DIGITAL SOIL MAPPING OF THE PURDUE AGRONOMY CENTER FOR RESEARCH AND EDUCATION

by

Shams R. Rahmani

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Agronomy

West Lafayette, Indiana

May 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Darrell G. Schulze, Chair

Department of Agronomy

Dr. Jason P. Ackerson

Department of Agronomy

Dr. Zamir Libohova

USDA-NRCS National Soil Survey Center

Dr. Ningning N. Kong

Department of Forestry and Natural Resources

Dr. Kevin T McNamara

Department of Agricultural Economics

Approved by:

Dr. Ronald F. Turco, Head, Department of Agronomy

This dissertation is dedicated to my loving parents, family, Professor Ab. Ghani Ayubi, Dr. Darrell G. Schulze, and Dr. M. Ashraf Ghani president of Afghanistan.

ACKNOWLEDGMENTS

There is a saying, “It takes a village to raise a child.” One could also say, it takes a community to make a Ph.D. It is clear that the work presented in this dissertation could never have been done on my own. The support and contributions that I received through my PhD journey have made it possible for me to reach this great achievement. Therefore, I would like to express my gratitude to those who directly and indirectly helped me along the way.

My deepest appreciation goes to my advisor, Dr. Darrell Schulze, for his endless support, inspiration and guidance. I cannot express how fortunate I am to have you as my advisor. You were always available for discussions and meetings, and I am thankful for all your time and patience. The one thing that I have learned and will strive for is your commitment to quality and the high standards of science. I hope someday to follow in your footsteps and be a scientist and mentor as you.

My academic committee members, Dr. Jason Ackerson, Dr. Zamir Libohova, Dr. Nicole Kong, and Dr. Kevin McNamara, are especially deserving of my gratitude. Dr. Ackerson, thank you for the weekly meetings and introducing me to the R programing language and instilling in me a passion for coding and statistics. Dr. Libohova, thank you for extensive, lengthy, and comprehensive email conversations and fast replies. I am also grateful to you and your family for hosting me while working closely with you on my research. Dr. Kong and Dr. McNamara, you encouraged me to think broadly and motivated me to participate in extracurricular and leadership activities.

I owe special thanks to Dr. Eileen Kladivko, Dr. Darrell Schulze, Dr. George Van Scoyoc, Dr. John Graveel, and Dr. Michael Mashtare for permitting me to serve as a teaching assistant for their classes. It was an immense pleasure and an absolute privilege to work with all of you and with Sherry Fulk-Bringman, the lab coordinator for the Introductory Soil Science course.

I am especially grateful to Purdue University, particularly to the Agronomy Department, for providing countless opportunities and support to attend meetings and conferences and to acquire various certificates and leadership skills. Thanks to both the former and current Agronomy Department heads, Dr. Joseph Anderson and Dr. Ronald Turco, and administrative and clerical staff, Karen Clymer, Patricia (Patti) Oliver, Cheryl Long, Connie Foster, and Dawn Bull, for taking care all the administrative and financial issues and letting me concentrate on my research. The help

of Suzanne Cunningham with language editing is gratefully acknowledged. Thanks to Dr. Gebisa Ejeta and Mr. Gary Burniske for allowing me to participate in the Borlaug Summer Institution on Global Food Security, and Dr. Bruce Erickson for permitting me to participate in the Precision Agriculture program. It is an honor to be part of the Purdue Agronomy Department.

I would like to thank Dr. Phillip Owens, my master's advisor, for providing me the opportunity to come back to Purdue University for my PhD. I also deeply thank my Afghan colleagues for letting me pursue my PhD. I am specifically thankful to Dr. Abdul Qahar Samin whom has now sadly passed, and to Professor Abdul Ghani Ayubi for instilling in me a passion for soil science. It was a great honor working with you. I furthermore extend my gratitude to Abdul Walid Salik, a colleague and a close friend, for taking care of all my academic issues at Kabul University.

Much appreciation goes to Dr. Gary Steinhardt, Michael Wigginton, Nasrat Wardag and Joe Rorick for helping with the fieldwork at the Agronomy Center for Research and Education (ACRE). We had so much fun and plenty of laughs. A whole-hearted thanks to Mr. James (Jim) Beaty, the ACRE superintendent, who provided appropriate tools for the fieldwork and who shared his knowledge of the farm, particularly for the tile drainage mapping project. Thanks also to Jason Adams, manager of the Indiana Corn and Soybean Innovation Center at ACRE for providing the RTK GPS. I would also like to thank Dr. Beth Hall, Indiana State Climatologist, and the staff of the USDA Natural Resources Conservation Service, Lafayette Service Center, and others, for sharing and guiding me to the right data sources.

The daily work would not have been as enjoyable without having my lab-mates, Mercy W. Ngunjiri, Joshua Minai; past and present officemates, Minerva Dorantes, Cheng-Hsien Lin, Richard Smith, Nathan Slavens, Kathryn Kamman, Stephen Boersma, and Nick Roysdon; and friends, Stefanie Griebel and Blake Russell. You colored my life by sharing fun moments and your experiences.

I thank Afghan families, Dr. Zarjon Baha, Dr. Nasser Shinwari, Professor Guljon Saber, Najibullah Wardag, Hakim Hassan, Dr. Abdullah Mirzoy, Saber Atmar, and Ashiqullah Bandawal, and friends, Nasir Wardag, Idris Noor, Nasrat Wardag, Anayat Wardag, and Khoshal Saber, for your encouragement, true friendship, and hospitality during my stay in West Lafayette, Indiana. We had plenty of laughs and enjoyable moments. I will forever carry these memories with me. I am also thankful to Meena Saber, Mariam Alamyar, Lashta Saber, Menna Hassan, Leila Hassan,

and Lema Saber for helping my family. I would also like to acknowledge the members of the Afghan Student Association of Purdue (ASAP) for helping with event planning and organization when I was in charge of the ASAP.

Much love and appreciation go to my parents, father-in-law, mother-in-law, siblings, uncles, aunts, brothers-in-law, sisters-in-law, and close relatives for their moral support. For sure, it was hard to be away from you all, but your love and best wishes always made my life easy. I specifically would like to thank my beloved father, Gul R. Rahmani, and my older brother, Wahid R. Rahmani, for their unconditional support and encouragements. I cannot be thankful enough to have you by my side.

My heart goes to my wife, Zahera Rahmani, for her constant support and shouldering family and parenting responsibilities. Words cannot express how much you helped. I owe you everything. My son Edris Rahmani and daughters Aisha and Hawa, thank you for keeping me refreshed. You are the loves of my life and I am truly blessed to have you all in my life.

Finally, I acknowledge the invaluable financial support of the Office of Agriculture Research and Graduate Education at Purdue University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

ACRE	Agronomy Center for Research and Education
ANN	Artificial Neural Network
ASAP	Afghan Student Association of Purdue
CAD	Computer Aid Design
CART	Classification and Regression Tree
CEC	Cation Exchange Capacity
cLHS	Conditioned Latin Hypercube Sampling
cm	Centimeters
CND	Channel Network Distance
CSC	Cross Sectional Curvature
CTI	Compound Topographic Index
DEM	Digital Elevation Model
DSM	Digital Soil Mapping
est err	Estimation or Prediction Error
ft	Feet
GCS	Geographic Coordinate System
GDG	Geospatial Data Gateway
GIS	Geographic Information System
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
IGWS	Indiana Geological and Water Survey
IN	Indiana
InGCS	Indiana Geospatial Coordinate System
ISDP	Indiana Spatial Data Portal
Isee	Integrating Spatial Educational Experiences
LCCC	Lin's Concordance Correlation Coefficient
LCD	Land Conservation Department
LOI	Loss on Ignition

LOOCV	Leave One Out Cross Validation
ME	Mean Error
MLA	Machine Learning Algorithm
mm	Millimeter
MNLR	Multinomial Logistic Regression
MrRTF	Multi Resolution Ridge Top Flatness Index
MrVBF	Multi Resolution Valley Bottom Flatness Index
MSE	Mean Square Error
MW	Moderately Well Drained
NAD	North American Datum
NAIP	National Agriculture Imagery Program
NCSS	National Cooperative Soil Survey
NRCS	Natural Resources Conservation Service
OIF	Optimal Index Factor
OM	Organic Matter Content
PD	Poorly Drained
pdf	Portable Document Format
pH	Potential Hydrogen
ppm	Parts Per Million
PURR	Purdue University Research Repository
RF	Random Forest
RK	Regression Kriging
RMSE	Root Mean Square Error
RSP	Relative Slope Position
RTK	Real Time Kinematics
SAGA	System for Automated Geoscientific Analysis
SH	Slope Height
SSURGO	Soil Survey Geographic Database
STATSGO	State Soil Geographic Database
SWP	Somewhat poorly drained
T	Trace

TPI	Topographic Position Index
TWI	Topographic Wetness Index
UK	Universal Kriging
USDA	United States Department of Agriculture
USGS	United States Geological Survey
UTM	Universal Transverse Mercator
VIF	Variance Inflation Factor
VPD	Very Poorly Drained
WGS	World Geodetic System
WQFS	Water Quality Field Station

ABSTRACT

The Purdue Agronomy Center for Research and Education (ACRE) has become a hub for state-of-the-art research using automated field-based plant phenotyping. Soils are critical in field-based phenotyping because they vary spatially across a field due to soil forming factors, previous land use, and human modifications. The same genotype is likely to express a different phenotype in different parts of the same field because of soil variability.

The current, polygon-based USDA soil survey, while useful, is not detailed enough to support field-based phenotyping and precision agriculture. We used digital soil mapping (DSM) techniques to accurately predict soil variation across ACRE and produce higher resolution continuous soil maps.

We produced maps of soil organic matter content (OM), cation exchange capacity (CEC), natural soil drainage classes, and tile line locations to encompass both chemical and physical properties that influence crop growth. A lidar-based digital elevation model (DEM) was used to derive terrain attributes that capture topographic variation, the main driver of soil variation at ACRE. Some 178 soil samples were collected for OM and CEC determination and 154 locations were sampled for natural soil drainage class determination. For each dataset, 70% of the points were used for training and 30% were used for evaluation.

The spatial distributions of OM and CEC were determined by universal kriging (UK), Cubist, and random forest (RF) geostatistical models. Similarly, multinomial logistic regression (MNLR), C5.0 decision tree, RF, and artificial neural network (ANN) models were used to predict natural soil drainage classes.

All the DSM models produced similar OM, CEC, and soil drainage class predictions. For OM, R^2 ranged from 0.44 – 0.45, root mean square error (RMSE) ranged from 0.8 – 0.83,

concordance ranged from 0.56 – 0.58, and bias ranged from 0 – 0.22. For CEC, R^2 ranged from 0.39 – 0.44, RMSE ranged from 3.62 – 3.74, concordance ranged from 0.55 – 0.57, and bias ranged from 0 – 0.17. The overall accuracy of the four predictive DSM models for natural soil drainage classes ranged from 66 – 70% and kappa coefficient ranged from 0.53 – 0.59.

The results of the DSM models were also compared to the USDA Soil Survey Geographic (SSURGO) data. For OM and CEC, SSURGO was comparable to the DSM models, but SSURGO underestimated or overestimated both soil properties for a few map units. For natural soil drainage class predictions, the DSM models slightly outperformed SSURGO with an overall accuracy of 64% and kappa of 0.52.

The tile drainage lines map was based on visual interpretation of aerial photographs, physical paper maps, electronic as-built maps, and the knowledge of the ACRE superintendent. For 27 tile locations determined physically, the mapped locations occurred within ±1.23 m of the true tile locations.

CHAPTER 1. GENERAL INTRODUCTION AND MOTIVATION

With the Purdue Moves Plant Science Initiative, the Agronomy Center for Research and Education (ACRE) has become a state-of-the-art center for field-based phenotyping. Even though the main focus of plant phenotyping is on plants above ground, soil variability below ground cannot be ignored because of its influence on plant growth and development. Understanding soil variability is important and perhaps underappreciated in field-based phenotyping. The same genotype will likely to show slightly different phenotypic characteristics in different parts of the same field due to soil variation.

Researchers at ACRE currently rely on traditional soil survey maps from the U.S. Department of Agriculture, Natural Resources Conservation Service (NRCS). These maps are useful but do not provide sufficient spatial resolution to describe soil variability for areas less than one ha (Soil Science Division Staff, 2017). Since the traditional soil survey maps delineate soil map units as polygons, sharp changes in reported soil properties often occur at the polygon boundaries. Therefore, higher resolution and continuous soil maps are needed to capture field scale variability and soil functional properties that impact plant responses and hydrological processes. This dissertation focuses on digital soil mapping (DSM) techniques for capturing soil spatial variability at the scale of an individual farm (570 ha), the Purdue Agronomy Center for Research and Education.

1.1 Research Objectives and Hypothesis

The overall objective of this study is to explore field scale soil variability and develop spatially explicit digital soil class and soil property maps relevant for agronomic research,

particularly to supporting field-based phenotyping at ACRE. The information will be then made available for use by other projects at ACRE. Specific objectives are:

1. Develop continuous soil organic matter (OM) and cation exchange capacity (CEC) maps using terrain attributes generated from a high-resolution digital elevation model combined with data from soil samples collected in the field.

- Much of the soil variability at ACRE is due to wetness differences. Therefore, terrain attributes such as topographic wetness index (TWI), topographic position index (TPI), and others that quantify water flow and accumulation will be calculated from the DEM. These terrain attributes infer water redistribution by utilizing algorithms to describe topography by creating a unique index value for each pixel in the digital elevation model. The terrain indices will allow us to make more detailed soil maps.

2. Develop a soil drainage class map using terrain attributes generated from a high-resolution digital elevation model combined with field determinations of natural soil drainage class.

3. Prepare a map of the location of tile drainage lines based on aerial imagery, expert knowledge, available physical paper maps, and, when available, georeferenced, as-installed data.

- For this research objective, we will assemble all available aerial photographs for ACRE. Aerial photos are used for assessing field conditions and detecting soil or crop problems that might otherwise go unnoticed at ground level (Reising et al., 1988). Air photos acquired at different times are helpful in capturing different features. The light and dark patterns of bare ground images often correlate to soil differences. Some of the aerial photos show the location of tile lines and areas that pond regularly. We will also use the knowledge of the farm manager and physical paper maps to determine the locations

of tile lines for areas where the locations of the tiles cannot be determined from aerial imagery.

The main hypothesis of this study is that soil variability at ACRE is the result of topography and tile drainage and can be predicted by using terrain analysis and aerial imagery.

1.2 Organization and Outline

This dissertation consists of six chapters. Chapter 1 provides a general introduction to soil spatial variability, the justification for this study, the objectives, and the hypothesis of the research topics. Chapter 2 describes the principles and concepts of soil surveys and digital soil mapping and reviews previous studies related to this field of research. Chapters 3, 4 and 5 address objectives 1, 2 and 3, respectively. The final chapter, Chapter 6, provides information about the utilization and delivery of spatially explicit digital soil information. The appendix contains additional materials relevant to the study.

CHAPTER 2. LITERATURE REVIEW

2.1 Predicting and Mapping Soil Spatial Variability

Soil is a complex and non-renewable natural resource that sustains life on the earth by providing essential ecosystem functions. Soils are increasingly under threat. According to the European Commission (2002), soils need to be protected from the following eight major threats: erosion, organic matter reduction, soil compaction, soil sealing, soil contamination, floods and landslides, salinization, and soil biodiversity reduction. Furthermore, climate change and water scarcity are other factors that affect the ability of soils to optimally perform their functions. A comprehensive understanding of soils and their spatial distribution over a landscape is important for the proper use of soil resources and to protect them from degradation.

Soil surveys document how soils vary across landscapes. A common misconception is that a soil survey is equal to, or the same as, a soil map. A soil survey is more than just a soil map (Brady and Weil, 2002). *Soil surveys* describe the characteristics of the soils in a given location, classify the soils based on a standard system of taxonomy, delineate the soil boundaries on a map, store soil property information in an organized database, and make interpretations about the suitability and limitation of each soil for various uses, as well as likely responses to management systems (Soil Science Division Staff, 2017). The information assembled in a soil survey can be used for land use planning and evaluation of the impact of land use on the environment. The public and scientific community can use the information in soil surveys for informed decision making.

A soil survey has the same basic objective for all kinds of land. However, the number of soil mapping units and the spatial detail of the mapping vary based on the needs of end users and the complexity of the soil landscapes in a particular survey area. Soil surveys may be conducted at various levels of detail, referred to as the *order* of the soil survey. First order soil surveys provide

detailed information, usually over relatively small areas, while fifth order surveys provide reconnaissance data over large areas (i.e. regions or continental scale) (Soil Science Division Staff, 2017).

The three main phases in soil surveys are assessment, mapping, and interpretation. In the assessment phase, soil scientists determine which soil properties are important for that specific type of survey and for land evaluation and management recommendations. Soil mapping is the most widely recognized phase. Soil scientists delineate the boundaries of soils at a specified map scale based on information from soil sampling and/or an earlier research phase. In the interpretation stage, soil scientists provide information about land use potential, management practices, avoidance of hazards, and economic evaluations of soil data (Dent and Young, 1981; Soil Science Division Staff, 2017).

2.2 Conventional Soil Mapping and its Limitations

In conventional or traditional soil mapping, soil scientists develop a conceptual, mental soil-landscape model through intensive fieldwork (Fig. 2.1). Soil scientists delineate polygons of similar soils on aerial photographs based on their knowledge of the distribution of landscape units. In addition to aerial photographs, soil surveyors use Landsat images, and geology and topographic maps to identify landscape features. They then conduct field observations to verify and refine their concepts (USDA-NRCS, 1998; Soil Science Division Staff, 2017).

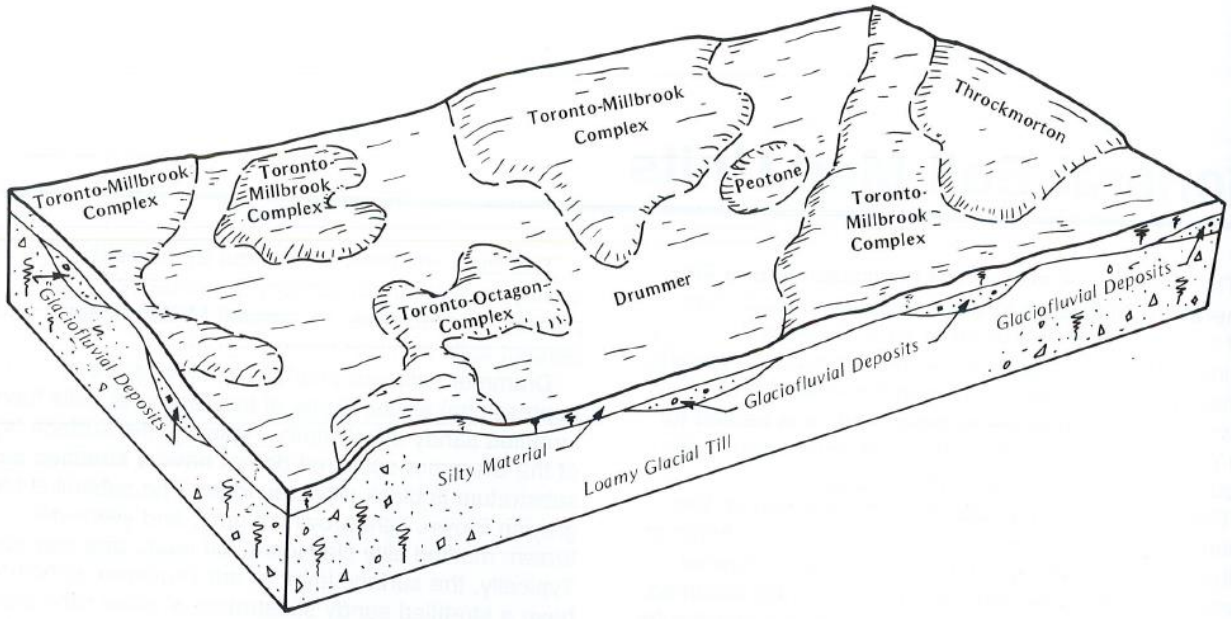

Figure 2.1: Block diagram of soil-landscape model developed by an expert soil scientist for the Drummer and Toronto-Millbrook Complex mapping units of Tippecanoe County, Indiana (USDA-NCSS, 1998).

The output and procedure of the conventional mapping approach has its limitations. For example, the USDA soil survey maps, while very useful, do not estimate soil variation for areas smaller than about one hectare (2.5 acres) (Soil Science Division Staff, 2017). Additionally, soil polygons represent soil classes or properties as spatially homogenous within the polygons. This often results in sharp transitions to adjacent polygons with different soil classes or properties. This method for modeling soil spatial variability does not account for the continuous spatial variation of soil. Furthermore, information generated through the conventional approach is qualitative, therefore it has a limited use in quantitative studies (Hartemink et al., 2010; Boettinger, 2010). Due to advances in computer technologies and various statistical models, it is now possible to capture soil variability on a more continuous and quantitative basis. Digital soil mapping (DSM)

is an emerging technology that can play a role in overcoming some of the limitations of conventional soil mapping (McKenzie and Ryan, 1999; Kempen et al., 2012).

2.3 Digital Soil Mapping

There have been a significant number of DSM projects conducted in various parts of the world and most of those conducted prior to 2002 are discussed by McBratney et al., (2003). Advancement in computer technology and freely available high-resolution data, such as digital elevation models and remotely sensed and proximally sensed data, are major reasons for the rapid growth of DSM methodologies. Digital soil mapping (DSM) is the creation of digital soil type and/or property maps based on spatially explicit environmental variables and measurements made in the field and laboratory (McBratney et al., 2003). DSM can be used to develop initial soil survey maps, update existing surveys, assess risk, and generate soil interpretations (Carré et al., 2007).

Conceptually there is no difference between DSM and conventional soil mapping. Both methods rely on soil-landscape models to predict soil properties at unobserved locations (Hudson, 1992). They both need soil and covariate data for model building. The main difference is how the models utilize, process, and display the soil information from the input data. Conventional soil mapping captures soil-landscape models qualitatively, while DSM quantifies the soil-landscape model numerically by establishing relationships between soil forming factors and soils (Kempen et al., 2012; Soil Science Division Staff, 2017).

Each digital soil mapping project is unique, and aspects of each project may vary with respect to the goal of the project, the availability of environmental variables, and the method of prediction. The stages and processes in producing a soil map, however, should be consistent in all digital soil mapping projects. Soil Science Division Staff (2017) provide a complete and detailed outline of the various stages and processes in DSM.

2.4 The CLORPT and SCORPAN Models

The properties of soil vary over space and this variation is not random. Jenny (1949) is credited with developing the conceptual model that ascribes soil variation (S) on a landscape as the result of climate (cl), living organisms (o), topography or relief (r), parent material (p), and time (t), mathematically expressed as:

$$S = f\,(cl, o, r, p, t) \quad [1]$$

The CLORPT model has been used in conventional soil mapping, but it is not spatially explicit nor quantitative (Soil Science Division Staff, 2017). In order to represent soil and its associated environmental variables in a spatial context, McBratney et al. (2003) proposed the SCORPAN model. The SCORPAN model is a quantitative and inference-based model that can be expressed in two general forms:

$$S_c = f\,(s, c, o, r, p, a, n) \text{ and } S_p = f\,(s, c, o, r, p, a, n) \quad [2]$$

In this model S_c refers to a soil class and S_p a soil property, while s represents other soil information at that point, c climate, o living organism (vegetation or flora, fauna, and human activity), r relief or topography, p parent material, a age or time, n the spatial position, and f the soil spatial prediction function. As opposed to the CLORPT model, the SCORPAN model uses soil itself as a covariate for soil prediction. McBratney et al. (2003) provide in-depth information on the various sources for obtaining the seven SCORPAN covariates for DSM. In the section below, we will only discuss the sources for obtaining r, the relief factor.

2.5 Data Sources for Relief and the Selection of Appropriate Terrain Predictors

McBratney et al. (2003) reviewed more than 130 papers about DSM and observed that among the seven SCORPAN variables, relief (r) was the most extensively used variable in DSM studies. Currently, relief, or topographic information, is mainly derived from digital elevation

models (DEMs). A DEM provides quantitative information about the continuous variation of the Earth's surface. Remotely sensed elevation data, digitized contour and stream lines data, and point measurements of elevation, either from conventional land surveys or vehicle-mounted high-resolution global positioning system (GPS) receivers, are different sources for acquiring and generating elevation data (McBratney et al., 2003).

Digital terrain analysis is a useful approach for acquiring topographic information from DEMs and provides information about elevation, stream networks and other terrain associated attributes, along with their geographic position (Moore et al., 1993; Wilson and Gallant, 2000). Terrain covariates are calculated from DEMs and have been widely applied in digital soil mapping. Terrain covariates can be broadly classified as first order and second order derivatives, also known as primary and secondary (or compound) attributes. The first order derivatives are directly calculated from a DEM, while second order derivatives result from combinations of first order attributes (Moore et al., 1991). Slope, aspect, plan and profile curvatures, and upslope contributing area are the major first order derivatives. Stream power index, sediment transport capacity index, and topographic wetness index are the major second order derivatives. For a thorough list of terrain attributes, see Wilson and Gallant (2002). Landscape classification, which has a strong link to soil properties such as organic matter (Pennock et al., 1987), can be easily generated using a DEM (MacMillan et al., 2003). The generated landform classes have been used in soil mapping as environmental predictors (Smith et al., 2012).

It is necessary to understand the details and functionality of each terrain attribute before generating a DSM. The terrain covariates based on hydrology need to be calculated using hydrological units like watersheds (Soil Science Division Staff, 2017). Deriving many terrain attributes from a DEM and collecting other ancillary data such as soil legacy data and indices

derived from remotely sensed data are relatively inexpensive and easy. Additionally, it is possible to use all of these attributes and indices in predictive models. However, selecting appropriate covariates is recommended to prevent model uncertainty and overfitting. Additionally, models with fewer covariates are easier to interpret and faster to compute (Soil Science Division Staff, 2017).

Pedological knowledge and various statistical methods such as: optimal index factor (OIF), variance inflation factor (VIF), Pearson's correlation coefficient (r), principle component analysis (PCA), and forward and backward selection, to name a few, can be used to select optimal terrain attributes and other ancillary data for DSM. The resulting digital soil map is more accurate when it is derived by an expert soil scientist. Therefore, for better DSM outcomes, it is recommended that expert soil knowledge be used along with statistical procedures in covariate selections (Kempen et al., 2009; Kuhn and Johnson, 2013).

2.6 Collecting Field Soil Point Observations

Sampling designs for collecting soil samples play a critical role in soil spatial prediction modeling. Different sampling designs result in different soil distributions and ultimately impact DSM accuracy (Brus et al., 2006; Van Groenigne et al., 2000; Heim et al., 2009). Taking into consideration that soil sampling is time consuming and resource intensive, selecting an efficient soil sampling scheme is important. Poor sampling design can introduce significant biases, which may result in over or under predictions by models (Congalton, 1991). Four common probabilistic sampling designs that are often used in environmental correlation are discussed briefly below.

2.6.1 Simple Random Sampling

Simple random sampling is used in areas where prior soil information is limited. In this sampling approach, each sampling location has an equal random chance of being selected. This sampling design provides unbiased estimates of the mean and variance but requires a large number of samples to reduce prediction errors (Lee et al., 2017; Palmer, 2003). Additionally, this design may cause spatial clustering of sample locations and may not provide good geographical coverage over a specific area (Avery and Burkhart, 1994; Yang et al., 2016). This method works best in small, homogenous areas. Howell et al. (2004) found that models using simple random sampling compared to models using purposive or subjective sampling were more accurate in predicting soil morphological features. Purposive sampling is typically used in conventional soil surveys when soil sampling locations are determined by the intuition of expert soil scientists. The poor performance of purposive sampling might be due to the subconscious bias of soil experts (Buol et al., 1997; McKenzie and Ryan, 1999).

2.6.2 Systematic Sampling Design

Systematic sampling design places a grid of sampling units over the sampling location. Squares, triangles and hexagons are commonly applied grid patterns used in systematic sampling designs. This sampling design will not be effective in irregularly shaped areas or in areas that have periodic distributions (systematic variations) (Brus, 2019). In such case, the collected samples are less precise than using a simple random sample design (Lark and Cullis, 2004; Sparks et al., 1996). If experts are aware of such periodicity, they can use a systematic sampling design but with higher care (Lark and Cullis, 2004).

2.6.3 Stratified Random Sample

In this sampling design, the area to be mapped is spatially sub-divided into different strata and a random sample is selected from each stratum. Strata are typically based on landscape characteristics such as landform, slope gradient, parent material, or land cover type. It is assumed that these strata have strong correlations with the target soil feature(s) (Soil Science Division Staff, 2017). Generally, a stratified random sample is more efficient than a simple random sample (Palmer, 2003). Stratification helps to prevent the spatial clustering of samples (Walvoort et al., 2010). Several studies (Gessler et al., 1995; McKenzie and Ryan, 1999; Park et al., 2001; Minasny and McBratney, 2006) developed various stratified random sampling schemes, all aimed at reducing the overall estimation errors by distributing sampling locations in feature and/or geographic space. Distributing sampling locations minimizes the spatial dependency among the model residuals by covering the multivariate distribution of ancillary data. Gessler et al. (1995) utilized a compound topographic index (CTI) to stratify a landscape into equal areas. In order to prevent redundant sampling locations, Gessler et al. (1995) developed a CTI variogram to determine the extent of spatial dependency or autocorrelation. Samples were randomly assigned to each CTI stratum at distances further apart than the extent of spatial autocorrelation.

2.6.4 Conditioned Latin Hypercube Sampling

The conditioned Latin hypercube sampling (cLHS) design has been widely used in DSM (Minasny and McBratney, 2006). This is a special type of stratified random sampling which uses ancillary landscape data for obtaining representative samples. Conditioned Latin hypercube sampling is considered an efficient method of soil sampling because it operates based on the combined powers of stratification, randomness, and the efficient allocation of samples from multivariate distributions (McKay et al., 1979; Minasny and McBratney, 2006; Worsham et al.,

2012, Silva et al., 2014; Kidd et al., 2015). In a comparative study, Minasny and McBratney (2006) found that cLHS is the most effective method for replicating the distribution of soil variables when compared to simple random sampling and equal area stratified random sampling methods.

2.7 Spatial Inference Models

Since we will discuss and compare several predictive models for soil property and soil class maps in subsequent chapters, only general information is provided about various DSM models here.

After selecting the optimal set of SCORPAN variables and collecting training data (Fig. 2.2), a model is needed to predict the soil classes or properties of interest. Various predictive models exist to quantify the relationships between soil data and related environmental factors in a spatial context. All models operate based on the equation: $S = f(Q) + e$; proposed by McBratney et al. (2003), where S stands for a soil attribute or soil class, Q represents the SCORPAN auxiliary environmental predictors, and e is the error of prediction. These predictive models can be classified using three main approaches: (1) soil survey, (2) geostatistical, and (3) data mining. We provide general information about each of the approaches in the sections below.

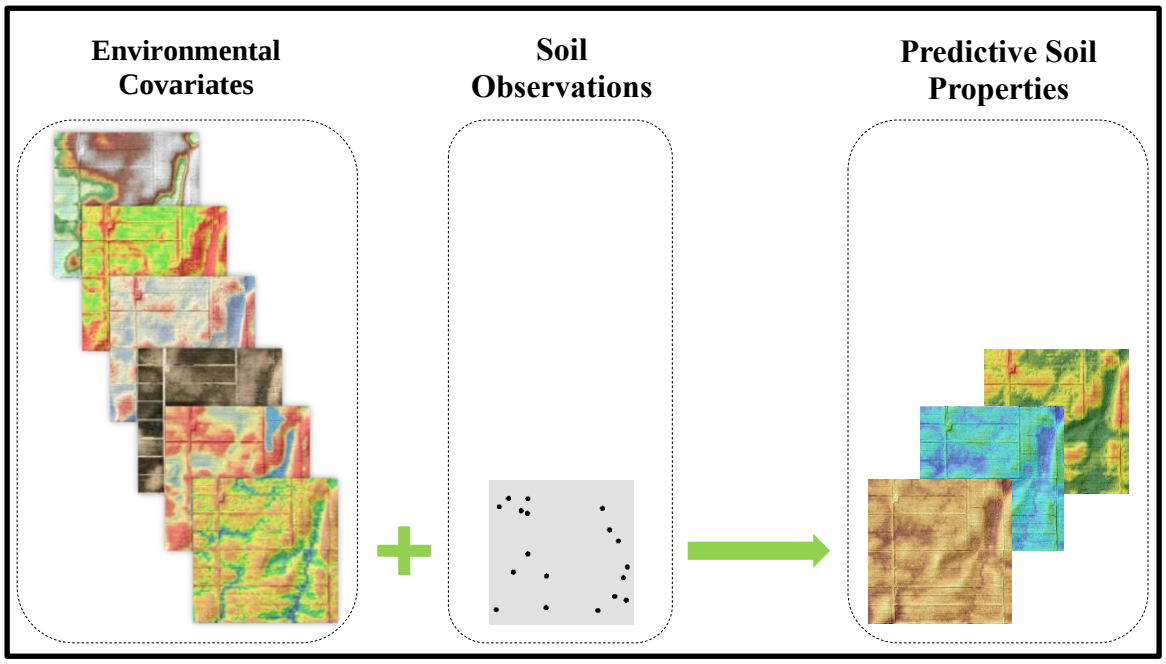

Figure 2.2: Principle and workflow of digital soil mapping.

There is no one unique model that will predict most accurately for any particular DSM project because the performance of each prediction model depends on the structure of individual datasets and the method used to select the covariates (Soil Science Division Staff, 2017). The most appropriate way to find an optimal prediction model is to apply several models and then select the one with the highest accuracy. Due to the ease of interpretation, simple models are preferred over complex models. Therefore, if the performance of a simple model is comparable to more complex models, based on the principle of Ockham's Razor, the simple model would likely be the one selected (Soil Science Division Staff, 2017).

2.7.1 Soil Survey Approach

The soil surveyor method is also known as the CLORPT method. This method uses the knowledge of expert soil surveyors to build the predictive soil-landscape model. Walter et al., (2006) experimented with various methodologies for capturing this expert knowledge. Some of

the methodologies used for modeling expert knowledge include, (1) translation of narrative modeling into a set of if-then rules (McKenzie and Ryan, 1999; Cole and Boettinger, 2006), (2) fuzzy inference systems (Zhu et al., 1996), and (3) conditional probabilities calculated from legacy soil maps (Lagacherie et al., 1995). In the next section, we will discuss only the fuzzy inference soil survey approach because it is a common method for modeling expert knowledge.

The Fuzzy Inference System

The Soil Land Inference Model (SoLIM) developed by Zhu et al., (1996), is a good example of using the fuzzy inference or fuzzy logic system for mapping qualitative soil-landscape relationships. In qualitative modeling, the fuzzy logic algorithm provides information about a soil's similarity to a particular class. With fuzzy logic, a soil will have partial membership in more than one class. The scale of soil membership in a class is set between 1 and 0, where 1 is perfect similarity or membership and 0 is no membership. The values of each terrain attribute associated with a soil class are used to define the membership function in SoLIM. Libohova (2010) provided an example for the Pekin soil map unit that has a slope between 2 – 12%, where a slope of 7% would be considered as the optimum membership value. The Pekin soil will have 100% membership for a raster pixel identified as having a 7% slope. The membership will decrease to 50% if the slope decreases to 2% or increases to 12%. The membership will further decrease if the slope goes beyond the slope range of the Pekin soil, meaning that environmental conditions are less than ideal for this soil.

Several studies have been conducted and utilized fuzzy set theory for soil-landscape modeling (Libohova et al., 2010; Grunwald et al., 2001; Lagacherie et al., 1997; Zhu et al., 1996; McBratney and de Gruijter, 1992). The fuzzy inference method of soil mapping, however, has not

been widely adopted (Grunwald and Lamsal, 2006). A major reason is the interpretation of the output is difficult because there are complexes of fuzzy outputs instead of one map.

2.7.2 The Geostatistical Approach

This approach is also known as pedometrics and uses statistical and mathematical approaches to predict soil properties of interest. Geostatistics is a data driven method and uses georeferenced point observations and gridded covariates to predict soil properties at unvisited locations (Hengl and MacMillan, 2019). Universal kriging (UK), which is analogous to regression kriging (RK) (Odeh et al., 1995), is a common geostatistical method in DSM, when auxiliary variables are spatially exhaustive (McBratney et al., 2003). It predicts soil properties based on sums of deterministic trends and the spatially autocorrelated stochastic residuals. The former (deterministic trend) is derived from the regression of auxiliary variables, while the interpolated residuals along with regression coefficients are derived from soil observations (Heuvelink et al., 2006).

In geostatistics, a semivariogram (Fig. 2.3) is used to describe the "law of geography," which indicates that close-by objects are more correlated and alike than more distant ones. In other words, a semivariogram shows how data are correlated with distance and it is used to model the residuals. Based on McBratney and Pringle (1999), the semivariogram is using the semivariance function (Eq. 3), which is the mean variance between two sampling points to measure the spatial autocorrelation. The semivariance uses the following equation (Eq. 3) to interpolate the residuals (Webster and Oliver, 2007):

$$\Upsilon(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \quad [3]$$

where $\Upsilon(h)$ is the semivariance, $N(h)$ represents the number of pairs of observations separated by a lag distance h, and $Z(x_i)$ and $Z(x_i + h)$ are values of regionalized variables at sites x_i and $x_i + h$, respectively.

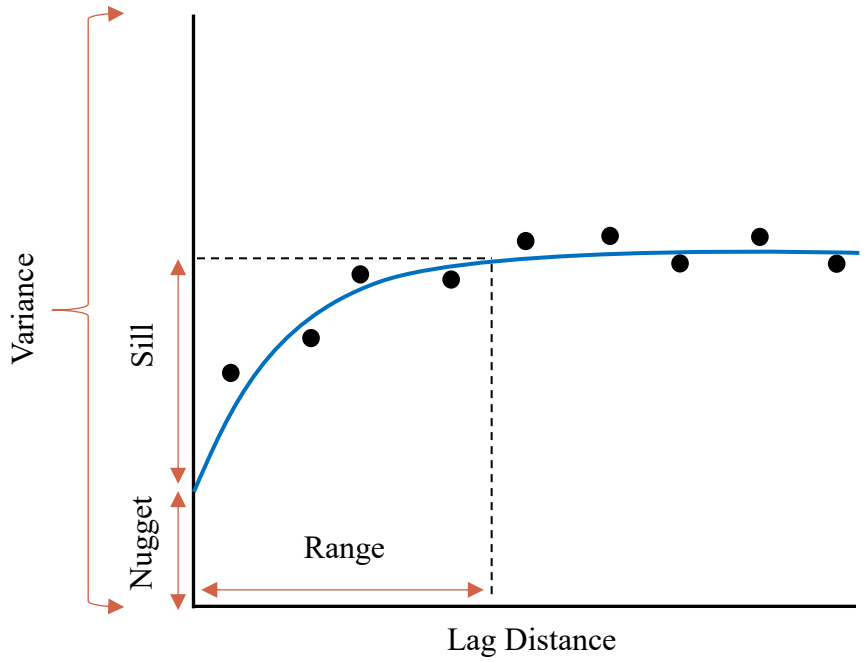

Figure 2.3: Theoretical variogram model.

After interpolating the residuals through semivariance, UK will use the derived information and predict the target properties based on the following equation:

$$\hat{Z}(x_0) = \hat{m}(x_0) + \hat{e}(x_0) = \sum_{k=0}^{p} \hat{\beta}_k * q_k(x_0) + \sum_{i=1}^{n} \lambda_i * e(x_i) \quad [4]$$

where $\hat{Z}(x_0)$ is the estimated value at an unobserved location and $\hat{m}(x_0)$ is the fitted deterministic part, which is not constant and varies within the surrounding neighborhoods to represent the local drift, $\hat{e}(x_0)$ is the estimated residual, $\hat{\beta}_k$ are deterministic model coefficients, λ_i are kriging weights, and $e(x_i)$ is the residual at location (x_i).

Bishop and McBratney (2001) concluded that UK outperformed other statistical and geostatistical models. However, Scull et al. (2005) found that UK did not provide better results

when compared to multiple linear regression. The main interpretation for such poor performances of UK was that soil observations on a landscape are not sampled at distances closer than the average range of spatial dependency.

2.7.3 Data Mining Approach

The data mining method, which is also known as the machine learning algorithm (MLA), hypothesizes that all the required information for a soil prediction is contained within the data (Dobos et al., 2006). There are various MLAs that are applied in DSM projects (Brungard et al., 2015; Heung et al., 2016). Decision tree models are common machine learning algorithms that have become increasingly popular methods for DSM. The outcomes of decision tree models are easily comprehensible and readily interpretable (Odgers, 2017). Linear regression will fail if no linear relationship exists between target and environmental variables, or if an interaction exists between the environmental variables. In such conditions, decision tree algorithms are recommended for use (Molnar, 2019). Decision tree models, unlike linear models, do not make any assumptions about the distribution of residuals (Hastie et al., 2009). One of the main flaws of decision tree models is that they are more susceptible to overfitting than linear models (Odgers, 2017).

In tree-based modeling, the data is split multiple times based on certain cutoff/threshold values of environmental features. This results in different subsets of the dataset. Subsets are either intermediate subsets, which are also known as intern nodes, or terminal subsets, which are also known as leaf nodes. The average outcome of the training data in a final node is used for prediction (Molnar, 2019). Cubist, artificial neural networks, random forest (RF), multinomial logistic regression, support vector machine, classification and regression trees (CART), and Quinlans's C5.0 algorithm are tree-based algorithms that have been widely used in DSM (Lacoste et al., 2014;

Heung et al., 2016; Junjun et al., 2018; Sharififar et al., 2019). The main differences among these algorithms are the structure of the tree (number of splits per node) and the criteria used for making the splits and/or when to stop the splitting (Molnar, 2019).

2.8 Assessing the Quality of Digital Soil Maps

Due to the quantitative (statistical) nature of predictive models, a digital soil map is not perfect and is subject to error. Quality assessment of a digital soil map is critical because the generated product will be used in decision making and risk assessment analysis. The quality of a soil map can be assessed with calibration and/or validation data. Using calibration or training data overestimates the actual accuracy (Refaeilzadeh, 2009). Therefore, it is preferable to evaluate the prediction accuracy using independent or validation data. Quality measures are quantified differently for soil properties and soil class (categorical) maps.

Mean square error (MSE), root mean square error (RMSE), and coefficient of determination (R^2) are common measures used for assessing continuous soil properties maps. MSE is a single value and measures the average squared difference between predicted and observed values. RMSE is the square root of MSE. Both MSE and RMSE provide information about the goodness of fit of the predictive models. The smaller the values of MSE and RMSE, the better the fit or more accurate the model. The coefficient of determination (R^2) provides information about the portion of the data explained by the models.

Various statistical measures exist to evaluate the quality of soil class or categorical soil maps. All of these statistical measures are based on an error or a confusion matrix (Brus et al., 2011). Overall accuracy or map purity, user's accuracy, producer's accuracy, and kappa coefficient of agreement are the most important measures for evaluating the quality of categorical soil maps (Congalton, 1991). Overall accuracy or map purity is the proportion of the correctly classified

observations in a given dataset. User's accuracy, also known as error of commission, indicates the probability that a pixel on the map truly matches the observed soil class on the ground. Producer's accuracy, also known as error of omission, shows the probability that an observed soil class on the ground is classified as such on the map (Congalton, 1991). The kappa coefficient of agreement measures the difference between the observed and expected agreements. It lies between -1 and 1, where 1 is a perfect agreement, 0 is exactly what would be expected by chance, and negative values show less agreement than chance (Viera and Garrett, 2005). Table 2.1 shows the scale for kappa coefficient agreement of categorical data (Landis and Koch, 1977).

Table 2.1: Interpretations of kappa values (Landis and Koch, 1977).

Kappa Value	Degree of Agreement
<0.00	Poor or less than chance agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

2.9 Validation Methods for Digital Soil Mapping

Internal validation, data splitting, cross-validation, and independent validation based on additional probability sampling are common validation approaches for assessing the accuracy and reliability of a prediction model (Brus et al., 2011; Soil Science Division Staff, 2017). For a true and an unbiased evaluation, independent validation is recommended over internal validation (Malone et al., 2017).

The data splitting method of validation, which is also known as the random holdback method, splits the data into two parts: calibration data and test data. The test data is held out and not used during model building. Typically, 10 to 30 percent of the available data is reserved for

validation (McBratney et al., 2003; Soil Science Division Staff, 2017). This type of validation avoids overlaps between the calibration and test data, thus improving accuracy estimations. A downside of this procedure is that not all of the available data is used for calibration, which is a particular concern if the point data is limited. The validation data may be valuable for model calibration and if it is held out, the prediction may suffer (Refaeilzadeh, 2009).

Similar to data splitting, cross-validation data is also divided into two segments (Efron and Tibshirani, 1994). The first segment is used to "learn" or train the model, and the second segment is used to validate the performance of the model. The main difference between this procedure and data splitting is that in cross-validation the data splitting is repeated, thus making cross-validation more efficient than data-splitting. K-fold cross-validation is the basic form of cross-validation. In k-fold cross-validation, the data is initially divided into k equally sized folds. Afterwards, k iterations of training and validation are performed such that each time a different segment or fold of the data is held-out for validation while the rest (k-1 folds) are used for model training. Leave-one-out cross-validation (LOOCV) is the most common form of k-fold cross-validation (Efron and Tibshirani, 1994). The LOOCV trains the model with n-1, and tests it with the one observation that was omitted. The accuracy estimate gained from LOOCV is almost unbiased but contains a high degree of variance (Efron, 1983). Leave-one-out cross-validation is extensively used when data are rare. If the initial collected data are biased, then a true prediction will not be captured by cross validation (Soil Science Division Staff, 2017).

Independent validation is the optimal way to capture true prediction accuracy. In order to avoid bias, collection of additional independent data based on a probabilistic sampling design is recommended (Stehman, 1999; De Gruijter et al., 2006). Similarly, in a review, Brus et al. (2011) concluded that validation based on probability sampling is preferred when compared to data

splitting (random holdback) and cross-validation because unbiased estimates of soil mapping quality are calculated with data that were collected free of model assumptions. Any soil sampling design that utilizes some form of random selection is known as a probability soil sampling design. Stratified random sampling, systematic sampling, and cLHs are considered probability-based sampling designs.

CHAPTER 3. HIGH RESOLUTION DIGITAL SOIL ORGANIC MATTER CONTENT AND CATION EXCHANGE CAPACITY MAPS

Abstract

Soil organic matter content (OM) and cation exchange capacity (CEC) are important soil properties in describing nutrient availability for plant growth. Accurate, high-resolution spatial information of OM and CEC are needed for high-resolution farm management (e.g. precision agriculture and sustainable land management). The objectives of this study were to: 1) determine the spatial distribution of soil OM and CEC in a relatively low relief area using only point measurements of OM and CEC and lidar elevation data, and 2) compare the prediction accuracy of OM and CEC maps created by universal kriging (UK), Cubist, and random forest (RF). For this study, 174 soil samples based on the conditioned Latin hypercube sampling (cLHS) method were collected from 0 to 10 cm depth. The topographic wetness index (TWI), topographic position index (TPI), multi resolution valley bottom flatness (MrVBF), and multi resolution ridge top flatness indices (MrRTF) generated from the lidar data were used as covariates in model predictions. Based on an independent evaluation, we found no major differences in the prediction performance of all selected models. For OM, the predictive models provide results with R^2 (0.44 – 0.45), RMSE (0.8 – 0.83), bias (0 – 0.22), and concordance (0.56 – 0.58). For CEC, the R^2 ranged from 0.39 – 0.44, RMSE ranged from 3.62 – 3.74, bias ranged from 0 – 0.17, and concordance ranged from 0.55 – 0.57. We also compared the results to the USDA Soil Survey Geographic (SSURGO) data. For both OM and CEC, SSURGO was comparable with our predictive models, however SSURGO overestimated or underestimated the selected properties for a few mapping units.

3.1 Introduction

Detailed and accurate spatial soil information is needed for agricultural and ecological decision-making. Conventional, polygon-based soil maps are the main data source for these applications. In the United States, the Soil Survey Geographic database (SSURGO) and the State Soil Geographic (STATSGO) database maintained by the Natural Resources Conservation Service (NRCS) are extensively used for many applications. These polygon-based maps were originally developed for land management and may not be suitable for quantitative modeling that needs more spatially accurate soil property data (Nauman et al., 2012). Map unit polygons often contain more than one major soil component as well as a number of minor soil components, which reduces the map unit purity of these conventional maps. For applications like precision crop management and high throughput phenomics research, more detailed maps are needed than what is available in the SSURGO database.

Digital soil mapping (DSM) is an approach for overcoming the limitations of traditional soil polygon maps and for improving the accuracy of soil property predictions at a finer resolution (McBratney et al., 2003). Digital soil maps are generated using statistical algorithms and stored within a geographic information system (GIS), which allows data to be used readily for further analysis and interpretation (Minasny et al., 2013).

In this study, our goal was to map organic matter content (OM) and cation exchange capacity (CEC) on the Purdue University Agronomy Center for Research and Education (ACRE). ACRE is a hub of agronomic research for more than 50 researchers conducting about 180 research projects (ACRE, 2020) and is a state-of-the-art research facility for automated, high-throughput, field-based phenotyping (ICSIC, 2020). Soil organic matter content and CEC are important for plant nutrient availability and soil hydraulic properties, (Brady and Weil, 2002) and are properties

that influence the phenotypic response and productivity of plants (Havlin et al., 2014; Brady and Weil, 2002; Grigal and Vance, 2000).

Several DSM prediction methods have been used to map soil OM and CEC using point samples, remote sensing indices, and terrain attributes derived from a digital elevation model as inputs (Lacoste et al., 2014; Grim et al., 2008; Minasny et al., 2006; Simbahan et al., 2006; Thompson et al., 2006; Thompson and Kolka, 2005; Hengl et al., 2004; Florinsky et al., 2002; McKenzie and Ryan, 1999; Arrouays et al., 1995; McKenzie and Austin, 1993; Moore et al., 1993). McKenzie and Austin (1993) predicted CEC based on a generalized linear model using environmental variables as prediction covariates. Linear and multiple linear regressions models have been widely used for spatial prediction of soil organic carbon due to their simplicity in application and ease of interpretation (Thompson et al., 2006; Thompson and Kolka, 2005; Florinsky et al., 2002; Arrouays et al., 1995; Moore et al., 1993). Other studies used universal kriging (Simbahan et al., 2006), co-kriging (Hengl et al., 2004), generalized linear models (McKenzie and Ryan, 1999), and machine learning algorithms such as artificial neural networks (Minasny et al., 2006), random forest (Grim et al., 2008), and Cubist (Lacoste et al., 2014) for predicting OM.

Mapping soil variation in low relief areas can be a challenge because soil forming factors, especially topography and vegetation, may not co-vary with soil conditions over space to the level at which they can be used effectively in DSM (Zhang et al., 2017; Zhu et al., 2010). Terrain parameters derived from high-resolution elevation data, however, are capable of capturing local soil spatial variation that is caused by the interaction of water flow and topography (Luca et al., 2017; Roecker, 2013; Moore et al., 1993; Beven and Kirkby, 1979). In this study, we rely on lidar elevation data because it is available for the study region and because the relationship between soil

distribution and the gentle relief of the study area (section 3.1) is already known. The predictions of digital soil mapping are expected to improve as more detailed environmental variables, for example, normalized difference vegetation index and enhance vegetation index, are utilized (Maynard and Johnson, 2014; Peng et al., 2015). However, use of vegetation-based covariates are problematic at our study site where there are many small field plots with heterogeneous experiments.

In this study, we used universal kriging (UK), Cubist, and random forest (RF) to predict the spatial trend in OM and CEC for a 570 ha research farm. We hypothesized that on a field scale, terrain-driven hydrological flow patterns are the dominant process responsible for soil OM and CEC differences in surface soils.

3.2 Materials and Methods

3.2.1 The Study Area

The Purdue Agronomy Center for Research and Education (ACRE) is a 570 ha agronomic field research station located in Tippecanoe County Indiana, USA (40° 29' N, 86° 59' W) (Fig. 3.1). ACRE is located on a low relief, gently undulating Wisconsin age till plain. The soils formed in ~50 cm of loess over loamy Wisconsin till and outwash. Most of the soils are poorly and somewhat poorly drained. ACRE is located at the transition between the Eastern Hardwood Forests to the east and the prairies of the Great Plains to the west. Mollisols occur over most of the study area, but Alfisols occur on the southern edge (USDA-NRCS, 1998). Corn and soybean are the major crops. Based on 30-year normals for 1981 to 2010, the mean annual temperature is 10° C and mean total annual precipitation is 970 mm (MRCC, 2013). The average summer temperature (June to August) is 22.2° C and average winter temperature (December to February) is -2.6° C (NWS-COOP, 2020). Climatically, the site is in the mesic soil temperature regime and the udic

soil moisture regime, but large areas of the study site have soils with an aquic soil moisture regime because of the presence of a seasonal high water table (USDA-NRCS, 1998).

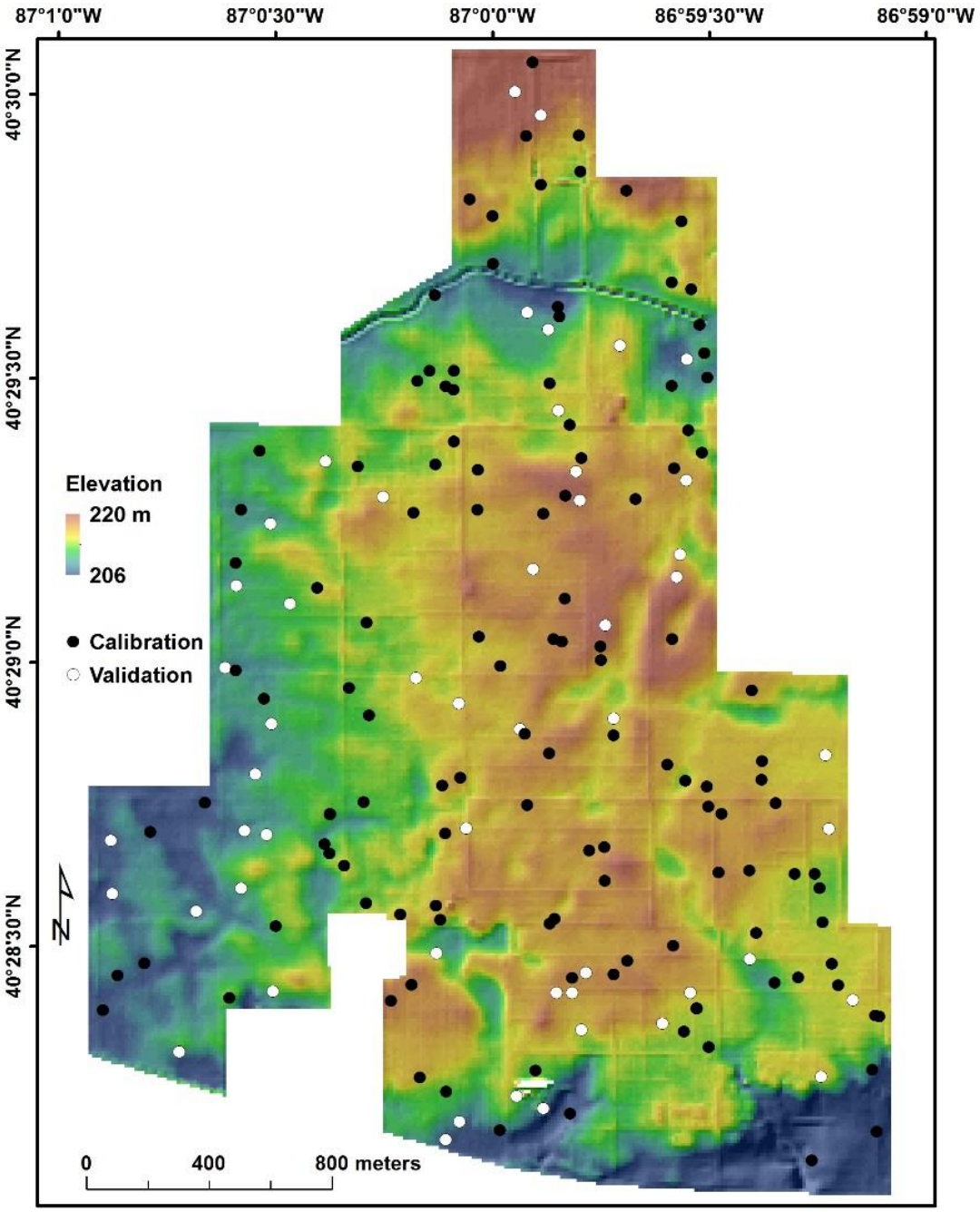

Figure 3.1: Study area and sampling locations over a lidar-derived hillshade base map. Seventy percent of the samples were used for calibration and 30 percent were used for validation.

3.2.2 Soil Sampling and Analysis

One hundred and seventy four (174) soil samples had been collected at ACRE as part of a previous, unpublished study. Sampling locations were selected using the conditioned Latin hypercube sampling (cLHS) algorithm (Minasny and McBratney, 2006) using the *clhs* package (Roudier, 2011) in R-software 3.5.1 (R Core Team, 2018) to generate the sampling locations. The cLHS method is a stratified random procedure that selects sampling locations based on the probability distribution of environmental covariates. Environmental covariates for cLHS were generated from terrain derivatives derived from a lidar-based digital elevation model. Unfortunately, records of the exact combination of environmental covariates used for cLHS sampling were lost and exact information about the specific covariates used is not available.

The samples were collected from 0 – 10 cm, oven-dried at 40° C, crushed, and passed through a 2 mm sieve. The samples were analyzed by A&L Great Lakes Laboratories, Inc, Fort Wayne, Indiana, following the soil test procedures for the North Central Region (NCR, 1998). Briefly, organic matter content (OM) was determined by loss on ignition at 360° C with a base factor of 0.97 and OM expressed on a weight percent basis (%), while CEC ($cmol_c\ kg^{-1}$) was measured by sum of cations displaced by 1 M ammonium acetate solution at pH 7.

For model building and spatial predictions, the data were randomly split, with 70% of the samples used for model calibration and 30% used for model evaluation (Fig. 3.1). Descriptive statistics are given in Table 3.1.

Table 3.1: Summary statistics of soil organic matter content (OM) and cation exchange capacity (CEC) data for the study area.

Statistical Index	OM			CEC		
	Whole	Calibration	Evaluation	Whole	Calibration	Evaluation
		-------- % --------			------- $cmol_c\ kg^{-1}$ -------	
Minimum	1.2	1.9	1.2	9.9	11.1	9.9
1st Quartile	3.5	3.5	3.2	16.2	16.5	14.9
Median	4.0	4.0	4.2	20.1	20.1	20.1
Mean	4.2	4.2	4.0	19.9	20.1	19.5
3rd Quartile	4.8	4.9	4.6	23.2	23.3	23.0
Maximum	7.2	7.0	7.2	30.1	30.1	29.3
Standard Deviation	1.1	1.2	1.1	4.6	4.5	4.9
Soil Samples (N)	174	123	51	174	123	51

3.2.3 Digital Elevation Model and Terrain Attributes

Digital Elevation Model

Digital elevation data for Tippecanoe County, IN acquired in 2013 at 1.5 x 1.5 m pixel resolution using lidar was downloaded from the Indiana Spatial Data Portal (http://gis.iu.edu/). The DEM was re-projected from the Indiana State Plane West Coordinate System (NAD_1983_StatePlane_Indiana_West_FIPS_1302_Feet) which uses dimensions (XY and Z) in feet, to the Indiana Geospatial Coordinate System (InGCS) for the Tippecanoe and White Counties (NAD_1983_2011_InGCS_Tippecanoe-White_(m)). The XY and Z dimensions in InGCS are in meters. The InGCS has lower grid vs. ground distortion (±2.6 ppm) when compared to the State Plane Coordinate System (±80 ppm) (INDOT, 2016) and thus is more appropriate for a small area such as ACRE.

Digital elevation models with pixel resolutions on the order of 1 – 2 m are often too detailed for modeling soil spatial variability (Smith et al., 2006; Shi et al., 2012; Maynard and Johnson, 2014; Lacoste et al., 2014). Winzeler et al. (2008) found that pixel resolutions from 5 – 10 m are sufficient to capture the topography for digital soil mapping of Northern Indiana's

glaciated landscapes. Our own initial evaluation showed that within ACRE, anthropogenic micro-topographic features such roads and field boundaries, that are on average 20 cm higher than the cultivated fields, unduly affected the DEM derived indices. We resampled the original 1.5 m DEM to 10 m using simple mean aggregation in ArcGIS 10.6 (https://esri.com) in order to smooth out most of the anthropogenic features.

A rectangular buffer that included the entire watershed contributing water to ACRE was defined using watershed boundaries and stream channels obtained from the United States Geological Survey – National Hydrography Dataset (USGS-NHD) downloaded from the United States Department of Agriculture (USDA) Geospatial Data Gateway (GDG) (https://datagateway.nrcs.usda.gov/) for Tippecanoe County, IN. The buffer was then used to clip the resampled DEM prior to further processing.

Terrain Attributes

It is possible to generate many terrain attributes from a DEM, but it is important to limit their number to avoid redundancy and model overfitting. We focused on those terrain attributes that have a close relationship to water redistribution across a landscape and are commonly used in DSM. We calculated the following terrain attributes using SAGA-GIS 2.1.4 (Conrad et al., 2015): topographic wetness index (TWI), topographic position index (TPI), multi-resolution valley bottom flatness index (MrVBF), multi-resolution ridge top flatness index (MrRTF), profile curvature, and plan curvature. Due to the low relief of the study area, both plan and profile curvatures displayed high levels of small-scale noise and did not capture field-level topographic variations. Thus, they were not included in subsequent calculations. Details of the four terrain attributes that were used in subsequent calculations, TWI, TPI, MrVBF, MrRTF (Fig. 3.2), follow.

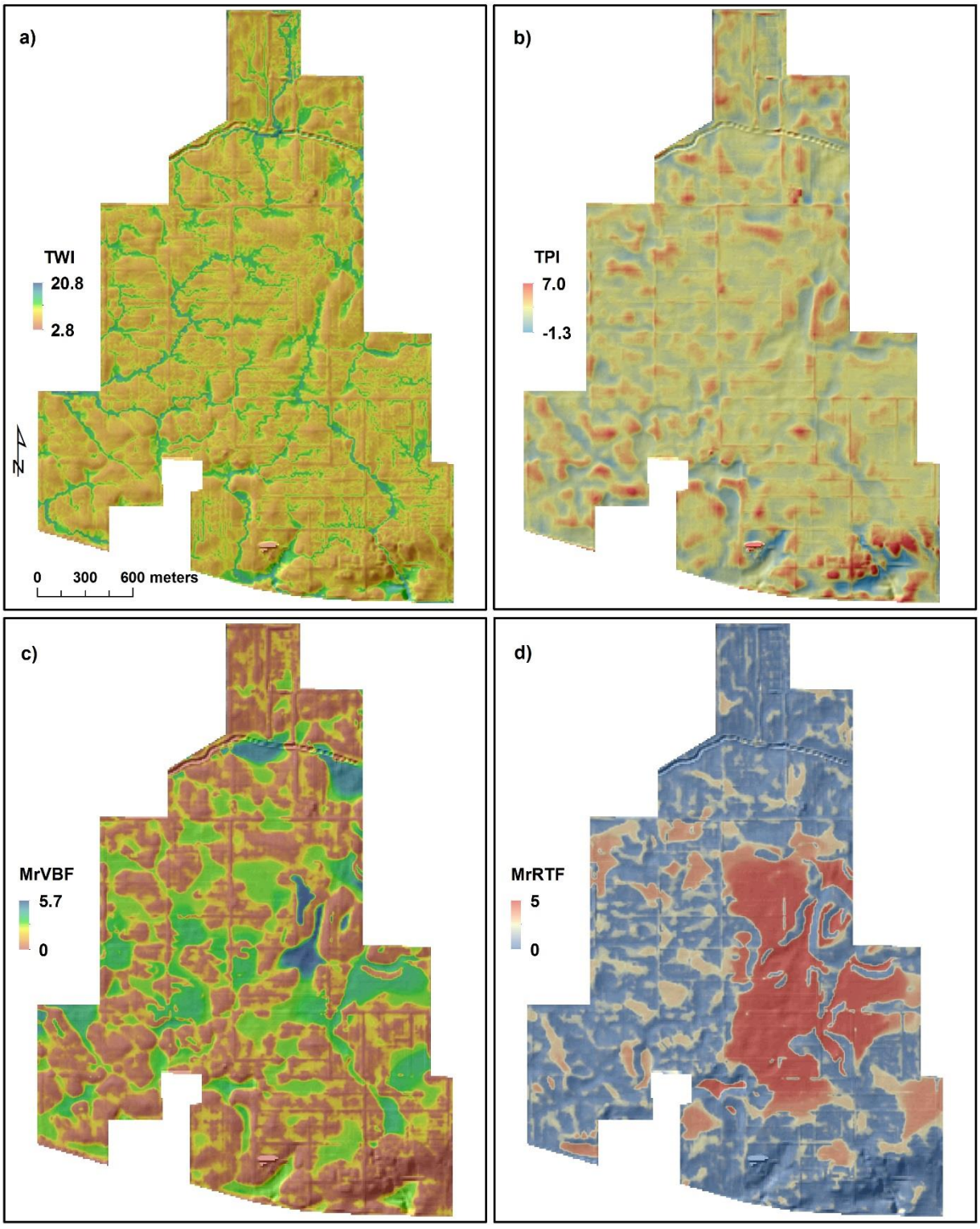

Figure 3.2: Terrain attributes calculated from the digital elevation model. (a) Topographic wetness index (TWI), (b) topographic position index (TPI), (c) multi resolution valley bottom flatness index (MrVBF), and (d) multi resolution ridge top flatness index (MrRTF).

Topographic Wetness Index

The topographic wetness index (TWI) is used to quantitatively present the relationship between topography and hydrological process, mainly surface runoff in a watershed. The SAGA multi-flow-direction algorithm of TWI was used in this study. Higher values of TWI represent areas that accumulate water, such as depressions and drainage ways, while lower values represent areas that shed water, such as crests and ridges.

Topographic Position Index

The topographic position index (TPI) (Weiss, 2001) compares the elevation of a cell (Z_0) to the average elevation of its surrounding cells (Z_α) in a specific area as defined by circles of arbitrary radius ($TPI = Z_0 - Z_\alpha$). Positive values of TPI represent ridges, and negative values represent valleys, while flat areas contain values close to zero. This index is scale dependent, and by using different radii it can delineate small hummocks and larger ridges, as well as small depressions and larger valleys. We evaluated different radii for our study area. The larger radii (150, 200, 300, and 500 m) resulted in smoothing the landscape features, while smaller radii (30 and 60 m) generated linear artifacts and divided the actual landforms into small pieces. Based on visual interpretation and familiarity with the study location, a radius of a 100 m was found to best represent the landscape units of the study site.

Multiresolution Valley Bottom Flatness and Multiresolution Ridge Top Flatness

The MrVBF algorithm (Gallant and Dowling, 2003), identifies valley bottoms by utilizing the lowness and flatness characteristics of them. The lowness parameter is measured by ranking elevation with respect to a circular neighborhood area, and the flatness parameter is measured using the inverse of slope.

The slope threshold is a critical parameter in MrVBF calculations, and it depends on the DEM resolution. The suggested slope threshold for a DEM of 250 m resolution is 4%, while for a DEM of 25 m it is 16%, and for a DEM of 8 m it is 32%. Slope thresholds of 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 16% (the default slope threshold of the algorithm) were compared. Based on our familiarity with the area and after generating MrVBF with different slope thresholds and checking the resulting terrain attributes in the field, a slope threshold of 2% was found to best represent the topography of the landscape. The terrain attribute of MrRTF is a separate index but complementary to the MrVBF. It is derived in a similar way to MrVBF, except it identifies the upper parts of the landscape. Similar to MrVBF, the same slope threshold value (2%) was selected for MrRTF.

3.2.4 Data from the Soil Survey Geographic Database (SSURGO)

For Tippecanoe County, the Soil Survey Geographic (SSURGO) database provides soil mapping information at a scale 1:15,840 (USDA-NRCS, 1998). In this study, three values of OM (low, representative, and high) and three values of CEC (low, mean, and high) from SSURGO (Table 3.2) were compared to the predictions from the DSM models. All of these SSURGO OM and CEC values were directly acquired from the Web Soil Survey website (Soil Survey Staff, 2020), except the mean value of CEC, which was calculated as the average of the low and high CEC values. The SSURGO values of soil properties have been derived from a combination of laboratory measured data and soil scientist expert knowledge (Libohova et al., 2016). Soil OM was determined using the Walkley-Black method and CEC was determined by summation of cations, which were displaced by ammonium acetate solution (Franzmeier et al., 1977; Soil Survey Staff, 2014). There is no universal conversion between Walkley-Black and loss on ignition method. Therefore, the loss on ignition OM values used for DSM prediction models are compared with the SSURGO OM predictions measured with Walkley-Black method.

Table 3.2: The soil survey geographic (SSURGO) organic matter content (OM) and cation exchange capacity (CEC) low, representative (Rep.), mean, and high values for 0 – 10 cm based on the spline function. Data is for the ACRE study site.

SSURGO Soil Map Unit		Area	OM			CEC		
			Low	Rep.	High	Low	Mean	High
		%	----- % -----			----- $cmol_c\ kg^{-1}$ -----		
Cm	Chalmers silty clay loam	33.90	3.35	4.91	6.46	18.20	28.50	38.80
CwB2	Crosby-Miami silt loams	0.40	0.99	2.43	3.11	2.47	7.00	12.16
Du	Drummer soils	17.60	3.26	4.86	6.47	22.40	29.23	36.16
Md	Mahalasville-Treaty complex	0.20	3.28	4.83	6.39	18.18	24.22	30.30
MsC2	Miami silt loam	0.20	1.05	2.14	3.24	4.72	9.24	16.00
Mu	Milford silty clay loam	4.20	3.47	5.72	6.71	25.40	31.10	36.70
Pg	Pella silty clay loam	1.60	4.81	4.65	7.05	22.90	28.90	34.90
Pk	Peotone silty clay loam	0.60	4.31	5.87	7.43	21.20	29.60	38.10
RcA	Raub-Brenton complex	22.10	1.97	2.88	4.24	12.55	17.38	22.26
RoB	Rockfield silt loam	4.60	1.07	1.60	2.14	6.73	12.20	17.60
SwA	Starks-Fincastle complex	3.60	0.95	2.11	2.89	7.10	12.81	18.59
TfB	Throckmorton silt loam	1.40	1.05	2.11	3.19	5.61	11.60	17.60
TmA	Toronto-Millbrook complex	8.70	1.95	2.85	3.77	10.96	16.71	22.46
Ua	Udorthents, loamy	0.90	-	-	-	-	-	-

The SSURGO data are based on a traditional method of soil sampling, meaning that each soil profile is divided into soil horizons based on morphological properties of the soil. A bulk sample is taken from each horizon and it is assumed to represent the average value of a soil attribute for the depth interval of that horizon. The analysis of samples collected in our study were based on 0 – 10 cm depths. We used the mass-preserving splines function (ea_spline) of the *ithir* package (Malone, 2018) in R-software 3.5.1, which predicts continuous soil properties both within the observed depths and among the depths where no observations were made (Malone et al., 2017). For detailed information and mathematical expression of the spline function, see Ponce-Hernandez et al. (1986), Bishop et al. (1999), and Malone et al. (2009).

Since a map unit may have two or more components, we derived the final values of a map unit based on the weighted mean of each component. For instance, CwB2 (Crosby-Miami silt loams, 2 to 4 percent slopes, eroded) contains 64% Crosby and 33% Miami and 3% other minor components. Based on the spline function for the 0 – 10 cm depth, the OM representative values

of Crosby was 2.67% and Miami was 2.18%. The final OM representative value of CwB2 for the 0 – 10 cm depth was derived as:

$$\text{CwB2 mapping unit mean value of OM} = (0.64 * 2.67) + (0.33 * 2.18) = 2.43\% \quad [1]$$

3.2.5 Spatial Prediction Models

Three different models (universal kriging, Cubist, and random forest) were used to predict soil organic matter content and cation exchange capacity. All model training and evaluation was performed in R-software 3.5.1 (R Core Team, 2018).

Universal kriging, also known as regression kriging (Odeh et al., 1995; Hengl et al., 2007), is a hybrid approach to modeling, meaning that the prediction of a desired variable is made based on deterministic and stochastic components. The deterministic part of the regression relies on the covariate information, while the stochastic part relies on the spatial auto-correlation of the residual based on a variogram (Malone et al., 2017). We ran backwards stepwise linear models to select appropriate covariates for the deterministic part of UK. The *gstat* package (Gräler et al., 2016) in R 3.5.1 environment was used for UK prediction of OM and CEC.

Cubist is a data mining tool that uses a rule-based regression algorithm for prediction (Quinlan, 1992). It operates based on *if, then, else* statements. If a condition is matched, the next step is a prediction of the desired soil property by using ordinary least squares regression from the covariates within that subset (Minasny and McBratney, 2008; Peng et al., 2015; Malone et al., 2017). However, if a condition is not met, then the next node of the tree is defined by the rule and the *if, then, else* sequence is repeated. The interpretation of a Cubist model is easy as it provides an explicit model stating the relative importance of the predictors. The *Cubist* function (Kuhan and Quinlan, 2018) in R 3.5.1 environment was used to predict OM and CEC of the study area.

The Random Forest (RF) model developed by Breiman (2001), is a type of ensemble machine learning algorithm. The RF model predicts the property of interest based on covariates by creating multiple decision trees. The outcomes of the decision trees are then aggregated to provide the final prediction. A random and independent bootstrap sample of the training data is used to train each tree in the forest. From the bootstrap sample, a random subset is selected for training and the remaining points, known as "out of bag," are used for validating the tree. Additionally, a random subset of the variables is selected to split the nodes of each tree (Forkuor et al., 2017). In summary, the RF decision trees are developed based on a random selection of data (bootstrap sample) and a random selection of variables. Further details about RF and the underlying theory can be found in Breiman (2001) and Grimm et al. (2008). The *randomForest* package (Liaw and Wiener, 2002) was used in the R 3.5.1 environment to predict both OM and CEC.

The semi-variogram, which is commonly referred to simply as the "variogram," is used to define the spatial autocorrelation or spatial dependency of the observed sample points. Linear, spherical, exponential, and Gaussian are four common variogram models (Malone et al., 2017). Kriging of residuals may capture spatial variability that was not estimated by deterministic or linear models of UK. For UK, a spherical variogram, and for Cubist and RF, exponential variograms were fitted to krige the residual of OM. For kriging the CEC residuals, a spherical variogram was fitted for all three predictive models. For residual kriging, we used the *gstat* package (Gräler et al., 2016) in R 3.5.1 environment. The final estimates of OM and CEC were derived based on a combination of the kriged residuals and the predicted values from the models.

3.2.6 Evaluation of Model Performance

All the predictive models were first evaluated with the calibration dataset from which they were generated (internal evaluation). A second random-hold-back independent evaluation was conducted using 30% of the data for testing the prediction performance of each digital soil mapping methodology. The prediction quality of the models was evaluated with root mean square error (RMSE), mean error (ME) or bias, coefficient of determination (R^2), and Lin's concordance correlation coefficient (LCCC), respectively equated as:

$$RMSE = \sqrt{\left(\frac{\sum_{i=1}^{n}(obs_i - pred_i)^2}{n}\right)} \quad [2]$$

$$ME\ (i.e.\ bias) = \frac{\sum_{i=1}^{n} obs_i - pred_i}{n} \quad [3]$$

$$r = \frac{\sum_{i=1}^{n}(obs_i - \overline{obs})\ (pred_i - \overline{pred})}{\sqrt{\sum_{i=1}^{n}(obs_i - \overline{obs})^2}\sqrt{\sum_{i=1}^{n}(pred_i - \overline{pred})^2}} \quad [4]$$

$$LCCC = \frac{2\rho\sigma_{obs}\rho\sigma_{pred}}{\sigma_{pred}^2 + \sigma_{obs}^2 + (\mu_{pred} + \mu_{obs})^2} \quad [5]$$

where obs_i are the observed values and $pred_i$ are the predicted values of the soil properties at location i, μ_{obs} is the mean of the observed values, μ_{pred} is the mean of the predicted values, σ^2_{obs} is the variance of the observed values, σ^2_{pred} is the variance of predicted values, n is the number of the sampling locations, and ρ is the correlation coefficient among the observations and predictions (Malone et al., 2017).

The RMSE shows the accuracy of the prediction. Smaller values, which show higher accuracy, are preferred. Bias shows the mean error of the prediction and an unbiased prediction has a bias of zero. The R^2, which is the square of Pearson's correlation coefficient (r), measures the precision of the relationship between the predicted and observed values. The LCCC (Lawrence and Lin, 1989), is a single statistic that measures both the precision and the accuracy of the relationship. LCCC is also known as the goodness of fit along a 45° line (1:1 line). The value of

LCCC falls between -1 and +1. A value of -1 indicates perfect negative agreement, while a value of +1 indicates perfect positive agreement between the predicted and observed values. An LCCC value of zero shows that there is no agreement at all (Malone et al., 2017; Santra and Panwar, 2017). The strength of the agreement was evaluated based on the proposed scale from McBride (2005). Lin's concordance correlation coefficient is considered poor (<0.90), moderate (0.90–0.95), substantial (0.95–0.99) and almost perfect (>0.99). The *goof* function of the *ithir* package (Malone, 2018) was used in the R 3.5.1 to compute these evaluation indices.

3.3 Results and Discussion

3.3.1 Spatial Trend Modeling

Each model utilized different environmental covariates for OM and CEC predictions. A backwards stepwise linear model selection was used for the UK model. Based on the following equations (Eq. 6 and 7), the backwards stepwise model selected TWI, TPI, and MrRTF for OM, and TPI and MrVBF for CEC predictions.

$$OM = 3.37 + 0.11 * TWI - 2.20 * TPI - 0.09 * MrRTF \quad [6]$$

$$CEC = 18.41 - 9.36 * TPI + 1.02 * MrVBF \quad [7]$$

From a pedological standpoint, Eqs. 6 and 7 reveal meaningful relationships between terrain and OM or CEC. Eq. 6 shows that OM is positively correlated to TWI or wet/low-lying areas of the landscape, while it is negatively correlated to TPI and MrRTF or higher/steeper areas of the landscape. Similarly, CEC (Eq. 7) is negatively correlated with TPI, but positively correlated with MrVBF or lower landscape positions.

Cubist utilized all four covariates for OM, and only TPI and MrVBF for CEC predictions. Out of ten models generated by Cubist for OM and CEC predictions (Appendix A and B), we selected the models with the lowest prediction error. For example, the OM model (Eq. 8) was only

applicable in 113 locations where the average OM was 4.09%. The prediction error of this model was 0.70%. The CEC model (Eq. 9) was applied to all 123 training locations which has a mean value of 20.12 $cmol_c\ kg^{-1}$. The predication error of this model is 2.84 ($cmol_c\ kg^{-1}$). The Cubist model provided slightly different models for OM prediction based on the combination of four terrain attributes but produced identical models for CEC. Following are examples of Cubist models for OM and CEC predictions.

$$If\ \ TWI <= 13.76\ then\ OM = 3.79 + 0.21 * MrVBF - 1.17 * TPI - 0.11 * MrRTF \quad [8]$$

$$CEC = 18.84 - 9.30 * TPI + 1.03 * MrVBF \quad [9]$$

The Cubist model also provided the relative usage and relative importance of the covariates, which shows the usage of covariates in multivariate linear models and importance of covariate(s) in developing conditions rules (*if then else* rules). In OM prediction, the relative usage of the four terrain attributes was 54 (TPI), 37 (MrRTF), 35 (TWI), and 34 (MrVBF). The Cubist model only showed a relative importance of 54% for TWI, meaning that TWI was the only predictor that appeared in rule conditions. Therefore, TWI is the best predictor for the Cubist model for OM prediction. For CEC, the Cubist model did not provide relative importance for any of the covariates, but it provided a relative usage of 100% for both TPI and MrVBF.

Random forest used all four covariates for predicting both OM and CEC. The *varImpPlot* function in the *randomForest* package (Liaw and Wiener, 2002) shows the importance of covariates in OM and CEC predictions. For the RF predictions the most important covariates were TPI and TWI for OM predictions, and TPI and MrVBF for CEC predictions (Fig. 3.3). Overall, TPI was the most important variable and MrRTF was the least important variable in all selected models.

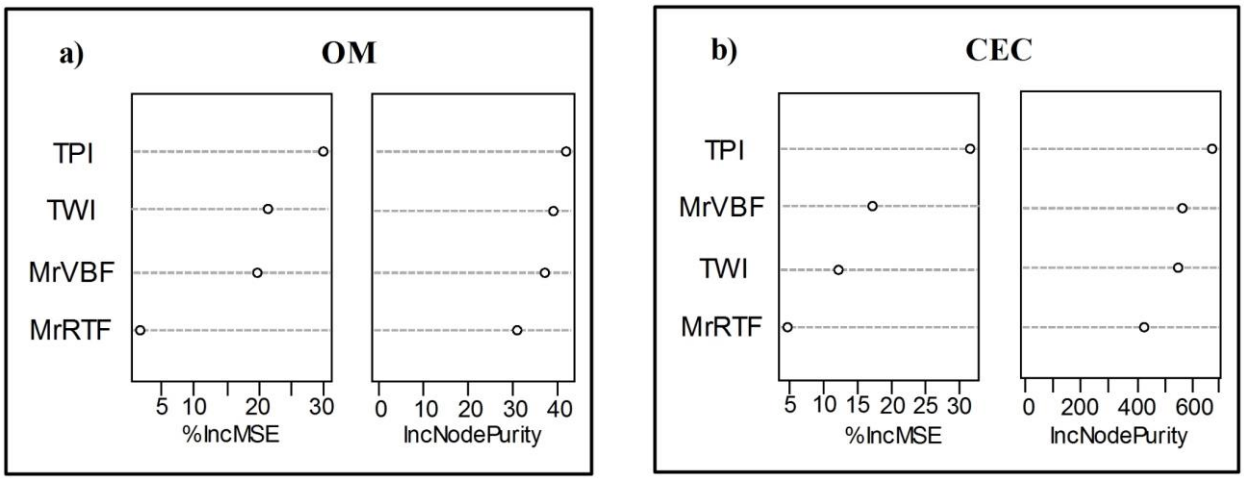

Figure 3.3: Random forest generated importance plots of covariates, a) for organic matter content (OM) and b) for cation exchange capacity (CEC) prediction. The %IncMSE shows the mean decrease in accuracy. The IncNodePurity shows the decrease in node purity at the end of the tree. The higher the %IncMSE and IncNodePurity show that a particular variable is highly important and if removed the prediction accuracy and node purity will be affected.

The CEC prediction equations of the UK (Eq. 7) and Cubist (Eq. 9) models were almost identical. Both the UK and Cubist models show that OM and CEC increase as TWI and MrVBF increase and decrease as TPI and MrRTF values increase. In other words, CEC and OM have positive relation with TWI and MrVBF and negative with TPI and MrRTF. Even though, RF utilized all four covariates for OM and CEC predictions, similar to UK and Cubist, TPI and MrVBF were the most important variables for CEC predictions in RF. Therefore, as expected, all the predictive models produced similar results for CEC prediction (Table 3.3).

Table 3.3: Universal kriging (UK), Cubist, and random forest (RF) accuracy assessment for organic matter content (OM) and cation exchange capacity (CEC) predictions with calibration and evaluation datasets.

Prediction Model		OM				CEC			
		R^2	Bias	RMSE	Concordance	R^2	Bias	RMSE	Concordance
			%				$cmol_c\ kg^{-1}$		
UK	Calibration	0.50	0.00	0.80	0.60	0.60	0.00	2.80	0.70
	Evaluation	0.44	0.22	0.83	0.56	0.39	0.05	3.74	0.55
Cubist	Calibration	0.50	0.00	0.80	0.70	0.60	0.00	2.80	0.70
	Evaluation	0.45	0.17	0.80	0.58	0.41	0.00	3.68	0.57
RF	Calibration	0.90	0.00	0.40	0.90	0.90	0.00	1.40	0.90
	Evaluation	0.45	0.17	0.80	0.56	0.44	0.17	3.62	0.56

3.3.2 Predictive Model Performance

Based on the four evaluation metrics (Table 3.3) and scatter plots of measured versus predicted OM (Fig. 3.4) and CEC (Fig. 3.5), we found no major differences in the prediction performance of all three models.

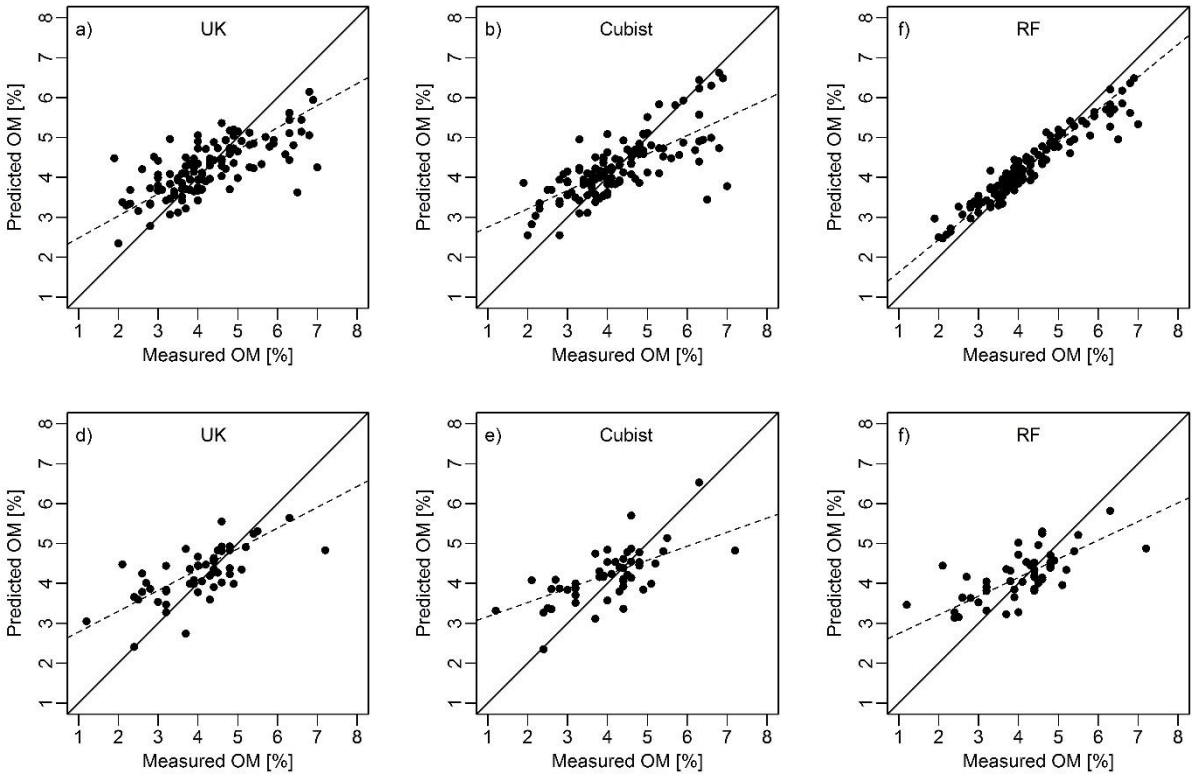

Figure 3.4: Scatter plots of measured vs predicted organic matter content (OM) based on calibration and evaluation data. (a) Universal kriging (UK), (b) Cubist, and (c) random forest (RF) scatter plots are based on the calibration data and (d) universal kriging (UK), (e) Cubist, and (f) random forest (RF) scatter plots are based on the evaluation data. The solid line indicates a line of concordance or a 1:1 relationship. The dashed line indicates the line of best fit.

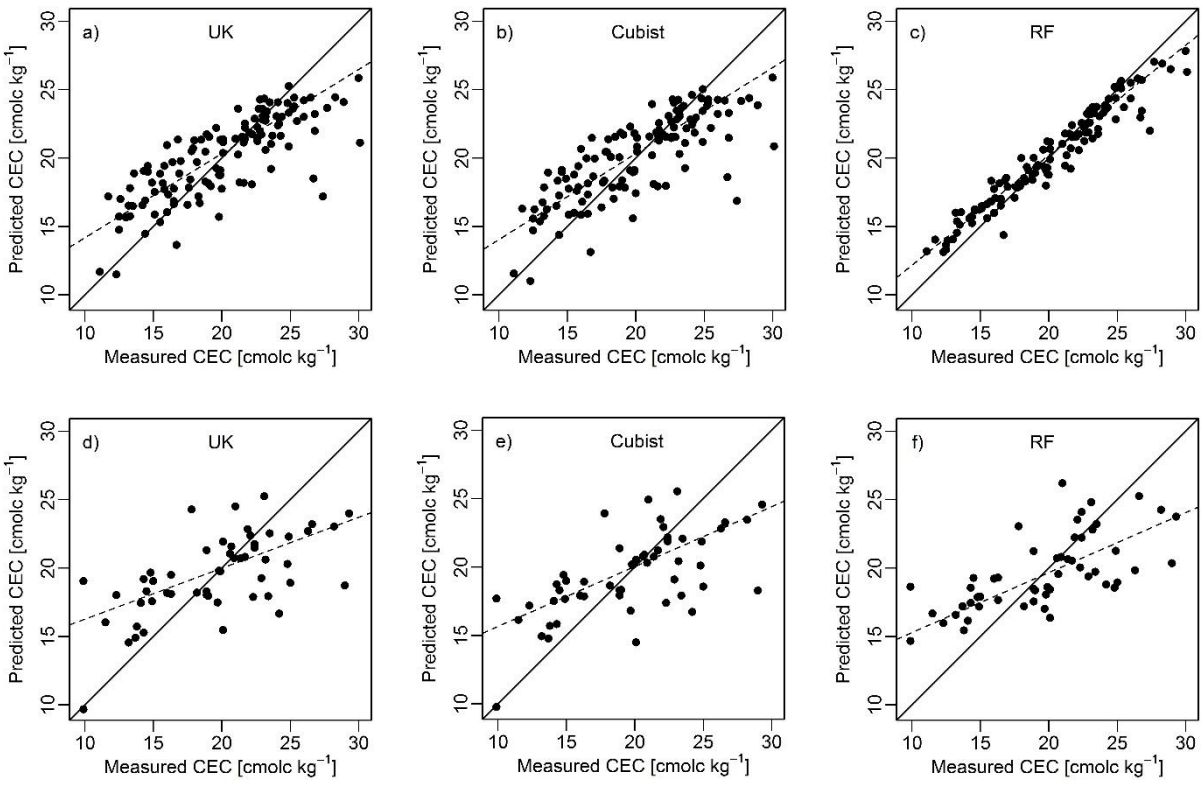

Figure 3.5: Scatter plots of measured vs predicted CEC based on the calibration and evaluation data. (a) Universal kriging (UK), (b) Cubist, and (c) random forest (RF) scatter plots are based on the calibration data and (d) universal kriging (UK), (e) Cubist, and (f) random forest (RF) scatter plots are based on the evaluation data. The solid line indicates a line of concordance or a 1:1 relationship. The dashed line indicates the line of best fit.

When the performance of the models was tested with the calibration data, RF had lower RMSE, lower bias, higher R^2, and higher concordance than UK and Cubist for both soil OM and CEC predictions. With RF, this was expected due to the ensemble approach, which can result in low bias and variance (Nabiollahi et al., 2019). Additionally, RF used all four predictors, which may result in overfitting (Nussbaum et al., 2018; Statnikov et al., 2008). For the RF models, however, there was a significant change in model performance between calibration and evaluation data. For example, for OM predictions, the R^2 of the RF prediction were 0.90 for the calibration dataset and 0.45 for the evaluation dataset. This significant change between calibration and

evaluation performance is strong evidence that the RF models may be over fitted. This highlights one of the dangers of RF for DSM: if RF models are not evaluated rigorously (i.e. using independent evaluation rather than leave-one-out), model performance estimates may be overly optimistic.

According to the scatter plots for OM (Fig. 3.4 a, b, c) and CEC (Fig. 3.5 a, b, c), all three models tended to over predict at low values and under predict at high values. This behavior is less pronounced for RF models on the calibration data, but it is apparent for all models on the evaluation dataset. This lack of performance may be due to several factors, which are discussed below.

One reason for this poor correlation may be the due to the study location itself. ACRE serves as a research and education facility and consists of many smaller individual fields that are managed under highly variable practices (e.g. multiple tillage systems, nutrient application rates, and crop rotations). This high variation in management may lead to higher soil variability from field to field than expected. Training models using samples from highly variable fields can limit their predictive performance outside the sample areas (Thomasson et al., 2001; Rossel et al., 2006). To account for the effects of variable management history, we would need to incorporate environmental covariates that describe previous management of each field into our modeling framework. Further research is needed to identify suitable covariates that describe agriculture management history.

Another possible cause for poor correlations was the relatively small changes in environmental covariates across the study area. The study site is characterized by relative flat topography with subtle topographic variation (on average 1% slope based on a 3 x 3 pixel window). In many environments, chemical properties of surface soil and OM are highly variable spatially, and distinct variations are often found within short distances of meters and/or decimeters

(Trangmar et al., 1985; Schöning et al., 2006; Wiesmeier et al., 2009). A further complication is that ACRE is crisscrossed by a grid of roads and grassed field boundaries that are on average 20 cm higher than the adjoining fields, and by a dense network of underground drainage tiles. The impact of the roads on terrain attributes is evident as linear features in the covariate maps (Fig. 3.2). All of these factors may have diluted the influence of terrain in the distribution of OM and CEC and lead to poorer than expected model performance.

Generally, the values of OM (Fig. 3.4 d, e, f) were closer to the 1:1 line than the CEC values (Fig. 3.5 d, e, f). This difference in predictions of OM compared to CEC might be due to the existence of higher variation in CEC data, as well as the number and type of predictors selected by these models. For instance, UK used three predictors for OM and two for CEC predictions. Similarly, Cubist used all four predictors for OM and only TPI and MrVBF for CEC predictions.

In general, the correlations between terrain attributes and soil properties were fair for all models in our study. Nonetheless, the R^2 values in this study were comparable with other studies that considered terrain/climatic data only (Mason and Sulaeman, 2016; Forkuor et al., 2017; Pei et al., 2010).

3.3.3 Organic Matter Content and Cation Exchange Capacity Distribution in the Landscape

DSM model predictions of OM (Fig. 3.6) and CEC (Fig. 3.7) were consistent with the theoretical, pedological distributions of OM and CEC within the landscape. The maps of predicted OM and CEC for all three models indicate higher values for OM and CEC in lower landscape positions (i.e. foot and toeslopes), and lower values at higher and steeper landscape positions (i.e. shoulders and summits). Lower areas receive more overland flow of nutrients and crop residue from the steeper areas leading to an increase in OM and CEC. The steeper regions, due to less

vegetative cover, are subjected to erosion and lose of nutrients to the lower parts of the landscape. On the other hand, waterlogging in lower areas reduces the rate of OM decomposition and results in higher OM and nutrient accumulation (Brady and Weil, 2002; Starr et al., 2000).

At ACRE, it is not surprising to find vice-versa results, meaning that lower landscape positions might have lower OM and CEC values when compared to upper landscape positions. This can happen for various reasons. First, the OM content in the depressions is diluted by erosion that carries lower OM soil from upslope. In this landscape, it is not unusual to find buried surface horizons. Second, prior to the European settlement and the large-scale drainage that took place since then, areas that were once ponded for long periods are now much better drained.

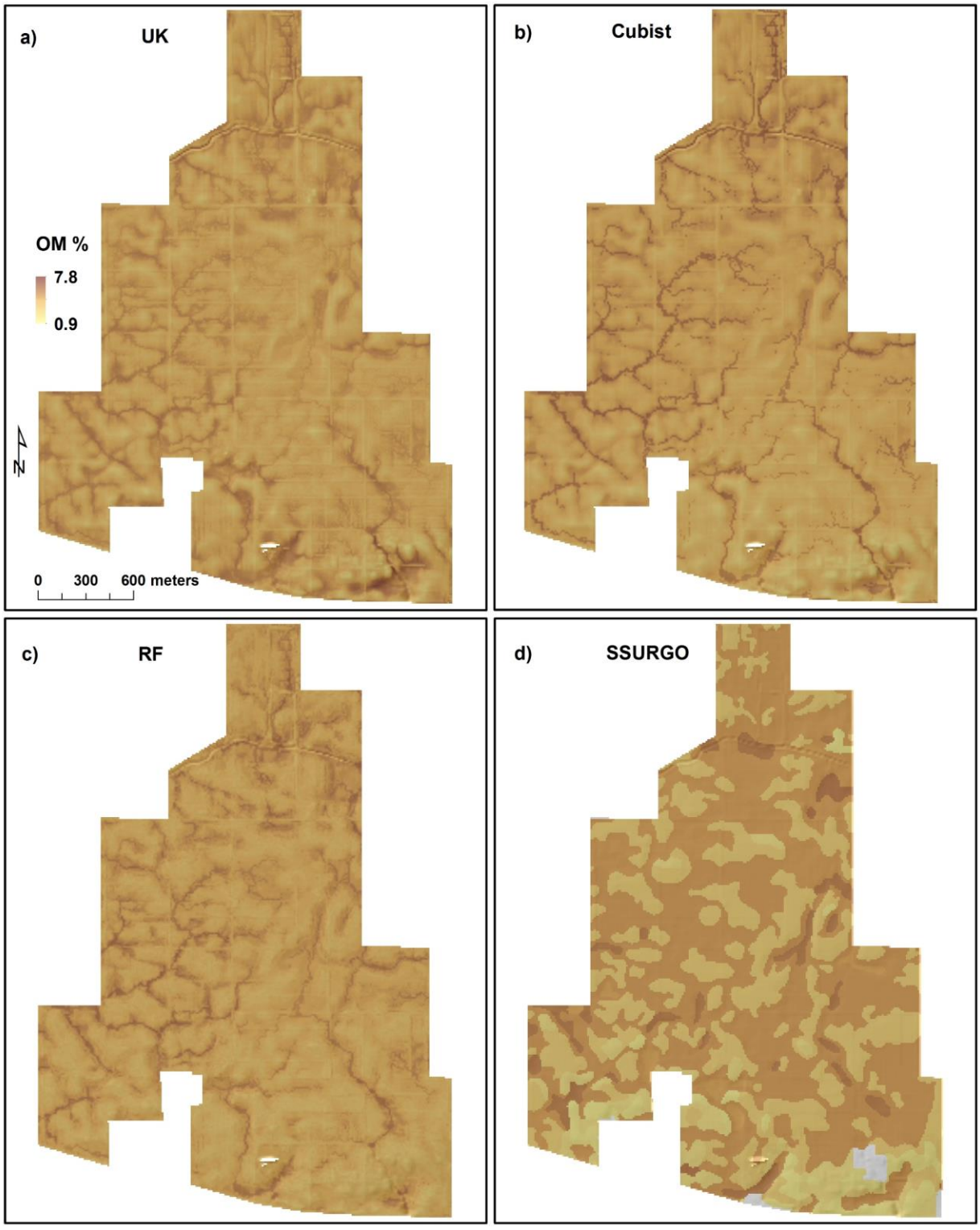

Figure 3.6: Organic matter content (OM) prediction. a) Universal kriging (UK), b) Cubist, c) random forest (RF), and d) soil survey geographic (SSURGO).

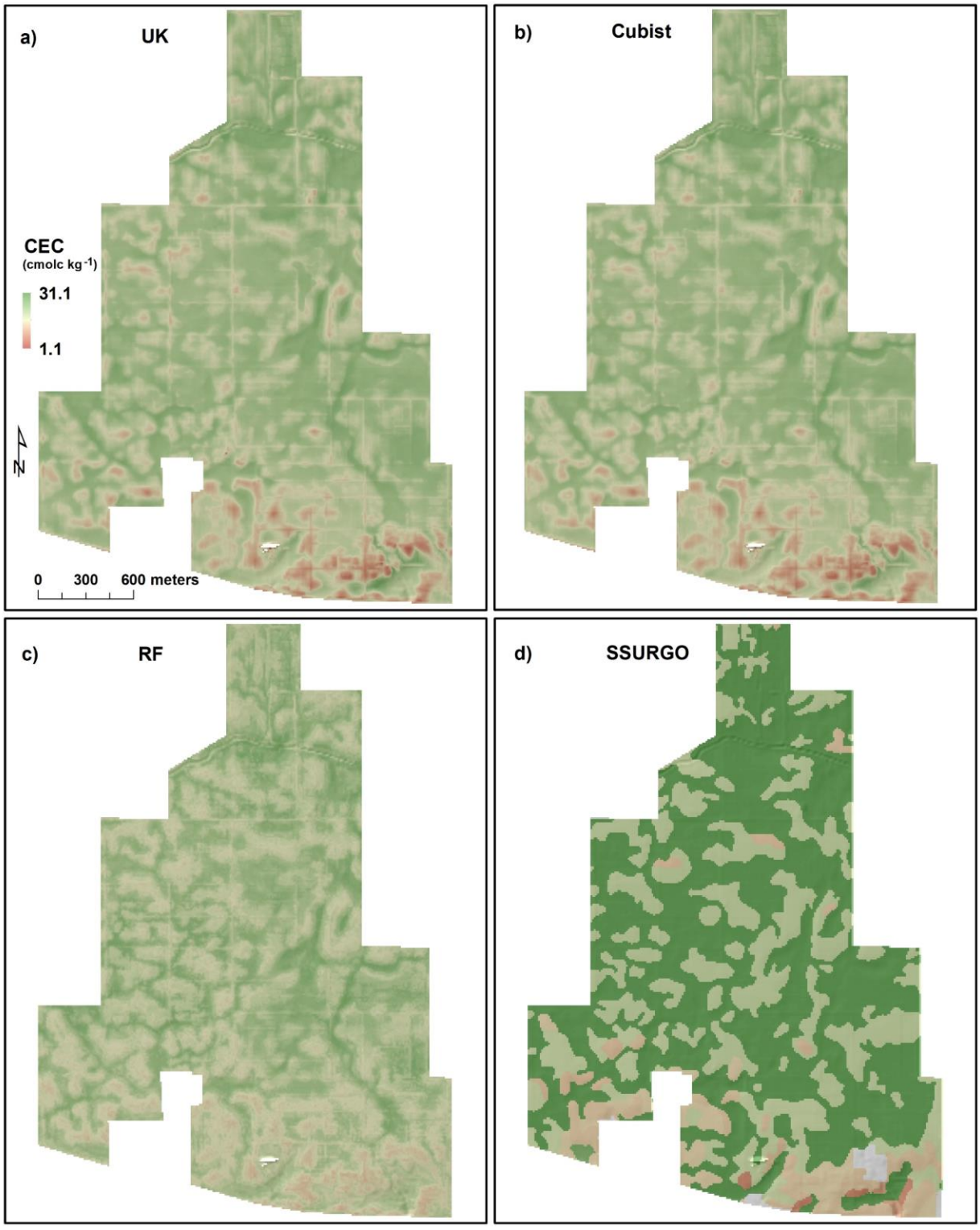

Figure 3.7: Cation exchange capacity (CEC) prediction. a) Universal kriging (UK), b) Cubist, c) random forest (RF), and d) soil survey geographic (SSURGO).

3.3.4 Predictive Models versus SSURGO

When comparing maps of SSURGO OM and CEC to DSM maps, all maps generally show a similar trend: high CEC and OM occurred on lower landscape positions (Fig. 3.6 and 3.7). Where these maps differ is in the extent of regions of high CEC and OM. In SSURGO, the regions of high OM and CEC are much larger in extent compared to the DSM maps. Generally, SSURGO overestimates the areas with high OM and CEC. For example, SSURGO representative values had a median OM content of 4.9% compared to 4.1 and 4.2% for DSM maps (Table 3.4). Similarly, the SSURGO mean values for CEC had a median of 28.5% while DSM maps had a median between 19.4 and 20.0% (Table 3.4). Additionally, the standard deviation of SSURGO OM (1.2%) and CEC (6.6%) maps are higher compared to DSM Maps, which is less than 0.8% for OM and 2.8 – 3.3% for CEC (Table 3.4). This means that SSURGO OM and CEC values are more diffuse and spread-out when compared to the DSM values.

Table 3.4: Summary statistics of universal kriging (UK), Cubist, random forest (RF), and soil survey geographic (SSURGO) organic matter content (OM) and cation exchange capacity (CEC) maps.

Statistical Index	OM				CEC			
	UK	Cubist	RF	SSURGO	UK	Cubist	RF	SSURGO
	---------- % ----------				---------- $cmol_c\ kg^{-1}$ ----------			
Minimum	0.9	1.9	2.4	1.6	1.1	5.2	11.9	7.0
1st Quartile	3.8	3.8	3.7	2.9	17.7	17.6	17.5	17.4
Median	4.2	4.2	4.1	4.9	20.0	19.9	19.4	28.5
Mean	4.2	4.2	4.2	4.0	19.6	19.6	19.7	23.5
3rd Quartile	4.7	4.6	4.6	4.9	22.0	22.1	21.8	28.5
Maximum	7.8	7.2	6.5	5.9	28.6	28.1	27.8	31.1
Standard Deviation	0.7	0.8	0.7	1.2	3.2	3.3	2.8	6.6

One interesting area of agreement between SSURGO and DSM maps is for CEC predictions in the southern quarter of the study area. In this area, both SSURGO and DSM models predicted the lowest CEC values. Despite the fact that sampling points were not concentrated at this part of the study site (Fig. 3.1), DSM models still managed to predict these regions of low CEC. Low DSM-derived CEC predictions likely resulted from the low TPI in the study areas (see section 4.1). While SSURGO was not developed using TPI specifically, SSURGO mapping did rely heavily on relationships between soils and landscape positions. Agreement between DSM-predicted CEC and SSURGO maps highlights the importance of soil-landscape relationships in soil spatial distributions.

We compared OM and CEC predicted by DSM techniques to OM and CEC from the SSURGO soil map. Specifically, we compared OM and CEC contents predicted by DSM to the OM and CEC contents within each map unit from SSURGO (Fig. 3.8). Both OM and CEC show that the three predictive models follow similar prediction trends in each of the SSURGO mapping units. The results of our models for OM are consistent with the estimates from eight SSURGO mapping units; exceptions were CwB2, McS2, RoB, SwA, TfB, and TmA. SSURGO underestimated the OM for these map units while the other models predicted greater concentrations of OM. Generally, SSURGO had a wider range in OM and CEC values when compared to the prediction models. This was seen particularly for CEC estimates. The prediction of our models for CEC is consistent with a few of the SSURGO mapping units see: RcA, RoB, SW, TfB, and TmA. However, for most of the mapping units, our models either over- or under-predicted CEC.

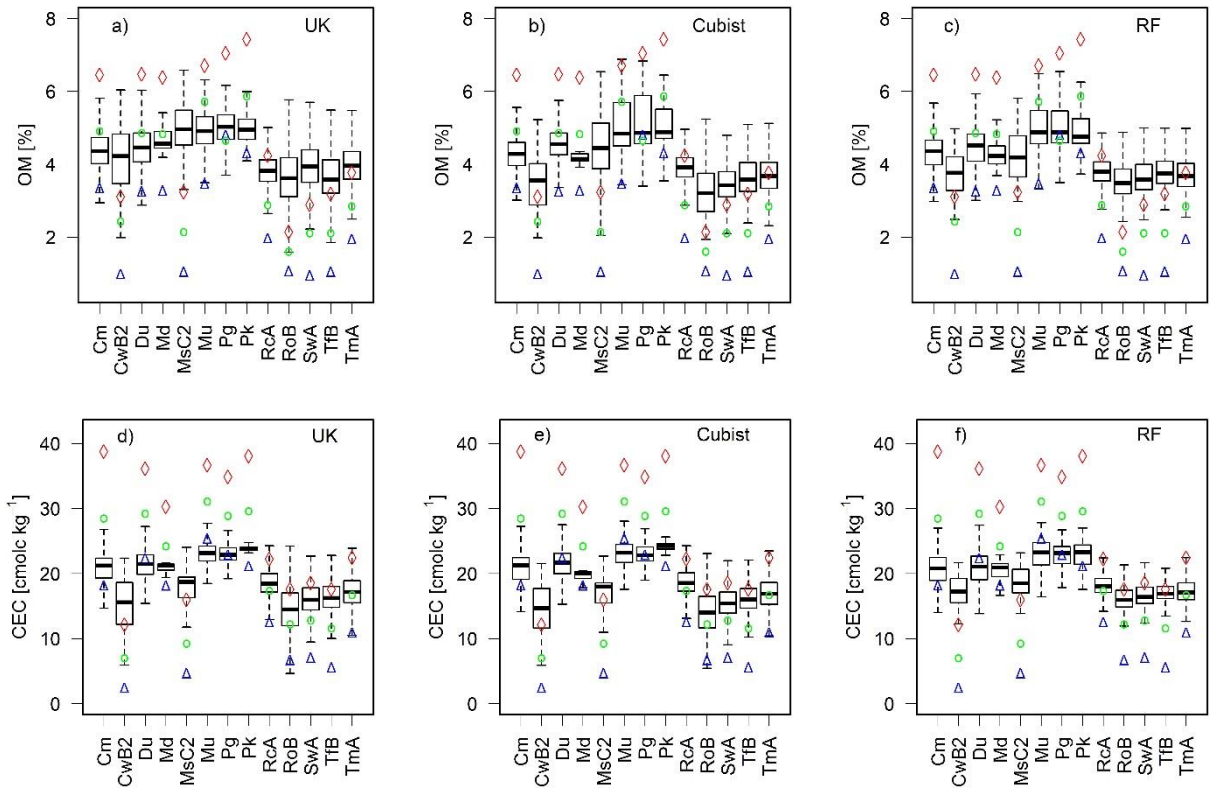

Figure 3.8: Comparison of predictive performance of DSM models with soil survey geographic (SSURGO) organic matter content (OM) and cation exchange capacity (CEC) estimates. The boxplots of a), b) and c) show OM and d), e), and f) show CEC prediction based on universal kriging (UK), Cubist, and random forest (RF) respectively. Triangles show low values and rhombuses show high values. Circles show representative OM and CEC mean values.

There are several reasons for the inconsistencies of model predictions with SSURGO. First, SSURGO has inherent limitations; the soil variability is represented using aggregated polygon map units with one to four named components plus inclusions of other soils or non-soils areas that do not explicitly capture the underlying spatial variability of soils within polygons (Nauman and Thompson, 2014). Thus, these inclusions reduce the purity of the map units and impact interpretation and modeling (Geza and McCray, 2008). Second, the procedure for OM analysis differed between the datasets. The Walkely-Black method was used for the SSURGO data, while the loss-on-ignition (LOI) method was used for our collected data. Due to incomplete digestion of

soil organic carbon, the Walkley-Black method usually underestimates OM (De Vos et al., 2007; Conyers et al., 2011). Additionally, the SSURGO values might have been impacted by errors introduced by the spline interpolation. Third, the SSURGO database was developed based on historical soil survey data and may not accurately reflect the current status of soil properties, particularly OM and CEC, which are relatively dynamic and altered by various factors such as land management, climate change, and wild fires to name a few (Schoonover and Crim, 2015; Bot and Benites, 2005; Grigal and Vance, 2000). Additionally, the data were produced on different dates and therefore inherit inconsistencies (Nauman et al., 2012). A fourth reason for the inconsistency is that the surveyors who collected data for SSURGO may not have had enough soil observations for building their mental models of soil formation at this small scale. A fifth reason for the inconsistency is that SSURGO values are not purely derived from laboratory analysis, instead the data may have resulted from a combination of laboratory measurements and field observations of expert soil scientists (Libohova et al., 2016). Due to these shortcomings, using SSURGO data in quantitative modeling and/or for monitoring soil carbon stocks sequestration is misleading, particularly at the farm scale for highly variable soils in a glaciated landscape.

3.4 Conclusion

The prediction performance of all three models (UK, Cubist, and RF) was similar for both OM content and CEC estimation. Random forest (RF) and Cubist, however, slightly outperformed UK for both OM and CEC properties prediction on an independent evaluation dataset. Universal Kriging (UK), however, due to the simplicity, faster computation, and more interpretable forms is favored over data mining or machine learning algorithms such as RF and Cubist and is recommended for future studies, at least for ACRE.

All three predictive models showed similar spatial predicting trends that were comparable to the SSURGO map units. Overall SSURGO had a wider range and/or either slightly under or over predicted soil properties when compared to the other models. Considering the high variability in farm management practices and nutrient application, the prediction accuracies were considered reasonable. The results demonstrate that lidar data alone can be used to adequately predict soil OM and CEC at the farm scale in this glaciated soil landscape.

CHAPTER 4. SPATIAL PREDICTION OF NATURAL SOIL DRAINAGE CLASSES USING DIGITAL SOIL MAPPING TECHNIQUES

Abstract

Accurate spatial prediction of natural soil drainage condition is not only important for agriculture and hydrological modeling but also for installing subsurface drainage and onsite waste disposal systems. For this research, 154 sites were selected based on a stratified random sampling method. For each site, drainage class was identified based on visual examination of soil cores. A digital elevation model developed from lidar data was used to derive seven terrain indices. Terrain indices were used to predict drainage class using four prediction models: multinomial logistic regression and three machine learning algorithms (random forest, C5.0, and artificial neural network). Based on 30% random hold-back validation data, all digital soil mapping (DSM) models provided similar results. The overall accuracy ranged between 66 – 70% and kappa coefficient ranging between 0.53 and 0.59. The DSM models slightly outperformed SSURGO, which had an overall accuracy of 64% and kappa of 0.52.

4.1 Introduction

Natural soil drainage class is an important soil property that influences crop growth and phenotypic response through aeration, nutrient, and water distribution. It also affects water flow and solute transport through soils (Kravchenko et al., 2002). Accurate maps of soil drainage classes are needed for soil and land management (i.e. tile drain installation and site selection for onsite septic system installation), and hydrological and environmental modeling.

The Soil Survey Geographic (SSURGO) database (Soil Survey Staff, 2020) currently contains the best available information on soil drainage classes for the U.S. The most detailed

mapping in the SSURGO database is at a scale of 1:15,840 (Soil Science Division Staff, 2017). More detailed maps are, however, needed for sustainable land management, precision agriculture, and plant phenotyping. The Purdue University Agronomy Center for Research and Education (ACRE) research farm is utilized more than 50 researchers and provides plots for ~180 research projects (ACRE, 2019). Since the announcement of the Purdue Plant Science Initiative in 2013 (Robinson, 2013), ACRE has become the centerpiece for high-tech, field-based phenotyping research. Natural soil drainage class has a marked effect on plant growth and phenotypic characteristics, therefore, methods to generate explicit, accurate, consistent, spatially realistic, and inexpensive soil drainage class maps are needed to support field-based phenotyping studies at ACRE.

There is a strong correlation between natural soil drainage and hillslope hydrological processes. Therefore, research scientists utilize numerous analytical methods to map soil drainage classes based on terrain attributes derived from digital elevation models (DEMs). Kravchenko et al. (2002) applied discriminate analysis and geostatistics to map three soil drainage classes in central Illinois, U.S.A. based on topographic and soil electrical conductivity data. Liu et al. (2008) used multivariate discriminant analysis to map three soil drainage classes in Ontario, Canada based on topographic variables, remotely sensed images, and apparent soil electrical conductivity. Cialella et al. (1997) used a decision tree classification method to map five drainage classes for a 24 km^2 area in Howland, Maine, USA based on topographic covariates and remote sensing images. Niang et al. (2012) also applied a decision tree model to predict soil drainage classes for a 167 km^2 area in Quebec, Canada, as did Møller et al. (2019) in order to develop a soil drainage class map in Denmark. Campling et al. (2002) integrated topographic and vegetation indices to develop a

probability map of drainage classes using a logistic model, while Zhao et al. (2013) predicted seven soil drainage classes in Nova Scotia, Canada using artificial neural networks (ANN).

In this study, multinomial logistic regression (MNLR), C5.0 decision tree, random forest (RF), and artificial neural network (ANN) models were used to predict natural soil drainage classes based on a high-resolution digital elevation model (DEM) derived from light detecting and ranging (lidar) data. In many studies utilizing digital soil mapping to map natural soil drainage classes, a DEM and terrain attributes are used in conjunction with other covariates (e.g. Bell et al. 1992 & 1994; Cialella et al. 1997; Lemercier et al. 2012). In this study, we focused on the use of lidar data alone for the following reasons. (1) For our study area there are challenges associated with utilizing vegetation indices for mapping soil drainage classes because of the many small research fields with heterogeneous experiments. Thus, crop and soil reflectance will differ between fields mainly due to crop residue management, different crops, and other factors that are not necessarily related to soil forming factors. (2) High quality lidar data is available for the area. (3) There is a close correlation between landscape position and natural soil drainage class that has been used by soil mappers in the area for many years.

The objective of this study was to predict the spatial distribution of natural soil drainage classes across the study area at a greater level of detail than the current SSURGO soil map. The specific objectives were to: 1) evaluate the relationship between soil drainage classes and DEM derived topographic indices, 2) compare the prediction accuracy of soil drainage class maps developed by MNLR and machine learning or decision tree models (C5.0, RF, and ANN), and 3) compare the predictive performance of digital soil mapping (DSM) models to the traditional soil map (i.e. the SSURGO map).

4.2 Materials and Methods

4.2.1 Study Site Descriptions

The study site is located at the Purdue Agronomy Center for Research and Education (ACRE) in Tippecanoe County, Indiana, USA (Fig. 4.1). The study site comprises 570 hectares and has gently undulating, low relief topography (on average 1% slope based on a 3 x 3 pixel window) (ACRE, 2019). Soils at this site are formed in about 50 cm of loess over loamy Wisconsin glacial till and outwash. ACRE is located at the transition between the Eastern Hardwood Forests and the prairies of the Great Plains, and Alfisols and Mollisols are the two most common soil orders (USDA-NRCS, 1998). There are 14 different soil mapping units and 18 soil series at ACRE (USDA-NRCS, 1998). For the 30-year period from 1981 – 2010, the mean total annual precipitation is 970 mm and the mean annual temperature is 10° C (MRCC, 2013), while the mean winter temperature is -2.6° C and mean summer temperature is 22.2° C (NWS-COOP, 2020). The area is in the udic soil moisture regime and the mesic soil temperature regime (USDA-NRCS, 1988). Soil drainage classes range from very poorly drained (VPD) to moderately well drained (MW). The major crops at ACRE are corn (*Zea mays* L.) and soybean (*Glycine max* (L.) Merr.).

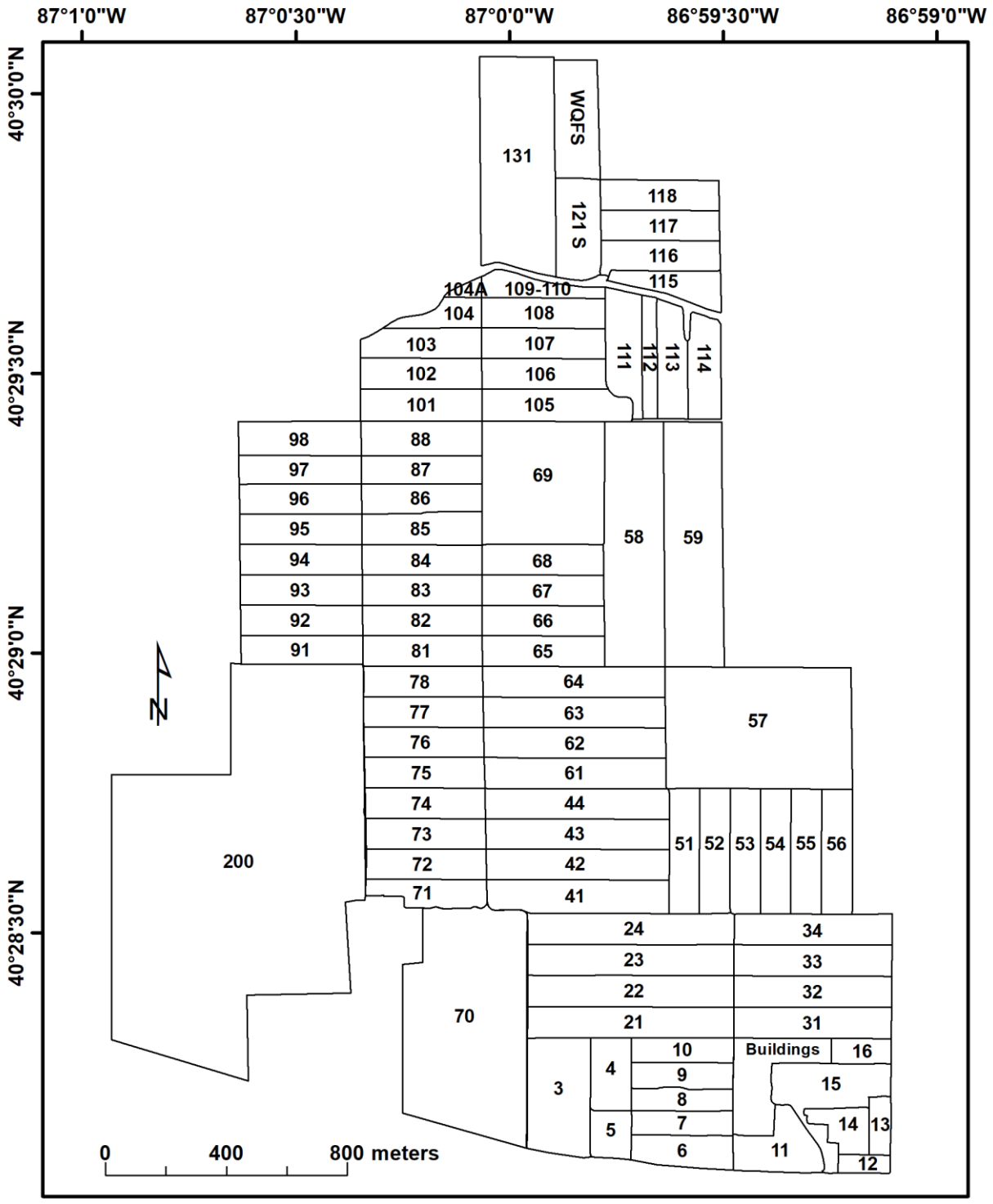

Figure 4.1: Geographic location of the study site and field plot layout of the Purdue Agronomy Center for Research and Education (ACRE). WQFS: water quality field station.

4.2.2 Field Data Collection

We collected 154 field observations of natural soil drainage class at ACRE (Fig. 4.2). Sampling locations were selected using a stratified random sampling design. Stratification was based on the drainage class from the SSURGO map, with 40 sampling locations randomly selected in each of the drainage classes. Drainage classes in SSURGO included very poorly drained (VPD), poorly drained (PD), somewhat poorly drained (SWP), and moderately well drained (MW). Out of 160 sampling points, six were in areas of disturbed soil (e.g. buildings or parking lots) and, thus they were excluded. The drainage class of each sampling location was determined by visual examination of cores obtained with a Dutch auger. We used the criteria (Fig. 4.3) described by Franzmeier et al. (2001) to define the drainage class at each sampling location (Fig. 4.4). Once the drainage class at a given location was determined by two field experts, the coordinates of the location were recorded using a Bad Elf Global Navigation Satellite System (GNSS) Surveyor receiver accurate to ±1 m (Bad Elf, 2020). The samples were collected from south to north. Most of the MW soils are located in southern part of the study area and thus were collected first. While, the rest of the points were collected based on geographic proximity and ease of access. The field data were split into training or calibration data (70%) and testing or validation data (30%) based on a stratified random split to maintain equal proportions of drainage classes in both datasets (Table 4.1).

Table 4.1: Soil drainage classes and number of collected field samples in the whole dataset, calibration dataset, and validation dataset.

Data Type	Number of soil samples in each drainage class				Total
	VPD	PD	SWP	MW	
Calibration	30	27	35	15	107
Validation	13	12	15	7	47
Total	43	39	50	22	154

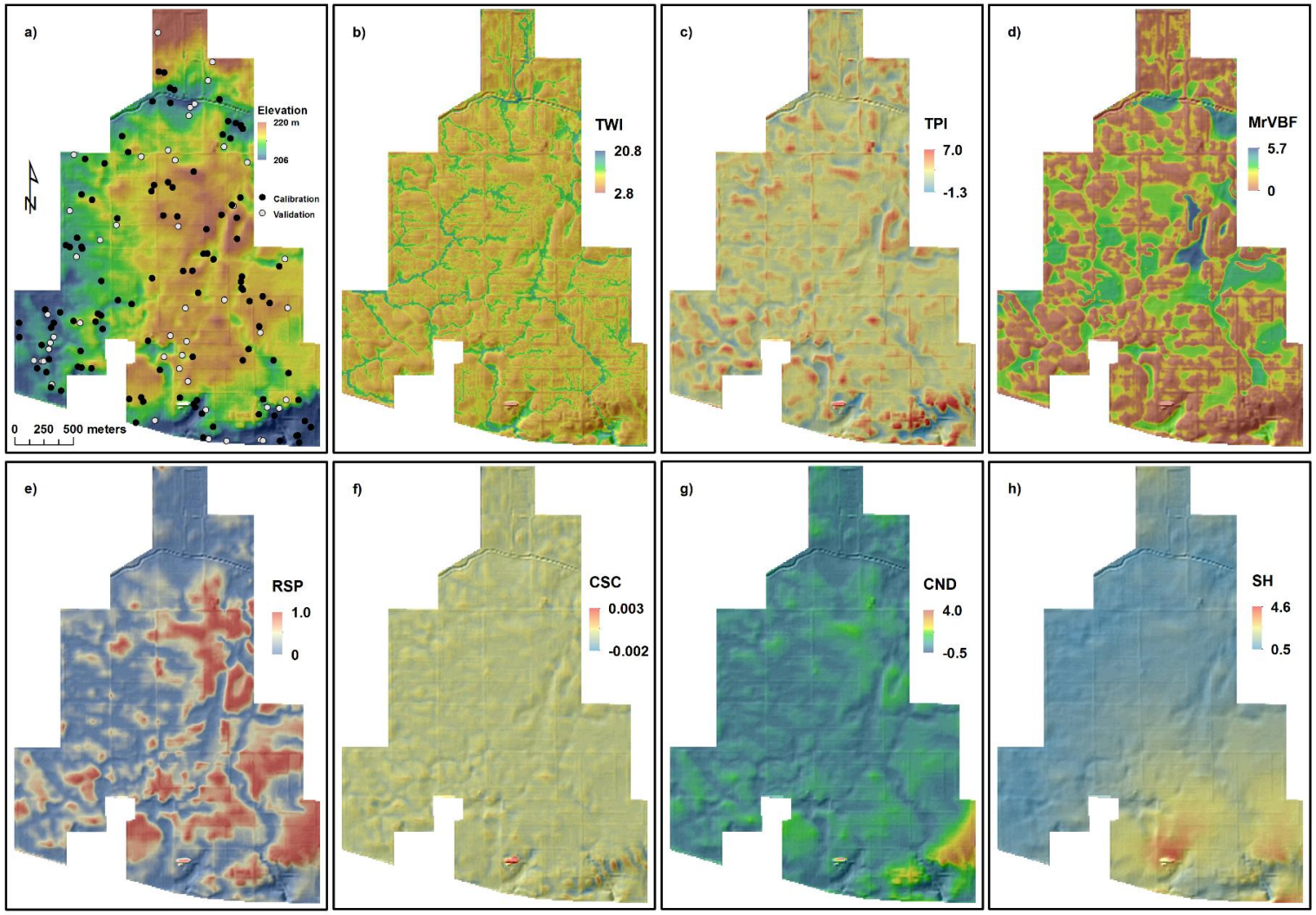

Figure 4.2: Lidar digital elevation model and derived terrain covariates for the study area. **(**a) Elevation map with soil drainage class sampling locations, (b) topographic wetness index (TWI), (c) topographic position index (TPI), (d) multi resolution valley bottom flatness index (MrVBF), (e) relative slope position (RSP), (f) cross sectional curvature (CSC), (g) channel network distance (CND), and (h) slope height (SH).

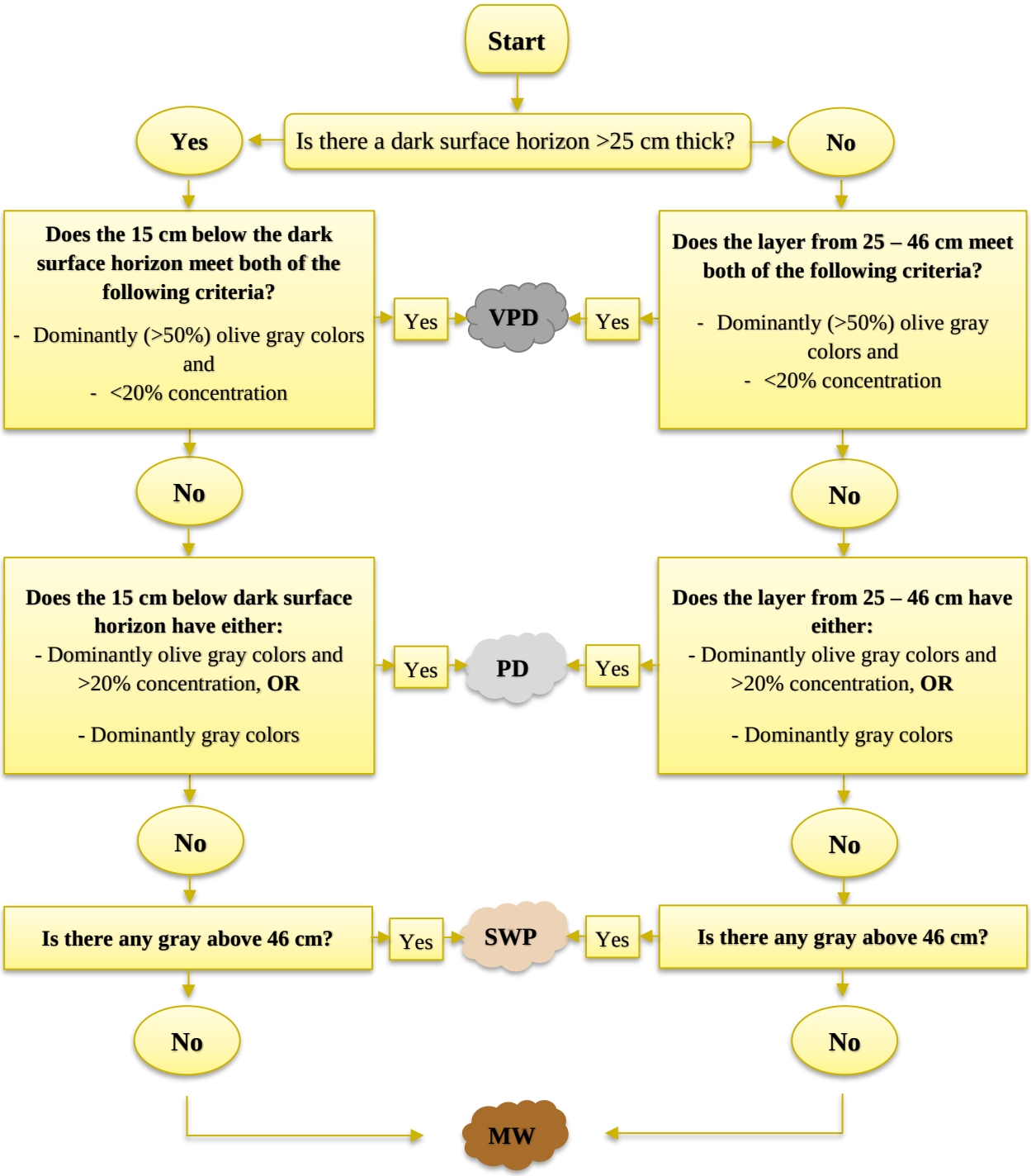

Figure 4.3: Steps for determining the soil drainage classes in the field. **Dark:** value ≤ 3 and chroma ≤ 3. **Gray:** hue = any, value ≥ 4 and chroma ≤ 2. **Olive gray:** hue = 2.5Y or 5Y, value ≥ 4, and chroma ≤ 2 (Adapted from Franzmeier et al. 2001).

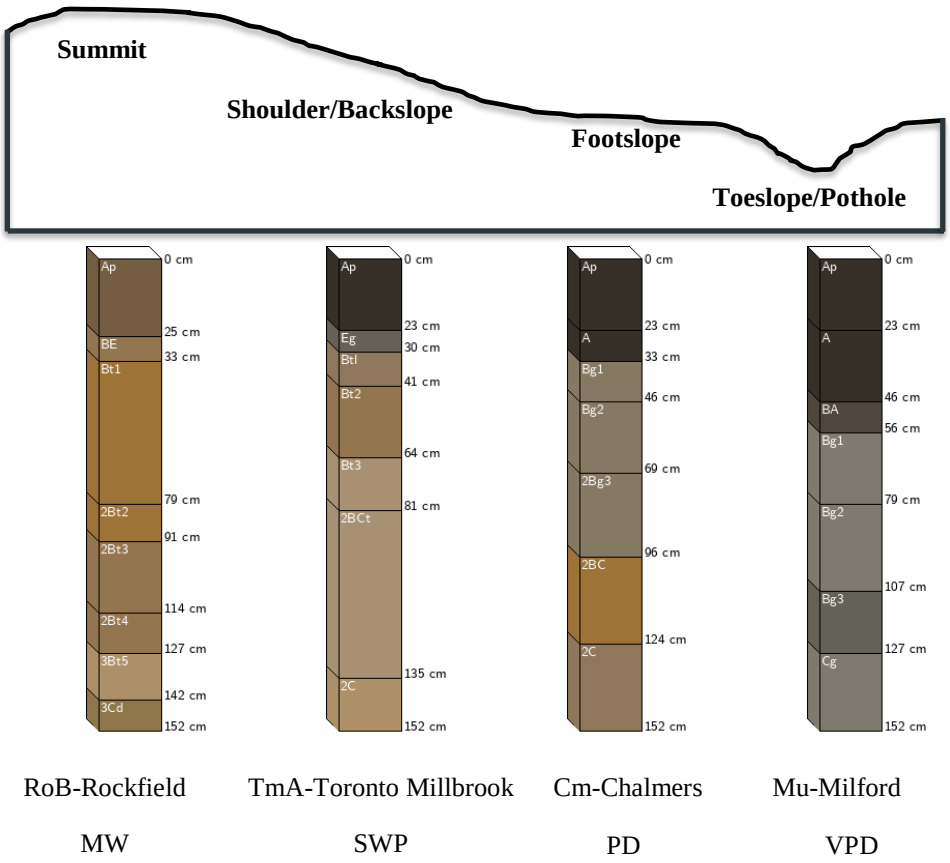

Figure 4.4: Occurrence of soil drainage classes on the landscape positions of the study site. The bottom two lines show soil mapping unit and soil drainage classes, moderately well drained (MW), somewhat poorly drained (SWP), poorly drained (PD), and very poorly drained (VPD). Soil profiles were obtained from Soil-Web (Beaudette and O'Geen, 2009).

4.2.3 Environmental Covariate Data

The 1.5 m pixel resolution lidar-derived digital elevation model acquired in 2013 for Tippecanoe County, Indiana was obtained from the Indiana Spatial Data Portal website (http://gis.iu.edu/). The DEM was re-projected from the Indiana State Plane West Coordinate System to the Indiana Geospatial Coordinate System (InGCS) for Tippecanoe and White Counties developed by the Indiana Department of Transportation (INDOT, 2016) using ArcMap 10.6 (https://esri.com).

Initially terrain attributes were calculated based on the original 1.5-m DEM. However, roads and field boundaries (Fig. 4.1) that are ~20 cm higher, on average, than the cultivated fields interfered with the distribution patterns of the calculated terrain attributes. Fine resolution DEMs with pixel sizes on the order of 1 – 2 m are often too detailed and not desirable for mapping soil spatial variability (Smith et al., 2006; Winzeler et al., 2008; Shi et al., 2012; Maynard and Johnson, 2014; Lacoste et al., 2014). To smooth out these anthropogenic features, the original 1.5 m DEM was resampled to 10 m based on simple mean aggregation in ArcMap 10.6 (https://esri.com).

Topographic covariates are impacted by the extent of watershed. Therefore, the United States Geological Survey – National Hydrography Dataset (USGS-NHD) was obtained from the USDA, NRCS Geospatial Data Gateway (https://datagateway.nrcs.usda.gov/) and used to delineate the complete network of watersheds that flowing into and out of ACRE. Based on this channel network, a buffer around ACRE was created and the resampled DEM was clipped to the buffer.

The algorithms in SAGA-GIS 2.1.4 (Conrad et al., 2015) were used to generate seven terrain attributes or environmental covariates (Fig. 4.2) from the resampled DEM: (1) relative slope position (RSP), (2) cross sectional curvature (CSC), (3) channel network distance (CND), (4) slope height (SH), (5) topographic wetness index (TWI), (6) topographic position index (TPI), and (7) multiresolution valley bottom flatness index (MrVBF). In the section below, we will only discuss the first four environmental covariates predictors. Information about the last three covariates (TWI, TPI and MrVBF) are presented in Chapter 3.

Relative Slope Position

Relative slope position (RSP), which is also known as relative hillslope position (Behrens et al., 2010), combines altitude below ridge lines with altitude above channel networks (Bock et

al., 2007). In other words, RSP measures the position of a given location relative to the slope of a ridge (crest) and valley using the following equation:

$$RSP = \frac{Z_i - Z_v}{Z_r - Z_v} \quad [1]$$

where, Z_i is the elevation of a given location, Z_v is the elevation of the adjacent valley, and Z_r is the elevation of the adjacent ridge. The RSP values range between zero (downslope or channel lines) and one (upslope or ridge lines) (Conrad et al., 2015).

Cross Sectional Curvature

Cross sectional curvature (CSC) shows the divergence and convergence of flow across the land surface. This index calculates the curvature perpendicular to the steepest slope direction (Pipaud and Lehmkuhl, 2017). A negative value of CSC indicates a concave slope in the cross-sectional direction where water converges. A positive CSC shows that the slope is convex along the cross-sectional direction and represents a ridge where water diverges. CSC values close to zero, show planar or flat areas (Ehsani and Malekain, 2011).

Channel Network Distance

The behavior of water flow is different in channels than in other areas. Therefore, channel network distance (CND), which quantifies the distance of each pixel to the nearest channel or stream network, is an important index that provides information about the hydrological characteristics of channel and non-channel cells (Olaya, 2004). Lower values of CND are found near channels and ground water and, thus, are characterized by water accumulation. In contrast, higher values of CND are found on the plateaus (planar uplands) and farther away from channels. Medium values of CND show material transfers on slopes (Boehner et al., 2002).

Slope Height

Slope height (SH) is defined as the relative height difference to the immediate nearby ridge line. In other words, SH is estimated based on calculating the vertical distance from the lower position of the hillslope to the crest of the hillslope (Malone et al., 2018).

4.2.4 SSURGO Data

For this study, SSURGO is used as the reference, conventional soil drainage class map. The SSURGO data was downloaded from the Web Soil Survey website (Soil Survey Staff, 2020). The SSURGO database provides detailed soil survey and mapping information for the Tippecanoe County at a predominant scale of 1:15,840 (USDA-NRCS, 1998). Four natural soil drainage classes, very poorly, poorly, somewhat poorly, and moderately well drained, occur in the study area. For each map unit, the natural soil drainage class of the dominant component was assigned as the drainage class for the entire map unit.

4.2.5 Spatial Inference Mapping Models

To develop digital soil maps of drainage classes for the study area we tested four prediction models: (1) multinomial logistic regression (MNLR), (2) the C5.0 decision tree model, (3) random forest (RF), and (4) artificial neural network (ANN). The last three models are considered decision tree or machine learning models. Each of the four models relates topographic information to the occurrence of soil drainage classes and quantifies the relationship, and spatially predicts drainage classes across the landscape. Each model is described briefly below.

Multinomial Logistic Regression

Multinomial logistic regression (MNLR) is a form of a generalized linear model and is used to predict response variables containing more than two categories based on a set of multiple independent variables (Hosmer and Lemeshow, 1989). The independent variables can be continuous, discrete or both. Soil drainage class is a categorical response variable, therefore MNLR may be suitable for estimating the occurrence of soil drainage classes from topographic covariates.

Unlike logistic regression that has only one logit or log odds equation, MNLR has multiple or N – 1 logit equations (Abdel-Kader, 2011). Logit is a logarithmic function that shows the ratio of probability (*p*) that a given pixel belongs to a specific category/class divided by the probability that it is not (1 – *p*) (Abdel-Kader, 2011) and it is expressed as:

$$logit_i = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_n X_n + \mathcal{E} \quad [2]$$

where β_0 is a constant (y – intercept), β_n is regression coefficients with n=1, 2, …, n–1, X_n is a vector of predictor variables, and $\mathcal{E}$ is random error. From the above equation we can determine the probability that a pixel belongs to a specific class (*k*) as follows:

$$\boldsymbol{p}(i = k) = \frac{exp^{(\beta_0 + \beta_n X_n)}}{1+\sum_1^{k-1} exp^{(\beta_0 + \beta_n X_n)}} + \mathcal{E} \quad [3]$$

The logit of one category (typically the first or last, or the value with the highest frequency) is not estimated because it is considered as the reference category. However, its probability of presence is determined using:

$$\boldsymbol{p}_r = \frac{1}{1+\sum_1^{k-1} exp^{(\beta_0 + \beta_n X_n)}} + \mathcal{E} \quad [4]$$

Through an exhaustive search, and based on external validation, we tested various assemblages of terrain attributes aiming to find a model with a higher accuracy and kappa

coefficient. MNLR was carried out in R 3.5.1 environment (R Core Team, 2018), using the *multinom* function of the *nnet* package (Venables and Ripley, 2002).

C5.0 Decision Tree Model

The C5.0 decision tree model also known as See5 is the successor of the C4.5 model and is a sophisticated data mining algorithm developed by Quinlan (1993). The C5.0 decision tree model (Fig. 4.5) is used to find patterns of categories from organized data, assemble these patterns into classifiers and finally make predictions (Quinlan, 1993). The C5.0 model splits the data based on the maximum information gain criteria. Therefore, for each node tree, the C5.0 decision tree model selects a covariate that results in providing more information to make the decision (Quinlan, 1993).

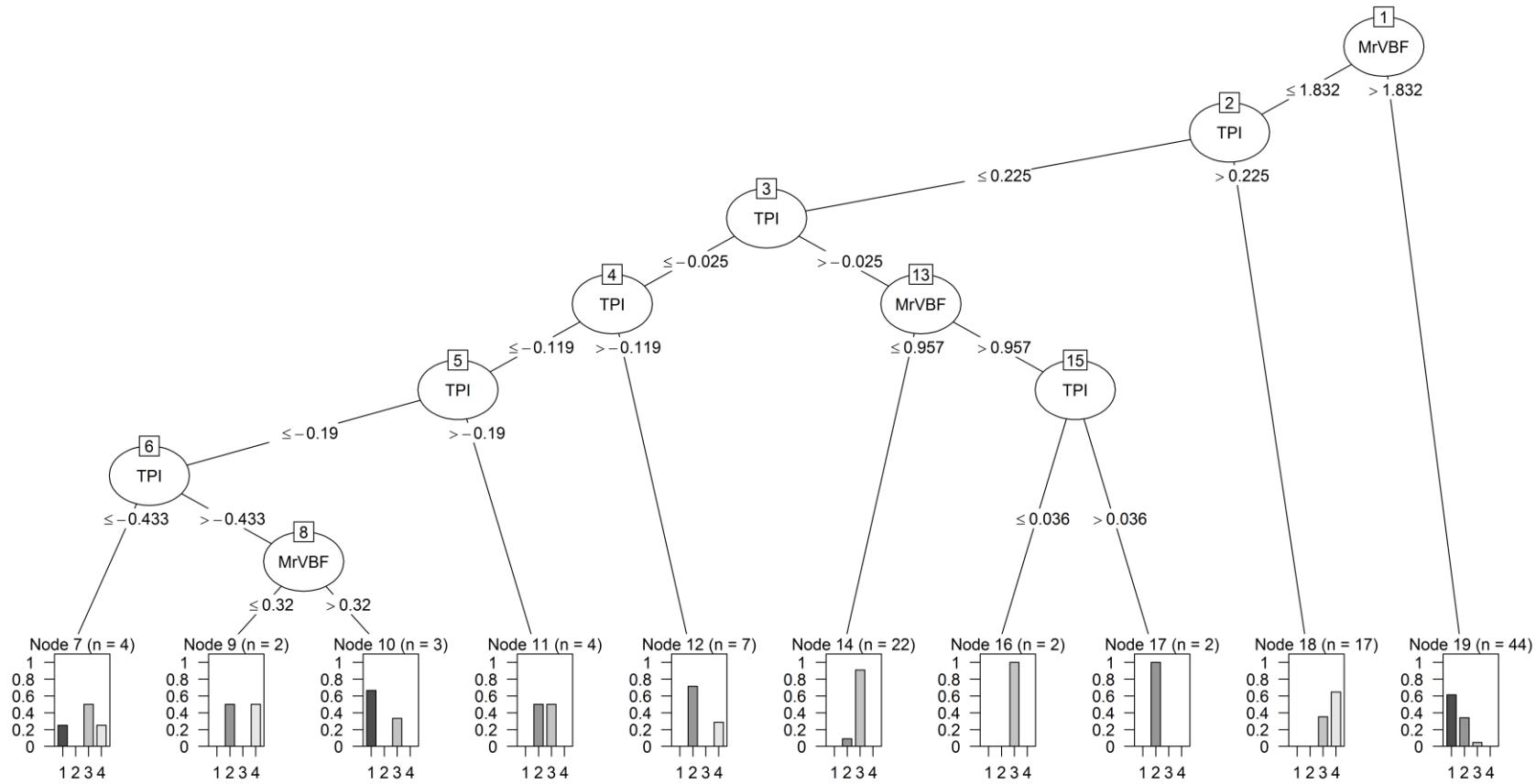

Figure 4.5: Graphical representation of the C5.0 decision tree model for the current study. MrVBF: multiresolution valley bottom flatness index; TPI: topographic position index. In the output layer, n shows the number of observations that is used to determine the final drainage class(es) and 1 represents very poorly drained, 2 represents poorly drained, 3 shows somewhat poorly drained, and 4 shows moderately well drained soil.

Pruning is necessary to removing parts of the generated tree that contribute little in classification. C5.0 initially generates a fully-grown tree to fit the data and afterwards prunes the tree by excluding parts with the highest error rates (Adhikari et al., 2014). The *C5.0* package (Kuhn and Quinlan, 2018) was used in the R 3.5.1 environment (R Core Team, 2018) to spatially predict drainage classes.

Random Forest

Random Forest (RF) is an ensemble modeling approach developed by Breiman (2001) as an extension of classification and regression trees (CART model) to enhance the prediction performance of the model (Wiesmeier et al., 2011). Random Forest has been widely adopted and has become a dominant decision tree model in digital soil mapping (Grimm et al., 2008; Stum et al., 2010; Wiesmeier et al., 2011; Forkuor et al., 2017; Adhikari et al., 2018).

RF generates decision trees based on strong predictors and by repeatedly drawing random and independent bootstrap samples from the training data (Stum et al., 2010; Forkuor et al., 2017). The rest of the data, which is known as out-of-bag, is used for validation of the generated trees. As a rule of thumb, for each decision tree RF takes approximately 2/3 of the training data for bootstrap sampling and 1/3 for out-of-bag validation (Peters et al. 2007).

The number of trees (ntree) and the number of covariates (mtry) are important input parameters determined by the user. Random forest randomly selects the strongest covariates (mtry) to split the nodes of each tree. The results of RF prediction improve by utilizing many predictive trees (Adhikari et al., 2018). In this study, we tested the RF performance using different numbers of trees (ntree) starting from 500 up to 1500 in increments of 100. While mtry can be manually specified, RF automatically attempts to optimize mtry (Malone et al., 2017). The *randomForest*

package (Liaw and Wiener, 2002) was used in the R 3.5.1 environment (R Core Team, 2018) to spatially predict drainage classes for the study site.

Artificial Neural Network

Artificial neural network (ANN) is a deep-learning technique (a powerful branch of machine learning) that mimics the processing of information in human brains (Hewitson and Crane 1994). ANN finds patterns and classifies new and unknown data based on associations between predictor variables and observation or training points (Zell et al., 1998). ANN can handle large (Chagas et al., 2013), noisy, and non-linear datasets (Rossel and Behrens, 2010). Additionally, it can handle both regression and classification problems.

In this study, we used a backpropagation neural network method (Rumelhart et al., 1985; Günther and Fritsch, 2010). This method reduces the overall learning error through a reverse direction (from the output layer to the input layer) (Gallant, 1993). The ANN model that was used has three interconnected layers (Fig. 6). First, an input layer that contains the terrain attributes. Second, a hidden layer that has three different artificial neurons and connects the input layer with the output layer. Similar to Bodaghabadi et al., 2015 and Ghaderi et al., 2019, the performance of ANN was evaluated with different numbers of hidden neurons staring from 2 and increasing one at a time up to 30. Optimal performance based on overall accuracy and kappa were achieved with three hidden layers. Third, an output layer with four neurons, with each neuron estimating a drainage class. The output of a neuron is derived from the following function (Eq. 5), which is basically a weighted sum of input variables (i.e. terrain attributes) plus the bias weight (Ciaburro and Venkateswaran, 2017).

$$\text{Output (drainge class)} = \sum(\text{weights} * \text{inputs or terrain attributes}) + \text{bias weight} \quad [5]$$

For the ANN model predictions, the *neuralnet* package (Fritsch et al., 2019) in R 3.5.1 environment (R Core Team, 2018) was used.

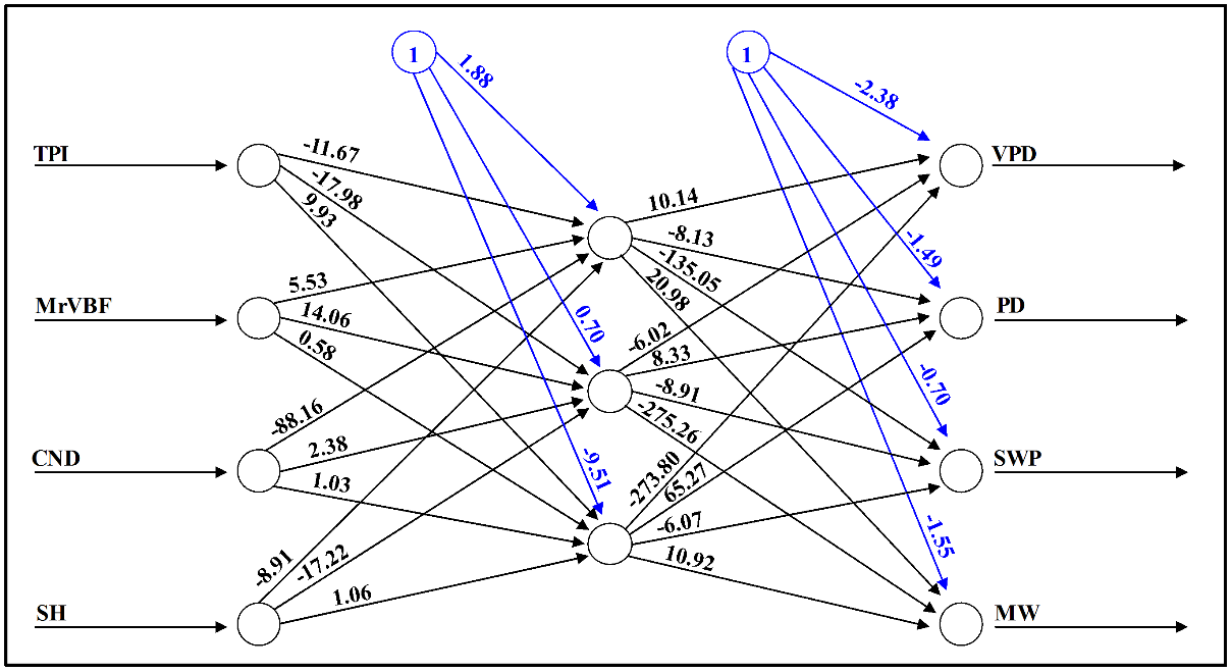

Figure 4.6: Graphical representation of the developed artificial neural network model (ANN) for the current study. TPI: topographic position index; MrVBF: multiresolution valley bottom flatness index; CND: channel network distance; SH: slope height. The blue lines represent the bias weight. In the output layer, VPD represents very poorly drained, PD represents poorly drained, SWP represents somewhat poorly drained, and MW represents moderately well drained soils.

4.2.6 Selection of Predictor Variables

For MNLR, C5.0, and RF models, the *varImp* function of the *caret* package (Kuhn, 2008) in R 3.5.1 environment (R Core Team, 2018) was used to determine the importance of each variable (Table 4.2). Similar to the *varImpPlot* function of the *randomForest* package (Liaw and Wiener, 2002), the *varImp* function determines the importance of a variable based on the Gini index (MeanDecreaseGini). The MeanDecreaseGini index shows the average decrease in node impurities across overall trees from splitting on the variable. Variables with higher values indicate greater importance in the model and if removed, greatly affect the node purity and the predictive

power of the model. Generally, the node impurity for decision trees or classification is calculated based on the Gini index, while for regression it is calculated based on the residual sum of squares (Malone et al., 2017).

Table 4.2: Terrain attributes and their overall importance.

Terrain Attributes	Overall Variable Importance			
	RF[*1]	**MNLR**[*2]	**C5.0**	**ANN**[*3]
Topographic Wetness Index	9			
Topographic Position Index	13		59	4813
Multiresolution Valley Bottom Flatness Index	13	5	100	-3749
Relative Slope Position	11			
Cross Sectional Curvature	11			
Channel Network Distance	12	26		-2492
Slope Height	9	1		4566

[*1]Random forest (RF), [*2]multinomial logistic regression (MNLR), [*3]artificial neural network (ANN).

The variable importance for the ANN model (Table 4.2) was determined based on Olden's function in the *NeuralNetTools* package (Beck, 2018) in R 3.5.1 environment (R Core Team, 2018). Olden's function evaluates the variable importance based on the sum of raw input-hidden and hidden-output connection weights between each input and output node or neuron (Olden, 2004). The derived importance values for Olden's algorithm resulted from the summed product of model weights, thus they were not rescaled (Beck, 2018).

4.2.7 Accuracy Assessment of the Predictive Models

The performance of the predictive models was assessed based on a stratified random-hold back independent validation using 30% of the collected field data. The Kappa coefficient (K), overall or observed accuracy (P_{obs}), and user's and producer's accuracies were used to assess the quality of the developed soil drainage class maps. The statistical indices are generated from the

confusion matrices. The *goofcat* function of the *ithir* package (Malone, 2018) was used in the R 3.5.1 environment (R Core Team, 2018) to compute these statistical validation indices.

$$\text{Kappa Coefficient } (K) = \frac{P_{obs} - P_{exp}}{1 - P_{exp}} \quad [6]$$

$$\text{Overall or Observed Accuracy } (P_{obs}) = \frac{\sum_{i=1}^{n} E_{ii}}{N} \quad [7]$$

$$P_{exp} = \frac{\sum_{i=1}^{n} X_{ij} Y_{ij}}{N^2} \quad [8]$$

$$\text{User's Accuracy } (U_{ac}) = \frac{X_{ii}}{\sum_{i=1}^{n} X_{ij}} \quad [9]$$

$$\text{Producer's Accuracy } (P_{ac}) = \frac{X_{ii}}{\sum_{i=1}^{n} Y_{ij}} \quad [10]$$

where P_{obs} shows the observed agreement between the prediction and the actual or reference data, in other words, P_{obs} shows the mean of pixels that classified correctly, P_{exp} shows the probability that agreement is due to chance, N is the total number of observations, n is the number of predicted soil drainage classes, E_{ii} is the sum of diagonal values, X_{ii} is the diagonal element of each class, X_{ij} is the sum of values in a row, and Y_{ij} is the sum of values in a column.

The Kappa coefficient measures the difference between the observed (actual) and expected (by chance) agreement. The value of the Kappa coefficient falls on a -1 to 1 scale, where a value of 1 indicates perfect agreement, 0 shows an agreement expected by chance, and negative values show less than chance agreements (Malone et al., 2017). The strength of the agreement of the predictive model was tested based on the scale proposed by Landis and Koch (1977). The Kappa coefficient shows less than chance agreement (K < 0), slight agreement (0.01 – 0.20), fair agreement (0.21 – 0.40), moderate agreement (0.41 – 0.60), substantial agreement (0.61 – 0.80), and almost perfect agreement (0.81 – 0.99).

User's accuracy is the probability that a predicted drainage class on the map actually represents that class in the field. Producer's accuracy indicates how well the model predicts the observed soil drainage class.

4.3 Results and Discussion

4.3.1 Important Predictor Variables

Overall, based on varImp and Olden's function, MrVBF and TPI were the most important predictors. Additionally, MrVBF was utilized by all four models, while TPI was used by three models. Taghizadeh-Mehrjardi et al. (2014) also found that MrVBF and wetness index were most effective for predicting soil classes, with MrVBF being particularly effective in relatively flat areas. One potential reason for the relatively lower importance of TWI in the RF model, and the fact that it was not being selected by other models, might be the collinearity (Pearson correlation coefficient values – 0.44 to 0.56) of TWI with other covariates (terrain attributes), except for SH and CND.

From a conceptual soil-landscape standpoint, there is a logical relationship between natural soil drainage classes and terrain attributes. For instance, the fitted MNLR equations (Eqs. 11 – 13) revealed that SH is negatively correlated to PD and positively correlated to SWP and MW. This underlines the fact that higher and steeper slope areas are drier compared to lower slope areas. Even though CND showed positive correlation with all soil drainage classes, the weight of coefficients was higher for SWP and MW compared to PD (Eqs. 11 – 13). On the other hand, MrVBF showed negative correlation with all soil drainage classes, though, strong positive correlation was expected with PD and negative correlation with SWP and MW. Nonetheless, the coefficient of MrVBF for PD is closer to zero compared to SWP and MW coefficients.

$$\mathbf{PD} = 0.04 - 0.29 * \mathrm{MrVBF} + 7.1 * \mathrm{CND} - 0.03 * \mathrm{SH} \qquad [11]$$

$$\mathbf{SWP} = 0.35 - 1.29 * \mathrm{MrVBF} + 8.95 * \mathrm{CND} + 0.28 * \mathrm{SH} \quad [12]$$

$$\mathbf{MW} = -1.36 - 3.27 * \mathrm{MrVBF} + 9.77 * \mathrm{CND} + 0.88 * \mathrm{SH} \quad [13]$$

Similar patterns between terrain attributes and soil drainage classes were also observed in the C5.0 model (Fig. 4.5), meaning that MrVBF has a positive correlation with VPD and PD (i.e. node 19), whereas, TPI has a positive relationship with SWP and MW (i.e. node 18).

4.3.2 Predictive Digital Soil Mapping Models

The performance of the predictive models is presented in Table 4.3. Based on overall accuracy and the Kappa coefficient, there was no great difference in the performances of all four models. The overall accuracy for MNLR, C5.0, and RF was 66%, while ANN resulted in a slightly higher overall accuracy (70%). The Kappa value of MNLR and RF was 0.53, and it was 0.54 for C5.0 and 0.59 for ANN. Even though the numbers of terrain attributes utilized by all models were different, some of the variables, particularly the most important variables (MrVBF and TPI), were common in all models. The exception to this was the MNLR model that did not utilize TPI. Therefore, it was expected that the models would show similar results (based on overall accuracy and Kappa coefficient) for the soil drainage class predictions. At the same study site, we found that MrVBF and TPI were the most important variables for cation exchange capacity (CEC) predictions based on universal kriging, Cubist and RF (see Chapter 3). Therefore, the models show similar predictive performances for CEC estimations. Based on overall accuracy and the Kappa coefficient it is also clear that the use of greater numbers of terrain attributes in a model does not necessarily result in higher accuracy. For instance, RF utilized all seven terrain attributes, while C5.0 only used MrVBF and TPI, but C5.0 still provided similar results when compared to RF.

Table 4.3: Producer's, user's, and overall accuracies, probability of chance agreement, and kappa coefficient of multinomial logistic regression (MNLR), C5.0, random forest (RF), artificial neural network (ANN) models and SSURGO database for very poorly drained (VPD), poorly drained (PD), somewhat poorly drained (SWP), and moderately well drained (MW) soils.

Model Type	Producer's Accuracy				User's Accuracy				Overall Accuracy	Chance Agreement	Kappa
	VPD	PD	SWP	MW	VPD	PD	SWP	MW			
MNLR	92	42	67	57	63	83	63	67	66	27	0.53
C5.0	92	33	60	86	52	67	90	75	66	26	0.54
RF	77	58	67	57	67	58	67	80	66	27	0.53
ANN	85	50	80	57	69	75	67	80	70	28	0.59
SSURGO	46	58	67	100	60	50	91	58	64	25	0.52

According to the user's accuracy of the confusion matrix (Table 4.3), SWP is the most accurately determined natural soil drainage class with a 90% user's accuracy, followed by PD with 84% and MW with 80% user's accuracy, while VPD had the lowest user's accuracy of 69%. The higher prediction accuracy of SWP by C5.0 might be attributed to the higher numbers of observations (35) for the training model. Møller et al. (2019) state that C5.0 decision tree models can handle missing values but prefer informative variables. Jafari et al. (2012) found that the number of the sampling points relative to the total area played an important role in mapping purity; hence smaller numbers of sampling points cause greater uncertainty.

Even though there were a limited number of training points (15) for MW soils, MW was still predicted with relatively high accuracy (67 – 80%) using all models. In contrast, VPD had the second highest number of observations (30) but estimated with relatively lower accuracy (53 – 69%). This high prediction of MW soils and lower prediction of VPD may be due to the soil relationship with the terrain attributes; both VPD and MW drainage classes are found in two distinct landscape positions. MW soils that evolved in a higher landscape position have a good relationship with terrain attributes and vice-versa in the case of VPD.

According to the producer's accuracy, all models underpredicted PD and MW soil drainage classes, except C5.0, which overpredicted MW. Møller et al. (2019) state that under-prediction of a soil class is due to its rarity in the training dataset or is related to the over-prediction of a majority class. The majority class contained the most cases of the underpredicted class. In this study, there were relatively lower number of training data for PD and MW compared to VPD and SWP soils (Table 4.1), however the majority class rule was the primary reason for the under-prediction of PD and MW in all models. Both MNLR and RF classified one third of the validation dataset cases of PD as VPD, while C5.0 classified two thirds and ANN classified one fourth of PD as VPD (Table

4.4). Similarly, almost half of the validation dataset cases of MW were classified as SWP by all models except the C5.0 model (Table 4.4). The main reason might be due to the close occurrence (both in geographic and feature spaces) of these underpredicted soil classes with their majority classes on the landscape. In other words, they are not only found geographically in close proximity, but also found on similar terrain. Additionally, ACRE has a relatively flat topography with low topographic variation (on average 1% slope based on a 3 x 3 pixel window), it is difficult for the models to differentiate between close occurrences of drainage classes. Even during field sampling, it was relatively hard to morphologically distinguish between VPD and PD soils. Furthermore, at ACRE, drainage class might change over a few meters. Therefore, observations based on a single boring may be misleading. It may be best practice to take several borings within a specific distance (i.e. 1 m radius) and assign drainage class based on the most common class within the area. Location accuracy of the individual samples is an additional source of error, but in our study the sampled locations were accurate to ±1 m, which is well below other sources of error. Another potential reason for under-predictions of PD by C5.0 might be the lower number of terrain attributes (MrVBF and TPI) utilized by C5.0 when compared to other models. The over-prediction of VPD by C5.0 might also show a close correlation between VPD and MrVBF, which is the most important variable in the C5.0 model.

Table 4.4: Confusion matrix for the drainage class determination for the predictive models and the SSURGO database.

		Actual or Reference Data																			
		Multinomial Logistic Regression				**C5.0 Decision Tree**				**Random Forest**				**Artificial Neural Network**				**SSURGO**			
		VPD	PD	SWP	MW	VPD	PD	SWP	MW	VPD	PD	SWP	MW	VPD	PD	SWP	MW	VPD	PD	SWP	MW
Predicted	VPD	12	4	3	0	12	8	3	0	10	4	1	0	11	3	2	0	6	4	0	0
	PD	1	5	0	0	1	4	1	0	2	7	3	0	2	6	0	0	7	7	0	0
	SWP	0	3	10	3	0	0	9	1	1	1	10	3	0	3	12	3	0	1	10	0
	MW	0	0	2	4	0	0	2	6	0	0	1	4	0	0	1	4	0	0	5	7

The misclassified drainage classes from all models were estimated within ±1 class of their actual drainage classes (Table 4.4). The exception to this was four SWP observations that were incorrectly estimated as a VPD soil by all models. Based on the DEM and terrain attributes, these four observations are in the lower landscape positions, however, the underlying parent material might be of glacial outwash that has coarser and sandy texture. In addition to the terrain covariates, there is a need for utilizing covariates that reflect underlying geology. Collecting geological data, however, is not easy. It is also possible that terrain attributes might predict potholes or depressions that were artificially created in the mid-slope positions as VPD instead of SWP.

Different models showed different strengths. According to the user's accuracy, ANN and RF provided better estimation for VPD and MW, while MNLR provided better results for PD and C5.0 provided better estimation for SWP. It is worth mentioning that RF did not predict PD and SWP with relatively high accuracies, but the results were consistent for producer and user accuracies. Additionally, RF and ANN resulted in similar user accuracies for all drainage classes except PD, which is predicted with lower user accuracy by RF when compared to ANN.

Overall, the performance attained from the prediction models is considered good in comparison with other studies predicting drainage classes (Kravchenko et al., 2002; Zhao et al., 2008). In our study, generally, good correlation was observed between soil drainage classes and terrain attributes in all predictive DSM models. The results achieved in this study are purely based on terrain covariates. Other studies utilized parent material and a geology layer, texture and/or clay and sand content, remote sensing and vegetation indices, land use, and wetlands, in combination with terrain attributes (Bell et al., 1992; Cialella et al., 1997; Campling et al., 2002; Peng et al., 2003; Liu et al., 2008; Zhao et al., 2008; Lemercier et al., 2012; Niang et al., 2012; Zhao et al., 2013; Møller et al. 2019). While the use of additional covariates proved to be efficient for

differentiating the drainage classes for the studies above, in heavily influenced anthropogenic landscapes, like ACRE, relationships between covariates and drainage classes may be obscured by management.

4.3.3 Comparison of Digital Soil Maps to SSURGO

There were similarities and difference between the digital soil map (DSM) and the traditional soil survey or a SSURGO map (Fig. 4.7 & 4.8). All DSM models, however, showed slightly higher accuracy and Kappa values than did SSURGO (Table 4.3). Additionally, the user accuracy of SSURGO for all drainage classes (except SWP) was lower than the DSM models. Based on producer accuracy, SSURGO, when compared to the DSM models, underpredicted VPD and overpredicted MW.

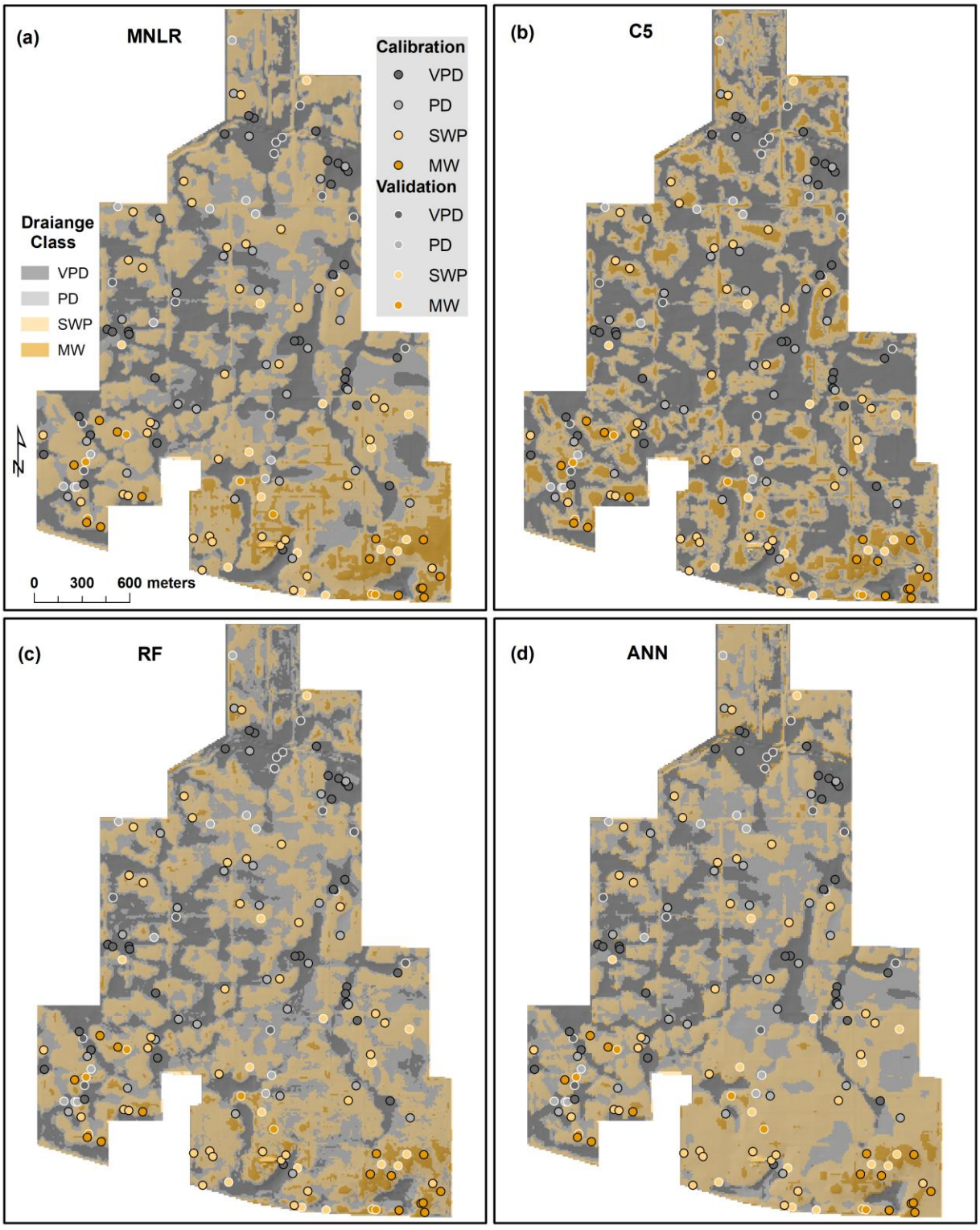

Figure 4.7: Prediction of natural soil drainage classes. (a) Multinomial logistic regression (MNLR), (b) C5, (c) random forest (RF), and (d) artificial neural network (ANN). The points with black rim represent calibration and points with white rim represent validation datasets.

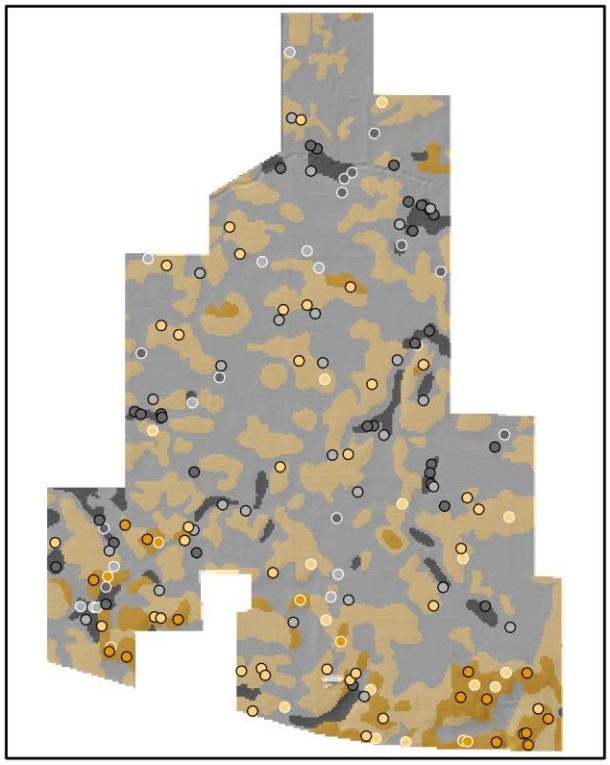

Figure 4.8: Prediction of natural soil drainage classes based on conventional soil survey (SSURGO data). The points with black rim represent calibration and points with white rim represent validation datasets.

Both DSM algorithms and SSURGO identified the occurrence of four drainage classes in the study site. Within the landscape, DSM and SSURGO were consistent with the conceptual pedological distribution of soil drainage classes. The predicted maps show that VPD and PD soils are found in lower landscape positions (i.e. foot and toe-slope and potholes), while SWP and MW (particularly MW) are found at higher and steeper landscape positions (i.e. summit, shoulder, and backslope). In chapter 3 we drew a similar conclusion for the organic matter and cation exchange capacity distribution within the landscape of this same study site (ACRE). Higher values of organic matter and cation exchange capacity were measured in lower landscape positions, while lower values occurred on higher and steeper landscape positions. Lower parts of the landscape have a

seasonal high-water table that is closer to the surface and receive overland flow, causing waterlogged conditions. The steeper and convex areas, due to the higher slope, shed water to the lower parts of the landscape.

The differences between the DSM and SSURGO maps are obvious in predicting the extent of VPD and PD soils (Fig. 4.7 & 4.8). SSURGO, when compared to the DSM maps, underpredicted the extent of VPD areas and overpredicted the extent of PD soils regions. Both methods, however, showed no great differences in the extent of SWP and MW soil regions. As mentioned in Section 4.2, VPD and PD soils have similar morphological characteristics and are found in similar landscape settings thus, it may be hard to differentiate between these drainage classes using terrain attributes alone. On the other hand, SWP and MW are relatively easy to differentiate on the landscape with SWP found at midslopes positions, while MW is found at upper slope positions.

4.3.4 Soil Drainage Class Probability Map

Chang and Burrough (1987) point out that more than one soil class can be found for certain landscape combinations, thus mapping uncertainty is necessary. The availability of an uncertainty map is important for determining the reliability of the predicted maps. These maps can also provide tools to soil surveyors to map the overlapping drainage classes or highly uncertain areas more efficiently.

Both MNLR and ANN algorithms showed similar patterns of uncertainties (Fig. 4.9 & 4.10). Generally, uncertainty maps following the theoretical positions of soil drainage classes on the landscape. For instance, VPD probability maps show lower uncertainty in low lying landscape positions and higher uncertainty in higher landscape positions. Similarly, MW probability maps show lower uncertainty in upper landscapes and higher uncertainty in depression areas. Higher uncertainty is observed in drainage classes that are located in between two other drainage classes.

For instance, PD and SWP are more uncertain than VPD and MW. These uncertainty maps show that the selected environmental variables are not sufficient to differentiate between closely occurrence drainage classes, particularly between VPD and PD soils.

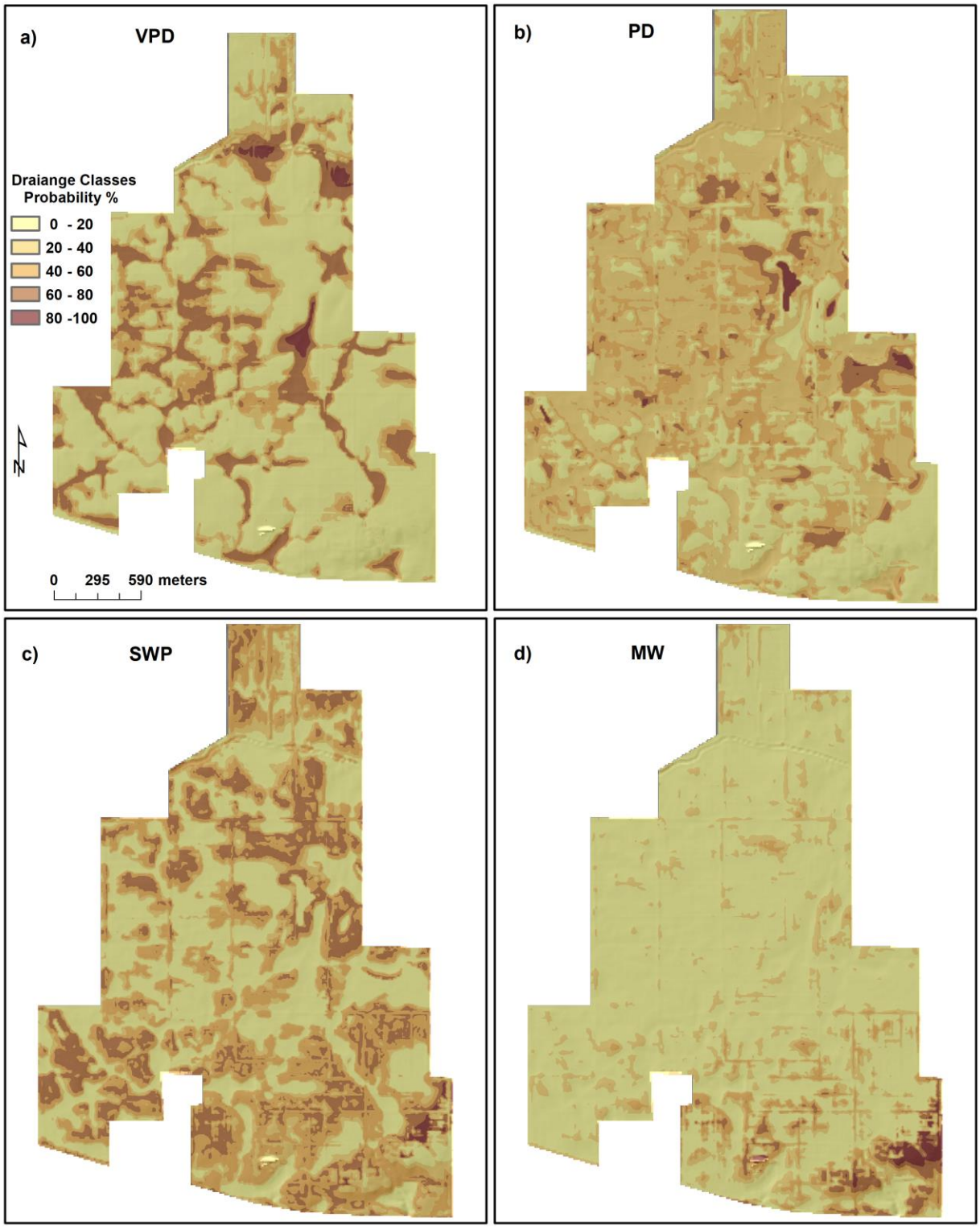

Figure 4.9: Maximum occurrence probability of soil drainage classes based on multinomial logistic regression. (a) Very poorly drained (VPD), (b) poorly drained (PD), (c) somewhat poorly drained (SWP), and moderately well drained (MW).

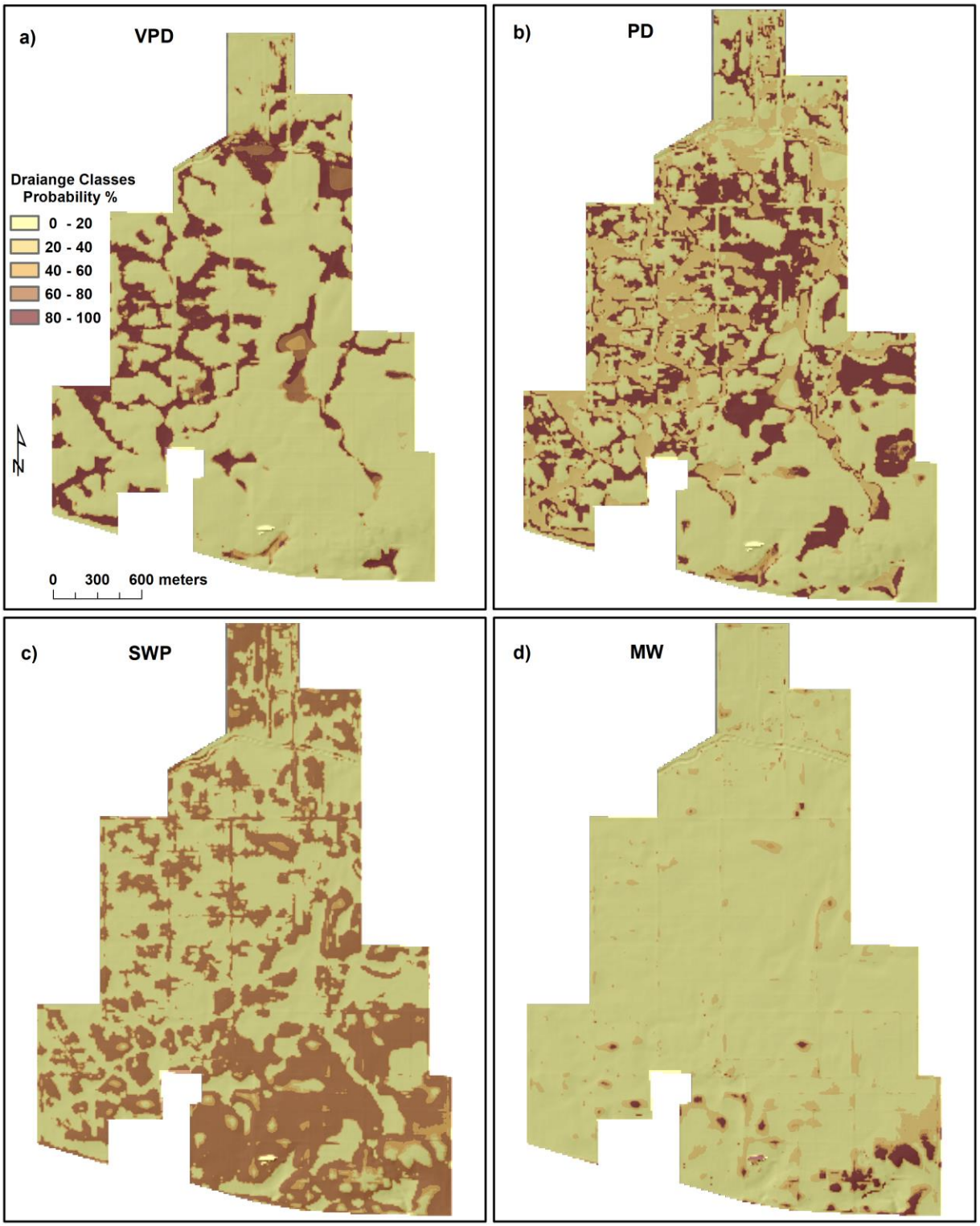

Figure 4.10: Maximum occurrence probability of soil drainage classes based on artificial neural network. (a) Very poorly drained (VPD), (b) poorly drained (PD), (c) somewhat poorly drained (SWP), and moderately well drained (MW).

4.3.5 Why Use Digital Soil Maps?

As described above, there is little difference between the prediction accuracy of DSM and conventional soil survey maps (SSURGO). The question may arise, 'why use DSM methodologies?' The answer to this question is provided in the following paragraphs.

Various studies have compared the soil class maps generated by DSM with conventional soil survey maps in terms of accuracy (Lorenzetti et al., 2015), cost and efficiency (Kempen et al., 2012; Zeraatpisheh et al., 2017), spatial correspondence and spatial details (Bazaglia Filho et al., 2013; Roecker et al., 2010). Overall, these studies concluded that DSM maps are more accurate, informative, detailed, and cost-efficient.

The three main limitations that are associated with conventional soil maps are the polygonal based product, the manual mapping process itself, and the lack of quantified documentation of the soil and landscape model. Due to scale issues (1:15,840), the SSURGO map provided information based on aggregated polygon map units (Soil Survey Staff, 2020). The smallest polygon in SSURGO is one hectare (Soil Science Division Staff, 2017). The SSURGO polygon may have one to four named components and include soils and non-soil areas, but all soils within the SSURGO polygon are considered homogenous. For instance, the CwB2 (Crosby-Miami silt loams) mapping unit of SSURGO in the study site has 64% Crosby with SWP, 33% Miami with MW, and 3% Treaty soils with PD conditions (Soil Survey Staff, 2020). However, SSURGO assigned the SWP class for CwB2 based on the dominant condition and/or component (Crosby). Additionally, soil variations occur along the boundary of polygons. Thus, the SSURGO maps provide simplified depiction of spatial variation of soils across the landscape. On the other hand, DSM methods provide continuous, consistent and potentially more realistic results. Additionally, DSM approaches (i.e. MNLR and ANN) are capable of providing probability maps that show the uncertainty associated with each allocated drainage class (Fig. 4.9 & 4.10).

In conventional maps (i.e. SSURGO) soil-landscape relationships are qualitatively documented with block diagrams and written descriptions of soil map units. In DSM however, soil-landscape relationships are documented in statistical models that use quantifiable digital inputs. Thus, decision criteria for DSM models can be easily documented and updated as more information becomes available over time. This has several advantages. First, the quantitative soil-landscape relationships from DSM is of use for further quantitative studies and models. Secondly, by changing the decision criteria or updating DSM models, new maps can be developed or updated. Because DSM models quantify the soil-landscape relationship, these updates and changes can be easily tracked and monitored.

4.4 Conclusions

This study predicted soil drainage classes based on field observations and terrain attributes generated from a lidar-based digital elevation model. MrVBF was the most important explanatory variable that was utilized by all models. According to the overall accuracy and kappa coefficient there is no major difference between MNLR, RF, C5.0, and ANN models. Artificial neural network (ANN), however, slightly outperformed MNLR, RF and C5.0 models. Multinomial logistic regression (MNLR) and C5.0 models, however, because of their interpretable forms, are preferred over RF and ANN models. Additionally, MNLR and C5.0 models are simpler and utilized fewer environmental covariates when compared with RF and ANN models. Furthermore, MNLR can predict both drainage classes and their associated uncertainties, while RF and C5.0 only predict soil drainage classes and ANN only provides the uncertainty maps. The MNLR and C5.0 models are recommended for future studies for areas like ACRE.

All DSM models showed slightly higher prediction performance when compared to the SSURGO data. Soil drainage classes predicted by both DSM and SSURGO correspond with the

soil and landscape model, meaning that VPD and PD were found in depressions and low-lying areas while, SWP and MW were predicted on higher and steeper landscapes. This study demonstrated that, on a farm scale that has similar characteristics as ACRE (i.e. gently undulating topography and glaciated landscape), the natural soil drainage classes can be adequately mapped using a high-resolution lidar DEM. For future studies, including covariates that capture underlying geology might improve the results and need to be considered.

CHAPTER 5. MAPPING SUBSURFACE TILE DRAINAGE LINES USING AERIAL PHOTO INTERPRETATION, PAPER MAPS, AND EXPERT KNOWELDGE

Abstract

Accurate maps of subsurface tile drainage lines are needed for agronomic and environmental research and the maintenance of current tile drainage systems. In this study, tile lines at the Agronomy Center for Research and Education (ACRE) were identified using a combination of visual aerial photo interpretation, expert knowledge, and paper construction drawings. The mapping accuracy was assessed using 27 points at which tile lines were located physically using a tile probe. Tile lines were correctly predicted 89% of the time with an average spatial accuracy of ±1.23 m of the true tile locations. This was better than a previous tile line locations map prepared by Naz and Bowling (2008) using an automated remote sensing method which had an average spatial accuracy of ±2.12 m.

5.1 Introduction

The Purdue University Agronomy Center for Research and Education (ACRE) has a long history of world-class research. More than 50 research scientists from various departments currently conduct research at ACRE (ACRE, 2020). Since the announcement of the Purdue Plant Sciences Initiative in 2013 (Robinson, 2013), ACRE has been transformed into a high-tech field phenotyping facility with a focus on collecting information at both the canopy and individual plant levels. Plants are monitored throughout the growing season using a combination of traditional in-field data collection, as well as with an array of different sensors mounted on unmanned aerial vehicles (UAVs) and on the PhenoRover, a ground-based mobile sensor platform (PU-IPS, 2020).

Most of the soils at ACRE are poorly and somewhat poorly drained (USDA-NRCS, 1998) and require subsurface drainage to remove excess water to provide better plant growth conditions. Subsurface tile drainage prevents crop drown out, minimizes soil erosion, and increases crop yields by preventing root damage caused by excess water and by providing better aeration (Fausey et al., 1987; Franzmeier et al., 2001). Tile drainage also allows farmers to access farmland to conduct timely farm operations (Franzmeier et al., 2001).

The presence of subsurface drainage tiles can greatly impact plant phenotypic respond through spatial redistribution of soil moisture, plant nutrients, soil pH, and rooting depth (Ritzema et al., 2008; Wang et al., 2006; Mathew et al., 2001; Rhoades et al., 1999). Therefore, accurate location of subsurface tile drainage lines is needed to support the above ground plant phenotyping research at ACRE.

Subsurface tile drainage systems are widely used in the Midwestern U.S. In 1985, ~12.5 million ha in the Midwest contained tile drainage (Pavelis, 1987). Since then, substantial additional areas have had subsurface drainage systems installed. Indiana, with approximately 50% artificially drained cropland, is the highest in the nation (Pavelis, 1987). Accurate maps of preexisting tile lines are not only important for agronomic and environmental research, but also for maintenance and repair of current drainage systems, and for reference during installation of new tile lines in previously tiled fields. In many cases, however, the locations of exiting tile lines are not known exactly because maps of their locations were not made or have been lost.

Due to the need to accurately locate subsurface tile lines, it is not surprising that much research has been conducted, and different approaches have been utilized, to identify existing tile lines. Geophysical and remote sensing are the two main methods of locating subsurface tile lines. Manual probing, trenching, and ground penetrating radar are common geophysical methods for

locating tile lines (Roy, 2014; Allred et al., 2018). Even though geophysical methods can accurately locate tile lines they are time consuming, labor intensive, expensive, and tedious, thus limiting their application for larger areas (Allred et al., 2004; Ale et al., 2007; Gökkaya et al., 2017). Ground penetrating radar does not work well for locating tile lines in high clay soils due to attenuation of the radar signal (Conyers and Goodman, 1997).

Various studies have shown that remote sensing, together with Geographic Information Systems (GIS), can be an effective approach to precisely mapping buried tile lines (Verma et al., 1996; Northcott et al., 2000; Varner et al., 2002; Naz and Bowling, 2008). Aerial imagery often captures spectral differences between wet and dry soils; thus it plays a critical role in locating tile lines using remote sensing. Two to three days after a heavy rain (25 mm within 24 hrs.), soil over tile lines often dries faster than soil in between tile lines. This results in higher reflectance of the drier soils in the visible and near infrared regions of the electromagnetic spectrum, and this difference can be captured by aerial imagery. Soil moisture, soil organic matter, soil texture, crop residue, and tillage practices, however, also affect the reflectance and, therefore, the accuracy of automated tile mapping using aerial imagery (Naz and Bowling, 2008; Naz et al., 2009, Andrade, 2013).

The objective of this research was to accurately map the locations of tile lines at ACRE and to develop a detailed attribute table with the type, material, status, and dimeter of the tile lines. To do so, we used a combination of visual photo interpretation, expert knowledge, and paper construction drawings. Additional objectives were to evaluate the new map against ground observations of tile line locations, to compare the new map to the map produced by Naz and Bowling (2008) using an automated method, to compare the results of this study with tile maps

generated by the tile installation companies, and to deliver this information to research scientists in a usable format.

5.2 Materials and Methods

5.2.1 Study Site

The Agronomy Center for Research and Education (ACRE) is located in Tippecanoe County, Indiana, USA (40° 28' 12" N, 86° 59' 31" W) (Fig. 5.1). ACRE was established in 1949 as a field research station for soils and crops research and currently consists of 570 hectares. ACRE is located on a low relief, gently undulating Wisconsin age till plain and contains fourteen different soil mapping units, with Chalmers (fine-silty, mixed, mesic Typic Hapludalfs), Raub (fine-silty, mixed, mesic Aquic Argiudolls), and Drummer (fine-silty, mixed, mesic Typic Haplaquolls) soils as the dominant soil types. Most of the soils at ACRE are poorly or somewhat poor drained, but a few locations are very poorly drained or moderately well drained (USDA-NRCS, 1998). Corn and soybean are the most extensive crops at ACRE. The average annual temperature is 10° C and the average total annual precipitation is 970 mm (30-year normals for 1981 to 2010) (MRCC, 2013). The mean winter temperature (December to February) is -2.6° C and the mean summer temperature (June to August) is 22.2° C (NWS-COOP, 2020).

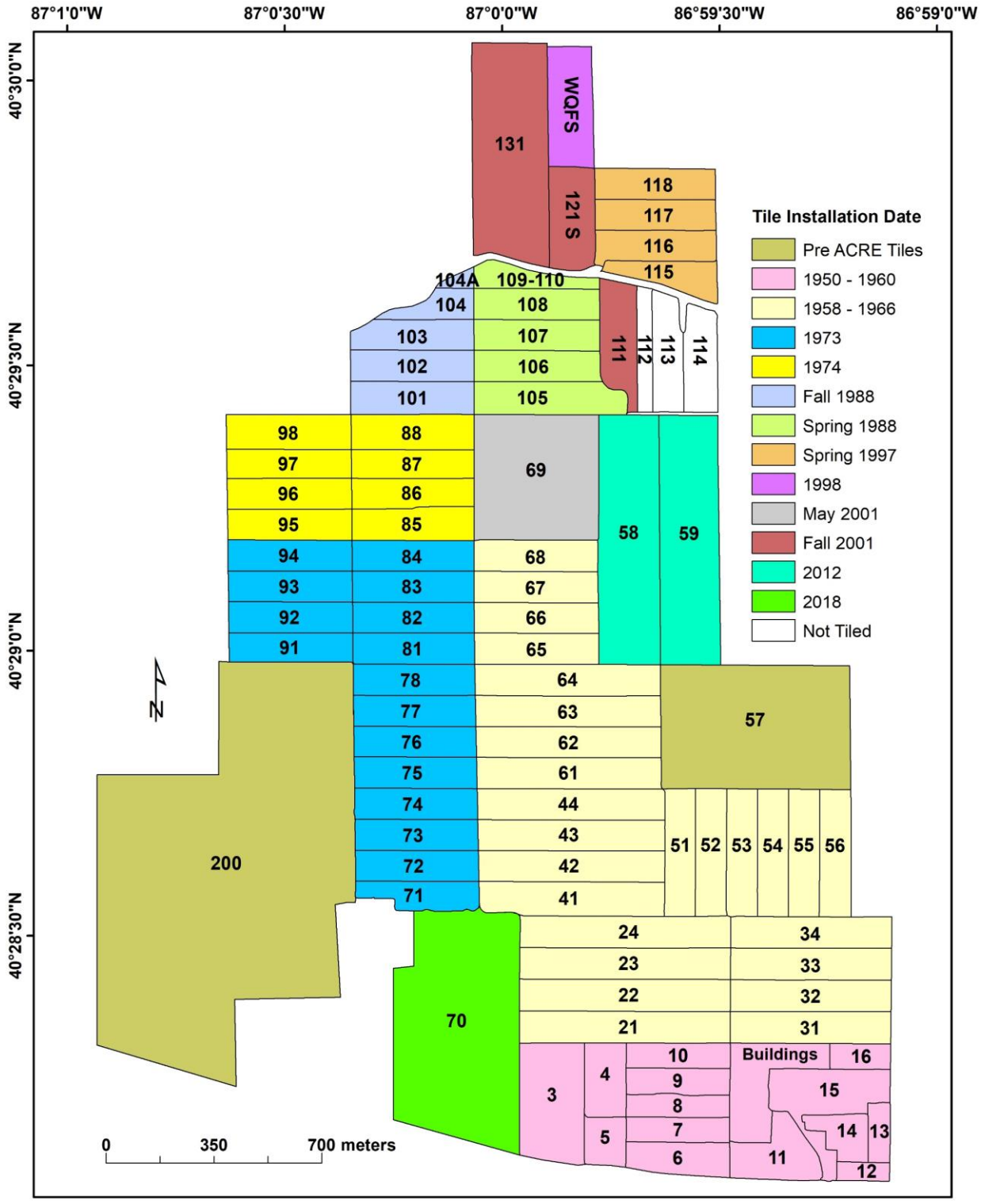

Figure 5.1: Map of the Purdue Agronomy Center for Research and Education (ACRE) showing the field boundaries and numbers and the dates of tile drainage installation up to spring 2020. Pre-ACRE = tiles installed prior to acquisition of the land for ACRE.

The tile lines at ACRE were typically installed at 0.9 m (3 ft) depth and on 20 m (66 ft) spacings. The lateral tile lines, which carry water from the fields to the larger main or sub-main tile lines, are typically 10 cm in diameter. The main tile lines vary from 15 cm to 60 cm in diameter. The tile lines were installed in different years (Fig. 5.1) and reflect the order of land acquisition and the availability of funds (J. Beaty, personal communication, April 5, 2018).

In the sections below, we provide information about input data, tile mapping procedure, and accuracy assessment. The main steps and methodological procedures of mapping tile lines are presented in Fig. 5.2.

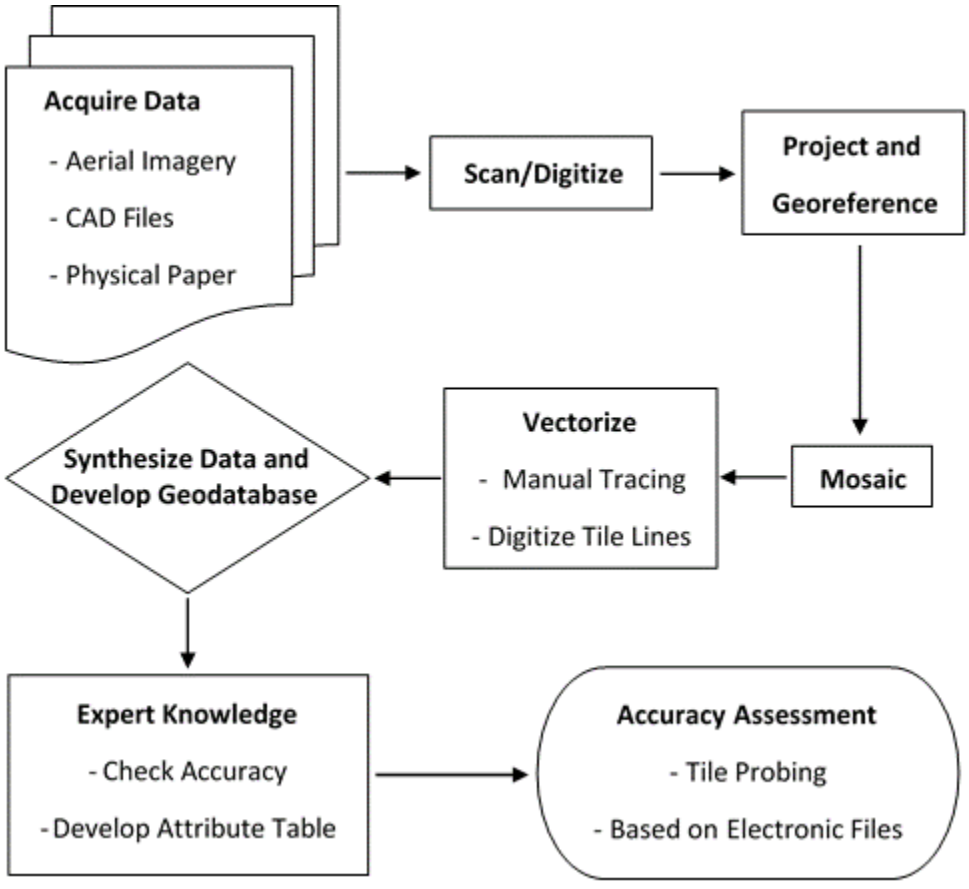

Figure 5.2: Methodological workflow and main steps in mapping subsurface tile lines at ACRE.

5.2.2 Input Data and Information

We used a combination of aerial imagery, paper and digital maps of planned or as-installed tile lines, and expert knowledge to locate the tile drainage lines at ACRE.

Aerial Imagery and Image Processing

Historic aerial imagery can provide information on the spatial location of tile lines, particularly when other sources of information such as installation drawings are not available. We obtained as much aerial imagery as we could find for the study site and ultimately assembled 24 images of the study area that spanned from 1939 to 2016 (Table 5.1). The 6 oldest datasets from 1939 through 1976 consisted of aerial photographs either downloaded from the Indiana Geological and Water Survey (IGWS) website (https://igws.indiana.edu/), or accessed as printed aerial photographs at the USDA Natural Resources Conservation Service, Lafayette Service Center, 1812 Troxel Dr., Suite C3, Lafayette, IN 47909. The imagery from 1939 and 1963 was available from IGWS already scanned, but not georeferenced. The aerial photographs from 1957, 1968, 1971, and 1976 accessed at the USDA Lafayette Service Center were scanned at 400 dpi (dots per inch) and stored as TIFFs (Tag Image File Format). The remaining 18 datasets spanning from 1998 through 2016 were downloaded from the Indiana Spatial Data Portal (ISDP) (https://gis.iu.edu) and were already georeferenced.

Table 5.1: Available aerial imagery to map tile lines at ACRE.

No	Original Dataset Name	Date	Resolution (m)	Original Datum and Projection	Source[*2]
1	1939 Aerial Imagery	04/13/1939	10	N/A	IGWS
2	1957 Aerial Imagery	09/04/1957	0.9	N/A	USDA
3	1963 Aerial Imagery	05/31/1963	5.5	N/A	IGWS
4	1968 Aerial Imagery	03/04/1968	0.3	N/A	USDA
5	1971 Aerial Imagery	06/17/1971	0.7	N/A	USDA
6	1976 Aerial Imagery	03/09/1976	1	N/A	USDA
7	1998-1999 USGS Digital Ortho Quarter-quad	1998-1999	1	NAD_1983_UTM_Zone_16N	ISDP
8	2003 National Agriculture Imagery Program	07/19/2003	1	NAD_1983_UTM_Zone_16N	ISDP
9	2004 National Agriculture Imagery Program	07/01/2004 08/15/2004	2	NAD_1983_UTM_Zone_16N	ISDP
10	2005 IndianaMap Color Infrared Photos	02/26/2005 05/29/2005	1	NAD_1983_UTM_Zone_16N	ISDP
11	2005 IndianaMap Natural Color Orthos: Orthophotography	March 2005 April 2005	0.15	NAD_1983_StatePlane_Indiana_West_FIPS_1302_Feet	ISDP
12	2005 IndianaMap Natural Color Orthos: Quarter quads	March 2005 April 2005	1	NAD_1983_UTM_Zone_16N	ISDP
13	2005 National Agriculture Imagery Program	07/01/2005 09/15/2005	2	NAD_1983_UTM_Zone_16N	ISDP
14	2006 IndianaMap Reflight Color Infrared	Spring 2006	1	NAD_1983_UTM_Zone_16N	ISDP
15	2006 IndianaMap Reflight Natural Color Orthophotography	Spring 2006	0.15	NAD_1983_StatePlane_Indiana_West_FIPS_1302_Feet	ISDP
16	2006 IndianaMap Reflight Natural Color Quarter-quads	Spring 2006	1	NAD_1983_UTM_Zone_16N	ISDP
17	2006 National Agriculture Imagery Program	07/06/2006 08/16/2006	2	NAD_1983_UTM_Zone_16N	ISDP
18	2007 National Agriculture Imagery Program	07/02/2007 08/13/2007	2	NAD_1983_UTM_Zone_16N	ISDP

Table 5.1: Continued

No	Original Dataset Name	Date	Resolution (m)	Original Datum and Projection	Source[*2]
19	2008 National Agriculture Imagery Program	06/24/2008 09/01/2008	1	NAD_1983_UTM_Zone_16N	ISDP
20	2010 National Agriculture Imagery Program	08/16/2010	1	NAD_1983_UTM_Zone_16N	ISDP
21	2012 National Agriculture Imagery Program	06/06/2012 06/19/2012	1	NAD_1983_UTM_Zone_16N	ISDP
22	2013 IndianaMap Data[*1]	02/14/2013 04/22/2013	0.3	NAD_1983_StatePlane_Indiana_West_FIPS_1302_Feet	ISDP
23	2014 NAIP Imagery	2014	1	NAD_1983_UTM_Zone_16N	ISDP
24	2016 NAIP Imagery	06/12/2016	0.6	NAD_1983_UTM_Zone_16N	ISDP

[*1] the 2013 imagery was considered the master image and all of the aerial imagery was georeferenced based on this master image.
[*2] IGWS = Indiana Geological and Water Survey, https://igws.indiana.edu; USDA = USDA Natural Resources Conservation Service, Lafayette Service Center, 1812 Troxel Dr., Suite C3, Lafayette, IN 47909, and ISDP = Indiana Spatial Data Portal, https://gis.iu.edu.

The 18 datasets that were already georeferenced were in a variety of projections (Table 5.1). As a common projection, we chose a projection from the Indiana Geospatial Coordinate System (InGCS) (INDOT, 2016). The InGCS is a set of low-distortion map projections that minimizing the horizontal linear (grid vs. ground) distortion across the design region, typically a 1-, 2-, or 3-county area within Indiana, and are defined in units of both meters and feet. The average grid vs. ground difference in InGCS is 0.014 feet per mile (±2.6 ppm). The average, grid vs. ground difference for other commonly used projections is considerably larger. For example, for the Indiana State Plane Coordinate System it is 0.42 feet per mile (±80 ppm), while for the Universal Transverse Mercator (UTM) Zone 16 North system it is 2.1 feet per mile (±400 ppm) (INDOT, 2016). Since ACRE is entirely in Tippecanoe County, IN, we used the (InGCS) for Tippecanoe and White Counties, which has an average grid vs. ground difference of 0.0159 feet per mile (3 ppm). In ArcGIS 10.6 software, this projection is listed as "NAD_1983_2011_InGCS_Tippecanoe-White_(m)" for units in meters. The Project Raster tool was used in ArcMap 10.6 to project all of the georeferenced images into "NAD_1983_2011_InGCS_Tippecanoe-White_(m)".

Of the available imagery, the 2013 IndianaMap Data image at 0.3 m resolution is georeferenced to the highest standard for use as a basemap for state-wide, county, and municipal projects. It was selected as the master image for georeferencing images that were obtained from USDA and IGWS. The Georeferencing tool in ArcMap 10.6 was used to georeference the USDA and IGWS images based on image-to-image registration.

For accurate georeferencing we used stable and visible benchmarks such as roads, intersections, driveways, railroad tracks, culverts, etc. as control points. Sufficient control points were added so that the spline transformation could be used. In general, about 15 control points

were used for each image. After georeferencing, the Mosaic tool was used to mosaic all the aerial photos from a single year into a seamless image that covers the entire study location.

Physical and Electronic Maps

Paper maps of tile line locations and the expert knowledge of the farm manager were also used to identify the tile lines at ACRE. Generally, these paper maps were large engineering drawings of planned tile lines made prior to installation. These maps were not scanned or digitized. The information in these maps, however, were used to map tile lines for parts of the farm where other data (e.g. aerial imagery and electronic maps) could not be used. The paper maps were also useful for determining which main or sub-main a particular lateral tile line was flowing into, and for developing the attribute table of tile line sizes and types.

For the Water Quality Field Station (WQFS) on the north end of the farm (Fig. 5.1), we have received a blueprint in pdf format. The quality of the blueprint was enhanced in Adobe Photoshop and saved as a TIFF file, which was then imported into ArcMap 10.6 and georeferenced to the 2013 master image.

We also acquired four electronic maps. The tile line map developed by Naz and Bowling (2008) was available as a shapefile. The Project tool in ArcMap 10.6 was used to re-project this shapefile from the NAD_1983_UTM_Zone_16N projection to the NAD_1983_2011_InGCS_Tippecanoe-White_(m) coordinate system. The as-installed maps of the tile lines for fields 58-59 and 70 (Fig. 5.1) were provided by the tile installation company (Schlatter's Inc, 16179 W 500 S, Francesville, IN 47946). These as-installed maps were generated using a Real-Time Kinematic (RTK) Global Navigation Satellite System (GNSS). The Project tool in ArcMap 10.6 was used to re-project these maps from GCS_WGS_1984 into the NAD_1983_2011_InGCS_Tippecanoe-White_(m) coordinate system. Finally, a computer aid

design (CAD) map of fields 3-16, 21, and 31 was provided by the Purdue University Physical Plant. After selecting the tile lines from the CAD file, the Export Data tool in ArcMap 10.6 was used to export them as a shapefile and the Define Projection tool was used to georeference the extracted shapefile to the NAD_1983_2011_InGCS_Tippecanoe-White_(m) coordinate system.

5.2.3 Mapping Tile Lines

While the use of automated and semi-automated remote sensing techniques for identifying tile lines would be attractive, we opted to develop the tile lines map using manual photo interpretation and manual digitization primarily because for a relatively small area such as ACRE (570 ha), development of an automated procedure would likely take longer than manual digitization. Automated procedures are likely to be impacted by the presence of other linear features in the field that are not tile lines and these features may be mapped as tile lines, and an automated model might work for one set of the images, but not for a different one.

The best aerial imagery to use for tile delineation needs to be taken 2 to 3 days after heavy rain (25 mm or greater within 24 hours) (Verma et al., 1996; Northcott et al., 2000; Varner et al., 2002). In the Midwestern US, April to late May are the best times to clearly see the tile patterns with minimal crop residue and crop canopy (ISUST-GISSRF, 2017). Most of the color aerial photography that we acquired for this study was taken well into the growing season and could not be used for tile delineation. Of the imagery we assembled, the imagery from 1963, 1976, 1998, 2012, and 2013 showed the locations of tile lines to varying degrees, and of these, the imagery from 1963 and 1976 was used most extensively for tile line delineation. All five aerial images have 1 m pixel resolution, except for 1963 which has 5.5 m pixel resolution, and 2013 which has 0.3 m pixel resolution. We could not determine the exact acquisition dates for the 1998, 2012, and 2013 imagery, but the older imagery contained explicit date stamps which allowed us to determine that

the 1963 imagery was acquired on May 31st and the 1976 imagery was acquired on March 9th. Table 5.2 shows the precipitation data for the two weeks prior to the acquisition of these two images. In 1963, one day of drying after two days of rainfall totaling 8 mm was sufficient to produce slightly drier soil with higher reflectance over the tile lines than between the tile lines. In 1976, three days of rain totaling 49 mm thoroughly wetted the soil and probably resulted in surface crusting, but after 2 days of drainage and drying, the soils over the tile lines had dried sufficiently that the surface was considerably more reflective than the still-wet soil between the tile lines.

Table 5.2: Daily precipitation for the two weeks prior to the acquisition of the 1963 and 1976 aerial imagery. Source: (MRCC, 2013).

	Date	**Precipitation[*1] (mm)**	**Date**	**Precipitation (mm)**
	05-17-1963	1.27	02-24-1976	0.00
	05-18-1963	4.57	02-25-1976	0.00
	05-19-1963	0.00	02-26-1976	0.00
	05-20-1963	3.30	02-27-1976	0.00
	05-21-1963	T[*2]	02-28-1976	0.00
	05-22-1963	0.00	02-29-1976	0.00
	05-23-1963	0.00	03-01-1976	0.00
	05-24-1963	0.00	03-02-1976	0.00
	05-25-1963	0.00	03-03-1976	23.11
	05-26-1963	0.00	03-04-1976	10.67
	05-27-1963	0.00	03-05-1976	15.49
	05-28-1963	4.57	03-06-1976	0.00
	05-29-1963	3.81	03-07-1976	0.00
	05-30-1963	T	03-08-1976	0.00
Image Acquisition Date	**05-31-1963**	0.00	**03-09-1976**	0.00

[*1] Precipitation is for rainfall only. No snow was recorded during these intervals.
[*2] T: Trace

The tile drainage network was manually drawn through tracing and heads-up digitizing on the aerial imagery using ArcMap 10.6. The tile lines were interpreted to exist in places where both

black and white and color aerial imagery clearly followed straight lines. Most of the tile lines on black and white aerial images were identified based on the spectral differences of light and dark colors of dry and moist soils, respectively (Fig. 5.3-1a). While, tile lines on the 2013 color aerial imagery were identified based on disturbed soils from tile trenches and installation (Fig. 5.3-2a). Due to the disturbance of topsoils during tile installation, the locations of tile lines are more distinct as compare to other features. For WQFS, similar to the aerial imagery, the tile lines were manually traced but based on the georeferenced blueprints.

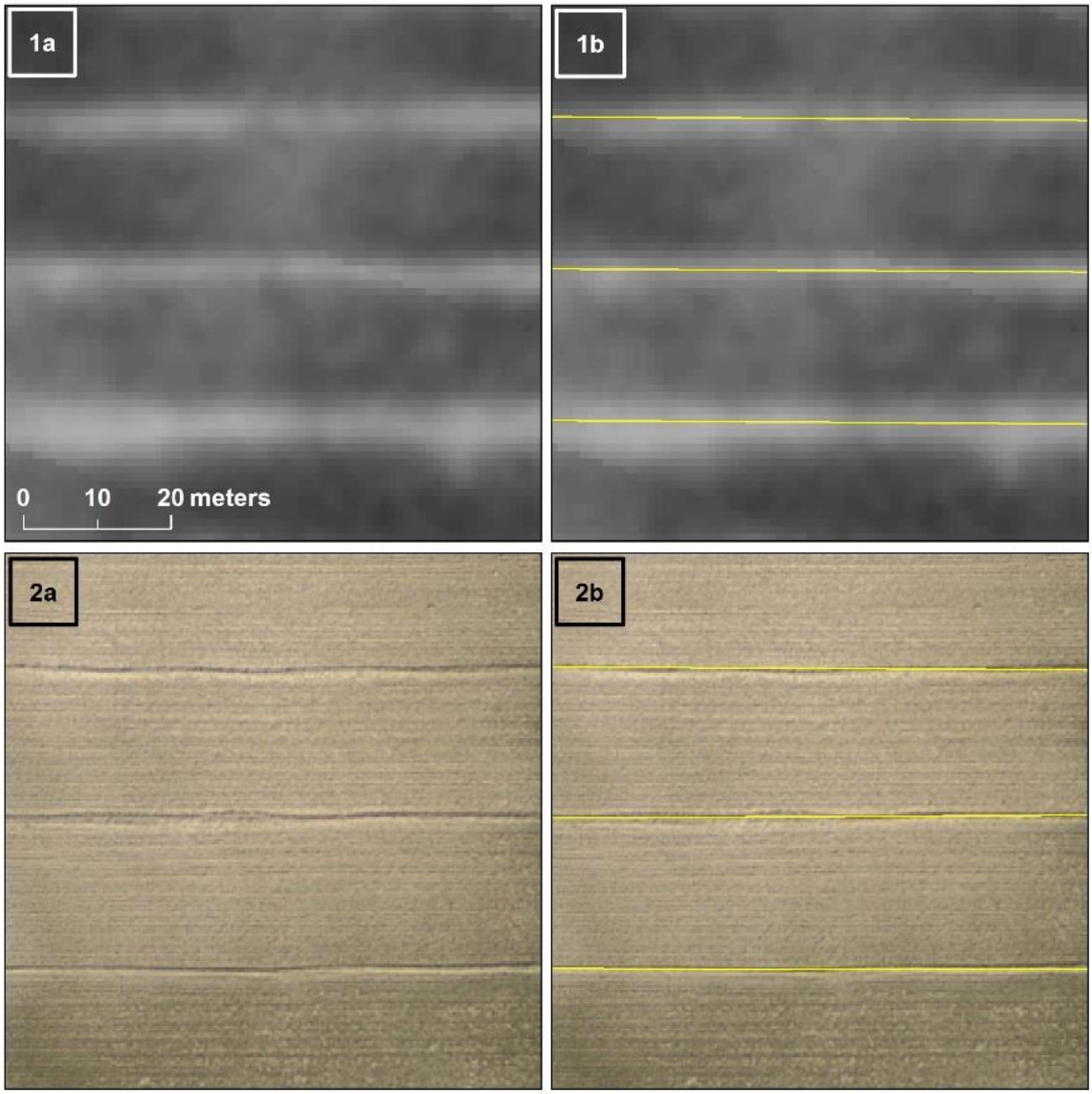

Figure 5.3: Identifying tile lines based on spectral differences and disturbed soils. The black and white images show the tile lines based on the spectral difference of light and dark colors due to the dry and moist soil condition. While, on the color image, tile lines were identified based on the disturb surface soil due to the tile installation. (a) Before locating tile lines (b) after locating tile lines. For actual locations of these images see Fig. 5.6.

As indicated above, where the locations of the tile lines could not be discerned from the aerial imagery, we relied on the expert knowledge of the farm manager and available paper maps. Large scale (1:3,500) draft paper maps were printed for review by the farm manager. Feedback from the farm manager and examination of the paper maps was also helpful in distinguishing tile

lines from other linear features such as tillage paths, surface drainage patterns, crop residue, grassed ways, and field dividers. In a few data poor locations, the farm manager was able to draw several tile lines on the paper maps based on the relative distance from a known or mapped tile or a field boundary. For instance, most of the lateral tile lines are placed at a fixed interval of 20.12 m or 66 ft. In some cases, the farm manager recognized that tile lines that were visible on older aerial photography were no longer active (i.e. under buildings) and these were removed from the map as well.

We created a geodatabase to assemble all acquired and generated tile drainage shapefiles for a complete tile drainage network of ACRE. We also developed an attribute table for the mapped tile lines. This attribute table was developed using original paper maps and the farm manager's knowledge. The attribute table provides information about type, material, status, and diameter of the tile lines. For a final approval, the developed tile map and its associated attribute table were once more checked by the farm manager.

5.2.4 Accuracy Assessment

After all tile lines were manually digitized, they were evaluated for accuracy. We used two different approaches to evaluate the accuracy of the mapped tile lines: (1) manually locating the tile lines at selected locations and, (2) comparison to the as-installed tile locations as provided by the installer.

For the first approach, we used a tile probe to locate tile lines in the field. A tile probe is a stainless steel rod that has a tee handle at one end and a pointed tip at other end (Fig. 5.4). Generally, tile probes are 1.2 m long and used for locating buried pipes, tiles, tanks, and utility lines. We went to the field without a preplanned design for ground truthing and randomly selected 27 locations

for field verification. Fields 41 – 44, however, were used for plant phenotyping research at the time of our evaluation, therefore we collected most of our *in situ* measurements in these fields.

Figure 5.4: Tile probe and investigating the location of a tile line based on a specific probing interval (~7 cm).

Of the 27 field locations we investigated, 24 of them were in areas where the tile lines had been digitized manually, two of them were in a field (field 70) that had as-built tile line locations from the installation contractor, and one point was in a field (field 58 – 59) that had both as-built and manually digitized maps. Generally, the ground validation was conducted close to the edges of the fields for ease of access and efficiency. Abandoned tile lines or tile lines installed prior to the establishment of ACRE were not validated at the field. Only a limited number of sites were manually assessed. First, it is difficult and tedious to distinguish tile lines from subsurface rocks, particularly in soils formed in glacial till and outwash as those at ACRE. Second, tile probes can easily enter corrugated, perforated plastic pipes, making it difficult to confirm the locations of these plastic tiles. Third, tile probing can cause corrugated plastic pipe to collapse, causing the tile line to cease to function as it should. Allred et al., (2018) noted similar problems associated with the use of a tile probe.

In order to take the tile line maps to the field, the base map and tile line locations were loaded into the Soil Explorer app for Apple iPad available in the Apple App Store. The Create Map Tile Package tool in ArcGIS 10.6 (https://esri.com) was used to prepare tile packages which were then loaded into the Soil Explorer app. When in the field, a dot shows the user's location on the map using the internal global positioning system (GPS) receiver in the iPad. This allowed us to determine the location of the tile line within the accuracy of the iPad GPS receiver, which is about ±5 m. Since the minimum diameter of a tile line is about 10 cm (4 in), the ground was probed with the tile probe at about 7 cm (3 in) intervals perpendicular to the axis as shown on the map on the iPad. The resistance increases when tile probe encounters a tile line, particularly a concrete tile. In addition to the resistance, the probe will also generate a sound when the tip hits a hard object. One can also feel when the probe tip penetrates into the plastic pipe. After locating what appeared

to be the tile line, we probed along what should be its axis to confirm its identification (Fig. 5.5a). Once the location of a tile line was confirmed by two field experts, the coordinates of the tile were recorded using a Trimble AgGPS 542 Real-Time Kinematic (RTK) base Global Navigation Satellite System (GNSS) receiver (Fig. 5.5b) accurate to ±0.8 cm horizontal and ±1.5 cm vertical.

Figure 5.5: Probing to identify the locations of a tile line in the field (a) Once the first probe line located what appeared to be the tile line, the second probe line was used to confirm the identification. (b) Recording the confirmed location of a tile line with an RTK GNSS receiver.

In the second approach, for field 58 – 59, we used the as-installed tile map generated by the tile installation company using an RTK GNSS system. In ArcMap 10.6, the as-installed tile lines were overlaid on the tile map produced by photo interpretation. The distance between 46 tile lines of the two methods were measured using the ArcMap 10.6 distance measure tool to determine how close the locations of the photo interpreted tile lines agreed with the as-installed tile line locations.

5.3 Results and Discussion

5.3.1 Accuracy Assessment Based on Tile Probing

The final drainage tile map is shown in Fig. 5.6. Using ground or *in situ* validation, our goal was to answer the following three questions. First, what is the overall prediction accuracy of the identified tile lines? In other words, what percent of the mapped tile lines are identified by ground validation? Second, how close are the predicted locations of the tile lines to their actual, ground validated locations? Third, how useful is the integration of expert knowledge and physical paper maps for predicting accurate tile lines?

The overall prediction accuracy was measured using the following equation:

$$\text{Detection or overall prediction accuracy} = \frac{\text{Number of detected tiles with tile probe}}{\text{Total existing tiles or investigated locations}} * 100 \qquad [1]$$

The above equation shows the percentage of predicted tile drains that were located by *in situ* detection. Out of the 27 locations that were investigated, tile lines were detected at 24 locations (Fig. 5.6), giving an overall predicted accuracy of 89%. One of the three undetected tile lines was in a field containing subsurface gravel and rocks, making it difficult to unambiguously distinguish and identify the tile line at this location. The other two locations were in field 70 where the tile lines were installed in 2018 and from which we received original, as-built shapefiles from the tile installer. This meant that according to the as-built map, these two tiles should be present, but we were unable to detect them using a field tile probe. We did not have a more recent image than 2016 to know exactly whether these two tiles were installed in 2018. It is very likely that these two undetected tiles are deeper than 1.2 m and therefore out of reach of the tile probe. Two other locations were investigated in the same field (field 70) and these tiles were detected with tile probe.

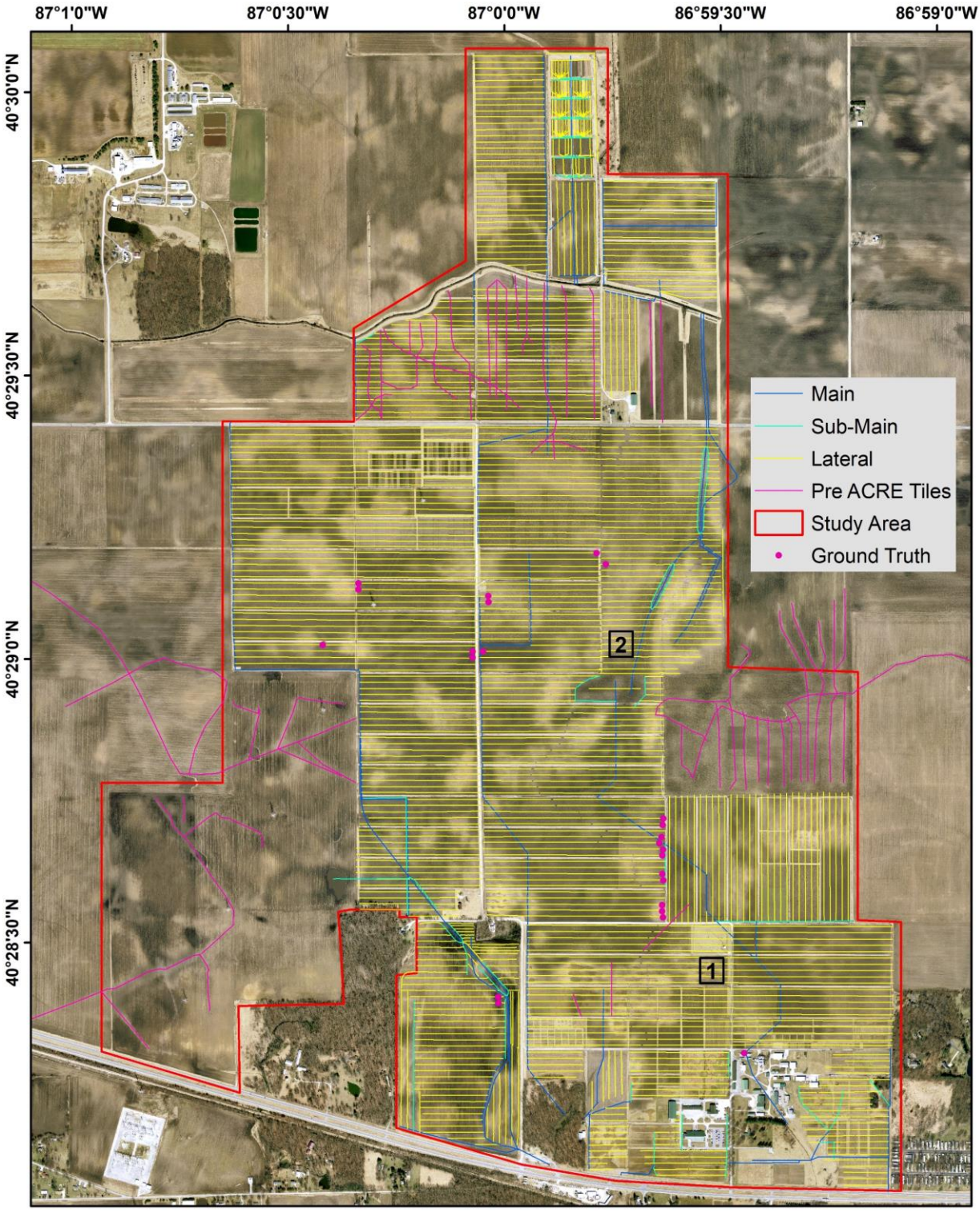

Figure 5.6: Final tile line map for ACRE. The red outline shows the study area with a few meters of buffer around the edge so that details near the edges are visible.

For the 24 locations where tiles were confirmed, Fig. 5.7 shows the perpendicular distance between the axis of the tile line as predicted by our map and the actual location determined with the tile probe and recorded by the RTK GNSS. On average, tile lines were predicted within ±1.23 m spatial accuracy. One third of the mapped tiles were estimated within ±0.5 m of the actual tile locations. For 21 tile probe locations we also evaluated the spatial prediction accuracy of the tile line map developed by Naz and Bowling (2008) (Fig. 5.8). We found that on average, the tile lines base on their automated method were estimated within ±2.12 m and one-fourth of the tiles have greater than ±2.5 m spatial prediction accuracy. Using an automated tile identification technique and relying on three aerial images from 1976, 1998, and 2002, and not using field expert knowledge might be the reasons for lower accuracy of the Naz and Bowling (2008) study.

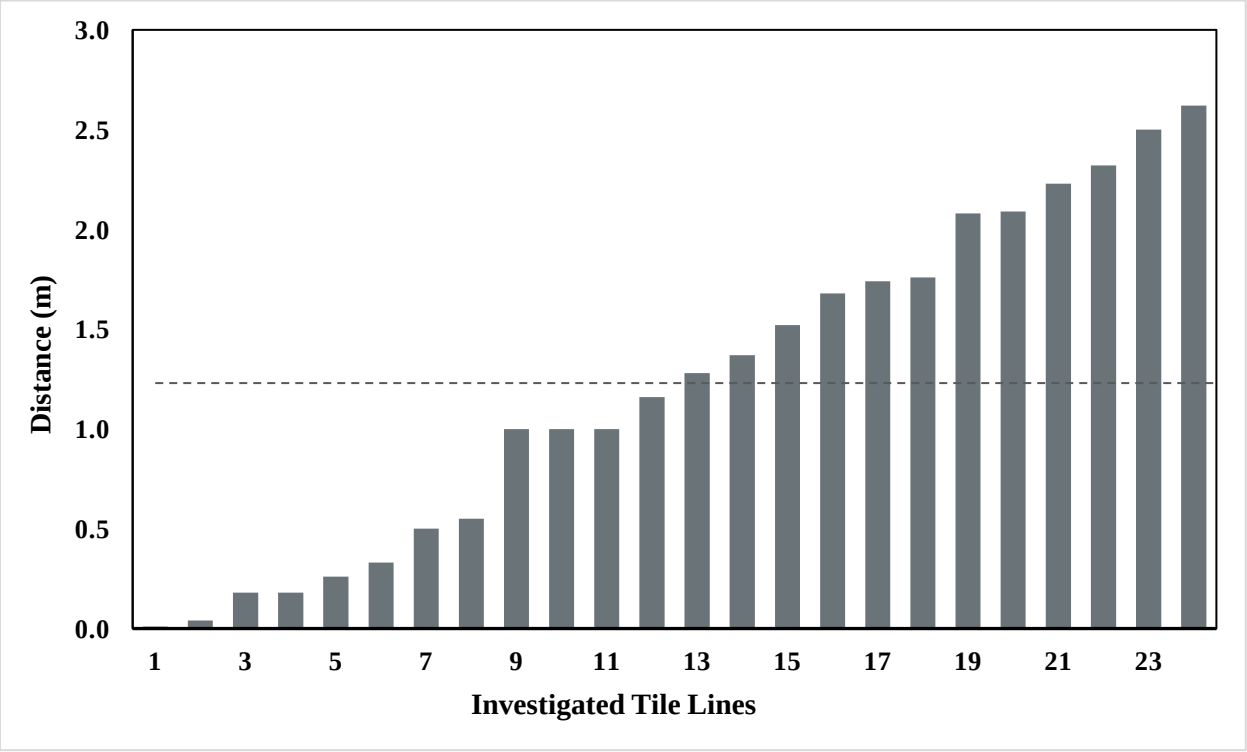

Figure 5.7: Prediction accuracy of the mapped tile lines based on tile probing in the field. The dashed line represents the average tile prediction accuracy of ±1.23 m.

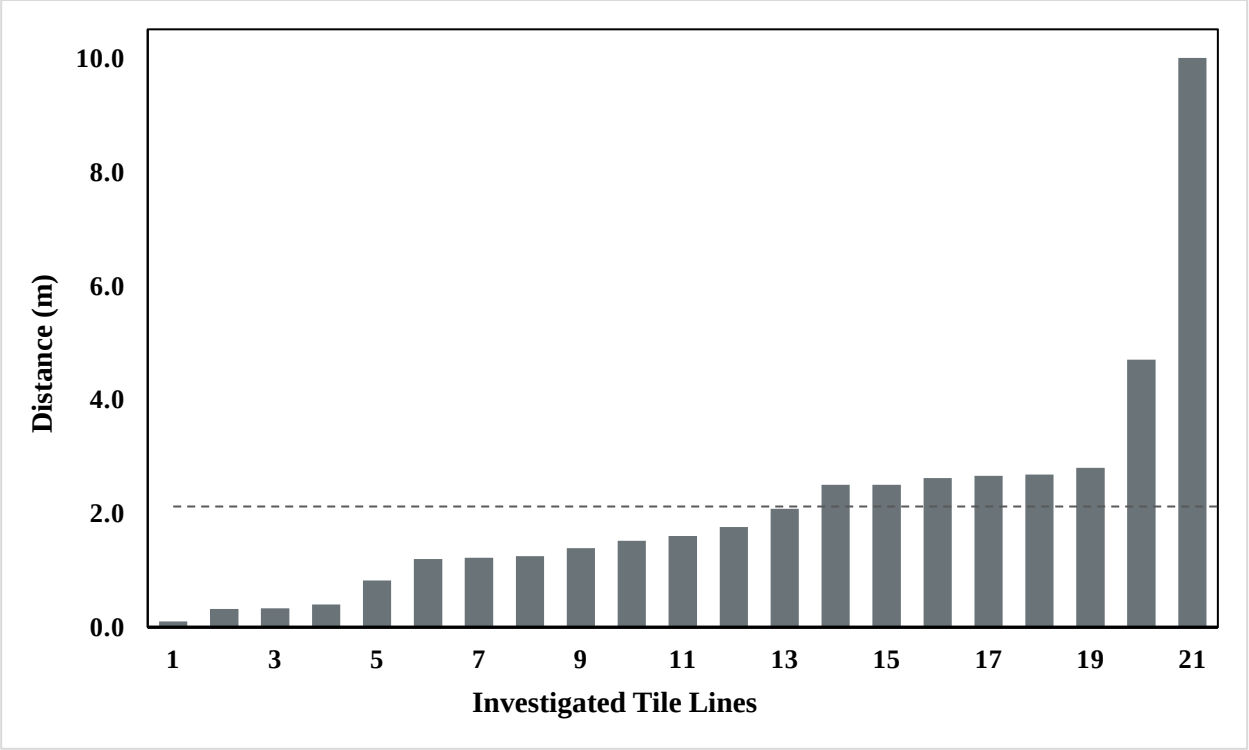

Figure 5.8: Spatial prediction accuracy of tile lines as mapped by Naz and Bowling (2008) based on tile probe location. The dashed line shows the average tile prediction accuracy.

5.3.2 Accuracy Assessment Based on As-Installed Maps

Out of the 48 tile lines in field 58-59 that were mapped during installation by the tile installation company, we identified 43 of them based on the disturbed soil visible in the 2013 aerial image. According to the overall accuracy equation (1), this will result in almost 90% overall prediction accuracy. The five undetected tiles are sub-mains and they are within 10 m of main tile lines. Compared to lateral tile lines, these sub-main and main tiles are located at deeper depths where water accumulation is higher and covers the disturb soil, thus making it harder to identify tile lines. According to the as-installed tile maps, the tile lines identified by disturbed soil on the 2013 image were predicted within ±1.02 m spatial accuracy.

5.3.3 Locating Tile Lines based on Expert Knowledge and Physical Paper Maps

As mentioned in section 3.3, the farm manager indicated that most of the tile lines were installed at a fixed spacing of 20.12 m (66 ft). In addition, most of the fields at ACRE are the same width and tiles from one field are located exactly across from the tiles of an adjacent field. These criteria were useful in predicting tiles lines when the locations could not be identified on the available aerial images. The distance between tile probe locations were measured in ArcMap 10.6 with distance measure tool. Based on 15 observations, it was confirmed that the spacing intervals between tile lines are 20.12 m or 66 ft.

5.3.4 Manual Digitization of Tile Mapping

The results from this study clearly demonstrate the utility of combing visual interpretation, expert knowledge, and physical paper maps data as an effective approach to accurately predicting tile line locations. This approach also rescues the expert knowledge and paper maps that are at risk of being lost or forgotten once the current farm manager retires. The finding of this study is in line with the previous study of Andrade (2013) that photo interpretation is a useful method to map unknown tile lines and provides better results than remote sensing.

5.4 Recommendations and Future Work

This study clearly shows the importance of quality aerial imagery, expert knowledge, and traditional tile information for predicting tile drains. However, availability and the accessibility of this data is a big challenge. If available, such data needs to be converted into a digital format for future use.

When mapping tile lines based on aerial imagery, acquire all the available aerial imagery and do not rely on data from only one year. However, finding quality images is difficult. To

overcome the complication of finding the right image for future studies, use one of the following reliable approaches for detecting tile lines. Acquire image(s) right after the tiles are installed to clearly show the disturbed soils from tile line installation. Conduct thermal infrared UAV surveys throughout the year to detect tile lines based on the lower soil heat capacity resulting from the reduction of moisture over tile lines (Allred and Rouse, 2018). This method, however, is costly and affected by tillage practices (Woo et al., 2009). Acquiring crop growth images and yield maps are other useful methods to locate tile lines because most of the time crops over tile lines have higher growth and yields in both wet and dry years (Ruark et al., 2009). Since tile drainage provides better conditions for plant emergence and early growth, it is expected that crops over tile lines will be clearly seen if higher resolution plant images are obtained early in the growing season.

There are less scientific methods to locate tile lines. For instance, air vent and surface inlets and outlets are features associated with tile drainage and are useful in detecting tile lines. Former landowner or local governmental agencies such as the Natural Resources Conservation Service (NRCS) might have tile maps or other essential information. However, this information needs to be ground validated.

Utilizing a combination approach will results in a higher degree of success in detecting subsurface tile lines. Before locating tile lines, it is important to identify tillage and harvesting patterns and field dividers, grassed waterways, and other features so that they can be avoided when mapping tile lines. After locating tile drains, it is important to generate accurate tile maps and keep copies in a secure file system. Modifications to the current drainage network and installation of new tile drains should be clearly documented and identified on the generated maps. The product of any tile detecting approach, particularly a remote sensing approach, should be matched with available paper maps, checked by experts, and ground validated.

For optimum utilization, tile drainage maps should be delivered in an easy to use format. The outcome of this study will be delivered through the Soil Explorer iOS app. We will also deposit the tile line map and other related information in the Purdue University Research Repository (PURR).

5.5 Conclusions

Locating buried tile drainage lines is important for incorporating the impact of these features on agronomic and environmental research. A combination approach of using aerial imagery, expert knowledge, and physical paper data was utilized to manually locate tile lines using ArcMap 10.6. A wealth of useful information about landforms, human influences, vegetation, and soils can be obtained simply by visual examination and interpretation of aerial imagery. Among the acquired aerial photographs, the 1976 and 2013 images provided useful information in this study, with the 1976 image being the most useful for identifying tile line locations. Tile lines at data poor sites were determined based on the ACRE farm manger's expert knowledge and original paper copies. This mapping approach resulted in ±1.23 m spatial accuracy. The results from this study are comparable to other studies (Naz and Bowling, 2008; Thompson. 2010). This approach not only accurately located buried agriculture tile lines, but also captured the expert knowledge and legacy data that otherwise was at risk of being lost. This method is efficient for use in a relatively small areas, but for larger soil regions (i.e. multiple-county level) it is likely to be too time consuming.

CHAPTER 6. UTILIZATION AND DELIVERY OF SPATIALLY EXPLICIT DIGITAL SOIL INFORMATION

The most important step after generating soil maps is to deliver the information to stakeholders for use. With today's technology it is possible to deliver soils information through a variety of user-friendly platforms. These technologies can deliver soils information as easy to use maps that can be zoomed, panned, and queried.

6.1 Using Digital Soil Maps

With current technology and on-the-go soil sensors, it is possible to precisely apply different agriculture inputs such as seed, fertilizer, pesticides, and irrigation to different sections of a field in response to different soil types or other variables. Therefore, detailed soil and/or yield maps are needed to successfully implement site specific management decisions regarding crop input applications. Soil and tile line maps show the location of properties that impact crop growth and yield and will be useful tools for analyzing plant phenotypic characteristics and for defining management zones and input decisions.

Most of the time, yield maps are correlated with soil maps and visually, both follow similar patterns (Adamchuk and Jasa, 2002; Georgi et al., 2018; Vallentin et al., 2019). This correlation is due to the variation in soil properties. The correlation of yield and soil maps often strongly depends on the amount of soil moisture available during the growing season. A side-by-side visual comparison of maps of deviation from the mean yield for two soybean research fields at ACRE and the soil and tile line maps described in previous chapters of this thesis (Fig. 6.1 and Fig. 6.2) show that the deviation from mean yield follows the soil property maps. The deviation from the

mean yield maps also show linear variations that appear to be due to the underlying tile lines (Fig. 6.1 b and Fig. 6.2 b).

In 2013, the higher yielding areas were generally located in lower landscape positions (i.e. yellow colors on the left side of field 43, Fig 6.1a), and lower yielding areas were usually found on higher landscape positions (i.e. red colors in the upper right corner and left side of field 44, Fig 6.1a). Similar patterns between yield and soil maps were observed in fields 63 and 64 in 2014 (Fig 6.2). The yield is generally higher in depression areas (i.e. yellow colors in the central part of field 63, Fig 6.2 a) with poorly drained soils with high organic matter content, as compared to the higher topographic positions (i.e. red colors in left side of field 63, Fig 6.2a) where soils are moderately well drained and organic matter contents are lower. The yield patterns also appear to follow the location of tile lines (Fig. 6.2b).

In summary, the maps generated by this research can be utilized for designing experiments, adjusting seeding rates and application rates of fertilizers and other crop inputs, and analyzing field phenotyping experiments. The maps can also be used to guide soil and plant sampling. Close collaboration, however, will be needed between soil scientists, crop scientists, and statisticians in order to utilize the soil and tile line maps to their fullest potential.

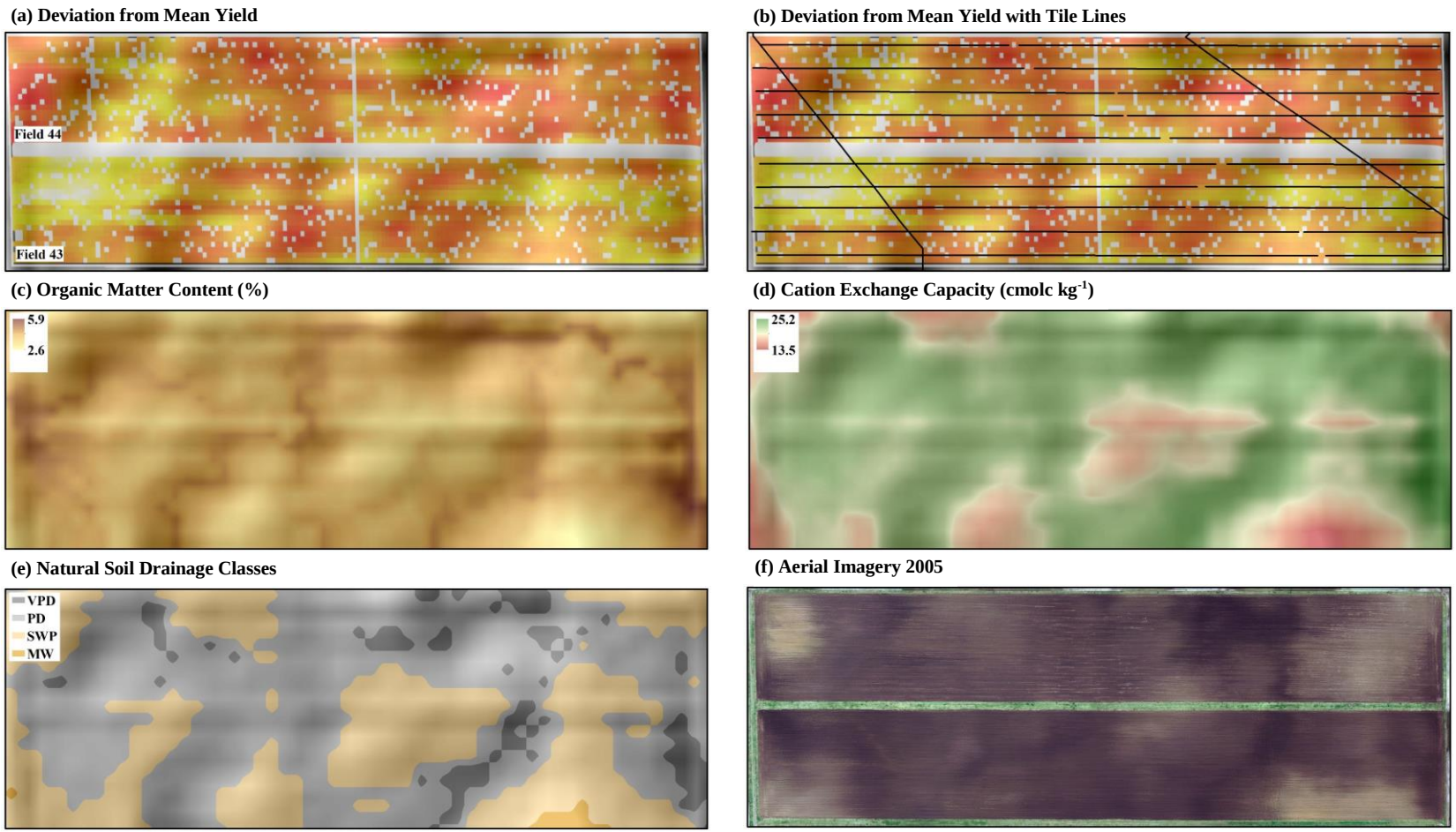

Figure 6.1: Visual comparison between soybean yield and soil maps for fields 43 and 44 at ACRE. (a) Deviation from mean yield of soybean in 2013, yellow colors represent higher yielding areas while red colors represent lower yielding areas, (b) deviation from mean yield with tile lines overlaid, (c) soil organic matter content, (d) cation exchange capacity, (e) soil drainage classes (VPD = very poorly drained, PD = poorly drained, SWP = somewhat poorly drained, and MW = moderately well drained soils), and (f) aerial imagery acquired in 2005. The map of deviation from mean yield was provided by Alencar Xavier and Katherine Rainey, Purdue University. The colored overlays are on top of a hillshade base map that shows where the high and low spots occur in the fields.

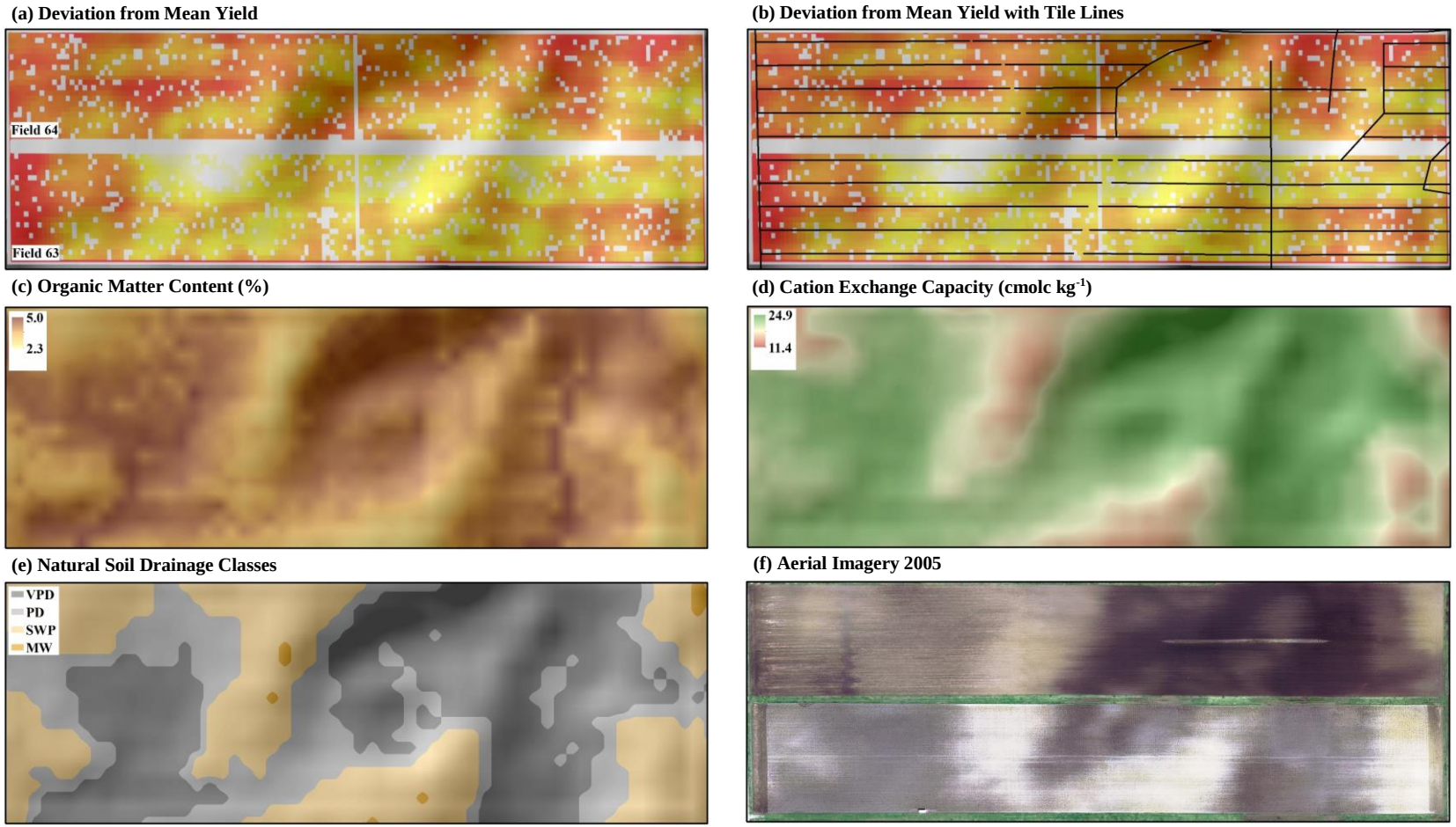

Figure 6.2: Visual comparison between soybean yield and soil maps for fields 63 and 64 at ACRE. (a) Deviation from mean yield of soybean in 2014, yellow colors represent higher yielding areas while red colors represent lower yielding areas, (b) deviation from mean yield with tile lines overlaid, (c) soil organic matter content, (d) cation exchange capacity, (e) soil drainage classes (VPD = very poorly drained, PD = poorly drained, SWP = somewhat poorly drained, and MW = moderately well drained soils), and (f) aerial imagery acquired in 2005. The map of deviation from mean yield was provided by Alencar Xavier and Katherine Rainey, Purdue University. The colored overlays are on top of a hillshade base map that shows where the high and low spots occur in the fields.

6.2 Delivery of Digital Soil Maps

ACRE is a research and teaching facility and there is a high demand for detailed soil maps. The outcomes of this research, along with other essential data and information, will be deposited in the Purdue University Research Repository (PURR). PURR is a research collaboration and data management platform for Purdue University researchers and their collaborators that facilitates publishing and archiving of research data.

Additionally, the maps and their associated products will be made available via the Soil Explorer mobile app (Isee Network, 2015-2020). The Soil Explorer app and website were developed as part of the Integrating Spatial Educational Experiences (Isee) project to leverage big data for teaching and learning. This user-friendly app can deliver soil spatial information on-the-go and works in both online and offline modes.

APPENDIX A. CUBIST MODELS FOR ORGANIC MATTER CONTENT PREDICTION

Model 1:

Rule 1/1: [123 cases, mean 4.25, range 1.9 to 7, est err 0.78]

outcome = 3.19 - 2.36 TPI + 0.105 TWI

Model 2:

Rule 2/1: [113 cases, mean 4.09, range 1.9 to 7, est err 0.70]

if
TWI <= 13.7557
then
outcome = 3.79 + 0.209 MrVBF - 1.17 TPI - 0.106 MrRTF

Rule 2/2: [10 cases, mean 6.11, range 5 to 6.9, est err 1.32]

if
TWI > 13.7557
then
outcome = 7.74 - 0.317 MrRTF + 0.023 MrVBF

Model 3:

Rule 3/1: [123 cases, mean 4.25, range 1.9 to 7, est err 0.80]

outcome = 4.21 - 3.55 TPI

Model 4:

Rule 4/1: [113 cases, mean 4.09, range 1.9 to 7, est err 0.81]

if
TWI <= 13.7557
then
outcome = 3.54 + 0.031 TWI

Rule 4/2: [10 cases, mean 6.11, range 5 to 6.9, est err 1.40]

if

TWI > 13.7557
then
outcome = 7.65

Model 5:

Rule 5/1: [123 cases, mean 4.25, range 1.9 to 7, est err 0.86]

outcome = 4.6 - 5 TPI - 0.152 MrRTF

Model 6:

Rule 6/1: [113 cases, mean 4.09, range 1.9 to 7, est err 0.82]

if
TWI <= 13.7557
then
outcome = 3.85

Rule 6/2: [10 cases, mean 6.11, range 5 to 6.9, est err 1.41]

if
TWI > 13.7557
then
outcome = 2.94 + 0.277 TWI

Model 7:

Rule 7/1: [123 cases, mean 4.25, range 1.9 to 7, est err 0.87]

outcome = 4.59 - 5.24 TPI - 0.146 MrRTF

Model 8:

Rule 8/1: [113 cases, mean 4.09, range 1.9 to 7, est err 0.82]

if
TWI <= 13.7557
then
outcome = 3.85

Rule 8/2: [10 cases, mean 6.11, range 5 to 6.9, est err 1.10]

if
TWI > 13.7557

then
outcome = 0.84 + 0.402 TWI

Model 9:

Rule 9/1: [123 cases, mean 4.25, range 1.9 to 7, est err 0.86]

outcome = 4.15 - 4.45 TPI - 0.159 MrRTF + 0.199 MrVBF

Model 10:

Rule 10/1: [113 cases, mean 4.09, range 1.9 to 7, est err 0.85]

if
TWI <= 13.7557
then
outcome = 2.61 + 0.346 MrVBF + 0.041 TWI

Rule 10/2: [22 cases, mean 4.15, range 2 to 7, est err 1.46]

if
TWI <= 7.0683
then
outcome = 4.78

Rule 10/3: [101 cases, mean 4.28, range 1.9 to 6.9, est err 0.93]

if
TWI > 7.0683
then
outcome = 0.78 + 0.381 TWI + 0.128 MrVBF

Evaluation on training data (123 cases):

Average |error|	0.93
Relative |error|	1.00
Correlation coefficient	0.44

Attribute usage:

Conds	Model	
54%	35%	TWI
	54%	TPI
	37%	MrRTF
	34%	MrVBF

APPENDIX B. CUBIST MODELS FOR CATION EXCHANGE CAPACITY PREDICTION

Model 1:

Rule 1/1: [123 cases, mean 20.12, range 11.1 to 30.1, est err 2.84]

outcome = 18.85 - 9.4 TPI + 1.02 MrVBF

Model 2:

Rule 2/1: [123 cases, mean 20.12, range 11.1 to 30.1, est err 2.84]

outcome = 18.84 - 9.3 TPI + 1.03 MrVBF

Model 3:

Rule 3/1: [123 cases, mean 20.12, range 11.1 to 30.1, est err 2.84]

outcome = 18.85 - 9.4 TPI + 1.02 MrVBF

Model 4:

Rule 4/1: [123 cases, mean 20.12, range 11.1 to 30.1, est err 2.84]

outcome = 18.84 - 9.3 TPI + 1.03 MrVBF

Model 5:

Rule 5/1: [123 cases, mean 20.12, range 11.1 to 30.1, est err 2.84]

outcome = 18.85 - 9.4 TPI + 1.02 MrVBF

Model 6:

Rule 6/1: [123 cases, mean 20.12, range 11.1 to 30.1, est err 2.84]

outcome = 18.84 - 9.3 TPI + 1.03 MrVBF

Model 7:

Rule 7/1: [123 cases, mean 20.12, range 11.1 to 30.1, est err 2.84]

outcome = 18.85 - 9.4 TPI + 1.02 MrVBF

Model 8:

Rule 8/1: [123 cases, mean 20.12, range 11.1 to 30.1, est err 2.84]

outcome = 18.84 - 9.3 TPI + 1.03 MrVBF

Model 9:

Rule 9/1: [123 cases, mean 20.12, range 11.1 to 30.1, est err 2.84]

outcome = 18.85 - 9.4 TPI + 1.02 MrVBF

Model 10:

Rule 10/1: [123 cases, mean 20.12, range 11.1 to 30.1, est err 2.84]

outcome = 18.84 - 9.3 TPI + 1.03 MrVBF

Evaluation on training data (123 cases):

Average |error|	3.25
Relative |error|	0.86
Correlation coefficient	0.51

Attribute usage:

Conds	Model
100%	TPI
100%	MrVBF

REFERENCES

Abdel-Kader, F. H. (2011). Digital soil mapping at pilot sites in the northwest coast of Egypt: A multinomial logistic regression approach. *The Egyptian Journal of Remote Sensing and Space Science*, 14, 29-40. https://doi.org/10.1016/j.ejrs.2011.04.001

ACRE (2020). Agronomy Center for Research and Education (ACRE), Purdue University – College of Agriculture. https://ag.purdue.edu/agry/acre/Pages/default.aspx (accessed 11 March, 2020).

Adamchuk, V. I., & Jasa, P. J. (2002). Precision agriculture: On-the-go vehicle-based soil sensors. University of Nebraska-Lincoln Extension EC02-178. https://cropwatch.unl.edu/documents/On-the-Go%20Vehicle-Based%20Soil%20Sensors%20-%20EC178.pdf (accessed 29 March, 2020).

Adhikari, K., Minasny, B., Greve, M. B., & Greve, M. H. (2014). Constructing a soil class map of Denmark based on the FAO legend using digital techniques. *Geoderma*, 214, 101-113. https://doi.org/10.1016/j.geoderma.2013.09.023

Adhikari, K., Owens, P. R., Ashworth, A. J., Sauer, T. J., Libohova, Z., Richter, J. L., & Miller, D. M. (2018). Topographic controls on soil nutrient variations in a silvopasture system. *Agrosystems, Geosciences & Environment*, 1, 1-15. https://doi.org/10.2134/age2018.04.0008

Ale, S., Naz, B. S., & Bowling, L. C. (2007). Mapping of tile drains in Hoagland watershed for simulating the effects of drainage water management. ASABE Paper No. 072144. St. Joseph, Mich.: ASABE. https://elibrary.asabe.org/pdfviewer.asp?param1=s:/8y9u8/q8qu/tq9q/5tv/L/2y3IGGN/GNIHKK.5tv&param2=I/O/IGIG&param3=HIO.IHG.HII.IGH&param4=23549 (accessed 11 March, 2020).

Allred, B. J., Fausey, N. R., Peters, L., Chen, C. C., Daniels, J. J., & Youn, H. S. (2004). Detection of buried agricultural drainage pipe with geophysical methods. *Applied Engineering in Agriculture*, 20, 307-318. https://doi.org/10.13031/2013.16067

Allred, B., and G. Rouse. 2018. Using drones to find drainage pipes. Ohio Country Journal. 26:27-29. https://www.ocj.com/2018/07/using-drones-to-find-drainage-pipes/ (accessed 05 March, 2020)

Allred, B., Wishart, D., Martinez, L., Schomberg, H., Mirsky, S., Meyers, G., ... Charyton, C. (2018). Delineation of agricultural drainage pipe patterns using ground penetrating radar integrated with a real-time kinematic global navigation satellite system. *Agriculture*, 8, 1-14. https://doi.org/10.3390/agriculture8110167

Andrade, C. (2013). An exploratory study on heads up photo interpretation of aerial photography as a method for mapping drainage tiles. Papers in Resource Analysis at Saint Mary's University of Minnesota, Winona, MN. http://www.gis.smumn.edu/gradprojects/andradec.pdf (accessed 23 March, 2020)

Arrouays, D., Vion, I., & Kicin, J. L. (1995). Spatial analysis and modeling of topsoil carbon storage in temperate forest humic loamy soils of France. *Soil Science*, 159, 191-198.

Avery, T. E., & Burkhart, H.E. (1994). *Forest measurements* (4th ed.). McGraw-Hill, Boston, Massachusetts.

Bad Elf. (2020). Accuracy statement of Bad Elf GNSS Surveyor. https://bad-elf.com/pages/be-gps-3300-detail (accessed 27 April, 2020).

Bazaglia Filho, O., Rizzo, R., Lepsch, I. F., Prado, H. D., Gomes, F. H., Mazza, J. A., & Demattê, J. A. M. (2013). Comparison between detailed digital and conventional soil maps of an area with complex geology. *Revista Brasileira de ciência do solo*, 37, 1136-1148. https://doi.org/10.1590/S0100-06832013000500003

Beaudette, D. E., & O'Geen, A. T. (2009). Soil-Web: An online soil survey for California, Arizona, and Nevada. *Computers & Geosciences*, 35, 2119-2128. https://doi.org/10.1016/j.cageo.2008.10.016

Beck, M. W. (2018). NeuralNetTools: Visualization and analysis tools for neural networks. *Journal of Statistical Software*, 85, 1-20. https://doi.org/10.18637/jss.v085.i11

Behrens, T., Zhu, A. X., Schmidt, K., & Scholten, T. (2010). Multi-scale digital terrain analysis and feature selection for digital soil mapping. *Geoderma*, 155, 175-185. https://doi.org/10.1016/j.geoderma.2009.07.010

Bell, J. C., Cunningham, R. L., & Havens, M. W. (1992). Calibration and validation of a soil-landscape model for predicting soil drainage class. *Soil Science Society of America Journal*, 56, 1860-1866. https://doi.org/10.2136/sssaj1992.03615995005600060035x

Bell, J. C., Cunningham, R. L., & Havens, M. W. (1994). Soil drainage class probability mapping using a soil-landscape model. *Soil Science Society of America Journal*, 58, 464-470. https://doi.org/10.2136/sssaj1994.03615995005800020031x

Beven, K. J., & Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Journal*, 24, 43-69. https://doi.org/10.1080/02626667909491834

Bishop, T. F. A., & McBratney, A. B. (2001). A comparison of prediction methods for creation of field-extent soil property maps. *Geoderma* 103, 149-160. https://doi.org/10.1016/S0016-7061(01)00074-X

Bishop, T. F. A., McBratney, A. B., & Laslett, G. M. (1999). Modelling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma*, 91, 27-45. https://doi.org/10.1016/S0016-7061(99)00003-8

Bock, M., Böhner, J., Conrad, O., Köthe, R., & Ringeler, A. (2007). XV. Methods for creating functional soil databases and applying digital soil mapping with SAGA GIS. JRC Scientific and technical Reports, Office for Official Publications of the European Communities, Luxemburg (pp. 149-163).

Bodaghabadi, M. B., Martínez-Casasnovas, J. A., Salehi, M. H., Mohammadi, J., Borujeni, I. E., Toomanian, N., & Gandomkar, A. (2015). Digital soil mapping using artificial neural networks and terrain-related attributes. *Pedosphere*, 25, 580-591. https://doi.org/10.1016/S1002-0160(15)30038-2

Boehner, J., Koethe, R., Conrad, O., Gross, J., Ringeler, A., & Selige, T. (2002). Soil regionalization by means of terrain analysis and process parameterisation. European Soil Bureau (No. 7, p. 10). Research Report.

Boettinger, J. L. (2010). Environmental covariates for digital soil mapping in the Western USA. In J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, & S. Kienast-Brown (Eds.), *Digital soil mapping: Bridging research, environmental application, and operation* (pp. 17-27). Dordrecht: Springer. https://doi.org/10.1007/978-90-481-8863-5

Bot, A., & J. Benites, J. (2005). The importance of soil organic matter: Key to drought-resistant soil and sustained food production. FAO Soils Bulletin 80. Food and Agriculture Organization of the United Nations, Rome.

Brady, N. C., & Weil, R. R. (2002). *The nature and properties of soils* (13th ed.). Upper Saddle River, New Jersey: Pearson Prentice Hall.

Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32. https://doi.org/10.1023/A:1010933404324

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., & Edwards Jr. T. C. (2015). Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma*, 239, 68-83. https://doi.org/10.1016/j.geoderma.2014.09.019

Brus, D. J. (2019). Sampling for digital soil mapping: A tutorial supported by R scripts. *Geoderma*, 338, 464-480. https://doi.org/10.1016/j.geoderma.2018.07.036

Brus, D. J., De Gruijter, J. J., & Van Groenigen J. W. (2006). Designing spatial coverage samples using the k-means clustering algorithm. *Developments in Soil Science*, 31, 183-192. https://doi.org/10.1016/S0166-2481(06)31014-8

Brus, D. J., Kempen, B., & Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. *European Journal of Soil Science*, 62, 394-407. https://doi.org/10.1111/j.1365-2389.2011.01364.x

Buchanan, B. P., Fleming, M., Schneider, R. L., Richards, B. K., Archibald, J., Qiu, Z., & Walter, M. T. (2014). Evaluating topographic wetness indices across central New York agricultural landscapes. *Hydrology and Earth System Sciences*, 18, 3279-3299. https://doi.org/10.5194/hess-18-3279-2014

Buol, S. W., Hole, F. D., McCracken, R. J., & Southard, R. J. (1997). *Soil genesis and classification* (4th ed.). Ames, Iowa: Iowa State University Press.

Campling, P., Gobin, A., & Feyen, J. (2002). Logistic modeling to spatially predict the probability of soil drainage classes. *Soil Science Society of America Journal*, 66, 1390-1401. https://doi.org/10.2136/sssaj2002.1390

Carré, F., McBratney, A. B., Mayr, T., & Montanarella, L. (2007). Digital soil assessments: Beyond DSM. *Geoderma*, 142, 69-79. https://doi.org/10.1016/j.geoderma.2007.08.015

Castrignano, A., Maiorana, M., Fornaro, F., & Lopez, N. (2002). 3D spatial variability of soil strength and its change over time in a durum wheat field in Southern Italy. *Soil and Tillage Research*, *65*, 95-108. https://doi.org/10.1016/S0167-1987(01)00288-4

Chagas, C. D. S., Vieira, C. A. O., & Fernandes Filho, E. I. (2013). Comparison between artificial neural networks and maximum likelihood classification in digital soil mapping. *Revista Brasileira de Ciência do Solo*, 37, 339-351. https://doi.org/10.1590/S0100-06832013000200005

Chang, L., & Burrough, P. A. (1987). Fuzzy reasoning a new quantitative aid for land evaluation. *Soil Survey and Land Evaluation*, 7, 69-80.

Christensen, R. (1996). *Plane answers to complex questions: The theory of linear models* (2nd ed.). New York: Springer.

Ciaburro, G., & Venkateswaran, B. (2017). *Neural Networks with R: Smart models using CNN, RNN, deep learning, and artificial intelligence principles*. Birmingham, UK: Packt Publishing Ltd.

Cialella, A. T., Dubayah, R., Lawrence, W., & Levine, E. (1997). Predicting soil drainage class using remotely sensed and digital elevation data. *Photogrammetric Engineering and Remote Sensing*, 63, 171-177.

Cole, N. J., & Boettinger, J. L. (2006). Pedogenic understanding raster classification methodology for mapping soils, Powder River Basin, Wyoming, USA. *Developments in Soil Science*, *31*, 377-388. Amsterdam: Elsevier. https://doi.org/10.1016/S0166-2481(06)31027-6

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote sensing of environment*, 37, 35-46. https://doi.org/10.1016/0034-4257(91)90048-B

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., ... Böhner, J. (2015). System for automated geoscientific analyses (SAGA) v. 2.1. 4. *Geoscientific Model Development Discussions*, 8, 2271-2312. https://doi.org/10.5194/gmdd-8-2271-2015

Conyers, L. B. (1997). Ground-penetrating radar. In L. B. Conyers & D. Goodman (Eds.), *Ground-penetrating radar: An introduction for archaeologist* (pp. 131-159). Walnut Creek, CA: AltaMira Press.

Conyers, M. K., Poile, G. J., Oates, A. A., Waters, D., & Chan, K. Y. (2011). Comparison of three carbon determination methods on naturally occurring substrates and the implication for the quantification of 'soil carbon'. *Soil Research*, 49, 27-33. https://doi.org/10.1071/SR10103

De Gruijter, J., Brus, D. J., Bierkens, M. F., & Knotters, M. (2006). *Sampling for natural resource monitoring*. Berlin: Springer Science & Business Media.

De Vos, B., Lettens, S., Muys, B., & Deckers, J. A. (2007). Walkley–Black analysis of forest soil organic carbon: recovery, limitations and uncertainty. *Soil Use and Management*, 23, 221-229. https://doi.org/10.1111/j.1475-2743.2007.00084.x

Dent, D., & Young, A. (1981). *Soil survey and land evaluation*. London: George Allen & Unwin.

Dobos, E., Carré, F., Hengl, T., Reuter, H. I., & Tóth, G. (2006). Digital Soil Mapping as a support to production of functional maps. EUR 22123 EN, Office for Official Publications of the European Communities, Luxemburg.

Draper, N. R., and Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: Wiley.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American statistical association*, 78, 316-331. https://doi.org/10.1080/01621459.1983.10477973

Efron, B., and R.J. Tibshirani. (1994). An introduction to the bootstrap. Boca Raton, Florida: CRC Press.

Ehsani, A. H., & Malekian, A. (2011). Landforms identification using neural network-self organizing map and SRTM data. *Desert*, 16, 111-122.

European Commission. (2002). Communication from the Commission to the Council, the European Parliament, the Economic and Social Committee and the Committee of the Regions. Towards a thematic strategy for soil protection. COM (2002) 179 final.

Fausey, N.R., Doering, E. J., & Palmer, M. L. (1987). Purposes and benefits of drainage. In G. A. Pavelis (Ed.), *Farm drainage in the United States: History, status, and prospects* (pp. 48–51). Economic Research Service (DOA), Washington, D.C., Misc. Pub. No. 1455.

Florinsky, I. V., Eilers, R. G., Manning, G. R., & Fuller, L. G. (2002). Prediction of soil properties by digital terrain modelling. *Environmental Modelling & Software*, 17, 295-311. https://doi.org/10.1016/S1364-8152(01)00067-6

Forkuor, G., Hounkpatin, O. K., Welp, G., & Thiel, M. (2017). High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: A comparison of machine learning and multiple linear regression models. *PloS One*, 12, 1–21. https://doi.org/10.1371/journal.pone.0170478

Franzmeier, D. P., Steinhardt, G. C., Crum, J. R., & Norton, L. D. (1977). Soil characterization in Indiana: I. Field and Laboratory Procedures. Department of Agronomy, Agriculture Experiment Stations, Purdue University. West Lafayette, Indiana with cooperation of the Soil Conservation Service, U.S. Department of Agriculture. Research Bulletin No. 943

Franzmeier, D.P., Kladivko, E. J., & Jenkinson B. J. (2001). Drainage and wet soil management: Wet soils of Indiana. AY– 301. Purdue Extension. https://engineering.purdue.edu/SafeWater/Drainage/AY301.pdf (accessed 23 March, 2020).

Fritsch, S., G. Frauke, and M.N. Wright. 2019. neuralnet: Training of Neural Networks. R package version 1.44.2. https://CRAN.R-project.org/package=neuralnet

Gallant, J. C., & Dowling, T. I. (2003). A multiresolution index of valley bottom flatness for mapping depositional areas. *Water Resources Research*, 39, 4.1-4.13. https://doi.org/10.1029/2002WR001426

Gallant, S. I. (1993). *Neural network learning and expert systems*. Cambridge, Massachusetts: MIT press.

Georgi, C., Spengler, D., Itzerott, S., & Kleinschmit, B. (2018). Automatic delineation algorithm for site-specific management zones based on satellite remote sensing data. *Precision Agriculture*, 19, 684-707. https://doi.org/10.1007/s11119-017-9549-y

Gessler, P. E., Moore, I. D., McKenzie, N. J., & Ryan, P. J. (1995). Soil-landscape modelling and spatial prediction of soil attributes. *International journal of geographical information systems*, 9, 421-432. https://doi.org/10.1080/02693799508902047

Geza, M., & McCray, J. E. (2008). Effects of soil data resolution on SWAT model stream flow and water quality predictions. *Journal of environmental management*, 88, 393-406. https://doi.org/10.1016/j.jenvman.2007.03.016

Ghaderi, A., Shahri, A. A., & Larsson, S. (2019). An artificial neural network based model to predict spatial soil type distribution using piezocone penetration test data (CPTu). *Bulletin of Engineering Geology and the Environment*, 78, 4579-4588. https://doi.org/10.1007/s10064-018-1400-9

Gökkaya, K., Budhathoki, M., Christopher, S. F., Hanrahan, B. R., & Tank, J. L. (2017). Subsurface tile drained area detection using GIS and remote sensing in an agricultural watershed. *Ecological Engineering*, 108, 370-379. https://doi.org/10.1016/j.ecoleng.2017.06.048

Gräler, B., E. Pebesma, and G. Heuvelink. (2016). Spatio-Temporal Interpolation using gstat. *RFID Journal* 8, 204-218.

Grigal, D. F., & Vance, E. D. (2000). Influence of soil organic matter on forest productivity. *New Zealand Journal of Forestry Science*, 30, 169-205.

Grimm, R., Behrens, T., Märker, M., & Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis. *Geoderma*, 146, 102-113. https://doi.org/10.1016/j.geoderma.2008.05.008

Grunwald, S., & Lamsal, S. (2016). The impact of emerging geographic information technology on soil-landscape modeling. In S. Grunwald (Ed.), *Environmental soil-landscape modeling: Geographic information technologies and pedometrics* (pp. 127-154). Boca Raton, Florida, CRC Press.

Grunwald, S., McSweeney, K., Rooney, D. J., & Lowery, B. (2001). Soil layer models created with profile cone penetrometer data. *Geoderma*, 103, 181-201. https://doi.org/10.1016/S0016-7061(01)00076-3

Günther, F., and S. Fritsch. 2010. neuralnet: Training of neural networks. The R journal, 2, 30-38.

Hartemink, A. E., Hempel, J., Lagacherie, P., McBratney, A., McKenzie, N., MacMillan, R. A., ... Walsh, M. (2010). GlobalSoilMap.net–A new digital soil map of the world. In J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, & S. Kienast-Brown (Eds.), *Digital soil mapping: Bridging research, environmental application, and operation* (pp. 423-428). Dordrecht: Springer. https://doi.org/10.1007/978-90-481-8863-5_33

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer Science & Business Media.

Havlin J. L., Tisdale, S. L., Nelson, W. L., & Beaton, J. D. (2014). *Soil fertility and fertilizers: An introduction to nutrient management* (8th ed.). New York: Pearson.

Heim, A., Wehrli, L., Eugster, W., & Schmidt, M. W. I. (2009). Effects of sampling design on the probability to detect soil carbon stock changes at the Swiss CarboEurope site Lägeren. *Geoderma*, 149, 347-354. https://doi.org/10.1016/j.geoderma.2008.12.018

Hengl, T., & MacMillan, R. A. (2019). Predictive soil mapping with R. OpenGeoHub Foundation, Wageningen, The Netherlands, www.soilmapper.org, ISBN: 978-0-359-30635-0.

Hengl, T., Heuvelink, G. B., & Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. *Computers & Geosciences*, 33, 1301-1315. https://doi.org/10.1016/j.cageo.2007.05.001

Hengl, T., Heuvelink, G. B., & Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, 120, 75-93. https://doi.org/10.1016/j.geoderma.2003.08.018

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma*, 265, 62-77. https://doi.org/10.1016/j.geoderma.2015.11.014

Heuvelink, G. B., Brus, D. J., & de Gruijter, J. J. (2006). Optimization of sample configurations for digital mapping of soil properties with universal kriging. *Developments in soil science*, 31, 137-151. https://doi.org/10.1016/S0166-2481(06)31011-2

Hewitson, B. C., & Crane, R. G. (Eds.). (1994). *Neural nets: Applications in geography*. Dordrecht: Kluwer Academic Publishers.

Hosner, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.

Howell, D., Kim, Y., Haydu-Houdeshell, C., Clemmer, P., & Almaraz, R. (2004). Soil property distribution models from point data for soil survey. 24th Annual ESRI International User Conference, August 9–13, 2004. https://proceedings.esri.com/library/userconf/proc04/abstracts/a2018.html (accessed 16 March, 2020).

Hudson, B.D. (1992). The soil survey as paradigm-based science. *Soil Science Society of American Journal*, 56, 836-841. https://doi.org/10.2136/sssaj1992.03615995005600030027x

Indiana Corn and Soybean Innovation Center (ICSIC). (2020). Purdue University – College of Agriculture. https://ag.purdue.edu/icsc/ (accessed 16 March, 2020).

INDOT. (2016). Indiana Geospatial Coordinate System, Ver. 1.05. Indiana Department of Transportation (INDOT), Aerial and Land Survey Office, Indianapolis, IN.

Isee Network (2015 – 2020). Soil Explorer mobile app. Online at http://SoilExplorer.net (accessed 30 April, 2020).

ISUST-GISSRF. 2017. Iowa State University of Science and Technology (ISUST), Geographic Information Systems Support and Research Facility (GISSRF). Tutorial 3: Identifying and mapping tile drainage tutorial for Wright County, Iowa. https://www.iowaview.org/wp-content/uploads/2018/03/Tutorial_3_TileMapping.pdf (accessed 26 Feb, 2020).

Jafari, A., Finke, P. A., Vande Wauw, J., Ayoubi, S., & Khademi, H. (2012). Spatial prediction of USDA-great soil groups in the arid Zarand region, Iran: Comparing logistic regression approaches to predict diagnostic horizons and soil types. *European Journal of Soil Science*, 63, 284-298. https://doi.org/10.1111/j.1365-2389.2012.01425.x

Junjun, Z. H. I., Zhang, G., Renmin, Y. A. N. G., Fei, Y. A. N. G., Chengwei, J. I. N., Feng, L. I. U., ... Decheng, L. I. (2018). An insight into machine learning algorithms to map the occurrence of the soil mattic horizon in the Northeastern Qinghai-Tibetan Plateau. *Pedosphere*, 28, 739-750. https://doi.org/10.1016/S1002-0160(17)60481-8

Kempen, B., Brus, D. J., Heuvelink, G. B., & Stoorvogel, J. J. (2009). Updating the 1: 50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. *Geoderma*, 151, 311-326. https://doi.org/10.1016/j.geoderma.2009.04.023

Kempen, B., Brus, D. J., Stoorvogel, J. J., Heuvelink, G., & de Vries, F. (2012). Efficiency comparison of conventional and digital soil mapping for updating soil maps. *Soil Science Society of America Journal*, 76, 2097-2115. https://doi.org/10.2136/sssaj2011.0424

Kidd, D. B., Malone, B. P., McBratney, A. B., Minasny, B., & Webb, M. A. (2014). Digital mapping of a soil drainage index for irrigated enterprise suitability in Tasmania, Australia. *Soil Research*, 52, 107-119. https://doi.org/10.1071/SR13100

Kidd, D., Malone, B., McBratney, A., Minasny, B., & Webb, M. (2015). Operational sampling challenges to digital soil mapping in Tasmania, Australia. *Geoderma Regional*, 4, 1-10. https://doi.org/10.1016/j.geodrs.2014.11.002

Kravchenko, A. N., Bollero, G. A., Omonode, R. A., & Bullock, D. G. (2002). Quantitative mapping of soil drainage classes using topographical data and soil electrical conductivity. *Soil Science Society of America Journal*, 66, 235-243. https://doi.org/10.2136/sssaj2002.2350

Kuhn, M. & Quinlan, R. (2018). Cubist: Rule and instance based regression modeling. R package version 0.2.2. https://CRAN.R-project.org/package=Cubist (accessed, 23 March, 2020)

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28, 1-26.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.

Kuhn, M., & Quinlan, R. (2018). C50: C5.0 Decision Trees and Rule-Based Models. R package version 0.1.2. https://CRAN.R-project.org/package=C50

Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., & Walter, C. (2014). High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. *Geoderma*, 213, 296-311. https://doi.org/10.1016/j.geoderma.2013.07.002

Lagacherie, P., Cazemier, D. R., van Gaans, P. F., & Burrough, P. A. (1997). Fuzzy k-means clustering of fields in an elementary catchment and extrapolation to a larger area. *Geoderma*, 77, 197-216. https://doi.org/10.1016/S0016-7061(97)00022-0

Lagacherie, P., Legros, J. P., & Burfough, P. A. (1995). A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. *Geoderma*, 65, 283-301. https://doi.org/10.1016/0016-7061(94)00040-H

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.

Lark, R. M., & Cullis, B. R. (2004). Model-based analysis using REML for inference from systematically sampled data on soil. *European Journal of Soil Science*, 55, 799-813. https://doi.org/10.1111/j.1365-2389.2004.00637.x

Lawrence, I., & Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45, 255-268. https://doi.org/10.2307/2532051

Lee, L., Srivastava, P. K., & Petropoulos, G. P. (2017). Overview of sensitivity analysis methods in earth observation modeling. In G. P. Petropoulos & P. K. Srivastava (Eds.), *Sensitivity analysis in earth observation modelling* (pp. 3-24). Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-12-803011-0.00001-X

Lemercier, B., Lacoste, M., Loum, M., & Walter, C. (2012). Extrapolation at regional scale of local soil knowledge using boosted classification trees: A two-step approach. *Geoderma, 171*, 75-84. https://doi.org/10.1016/j.geoderma.2011.03.010

Liaw, A. & Wiener, M. (2002). Classification and regression by randomForest. R News 2, 18-22.

Libohova, Z. (2010). *Terrain attribute soil mapping for predictive continuous soil property maps* (Doctoral dissertation, Purdue University, Order No. 3444689). Available from Dissertations & Theses @ CIC Institutions; ProQuest Dissertations & Theses Global. (859009455). Retrieved from https://search.proquest.com/docview/859009455?accountid=13360.

Libohova, Z., Odgers, N. P., Ashtekar, J., Owens, P. R., Thompson, J. A., & Hempel, J. (2016). Some challenges on quantifying soil property predictions uncertainty for the GlobalSoilMap using legacy data. In G. L. Zhang , D. Brus, F. Liu, X. D. Song, & P. Lagacherie (Eds.), *Digital soil mapping across paradigms, scales and boundaries* (pp. 131-140). Singapore: Springer. https://doi.org/10.1007/978-981-10-0415-5_11

Libohova, Z., Winzeler, E. H., & Owens, P. R. (2010). Developing methods for a terrain attribute derived soil map. *Soil Survey Horizons*, 51, 37-40.

Liu, J., Pattey, E., Nolin, M. C., Miller, J. R., & Ka, O. (2008). Mapping within-field soil drainage using remote sensing, DEM and apparent soil electrical conductivity. *Geoderma*, 143, 261-272. https://doi.org/10.1016/j.geoderma.2007.11.011

Lorenzetti, R., Barbetti, R., Fantappiè, M., L'Abate, G., & Costantini, E. A. (2015). Comparing data mining and deterministic pedology to assess the frequency of WRB reference soil groups in the legend of small scale maps. *Geoderma*, 237, 237-245. https://doi.org/10.1016/j.geoderma.2014.09.006

Luca, C., Si, B. C., & Farrell, R. E. (2007). Upslope length improves spatial estimation of soil organic carbon content. *Canadian journal of soil science*, 87, 291-300. https://doi.org/10.4141/CJSS06012

MacMillan, R. A., Martin, T. C., Earle, T. J., & McNabb, D. H. (2003). Automated analysis and classification of landforms using high-resolution digital elevation data: Applications and issues. *Canadian Journal of Remote Sensing*, 29, 592-606. https://doi.org/10.5589/m03-031

Malone, B. P. (2018). ithir: Soil data and some useful associated functions. R package version 1.0.

Malone, B. P., McBratney, A. B., & Minasny, B. (2018). Description and spatial inference of soil drainage using matrix soil colours in the Lower Hunter Valley, New South Wales, Australia. *PeerJ*, 6, 1-20. https://doi.org/10.7717/peerj.4659

Malone, B. P., McBratney, A. B., Minasny, B., & Laslett, G. M. (2009). Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma*, 154, 138-152. https://doi.org/10.1016/j.geoderma.2009.10.007

Malone, B. P., Minasny, B., & McBratney, A. B. (2017). *Using R for digital soil mapping*. Basel, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-44327-0

Mason, E., & Sulaeman, Y. (2016). Comparison of three models for predicting the spatial distribution of soil organic carbon in Boalemo Regency, Sulawesi. *Jurnal Ilmu Tanah dan Lingkungan*, 18, 42-48. https://doi.org/10.29244/jitl.18.1.42-48

Mathew, E. K., Panda, R. K., & Nair, M. (2001). Influence of subsurface drainage on crop production and soil quality in a low-lying acid sulphate soil. *Agricultural Water Management*, 47, 191-209. https://doi.org/10.1016/S0378-3774(00)00110-4

Maynard, J. J., & Johnson, M. G. (2014). Scale-dependency of LiDAR derived terrain attributes in quantitative soil-landscape modeling: Effects of grid resolution vs. neighborhood extent. *Geoderma*, 230, 29-40. https://doi.org/10.1016/j.geoderma.2014.03.021

McBratney, A. Á., & Pringle, M. J. (1999). Estimating average and proportional variograms of soil properties and their potential use in precision agriculture. *Precision Agriculture*, 1, 125-152. https://doi.org/10.1023/A:1009995404447

McBratney, A. B., & De Gruijter, J. J. (1992). A continuum approach to soil classification by modified fuzzy k-means with extragrades. *Journal of Soil Science*, 43, 159-175. https://doi.org/10.1111/j.1365-2389.1992.tb00127.x

McBratney, A. B., Santos, M. M., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, 117, 3-52. https://doi.org/10.1016/S0016-7061(03)00223-4

McBride, G.B. (2005). A Proposal for Strength-of-Agreement Criteria for Lin's. Concordance Correlation Coefficient. National Institute of Water and Atmospheric Research Ltd, Hamilton, New Zealand. Retrieved from https://www.medcalc.org/download/pdf/McBride2005.pdf (accessed 24 March 2020)

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21, 239-245. https://doi.org/10.1080/00401706.1979.10489755

McKenzie, N. J., & Austin, M. P. (1993). A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. *Geoderma*, 57, 329-355. https://doi.org/10.1016/0016-7061(93)90049-Q

McKenzie, N. J., & Ryan, P. J. (1999). Spatial prediction of soil properties using environmental correlation. *Geoderma*, 89, 67-94. https://doi.org/10.1016/S0016-7061(98)00137-2

Meul, M., & Van Meirvenne, M. (2003). Kriging soil texture under different types of nonstationarity. *Geoderma*, 112, 217-233. https://doi.org/10.1016/S0016-7061(02)00308-7

Miller, B. A., & Schaetzl, R. J. (2014). The historical role of base maps in soil geography. *Geoderma*, 230, 329-339. https://doi.org/10.1016/j.geoderma.2014.04.020

Minasny, B., & McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, 32, 1378-1388. https://doi.org/10.1016/j.cageo.2005.12.009

Minasny, B., & McBratney, A. B. (2008). Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 94, 72-79. https://doi.org/10.1016/j.chemolab.2008.06.003

Minasny, B., McBratney, A. B., Malone, B. P., & Wheeler, I. (2013). Digital mapping of soil carbon. *Advances in Agronomy*, 118, 1-47.

Møller, A. B., Iversen, B. V., Beucher, A., & Greve, M. H. (2019). Prediction of soil drainage classes in Denmark by means of decision tree classification. *Geoderma*, 352, 314-329. https://doi.org/10.1016/j.geoderma.2017.10.015

Molnar, C. (2019). Interpretable machine learning: A guide for making black box models explainable. Christoph Molnar, Leanpub. https://christophm.github.io/interpretable-ml-book/ (accessed, 22 March, 2020).

Moore, I. D., Gessler, P. E., Nielsen, G. A. E., & Peterson, G. A. (1993). Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal*, 57, 443-452. https://doi.org/10.2136/sssaj1993.03615995005700020026x

Moore, I. D., Grayson, R. B., & Ladson, A. R. (1991). Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. *Hydrological Processes*, 5, 3-30. https://doi.org/10.1002/hyp.3360050103

MRCC (Midwestern Regional Climate Center). (2013). *(1981-2010) Normal Annual Snowfall – Midwestern States*. Retrieved from https://mrcc.illinois.edu/ (accessed 6 May, 2020).

Nabiollahi, K., Eskandari, S., Taghizadeh-Mehrjardi, R., Kerry, R., & Triantafilis, J. (2019). Assessing soil organic carbon stocks under land-use change scenarios using random forest models. *Carbon Management*, 10, 63-77. https://doi.org/10.1080/17583004.2018.1553434

Nauman, T. W., & Thompson, J. A. (2014). Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. *Geoderma*, 213, 385-399. https://doi.org/10.1016/j.geoderma.2013.08.024

Nauman, T. W., Thompson, J. A., Odgers, N. P., & Libohova, Z. (2012). Fuzzy disaggregation of conventional soil maps using database knowledge extraction to produce soil property maps. In B. Minasny, B. P. Malone, & A. B. McBratney (Eds.), *Digital soil assessments and beyond* (pp. 203-208). London: Taylor & Francis Group of CRC Press.

Naz, B. S., & Bowling. L. C. (2008). Automated identification of tile lines from remotely sensed data. *Transactions of the ASABE*, 51:1937-1950. https://doi.org/10.13031/2013.25399

Naz, B. S., Ale, S., & Bowling, L. C. (2009). Detecting subsurface drainage systems and estimating drain spacing in intensively managed agricultural landscapes. *Agricultural water management*, 96, 627-637. https://doi.org/10.1016/j.agwat.2008.10.002

Niang, M. A., Nolin, M., Bernier, M., & Perron, I. (2012). Digital mapping of soil drainage classes using multitemporal RADARSAT-1 and ASTER images and soil survey data. *Applied and Environmental Soil Science*, 2012, 1-17. https://doi.org/10.1155/2012/430347

North Central Region (NCR). (1998). Recommended chemical soil test procedure for the North Central Region. North Central Region publication No. 221. Missouri Agriculture Experiment Station.

Northcott, W. J., Verma, A. K., and Cooke, R. A. (2000, July). Mapping subsurface drainage systems using remote sensing and GIS. Paper presented at 2000 ASAE Annual international meeting, Milwaukee, WI.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., ... Papritz, A. J. (2018). Evaluation of digital soil mapping approaches with large sets of environmental covariates. *Soil*, 4, 1-22. https://doi.org/10.5194/soil-4-1-2018

NWS-COOP, 2020. National Weather Service's Cooperative Observer Program (NWS-COOP). https://www.weather.gov/coop/ (accessed 29 January, 2020).

Odeh, I. O., McBratney, A. B., & Chittleborough, D. J. (1995). Further results on prediction of soil properties from terrain attributes: Heterotopic cokriging and regression-kriging. *Geoderma*, 67, 215-226. https://doi.org/10.1016/0016-7061(95)00007-B

Odgers, N. (2017). Digital soil mapping with covariates. http://pierreroudier.github.io/teaching/20171014-DSM-Masterclass-Hamilton/2017-10-09-dsm-with-covariates.html (accessed 22 March, 2020).

Olaya, Vicotr. (2004). A gentle introduction to SAGA GIS. The SAGA User Group eV, Gottingen, Germany.

Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, 178, 389-397. https://doi.org/10.1016/j.ecolmodel.2004.03.013

Palmer, L. J., Scurrah, K. J., Tobin, M., Patel, S. R., Celedon, J. C., Burton, P. R., & Weiss, S. T. (2003). Genome-wide linkage analysis of longitudinal phenotypes using σ^2_A random effects (SSARs) fitted by Gibbs sampling. *BMC Genetics*, 4, 1-5. https://doi.org/10.1186/1471-2156-4-S1-S12

Park, S. J., McSweeney, K., & Lowery, B. (2001). Identification of the spatial distribution of soils using a process-based terrain characterization. *Geoderma*, 103, 249-272. https://doi.org/10.1016/S0016-7061(01)00042-8

Patil, N., Lathi, R., & Chitre, V. (2012). Comparison of C5. 0 & CART classification algorithms using pruning technique. *International Journal of Engineering Research & Technology*, 1, 1-5.

Pavelis, G. A. (1987). Economic survey of farm drainage. In G. A. Pavelis (Ed.), *Farm drainage in the United States: History, status, and prospects* (pp. 110–136). Economic Research Service (DOA), Washington, D.C., Misc. Pub. No. 1455.

Pei, T., Qin, C. Z., Zhu, A. X., Yang, L., Luo, M., Li, B., & Zhou, C. (2010). Mapping soil organic matter using the topographic wetness index: A comparative study based on different flow-direction algorithms and kriging methods. *Ecological Indicators*, 10, 610-619. https://doi.org/10.1016/j.ecolind.2009.10.005

Peng, W., Wheeler, D. B., Bell, J. C., & Krusemark, M. G. (2003). Delineating patterns of soil drainage class on bare soils using remote sensing analyses. *Geoderma*, 115, 261-279. https://doi.org/10.1016/S0016-7061(03)00066-1

Peng, Y., Xiong, X., Adhikari, K., Knadel, M., Grunwald, S., & Greve, M. H. (2015). Modeling soil organic carbon at regional scale by combining multi-spectral images with laboratory spectra. *PloS One*, 10, 1-22. https://doi.org/10.1371/journal.pone.0142295

Pennock, D. J., Zebarth, B. J., & De Jong, E. (1987). Landform classification and soil distribution in hummocky terrain, Saskatchewan, Canada. *Geoderma*, 40, 297-315. https://doi.org/10.1016/0016-7061(87)90040-1

Peters, J., De Baets, B., Verhoest, N. E., Samson, R., Degroeve, S., De Becker, P., & Huybrechts, W. (2007). Random forests as a tool for ecohydrological distribution modelling. Ecological Modelling, 207, 304-318. https://doi.org/10.1016/j.ecolmodel.2007.05.011

Pipaud, I., & Lehmkuhl, F. (2017). Object-based delineation and classification of alluvial fans by application of mean-shift segmentation and support vector machines. *Geomorphology*, 293, 178-200. https://doi.org/10.1016/j.geomorph.2017.05.013

Ponce-Hernandez, R., Marriott, F. H. C., & Beckett, P. H. T. (1986). An improved method for reconstructing a soil profile from analyses of a small number of samples. *Journal of Soil Science*, 37, 455-467. https://doi.org/10.1111/j.1365-2389.1986.tb00377.x

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9, 181-199. https://doi.org/10.1007/s10021-005-0054-1

PU-IPS. 2020. Purdue University – Institute for Plant Sciences (PU-IPS), Field Phenomics. https://ag.purdue.edu/plantsciences/field-phenomics/ (accessed 4 February, 2020).

Quinlan, J. R. (1992). Learning with continuous classes. In A. Adams & L. Sterling (Eds.), *5th Australian Joint Conference on Artificial Intelligence* (pp. 343-348). Hobart, Tasmania.

Quinlan, J. R. (Ed.). (1993). *C4.5: Programs for machine learning*. San Mateo, CA, USA: Morgan Kaufmann.

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Refaeilzadeh, P., Tang, L., & Liu, H. (2009) Cross-Validation. In L. Liu & M. T. ÖZSU (Eds.), *Encyclopedia of database systems* (pp. 532-538). Boston, MA: Springer. https://doi.org/10.1007/978-0-387-39940-9_565

Rhoades, J. D., Chanduvi, F., & Lesch, S. (1999). Soil salinity assessment: Methods and interpretation of electrical conductivity measurements. FAO Irrigation and Drainage Paper 57. Food and Agriculture Organization of the United Nations. http://www.fao.org/3/x2002e/x2002e.pdf (accessed 22 July, 2019).

Ritzema, H. P., Satyanarayana, T. V., Raman, S., & Boonstra, J. (2008). Subsurface drainage to combat waterlogging and salinity in irrigated lands in India: Lessons learned in farmers' fields. *Agricultural Water Management*, 95, 179-189. https://doi.org/10.1016/j.agwat.2007.09.012

Robinson, K. (2013). Purdue Ag. to receive major funding for plant sciences. https://www.purdue.edu/newsroom/releases/2013/Q3/purdue-ag-to-receive-major-funding-for-plant-sciences.html (accessed 17 March, 2020).

Roecker, S. M. (2013). *Solving for y: Digital soil mapping using statistical models and improved models of land surface geometry* (Master's Thesis, West Virginia University, Morgantown, West Virginia). Retrieved from https://researchrepository.wvu.edu/cgi/viewcontent.cgi?article=4628&context=etd (accessed 24 March, 2020).

Roecker, S. M., Howell, D. W., Haydu-Houdeshell, C. A., & Blinn, C. (2010). A qualitative comparison of conventional soil survey and digital soil mapping approaches. In J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, & S. Kienast-Brown (Eds.), *Digital soil mapping: Bridging research, environmental application, and operation* (pp. 369-384). Dordrecht: Springer. https://doi.org/10.1007/978-90-481-8863-5_29

Rossel, R. V., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, 158, 46-54. https://doi.org/10.1016/j.geoderma.2009.12.025

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

Roudier, P. (2011). clhs: a R package for conditioned Latin hypercube sampling.

Roy, S. (2014). *Remote sensing and GIS applications for drainage detection and modeling in agricultural watersheds* (Master's Thesis, Indiana University, Bloomington, Indiana). Retrieved from https://scholarworks.iupui.edu/bitstream/handle/1805/4086/Roy_Thesis_Scholar.pdf?sequence=1&isAllowed=y (accessed 23 March, 2020)

Ruark, M. D., Panuska, J. C., Cooley, E. T., & Pagel, J. (2009). Tile drainage in Wisconsin: Understanding and locating tile drainage systems. University of Wisconsin – Madison. Retrieved from https://fyi.extension.wisc.edu/drainage/files/2012/06/Understanding-and-Locating-Tile-Drain-Systems-Update.pdf (accessed 23 March, 2020).

Rumelhart, D. E., Hinton G. E., & Williams, R. J. (1985). Learning internal representations by error propagation (No. ICS-8506). California Univ San Diego La Jolla Inst for Cognitive Science.

Santra, P., Kumar, M., & Panwar, N. (2017). Digital soil mapping of sand content in arid western India through geostatistical approaches. *Geoderma Regional*, 9, 56-72. https://doi.org/10.1016/j.geodrs.2017.03.003

Schöning, I., Totsche, K. U., & Kögel-Knabner, I. (2006). Small scale spatial variability of organic carbon stocks in litter and solum of a forested Luvisol. *Geoderma*, 136, 631-642. https://doi.org/10.1016/j.geoderma.2006.04.023

Schoonover, J. E., & Crim, J. F. (2015). An introduction to soil concepts and the role of soils in watershed management. *Journal of Contemporary Water Research & Education*, 154, 21-47. https://doi.org/10.1111/j.1936-704X.2015.03186.x

Scull, P., Okin, G., Chadwick, O. A., & Franklin, J. (2005). A comparison of methods to predict soil surface texture in an alluvial basin. *The Professional Geographer*, 57, 423-437. https://doi.org/10.1111/j.0033-0124.2005.00488.x

Sharififar, A., Sarmadian, F., Malone, B. P., & Minasny, B. (2019). Addressing the issue of digital mapping of soil classes with imbalanced class observations. *Geoderma*, 350, 84-92. https://doi.org/10.1016/j.geoderma.2019.05.016

Shi, X., Girod, L., Long, R., DeKett, R., Philippe, J., & Burke, T. (2012). A comparison of LiDAR-based DEMs and USGS-sourced DEMs in terrain analysis for knowledge-based digital soil mapping. *Geoderma*, 170, 217-226. https://doi.org/10.1016/j.geoderma.2011.11.020

Silva, S. H. G., Owens, P. R., Duarte de Menezes, M., Santos, R., Junior, W., & Curi, N. (2014). A technique for low cost soil mapping and validation using expert knowledge on a watershed in Minas Gerais, Brazil. *Soil Science Society of America Journal*, 78, 1310-1319. https://doi.org/10.2136/sssaj2013.09.0382

Smith, C. A. S., Daneshfar, B., Frank, G., Flager, E., & Bulmer, C. (2012). Use of weights of evidence statistics to define inference rules to disaggregate soil survey maps. In B. Minasny, B. P. Malone, & A. B. McBratney (Eds.), *Digital Soil Assessment and Beyond* (pp. 215-220). Leiden, Netherlands, CRC press.

Smith, M. P., Zhu, A. X., Burt, J. E., & Stiles, C. (2006). The effects of DEM resolution and neighborhood size on digital soil survey. *Geoderma*, 137, 58-69. https://doi.org/10.1016/j.geoderma.2006.07.002

Soil Science Division Staff. (2017). Soil survey manual. In C. Ditzler, K. Scheffe, & H. C. Monger (Eds.), *USDA Handbook 18*. Government Printing Office, Washington, D.C.

Soil Survey Staff. 2020. Natural Resources Conservation Service, United States Department of Agriculture. Web Soil Survey. Available online at https://websoilsurvey.nrcs.usda.gov/ (accessed 16 March, 2020).

Soil Survey Staff. 2014. Kellogg Soil Survey Laboratory Methods Manual. Soil Survey Investigations Report No. 42, Version 5.0. R. Burt and Soil Survey Staff (ed.). U.S. Department of Agriculture, Natural Resources Conservation Service.

Sparks, D. L., Page, A. L., Helmke, P. A., Loeppert, R. H., Soltanpour, P. N., Tabatabai, M. A., ... Sumner, M. E. (1996). Methods of soil analysis, part 3. Chemical methods. Madison, WI: SSSA. Inc.

Starr, G. C., Lal, R., Malone, R., Hothem, D., Owens, L., & Kimble, J. (2000). Modeling soil carbon transported by water erosion processes. *Land Degradation & Development*, 11, 83-91. https://doi.org/10.1002/(SICI)1099-145X(200001/02)11:1<83::AID-LDR370>3.0.CO;2-W

Statnikov, A., Wang, L., & Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, 9, 1-10. https://doi.org/10.1186/1471-2105-9-319

Stehman, S. V. (1999). Basic probability sampling designs for thematic map accuracy assessment. *International Journal of remote sensing*, 20, 2423-2441. https://doi.org/10.1080/014311699212100

Stum, A. K., Boettinger, J. L., White, M. A., & Ramsey, R. D. (2010). Random forests applied as a soil spatial predictive model in arid Utah. In J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, & S. Kienast-Brown (Eds.), *Digital soil mapping: Bridging research, environmental application, and operation* (pp. 179-190). Dordrecht: Springer. https://doi.org/10.1007/978-90-481-8863-5_15

Taghizadeh-Mehrjardi, R., Sarmadian, F., Minasny, B., Triantafilis, J., & Omid, M. (2014). Digital mapping of soil classes using decision tree and auxiliary data in the Ardakan region, Iran. *Arid Land Research and Management*, 28, 147-168. https://doi.org/10.1080/15324982.2013.828801

Thomasson, J. A., Sui, R., Cox, M. S., & Al–Rajehy, A. (2001). Soil reflectance sensing for determining soil properties in precision agriculture. *Transactions of the ASAE*, 44, 1445-1453. https://doi.org/10.13031/2013.7002

Thompson, J. (2010). *Identifying subsurface tile drainage systems utilizing remote sensing techniques* (Master's Thesis, The University of Toledo, Toledo, Ohio). Retrieved from https://etd.ohiolink.edu/!etd.send_file?accession=toledo1290141705&disposition=inline (accessed 23 March, 2020).

Thompson, J. A., & Kolka, R. K. (2005). Soil carbon storage estimation in a forested watershed using quantitative soil-landscape modeling. *Soil Science Society of America Journal*, 69, 1086-1093. https://doi.org/10.2136/sssaj2004.0322

Thompson, J. A., Pena-Yewtukhiw, E. M., & Grove, J. H. (2006). Soil–landscape modeling across a physiographic region: Topographic patterns and model transportability. *Geoderma*, 133, 57-70. https://doi.org/10.1016/j.geoderma.2006.03.037

Trangmar, B. B., Yost, R. S., & Uehara, G. (1985). Application of geostatistics to spatial studies of soil properties. *Advances in Agronomy*, 38, 45-94. https://doi.org/10.1016/S0065-2113(08)60673-2

USDA – NRCS. (1998). Soil Survey of Tippecanoe County, Indiana. United States Department of Agriculture, Natural Resources Conservation Service.

Vallentin, C., Dobers, E. S., Itzerott, S., Kleinschmit, B., & Spengler, D. (2019). Delineation of management zones with spatial data fusion and belief theory. *Precision Agriculture*, 1-29. https://doi.org/10.1007/s11119-019-09696-0

Van Groenigen, J. W., Gandah, M., & Bouma, J. (2000). Soil sampling strategies for precision agriculture research under Sahelian conditions. *Soil Science Society of America Journal*, 64, 1674-1680. https://doi.org/10.2136/sssaj2000.6451674x

Varner, B.L., Gress, T., Copenhaver, K., & White, S. (2002). The effectiveness and economic feasibility of image based agricultural tile maps. Inst. of Tech., Champaign, IL. Final Report to NASA ESAD 2002.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. (4th ed.). New York: Springer.

Verma, A. K., Cooke, R. A., & Wendte, L. (1996). Mapping subsurface drainage systems with color infrared aerial photographs: Proceedings of the American Water Resource Association's 32nd Annual Conference and Symposium 'GIS and Water Resources'. Ft. Lauderdale, Florida.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37, 360-363

Walter, C., Lagacherie, P., & Follain, S. (2006). Integrating pedological knowledge into digital soil mapping. *Developments in Soil Science*, *31*, 281-300. https://doi.org/10.1016/S0166-2481(06)31022-7

Waltman, S. W., Olson, C., West, L., Moore, A., & Thompson, J. (2010). Preparing a soil organic carbon inventory for the United States using soil surveys and site measurements: Why carbon stocks at depth are important. In *19th World Congress of Soil Science* (pp. 32-35). Brisbane, Australia.

Walvoort, D. J., Brus, D. J., & De Gruijter, J. J. (2010). An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. *Computers & geosciences*, 36, 1261-1267. https://doi.org/10.1016/j.cageo.2010.04.005

Wang, X., Mosley, C. T., Frankenberger, J. R., & Kladivko, E. J. (2006). Subsurface drain flow and crop yield predictions for different drain spacings using DRAINMOD. *Agricultural Water Management*, 79, 113-136. https://doi.org/10.1016/j.agwat.2005.02.002

Weber, D. D., & Englund, E. J. (1994). Evaluation and comparison of spatial interpolators II. *Mathematical Geology*, 26, 589-603. https://doi.org/10.1007/BF02089243

Webster, R., & Oliver, M. A. (2007). *Geostatistics for environmental scientists* (2nd ed.). Chichester, England: John Wiley & Sons.

Weiss, A. (2001, July). *Topographic position and landforms analysis*. Poster presented at the annual meeting of ESRI user conference, San Diego, CA. Retrieved from http://www.jennessent.com/downloads/tpi-poster-tnc_18x22.pdf (accessed 24 March, 2020).

Wiesmeier, M., Barthold, F., Blank, B., & Kögel-Knabner, I. (2011). Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant Soil*, 340, 7-24. https://doi.org/10.1007/s11104-010-0425-z

Wiesmeier, M., Steffens, M., Kölbl, A., & Kögel-Knabner, I. (2009). Degradation and small-scale spatial homogenization of topsoils in intensively-grazed steppes of Northern China. *Soil and Tillage Research*, 104, 299-310. https://doi.org/10.1016/j.still.2009.04.005

Wilson, J. P., & Gallant, J. C. (Eds.). (2000). *Terrain analysis: Principles and applications*. New York: John Wiley & Sons.

Winzeler, H. E., Owens, P. R., Joern, B. C., Camberato, J. J., Lee, B. D., Anderson, D. E., & Smith, D. R. (2008). Potassium fertility and terrain attributes in a Fragiudalf drainage catena. *Soil Science Society of America Journal*, 72, 1311-1320. https://doi.org/10.2136/sssaj2007.0382

Woo, D. K., Song, H., & Kumar, P. (2019). Mapping subsurface tile drainage systems with thermal images. *Agricultural water management*, 218, 94-101. https://doi.org/10.1016/j.agwat.2019.01.031

Worsham, L., Markewitz, D., Nibbelink, N. P., & West, L. T. (2012). A comparison of three field sampling methods to estimate soil carbon content. *Forest Science*, 58, 513-522. https://doi.org/10.5849/forsci.11-084

Yang, L., Qi, F., Zhu, A., Shi, J., & An, Y. (2016). Evaluation of integrative hierarchical stepwise sampling for digital soil mapping. *Soil Science Society of America Journal*, 80, 637-651. https://doi.org/10.2136/sssaj2015.08.0285

Zell, A., Mamier, G., Vogt, M., Mache, N., Hübner, R., Döring, S., ... & Hatzigeorgiou, A. (1995). SNNS: Stuttgart Neural Network Simulator. User Manual, Version 4.2. Institute for Parallel and Distributed High Performance Systems, Technical Report, (6/95).

Zeraatpisheh, M., Ayoubi, S., Jafari, A., & Finke, P. (2017). Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran. *Geomorphology*, 285, 186-204. https://doi.org/10.1016/j.geomorph.2017.02.015

Zhang, G. L., Feng, L. I. U., & Song, X. D. (2017). Recent progress and future prospect of digital soil mapping: A review. *Journal of Integrative Agriculture*, 16, 2871-2885. https://doi.org/10.1016/S2095-3119(17)61762-3

Zhao, Z., Ashraf, M. I., & Meng, F. R. (2013). Model prediction of soil drainage classes over a large area using a limited number of field samples: A case study in the province of Nova Scotia, Canada. *Canadian Journal of Soil Science*, 93, 73-83. https://doi.org/10.4141/cjss2011-095

Zhao, Z., Chow, T. L., Yang, Q., Rees, H. W., Benoy, G., Xing, Z., & Meng, F. R. (2008). Model prediction of soil drainage classes based on digital elevation model parameters and soil attributes from coarse resolution soil maps. *Canadian Journal of Soil Science*, 88, 787-799. https://doi.org/10.4141/CJSS08012

Zhu, A. X., Band, L. E., Dutton, B., & Nimlos, T. J. (1996). Automated soil inference under fuzzy logic. *Ecological Modelling*, 90, 123-145. https://doi.org/10.1016/0304-3800(95)00161-1

Zhu, A. X., Liu, F., Li, B., Pei, T., Qin, C., Liu, G., ... Zhou, C. (2010). Differentiation of soil conditions over low relief areas using feedback dynamic patterns. *Soil Science Society of America Journal*, 74, 861-869. https://doi.org/10.2136/sssaj2008.0411

VITA

Shams R. Rahmani

EDUCATION

Present	**Purdue University** **PhD candidate, Soil Science** ***Advisor:*** Dr. Darrell G. Schulze ***Dissertation Title:*** Digital Soil Mapping of the Purdue Agronomy Center for Research and Education	West Lafayette, IN
December 2014	**Purdue University** **MS in Soil Science** ***Advisor:*** Dr. Phillip R. Owens ***Thesis Title:*** Creating an Initial Digital Soil Properties Map of Afghanistan	West Lafayette, IN
December 2009	**Kabul University** **BSc in Agronomy** ***Advisors:*** Dr. Abdul Q. Samin & Professor Ab. Ghani Ayubi ***Concentration:*** Soil and Water Science	Kabul, Afghanistan

PROFESSIONAL EXPERIENCE

July 2015 – Present	**Graduate Research Assistant** Department of Agronomy, Purdue University, West Lafayette, IN, USA
Dec. 2010 – Present	**Lecturer** Department of Soil Science and Irrigation, Kabul University, Kabul, Afghanistan. ***Courses:*** Taught Introductory Soil Science, Soil Fertility, and Soil and Water Conservation courses. ***Responsibilities:*** Prepare and conduct indoor labs and field trips; proctor and grade exams, lab reports, field reports, and quizzes; hold office hours.
Jan. 2020 – Present	**Graduate Teaching Assistant** ***Course:*** AGRY 270 – Forests Soils. Department of Agronomy, Purdue University

Responsibilities: Responsible for 12 weekly indoor and 3 outdoor labs for one section of the course (19 students). Duties include brief introduction and review of lecture concepts, help students make connections between the lecture and lab activities, lab preparation, teach lab procedures, supervise lab exercises, check the lab hand-ins, proctor and grade exams and quizzes, hold office hours. Most of this course is the same as Introductory Soil Science (AGRY 255), but with a focus on forest soils.

Aug. – Dec. 2019 **Graduate Teaching Assistant**
Course: AGRY 565 – Soils and Landscapes
Department of Agronomy, Purdue University
Responsibilities: Responsible for assisting instructor with all indoor and outdoor lab activities, including, lab preparation, supervision of lab activities, charging iPads, proctoring and grading exams, lab reports, and field trip reports. Number of students: 16.

Aug. – Dec. 2019 **Graduate Teaching Assistant**
Course: AGRY 255 – Introductory Soil Science
Department of Agronomy, Purdue University
Responsibilities: Same responsibilities as AGRY 270 (Forests Soils). Two discussion sections with 16 students each.

Jan. – May. 2019 **Graduate Teaching Assistant**
Course: AGRY 255 – Introductory Soil Science
Department of Agronomy, Purdue University
Responsibilities: Same responsibilities as AGRY 270 (Forests Soils). Two discussion sections with 16 and 14 students each.

Aug. – Dec. 2017 **Graduate Teaching Assistant**
Course: AGRY 560 – Soil Physics
Department of Agronomy, Purdue University
Responsibilities: Responsible for all indoor and outdoor lab duties, including brief introduction and review of lecture concepts and drawing connections between lecture and lab activities, lab preparation, teaching lab procedures, supervising lab projects, grading lab reports; proctoring exams; holding help sessions and office hours. Number of students: 12.

Aug. – Dec. 2016 **Graduate Teaching Assistant**
Course: AGRY 565 – Soil and Landscape
Department of Agronomy, Purdue University
Responsibilities: See above. Number of students: 16.

Aug. – Dec. 2014 **Graduate Teaching Assistant**
Course: AGRY 255 – Introductory Soil Science

	Department of Agronomy, Purdue University ***Responsibilities:*** See above. Number of students 10.
Aug. – Dec. 2014	**Graduate Research Assistant** Department of Agronomy, Purdue University
Mar. 2011 – Apr. 2012	**Lecturer in Horticulture Department** Afghanistan Technical Vocational Institute (ATVI – USAID) ***Courses:*** Taught Introductory Soil Science, Soil Fertility, and Soil and Water Conservation courses. ***Responsibilities:*** Presented class lectures, prepared and conducted indoor labs; proctored and graded exams, lab reports, and quizzes, and held office hours.

ORAL PAPERS AND POSTER PRESENTATIONS

Rahmani, S.R., J.P. Ackerson, Z. Libohova, D.G. Schulze. 2018. Mapping Soil Organic Matter (OM) and Cation Exchange Capacity (CEC) to Support Plant Phenotyping Research. ASA CSSA SSSA annual meeting, San Diego, California. Five-minute rapid oral plus poster presentations. January 6 – 9, 2018.

Rahmani, S.R., D.G. Schulze. 2018. Mapping Soil Spatial Variability for Site-Specific Management. Health and Disease. Science, Technology, Culture and Policy Research Poster Session. Purdue University wide poster session to promote interdisciplinary collaboration for health and disease research. Purdue University. West Lafayette, Indiana. Poster Presentation. March 1, 2018.

Rahmani, S.R., D.G. Schulze. 2018. Mapping Soil Spatial Variability at the Purdue Agronomy Center for Research and Education (ACRE). Indiana Academy of Science (IAS) Annual meeting, Indianapolis, Indiana. Oral Presentation. March 24, 2018.

Rahmani, S.R., M. Ngunjiri, J.O. Minai, P.R. Owens. D.G. Schulze. 2016. Predicting and Developing Soil Management Zones Based on Topography. Purdue Plant Science Symposium. Purdue University, West Lafayette, Indiana. Oral Presentation. August 4, 2014.

Rahmani, S.R., M. Ngunjiri, P.R. Owens, D.G. Schulze. 2016. Optimal Number of Terrain-Based Clusters for Knowledge-Based Inference Digital Soil Mapping. ASA CSSA SSSA Annual meeting, Phoenix, Arizona. Five-minute rapid oral plus poster presentations. November 6 – 9, 2016.

Rahmani, S.R., M. Ngunjiri, P.R. Owens. 2016. Predicting Spatial Variability of Soil Properties Across the Landscape Using Knowledge Based Inference Mapping Approach. Graduate

Student Welcome and Networking Event at Purdue University, West Lafayette, Indiana. Poster Presentation. September 2, 2016

Rahmani, S.R., M. Ngunjiri, P.R. Owens. 2016. Predicting and Mapping Soil Organic Carbon Using Environmental Covariates. Purdue University Corn Showcase, Beck Agriculture Center, West Lafayette, Indiana. Poster Presentation. July 26, 2016.

Rahmani, S.R., P.R. Owens. J.G. Graveel. 2014. Creating an Initial Digital Soil Map of Afghanistan. ASA CSSA SSSA Annual meeting, Long Beach, California. Oral Presentation. October 30 – November 6, 2014.

ACADEMIC HONORS, AWARDS, AND SCHOLARSHIPS

- Apr. 28, 2020 — Outstanding Teaching Award from Purdue University.
- Jan. 09, 2019 — First place in the Pedology Division and finalist in the society-wide 2019 Soil Science Society of America (SSSA) graduate student competition. San Diego, CA
- Nov. 01, 2018 — Second place in poster presentation at the Purdue GIS Day conference. This was a Purdue University wide conference.
- Mar. 01, 2018 — First place poster presentation at the Purdue Health Disease: Science, Technology, Culture and Policy Research poster session. This was a Purdue University wide poster session.
- Feb. 21, 2018 — Bronze award for poster presentation at the Purdue Chapter of Sigma Xi Scientific Research Society poster symposium. This was a Purdue University wide symposium.
- Feb. 2018 — First place in poster presentation at Purdue Agriculture and Biological Engineering graduate industrial & research symposium. This was a Purdue University Agriculture College wide symposium.
- Apr. 2017 — Featured in the Purdue Graduate Student Ag Research Spotlight https://ag.purdue.edu/arge/Documents/Spotlights/Grad%20Spotlight%20-%20Shams%20Rahmani.pdf
- Aug. 2016 — First place in poster presentation at the Purdue Plant Science Symposium. This was a Purdue University Agriculture College wide symposium.

- Aug. 2016 George D. Scarseth Travel Award to attend the annual meetings of the American Society of Agronomy. Phoenix, AZ
- Mar. 2016 Best poster award at the Purdue Chapter of Sigma Xi Scientific Research Society poster symposium. This was a Purdue University wide symposium.
- Aug. 2014 George D. Scarseth Travel Award to attend the annual meetings of American Society of Agronomy. Long Beach, CA
- Fall, 2013 Semester honors at Purdue University – Fall semester
- Spring, 2012 Dean's list at Purdue University – Spring semester
- Spring, 2012 Semester honors at Purdue University – Spring semester

PROFESSIONAL TRAINING AND CERTIFICATES

- Nov. – Jan., 2019 **Precision Agriculture Certificate.** Purdue University. This was a 12-week long program covering the following topics: Introduction to precision agriculture, global positioning system, differential correction, sensors and remote sensing, soil and water spatial variability, nutrient spatial variability, crop spatial variability, geographic information systems, automation, data analysis, telematics, and economics and adoption
- Sep. 25 – 27, 2018 **Scale Up Conference:** Effective approaches to scaling up agricultural technologies and innovations in the developing world. In this conference, I learned about obstacles in large-scale adoption of new technology and the driving factors of successful scale up.
- Mar. 09, 2018 **Graduate Teaching Certificate (GTC). Purdue University.**
  In order to achieve the GTC certificate, I have met the following criteria:
  - Taught a minimum of two, semester-long Purdue courses,
  - Participated in campus teaching orientation and micro-teaching sessions
  - Completed an additional nine hours of instructional development sessions
  - Utilized early feedback and end of semester evaluations

- Conducted two teaching observations and was observed while teaching
- Completed written self-analysis of the above listed activities.

- Mar. 22, 2018 **Introduction to R for Data Science.** This four-week online course was developed by Purdue University and was delivered through the Future Learn platform. In this course, I learned about managing data with the R platform.

- Mar. 13 – 14, 2017 **Advanced Phenomics Workshop.** Purdue University. In this two-day long workshop, I learned about the role of unmanned aerial vehicles (UAVs) in phenomics, effective ground truthing of plant phenotypes, soil spatial variability and plant phenotyping, and remote sensing for phenotyping.

- Nov. 03 – 05, 2016 **Soil Science Society of America (SSSA) Desert Pedology Tour from Tucson to Phoenix, Arizona.** In this two-day scientific tour, I learned about soil and water relationships in Sonoran Desert landscapes.

- Jun. 05 – 18, 2016 **Borlaug Summer Institute on Global Food Security**. Purdue University. In this two-week program, graduate students from various U.S. and international institutions learned about the challenges surrounding global food security.

- May 16 – 27, 2016 **Applied Management Principles (Mini – MBA)**. Krannert School of Management. Purdue University. This was a two-week long program where I learned about marketing, finance, strategy, negotiation, and problem solving.

- Dec. 2014 **Certificate of achievement for successful completion of the graduate MS program at Purdue University.** SAFF/USAID project.

- Apr. 2014 **Region 3 Colligate Soil Judging Contest**, Missouri. I was a member of Purdue University soil judging team. In this one-week program, we observed several soil practice pits and on the day of contest we competed against other soil judging teams.

- Oct. 30, 2014 – Nov. 01, 2014 **Soil Science Society of America (SSSA) Desert Pedology Tour from Las Vegas, Nevada to Long Beach, California.** In this two-day scientific tour, I learned about desert dust and dune processes, basalt

flows, the V master horizon, aeolian deposits on mountains, bio-crusts, and soil and water relationships in Mojave Desert landscapes.

- Oct. 2013 **Region 3 Colligate Soil Judging Contest,** Stevens Point, Wisconsin. As a member of the Purdue University soil judging team, I participated in this one-week program. After observing several soil practice pits, we then competed against other soil judging teams.

MEMBERSHIP IN PROFESSIONAL, HONORARY, AND SOCIAL SOCITIES

- American Society of Agronomy (ASA)
- Soil Science Society of America (SSSA)
- Crop Science Society of America (CSSA)
- Golden Key International Honor Society
- Soil and Water Conservation Society (SWCS)
- Indiana Academy of Science (IAS)
- Afghan Student Association of Purdue University (ASAP)

SERVICE AND LEADERSHIP

May, 2019 – Dec. 2019 **Member, Purdue GIS Day Conference Planning Committee**
The committee invited speakers, developed the online registration form, handled space reservation, provided poster printing for the presenters, planned lunch and coffee for the participants, and arranged the agenda.
https://www.lib.purdue.edu/gis/gisday/gisday_2019_college_program

Dec. 2018 – Dec. 2019 **Treasurer, Afghan Student Association of Purdue University**
Applied for a Purdue University Student Fee Advisory Board (SFAB) grant and received $15,000 for the Afghan New Year (Nowruz) event. In this event, we provided free Afghan food for more than 300 participants. We also invited a well-known Afghan musician, Mr. Homayoon Sakhi.

Nov. 4 – 7, 2018 **Purdue Agronomy Department hiring table during the Agronomy Society of America (ASA)/Crop Science Society of America (CSSA) meeting in Baltimore, Maryland**.
I greeted visitors and provided answers to their questions. Based on their field of interest, I tried to connect them with the appropriate Agronomy professors.

Jan. 2018 – Aug. 2018	**Member, Purdue Graduate Students Plant Science Symposium Planning Committee.** The symposium was funded by DuPont Pioneer under the Future of Food Security theme. We had several meetings regarding budgeting, inviting speakers and participants, designing the online registration form, providing scholarships to non-Purdue students, reserving and decorating the space, providing for live broadcast of the event, moderating the program, and providing lunch and transportation for the participants.
Dec. 2017 – Dec. 2018	**President, Afghan Student Association of Purdue University** We organized a number of large social events. We applied for a Purdue University Student Organization Allocation Grant (SOGA) and received $7,800 for the Nowruz event. In this event, we provided free Afghan food for 200 participants, Afghan music, live performances, and other cultural activities.
Dec. 2016 – Dec. 2017	**Vice president, Afghan Student Association of Purdue University**
Jan. 2016 – Jan. 2017	**Purdue Agronomy Graduate Students Representative** I was one of 6 graduate student representatives. We were responsible for organizing a number of social events for the Agronomy graduate students throughout the year.

LANGUAGES SKILLS

- English
- Pashto
- Persian

SOFTWARE AND COMPUTER SKILLS

- **GIS, Remote Sensing, Geostatistics tools and software:**

  ArcGIS, QGIS, SAGA, SoLIM, ERDAS IMAGINE, GRASS GIS, RTK-GPS
- **Statistical analysis packages:** R, SAS, SPSS
- **Data mining and machine learning tools:**

  Random Forest, Cubist, Regression, Decision Trees, CART
- **Programming language:** R
- **Microsoft office suite**