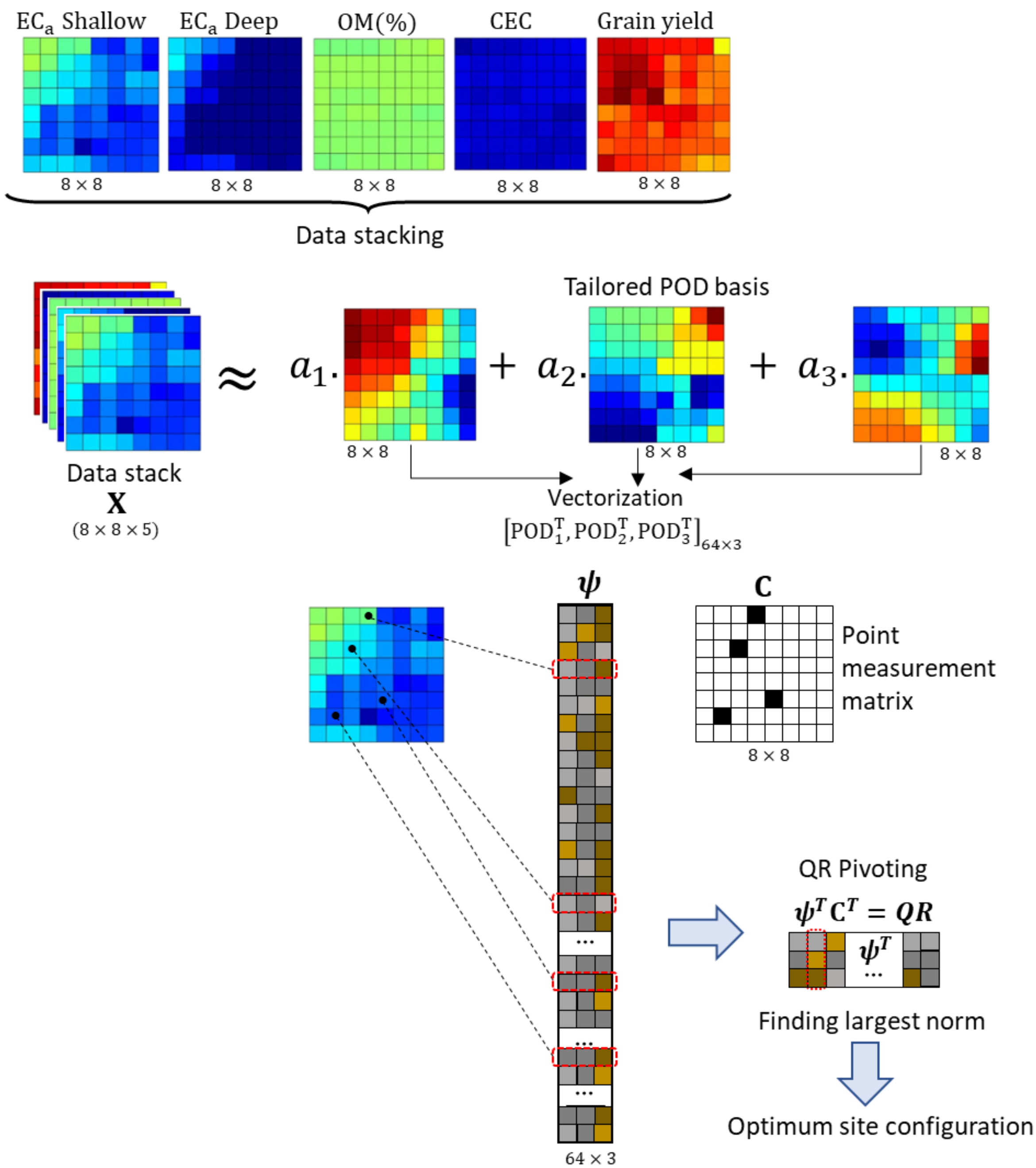


Figure 3.4. Schematic diagram for Sparse Sensor Placement Optimization for Reconstruction based optimum sampling site configuration. For illustration, 4 sites were selected (to be optimized using conditioned Latin Hypercube). The input data raster have dimensions of 8×8, and 5 features (i.e., Eca shallow, Eca deep, OM, CEC, and grain yield). Only three proper orthogonal decomposition bases are used for instance.

The optimization of sampling locations was undertaken with the objective of reconstructing observation data matrix (X) from point measurements with POD bases. Effective locations (n) results in a point measurement matrix C that was optimized to recover the modal mixture from observed vector y. The matrix C was presented as:

$$C = [e_{\gamma 1} \quad e_{\gamma 2} \quad \cdots \quad e_{\gamma n}]^T \quad , \qquad (3.7)$$

where, e_j are canonical basis vectors with a unit entry at index j and zeros elsewhere. The $\gamma \subset \{1, \cdots, n\}$ denotes the index set of sampling locations. The measurement matrix resulted in the linear system as:

$$y_i = \sum_{j=1}^{n} C_{ij} x_j = \sum_{j=1}^{n} C_{ij} \sum_{k=1}^{m} a_k \psi_{jk} \,. \quad (3.8)$$

For each sampling location, the data were reconstructed by approximating the unknown POD basis coefficients a_k with the Moore–Penrose pseudoinverse,

$$a_k = \Theta^{-1} y = (C\psi)^{-1} y \quad . \qquad (3.9)$$

Equivalently, the reconstruction was obtained using:

$$\hat{x} = \psi (C\psi)^{-1} y \qquad . \qquad (3.10)$$

The optimal sampling locations are those that permit the best possible reconstruction $\hat{x}$. Thus, the sampling location problem seeks rows of ψ corresponding to point locations in the space that optimally condition inversion of the matrix Θ. The sum (trace) of determinant of Θ was optimized as:

$$\gamma_* = arg \max_{\gamma} tr(\Theta^T \Theta) \qquad (3.11)$$

As the γ determines the structure of C, representing the sampling locations, thereby impacting the condition numbers of Θ (Manohar et al., 2018). The optimization of this criteria was performed using an approximate greedy solution given by the matrix QR factorization. The

SSPOR approach for optimum sampling locations were performed using python ‘pysensors’ package (de Silva et al., 2021). The number of sampling locations were set as the optimum number of locations derived from cLHS-KL divergence technique (Sec. 3.3.6).

3.3.10 Soil Sampling Sites Configuration with cLHS and Hybrid cLHS-SSPOR

Soil sampling sites were configured using cLHS and hybrid cLHS-SSPOR methods. The spatial coordinates of these sites were overlaid on interpolated maps of organic matter (OM) content and crop yield to visualize their distribution. This overlay was performed for siteyear-1 and site year-2, generated four distinct spatial data configurations for analysis. To discern patterns of clustering or dispersion and compare yield covariates on both methods, Ripley's K function was computed (Brungard & Boettinger, 2010; Y. P. Lin et al., 2014). The Ripley's K-function evaluates feature spatial distribution in relation to Complete Spatial Randomness (CSR). It allows determining whether a point pattern deviates from CSR at a given distance. Ripley's K function was used to compare soil sampling sites obtained by proximal soil and proximal soil + yield covariates using cLHS and hybrid cLHS-SSPOR methods. Both methods were evaluated in their ability to incorporate additional yield covariates. The assumption of CSR has a value of πr^2 for an expected value of K(r). It serves as a baseline for identifying deviations indicative of spatial patterns in observed K(r) in real data sets. At a certain distance r, K(r) from observed data exceeding πr^2 (of CSR) indicate clustering, while lower values suggest regular spacing. Confidence intervals were estimated via Monte Carlo simulations to accommodate spatial randomness, discerning non-random spatial patterns.

3.3.11 Validation of Soil Sampling Design

Sample sizes identified and configured with cLHS and hybrid cLHS-SSPOR techniques were validated with soil samples that were collected and analyzed for Organic Matter, Phosphorus and Potassium (OM, P and K). Raster grids were first generated from point samples of these soil properties. For each property map, centroids of all grid cells were considered as sampling population. These centroids can be used as all possible sampling points. The validation of soil sampling design was conducted with quantitative comparison between sampled data points (from cLHS and hybrid cLHS-SSPOR) and all possible sampling data points (population) on soil property (e.g., OM) grids. To quantitatively assess the similarity between optimized soil sampling site and the overall population, KL divergence metric was implemented. The comparison was critical to evaluate representativeness of the sampling design in addition to covering the whole extent of the field. Effective representative soil sampling design could be evaluated with how well variability in geographic space and feature space were characterized. Results of the KL divergence analysis, represented by the means of divergences provides a complete view of the sampling design's effectiveness. KL divergence value closer to zero indicates that the optimized soil sample design is representative of the overall population distribution.

3.4 Result and Discussion

3.4.1 Sample Size Determination

A cLHS algorithm minimizes the objective function to provide a maximally stratified sampling design in feature space (covariates). All given sample size for cLHS minimizes the objective function and thus represent variability of covariates. This study used KL divergence to compare two empirical distribution function that results an optimal sample size. A monotonic decline in mean KL divergence values with increasing sample size is apparent in Figure 3.5 a. The trend agrees with the premise that larger sample sizes tend to provide a more accurate representation of the population. Nevertheless, the observed stabilization of the KL divergence beyond a certain sample size does not substantially improve spatial soil properties representation. The identification of elbow point to ascertain optimum size was performed with second derivative of mean KL divergence as shown in Figure 3.5 b. The change of sign of the second derivative from positive to negative occurred at 20 sample size for site year-2.

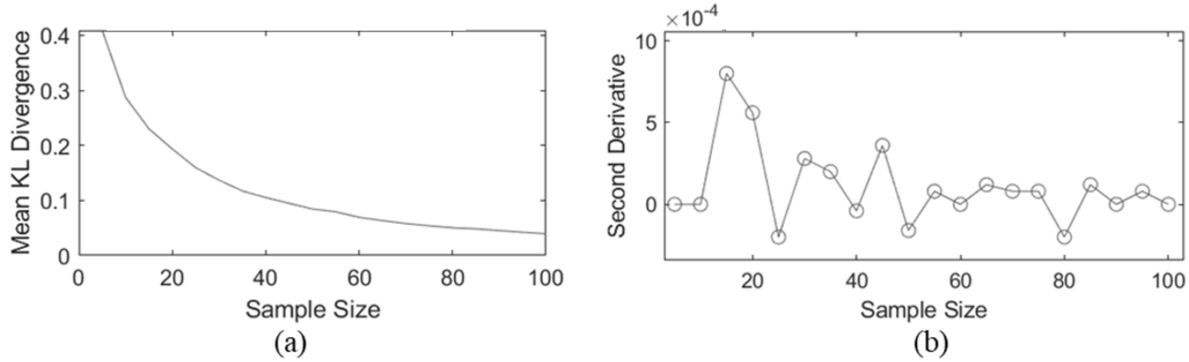


Figure 3.5. Optimum sampling size determination with mean KL divergence. (a) KL divergence identifier as a function of sample size, and (b) second derivative of KL divergence for Site year-2. Proximal soil data solely used for sample size determination.

Sample sizes determined from proximal soil, and proximal soil combined with yield data (proximal soil + yield) are presented in Table 3.2. Inclusion of yield covariates reduced the number of soil sample sizes, for site year-1 and site year-2. For site year-1, soil sampling size reduced from 15 to 11 in numbers and for site year-2, it was reduced from 20 to 19 sample size.

Table 3.2. Optimized soil sampling size for site year-1 and site year-2 with proximal soil and proximal soil + yield data.

Sample	Proximal soil based optimal sample size		Proximal soil + yield based optimal sample size	
	Site year-1	Site year-2	Site year-1	Site year-2
Number of samples	15	20	11	19
Samples ha^{-1}	8	4	6	4

In cLHS, proximal data such as apparent electrical conductivity (ECa) are frequently utilized covariates for soil sampling design (Minasny & McBratney, 2010; K. Schmidt et al., 2014). However, the inclusion of ancillary data, specifically grain yield data in this context resulted a better soil sampling size optimization. Crop yield data, represents the cumulative output of interaction between soil properties and agronomic practices. The addition of yield data as covariate provided a better perspective in capturing the feature space variability, which resulted in reduction of soil sample size. Incorporating crop yield data into proximal soil provide additional attributes in the hypercube that can guide the selection of sampling locations, ensuring selections that areas of distinct or unique conditions from new set of feature spaces. In similar manner, a study by Lin et al., (2011) introduced a covariates of vegetation index data that optimized soil sampling size. Optimized sampling size with covariate representation reduces soil sampling collection and analysis costs (Minasny & McBratney, 2010). The determined soil sample sizes ranged from 4 to 6 samples ha^{-1} are sufficient to generate soil fertility maps. A study by Vašát et al. (2012) reported that a single soil sample collected ha^{-1} was enough to develop a quality interpolated soil fertility map. In contrast, grid soil sampling with 1 sample per ha resulted in a poor soil fertility map due to low prediction efficiencies (Mueller et al., 2001). For soil total carbon, a significant change in spatial distribution was not noticed when soil sample size for interpolation increased beyond 4 to 6 samples ha^{-1} (Saurette et al., 2022b). Soil sampling design include both sampling size and site

(Heuvelink et al., 2006), and the optimization may benefit from considering both the geographic space and feature space (Hengl, 2007). Therefore, this study further optimized soil sampling design including configuration of sampling sites with cLHS and hybrid cLHS-SSPOR techniques (See Sec. 3.4.3).

3.4.2 Soil Sampling Site Configuration with cLHS and Hybrid cLHS-SSPOR

The spatial distribution of soil samples sites generated using cLHS optimization at two distinct sites, site year-1 and site year-2, is shown in Figure 3.6 and Figure 3.7, respectively. The soil sampling sites (black dots) were overlaid on OM and crop yield interpolated spatial maps for both study sites. For the site year-1 and site year-2, the distribution of sampling sites generated with proximal soil covariates were uniformly scattered across the field (Figure 3.6 and Figure 3.7). The qualitative analysis indicated that the cLHS optimization technique has identified representative sampling sites for both proximal soil and proximal soil+yield data with a visible stratification of the feature space. However, cLHS does not take into consideration the geographic space variability of sampling sites.

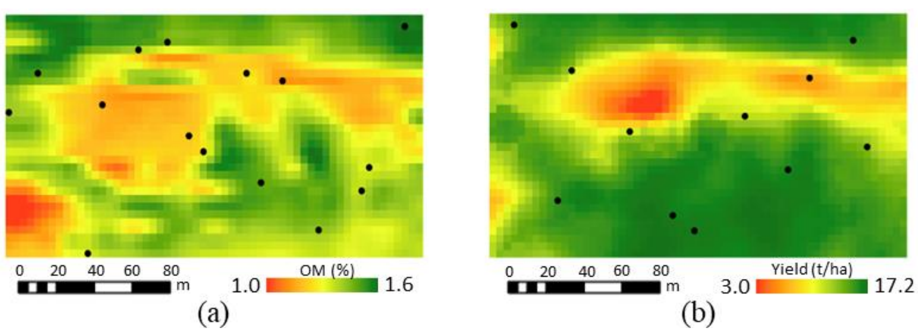


Figure 3.6. Soil sample site configuration using conditioned Latin Hypercube Sampling with (a) proximal data overlaid on organic matter (OM%) map; and (b) proximal soil+yield data overlaid on 2021 yield map of Kansas River Valley Topeka site. Configured sampling sites are highlighted with black points.

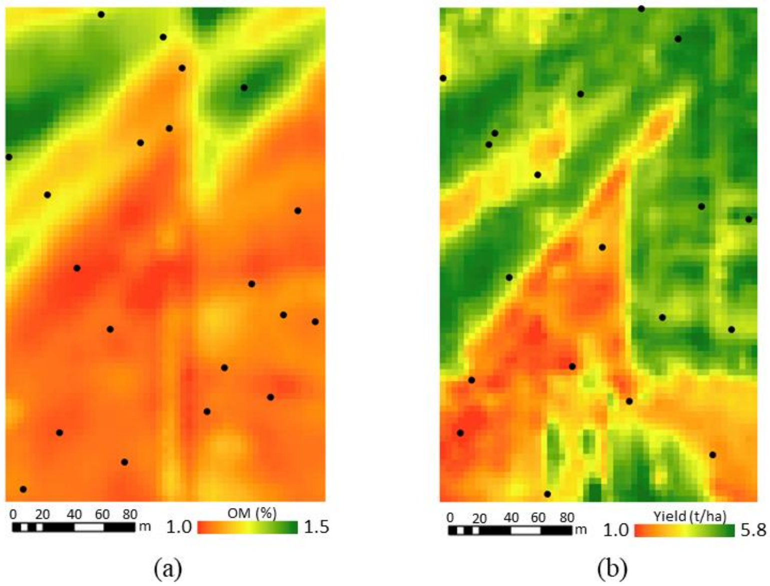


Figure 3.7. Soil sample site configuration using cLHS optimization with (a) proximal soil data overlaid on organic matter (OM%) map; and (b) proximal soil+yield data overlaid on 2021 yield map of Site year-2 site. Sampling sites are highlighted with black points.

For site year-1, with cLHS soil sampling configuration methods, the point pattern analysis of sampling sites showed a clustering at small scale (between to 10 to 15 meters) using both proximal soil and proximal soil + yield covariates. The observed K values at larger scale (more than 15 meter for this case) presented a regular spacing of soil sampling sites. The inclusion of yield covariates on proximal soil shows clustering at similar scale compared to proximal soil covariates sampling configuration (Appendix Figure K.1 a to d). The observed K values shows there was no clustering at all for site year-2 with soil sampling configuration using cLHS method and proximal soil covariates (Appendix Figure K.1 c). However, clustering occurred at less than 10 meters when yield covariates were included (Appendix Figure K.1 d). This implies that cLHS soil sampling site configuration for site year-2 was changed due to inclusion of yield covariates and spatial variability were more at smaller scales. Crop yield exhibits both small- and large-scale spatial dependence based on multiple soil and environmental factors. A soil sampling strategy with a distance closer to spatial dependence of soil properties captures variation in soil properties

(Khosla et al., 2010). Observation of soil sampling sites clustering at a distance of less than 30 meters guarantees that majority of spatial variability being captured with cLHS sampling design that uses proximal soil and proximal soil + yield covariates. Spatial dependence of soil properties ranges between 20 to 30 meters. Inclusion of yield covariates to proximal sensed soil parameters (ECa) showed a more clustered soil sampling configuration (Figure 3.7 b). A visual observation analysis was made to understand distribution of soil sampling site regarding OM and yield interpolated maps.

The spatial pattern of soil sampling sites generated with hybrid cLHS-SSPOR for two site years were presented in Appendix Figure K.1. The observed K values for proximal soil and proximal soil + yield using hybrid cLHS-SSPOR shows a strong clustering effect at short distance, less than 10 meters (Appendix Figure K.1 e to h). The observed K function values surpasses the CSR at shorter distances regardless of inclusion of yield as an additional covariate. Yield is a spatially structured nonrandom covariate; addition of yield covariate introduced more deviation from CSR to a more clustering effect (Appendix Figure K.1 g). The point pattern analysis findings were supported with a visual observation showing an overlaid of configured sampling points on OM and crop yield maps for site year-1 and site year-2 study areas as presented in Figure 3.8 and Figure 3.9, respectively. Sampling sites generated with proximal soil covariates shows more dispersed sites, as illustrated with points overlaid on OM maps (Figure 3.8 a and Figure 3.9 a). On the contrary, sampling sites configured with proximal soil+yield covariates show a clustered sampling point within a short distance.

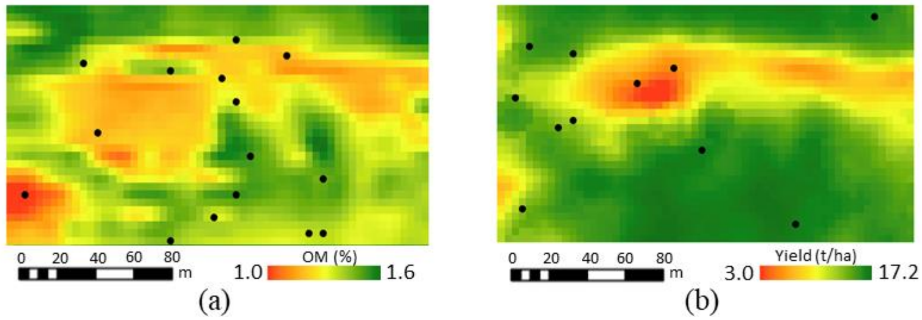


Figure 3.8. Soil sample site configuration using hybrid cLHS-SSPOR optimization with (a) proximal soil data overlaid on organic matter (OM%) map; and (b) proximal soil+yield data overlaid on 2021 yield map of Kanas River Valley Topeka site. Sampling sites are highlighted with black points.

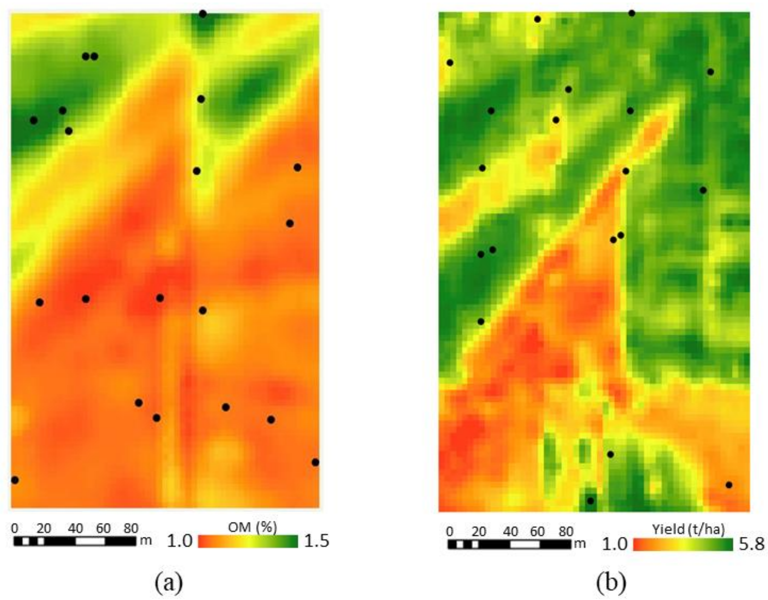


Figure 3.9. Soil sample site configuration using hybrid cLHS-SSPOR optimization with (a) proximal data overlaid on organic matter (OM%) map; and (b) proximal soil+yield data overlaid on 2021 yield map of Kanas River Valley Rossville site. Sampling sites are highlighted with black points.

The cLHS technique employs a strategy of generating sampling sites randomly within each stratum based on a multi-variate distribution of ancillary data. However, the introduction of independence through random sampling raises concerns about neglecting spatial autocorrelation, especially when the sampling interval exceeds the range. This concern is particularly relevant for cLHS, which focuses solely on attribute feature space, omitting consideration of geographic space (Wan et al., 2021). The "spatial statistic trinity" concept highlights that an effective sampling design is attainable when optimization is well-aligned with the properties of the entire population (J. Wang et al., 2020). Incorporating diverse information sources during sampling design not only maximizes information gain but also enhances representativeness (Mulder et al., 2013). In contrast to cLHS, the SSPOR-based approach employs a grid/pixel-based learning opportunity (reconstruction method), offering a more nuanced understanding of spatial relationships.

3.4.3 Soil Sampling Design Validation

KL divergence metrics were used to compare cLHS and hybrid cLHS-SSPOR soil sampling design with soil properties (OM, P, and K) as data point. The mean KL divergence result is shown in Table 3.3. The KL divergence values vary across different soil properties and different soil sampling design techniques.

Table 3.3. KL divergence values for cLHS and hybrid cLHS-SPOR optimization based soil sampling design using ancillary proximal soil and yield data. Soil parameter space was derived from validation soil samples collected in 2023.

Site	Optimization method	Data source	Soil parameter space			
			OM	P	K	Average
Site year-1	cLHS	Proximal soil	0.375	0.108	0.393	0.292
		Proximal soil + yield	0.421	0.163	0.266	0.283
	Hybrid cLHS-SSPOR	Proximal soil	0.447	0.255	0.333	0.334
		Proximal soil + yield	0.597	0.222	0.432	0.417
Site year-2	cLHS	Proximal soil	0.299	0.377	0.318	0.331
		Proximal soil + yield	0.258	0.397	0.352	0.335
	Hybrid cLHS-SSPOR	Proximal soil	0.243	0.365	0.359	0.322
		Proximal soil + yield	0.282	0.301	0.313	0.298

The lowest mean KL divergence values of 0.283 for cLHS and 0.298 for hybrid cLHS-SSPOR were resulted in using proximal soil + yield data sources. Low KL divergence indicates higher similarity of probability distribution between point data to reference raster data. Coupling of proximal soil and yield data source was effective in reducing the KL divergence regardless of the soil sampling design technique. This indicates that the inclusion of yield covariates in sampling design improved representative of sample sites configured.

Hybrid cLHS-SSPOR technique-based KL divergence values across different soil properties were not showing a similar trend (as compared to mean KL). For site year-2, soil parameter space indicates a decrease in KL divergence from proximal soil covariates to proximal soil+yield covariates, except for OM. However, for site year-1 soil parameter space indicates an increase in KL divergence from proximal soil data to proximal soil+yield data, except for soil P. Soil parameter space of cLHS indicated a mix of increasing and decreasing trend of KL divergence values from proximal soil covariates to proximal soil+yield covariates for both site year-1 and KRV Roseville. For site year-1, the validation soil sampling design with inclusion of yield covariates resulted a decrease of KL divergence from 0.292 to 0.283 for cLHS and an increase from 0.335 to 0.331 for hybrid cLHS-SSPOR. The addition of yield covariates does not necessarily bring OM samples closer to the reference OM data. Non-uniform sampling sites distributions were exhibited on varying concentration of OM with cLHS and hybrid cLHS-SSPOR (with inclusion of yield) based soil sampling sites (Figure 3.10 a and b). On the other hand, site year-1 shows an increase in KL divergence of soil phosphorus from using proximal soil covariates to proximal soil+yield covariates for cLHS, but a decrease for hybrid cLHS-SSPOR. The hybrid cLHS-SSPOR did not show a consistent KL divergence trend across soil properties. However, it showed a greater consistency in site year-2 when incorporating proximal soil + yield covariates. This highlights importance of tailored soil sampling design in relation to unique soil spatial properties. Hybrid cLHS-SSPOR technique with proximal soil + yield data source was effective in designing soil sampling design that captures both geographical space and feature space for site year-2 study area. On the contrary, cLHS effectively managed to design a soil sampling design for site year-1 using proximal soil+yield data. It suggested the necessity of customizing soil sampling designs to specific study sites. Future research is required to develop refined sampling design techniques that can further reduce the KL divergence and thus improve the representativeness of soil samples.

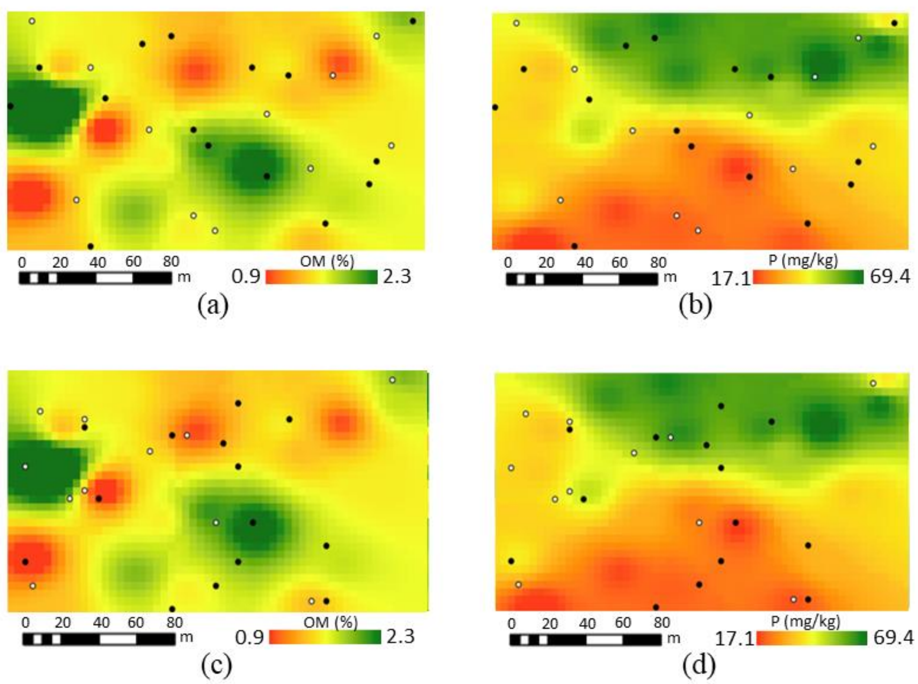


Figure 3.10. Soil sample site configuration using cLHS optimization laid on (a) organic matter (OM%) map; (b) Phosphorus map (mg kg-1); and hybrid cLHS-SSPOR optimization laid on (c) organic matter (OM%); and (d) Phosphorus map (mg kg-1). Sampling sites with proximal soil covariates are highlighted with black points. Sampling sites with proximal soil + yield covariates are highlighted with white points.

3.5 Conclusion and Perspectives

Initially, the soil sample sizes were 15 and 20 samples for site year-1 and site year-2 study areas, respectively. The cLHS technique reduced soil sampling size by integrating crop yield data. site year-1 and site year-2 require a minimum of optimized 11 and 19 soil sampling sizes. The optimal sample sizes determined for site year-1 and site year-2 sites highlighted the necessity of a tailored soil sample size determination. KL divergence was used to compare the effectiveness of cLHS and hybrid cLHS-SSPOR soil sampling design.

The analysis of configured soil sampling site distributions with cLHS and hybrid cLHS-SSPOR were examined by Ripley's K function. Both methods resulted in a clustering of sampling sites, as proved by Ripley's K values exceeding theoretical CSR. Observed K values from proximal soil covariates resulted in clustering for all site years except cLHS method. Inclusion of yield covariates exhibited a clustering of soil sampling sites for site year-1 and site year-2 under both cLHS and hybrid cLHS-SSPOR methods. This suggests that yield covariates introduced more localized spatial variability compared to solely proximal soil covariates. These specific soil sampling site configuration approaches demonstrated the complicated relationship of proximal soil and yield covariate and their different scale spatial variability.

KL divergence was used to validate cLHS and hybrid cLHS-SSPOR soil sampling design. Lower KL divergence values imply greater similarity between sampled distribution and actual soil property value distribution, indicating more representative sampling design. The low mean KL divergence values obtained with proximal soil+yield covariates suggest a potential benefit of integrating yield covariates into proximal soil covariates for a more effective soil sampling design. This concluded that additional ancillary data increase the effectiveness of soil sampling design. However, for site year-1 soil parameter space indicated an increase in KL divergence from proximal soil data to proximal soil + yield data, except for soil P. This reveals the complexity of applying a uniform sampling technique across different soil properties. The variability in KL divergence values for site year-1 and site year-2 sites underscores the need for site-specific soil sampling design. It underscores the importance of tailoring soil sampling designs to the unique conditions of each study site, considering both geographical and feature space, to enhance the representativeness and accuracy of soil sampling.

KL divergence of hybrid cLHS-SSPOR-based technique indicated its capability in representing local spatial heterogeneity compared to cLHS. However, evaluating the efficacy of the hybrid cLHS-SSPOR method in identifying variations in actual field management practices necessitates an extensive series of sample observations over multiple years. More studies should be conducted to create a soil sampling design techniques that can accurately represent the distribution of a discretely distributed population in both geographical and feature space.

Chapter 4 - Conclusion and Future Work

4.1 Conclusion

The major outcomes of this study were:

1. Relatively important and influential soil and crop variables were selected with PLSR and GBRT, respectively. This study reported that OM and ECa shallow showed higher relative importance on spatial variability of grain yield. OM was relatively the highest important soil variable with value of 32.54% and 33.20%, for site year-1 and 2, respectively. ECa shallow emerged as the most influential soil variable for site year-3 and 4. It was also revealed that, crop variable (NDRE) was the primary influencer of spatial variability of grain yield for 3 site years out of 4 site-years, with a relative influence ranging from 33.86% to 76.05%. As a result, the relative influences values of crop variable were used as weight factor to generate weighted Getis-Ord Gi* z-score values.

H-FIS based SSMG were delineated for all site years using the selected soil and crop variables. SSMG were delineated with solely using soil variables that represent macro-scale variability in soil and coupled soil and crop variables representing macro-scale variability in soil and micro-scale variability in crop, respectively. The kappa coefficients demonstrated dissimilarities between SSMGs delineated using solely soil variables and coupled soil and crop variables. Coupled soil and crop variables accounted macro-scale variability in soil and micro-scale variability in crop.

2. Experiments to select effective spectral windows were conducted through pretreating soil samples with high, medium, and low concentration of $NaNO_3$. First derivative of VNIR spectral exhibited sensitivity for soil pre-treated with medium (62 ppm) and high (620 ppm) concentrations of $NaNO_3$. Sensitivity of NO_3^- ions were detected at three spectral windows of 1460 nm to 1690

nm, 1730 nm to 1780 nm, and 1940 nm to 2150 nm. PLSR and RFR model demonstrated to estimate soil NO_3-N using hyperspectral data. RFR model emerged as a better model in NO_3-N estimation with high R^2 values ranging from 0.83 to 0.97. Sampling data set were reduced with SAM technique, pair of soil samples that resulted smaller radians showed a similar characteristic. Non-imaging hyperspectral data with ASD FieldSpec3 has a spectral range from 350 nm to 2500 nm and manages to estimate soil NO_3-N under high and reduced soil sampling data set. The regression model, RFR resulted to the best model in estimating soil NO_3-N under high and reduced soil sampling data set compared to PLSR.

3. Optimal soil sampling size and site configuration were determined using cLHS and hybrid cLHS-SSPOR soil sampling design. The cLHS technique managed to reduced soil sampling size by integrating soil proximal data with crop yield data. Based on the examination of Ripley's K function, yield covariates data introduced a localized spatial variability compared to solely proximal soil covariates.

4.2 Future Work Recommendations

Future research may consider on validation and implementation of SSMGs developed using H-FIS method. The validation process could include diverse crop types. Additional investigation may be considered through extension of using ASD FieldSpec3 hyperspectral sensor to enable on-field Nitrate Nitrogen estimation at field conditions. Advanced soil sampling design techniques can be improved by integrating multiple years of yield and proximal soil covariates to design more optimized sampling strategy. The hybrid cLHS-SSPOR method has shown promise, but its efficacy could be improved by using multiple years of covariates.

References

Adamchuk, V.I., A. Dobermann, M.T. Morgan, and S.M. Brouder. 2002. Feasibility of On-the-go Mapping of Soil Nitrate and Potassium Using Ion-Selective Electrodes. 2002 Chicago, IL July 28-31, 2002. American Society of Agricultural and Biological Engineers, St. Joseph, MI

Adamchuk, V.I., J.W. Hummel, M.T. Morgan, and S.K. Upadhyaya. 2004. On-the-go soil sensors for precision agriculture. Comput Electron Agric 44(1): 71–91. doi: 10.1016/j.compag.2004.03.002.

Anom, S.W.I.M., S. Shibusawa, A. Sasao, K. Sakai, H. Sato, et al. 2000. Moisture, Soil Organic Matter and Nitrate Nitrogen Content Maps Using the Real-Time Soil Spectrophotometer. IFAC Proceedings Volumes 33(29): 307–312. doi: https://doi.org/10.1016/S1474-6670(17)36796-4.

Araújo, S.R., J. Wetterlind, J.A.M. Demattê, and B. Stenberg. 2014. Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. Eur J Soil Sci 65(5): 718–729. doi: 10.1111/ejss.12165.

Barrett, G.E., P.D. Alexander, J.S. Robinson, and N.C. Bragg. 2016. Achieving environmentally sustainable growing media for soilless plant cultivation systems – A review. Sci Hortic 212: 220–234. doi: 10.1016/j.scienta.2016.09.030.

Basso, B., G. Shuai, J. Zhang, and G.P. Robertson. 2019. Yield stability analysis reveals sources of large-scale nitrogen loss from the US Midwest. Sci Rep 9(1). doi: 10.1038/s41598-019-42271-1.

Ben-Dor, E., and A. Banin. 1995. Near-Infrared Analysis as a Rapid Method to Simultaneously Evaluate Several Soil Properties. Soil Science Society of America Journal 59(2): 364–372. doi: 10.2136/sssaj1995.03615995005900020014x.

Bökle, S., M. Karampoiki, D.S. Paraforos, and H.W. Griepentrog. 2023. Using an open source and resilient technology framework to generate and execute prescription maps for site-specific manure application. Smart Agricultural Technology 5. doi: 10.1016/j.atech.2023.100272.

Borenstein, A., R. Linker, I. Shmulevich, and A. Shaviv. 2006. Determination of Soil Nitrate and Water Content Using Attenuated Total Reflectance Spectroscopy. Appl Spectrosc 60(11): 1267–1272. doi: 10.1366/000370206778998969.

Breunig, M., S. Galvão, R. Dalagnol, L. Santi, P. Della Flora, et al. 2020. Assessing the effect of spatial resolution on the delineation of management zones for smallholder farming in southern Brazil. Remote Sens Appl 19. doi: 10.1016/j.rsase.2020.100325.

Bronson, K.F., J.D. Booker, S.J. Officer, R.J. Lascano, S.J. Maas, et al. 2005. Apparent Electrical Conductivity, Soil Properties and Spatial Covariance in the U.S. Southern High Plains.

Brungard, C.W., and J.L. Boettinger. 2010. Conditioned Latin Hypercube Sampling: Optimal Sample Size for Digital Soil Mapping of Arid Rangelands in Utah, USA. Digital Soil Mapping. Springer Netherlands. p. 67–75

Cao, Q., Y. Miao, G. Feng, X. Gao, B. Liu, et al. 2017. Improving nitrogen use efficiency with minimal environmental risks using an active canopy sensor in a wheat-maize cropping system. Field Crops Res 214: 365–372. doi: 10.1016/j.fcr.2017.09.033.

Chang, C.-W., D.A. Laird, M.J. Mausbach, and C.R. Hurburgh. 2001. Near-Infrared Reflectance Spectroscopy-Principal Components Regression Analyses of Soil Properties. Soil Science Society of America Journal 65(2): 480–490. doi: 10.2136/sssaj2001.652480x.

Clark, R.N., and T.L. Roush. 1984. Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. J Geophys Res 89(B7): 6329–6340. doi: 10.1029/JB089iB07p06329.

Clark, V., and SAS Institute. 2004. SAS/STAT 9.1 : user's guide. SAS Pub.

Clifford, D., J.E. Payne, M.J. Pringle, R. Searle, and N. Butler. 2014. Pragmatic soil survey design using flexible Latin hypercube sampling. Comput Geosci 67: 62–68. doi: 10.1016/j.cageo.2014.03.005.

Colin, B., S. Clifford, P. Wu, S. Rathmanner, and K. Mengersen. 2017. Using Boosted Regression Trees and Remotely Sensed Data to Drive Decision-Making. Open J Stat 07(05): 859–875. doi: 10.4236/ojs.2017.75061.

Cordero, E., L. Longchamps, R. Khosla, and D. Sacco. 2019. Spatial management strategies for nitrogen in maize production based on soil and crop data. Science of the Total Environment 697. doi: 10.1016/j.scitotenv.2019.133854.

Córdova, S.C., M.J. Castellano, R. Dietzel, M.A. Licht, K. Togliatti, et al. 2019. Soybean nitrogen fixation dynamics in Iowa, USA. Field Crops Res 236: 165–176. doi: 10.1016/j.fcr.2019.03.018.

Corwin, D.L., and S.M. Lesch. 2005. Apparent soil electrical conductivity measurements in agriculture. Comput Electron Agric 46(1-3 SPEC. ISS.): 11–43. doi: 10.1016/j.compag.2004.10.005.

Damian, J.M., A.L. Santi, M. Fornari, C.O. Da Ros, and V.L. Eschner. 2017. Monitoring variability in cash-crop yield caused by previous cultivation of a cover crop under a no-tillage system. Comput Electron Agric 142: 607–621. doi: 10.1016/j.compag.2017.11.006.

Damiano, M., G. Micela, and G. Tinetti. 2019. A Principal Component Analysis-based Method to Analyze High-resolution Spectroscopic Data on Exoplanets. Astrophys J 878(2): 153. doi: 10.3847/1538-4357/ab22b2.

Demattê, J.A.M., R.C. Campos, M.C. Alves, P.R. Fiorio, and M.R. Nanni. 2004. Visible-NIR reflectance: A new approach on soil evaluation. Geoderma 121(1–2): 95–112. doi: 10.1016/j.geoderma.2003.09.012.

Denora, M., M. Fiorentini, S. Zenobi, P.A. Deligios, R. Orsini, et al. 2022. Validation of Rapid and Low-Cost Approach for the Delineation of Zone Management Based on Machine Learning Algorithms. Agronomy 12(1). doi: 10.3390/agronomy12010183.

Derby, N.E., F.X.M. Casey, and D.W. Franzen. 2007. Comparison of Nitrogen Management Zone Delineation Methods for Corn Grain Yield. Agron J 99(2): 405–414. doi: 10.2134/agronj2006.0027.

Diker, K., D.F. Heermann, and M.K. Brodahl. 2004. Frequency Analysis of Yield for Delineating Yield Response Zones. Precis Agric 5(5): 435–444. doi: 10.1007/s11119-004-5318-9.

Ehsani, M.R., S.K. Upadhyaya, D. Slaughter, S. Shafii, and M. Pelletier. 1999. A NIR Technique for Rapid Determination of Soil Mineral Nitrogen. Precis Agric 1(2): 219–236. doi: 10.1023/A:1009916108990.

Elith, J., J.R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77(4): 802–813. doi: 10.1111/j.1365-2656.2008.01390.x.

Farrar, D.E., and R.R. Glauber. 1967. Multicollinearity in Regression Analysis: The Problem Revisited. Rev Econ Stat 49(1): 92–107. doi: 10.2307/1937887.

Fraisse, C.W., K.A. Sudduth, and N.R. Kitchen. 2001. Delineation of Site-Specific Management Zones by Unsupervised Classification of Topographic Attributes and Soil Electrical Conductivity. American Society of Agricultural Engineers 155(1): 155–166.

Gallardo-Romero, D.J., O.E. Apolo-Apolo, J. Martínez-Guanter, and M. Pérez-Ruiz. 2023. Multilayer Data and Artificial Intelligence for the Delineation of Homogeneous Management Zones in Maize Cultivation. Remote Sens (Basel) 15(12). doi: 10.3390/rs15123131.

Gavioli, A., E.G. de Souza, C.L. Bazzi, L.P.C. Guedes, and K. Schenatto. 2016. Optimization of management zone delineation by using spatial principal components. Comput Electron Agric 127: 302–310. doi: 10.1016/j.compag.2016.06.029.

Gebbers, R., M. Schirrmann, E. Kramer, and J. Seidel. 2011. Predicting lime requirements by fusion of proximal soil sensors. Proceedings of the 8th conference on Precision Agriculture, Czech Centre of Science and Society, Prague, Czech Republic. p. 11–14

Georgi, C., D. Spengler, S. Itzerott, and B. Kleinschmit. 2018. Automatic delineation algorithm for site-specific management zones based on satellite remote sensing data. Precis Agric 19(4): 684–707. doi: 10.1007/s11119-017-9549-y.

Getis, A., and J.K. Ord. 1992. The Analysis of Spatial Association by Use of Distance Statistics. Geogr Anal 24(3): 189–206. doi: 10.1111/j.1538-4632.1992.tb00261.x.

Gili, A., C. Álvarez, R. Bagnato, and E. Noellemeyer. 2017. Comparison of three methods for delineating management zones for site-specific crop management. Comput Electron Agric 139: 213–223. doi: 10.1016/j.compag.2017.05.022.

de Gruijter, J.J., D.J. Brus, M.F.P. Bierkens, and M. Knotters. 2006. Sampling for natural resource monitoring. Springer, Germany.

Guerrero, A., and A.M. Mouazen. 2021. Evaluation of variable rate nitrogen fertilization scenarios in cereal crops from economic, environmental and technical perspective. Soil Tillage Res 213. doi: 10.1016/j.still.2021.105110.

Guillaume, S. 2001. Designing fuzzy inference systems from data: An interpretability-oriented review. IEEE Transactions on Fuzzy Systems 9(3): 426–443. doi: 10.1109/91.928739.

Gundy, G.J., and J.A. Dille. 2022. Implementing variable-rate herbicide applications based on soil physical properties in grain sorghum. Precis Agric 23(3): 768–790. doi: 10.1007/s11119-021-09860-5.

Haghverdi, A., B.G. Leib, R.A. Washington-Allen, P.D. Ayers, and M.J. Buschermohle. 2015. Perspectives on delineating management zones for variable rate irrigation. Comput Electron Agric 117: 154–167. doi: 10.1016/j.compag.2015.06.019.

Hasan, M., T. Tsegaye, X. Shi, G. Schaefer, and G. Taylor. 2008. Model for predicting rainfall by fuzzy set theory using USDA scan data. Agric Water Manag 95(12): 1350–1360. doi: 10.1016/j.agwat.2008.07.015.

Heiß, A., D.S. Paraforos, G.M. Sharipov, P. Ullrich, J. Bruns, et al. 2023. Versatile and user-centered concept for temporally and spatially adapted nitrogen application based on multiple parameters. European Journal of Agronomy 145. doi: 10.1016/j.eja.2023.126792.

Hengl, Tomislav. 2007. A practical guide to geostatistical mapping of environmental variables. Publications Office.

Heuvelink, G.B.M., D.J. Brus, and J.J. de Gruijter. 2006. Chapter 11 Optimization of Sample Configurations for Digital Mapping of Soil Properties with Universal Kriging. In: Lagacherie, P., McBratney, A.B., and Voltz, M., editors, Digital Soil Mapping. Elsevier. p. 137–151

Hirsh, S.M., and R.R. Weil. 2019. Deep Soil Cores Reveal Large End-of-Season Residual Mineral Nitrogen Pool. Agricultural & Environmental Letters 4(1). doi: 10.2134/ael2018.10.0055.

Inman, D., R. Khosla, D.G. Westfall, and R. Reich. 2005. Nitrogen uptake across site specific management zones in irrigated corn production systems. Agron J 97(1): 169–176. doi: 10.2134/agronj2005.0169.

Insausti, M., A.A. Gomes, F. V. Cruz, M.F. Pistonesi, M.C.U. Araujo, et al. 2012. Screening analysis of biodiesel feedstock using UV-vis, NIR and synchronous fluorescence spectrometries and the successive projections algorithm. Talanta 97: 579–583. doi: 10.1016/j.talanta.2012.04.056.

Israeli, A., M. Litaor, M. Emmerich, and O.M. Shir. 2019. Statistical learning in soil sampling design aided by Pareto optimization. GECCO 2019 - Proceedings of the 2019 Genetic and Evolutionary Computation Conference. Association for Computing Machinery, Inc. p. 1198–1205

it03_greedy.

Jahn, B.R., P.A. Brooksby, and S.K. Upadhyaya. 2005. Wavelet-based spectral analysis for soil nitrate content measurement. Transactions of the ASAE 48(6): 2065–2071. doi: 10.13031/2013.20083.

Jahn, B.R., R. Linker, S.K. Upadhyaya, A. Shaviv, D.C. Slaughter, et al. 2006. Mid-infrared Spectroscopic Determination of Soil Nitrate Content. Biosyst Eng 94(4): 505–515. doi: 10.1016/j.biosystemseng.2006.05.011.

Janik, L.J., R.H. Merry, and J.O. Skjemstad. 1998. Can mid infrared diffuse reflectance analysis replace soil extractions? Aust J Exp Agric 38(7): 681–696. doi: 10.1071/EA97144.

Jia, S., H. Li, Y. Wang, R. Tong, and Q. Li. 2016. Recursive variable selection to update near-infrared spectroscopy model for the determination of soil nitrogen and organic carbon. Geoderma 268: 92–99. doi: 10.1016/j.geoderma.2016.01.018.

Jia, S., H. Li, Y. Wang, R. Tong, and Q. Li. 2017. Hyperspectral imaging analysis for the classification of soil types and the determination of soil total nitrogen. Sensors (Switzerland) 17(10). doi: 10.3390/s17102252.

Jin, Z., R. Prasad, J. Shriver, and Q. Zhuang. 2017. Crop model- and satellite imagery-based recommendation tool for variable rate N fertilizer application for the US Corn system. Precis Agric 18(5): 779–800. doi: 10.1007/s11119-016-9488-z.

Johnson, D.M. 2014. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. Remote Sens Environ 141: 116–128. doi: 10.1016/j.rse.2013.10.027.

Kawamura, K., Y. Tsujimoto, M. Rabenarivo, H. Asai, A. Andriamananjara, et al. 2017. Vis-NIR spectroscopy and PLS regression with waveband selection for estimating the total C and N of paddy soils in Madagascar. Remote Sens (Basel) 9(10). doi: 10.3390/rs9101081.

Kelp, M.M., S. Lin, J.N. Kutz, and L.J. Mickley. 2022. A new approach for determining optimal placement of PM2.5air quality sensors: Case study for the contiguous United States. Environmental Research Letters 17(3). doi: 10.1088/1748-9326/ac548f.

Khosla, R., K. Fleming, J.A. Delgado, T.M. Shaver, and D.G. Westfall. 2002. Use of site-specific management zones to improve nitrogen management for precision agriculture. J Soil Water Conserv 57(6): 513. http://www.jswconline.org/content/57/6/513.abstract.

Khosla, R., D.G. Westfall, R.M. Reich, J.S. Mahal, and W.J. Gangloff. 2010. Spatial Variation and Site-Specific Management Zones. In: Oliver, M.A., editor, Geostatistical Applications for Precision Agriculture. Springer Netherlands, Dordrecht. p. 195–219

Kim, H.J., J.W. Hummel, and S.J. Birrell. 2006. Evaluation of nitrate and potassium ion-selective membranes for soil macronutrient sensing. Trans ASABE 49(3): 597–606. doi: 10.13031/2013.20476.

Kitchen, N.R., K.A. Sudduth, S.T. Drummond, P.C. Scharf, H.L. Palm, et al. 2010. Ground-based canopy reflectance sensing for variable-rate nitrogen corn fertilization. Agron J 102(1): 71–84. doi: 10.2134/agronj2009.0114.

Koch, B., R. Khosla, W.M. Frasier, D.G. Westfall, and D. Inman. 2004. Economic feasibility of variable-rate nitrogen application utilizing site-specific management zones. Agron J 96(6): 1572–1580. doi: 10.2134/agronj2004.1572.

Kooistra, L., J. Wanders, G.F. Epema, R.S.E.W. Leuven, R. Wehrens, et al. 2003. The potential of field spectroscopy for the assessment of sediment properties in river floodplains. Anal Chim Acta 484(2): 189–200. doi: 10.1016/S0003-2670(03)00331-3.

Kucheryavskiy, S., S. Zhilin, O. Rodionova, and A. Pomerantsev. 2020. Procrustes Cross-Validation - A Bridge between Cross-Validation and Independent Validation Sets. Anal Chem 92(17): 11842–11850. doi: 10.1021/acs.analchem.0c02175.

Kuhn, M. 2008. Building Predictive Models in R Using the caret Package. J Stat Softw 28(5): 1–26. doi: 10.18637/jss.v028.i05.

Lark, R.M. 2016. Multi-objective optimization of spatial sampling. Spat Stat 18: 412–430. doi: 10.1016/j.spasta.2016.09.001.

Lawrence, P.G., W. Roper, T.F. Morris, and K. Guillard. 2020. Guiding soil sampling strategies using classical and spatial statistics: A review. Agron J 112(1): 493–510. doi: 10.1002/agj2.20048.

Leikam, D.F., R.E. Lamond, and D.B. Mengel. 2003. Soil Test Interpretations and Fertilizer Recommendations - MF2586.

Li, X., C. Sun, B. Zhou, and Y. He. 2015. Determination of Hemicellulose, Cellulose and Lignin in Moso Bamboo by Near Infrared Spectroscopy. Sci Rep 5. doi: 10.1038/srep17210.

Lin, Y.P., H.J. Chu, Y.L. Huang, C.H. Tang, and S. Rouhani. 2011. Monitoring and identification of spatiotemporal landscape changes in multiple remote sensing images by using a stratified conditional Latin hypercube sampling approach and geostatistical simulation. Environ Monit Assess 177(1–4): 353–373. doi: 10.1007/s10661-010-1639-5.

Lin, Y.P., W.C. Lin, M.Y. Li, Y.Y. Chen, L.C. Chiang, et al. 2014. Identification of spatial distributions and uncertainties of multiple heavy metal concentrations by using spatial conditioned Latin Hypercube sampling. Geoderma 230–231: 9–21. doi: 10.1016/j.geoderma.2014.03.015.

Lin, H., D. Wheeler, J. Bell, and L. Wilding. 2005. Assessment of soil spatial variability at multiple scales. Ecol Modell 182(3): 271–290. doi: https://doi.org/10.1016/j.ecolmodel.2004.04.006.

Linker, R. 2004. Waveband Selection for Determination of Nitrate in Soil Using Mid-Infrared Attenuated Total Reflectance Spectroscopy. Appl Spectrosc 58(11): 1277–1281. doi: https://doi.org/10.1366/0003702042475394.

Linker, R., I. Shmulevich, A. Kenny, and A. Shaviv. 2005. Soil identification and chemometrics for direct determination of nitrate in soils using FTIR-ATR mid-infrared spectroscopy. Chemosphere 61(5): 652–658. doi: 10.1016/j.chemosphere.2005.03.034.

Longchamps, L., R. Khosla, R. Reich, and D.W. Gui. 2015. Spatial and Temporal Variability of Soil Water Content in Leveled Fields. Soil Science Society of America Journal 79(5): 1446–1454. doi: 10.2136/sssaj2015.03.0098.

Ludwig, B., R. Murugan, V.R.R. Parama, and M. Vohland. 2018. Use of different chemometric approaches for an estimation of soil properties at field scale with near infrared spectroscopy. Journal of Plant Nutrition and Soil Science 181(5): 704–713. doi: 10.1002/jpln.201800130.

Lunt, I.A., S.S. Hubbard, and Y. Rubin. 2005. Soil moisture content estimation using ground-penetrating radar reflection data. J Hydrol (Amst) 307(1–4): 254–269. doi: 10.1016/j.jhydrol.2004.10.014.

Madari, B.E., J.B. Reeves, P.L.O.A. Machado, C.M. Guimarães, E. Torres, et al. 2006. Mid- and near-infrared spectroscopic assessment of soil compositional parameters and structural indices in two Ferralsols. Geoderma 136(1–2): 245–259. doi: 10.1016/j.geoderma.2006.03.026.

Magdoff, F. 1991. Managing nitrogen for sustainable corn systems: Problems and possibilities. American Journal of Alternative Agriculture 6(1): 3–8. http://www.jstor.org/stable/44503636.

Malone, B.P., B. Minansy, and C. Brungard. 2019. Some methods to improve the utility of conditioned Latin hypercube sampling. PeerJ 2019(2). doi: 10.7717/peerj.6451.

Manohar, K., B.W. Brunton, J.N. Kutz, and S.L. Brunton. 2018. Data-Driven Sparse Sensor Placement for Reconstruction: Demonstrating the Benefits of Exploiting Known Patterns. IEEE Control Syst 38(3): 63–86. doi: 10.1109/MCS.2018.2810460.

Martin, K.L., K. Girma, K.W. Freeman, R.K. Teal, B. Tubańa, et al. 2007. Expression of variability in corn as influenced by growth stage using optical sensor measurements. Agronomy Journal. p. 384–389

Masclaux-Daubresse, C., F. Daniel-Vedele, J. Dechorgnat, F. Chardon, L. Gaufichon, et al. 2010. Nitrogen uptake, assimilation and remobilization in plants: Challenges for sustainable and productive agriculture. Ann Bot 105(7): 1141–1157. doi: 10.1093/aob/mcq028.

McKay Fletcher, D., S. Ruiz, K. Williams, C. Petroselli, N. Walker, et al. 2022. Projected Increases in Precipitation Are Expected to Reduce Nitrogen Use Efficiency and Alter Optimal Fertilization Timings in Agriculture in the South East of England. ACS ES and T Engineering 2(8): 1414–1424. doi: 10.1021/acsestengg.1c00492.

Miao, Y., D.J. Mulla, and P.C. Robert. 2018. An integrated approach to site-specific management zone delineation. Front Agric Sci Eng 5(4): 432–441. doi: 10.15302/J-FASE-2018230.

Minasny, B., and A.B. McBratney. 2006a. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput Geosci 32(9): 1378–1388. doi: 10.1016/j.cageo.2005.12.009.

Minasny, B., and A.B. McBratney. 2006b. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput Geosci 32(9): 1378–1388. doi: 10.1016/j.cageo.2005.12.009.

Minasny, B., and A.B. McBratney. 2010. Conditioned Latin Hypercube Sampling for Calibrating Soil Sensor Data to Soil Properties. Proximal Soil Sensing. Springer Netherlands. p. 111–119

Moral, F.J., F.J. Rebollo, and J. Serrano. 2023. Using a Non-Contact Sensor to Delineate Management Zones in Vineyards and Validation with the Rasch Model. Sensors 23(22): 9183. doi: 10.3390/s23229183.

Morellos, A., X.E. Pantazi, D. Moshou, T. Alexandridis, R. Whetton, et al. 2016. Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. Biosyst Eng 152: 104–116. doi: 10.1016/j.biosystemseng.2016.04.018.

Morris, T.F., T.S. Murrell, D.B. Beegle, J.J. Camberato, R.B. Ferguson, et al. 2018. Strengths and limitations of Nitrogen rate recommendations for corn and opportunities for improvement. Agron J 110(1): 1–37. doi: 10.2134/agronj2017.02.0112.

Mueller, T.G., F.J. Pierce, O. Schabenberger, and D.D. Warncke. 2001. Map Quality for Site-Specific Fertility Management. Soil Science Society of America Journal 65(5): 1547–1558. doi: 10.2136/sssaj2001.6551547x.

Mulder, V.L., S. de Bruin, and M.E. Schaepman. 2013. Representing major soil variability at regional scale by constrained Latin Hypercube Sampling of remote sensing data. International Journal of Applied Earth Observation and Geoinformation 21(1): 301–310. doi: 10.1016/j.jag.2012.07.004.

Mulla, D.J. 2013. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. Biosyst Eng 114(4): 358–371. doi: 10.1016/j.biosystemseng.2012.08.009.

Mulla, D., and R. Khosla. 2015. Historical Evolution and Recent Advances in Precision Farming. p. 1–36

Mzuku, M., R. Khosla, R. Reich, D. Inman, F. Smith, et al. 2005. Spatial Variability of Measured Soil Properties across Site-Specific Management Zones. Soil Science Society of America Journal 69(5): 1572–1579. doi: 10.2136/sssaj2005.0062.

Naderloo, L., R. Alimardani, M. Omid, F. Sarmadian, P. Javadikia, et al. 2012. Application of ANFIS to predict crop yield based on different energy inputs. Measurement (Lond) 45(6): 1406–1413. doi: 10.1016/j.measurement.2012.03.025.

Nie, P.C., T. Dong, Y. He, and F. Qu. 2017. Detection of soil nitrogen using near infrared sensors based on soil pretreatment and algorithms. Sensors (Switzerland) 17(5). doi: 10.3390/s17051102.

Niu, X., Z. Song, C. Xu, C. Wu, Q. Luan, et al. 2023. Prediction of Needle Physiological Traits Using UAV Imagery for Breeding Selection of Slash Pine. Plant Phenomics 5. doi: 10.34133/plantphenomics.0028.

Oglesby, C., A.A.A. Fox, G. Singh, and J. Dhillon. 2022. Predicting In-Season Corn Grain Yield Using Optical Sensors. Agronomy 12(10). doi: 10.3390/agronomy12102402.

Ohana-Levi, N., I. Bahat, A. Peeters, A. Shtein, Y. Netzer, et al. 2019. A weighted multivariate spatial clustering model to determine irrigation management zones. Comput Electron Agric 162: 719–731. doi: 10.1016/j.compag.2019.05.012.

Omara, P., L. Aula, F. Oyebiyi, and W.R. Raun. 2019. World Cereal Nitrogen Use Efficiency Trends: Review and Current Knowledge. Agrosystems, Geosciences and Environment 2(1): 1–8. doi: 10.2134/age2018.10.0045.

Ortega, R.A., and O.A. Santibáñez. 2007. Determination of management zones in corn (Zea mays L.) based on soil fertility. Comput Electron Agric 58(1): 49–59. doi: 10.1016/j.compag.2006.12.011.

Paiao, G.D., F.G. Fernández, J.A. Spackman, D.E. Kaiser, and S. Weisberg. 2021. Integrating canopy sensing and soil nitrogen for improved corn nitrogen management. Agron J 113(2): 1996–2005. doi: 10.1002/agj2.20566.

Peeters, A., M. Zude, J. Käthner, M. Ünlü, R. Kanber, et al. 2015. Getis-Ord’s hot- and cold-spot statistics as a basis for multivariate spatial clustering of orchard tree data. Comput Electron Agric 111: 140–150. doi: 10.1016/j.compag.2014.12.011.

Peralta, N.R., J.L. Costa, M. Balzarini, M. Castro Franco, M. Córdoba, et al. 2015. Delineation of management zones to improve nitrogen management of wheat. Comput Electron Agric 110: 103–113. doi: 10.1016/j.compag.2014.10.017.

Qaddoum, K., E. Hines, and D. Illiescu. 2011. Adaptive neuro-fuzzy modeling for crop yield prediction. Proceedings of the 10th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA. p. 199–204

R. Jahn, B., P. A. Brooksby, and S. K. Upadhyaya. 2005. WAVELET-BASED SPECTRAL ANALYSIS FOR SOIL NITRATE CONTENT MEASUREMENT. Transactions of the ASAE 48(6): 2065–2071. doi: https://doi.org/10.13031/2013.20083.

Rasmussen, J., S. Azim, S.K. Boldsen, T. Nitschke, S.M. Jensen, et al. 2021. The challenge of reproducing remote sensing data from satellites and unmanned aerial vehicles (UAVs) in the context of management zones and precision agriculture. Precis Agric 22(3): 834–851. doi: 10.1007/s11119-020-09759-7.

Rasskazov, I.L., R. Singh, P.S. Carney, and R. Bhargava. 2019. Extended Multiplicative Signal Correction for Infrared Microspectroscopy of Heterogeneous Samples with Cylindrical Domains. Appl Spectrosc 73(8): 859–869. doi: 10.1177/0003702819844528.

Reeves, J.B., G.W. McCarty, and W.D. Hively. 2010. Mid- Versus Near-Infrared Spectroscopy for On-Site Analysis of Soil. Proximal Soil Sensing. Springer Netherlands. p. 133–142

Ren, Z., L. Sun, and Q. Zhai. 2020. Improved k-means and spectral matching for hyperspectral mineral mapping. International Journal of Applied Earth Observation and Geoinformation 91. doi: 10.1016/j.jag.2020.102154.

Rossel, R.A.V., T. Behrens, E. Ben-Dor, D.J. Brown, J.A.M. Demattê, et al. 2016. A global spectral library to characterize the world's soil. Earth Sci Rev 155: 198–230. doi: 10.1016/j.earscirev.2016.01.012ï.

Rossi, R., A. Pollice, G. Bitella, R. Labella, R. Bochicchio, et al. 2018. Modelling the non-linear relationship between soil resistivity and alfalfa NDVI: A basis for management zone delineation. J Appl Geophy 159: 146–156. doi: 10.1016/j.jappgeo.2018.08.008.

Saifuzzaman, M. 2020. Optimization of Geospatial Data Modelling for Crop Production by Integrating Proximal Soil Sensing and Remote Sensing Data.

Saito, Y., T. Nonomura, K. Yamada, K. Nakai, T. Nagata, et al. 2021. Determinant-Based Fast Greedy Sensor Selection Algorithm. IEEE Access 9: 68535–68551. doi: 10.1109/ACCESS.2021.3076186.

Santos, U.J. dos, J.A. de M. Demattê, R.S.C. Menezes, A.C. Dotto, C.C.B. Guimarães, et al. 2020. Predicting carbon and nitrogen by visible near-infrared (Vis-NIR) and mid-infrared (MIR) spectroscopy in soils of Northeast Brazil. Geoderma Regional. doi: 10.1016/j.geodrs.2020.e00333.

Saurette, D.D., A.A. Berg, A. Laamrani, R.J. Heck, A.W. Gillespie, et al. 2022a. Effects of sample size and covariate resolution on field-scale predictive digital mapping of soil carbon. Geoderma 425. doi: 10.1016/j.geoderma.2022.116054.

Saurette, D.D., A.A. Berg, A. Laamrani, R.J. Heck, A.W. Gillespie, et al. 2022b. Effects of sample size and covariate resolution on field-scale predictive digital mapping of soil carbon. Geoderma 425. doi: 10.1016/j.geoderma.2022.116054.

Savitzky, A., and M. ~J. ~E. Golay. 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal Chem 36: 1627–1639. doi: 10.1021/ac60214a047.

Schepers, A.R., J.F. Shanahan, M.A. Liebig, J.S. Schepers, S.H. Johnson, et al. 2004. Appropriateness of Management Zones for Characterizing Spatial Variability of Soil Properties and Irrigated Corn Yields across Years. Agron J 96(1): 195–203. doi: 10.2134/agronj2004.1950.

Schmidt, K., T. Behrens, J. Daumann, L. Ramirez-Lopez, U. Werban, et al. 2014. A comparison of calibration sampling schemes at the field scale. Geoderma 232–234: 243–256. doi: 10.1016/j.geoderma.2014.05.013.

Schmidt, J.P., D. Markewitz, F. de A. Oliveira, A. Sila, and A.H. Joslin. 2017. Using vNIR spectroscopy to assess changes in Ultisols of Pará State, Brazil. Geoderma Regional 11: 71–77. doi: 10.1016/j.geodrs.2017.09.003.

Serrano, L., J. Peñuelas, and S.L. Ustin. 2015. Remote sensing of nitrogen and lignin in Mediterranean vegetation from AVIRIS data: Decomposing biochemical from structural signals. www.elsevier.com/locate/rse.

Shahriari, H., H. Ranjbar, M. Honarmand, and E.J.M. Carranza. 2014. Selection of less biased threshold angles for SAM classification using the real value-area fractal technique. Resource Geology 64(4): 301–315. doi: 10.1111/rge.12042.

Shanahan, J.F., N.R. Kitchen, W.R. Raun, and J.S. Schepers. 2008. Responsive in-season nitrogen management for cereals. Comput Electron Agric 61(1): 51–62. doi: 10.1016/j.compag.2007.06.006.

Shaner, D.L., R. Khosla, M.K. Brodahl, G.W. Buchleiter, and H.J. Farahani. 2008. How well does zone sampling based on soil electrical conductivity maps represent soil variability? Agron J 100(5): 1472–1480. doi: 10.2134/agronj2008.0060.

Shashikumar, B.N., S. Kumar, K.J. George, and A.K. Singh. 2023. Soil variability mapping and delineation of site-specific management zones using fuzzy clustering analysis in a Mid-Himalayan Watershed, India. Environ Dev Sustain 25(8): 8539–8559. doi: 10.1007/s10668-022-02411-6.

Shatar, T.M., and A.B. Mcbratney. 1999. Empirical Modeling of Relationships Between Sorghum Yield and Soil Properties. Precis Agric 1(3): 249–276. doi: 10.1023/A:1009968907612.

Shibusawa, S. 2003. On-line real time soil sensor. Proceedings 2003 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM 2003). p. 1061–1066 vol.2

de Silva, B., K. Manohar, E. Clark, B. Brunton, J. Kutz, et al. 2021. PySensors: A Python package for sparse sensor placement. J Open Source Softw 6(58): 2828. doi: 10.21105/joss.02828.

Sishodia, R.P., R.L. Ray, and S.K. Singh. 2020. Applications of remote sensing in precision agriculture: A review. Remote Sens (Basel) 12(19): 1–31. doi: 10.3390/rs12193136.

Smith, E.M., R. Vargas, M. Guevara, T. Tarin, and R. V. Pouyat. 2022. Spatial variability and uncertainty of soil nitrogen across the conterminous United States at different depths. Ecosphere 13(7). doi: 10.1002/ecs2.4170.

Soil Survey Staff. 2024. Natural Resources Conservation Service, United States Department of Agriculture. Web Soil Survey. https://websoilsurvey.sc.egov.usda.gov/.

Song, X., Y. Gao, Z. Liu, M. Zhang, Y. Wan, et al. 2019. Development of a predictive tool for rapid assessment of soil total nitrogen in wheat-corn double cropping system with hyperspectral data. Environmental Pollutants and Bioavailability 31(1): 272–281. doi: 10.1080/26395940.2019.1679041.

Sood, K., S. Shamsher, M. Thirumalaisamy, P. Tyagi, and N. Suvarna. 2022. A Novel Methodology based Soil characteristic Analysis using Machine Learning Techniques. 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2022. Institute of Electrical and Electronics Engineers Inc. p. 965–973

Sudduth, K.A., and S.T. Drummond. 2007. Yield Editor: Software for Removing Errors from Crop Yield Maps. Agron J v. 99.

Summers, D., M. Lewis, B. Ostendorf, and D. Chittleborough. 2011. Visible near-infrared reflectance spectroscopy as a predictive indicator of soil properties. Ecol Indic 11(1): 123–131. doi: 10.1016/j.ecolind.2009.05.001.

Taghavifar, H., and A. Mardani. 2014. Prognostication of vertical stress transmission in soil profile by adaptive neuro-fuzzy inference system based modeling approach. Measurement (Lond) 50(1): 152–159. doi: 10.1016/j.measurement.2013.12.035.

Temlyakov.bib.

Torino, M.S., B. V. Ortiz, J.P. Fulton, K.S. Balkcom, and C.W. Wood. 2014. Evaluation of vegetation indices for early assessment of corn status and yield potential in the southeastern United States. Agron J 106(4): 1389–1401. doi: 10.2134/agronj13.0578.

Tripathi, R., A.K. Nayak, M. Shahid, B. Lal, P. Gautam, et al. 2015. Delineation of soil management zones for a rice cultivated area in eastern India using fuzzy clustering. Catena (Amst) 133: 128–136. doi: 10.1016/j.catena.2015.05.009.

Urretavizcaya, I., L.G. Santesteban, B. Tisseyre, S. Guillaume, C. Miranda, et al. 2014. Oenological significance of vineyard management zones delineated using early grape sampling. Precis Agric 15(1): 111–129. doi: 10.1007/s11119-013-9328-3.

USDA, and FAS. 2023. World Agricultural Production.

Vallentin, C., E.S. Dobers, S. Itzerott, B. Kleinschmit, and D. Spengler. 2020. Delineation of management zones with spatial data fusion and belief theory. Precis Agric 21(4): 802–830. doi: 10.1007/s11119-019-09696-0.

Vašát, R., L. Bor_vka, and O. Jak_ík. 2012. Number of sampling points influences the parameters of soil properties spatial distribution and kriged maps. p. 251–256

Vibhute, A.D., K. V. Kale, S. V. Gaikwad, and R.K. Dhumal. 2020a. Estimation of soil nitrogen in agricultural regions by VNIR reflectance spectroscopy. SN Appl Sci 2(9). doi: 10.1007/s42452-020-03322-9.

Vibhute, A.D., K. V. Kale, S. V. Gaikwad, and R.K. Dhumal. 2020b. Estimation of soil nitrogen in agricultural regions by VNIR reflectance spectroscopy. SN Appl Sci 2(9). doi: 10.1007/s42452-020-03322-9.

Villarini, G., C.S. Jones, and K.E. Schilling. 2016. Soybean Area and Baseflow Driving Nitrate in Iowa's Raccoon River. J Environ Qual 45(6): 1949–1959. doi: 10.2134/jeq2016.05.0180.

Viscarra Rossel, R.A., D.J.J. Walvoort, A.B. McBratney, L.J. Janik, and J.O. Skjemstad. 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131(1–2): 59–75. doi: 10.1016/j.geoderma.2005.03.007.

Vohland, M., M. Ludwig, S. Thiele-Bruhn, and B. Ludwig. 2014. Determination of soil properties with visible to near- and mid-infrared spectroscopy: Effects of spectral variable selection. Geoderma 223–225(1): 88–96. doi: 10.1016/j.geoderma.2014.01.013.

Wan, C., Y. Kuzyakov, C. Cheng, S. Ye, B. Gao, et al. 2021a. A soil sampling design for arable land quality observation by using SPCOSA–CLHS hybrid approach. Land Degrad Dev 32(17): 4889–4906. doi: 10.1002/ldr.4077.

Wan, C., Y. Kuzyakov, C. Cheng, S. Ye, B. Gao, et al. 2021b. A soil sampling design for arable land quality observation by using SPCOSA–CLHS hybrid approach. Land Degrad Dev 32(17): 4889–4906. doi: 10.1002/ldr.4077.

Wang, J., B. Gao, and A. Stein. 2020. The spatial statistic trinity: A generic framework for spatial sampling and inference. Environmental Modelling and Software 134. doi: 10.1016/j.envsoft.2020.104835.

Wang, Q., H. Zhang, F. Li, C. Gu, Y. Qiao, et al. 2021. Assessment of calibration methods for nitrogen estimation in wet and dry soil samples with different wavelength ranges using near-infrared spectroscopy. Comput Electron Agric 186. doi: 10.1016/j.compag.2021.106181.

Wold, S., K. Esbensen, and P. Geladi. Principal Component Analysis. 1987.

Wu, H., M. Siege, R. Stiefelhagen, and J. Yang. 2002. Sensor Fusion Using Dempster-Shafer Theory 11: Static Weighting and Kalman Filter-like Dynamic Weighting. EEE Instrumentation and Measurement. p. 21–23

Xiao, S., Y. He, T. Dong, and P. Nie. 2018. Spectral analysis and sensitive waveband determination based on nitrogen detection of different soil types using near infrared sensors. Sensors (Switzerland) 18(2). doi: 10.3390/s18020523.

Xie, C., C. Yang, A. Hummel Jr, G. A Johnson, and F. T Izuno. 2018. Spectral reflectance response to nitrogen fertilization in field grown corn. International Journal of Agricultural and Biological Engineering 11(4): 102–109. doi: 10.25165/j.ijabe.20181104.2960.

Xu, Y., X. Wang, J. Bai, D. Wang, W. Wang, et al. 2020. Estimating the spatial distribution of soil total nitrogen and available potassium in coastal wetland soils in the Yellow River Delta by incorporating multi-source data. Ecol Indic 111. doi: 10.1016/j.ecolind.2019.106002.

Yan, T., S.L. Shen, and A. Zhou. 2023. Identification of geological characteristics from construction parameters during shield tunnelling. Acta Geotech 18(1): 535–551. doi: 10.1007/s11440-022-01590-w.

Yanl, L.I., S. Zhou2, W.U. Ci-Fang’, H.-Y. Li, and L.I. Feng3. 2007. Improved Prediction and Reduction of Sampling Density for Soil Salinity by Different Geostatistical Methods.

Zeng, R., D.G. Rossiter, Y.G. Zhao, D.C. Li, F. Liu, et al. 2023. The choice of spectral similarity algorithms influences suspected soil sample provenance. Forensic Sci Int 347. doi: 10.1016/j.forsciint.2023.111688.

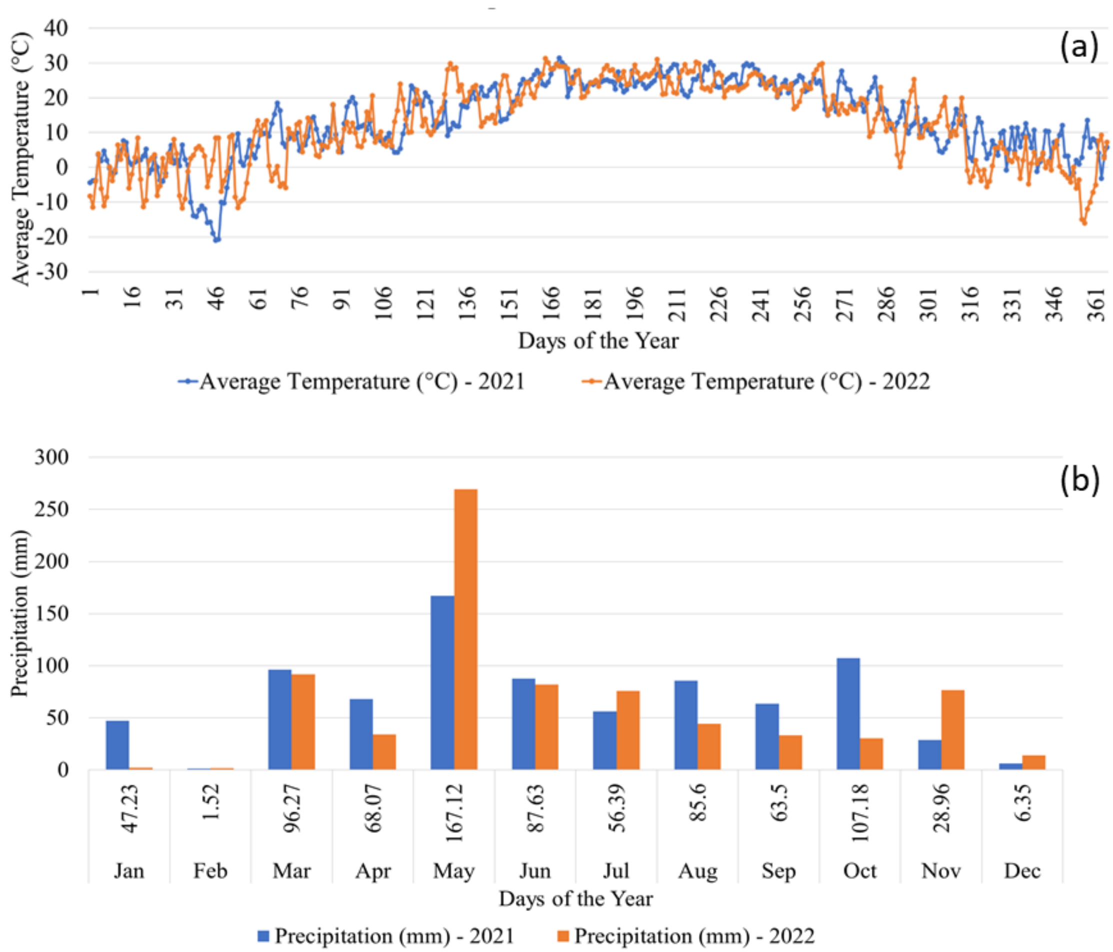
Zhang, Y., D.D. Saurette, T.H. Easher, W. JI, V.I. Admachuk, et al. 2022. Comparison of sampling designs for calibrating digital soil maps at multiple depths. Pedosphere 32(4): 588–601. doi: 10.1016/S1002-0160(21)60055-3.

Appendix A - Study Area

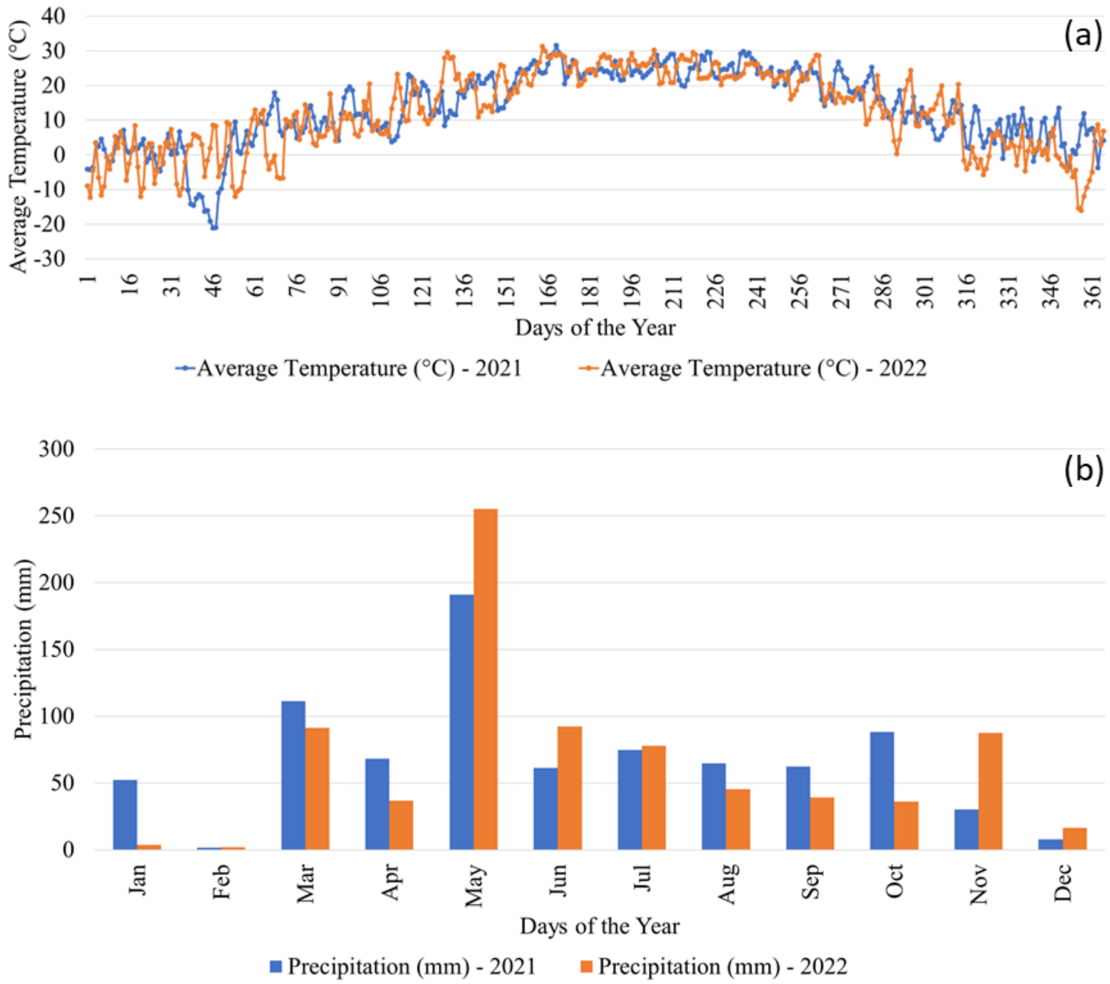
Appendix Table A.1. Study sites in Kansas and Colorado

Study site	**Year**	**Site Year**	**County**	**Location**	**Area (ha)**
Kansas Research Valley Topeka	2021	Site Year-1	Shawnee	39° 4’ 34.04" N, 95° 46’ 11.98" W	3.36
Kansas Research Valley Rossville	2021	Site Year-2	Shawnee	39° 7’ 6.36" N, 95° 55’ 37.35" W	7.76
Kansas Research Valley Rossville	2022	Site Year-3	Shawnee	39° 7’ 6.36" N, 95° 55’ 37.35" W	7.76
Lonsinger research site	2021	Site Year-4	Osborne	39° 29’ 50.57" N, 99° 00’ 54.53" W	16.18
Fort Collins	2021	Site Year-5	Larimer	40° 39' 57.4" N, 104° 59' 53.1" W	8.85
Fort Collins	2022	Site Year-6	Larimer	40° 39' 57.4" N, 104° 59' 53.1" W	3.91

Appendix B - Weather information for all four site years.



Appendix Figure B.1. Weather station data for year 2021 and 2022 at Kansas Research Valley Topeka experiment site (a) daily average temperature in °C; and (b) monthly precipitation in mm.



Appendix Figure B.2. Weather station data for year 2021 and 2022 at Kansas Research Valley Rossville experiment site (a) daily average temperature in °C; and (b) monthly precipitation in mm.

Appendix C - Crop cultivation and management practices

Appendix Table C.1. Crop cultivation and management practices

Site year	Crop	Variety	Planting date	Seed rate (seeds per ha)
Site year-1	Corn (Zea mays L.)	Corn DK 65-95	April 26th, 2021	79,040
Site year-2	Soybean (Glycine max)	AG37X71	May 23rd, 2022	353,200
Site year-3	Soybean (Glycine max)	NK 39-62X	May 5th, 2021	345,800
Site year-4	Corn (Zea mays L.)	DK 65-95	April 22nd, 2022	79,040

Appendix D - UAV flight plan preparation

Flight plans were prepared using UgCS v.4.18 for DJI. The polygon shapefile for target area was prepared with ArcGIS Pro 3.1.1 (ArcGIS, ESRI, Redlands, CA, USA). Flight parameters in relation to flight mission were set on UgCS v.4.18 photogrammetry tool. Flight mission was created with flight speed of 6 m/s, ground resolution (GSD) of 3.34 cm, a flight altitude of 50 meters above mean sea level and forward and side overlap of 75% parameters. The flight was monitored on work laptop and cell phone devices. The operation was at auto-pilot mode, and controlled on UgCS telemetry window with, allowing for real-time updates on drone flight status and position. Flights were conducted on a no partial cloudy days and within ± of two hours from solar noon.

Appendix E - UAV imagery collection

Appendix Table E.1. UAV imagery collection dates for four site years at Kansas River Valley (KRV) Topeka and Rossville research sites

Year	Crop status	KRV Topeka site	KRV Rossville site
2021	Vegetative stage	28-Jun	2-Jul
	Vegetation stage	12-Jul	13-Jul
	Closed canopy	27-Jul	27-Jul
	Near maturity	16-Aug	16-Aug
2022	Vegetation stage	9-Jun	18-Jun
	Vegetation stage	7-Jul	13-Jul
	Closed canopy	29-Jul	29-Jul
	Near maturity	10-Aug	10-Aug
	Near maturity	23-Aug	-

Appendix F - Soil parameters estimated from samples collected

The soil samples were analyzed for physical and chemical properties as enlisted in Appendix Table F.1 , at AgSource Laboratories in Lincoln, Nebraska, USA.

Appendix Table F.1. Analyzed soil parameters on the soil samples collected from all study sites.

Soil parameters	**Unit**	**Methodology**
Organic Matter	%	Loss on Ignition (LOI)
Nitrate-Nitrogen	mg/kg	Cadmium Reduction
Phosphorus, Bray 1	mg/kg	Bray 1 Extraction, If pH < 7.2
Phosphorus, Olsen	mg/kg	Olsen Extraction, If pH >7.1
Potassium	mg/kg	Ammonium Acetate Extraction
Calcium, Magnesium, and Sodium	mg/kg	Ammonium Acetate Extraction
Sulfur	mg/kg	Monocalcium Phosphate Extraction
Copper, Iron, Manganese, and Zinc	mg/kg	DTPA Extraction
Soil pH	--	1:1 Soil/Water Slurry
Buffer pH	--	Sikora Buffer Solution
Soluble Salts (EC)	mmhos/cm	1:1 Soil/Water Slurry
CEC	meq/100g	--
Base Saturation (K^+, Ca^{2+}, Mg^{2+}, Na^+, and H^+)	%	--

Appendix G - Descriptive statistics of soil properties

Appendix Table G.1. Descriptive statistics of soil properties for Kansas Research Valley Topeka, year 2021.

Soil Parameters	Mean	Min	Max	Median	S.D.
Organic matter (%)	1.1	0.4	2.2	1.1	0.5
N (mg/kg)	11.0	1.0	27.8	9.3	8.1
Bray P (mg/kg)	18.7	3.0	85.0	8.5	22.9
Na (mg/kg)	20.0	10.0	31.0	20.5	7.1
K (mg/kg)	126.6	69.0	396.0	108.0	81.7
Mg (mg/kg)	146.2	58.0	331.0	143.5	70.0
Ca (mg/kg)	2109.7	1157.0	3488.0	1963.5	789.8
S (mg/kg)	8.2	5.0	14.0	8.0	3.1
pH	7.0	6.3	8.0	6.7	1.8
Soluble salts (mmhos/cm)	0.2	0.2	0.3	0.2	0.1
CEC (m eq/100 g)	12.4	7.4	23.5	11.3	4.9

S.D = Standard deviation, N = Nitrogen (mg/kg), P = Phosphorus (Bray P) (mg/kg), K = Potassium (mg/kg), Mg = Magnesium (mg/kg), Ca = Calcium (mg/kg), S = Sulfur (mg/kg), CEC = Cation exchange capacity (meq/100 g soil).

Appendix Table G.2. Descriptive statistics of soil properties for Kansas Research Valley Rossville, year 2021.

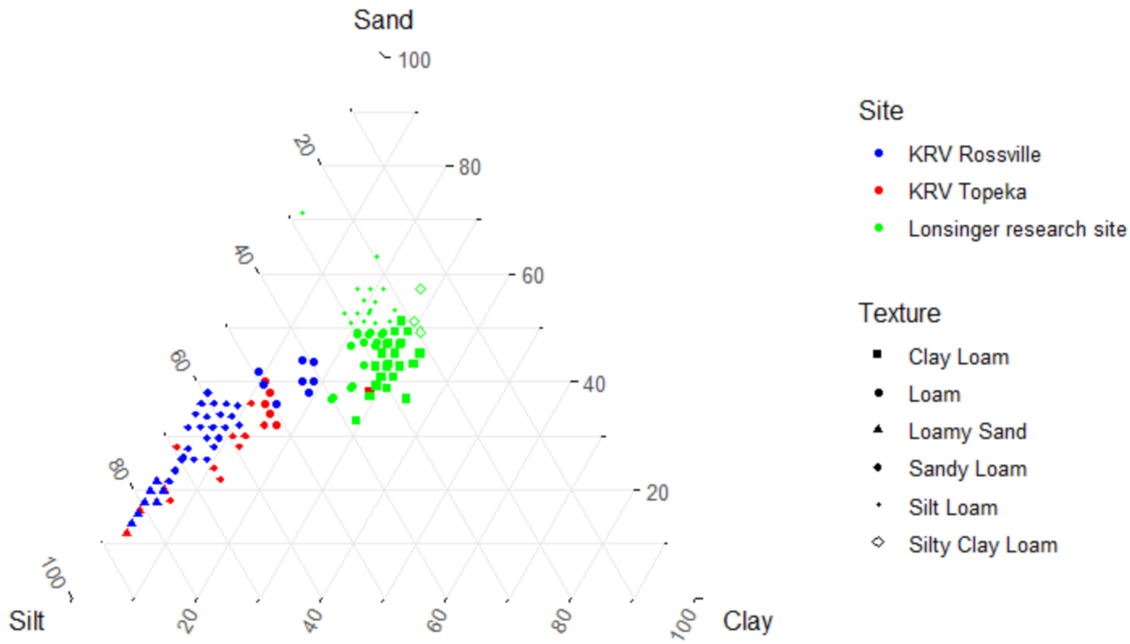
Soil Parameters	Mean	Min	Max	Median	S.D.
Organic matter (%)	1.3	0.7	2.3	1.2	0.4
N (mg/kg)	2.2	1.0	6.8	1.6	1.3
Bray P (mg/kg)	20.4	3.0	64.0	16.0	13.0
Na (mg/kg)	24.0	9.0	44.0	23.0	9.4
K (mg/kg)	180.1	86.0	362.0	166.0	59.8
Mg (mg/kg)	113.8	67.0	211.0	109.0	32.7
Ca (mg/kg)	1337.7	650.0	2458.0	1243.0	452.4
S (mg/kg)	8.9	4.0	14.0	9.0	2.7
pH	6.8	5.6	7.8	6.8	0.5
Soluble salts (mmhos/cm)	0.2	0.1	0.2	0.2	0.0
CEC (m eq/100 g)	8.7	4.8	15.1	8.3	2.5

S.D = Standard deviation, N = Nitrogen (mg/kg), P = Phosphorus (Bray P) (mg/kg), K = Potassium (mg/kg), Mg = Magnesium (mg/kg), Ca = Calcium (mg/kg), S = Sulfur (mg/kg), CEC = Cation exchange capacity (meq/100 g soil).

Appendix Table G.3. Descriptive statistics of soil properties for Lonsinger research site, year 2021.

Soil Parameters	Mean	Min	Max	Median	S.D.
Organic matter (%)	2.1	1.6	3.1	2.0	0.4
N (mg/kg)	9.1	3.7	17.3	8.9	2.9
Bray P (mg/kg)	15.3	1.0	52.0	13.0	11.0
Na (mg/kg)	33.1	7.0	293.0	15.0	54.1
K (mg/kg)	510.6	179.0	772.0	509.0	116.0
Mg (mg/kg)	281.5	104.0	645.0	275.5	113.7
Ca (mg/kg)	3904.9	1497.0	6543.0	4009.0	1395.1
S (mg/kg)	5.7	4.0	19.0	5.0	2.3
pH	7.1	5.4	8.1	7.5	0.9
Soluble salts (mmhos/cm)	0.4	0.3	0.6	0.4	0.1
CEC (m eq/100 g)	24.5	14.4	36.6	25.3	5.3

S.D = Standard deviation, N = Nitrogen (mg/kg), P = Phosphorus (Bray P) (mg/kg), K = Potassium (mg/kg), Mg = Magnesium (mg/kg), Ca = Calcium (mg/kg), S = Sulfur (mg/kg), CEC = Cation exchange capacity (meq/100 g soil).



Appendix Figure G.1. Soil texture distribution by location

The plot represents the distribution of soil texture classes among three sites: Kansas River Valley (KRV) Rossville (blue), Kansas River Valley (KRV) Topeka (red), and Lonsinger research site (green). Texture classes are represented as follows: square for Clay Loam, circle for Loam, triangle pointing up for Loamy Sand, bigger diamond for Sandy Loam, smaller diamond for Silt Loam, and diamond outline for Silty Clay Loam. Each point represents a soil sample's percentage composition of sand, silt, and clay, as measured in respective soil texture analyses**.**

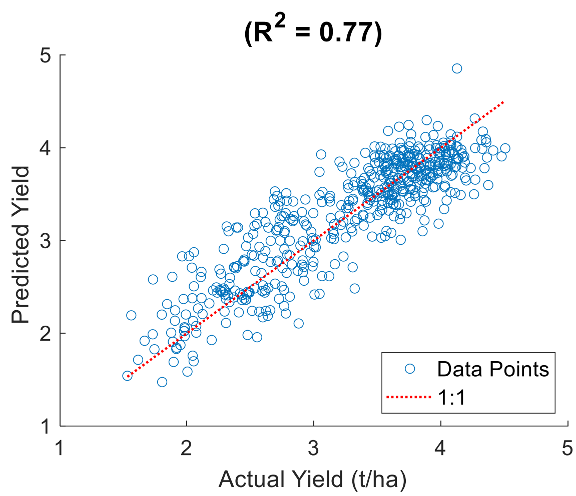
Appendix H - Grain yield monitor data

Appendix Table H.1. Summarized yield monitor data for all site years.

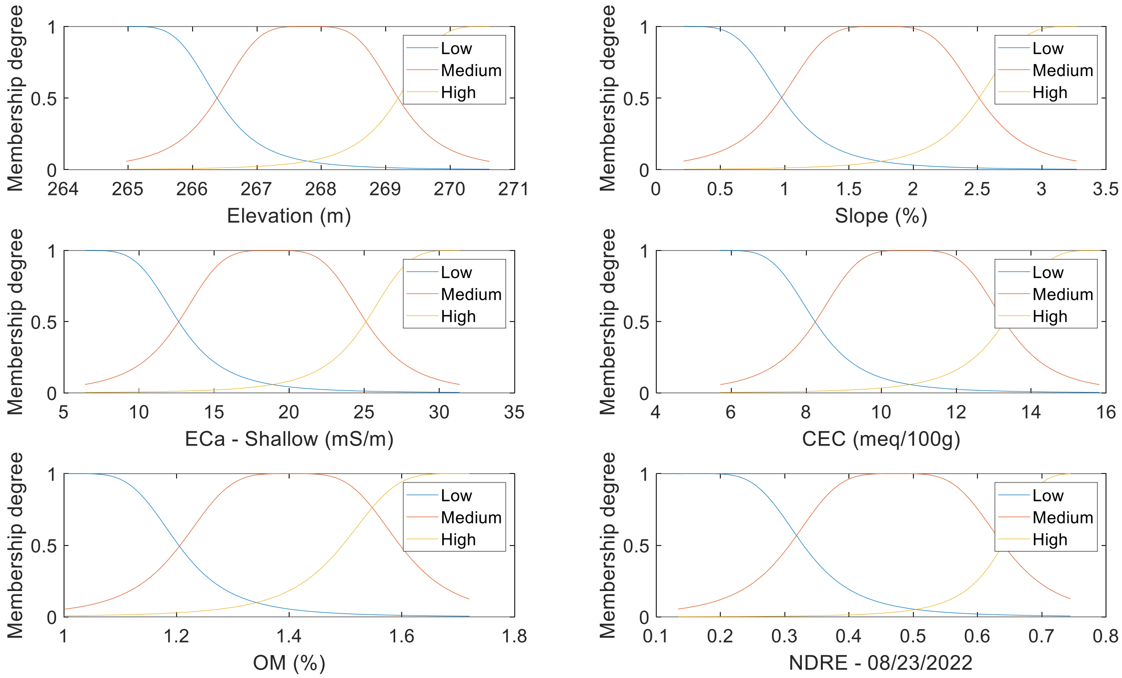
Year	Study site	Crop	Area (ha)	Harvesting date	Estimated average dry yield (t/ha)
2021	Site year-1	Corn	3.36	15-Sep	14.91
	Site year-2	Soybeans	7.73	18-Oct	3.78
2022	Site year-1	Soybeans	3.36	10-Oct	3.62
	Site year-2	Corn	7.73	27-Sep	14.64

Appendix I - Membership functions and ANFIS output

For site year-2, the ANFIS model was set up with six inputs (Elevation, slope, ECa at shallow depth, OM, CEC, NDRE - 08/23/2022) and Soybean yield. A minimum training root means squared error (RMSE) of 0.29 t/ha and a validation RMSE of 0.51 t/ha by the 70th epoch, which is where the training process was completed. The ANFIS model showed a promising model fit, but it can be further improved by adjusting the membership function and re-tune it with a different set of training data to acquire a better prediction of Soybean yield for site year-2. The model's performance was further quantified by the coefficient of determination, R^2, which stands at 0.82. The model effectively captured a significant portion of the variations in the actual Soybean yield.

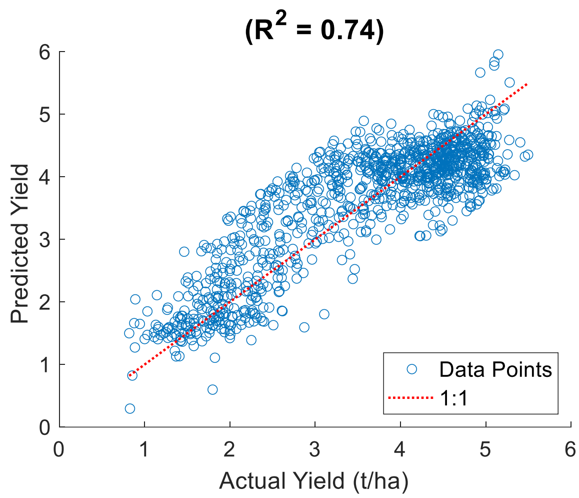


Appendix Figure I.1. Scatter plot of actual vs. Adaptive Neuro-Fuzzy Inference System (ANFIS)-predicted soybean yield for site year-2

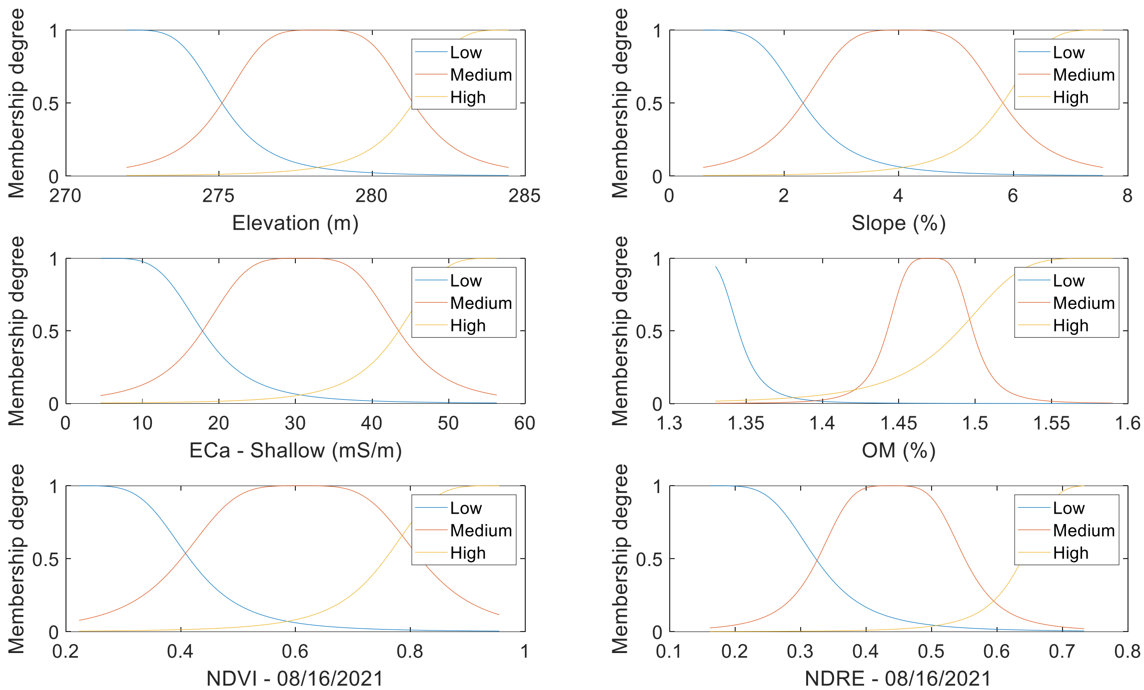


Appendix Figure I.2. Membership function of soil and crop variables for site year-2.

Variables like, elevation, slope, ECa at shallow, OM, NDRE on 08/16/2021 and NDVI on 08/16/2021 were used to train the model for site year-3. Soybean yields were used as the output variable for site year-3. The model training stopped in the 70th epoch, where the minimum training root mean squared error (RMSE) of 0.60 t/ha and a validation RMSE of 0.77 t/ha reached. The ANFIS model was effectively trained and fine-tuned for predicting soybean yield for site year. The scatter plot showed that the predicted soybean yield mirrors the observed yield values, with a coefficient of determination, R^2 of 0.74. The proportion of variance captured by the model in the actual soybean yield was effective.

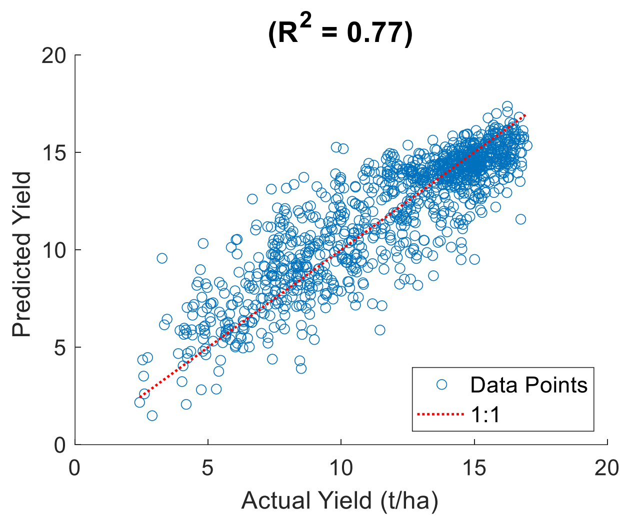


Appendix Figure I.3. Scatter plot of actual vs. Adaptive Neuro-Fuzzy Inference System (ANFIS)-predicted soybean yield for site year-3.

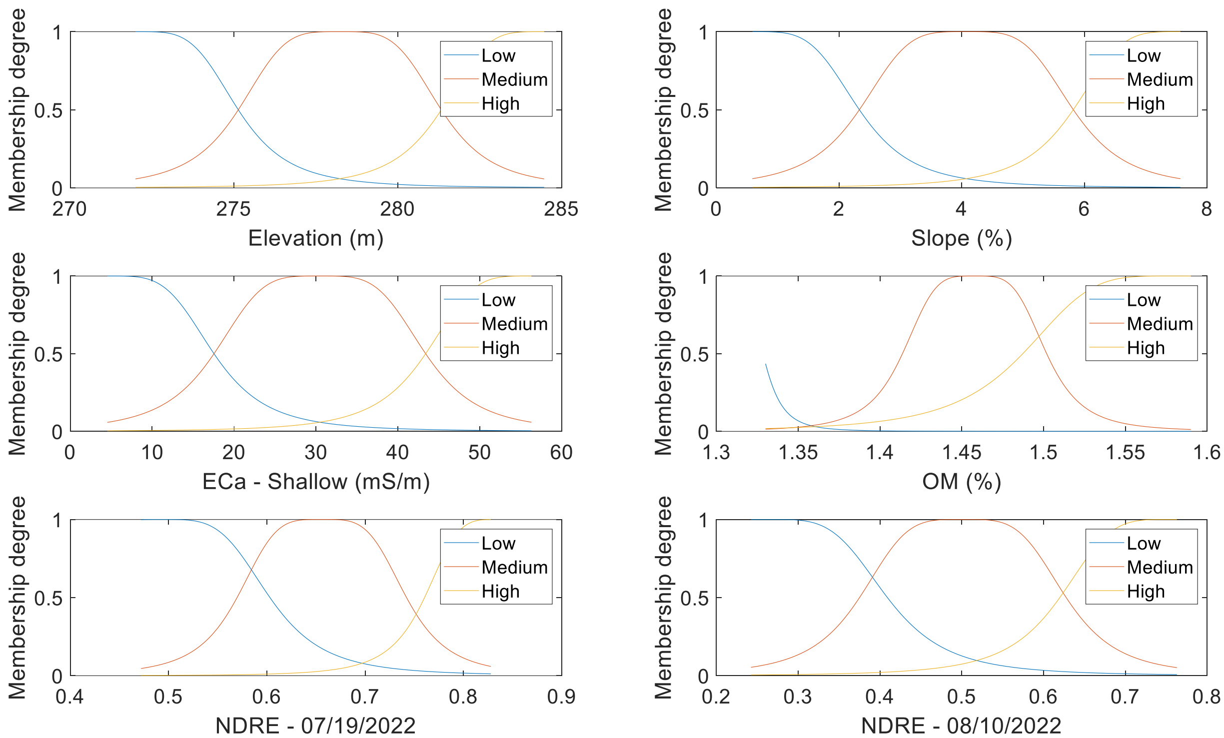


Appendix Figure I.4. Membership function of soil and crop variables for site year-3.

The ANFIS model training for site year-4 demonstrated an improving trend of predictive performance over 70 epochs. Over 70^{th} training epochs, a decrease in the Root Mean Square Errors (RMSEs) was observed for both the training and validation data. A minimum RMSE for the training data and validation data was 1.41 t/ha and 2.05 t/ha, respectively. The coefficient of determination, R^2, stands at 0.77 indicates that the model performed well in predicting corn yield during site year-4.



Appendix Figure I.5. Scatter plot of actual vs. Adaptive Neuro-Fuzzy Inference System (ANFIS)-predicted corn yield for site year-4.



Appendix Figure I.6. Membership function of soil and crop variables for site year-4.

Appendix J - Semivariogram model

Appendix Table J.1. Semivariogram model for measured soil Nitrate-Nitrogen (NO_3-N)

Site years	Nugget	Sill	Nugget/Sill	Spatial Class	Range (m)	RMSE	MSE	RMSSE	ASE	R^2
Site year-1	0.18	0.6	0.30	Moderate	92	6.33	0.06	0.89	6.82	0.47
Site year-2	0.12	1.1	0.11	Strong	277	0.85	-0.03	1.04	0.86	0.58
Site year-3	0.59	1.17	0.50	Moderate	115	11.1	0.016	0.96	11.53	0.13
Site year-4	0.12	1.06	0.11	Strong	229	1.7	-0.006	0.99	1.83	0.65

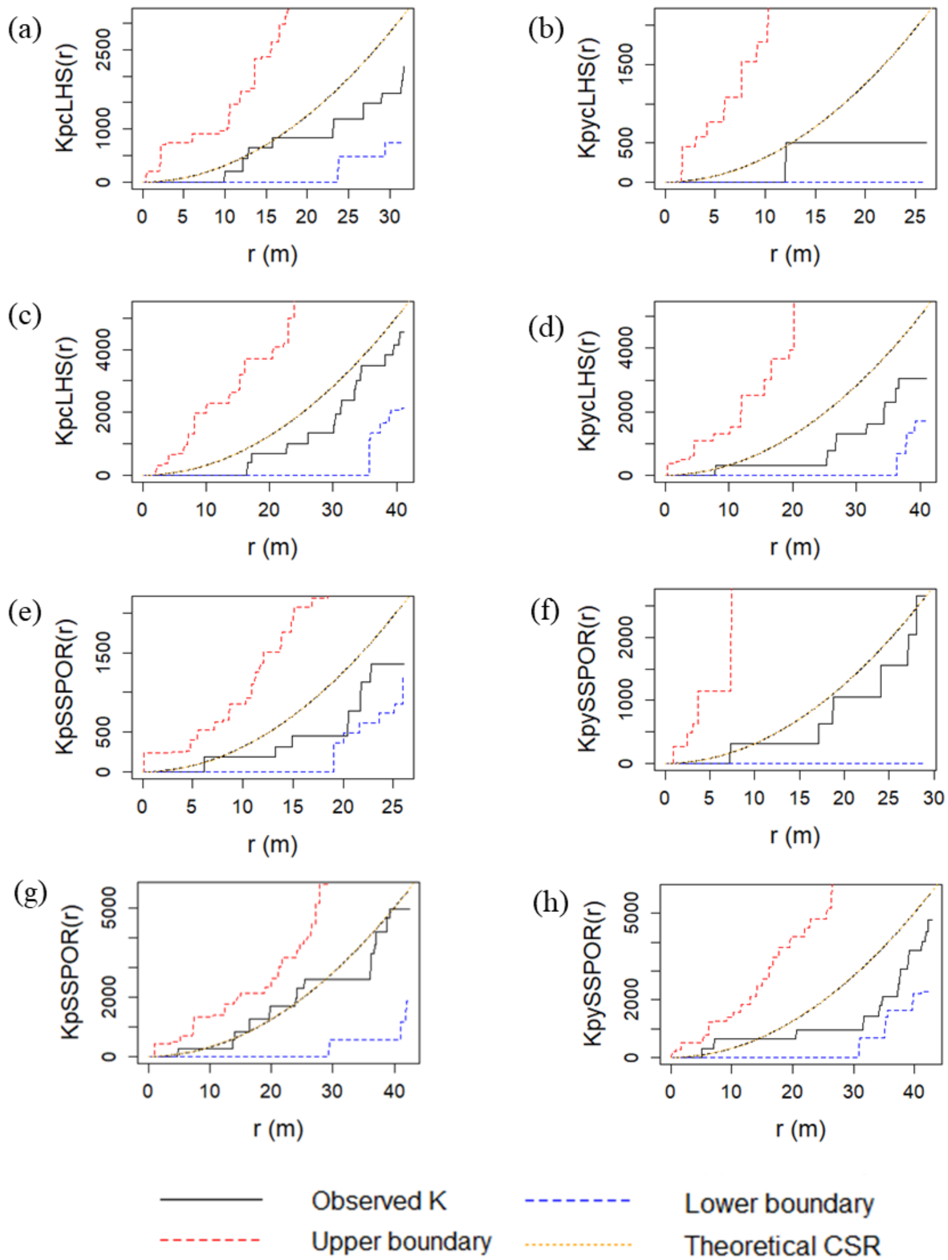
RMSE = Root Mean Square Error, MSE = Mean Square Error, RMSSE = Root Mean Square Standardized Error, ASE = Average Standard Error, R^2 = coefficient of determination

Appendix Table J.2. Semivariogram model for Random Forest (RF) based estimated soil NO_3-N

Site years	Nugget	Sill	Nugget/Sill	Spatial Class	Range (m)	RMSE	MSE	RMSSE	ASE	R^2
Site year-1	0.17	0.81	0.21	Strong	85	4.56	0.01	1.07	4.32	0.15
Site year-2	0.14	1.43	0.10	Strong	261	0.62	0.06	0.79	0.74	0.65
Site year-3	0.01	1.35	0.01	Strong	70	6.11	-0.01	0.94	6.41	0.11
Site year-4	0.01	1.2	0.01	Strong	226	0.95	0.04	0.79	1.16	0.68

RMSE = Root Mean Square Error, MSE = Mean Square Error, RMSSE = Root Mean Square Standardized Error, ASE = Average Standard Error, R^2 = coefficient of determination

Appendix K - Estimated Ripley's K(r) values



Appendix Figure K.1. Estimated Ripley's K (r) values. (a) Site year-1 (Topeka 2021), proximal soil cLHS; (b) Site year-1 (Topeka 2021), proximal soil + yield cLHS; (c) Site year-2 (Rossville 2021), proximal soil cLHS; (d) Site year-2 (Rossville 2021),), proximal soil + yield cLHS; (e) Site year-1 (Topeka 2021), proximal soil SSPOR; (f) Site year-1 (Topeka 2021), proximal soil + yield SSPOR; (g) Site year-2 (Rossville 2021), proximal soil SSPOR; (h) Site year-2 (Rossville 2021),), proximal soil + yield SSPOR

The spatial distribution of soil sampling site configured with cLHS and hybrid cLHS-SSPOR were tested with Ripley's K function to depict randomness or clustering of soil sampling sites. As shown in Appendix H (Appendix Figure K.1), the observed K values indicated a spatial pattern comparing to distribution of soil sampling sites. The solid black line (Observed K) represented the calculated K values, the dotted line (Theoretical CSR) represented expected K values under the hypothesis of complete spatial randomness, dashed blue line (Lower boundary) likely represented a confidence envelope or boundary for lower end, and dash-dot red line (Upper boundary) likely represented a confidence envelope or boundary for the higher end. Ripley's K values resulted a random distribution for all site-years because it spans within the confidence interval boundaries.